

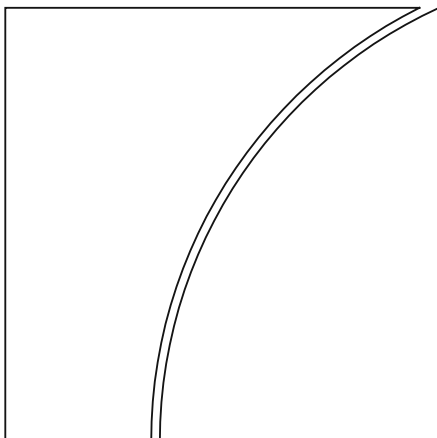
Irving Fisher Committee on Central Bank Statistics

IFC Bulletin

No 59

Data science
in central banking:
applications and tools

October 2023



Irving Fisher Committee on
Central Bank Statistics



Contributions in this volume were prepared for the proceedings of the IFC-Bank of Italy Workshop on “Data science in central banking: applications and tools”, organised by the BIS as a virtual event on 14-17 February 2022.

The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the Bank of Italy, the IFC and its members and the other institutions represented at the meeting.

This publication is available on the BIS website (www.bis.org).

© *Bank for International Settlements 2023. All rights reserved. Brief excerpts may be reproduced or translated provided the source is stated.*

ISSN 1991-7511 (online)

ISBN 978-92-9259-679-8 (online)

Data science in central banking: applications and tools

IFC Bulletin No 59

October 2023

Proceedings of the Irving Fisher Committee on Central Bank Statistics (IFC)-Bank of Italy Workshop on “Data science in central banking: applications and tools”

BIS Basel, virtual event, 14–17 February 2022

Overview

Data science in central banking: applications and tools

Douglas Araujo, Economist, Statistics and Research Support, Monetary and Economic Department (MED), Bank for International Settlements (BIS)

Giuseppe Bruno, Director, Economics and Statistics Directorate, Bank of Italy

Juri Marcucci, Economist, Economics and Statistics Directorate, Bank of Italy

Rafael Schmidt, Head of MED IT, Statistics and Research Support, BIS

Bruno Tissot, Head of Statistics and Research Support, BIS, and Head of IFC Secretariat

Keynote speech

Artificial intelligence in finance – quo vadis

Joerg Osterrieder, Professor of Finance and Artificial Intelligence, ZHAW School of Engineering, Switzerland, University of Twente, Netherlands

1. Modern data architectures and tools

Swimming in the data lake: an application to NielsenIQ Homescan

Minnie H Cui, Gene (Fa Gui) Jiang and Botlhale Mosweu, Bank of Canada

Implementing a multi-tenant big data platform – challenges and approaches taken in the BIS

Hiren Jani and Anand Kannan, BIS

Modern computing platforms as key technology for central banks, financial supervisors, and regulators

John Ashley and Jochen Papenbrock, NVIDIA

The data analytics lab – from innovation to products

Mona Amer, Hiren Jani and Mathieu Le Cam, BIS

Containerisation for research collaboration: platform-independent economics

Venkat Balasubramanian and Kim Huynh, Bank of Canada;

Danielle Handel, Stanford University; Anson Ho, Ted Rogers School of Management

David Jacho-Chávez and Carson Rea, Emory University

Banco de Portugal data centric-strategy for the adoption of a modern data architecture

Caio Costa, Guilherme de Sousa and Hugo Matos, Banco de Portugal

Microdata utility – data loading with automated data parsing and data structure creation

Marcus Jellinghaus, BIS

Using non-traditional point of interest data as merchant survey sample frames

Angelika Welte and Joy Wu, Bank of Canada, and Marcel Voia, University of Orléans

gingado: a machine learning library focused on economics and finance

Douglas Araujo, BIS

A multi-layer dynamic network for significant European banking groups

Annalaura Ianaro and Joerg Reddig, European Central Bank (ECB)

EU single resolution board system for data collection, transformation and analysis

Michał Piechocki, Karol Minczyński and Marta Kuczyńska, Business Reporting - Advisory Group (BR-AG)

2. ML/Big data analytics supporting central bank activities

A network analysis of the JGB repo market

Yasufumi Gemma, Takumi Horikawa and Yujiro Matsui, Bank of Japan

Data science and statistics: a network analysis to understand the foreign investment

João Falcão Silva, Banco de Portugal, Flávio Pinheiro and Bojan Stavrik, NOVA Information Management School

Cross-currency swap market through the lens of OTC derivative transaction data – impact of Covid-19 and subsequent recovery

Kazuaki Washimi and Rinto Maruyama, Bank of Japan

Changes in the lending activity of banks in Poland, including the portfolio of non-financial corporate loans

Aneta Kosztowniak, Narodowy Bank Polski

Is mobile money part of money? Understanding the trends and measurement / Evaluating mobile money access and use with non-traditional data sources

Kazuko Shirono, Bidisha Das, Yingjie Fan, Esha Chhabra and Hector Carcel-Villanova, International Monetary Fund

3. Dealing with textual information

Sentiment analysis of tourist reviews from online travel forum for improving Indonesia tourism sector

Muhammad Abdul Jabbar, Arinda Dwi Okfantia, Anggraini Widjanarti and Alvin Andhika Zulen, Bank Indonesia

Mailbot – optimising the process of answering statistical queries

Daphne Aurouet, Nina Blatnik, Samo Boh, Andrea Colombo, Almir Delic, Jordi Gutiérrez, Gavril Petrov and Kristine Rikova, ECB

Central bank communication: what can a machine tell us about the art of communication? One size does not fit all

Joan Huang and John Simon, Reserve Bank of Australia

News and banks' equities: do words have predictive power?

Valerio Astuti, Giuseppe Bruno, Sabina Marchetti and Juri Marcucci, Bank of Italy

Natural language processing for risk management

Bijan Sahamie, Deutsche Bundesbank

Integrating natural language processing technologies to central bank operations at Bank of Thailand

Jiradett Kerd Sri and Pucktada Treeratpituk, Bank of Thailand

Machine learning for measuring central bank policy credibility and communication from news

Muhammad Abdul Jabbar, Okiriza Wibisono, Anggraini Widjanarti and Alvin Andhika Zulen, Bank Indonesia

A machine learning approach to narrative retrieval in economic news: the case of oil price uncertainty

Donald Jay Bertulfo, Delft University of Technology

Creation of a structured sustainability database from company reports – a web application prototype for information retrieval and storage

Eugenia Koblents and Alejandro Morales, Bank of Spain

Measuring text-based sentiments from monetary policy statements – a Malaysian case study using natural language processing

Eilyn Chong and Sui-Jade Ho, Central Bank of Malaysia

4. Data visualisation and quality assurance processes

Interactive visualisation tool: outlier detection in large multi-dimensional datasets

Christoph Leitner, Thomas Kemetmueller and Philipp Reisinger, National Bank of the Republic of Austria

Spot the flaw – using power BI for quality control: an application to non-financial corporations' data

José Alexandre Neves, Tiago Pinho Pereira and Ana Bárbara Pinto, Banco de Portugal

Leveraging the power of visualisation for data exploration and insights communication – visual analytics with Tableau

Zunaira Rasheed, BIS

Defining business transformation rules in a standardised format – a practical case

Daniela Arru and Giulia Oddone, ECB

Anomaly intersection: disentangling data quality and financial stability developments in a scalable way

Gemma Agostoni, ECB, Louis de Charsonville, McKinsey & Company, Marco D'Errico, ECB, ESRB Secretariat, Cristina Leonte, BIS, Grzegorz Skrzypczynski, ECB

Introducing explainable supervised machine learning into interactive feedback loops for statistical production systems

*Thomas Gottron, Georgios Kanellos and Johannes Micheler, ECB
José Martínez, Solenix, Carlos Mougan, University of Southampton*

Stacking machine learning models for anomaly detection: comparing AnaCredit to other banking datasets

Davide Nicola Continanza, Andrea del Monaco, Marco di Lucido, Daniele Figoli, Pasquale Maddaloni, Filippo Quarta and Giuseppe Turturiello, Bank of Italy

5. Nowcasting and modelling the economy

Nowcasting business and financial cycles using low- and high-frequency data

Alberto Americo, Frederik Hering and Rukmani Vaithianathan, BIS

Nowcasting economic activity with mobility data

Koji Takahashi, BIS, Kohei Matsumura, Yusuke Oh and Tomohiro Sugo, Bank of Japan

Extracting economic sentiment from news articles: the case of Korea

Younghwan Lee and Beomseok Seo, Bank of Korea

Big data analytics on real-time gross settlement data for tracking corporate activity

Mohammad Khoirul Hidayat, Amin Endah Sulistiawati, and Alvin Andhika Zulen, Bank Indonesia

Deep vector autoregression for macroeconomic data

Marc Agustí and Ignacio Vidal-Quadras Costa, ECB; Patrick Altmeyer, Delft University of Technology

What should be the optimal financial structure of the FDI inflows to Poland in stimulating growth processes?

Aneta Kosztowniak, Narodowy Bank Polski

Big data analytics on payment system data for measuring household consumption in Indonesia

Muhammad Abdul Jabbar, Mohammad Khoirul Hidayat and Alvin Andhika Zulen, Bank Indonesia

Tracking the economy during the Covid-19 pandemic: the contribution of high frequency indicators

Jérôme Coffinet, Jean-Brieux Delbos, Jean-Noël Kien, Étienne Kintzler, Ariane Lestrade, Michel Mouliom, Théo Nicolas and and Wojtech Kaiser, Bank of France

Data science in central banking: applications and tools

Douglas Araujo, Giuseppe Bruno, Juri Marcucci, Rafael Schmidt, Bruno Tissot¹

Executive summary

The Irving Fisher Committee on Central Bank Statistics (IFC) periodically organises workshops on “Data science in central banking” with a diverse audience of practitioners and technicians. The most recent one took place in 2022 and focused on **the broad spectrum of data science applications/tools used in central banks**.

The concept of data science refers to the study of data and therefore includes the various techniques for extracting insights from them. **Yet data science is fundamentally different from traditional data analysis, as it typically applies to large, complex and/or unstructured information sets.**

A key factor supporting the development of central banks’ data science projects in recent years has been **the sheer volume and complexity of financial data available** in today’s societies. This requires more sophisticated techniques for data management and analysis, a trend reinforced by the new opportunities opened up by artificial intelligence (AI) and machine learning (ML). Another factor has been the greater focus on real-time, evidence-based policymaking, which requires authorities to rely on better analytical and forecasting capacities to support their decisions. Additionally, advances in statistical computing infrastructure and enhanced training have enhanced the data skills of official sector staff.

These elements have accelerated efforts to advance data science, helping central banks to quickly adapt to the swiftly evolving financial landscape. In this endeavour, the **role of data scientists lies at the intersection of three areas: information technology (IT); mathematical and statistical methods; and business, or “subject-matter” expertise.**

From the outset, IT has absorbed a great deal of attention and resources. Central banks are increasingly aware that a modern IT architecture is crucial in reliably

¹ Respectively, Economist, Monetary and Economic Department, Bank for International Settlements (BIS) (Douglas.Araujo@bis.org); Director, Economics and Statistics Directorate, Bank of Italy (Giuseppe.Bruno@bancaditalia.it); Economist, Economics and Statistics Directorate, Bank of Italy (Juri.Marcucci@bancaditalia.it); Head of IT, Monetary and Economics Department, BIS (Rafael.Schmidt@bis.org); and Head of Statistics and Research Support, BIS, and Head of the Secretariat of the Irving Fisher Committee on Central Bank Statistics (IFC) (Bruno.Tissot@bis.org).

The views expressed here are those of the authors and do not necessarily reflect those of the Bank of Italy, the BIS, the IFC or any of the institutions represented at the workshop.

We thank Olivier Sirello for helpful comments and suggestions.

and securely dealing with data. A key objective is to facilitate access to a large and diverse range of sources as well as relevant IT software and tools in a user-friendly way. But implementing such an IT architecture can be challenging, calling as it does for careful implementation, clear governance frameworks, and the application of common standards. The emphasis is on adopting advanced IT tools and engineering practices – including cloud computing, software containers, automation tools, and continuous software integration and delivery pipelines. In particular, there has been a growing interest among central banks on using and producing software that can be shared as open source, either with their peers or with the general public. Such an open source software (OSS) strategy can be instrumental for honing their own IT development and strengthening their data science capabilities.

Once the IT infrastructure is able to support the development and deployment of data science applications, **the focus is on performing the various mathematical and statistical operations that are needed to deal with the raw data**. Data scientists need not only to access very large and complex information sets, but also to compile statistics via multiple sequential tasks (signal extraction, quality management, dissemination) before using them to extract relevant insights. Many different AI techniques can be used for this purpose, including for conducting textual analysis, reflecting the increasing opportunities offered by natural language processing (NLP) tools and large language models (LLMs).

A third lesson is that data science projects require a good understanding of the business cases and therefore a close cooperation with subject-matter experts. One obvious reason is that economic indicators such as GDP are more than just numbers: analysing them calls for an understanding of the way the statistics have been compiled as well as the complex factors that drive them – say, fiscal policy or geopolitical tensions – and their real-world implications. Moreover, this expertise is essential to support informed policy decisions: in particular, translating data insights into actionable recommendations for central banks cannot be communicated as a “black box” and requires transparent explanations to the various stakeholders involved – from other authorities to the general public. This is even more important with data science applications that may need to be adapted when used to answer economic questions, for instance when analysing causal relationships. Finally, business area expertise can help central banks prioritise effectively between concurrent data science initiatives especially in view of resources constraints.

1. Introduction: data science to support central bank operations

The scope of data science

Central banks have been actively reviewing the ongoing adoption of data science as well as developments in the big data ecosystem in recent years (IFC (2017)). A number of projects have been launched on a pilot basis, of which some have already become permanent processes supporting current operations. In this context, central banks have realised the importance of sharing experience, not least to develop in-house knowledge and reduce reliance on external services providers.

To support these initiatives, the IFC has organised recurrent workshops on “Data science in central banking” with a broad audience of practitioners and technicians.² The most recent one, organised on a virtual basis with the Bank of Italy at the BIS in 2022, focused on the **broad spectrum of data science applications/tools used in central banks**. Several hundred participants representing more than 120 institutions took this opportunity to present and discuss novel data science applications of particular relevance that are under development or already in production.

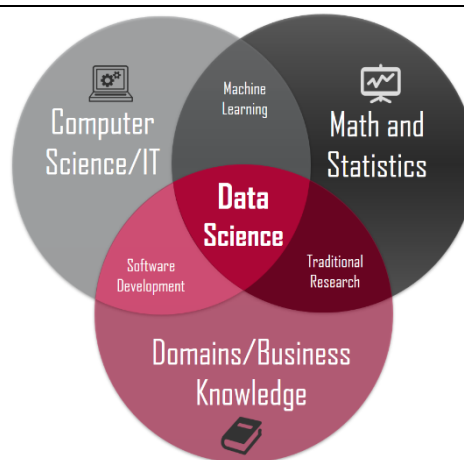
But what is data science? The concept **refers to the study of data, and therefore includes the various techniques for extracting insights from them**. Yet in practice its scope is elusive and continuously evolving. As a starting point, data science should cover all the features related to *data analysis*, including the “procedures for analysing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data [...], and all the machinery and results [...] which apply to analysing data” (Tukey (1962)). From this perspective, it requires the performance of multiple tasks, such as data-gathering, preparation and exploration; data representation and transformation; computing with data; visualisation and presentation; data modelling; and using data to study science itself (Donoho (2017)). The goal of data science is therefore to turn *data analysis* into *actionable knowledge*, whether that means better decision-making, identifying new patterns and trends, optimising processes, or even creating new methods of data-driven research.

Data science is, however, fundamentally different from traditional analysis as it typically applies to large or complex data, including different types of unstructured information such as text (Bholat (2020)). In fact, as argued by Jörg Osterrieder in his keynote address, the AI “revolution” may not be as new as people usually think.³ But what is new is the current combination of greater computing power, the increased availability of big data sources and advanced analytical techniques, and the multiple business applications of interest for which these data and tools can bring useful insights.

These considerations mean that the range of methods potentially covered by data science is considerable and requires diverse fields of expertise. As a result, an effective approach to data science should encompass the large variety of relevant business cases as well as multiple types of data. Data science could thus be usefully defined as an “*interdisciplinary field* that uses scientific method, processes, algorithms and systems to extract knowledge and insights from data in various forms” (Walczak (2019)). In view of these interrelationships, the multifaceted role of data scientists would appear to sit at the intersection of **information technology; mathematical and statistical methods; and business, or “subject-matter” expertise** (Graph 1).

² The material from the last workshops can be found at www.bis.org/ifc/events.htm.

³ The term was first coined in the academia during the 1950s (Moor (2006)).



Source: N Cheng, "Data analyst primer: the essential guide", Medium, 27 Aug 2020, after D Conway, "The data science Venn diagram", 30 September 2010.

Data science in central banking

The approach to data science outlined above would seem to be well suited for central banks, whose activities range from statistical production to economic analysis, monitoring and policymaking. Recent projects have in fact involved the three dimensions at the intersection of data science. First, significant investment in software engineering tools, especially to support big data analytics, has improved the underlying IT and data infrastructure. Second, more diverse and mature econometric and data visualisation tools have been deployed to leverage mathematical/statistical knowledge and extract more actionable and timely insights from data. And third, deep subject-matter expertise in central banks (eg on economic and financial information) has ensured that the business needs and related priorities are fully understood.

A number of factors have supported these developments. First is the **skill set that is increasingly available internally**. Reflecting the variety of the tasks they perform, central banks have at their disposal a large variety of staff competencies, a key condition for ensuring the involvement of multidisciplinary teams and making the most of the new data landscape.⁴ Moreover, while central banks have traditionally had a large number of statisticians and subject-matter experts in their ranks, their staff have become increasingly proficient in IT tools and techniques in recent years. The upskilling of internal staff, combined with fresh hires with new IT abilities, has helped the adoption of best practices from computer science fields, in particular software engineering. In addition, the methodology of prototyping tools in business areas, experimenting quickly and then developing them over time as their usage

⁴ See the Banco de Portugal's aim to have specialised staff, represented by different "colours", comprising a blend of business and technology skills and working together in composite groups of "purple people", ie specialised staff whose skills overlap (Teles Dias (2021)).

matures has proved a pragmatic approach to further increasing data science usage in existing teams (Araujo et al (2022)).

A second supportive factor has been the **decisive action of central banks to enhance the underlying global statistical infrastructure** to make it more flexible and efficient, for instance by developing global registers and identifiers and by promoting data-sharing and access to new sources of information. One particularly important dimension in supporting data-based applications in this context has been the development of global initiatives to standardise statistical information, eg the Statistical Data and Metadata eXchange (SDMX; IFC (2016)) or the Legal Entity Identifier (LEI). Such standardisation initiatives provide a common language to deal with the data of interest, on top of which more comprehensive applications can be developed. And, in fact, they can be instrumental for setting up advanced data architectures. For instance, the SDMX information model allows organisations to automate data processing and collection, based on a metadata-driven approach that strengthens their data quality management – a valuable benefit for central banks that are key producers of economic and financial statistics.

A third factor has been the **general push for developing innovation in central banking** to better address the rapidly changing financial landscape, as observed for instance by the creation of the BIS Innovation Hub.⁵ This research initiative aims to develop public goods in the technology space to support central banks and improve the functioning of the financial system. These goals are realised in the form of specific projects, for instance Projects Rio and Ellipse, which, respectively, facilitate market monitoring and the development of early warning indicators based on big data analytics. More generally, innovation has proved essential to support central banks not only as statistical producers but also as users of the data needed for conducting their activities, depending on their mandates and specific requirements.

Drawing on the various central bank contributions included in this IFC Bulletin, this overview sheds light on the **three main dimensions at the intersection of data science**. Section 2 reviews the IT infrastructure elements that can support data science projects in central banks, such as the setup of a data platform, the provision of the software and tools for data management, and the new opportunities offered by modern IT tools and engineering practices, including cloud computing and the use of software containers. Section 3 discusses the role played by mathematical and statistical techniques and access to new information sources that allow central banks to extract insights from more data, including unstructured ones eg text. Lastly, Section 4 emphasises the importance of leveraging subject-matter expertise to support data science use cases in the economic and financial sphere.

2. Modern IT architectures

The first main area at the intersection of the multidisciplinary data science concept relates to IT. Central banks, as well as other public institutions and international organisations, are increasingly aware of the **importance of having a modern IT architecture** to deal with the data they need to fulfil their missions. This reflects the rapid pace of technological advance and the mission-critical nature of many of their data-intensive processes.

⁵ For more on the BIS Innovation Hub's projects, see www.bis.org/about/bisih/projects.htm.

The IT architecture has to cover the **various processes, software and hardware that underpin the multiple activities when dealing with data** – eg sourcing, collecting, storing, managing, analysing, sharing and making data available to the appropriate users. These “plumbing elements” have to be linked together in a way that is comprehensive and consistent with the organisational priorities that guide day-to-day activities and investments (OECD (2019)). Ideally, this integration would be organised in the context of a dedicated data governance framework supporting the institution’s overall strategy (Križman and Tissot (2022)). Yet the implementation of new IT platforms and software stacks can be complex, requiring considerable investment in terms of resources and time.

While no single solution can fit all organisations and use cases, **a number of useful insights can be derived from the recent projects implemented in the central banking community**. In particular, three points deserve to be highlighted. First, a modern data platform is essential for central banks that want to fully tap the potential from the information available. This means for example that, to be efficiently organised, data should be put in a common place available to all potential users subject to access rules. Second, the IT environment should provide all the IT tools and processes needed to manage data consistently and efficiently, and in a user-friendly way. Third, there is a premium on keeping this infrastructure up to date to allow the continuous deployment of novel applications and tools, which usually require a large quantity of information to be reliably sourced and delivered. One way to deal with this issue is to move some IT operations to a cloud environment, an option that is increasingly under consideration in the central banking community.

A common data platform?

A first lesson is that there can be **value in implementing a unique and powerful platform** to manage all the different types of information (including big data sets) of interest to the organisation. Such multi-tenant data platforms, which may be centralised (enterprise data lake) or decentralised (data mesh architecture), can bring several benefits, especially in terms of the ease of governance and economies of scope and scale offered by a common infrastructure. They can be instrumental when dealing with multiple business areas, as is often the case for financial stability (macroprudential) analyses, which typically draw on multiple information sources. It thus puts a premium on developing an institution-wide, unified data model – comprising for instance a set of common identifiers for financial institutions and instruments, consistent definitions and a comprehensive metadata schema.

A second lesson is that the **platform should be flexible enough to manage all the various types of data** of interest. For instance, a key requirement for the system for data collection, transformation and analyses implemented at the European Union’s Single Resolution Board (ESRB) was to be able to deal with both large, highly structured data (eg granular supervisory reporting from financial institutions) and unstructured information collected from narrative texts. This reflects the need to use advanced analytical tools for checking banks’ crisis resolution plans (ResTech; Loiacono and Rulli (2022)). Given the heterogeneity of the information at stake, quality was ensured by designing the data ecosystem on unified standards such as the LEI and the eXtensible Business Reporting Language (XBRL) global framework for exchanging business information. Similarly, the various processes followed by the BIS for the collection, production and dissemination of its own statistics – eg for

validating, transforming and mapping different types of macro as well as micro data sets – are based on a (SDMX) metadata-driven environment.

Experience shows that there are significant **challenges** when implementing an institutional data platform, namely the difficulty of dealing with diverse technological landscape across business units; their varied requirements, for instance as regards the trade-offs faced in terms of agility/robustness or security/innovation; and different organisational setups – with the key issue of adequately balancing the centralised governance of the platform against the provision of sufficient user freedom to design their (evolving) IT requirements.

A third lesson is to **offer sufficient IT self-service capacities to the business units**, so that they are adequately empowered with the resources to manage their data projects on their own rather through a centralised point. In fact, the development of the BIS centralised multi-tenant platform was accompanied by the creation of an analytical lab to facilitate the launch and maintenance of the various data-based initiatives. Certainly, one issue was the time and close collaboration required by the upskilling of the staff located in different units. But the implementation of self-service capacities proved to be a catalyst for innovation, stimulating the efficient design and delivery of the business areas' projects. Self-service was also a key element of the Banco de Portugal's information strategy to become a more data-driven institution. Under its advanced analytics pillar, a data science lab was set up to offer to the various units tools for code versioning, dedicated storage, grid computing and multiple coding languages used in data science and econometrics. These facilities let statisticians develop automatic quality control checks based on traditional techniques as well as ML approaches to identify erroneous reports in the credit registry. The central bank's strategy was to promote the automation and standardisation of the new projects to follow good practices in software development.

Making available efficient IT tools and processes

Modern IT architectures offer two key benefits to central banks. The first is to provide access to better-quality data, delivered more promptly to staff with the appropriate permissions to use them. The second is to **make available to each business unit the necessary tools to deal with this information**. The obvious reason is that data science work requires the use of advanced analytical tools and the necessary software so that users can (i) deal with a wide range of data types; (ii) have access to large spectrum of functionalities; and (iii) rely on efficient operational processes to save time and resources (Wibisono et al (2019)).

As regards the first aspect, the aim of a modern IT platform is to **facilitate users' access to a large and diverse range of data types**, ie structured or unstructured, coming from different sources (eg commercial vendors, reporting entities) and with different formats (eg generic spreadsheets "pushed" by reporters, data "pulled" from websites). Certainly, such diversity can raise practical difficulties, which are reinforced by the increasing demand for very granular information observed in recent years. In addition, the conditions for sharing/accessing these new data sets may require important legal and security work, for instance to deal with confidentiality and licensing issues.

With respect to IT aspects more specifically, the continuous **access to new information sources calls for setting up agile, structured and scalable data ingestion processes**. For instance, the BIS has developed a microdata ingestion utility

that allows the structured loading of millions of data points from a variety of commercial and public data providers. This utility usually requires minimal configuration to import a new data set, ensuring a fast turnaround for users. It has been in production since 2020, leveraging open source tools such as Python/Pandas and SQLAlchemy to download, parse and store data (files) on an SQL server. This tool and underlying software are available for interested central banks as part of the BIS's promotion of international software collaboration. Similarly, the ECB has developed a framework, based on Apache Spark distributed computing and other open source tools, to integrate granular banking data from multiple sources such as the European credit registry (Anacredit), the European Market Infrastructure Regulation (EMIR), money market statistical reporting (MMSR), the Central Securities Database (CSDB), and the securities holdings statistics (SHS). The aim is to analyse in a comprehensive and timely way the systemic importance of European banking groups by simultaneously considering different financial instruments and the related interconnections.

Second, making available more diverse and well managed information is not enough since business area users need sufficient capabilities to deal with it. This calls for **having a sufficiently varied palette of IT software, advanced analytical tools and accessible programming languages**. To address this demand, the free open-source library *gingado* has been created at the BIS to facilitate the internal use of ML in economic and finance use cases (Araujo (2023)). This package uses the SDMX standard to help users find and obtain high-quality data to augment their particular data set; provides convenient functions that train benchmark ML models; and promotes proper modelling documentation. And because this library is written in Python, it can also be easily used in other programming environments based in R, Stata and others. Relatedly, central banks' analytical capabilities can be improved more generally by both the use and production of OSS (see Box 1).

Box 1

Central banks as users and providers of open-source software

Douglas Araujo, Stratos Nikoloutsos, Rafael Schmidt, Olivier Sirello

In a world increasingly shaped by collaborative creativity, central banks are engaging more intensively with open-source software (OSS), ie software whose source code can be freely edited, reproduced and redistributed.^① As software users, they may extensively rely on OSS, including ML tools, to perform their operations.^② As providers, they can share publicly or with other central banks their source codes, such as macroeconomic models or tools for data dissemination. This box discusses how central banks can exploit the value added of OSS.

As a starting point, OSS can be defined against its opposite, "closed-source" or proprietary software whose source code is not disclosed. Private companies traditionally choose closed-source models, which can be the most straightforward option, especially if the source code they develop is a trade secret. Yet many firms, including big techs, have been increasingly open-sourcing the tools developed in-house, and a number of newer firms have even built their entire business models around OSS. The private sector experience shows that there is no single correct choice; each model has its own advantages and disadvantages.

As regards public sector entities, including central banks, they tend not to disclose their source code. Apart from the need to protect intellectual property rights, a key reason is privacy protection and confidentiality settings, since some codes might contain sensitive information (eg insights into policy). Hence, even the central banks that are most active in OSS generally

prefer to keep some of their source codes confidential. Still, a number of them see value in gradually open sourcing their software and are taking gradual steps in this direction.

Compared with the closed-source model, OSS can bring four main benefits to central banks, in terms of costs, customisation, security and usage base:

- First, from a financial perspective, OSS is virtually always free of charge, unlike most closed-source software. This is an important feature for central banks wishing to control costs while selecting software from a flexible external palette.
- Second, OSS enables central banks to tailor the software they use to their specific needs. In contrast, closed software developed by third parties typically prevents customers from sharing, modifying or using it beyond a narrow set of purposes.
- Third, the disclosure of the source code can make it easier for software users and producers to identify and correct any vulnerability. Indeed, it is usually more straightforward for central banks to test the security and robustness of an OSS than of closed-source programs, thereby reducing operational and security risks.
- Finally, OSS often benefits from a strong collaboration between developers and users. The code is typically shared on accessible “repositories”, which promotes software improvements thanks to collective programming and frequent feedback. In fact, many OSS have user communities that provide efficient support and share learning resources, especially for data science applications, in turn improving the quality of the tools made available. Alternatively, if a central bank decides to produce an internal application itself and make it open source, it can more easily share the related development burden with peer institutions, in turn helping to build collective capacity and promote best practices.

Reflecting these strengths, OSS solutions have proved able to address a variety of needs for software users in central banks. Well known use cases include tools for data management (eg DuckDB, MySQL, Apache Cassandra), and big data (eg Hadoop), data analysis (eg pandas, tidyverse), and data visualisation (eg Shiny, Plotly) as well as more advanced data science applications such as ML and deep learning (eg scikit-learn, PyTorch, TensorFlow, Keras). Furthermore, central banks also use OSS for source control (eg git), integrated development environment (IDE) (eg RStudio, Spyder, VS Code) and programming languages (such as Python, R and Julia). Central banks’ growing interest in these examples stems from the increasingly dominant role that OSS plays in software development, with virtually all commercial applications now having one OSS alternative. In addition, central banks’ use of OSS can help them attract data scientists and other technology experts.

Moreover, a growing number of central banks act as providers of OSS, reflecting their higher degree of technological maturity. For instance, the Bank of England publishes the source code behind its research publications,^③ and the Central Bank of Brazil does the same for the application programming interfaces (APIs) that support its fast retail payment system PIX.^④ Further, various central banks disclose the code of their macroeconomic models^⑤ or of their data management processes, as in the case of De Nederlandsche Bank for the quality rules used to improve supervisory reports.^⑥ Lastly, some institutions also share the codes of their prototype central bank digital currencies (CBDCs). Notable examples are the Federal Reserve Bank of Boston’s partnership with MIT on the OpenCBDC project and the Central Bank of Norway CBDC sandbox.^⑦

However, central banks face important challenges in using and even more in producing OSS. One key issue is security: just as OSS is easier to audit than is closed-source software, malicious actors may be better able to understand how an OSS application could be compromised. Another important drawback is that making a software open source can require considerable financial and human resources. For example, maintaining an OSS in a fully fledged public repository will typically involve dedicated staff to address changing user needs and/or to adapt the software to the evolving technological environment.

Despite such challenges, the rising number of central banks adopting new open source tools confirms that OSS benefits are greater than their limitations. One main factor supporting this trade-off is that open source is not only about code-sharing for development purposes, as it also drives resource-sharing across the user community and active collaboration for the enhancement and development of new projects. In the age of accelerated innovation in the financial space, source code can thus be considered as one of the new frontiers of international cooperation.

The BIS has taken several steps to promote OSS. First, the BIS supports sdmx.io, a platform for managing and exchanging statistical data and metadata.^① This platform has become an ecosystem of OSS tools and components that allow organisations to fully exploit the SDMX open standard^② for collecting, producing and disseminating statistics. A key software made available by the BIS on this platform is the Fusion Metadata Registry (FMR), which enables the management and sharing of SDMX metadata and is built on open-source components shared with other organisations. Another OSS partnering with sdmx.io is the [Stat Suite](https://stat.suite.org), a platform for the efficient production and dissemination of high-quality statistical data, which was developed in partnership with the OECD, Eurostat and the SIS-CC community. Looking forward, sdmx.io seeks to onboard more tools and components in collaboration with interested partner organisations, including the engine developed for the SDMX-based [Validation and Transformation Language \(VTL\)](https://vltl.org) by the Bank of Italy.

Another BIS contribution has been its [Open Tech](https://opentech.bis.org) initiative to promote and support the development and adoption of open-source technology in official statistics and the financial sector. This initiative aims to address the growing demand expressed by central banks, commercial banks and technology providers, with the aim of collaborating together on the development and implementation of open-source solutions as public goods so as to enhance efficiency, security, best practices, knowledge-sharing and innovation. The first BIS Open Tech project was the Project Ellipse integrated regulatory data and analytics platform, which was launched in 2021 by the BIS Innovation Hub in collaboration with the Monetary Authority of Singapore.

Overall, OSS has been already adding value to central banks and is paving the way for greater innovation as well as technical collaboration with their main stakeholders.

① The Open Source Initiative (OSI) specifies internationally recognised criteria for OSS, available at opensource.org/osd. ② D Araujo, G Bruno, J Marcucci, R Schmidt and B Tissot “Machine learning applications in central banking”, *IFC Bulletin*, no 57, 2022. ③ github.com/Bank-of-England. ④ github.com/bacen/pix-api. ⑤ D Araujo, *Open-sourced macroeconomic models*, 2023, github.com/dkgaraujo/OpenSourcedMacroModels. ⑥ github.com/DeNederlandscheBank/data-quality-rules. ⑦ Respectively available at github.com/mit-dci/opencbdc-tx and github.com/norges-bank/cbdc-sandbox-frontend. ⑧ The platform provides tools for cleaning, transforming, and publishing data to make them more easily accessible and usable, www.sdmx.io. ⑨ SDMX is a data exchange standard used by international organisations and national statistical systems, and the platform aims to make working with this standard more user-friendly and efficient, sdmx.org.

A third important area is to **ensure that data scientists’ operations are based on sound and efficient IT processes**, especially when they rely on self-service capacities. For instance, and as mentioned above, a key benefit of the *gingado* package is indeed to foster the dissemination of good practices in ML use cases. It also promotes efficiency through its consistent and simple application programming interface (API) that makes it easy to plug into existing code or to reuse ML code by other teams. In addition, other best practices in software engineering are finding their way into central bank processes. One key example is the use of containers – ie a fully self-contained operating environment on which a user application can be run in an

isolated and portable way.⁶ Containerisation ensures that specific analyses can be “packaged” and therefore run on different computers. It thus provides important opportunities to make business operations more efficient, fosters the reproducibility and portability of the projects involved (for instance in economic research; Vilhuber (2021)), and promotes internal collaboration as well as cooperation with academia and external organisations. These benefits were highlighted by the recent containerisation project (based on the Docker tool) undertaken by the Bank of Canada with Emory University, Ryerson University and Stanford University.

Using cloud computing services?

Cloud computing has emerged as one of the key technological innovations in data architecture, offering advantages such as “scalability” – ie the ability to quickly adapt the IT hardware to perform well a larger number of analyses and processes, operational efficiency, the possibility to use a wider range of tools and computing resources, and continuity – meaning the possibility to keep the software stack up to date (IFC (2020)).

Certainly, **a number of central banks remain sceptical**. In a recent survey of senior IT managers at 25 central banks, Edmond et al (2022) report a reluctance to migrate to the cloud due to concerns about data protection and privacy, security and other operational risks; and even when a cloud strategy exists, a private cloud environment will often be favoured over a public cloud.

Nevertheless, the increasing experience in using cloud computing services has highlighted **a number of important benefits for central banks. One relates to more efficient data ingestion processes**. A cloud-based environment can flexibly allow for the onboarding of large data volumes at speeds that are multiples of those seen with more traditional IT infrastructure. For instance, the cloud-based data lake solution explored by the Bank of Canada accelerated the loading of a data set with millions of observations. One reason was the reduced need for computing memory allowed by the common platform, compared with the previous scenario where each user would require a copy of the data. In another example, the Bank of Canada has been using cloud services to automatically ingest on a weekly basis expenditure information at commercial addresses collected by SafeGraph, a private data vendor. The data set appears to usefully complement other official statistical sources, especially in terms of timeliness and details. For instance, it provides information on businesses openings and closures, facilitating the organisation of the Bank of Canada’s payments surveys.

A second main benefit provided by cloud environments relates to the manipulation of the data once they are ingested. This takes advantage of massive parallel computing, which is the ability to “divide-and-conquer” calculations to accomplish tasks much more quickly than if they were done sequentially. The resulting gains in speed and productivity can be observed for many data management processes, such as data pre-processing and visualisation, but also for analytical tasks. The example provided by Nvidia shows that the parallelised calculation capacity offered in the cloud through graphics processing units (GPU) chips have enabled recent breakthrough advances in sustainable finance, LLMs and

⁶ In other words, containers encapsulate all dependencies of a certain application to ensure it runs the same way in different environments (Mouat (2015)).

other fields relying on the use of AI/ML tools – noting that the training of most of these tools requires considerable computing power.

Further, and despite recurrent security concerns as mentioned above, **cloud-based solutions can also help to maintain or increase the security posture**. First, many cloud providers offer advanced threat detection and response capabilities, which can automatically identify and mitigate potential security breaches, something that might be resource-intensive for a central bank to manage on its own infrastructure. A related example of cloud usage for its security focus is to enable secure collaboration with external researchers. Moreover, a cloud environment can, generally speaking, provide efficient processes for securing software, such as by implementing frequent updates and patching to close any known vulnerabilities. Of course, this does not mean that central banks should entirely outsource cyber security management, but that cloud service providers can be important partners in maintaining an efficient and secure working environment.

In any cases, there are substantial training and other costs associated with the implementation of cloud platforms, although the efficiencies gained might offset these, as discussed by Handel et al (2022). Hence it is important to rigorously **evaluate the resources implications before making any decision to transition to the cloud**. One key element is that on-premise solutions often demand significant expenditures, encompassing hardware acquisition, the setup and maintenance of a data centre, and the associated staff costs. In contrast, cloud solutions usually adopt an operational expenditure model, implying that the recurrent costs supported by client institutions are determined on an ongoing basis and aligned with their usage patterns. A second important point is that, as time progresses, on-premise IT setups would experience higher costs due to hardware ageing and related upgrade requirements. This is less of an issue for cloud infrastructures, which leverage economies of scale and continuous innovation and do usually not pass direct incremental hardware costs to user institutions. Nevertheless, while cloud-based solutions may appear advantageous by smoothing user costs and keeping available IT tools updated over time, it remains crucial to take a comprehensive view of the total expenses involved as well as to keep other strategic considerations in mind.

3. Leveraging statistical and mathematical tools to extract data insights

The second main dimension of data science relates to the **ability to perform various mathematical and statistical operations to deal with the raw data available**. From this perspective, a sound IT infrastructure that supports the required analytical tools and software (see Section 2) is a necessary but not sufficient condition. Data scientists must have the skills to apply various scientific methods and algorithms to access very large and complex data sets, prepare them for further analyses, and conduct inference exercises.

Accessing new, big data sources

Accessing new large and multifaceted “big data” sources has become increasingly important for central banks. One key reason is that, in an increasingly complex financial environment, policy institutions require more information and more

promptly (“actionable knowledge”) if they are to fulfil their mandates – see IFC (2020; 2021a). Fortunately, central banks’ thirst for data can be effectively quenched by leveraging new information sources, ie so-called alternative data. Their usage has increasingly supported a wider range of monitoring and policymaking tasks, as was evident during the Covid-19 pandemic. Looking ahead, these new sources are likely to continue to provide useful value as a complement to the traditional ones in the “new normal” landscape for official statistics (Jahangir-Abdoelrahman and Tissot (2023)).

A key factor, as observed by the Bank of France and the French Prudential Supervision and Resolution Authority (ACPR), is that **alternative sources can bring more timely and higher-frequency data**, which can be very useful in uncertain and fast-moving situations such as a financial crisis. Moreover, the new types of source considered, ranging from Google searches and social media posts to satellite images or mobile phone data, may offer unique insights. For instance, compared with “traditional” statistics, they can provide an almost real-time view of economic developments, often with high granularity in terms of sectors and/or geographic locations. One example has been the use by the International Monetary Fund of payments data from M-Pesa, one of the largest mobile money services in Africa, to analyse recent trends in mobile money – eg the relative importance of peer-to-peer (P2P) transactions and international money transfers and the impact on broad money (Shirono et al (2021)). Another example has been the use of Google Maps data to geolocate financial access points such as ATMs and mobile money agents in Kenya, complementing information collected through more traditional financial inclusion surveys.

A second factor supporting the access to alternative data sets is that they often represent “low-hanging fruit”, since they usually exist as an organic by-product of existing processes and activities. In particular, authorities are increasingly realising the potential value for economic and financial analyses of the wide range of administrative registers that are generated by public sector activities. For instance, electronic payments data have been used in Portugal to study consumer behaviour (Carvalho et al (2020)); similarly, cargo ship identification information was helpful to better track global supply chain activity in real time (Cerdeiro et al (2020)).

Needless to say, data scientists need to be able to **deploy various techniques and tools to efficiently make use of all the various data sources of interest**. This is needed for data collation, ie to properly query, integrate, store, clean, prepare and process the raw data. Cases in point relate to SQL for querying relational databases; the usage of Hadoop for processing vast amounts of data; applying packages such as tidyverse in R or pandas for data manipulation in Python; using integration tools like Talend or Apache NiFi to combine disparate data sources; and leveraging Tableau Prep, OpenRefine or other modern business intelligence (BI) software for cleaning and preparing the data (IFC (2019)). All these steps are essential to make the information ready for further use. And, once the raw data are properly organised and prepared, various platforms such as Jupyter Notebooks, RStudio and Quarto can be mobilised for conducting deeper – and reproducible – analysis.

Preparing raw data as the basis for further analyses

Once the data are accessible, the next step is to prepare them for knowledge extraction. Advanced analytical tools can be instrumental from this perspective,

especially as regards three main tasks inherent to central banks' statistical activities: the selection of the data features that are relevant; the management of data quality; and the dissemination of data to users.

As regards **signal extraction**, modern analytical tools are needed to select the indicators of interest, especially when information is buried in a sea of granular data points. For instance, Bank Indonesia has extracted over 600,000 user comments on the TripAdvisor platform, covering more than 1,000 touristic destinations in the country. This information was used to better measure tourism dynamics, for instance to assess the impact of Covid-19 pandemic. In parallel, another language-based project was related to the compilation of government expenditure statistics, by automating the classification of individual payment transactions.

Meanwhile, the Bank of Spain has developed a specific application to create a database of sustainability information on Spanish corporations. The tool extracts a structured database from the vast amount of environmental, social and governance indicators disclosed by corporates so as to enable regulators to better monitor them. Certainly, the underlying components of the application required significant open source inputs, starting with Streamlit,⁷ an open-source application for ML work, and Dash, a Plotly Python framework for creating interactive web applications.⁸ But in addition to the IT side, the overall data science initiative also required the involvement of statisticians to select the information of relevance. Similarly, the Banco do Portugal has designed a web application to ingest IMF data on the Coordinated Direct Investment Survey (CDIS) and the Coordinated Portfolio Investment Survey (CPIS) to conduct network analysis and derive countries' influence in global direct investment and portfolio flows. The aim was also to allow users to interact with the data, for instance to track individual countries' positions in global investment networks over time.

The second area, **data quality management**, is an imperative for central banks that are important producers of top-quality statistics. But they are also heavy users of data that need to be reliable to support their analyses and ultimately policy decisions. Hence, it is essential for them to detect any quality problems adequately and on time, especially when confronted with spurious data reports. One example was during the financial turmoil induced by the Covid-19 pandemic in March 2020, when central counterparties' (CCPs) initial margins increased dramatically (Boudiaf et al (2023)). Authorities had to analyse whether this reflected a real development, ie the demand for additional protection against a potential default of CCPs' clearing members, or was simply a matter of misreported data.

Fortunately, **advanced analytical tools can be a great help in ensuring the quality of the data sets compiled**, even when these are highly granular and so large that they cannot be manually checked. For instance, the supervisory (big) data set collected from trade repositories (TRs) under the auspices of EMIR presents significant quality challenges due to the high volume of data as well as to obstacles in aggregating inputs across TRs, since a similar transaction can be reported multiple times and in different ways (ESMA (2021); IFC (2018)). The ECB's strategy has been to automate quality controls, allowing for the production of timely analyses. Another example relates to the ECB CSDB, where data on all individual securities in Europe are

⁷ www.streamlit.io.

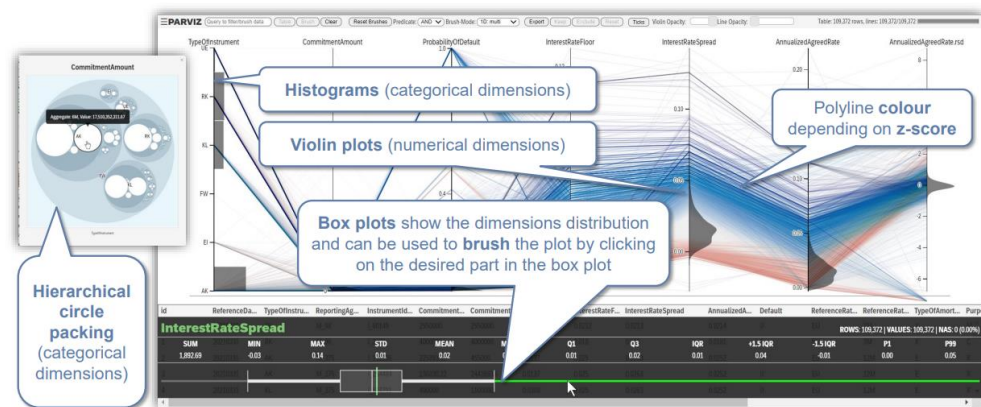
⁸ plotly.com/dash/.

compiled from multiple sources. Various algorithms – eg logistic regression, decision trees with the use of the ML gradient-boosting algorithm CatBoost (Prokhorenkova et al (2018)) – are used to select the data points that would require intervention by a data quality manager. The project has also focused on making the results explainable to end users (by using Shapley values for interpreting the ML modelling) and ensuring their relevance over time as new data come in (by checking the model performance over different periods). The Bank of Italy has also leveraged ML algorithms for data quality controls applied to its credit registry, with the combination of advanced techniques including an autoencoder, a type of artificial neural network to predict data anomalies.

The above examples underline the key contribution of data science in supporting the efficient production of statistics by central banks, helping to increase timeliness while respecting resource constraints and without compromising on data quality. It is also worth noting that **the new tools used for quality management can help authorities liaise with data reporters so that they improve their data submissions**. The ECB, for example, has worked closely with the industry to create the Banks' Integrated Reporting Dictionary (BIRD), a data dictionary to support the standardisation of data transmissions for regulatory or statistical purposes. This has helped to reduce banks' reporting burden, improved the data collected thanks to a better understanding of the reporting guidelines, and facilitated statistical calculations. In addition, regular quality feedback reports are sent to the original data owners of the TR data collected in the EMIR context.

Visual detection of outlier observations in credit data collected by the Austrian central bank

Graph 2



Source: P Reisinger, T Kemetmüller and C Leitner, "Interactive visualisation tool: outlier detection in large multidimensional data sets", *IFC Bulletin*, no 59, October 2023.

Another important comment is that **data quality management can be facilitated by statistical visualisation tools**, particularly to detect and analyse specific data points of interest. For instance, the Banco de Portugal has been using PowerBI, a BI visualisation tool, to check the information contained in its repository of balance sheet data for non-financial corporations. The tool allows for different data sources to be rapidly connected, as well as for consistency checks and drilling down into aggregates so that analysts can better spot individual data points that might need further investigation. In a similar vein, the Central Bank of the Republic of Austria

has developed a tool-based framework to visualise and detect outliers in its credit register. The approach is dynamic as it facilitates the visualisation of the various dimensions of the data (more than 100 variables) while also allowing users to change their search criteria (**Error! Reference source not found.**). Similarly, the BIS makes use of the Tableau dynamic visualisation tool to support data management tasks. The approach is multidisciplinary, involving IT technical features and advanced statistical methodologies, as well as communication aspects (eg as regards the cognitive content of the graphs and related visual perceptions); hence a key lesson of the project was the need to involve the various community of data scientists from different units.

Data science applications have also found their way into the **third main area of data dissemination**. For instance, central banks often receive information requests associated with the data published, eg from the media and the general public. Addressing these requests can be resource-intensive, not least because they need to be properly assigned to the business areas in charge so that they can respond. The ECB's mail robot (MailBot) system was developed to automate the sending of such messages and replace manual intervention. The application allows the classification of inquiries by business areas, the identification of similar queries to avoid duplication of work and the automation of some reply processes. The model was trained on previous requests and answers provided by the business areas, using a specific classification algorithm (extremely randomised trees; Geurts et al (2006)) and estimating the degree of similarity between text queries through their vectoral representations.

Extracting analytical insights from data...

Much of what is usually considered as "data science" encompasses the **various mathematical methods that can be used to extract relevant insights from the data once they have been properly collected and prepared**.

In general terms, these methods are based on AI, ie "the various computer systems that can perform tasks that traditionally have required human intelligence" (FSB (2017)). This includes the important subset of ML, "a method of designing a sequence of actions to solve a problem that optimise automatically through experience and with limited or no human intervention". There is also a growing interest in so called generative AI, ie AI models learning from their input ("training data") and able to generate new data with similar characteristics (for instance new texts produced with LLMs).⁹ The approach can be top-down, with humans designing how the data should be processed by the systems that can mimic human-like calculations, but much faster and on a vastly greater scale. It can also be bottom-up, using algorithms adapted to fit the data and with a focus on optimising the intrinsic performance of the model instead of mimicking human behaviour.

In practice, **data insights can be extracted in multiple ways**. The diversity of ML algorithms provides a large number of practical alternatives for exploring data (Athey and Imbens (2019)). When the task can be performed with the help of a particular subsample of the data set for which the outcome is known (and on which the model can be trained), "supervised learning" algorithms can be used, such as

⁹ Examples of LLMs include OpenAI's GPT models (used in ChatGPT), Google's PaLM (used in its conversational AI tool Bard), Meta's Llama as well as BigScience's open model BLOOM; see Box 2.

regularised regressions, random forests, gradient boosting trees and regression or classification neural networks (LeCun et al (2015)). In other cases – for instance when the explored data points need to be grouped automatically or summarised into fewer dimensions, or for other cases where the algorithm has to identify the patterns in the data autonomously – then unsupervised learning ML models such as clustering techniques and manifold learning can come into play. Other ML models, less used in data science currently but with arguably considerable potential in supporting decision-making, are part of reinforcement learning, ie ML algorithms that follow an optimisation rule to maximise a specific objective. Importantly, models of the same or different types can be combined in more or less sophisticated ways, as shown in the examples discussed by Araujo et al (2022).

... including unstructured data like text

An increasing number of central banks are working on the use of NLP techniques to support a whole range of applications when dealing with new information sources (eg Gentzkow et al (2019); Araujo et al (2022)). This interest reflects **the large amount of text data, often unstructured, that are available to them as part of their routine activities**. Another important development has been technological advance, which has facilitated the development of NLP models based on multiple languages and thereby their global use.¹⁰ Moreover, NLPs' capabilities have been expanding, spurring central banks' interest in LLMs (see Box 2).

Box 2

Harnessing recent breakthroughs in large language models (LLMs)

Douglas Araujo, Stephan Probst, Rafael Schmidt, Boris Vitez, Markus Zoss

In recent years, the field of natural language processing (NLP) has made significant progress, primarily due to the emergence of LLMs (Graph 3) such as ChatGPT – the language model developed by OpenAI, which is capable of generating text based on context and past conversations. This box discusses some of the ways LLMs might be helpful for central banks and the associated challenges.

These models are built with a ground-breaking neural network architecture known as the “transformer”,^① which enhances models' ability to grasp nuances of word meanings in their context and enables them to expand in size and process vast amounts of text. As a result, LLMs now achieve human-like proficiency in a variety of tasks that involve language, such as generating various text formats, text summarisation, sentiment analysis, translation etc. Sophisticated post-training methods, such as supervised fine-tuning and reinforcement learning from human feedback, have further refined the more recent LLMs, providing them with more advanced reasoning capabilities. These developments open up new potential use cases.

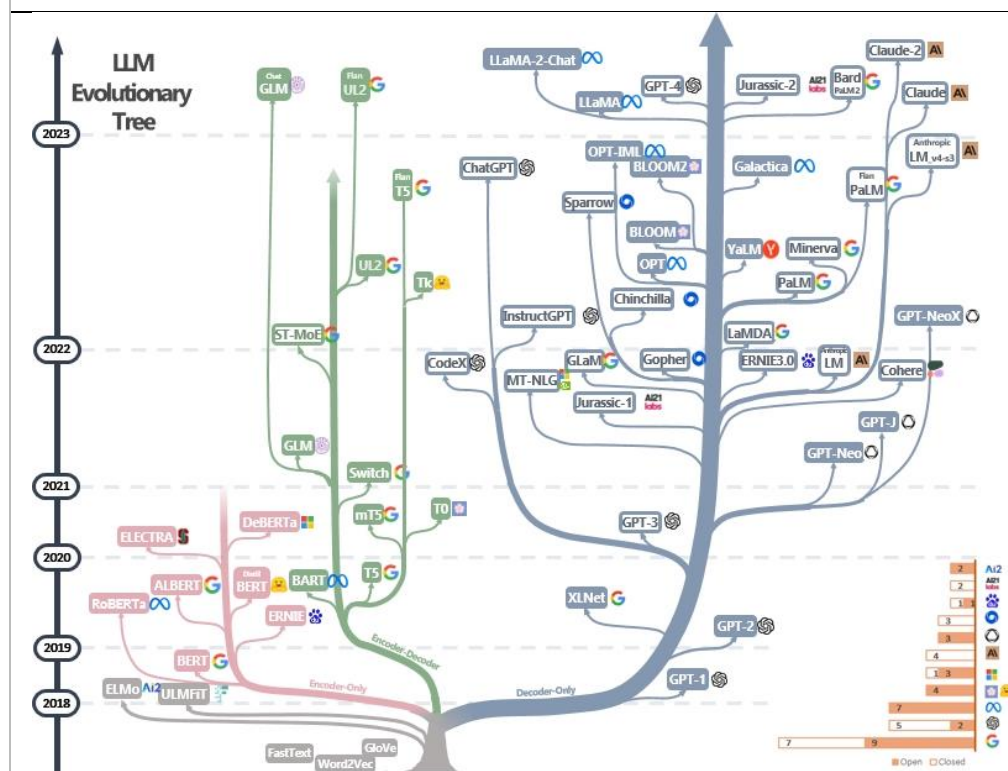
¹⁰ Initially, the development of large NLP models concentrated in the English language and the availability of advanced NLP tools was more limited for non-English languages. To address this issue, the Bank of Thailand has developed three AI-powered NLP applications so that it can analyse documents written in the Thai language. Looking ahead, the growing availability of various models even for languages with relatively fewer speakers, such as Google's MADLAD-400 (Kudugunta et al (2023)), is likely to spur central banks' exploration of various applications using local language(s).

A combination of factors contributes to the superiority of LLMs over other language models. The first is their architecture: unlike previous models, which look at text data sequentially, transformers can capture long-range dependencies and better differentiate the meaning and importance of words in different contexts. Also, this architecture can be significantly scaled up, leading to the second factor behind LLMs' success: their size. These models are much larger, allowing them to learn more intricate patterns and representations from the massive amounts of text data they are trained with. As a result, their output better resembles general language and specialised areas of written knowledge. A third factor is how they are trained. While generally all LLMs follow lengthy "classical" training processes, albeit with a larger data set, more recent models add more human-intensive steps that seek to improve their helpfulness, harmlessness and truthfulness. With this, LLMs can more easily adapt to new tasks after being given a few examples or even without any examples. These factors support enhanced performance and, crucially, the ability to transfer knowledge from one task to another, which is a key advantage of LLMs.

The evolutionary tree of large language models (LLMs)

Updated as of 6 August 2023

Graph 3



Source: J Yang, Q Feng, X Han, X Hu, H Jiang, H Jin, R Tang and B Yu, "Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond", arXiv, April 2023.

LLMs have led to breakthroughs in various NLP applications, such as in conversational AI, leading to significant quality improvements. For instance, new computer dialogue systems can provide more coherent and contextually relevant responses to human inquiries. For text summarisation, LLMs generate accurate and concise summaries of lengthy documents in a way that can be calibrated by the user to reach different audiences (eg a technical abstract, or a higher-level executive summary). Further, these models can accurately perform sentiment analysis of texts, by capturing nuances and contextual information better than previous models. Other examples are automated translation between languages, where LLMs can produce more fluent and contextually accurate output, and software engineering, where these models may help design, test and document code.

Because of these capabilities, LLMs are becoming useful tools in economic research, echoing the achievements they have brought in other domains. At the inception of research projects, they may assist in ideation by providing brainstorming possibilities and informed counterarguments. In the writing phase, LLMs can aid in text synthesis, editing and crafting compelling headlines, making the dissemination of research more impactful. Background research could also be streamlined with LLMs by supporting literature reviews, reference formatting, and even explaining complex concepts in simpler terms. On the technical front, LLMs can help in coding and debugging, data reformatting, or data classification. The versatility of LLMs may therefore lead to important possibilities for reshaping economic research looking forward.^②

Central banks are also increasingly recognising the potential of LLMs to support their various activities.^③ Economic data analysis is a major area of focus, as LLMs can help to parse vast amounts of economic reports, documents and news to efficiently extract, summarise and present information. In particular, the analysis of financial news, social media and other public fora facilitates the gauging of public sentiment, for instance as regards current and future economic and financial conditions, or the conduct of specific policies. The models also offer valuable support for the various internal exercises (eg risk assessments, economic forecasts) as required by the conduct of monetary and financial stability policies. Lastly, some central banks have started deploying LLMs in public relations tasks, for instance to handle inquiries from the public or financial institutions or to explain complex economic topics, regulatory requirements as well as policy decisions.

Looking ahead, the range of central banking applications that could benefit from LLMs is likely to continue to expand, especially in the policy area. For example, these models could support more complex financial monitoring exercises by tracking in a granular way market segments, institutions or instruments that show signs of instability or policy concerns. Regulators could leverage LLMs to facilitate compliance monitoring, eg by analysing internal reports to assess the implementation of financial regulations. Similarly, supervisors could detect suspicious patterns and potential misbehaviour or illicit activities by scrutinising transaction-level data and related textual content. Lastly, LLMs might assist in training central bank staff as well as the broader financial community and the public at large, by making complex topics more accessible, enhancing financial literacy, and supporting data queries at scale through more intuitive interfaces.

However, with great potential comes great responsibility. Central banks are rightfully cautious as they integrate these technologies. Transparency is paramount in understanding how the new models derive their conclusions and in communicating to the public the reasoning behind policy actions. A related issue is algorithmic unfairness, ie the risk that (non-explicit) unfair discrimination or bias entailed in pre-existing material used for training the model can be systematically perpetuated or even amplified, especially as it interfaces directly or indirectly with people. The trust in high-quality official statistics hinges upon addressing this pivotal question.^④ Consequently, a number of authorities have worked recently on documenting ethical considerations for guiding ML development and usage.^⑤ Data privacy also remains a top concern, given the risk that sensitive information could be inadvertently processed by LLMs and then disseminated in unexpected ways. Lastly, as for other AI-based tools, there is always the danger of overreliance on automated processes, with the need to strike the right balance between machine assistance and human judgment, especially for policy decisions.

In summary, LLMs have already ushered in a new era of automation and experimentation in central banking operations. Their ability to understand language and context in a massive and automated way has led to breakthroughs in several real-world applications. These advances can hold great promise in the central banking domain. But, as AI's role evolves, continuous research, collaboration and transparency will be essential to harness its benefits responsibly.

^① A Gomez, L Jones, Ł Kaiser, N Parmar, I Polusukhin, N Shazeer, J Uszkoreit and A Vaswani, "Attention is all you need", *NeurIPS Proceedings*, 2017. ^② A Korinek, "Generative AI for economic research: use cases and

implications for economists”, September version submitted to the *Journal of Economic Literature*, 2023. ③ D Araujo, G Bruno, J Marcucci, R Schmidt and B Tissot, “Machine learning applications in central banking”, *IFC Bulletin*, no 57, 2022. ④ C Julien, “*Machine Learning project report*”, UNECE, September, 2020. ⑤ UK Statistics Authority, “*Ethical considerations in the use of Machine Learning for research and statistics*”, October 2021, <https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-in-the-use-of-machine-learning-for-research-and-statistics/>.

As a result, various **NLP-based data science projects aim to leverage the considerable amount of texts available from various different sources**. This material is being increasingly used to support internal processes in central banks, reflecting the fact that they routinely analyse very large volumes of textual information, usually in unstructured format – ranging from internal documents (eg financial stability or monetary policy reports, governors’ speeches), reports from supervised entities (eg board documents, loan documentations), or external publications (eg social media, financial press, economic papers) – that often need to be analysed by various teams with different skills and background in parallel.

Central banks are taking these opportunities to explore and eventually reap benefits from a variety of texts, including regulatory and market data. Crucially, these data sets can be used by themselves, or merged with traditional data, to obtain new policy insights, eg by offering a more timely and granular assessment of the economy and by allowing a better understanding of how market participants are affected by shocks in financial networks. These use cases have benefited from the **recent breakthrough provided by language-based ML models**, which cover a broad range of applications, both for internal processes and to directly support policy. But of course, LLMs also have downsides, including questions about how to ensure the suitability of training data, the need for significant investments to preserve security, their high energy consumption during training, and the often unsatisfactory levels of transparency (Chen et al (2023)) – a point of particular importance for public authorities. The more experience central banks gain in exploring these issues, the more they will be in a position to address the associated challenges while keeping risks under control.

Experience shows that **there is no one-size-fits-all approach**. Advanced NLP models can address a wider range of use cases, but ML techniques can also allow for more nuanced, quantitative analyses of textual information. Moreover, simpler techniques can also deliver important analytical benefits. This puts a premium on combining various techniques. One example is the Bank of Italy’s assessment of whether information from newspapers can be useful for forecasting the stock market index and the relative performance of Italian banks. A sentiment analysis dictionary was developed to assess the predictive power of news, as broken down by topics using the Lasso regression analysis method (Tibshirani (1996)). The approach underlined the importance of using local language dictionaries in text analyses and also the fact that **interesting results could be obtained with simpler techniques such as dictionary-based search analyses compared with more sophisticated and resource-intensive NLP models**.

4. Incorporating subject matter expertise to support well defined analytical use cases

The use of advanced analytics, to be fully efficient, **requires a good understanding of the projects involved and therefore a close interaction with the business areas, including their subject-matter experts and statistical methodologists**. In other words, a data science process developed in a specific domain cannot be blindly applied to another completely different area. That third dimension of data scientists' work is essential, not least because it is key to ensure that the insights extracted from existing indicators represent useful knowledge (Drozdova (2017)). This aspect is even more important in policy institutions such as central banks, since their actions are taken on the basis of available data and have to be clearly explained to the public.

To illustrate this point, **three main areas deserve to be highlighted among the various data science projects undertaken by central banks**. First, the new approaches can help to make a better sense of the vast amount of granular information available in today's societies.¹¹ Second, they tend to provide useful insights on the state of the economy and its prospects, which is a key input for central banks in pursuing their mandates in the areas of monetary and financial stability. Third, they may help to assess how their policies are communicated and can be made more effective.

Making sense of the wealth of granular financial information

An important issue for central banks and financial supervisors is to deal with the vast amount of granular data collected on the financial system and to extract relevant indicators (IFC (2021b)). The problem here is to be able to detect signals at a very detailed level (eg a specific institution or a market segment) without being overwhelmed by the vast amount of data points to be considered. The key is to see "the forest as well as the trees" (Borio (2013)) which calls for both automated analytical techniques and a good understanding of user information needs.

Network analysis methods can be useful in addressing these aspects, as shown by the Bank of Japan in monitoring transactions in the Japanese government bond (JGB) repo market. The approach helped to extract two main insights from the large and granular data set considered: (i) transaction relationships are built around a small number of leading institutions acting as market intermediaries; and (ii) the Covid-19 pandemic significantly reduced the importance of securities lenders relative to other intermediaries. In another study, the Bank of Japan assessed the role of currency swaps as a source of US dollar funding based on transaction-level data collected on the over-the-counter derivatives market. The approach proved helpful in analysing the characteristics of the cross-currency swap market in Japan; and in monitoring market liquidity and the trading behaviour of market participants in a timely and detailed way and at a high frequency. Network analysis and NLP techniques have also been applied in a number of jurisdictions to identify fraud or misconduct. The aim is typically to extract from the vast amount of information

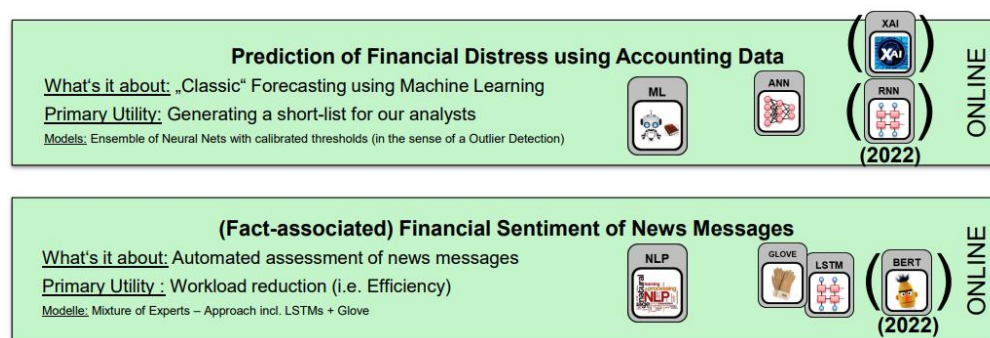
¹¹ In addition to the basic data editing tasks that can greatly benefit from data science applications when dealing with large data sets, as argued in Section 3.

(including text) generated by financial activities signals that need to be further analysed by experts and that may warrant on-site examination by supervisors.

Other techniques can also be applied in isolation or as a complement to facilitate micro-level investigations. For example, the Deutsche Bundesbank has in production a number of tools for risk management purposes. One key objective was to improve the assessment of the credit risk posed by its market counterparties, using ML-based forecasts of potential financial distress (estimated from accounting data) complemented by an automatic NLP-based algorithm so as to select incoming news for further investigation (**Error! Reference source not found.**). In a similar way, the Bank of Thailand's supervision of domestic financial institutions has been facilitated by an NLP-based analysis of the meeting minutes of boards of directors – using a combination of tools to support word segmentation, topic modelling and name-entity extraction (**Error! Reference source not found.**).

Use cases in production for machine learning algorithms at the Deutsche Bundesbank

Graph 4



Source: B Sahamie, “Natural language processing for risk management: discussion of use cases”, *IFC Bulletin*, no 59, October 2023.

Nowcasting and modelling the economy

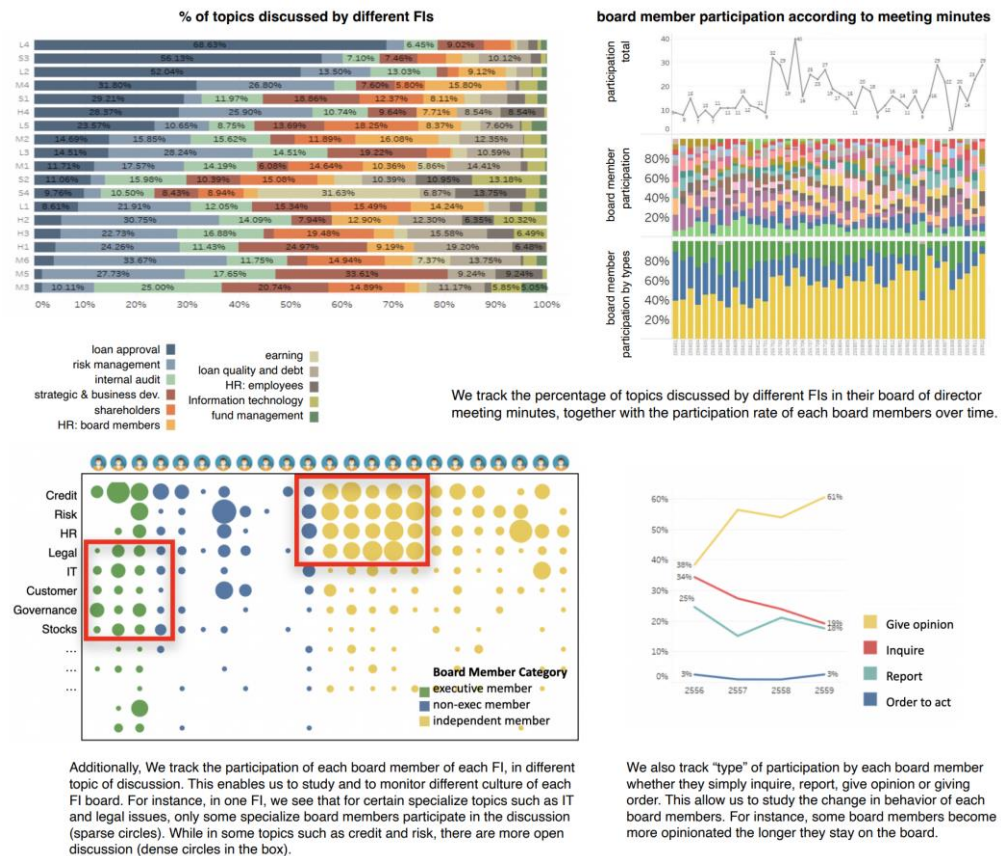
In pursuing policy goals such as monetary and financial stability, central banks have to rely on a sound analysis of the state of the economy (eg GDP growth, inflation) and potential developments. **Nowcasting techniques have therefore been increasingly used for supporting the associated monitoring and forecasting exercises** (see Giannone et al (2008); Bańbura and Modugno (2014); Kronenberg et al (2023)), with typically two main approaches.

The first type of approach is data-driven, by agnostically “letting the data speak” before validating the results with expert judgment. A large number of variables will often be considered by the model, with due consideration for the different frequencies available and/or data vintages. For instance, the BIS has an initiative to nowcast countries’ business and financial cycles based on a fully automated infrastructure so as integrate a large pool of indicators (ie various macroeconomic and financial time series as well as social media text) as they come in. Similarly, an ECB and Delft University of Technology project has combined traditional vector autoregression (VAR) modelling with the use of a deep learning architecture based on neural networks to estimate relationships in the economy. The results suggest that

there are important non-linearities in the interactions between variables and across time periods that could be better explored for macroeconomic modelling.¹²

Analyses of discussions of boards of directors at financial institutions supervised by the Bank of Thailand

Graph 5



Source: Treeratpituk, P and J Kerd Sri, "Integrating natural language processing to central bank operations at Bank of Thailand", *IFC Bulletin*, no 59, October 2023. FIs = financial institutions; IT = information technology.

Input data can also be non-quantitative, such as text. For instance, the Bank of Korea has put together 450,000 news articles from the web to build a news sentiment index (NSI) for the Korean economy using neural networks techniques. This NSI has proved to be a good leading indicator for GDP growth, allowing policymakers to assess the economic situation in a timely manner and with little cost (in comparison, for instance, with business surveys). Another analysis by the Delft University of Technology showed that extracting textual information from economic and financial news can help to anticipate oil price uncertainty. Based on the Baker et al (2016) methodology, an index was built to reflect the degree of uncertainty involved in news texts and the topical characterisation of the spikes observed in oil price uncertainty. The project involved the embedding of the texts, by mapping their words into numerical vectors (using the neural network-based approach doc2vec; Le

¹² Lenza et al (2022) have also documented the importance of considering the non-linear relationship between inflation and its determinants in Europe, using a quantile regression forest approach.

and Mikolov (2014)), an assessment of the similarity between the articles considered, their classification in clusters based on the Louvain algorithm (Blondel et al (2008)), and the identification of each cluster by a specific topic using the Latent Dirichlet Allocation method (Blei et al (2003)).

A second approach is to select ex ante a set of specific variables and assess their usefulness for economic analyses. A case in point relates to payments data, which have been collected for many years by several central banks and are now increasingly tapped to extract relevant insights. For instance, Bank Indonesia leverages data on retail payments transactions collected from the National Clearing System and classified by specific keywords to construct a real-time measure of household consumption. One issue, however, is that the model performance is not a given and may vary significantly over time, even though it appears to have improved significantly since the pandemic. In parallel, Bank Indonesia also nowcasts sectoral corporate activity using transaction-level payments data between more than 100 corporations spanning nine sectors, as collected in its real-time gross settlement system. One challenge was to clean entity names to ensure that the same firm is identified even in cases where the spelling changes slightly across different data points. Another issue was the different model performance observed across economic sectors. A third example is the Bank of Japan's work with personal mobility data collected from smartphones to nowcast economic activity in the first stages of the Covid-19 pandemic. Business data with GPS information from Agoop was matched with administrative data sets to measure business activities in labour-intensive industries and the services sector.

Of course, **the two types of approach can be combined.** For instance, the Narodowy Bank Polski has used a vector error correction model (VECM) to model the non-performing loan (NPL) portfolios of Polish banks. The project first looked agnostically at the various potential explanatory variables that might drive NPL dynamics with a view to selecting the most important ones, such as economic growth. This was combined with a parallel study that was focused on the specific impact of foreign direct investment (FDI) on economic growth and NPLs.

In any case, a major lesson from the above examples is, first, that there are increasingly available information sources that one can tap to assess the evolving macroeconomic landscape in a timely way. Second, as for other data science applications, **nowcasting exercises require the combination of strong IT computing capacities, advanced analytical techniques, and deep expert judgment** – either to validate ex post the automated results obtained, or ex ante to select the variables of interest before testing their information content.

Policy communication

An important issue is to **assess the way policy is communicated to the public, an area to which central banks are paying increasing attention** (IFC (2022)). For instance, the Reserve Bank of Australia has analysed how various audiences perceive communication quality in terms of the degree of readability and reasoning attributed to the messages published. These were processed by an NLP model containing a part-of-speech (PoS) tagger to identify the contribution of each word to a particular part of the text, coupled with a syntax tree parser to take into account its grammatical structure. The processed messages were then classified with random forests depending on their characteristics such as their degree of readability or reasoning

and types of audience (eg economists vs others). The analysis showed that there could be some trade-offs, for instance that simplicity may improve the readability of a text but not enhance the clarity of its reasoning – suggesting that central banks may be better off by producing different texts, each tailored to a specific audience category.

Moreover, **language-based models have proved helpful for assessing the effectiveness of central bank policies**. As an example, the Central Bank of Malaysia has used automated content analysis to extract the sentiment from each of its monetary policy statements published between 2004 to 2020. The initiative required the development of a specific monetary policy dictionary to classify words as “dovish” or “hawkish”. This allowed the Bank to test the prediction content of its statements in terms of interest rate developments and other financial market movements. Another approach followed by Bank Indonesia aimed to capture public opinion, as expressed on social networks, on issues related to central bank activities. In particular, the project deployed language analytical tools both to measure public perception of the credibility of central bank monetary policy actions based on multiple daily articles collected from varied domestic news source and to test its relevance for analysing the level and stability of inflation expectations.

References

- Araujo, D (2023): "gingado: a machine learning library focused on economics and finance", *BIS Working Papers*, no 1122.
- Araujo, D, G Bruno, J Marcucci, R Schmidt and B Tissot (2022): "Machine learning applications in central banking", *Journal of AI, Robotics & Workplace Automation*, vol 2, issue 3, pp 271–93.
- Athey, S and G Imbens (2019): "Machine learning methods that economists should know about", *Annual Review of Economics*, vol 11, pp 685–725.
- Baker, S, N Bloom and S Davis (2016): "Measuring economic policy uncertainty", *Quarterly Journal of Economics*, vol 131, pp 1593–636.
- Bañbura, M and M Modugno (2014): "Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data", *Journal of Applied Economics*, vol 29, no 1, pp 133–60.
- Bholat, D (2020): "The impact of Covid on machine learning and data science in UK banking", *Bank of England Quarterly Bulletin*, Q4.
- Blei, D, M Jordan and A Ng (2003): "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, no 3, pp 993–1022.
- Blondel, V, J-L Guillaume, R Lambiotte and E Lefebvre (2008): "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, vol 10.
- Boudiaf, I, M Scheicher and F Vacirca (2023): "CCP initial margin models: A peek under the hood", *SUERF Policy Brief*, no 624, June.
- Borio, C (2013): "The Great Financial Crisis: setting priorities for new statistics", *Journal of Banking Regulation*, vol 14, pp 306–17. Also published as *BIS Working Papers*, no 408, April.
- Carvalho, B, S Peralta and J dos Santos (2020): "What and how did people buy during the Great Lockdown? Evidence from electronic payments", *Covid Economics*, Centre for Economic and Policy Research, vol 28.
- Cerdeiro, D, A Komaromi, Y Liu and M Saeed (2020): "World seaborne trade in real time: a proof of concept for building AIS-based nowcasts from scratch", *IMF Working Papers*, no 7.
- Chen, L, M Zaharia and J Zou (2023): "How is ChatGPT's behavior changing over time?", arXiv:2307.09009 [cs.CL], August, <https://doi.org/10.48550/arXiv.2307.09009>.
- Donoho, D (2017): "50 years of data science", *Journal of Computational and Graphical Statistics*, vol 26, no 4, pp 745–66.
- Drozдова, A (2017): "Modern informational technologies for data analysis: from business analytics to data visualisation", *IFC Bulletin*, no 43, March.
- Edmond, D, S Bawa, L Garg and V Prakash, (2022): "Adoption of cloud services in central banks: hindering factors and the recommendations for way forward", *Journal of Central Banking Theory and Practice*, vol 11, no 2, pp 123–43, May.
- European Securities and Markets Authority (ESMA) (2021): *EMIR and SFTR data quality report*.

Financial Stability Board (FSB) (2017): Artificial intelligence and machine learning in financial services: market developments and financial stability implications, www.fsb.org/wp-content/uploads/P011117.pdf.

Gentzkow, M, B Kelly and M Taddy (2019): "Text as data", *Journal of Economic Literature*, vol 57, no 3, pp 535–74.

Geurts, P, D Ernst, and L Wehenkel (2006): "Extremely randomized trees", *Machine Learning*, no 63, pp 3–42.

Giannone, D, L Reichlin and D Small (2008): "Nowcasting: the real-time informational content of macroeconomic data", *Journal of Monetary Economics*, vol 55, no 4, pp 665–76.

Handel, D, A Ho, K Huynh, D Jacho-Chavez and C Rea (2022): "Cloud computing research collaboration: an application to access to cash and financial services", *IFC Bulletin*, no 57, November.

Irving Fisher Committee on Central Bank Statistics (IFC) (2016): "Central banks' use of the SDMX standard", *IFC Report*, no 4.

——— (2017): "Big data", *IFC Bulletin*, no 44.

——— (2018): "Central banks and trade repositories derivatives data", *IFC Report*, no 7.

——— (2019): "Business intelligence systems and central bank statistics", *IFC Report*, no 9.

——— (2020): "Computing platforms for big data analytics and artificial intelligence", *IFC Report*, no 11.

——— (2021a): "Use of big data sources and applications at central banks", *IFC Report*, no 13.

——— (2021b): "Micro data for the macro world", *IFC Bulletin*, no 53.

——— (2022): "How central banks communicate on official statistics", *IFC Report*, no 15, February.

Jahangir-Abdoelrahman, S and B Tissot (2023): "The post-pandemic new normal for central bank statistics", *Statistical Journal of the IAOS*, vol 39, no 3, pp 559–72.

Križman, I and B Tissot (2022): "Data governance frameworks for official statistics and the integration of alternative sources", *Statistical Journal of the IAOS*, vol 38, no 3, pp 947–55.

Kronenberg, P, M Bannert, H Mikosch, S Neuwirth and S Thöni (2023): "The Nowcasting Lab: live out-of-sample forecasting and model testing", available at SSRN.

Kudugunta, S, A Bapna, I Caswell, C A Choquette-Choo, O Firat, X Garcia, A Kusupati, K Lee, R Stella, D Xin and B Zhang, (2023): "MADLAD-400: A multilingual and document-level large audited dataset", arxiv.org/pdf/2309.04662.pdf.

Le, Q and T Mikolov (2014): "Distributed representations of sentences and documents", *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pp 1188–96, Beijing.

LeCun, Y, Y Bengio and G Hinton (2015): "Deep learning", *Nature*, vol 521, pp 436–44, May.

- Lenza, M, I Moutachaker and J Paredes (2023): "Density forecasts of inflation: a quantile regression forest approach", *ECB Working Paper Series*, no 2830.
- Loiacono, G and E Rulli (2022): "ResTech: innovative technologies for crisis resolution", *Journal of Banking Regulation*, vol 23, pp 227–43.
- Moor, J (2006): "The Dartmouth College Artificial Intelligence Conference: the next fifty years", *AI Magazine*, vol 27, no 4, pp 87–91.
- Mouat, A (2015): "Using docker: developing and deploying software with containers", *O'Reilly Media, Inc.*
- Organisation for Economic Co-operation and Development (OECD) (2019): "The path to becoming a data-driven public sector", *OECD Digital Government Studies*, OECD Publishing, November.
- Prokhorenkova, L, A Dorogush, A Gulin, G Gusev and A Vorobev, (2018): "CatBoost: unbiased boosting with categorical features", *Advances in Neural Information Processing Systems (NeurIPS)*, vol 31.
- Shirono, K, H Carcel-Villanova, E Chhabra, B Das and Y Fan (2021): "Is mobile money part of money? Understanding the trends and measurement", *IMF Working Papers*, no 117.
- Teles Dias, L (2021): "Post-crisis skills landscape: the emergence of 'purple people'", *IFC Bulletin*, no 53, April.
- Tibshirani, R (1996): "Regression shrinkage and selection via the Lasso", *Journal of the Royal Statistical Society, Series B (Methodological)*, vol 58, no 1, pp 267–88.
- Tukey, J (1962): "The future of data analysis", *The Annals of Mathematical Statistics*, no 33, pp 1–67.
- Vilhuber, L (2021): "Use of Docker for reproducibility in economics", Office of the American Economic Association Data Editor, <https://aeadataeditor.github.io/posts/2021-11-16-docker>.
- Walczak, E (2019): "Data science in the Bank of England: who are we and what do we do?", lecture at Taras Shevchenko National University of Kyiv, May.
- Wibisono, O, H Ari, B Tissot, A Widjanarti and A Zulen (2019): "Using big data analytics and artificial intelligence: a central banking perspective", "Data analytics", *Capco Institute Journal of Financial Transformation*, 50th edition, pp 70–83.

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Keynote speech

Artificial intelligence in finance - quo vadis¹

Joerg R. Osterrieder,

Professor of Sustainable Finance, Bern Business School, Switzerland

Associate Professor of Finance and Artificial Intelligence, University of Twente, Netherlands

Chair of the European COST Action CA19130 on Fintech and Artificial Intelligence

Coordinator of the European MSCA Doctoral Network on Digital Finance

¹ This contribution was prepared for the workshop. The views expressed are those of the author and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

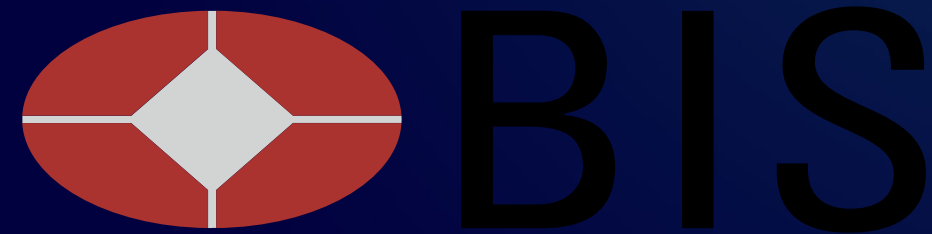
Artificial Intelligence in Finance Quo Vadis?

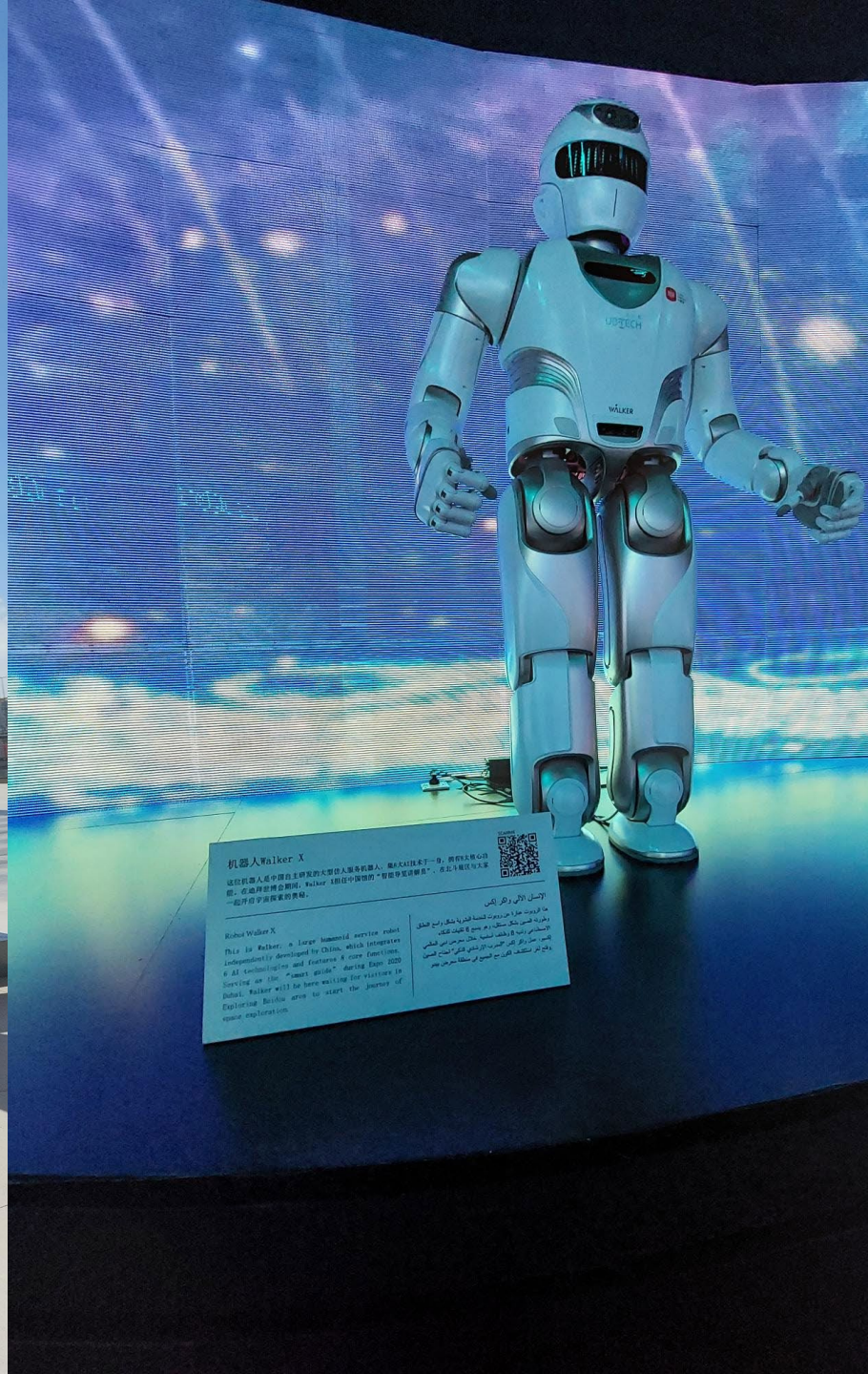


IFC workshop on Data science in central banking:
Applications and Tools
14 – 17 February, 2022

Irving Fisher Committee on Central Bank Statistics

Prof. Dr. Jörg Osterrieder





Prof. Dr. Jörg Osterrieder

- Professor of Finance and Risk Modelling, Zurich University of Applied Sciences, Switzerland
- Associate Professor of Finance and Artificial Intelligence, University of Twente, Netherlands
- Action Chair EU COST Action Fintech and Artificial Intelligence
- Senior Advisor to ING Group, Netherlands



Our Team



Jörg Osterrieder

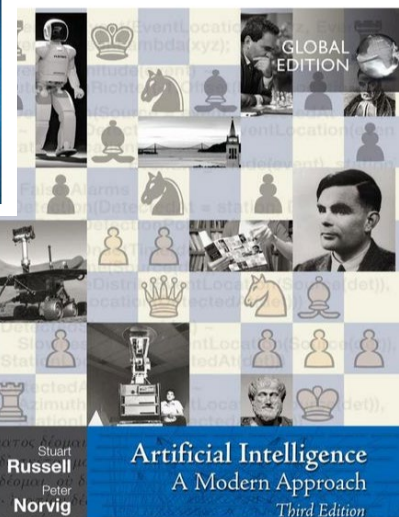
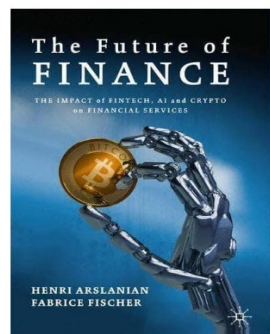


Branka Hadji Misheva

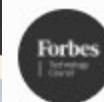


Piotr Kotlarz

AI has started a financial revolution - here's how



Preparing Your Business For The Artificial Intelligence Revolution



Dmitry Matskevich Forbes Councils Member
Forbes Technology Council COUNCIL POST | Paid Program
Innovation

POST WRITTEN BY
Dmitry Matskevich

CEO and Co-founder of **Dbrain**, a blockchain platform to collectively build AI apps.



Forbes

What Is The Artificial Intelligence Revolution And Why Does It Matter To Your Business?



IT STARTUPS
CHANGING THE RULES

FINANCIAL REVOLUTION

FUTURE DEVELOPMENT

Whoever leads in artificial intelligence in 2030 will rule the world until 2100

Financial Revolution: How IT Startups Change The Rules On Wall Street

Forbes

3,529 views | Jun 4, 2020, 03:16pm EDT

Is AI Overhyped?



Kathleen Walch Contributor
COGNITIVE WORLD Contributor Group

AI

Feb 22, 2019, 10:24am EST

The Fintech Revolution Is Here. Can It Help Build A Better Economy?



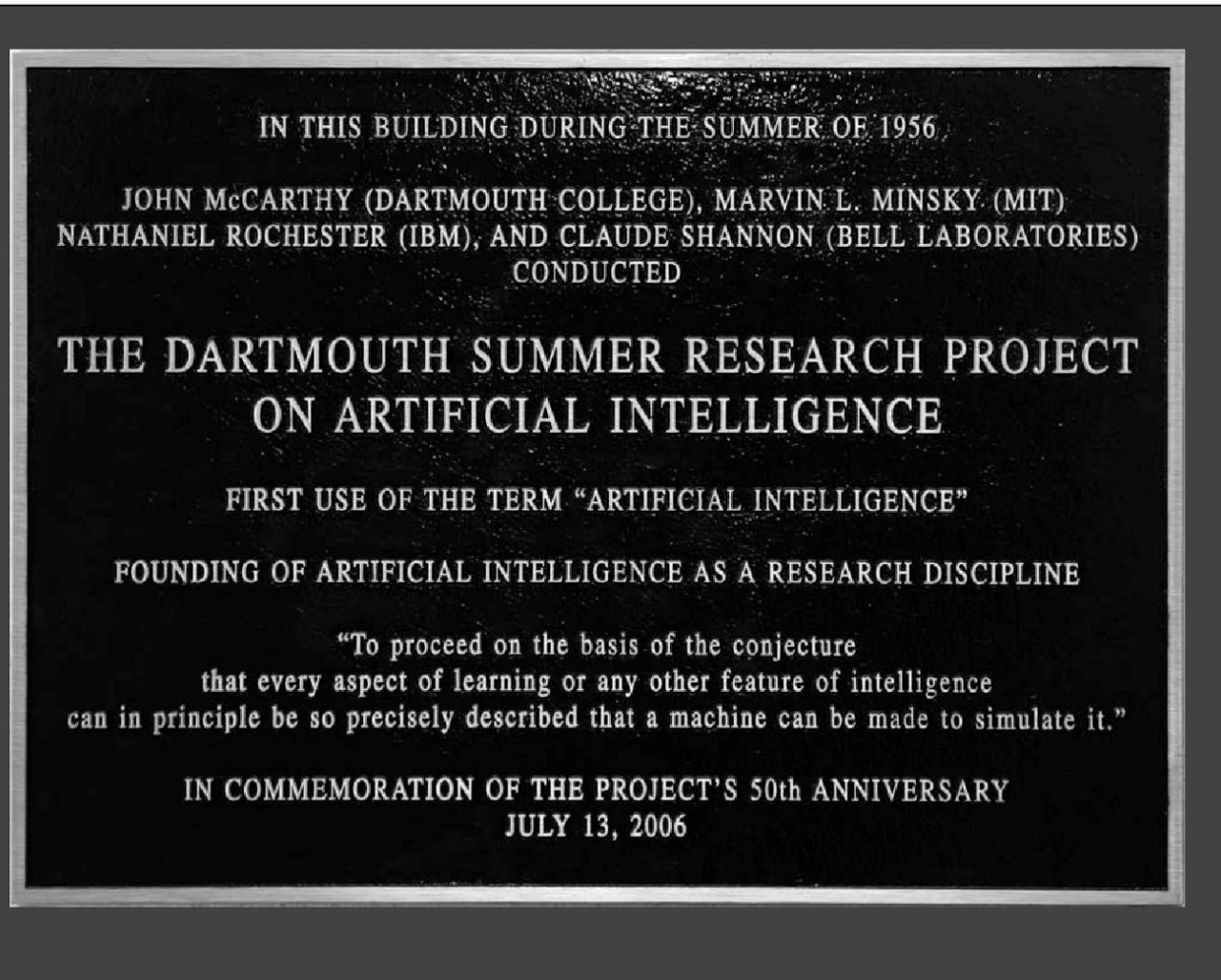
Jennifer Pryce Contributor
Hedge Funds & Private Equity
I connect communities and capital markets





**What is
Artificial Intelligence?**

Dartmouth Summer Research Project on Artificial Intelligence - Summer 1956



The study is to proceed on the basis of the conjecture that every aspect of learning or **any other feature of intelligence** can in principle be so precisely described that a **machine can be made to simulate it**.

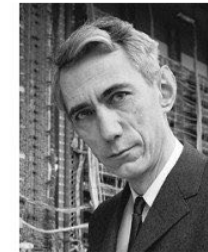
1956 Dartmouth Conference: The Founding Fathers of AI



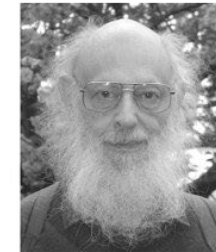
John McCarthy



Marvin Minsky



Claude Shannon



Ray Solomonoff



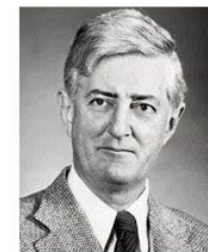
Alan Newell



Herbert Simon



Arthur Samuel



Oliver Selfridge



Nathaniel Rochester

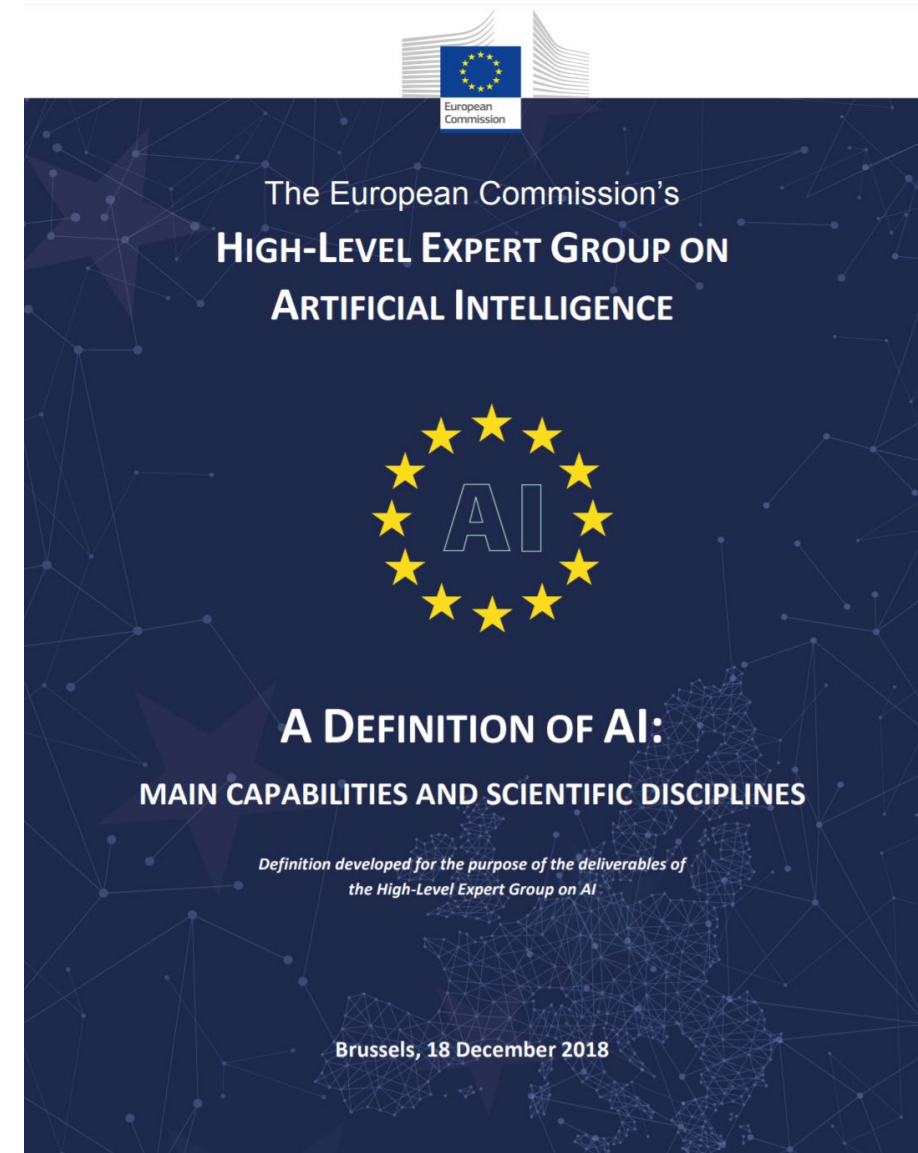


Trenchard More

Artificial Intelligence Definition - European Commission

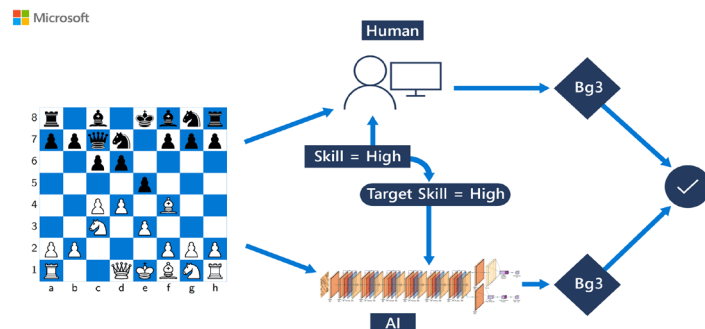
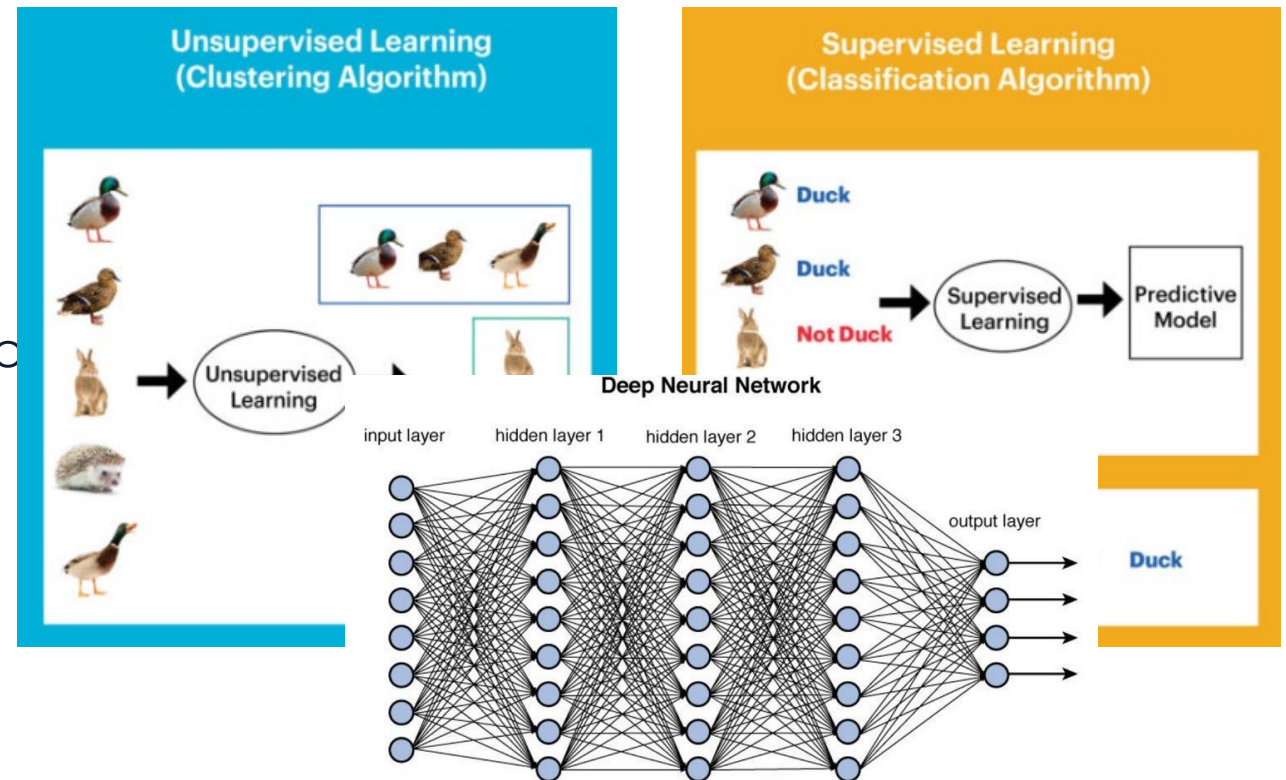
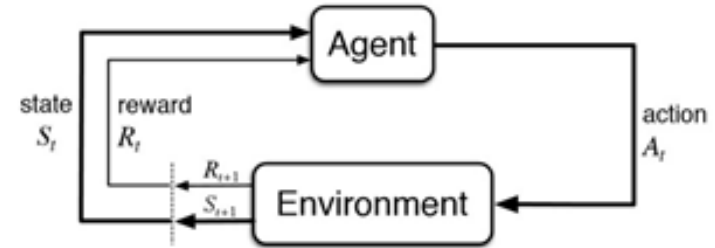
*Artificial intelligence refers to systems that display intelligent behaviour by analysing their environment and taking actions – **with some degree of autonomy** – to achieve specific goals.*

AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).



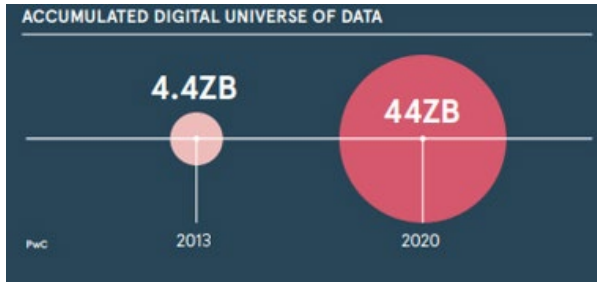
Machine Learning

- Machine learning → “a method of designing a sequence of actions **to solve a problem** that optimises automatically through experience and with limited or **no human intervention**” (FSB, 2017)
- Categories of machine learning:
 - Supervised machine learning**
 - Unsupervised machine learning**
 - Reinforcement learning**
 - Deep Learning**
- Few decades ago chess playing was a



What is Artificial Intelligence?

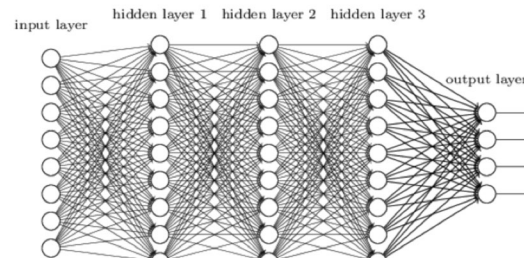
Data



A day in data

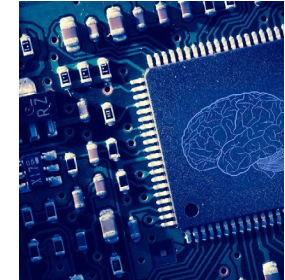
- 65 billion Whatsapp messages
- 10^{15} bytes generated by Facebook
- 500m tweets

The Mathematics



- Machine Learning
- Neural networks
- Numerical optimizations

Computing power



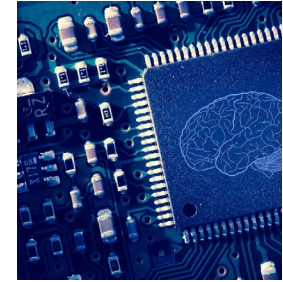
- 10^{21} FLOPS (floating point operations per second) globally available
- 10^{12} , one trillion, is 80 times the global GDP
- Cost of 1 GFLOP
 - 1945: 1800 trillion USD
 - 2000: 1500 USD
 - 2020: 0.04 USD

What is Artificial Intelligence?

You are Here

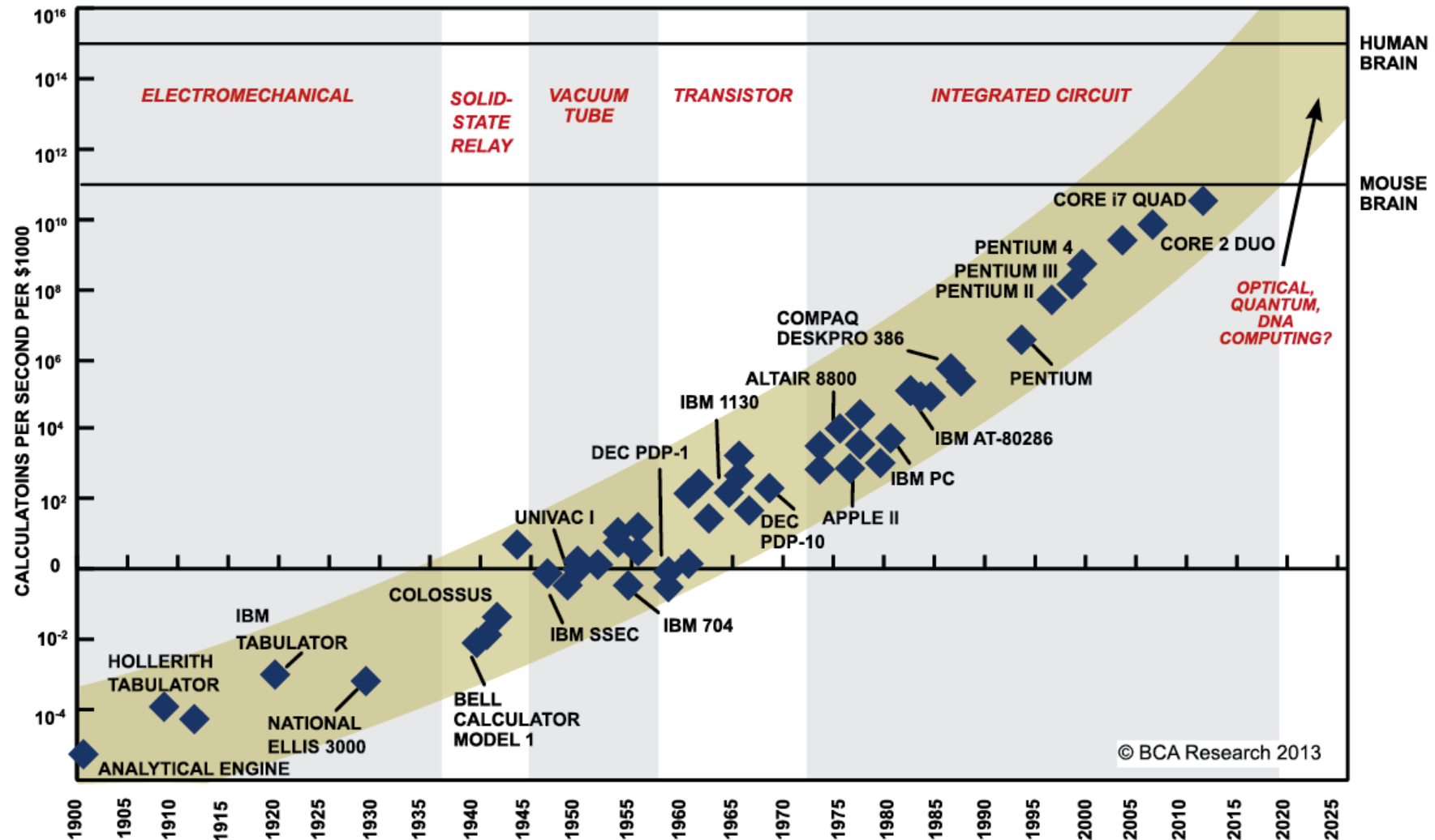
© 1997 Jerry Lodriguss

Computing power



- 10^{21} FLOPS (floating point operations per second) globally available
- 10^{12} , one trillion, is 80 times the global GDP
- Cost of 1 GFLOP
 - 1945: 1800 trillion USD
 - 2000: 1500 USD
 - 2020: 0.04 USD

Computing power



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, THE VIKING PRESS, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devices – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter



4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020

294bn

billion emails are sent

Radicati Group

3.9bn

people use emails

4TB

of data produced by a connected car

Intel

ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB

2013

44ZB

2020

PwC

DEMYSTIFYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

| Unit | Value | Size |
|---------------------|--------------------------|---|
| b bit | 0 or 1 | 1/8 of a byte |
| B byte | 8 bits | 1 byte |
| KB kilobyte | 1,000 bytes | 1,000 bytes |
| MB megabyte | 1,000 ² bytes | 1,000,000 bytes |
| GB gigabyte | 1,000 ³ bytes | 1,000,000,000 bytes |
| TB terabyte | 1,000 ⁴ bytes | 1,000,000,000,000 bytes |
| PB petabyte | 1,000 ⁵ bytes | 1,000,000,000,000,000 bytes |
| EB exabyte | 1,000 ⁶ bytes | 1,000,000,000,000,000,000 bytes |
| ZB zettabyte | 1,000 ⁷ bytes | 1,000,000,000,000,000,000,000 bytes |
| YB yottabyte | 1,000 ⁸ bytes | 1,000,000,000,000,000,000,000,000 bytes |

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook



Searches made a day

5bn

Searches made a day from Google

3.5bn

Smart Insights



463EB

of data will be created every day by 2025

idc

95m

photos and videos are shared on Instagram

Instagram Business



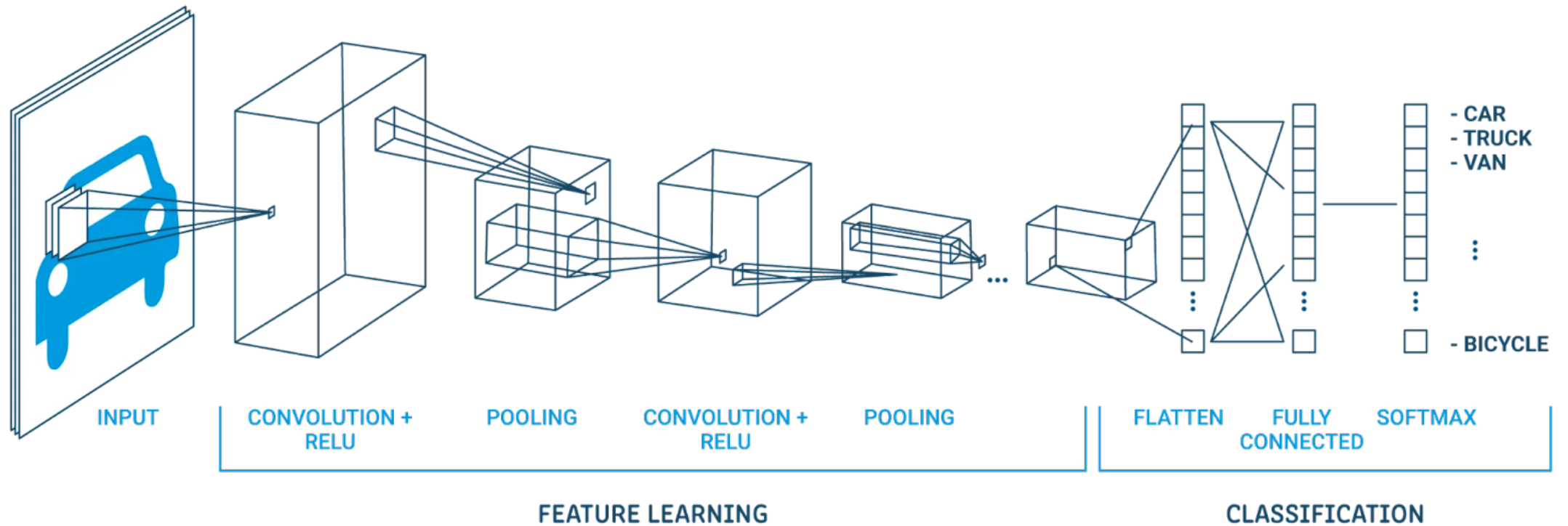
28PB

to be generated from wearable devices by 2020

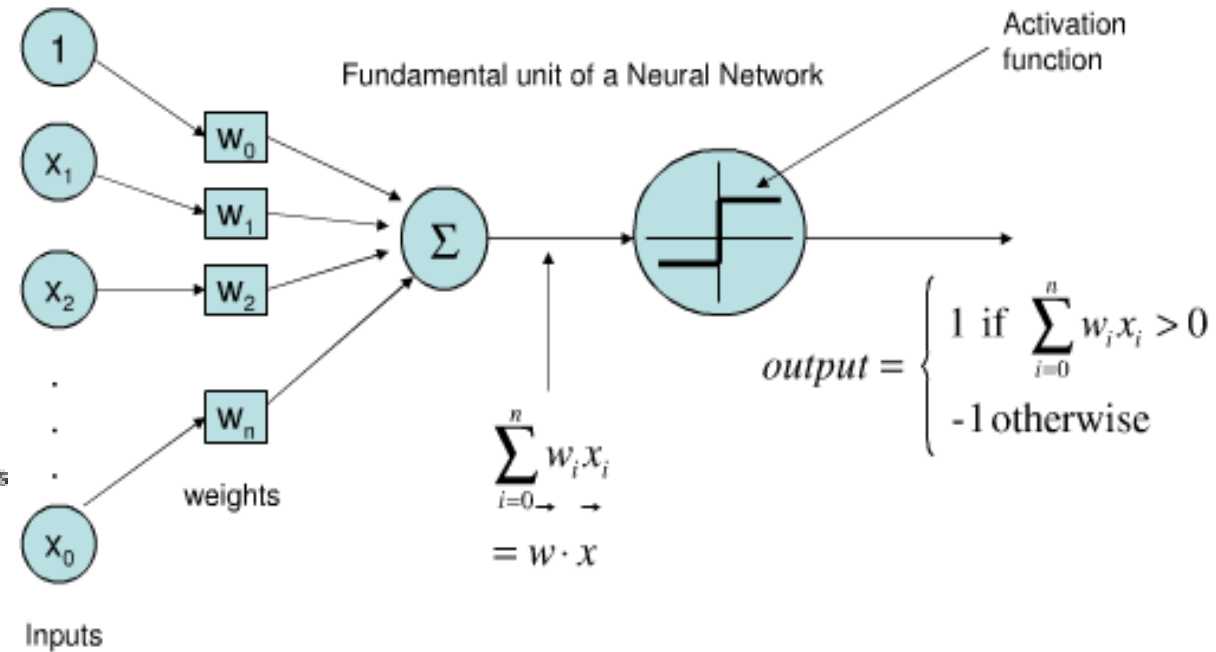
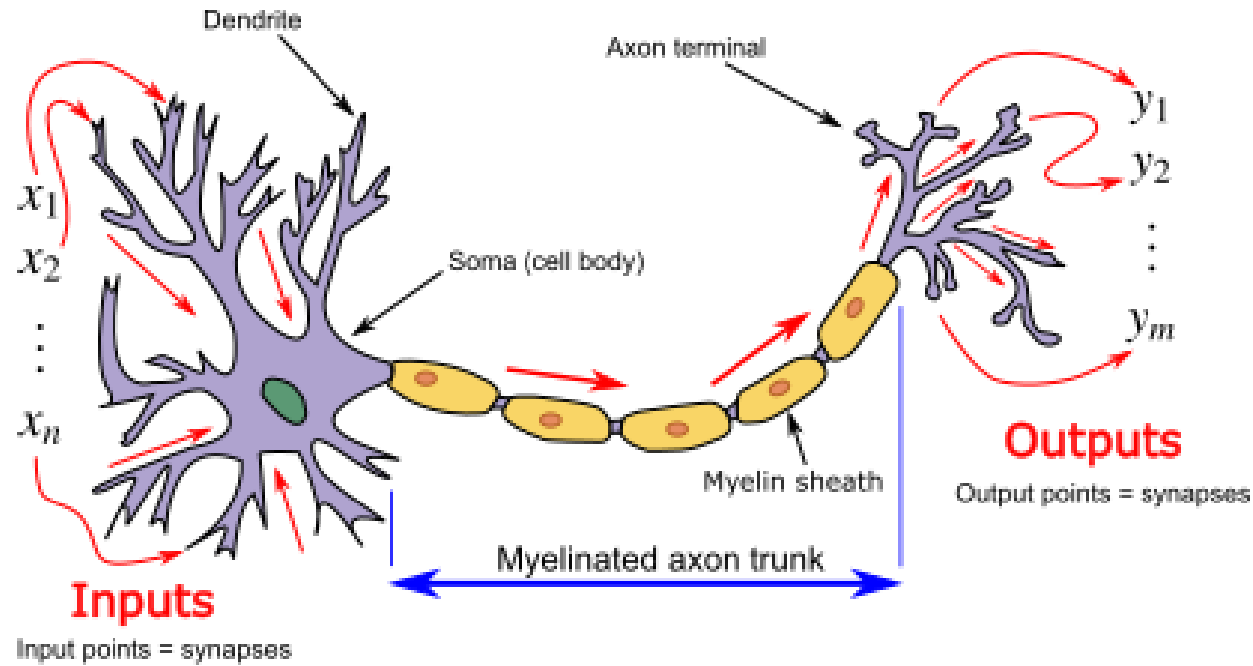
Statista



Artificial Intelligence techniques – The Mathematics



The Mathematics



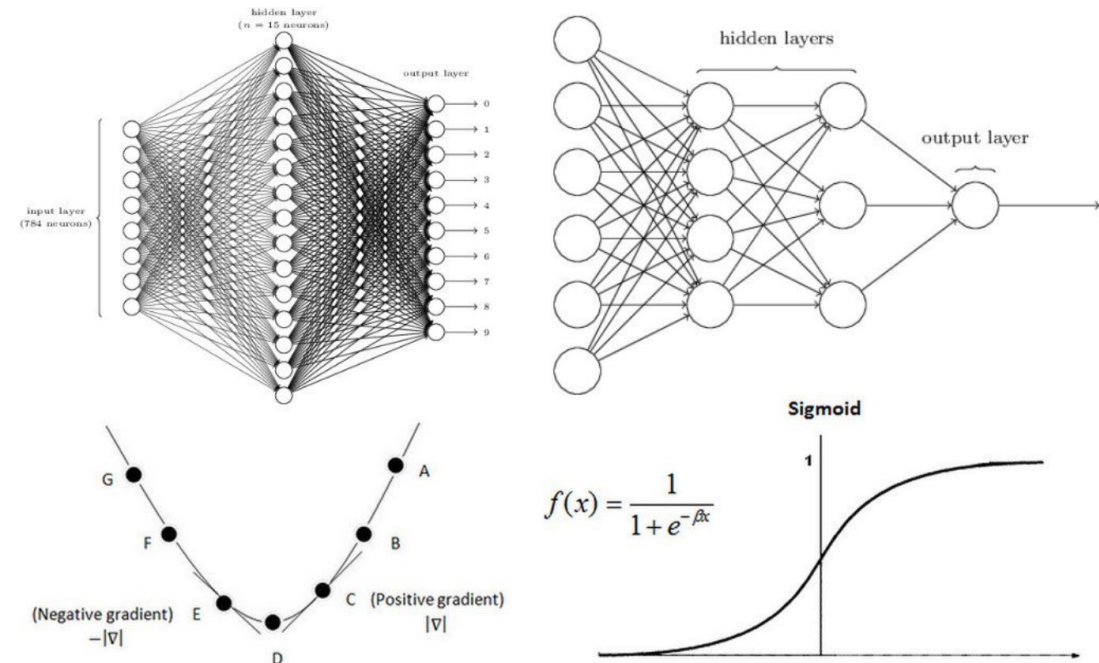
$$f: \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Neural networks (Artificial Intelligence) are functions

Neural networks and the universal approximation theorem

Neural networks can approximate (almost) arbitrary mathematical functions

Cybenko (1989) states that any continuous mathematical function on a compact domain can be **approximated with any precision** by an appropriate neural network with sufficient width and depth

















Neural networks are the most powerful functions we have ever had

Neural networks

A mostly complete chart of Neural Networks

©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org

-  Input Cell
-  Backfed Input Cell
-  Noisy Input Cell
-  Hidden Cell
-  Probabilistic Hidden Cell
-  Spiking Hidden Cell
-  Capsule Cell
-  Output Cell
-  Match Input Output Cell
-  Recurrent Cell
-  Memory Cell
-  Gated Memory Cell
-  Kernel
-  Convolution or Pool

Perceptron (P)



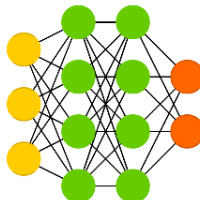
Feed Forward (FF)



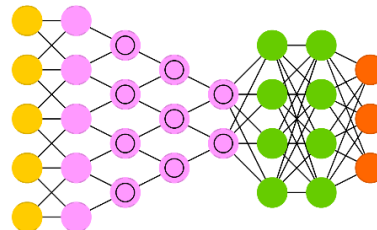
Radial Basis Network (RBF)



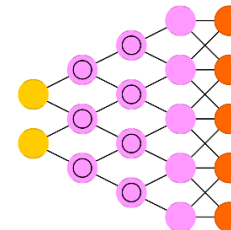
Deep Feed Forward (DFF)



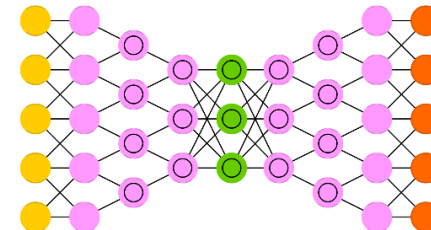
Deep Convolutional Network (DCN)



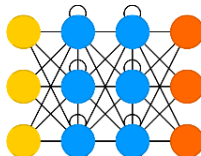
Deconvolutional Network (DN)



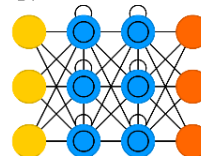
Deep Convolutional Inverse Graphics Network (DCIGN)



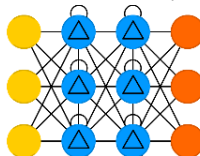
Recurrent Neural Network (RNN)



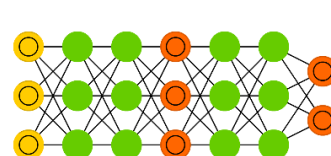
Long / Short Term Memory (LSTM)



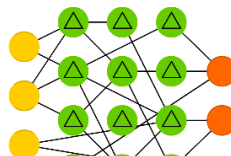
Gated Recurrent Unit (GRU)



Generative Adversarial Network (GAN)



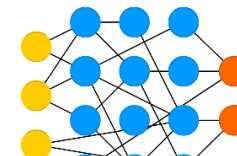
Liquid State Machine (LSM)



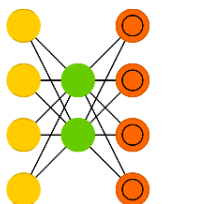
Extreme Learning Machine (ELM)



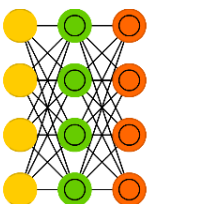
Echo State Network (ESN)



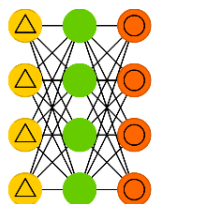
Auto Encoder (AE)



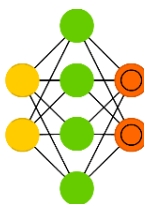
Variational AE (VAE)



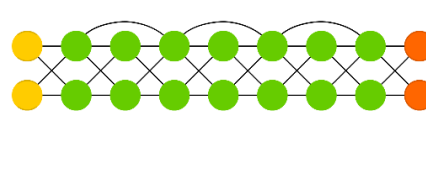
Denoising AE (DAE)



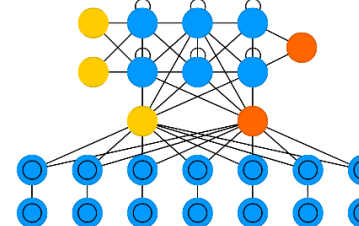
Sparse AE (SAE)



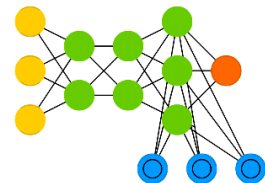
Deep Residual Network (DRN)



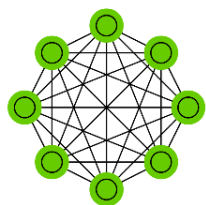
Differentiable Neural Computer (DNC)



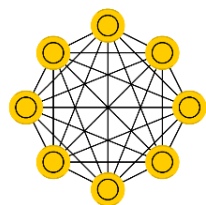
Neural Turing Machine (NTM)



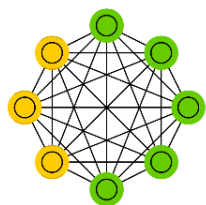
Markov Chain (MC)



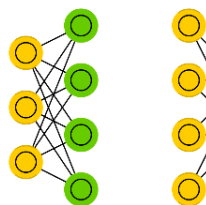
Hopfield Network (HN)



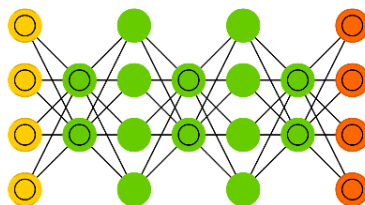
Boltzmann Machine (BM)



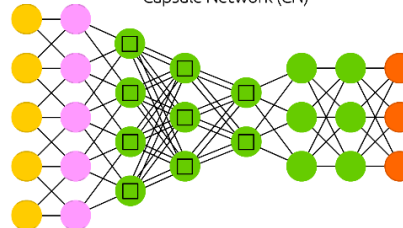
Restricted BM (RBM)



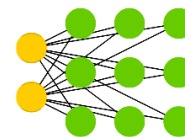
Deep Belief Network (DBN)



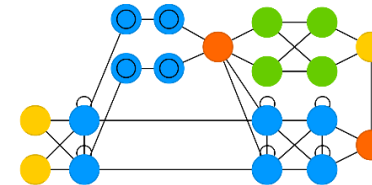
Capsule Network (CN)



Kohonen Network (KN)



Attention Network (AN)





History of Artificial Intelligence

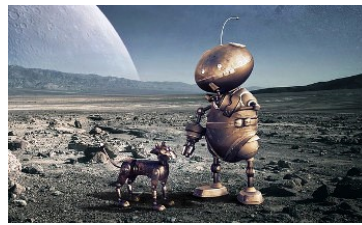
Antiquity

Greek myths
Sacred
mechanical
statues built
in Egypt and
Greece were
believed to
be capable
of wisdom
and emotion.



Symbolic AI 1956 – 1974

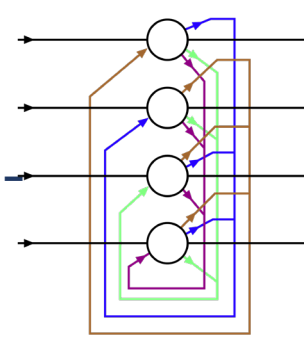
"Machines
will be
capable,
within twenty
years, of
doing any
work a man
can do."



Boom 1980 1987

The rise of
expert
systems

The money
returns: the
Fifth
Generation
project



AI 1993 – 2011

Milestones
and Moore's
law

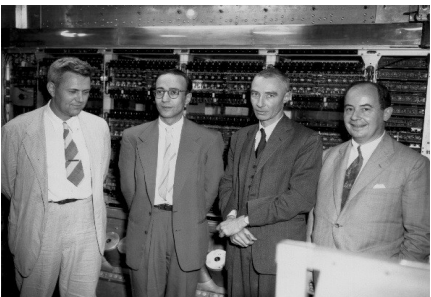
Intelligent
agents

AI behind the
scenes - AI
had solved a
lot of very
difficult
problems



The birth of AI 1952 – 1956

Turing's test
Dartmouth
Workshop
1956



The first AI winter 1974 – 1980

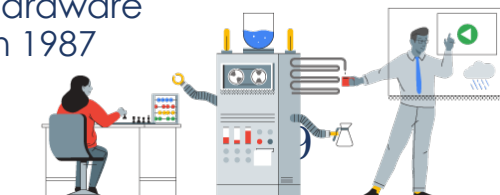
Limited
computer power
Intractability and
the
combinatorial
explosion

The end of
funding

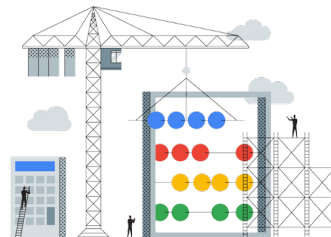


Bust: the second AI winter 1987 – 1993

Sudden
collapse of
the market
for
specialized
AI hardware
in 1987



Deep learning, big data and artificial intelligence: 2011 – present



AI Coins



A GLOBAL LOOK AT R&D SPENDING

The companies and nations that are leading the way in innovation and research



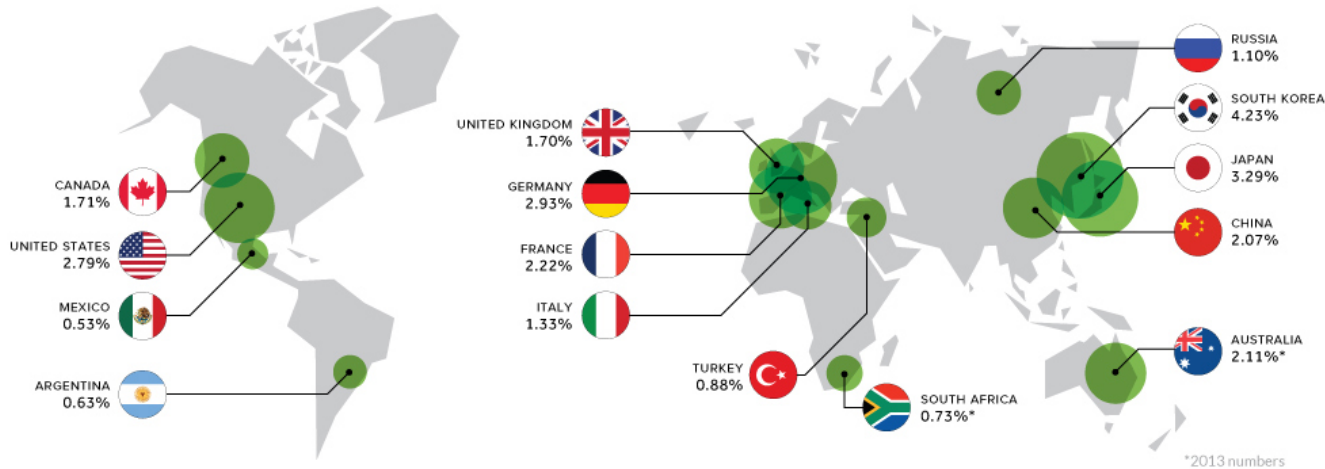
The G20 accounts for **92%** of global spending on research



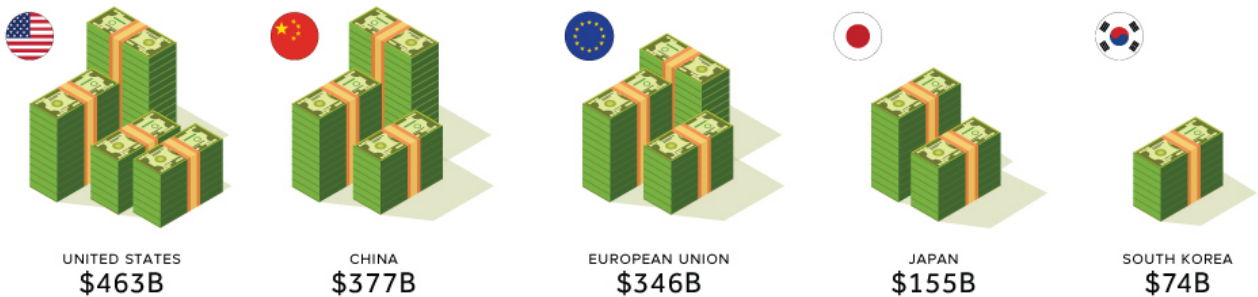
94% of patents granted by the US Patent and Trademark Office stem from G20 countries

R&D Expenditure as a percentage of GDP

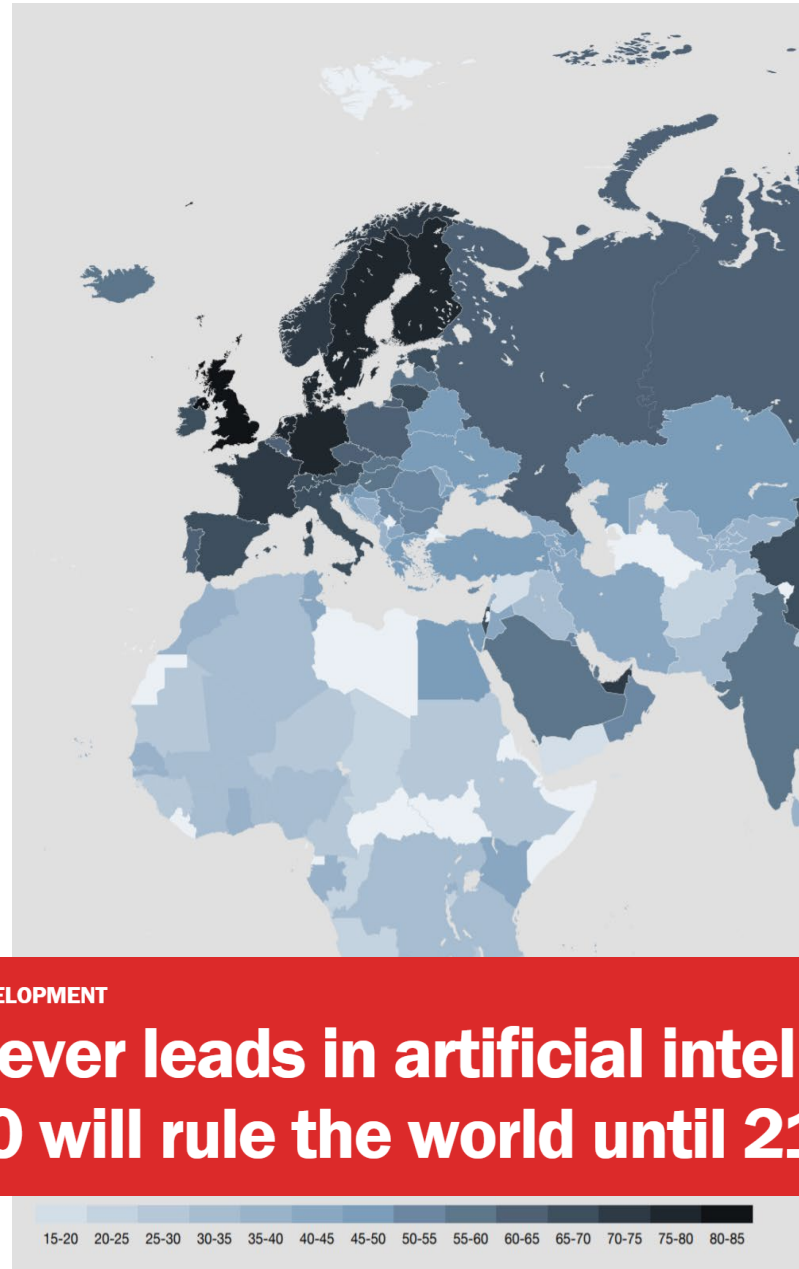
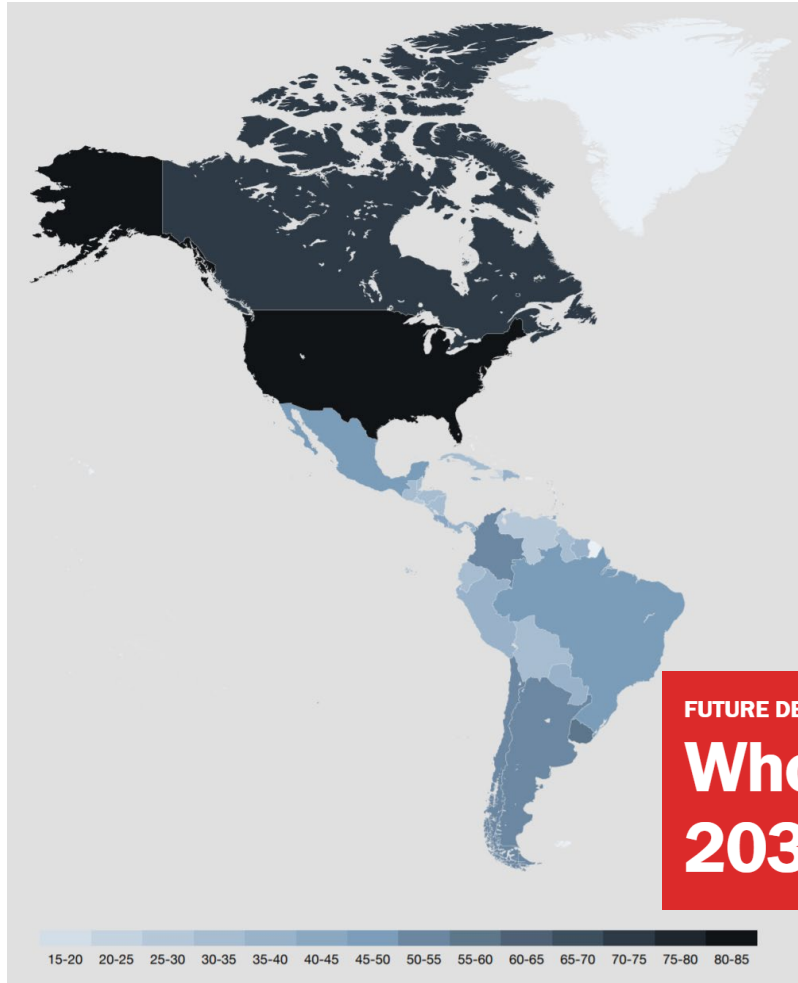
Select G20 countries; 2015



Top 5 Jurisdictions by R&D Expenditure (2015)



AI Readiness



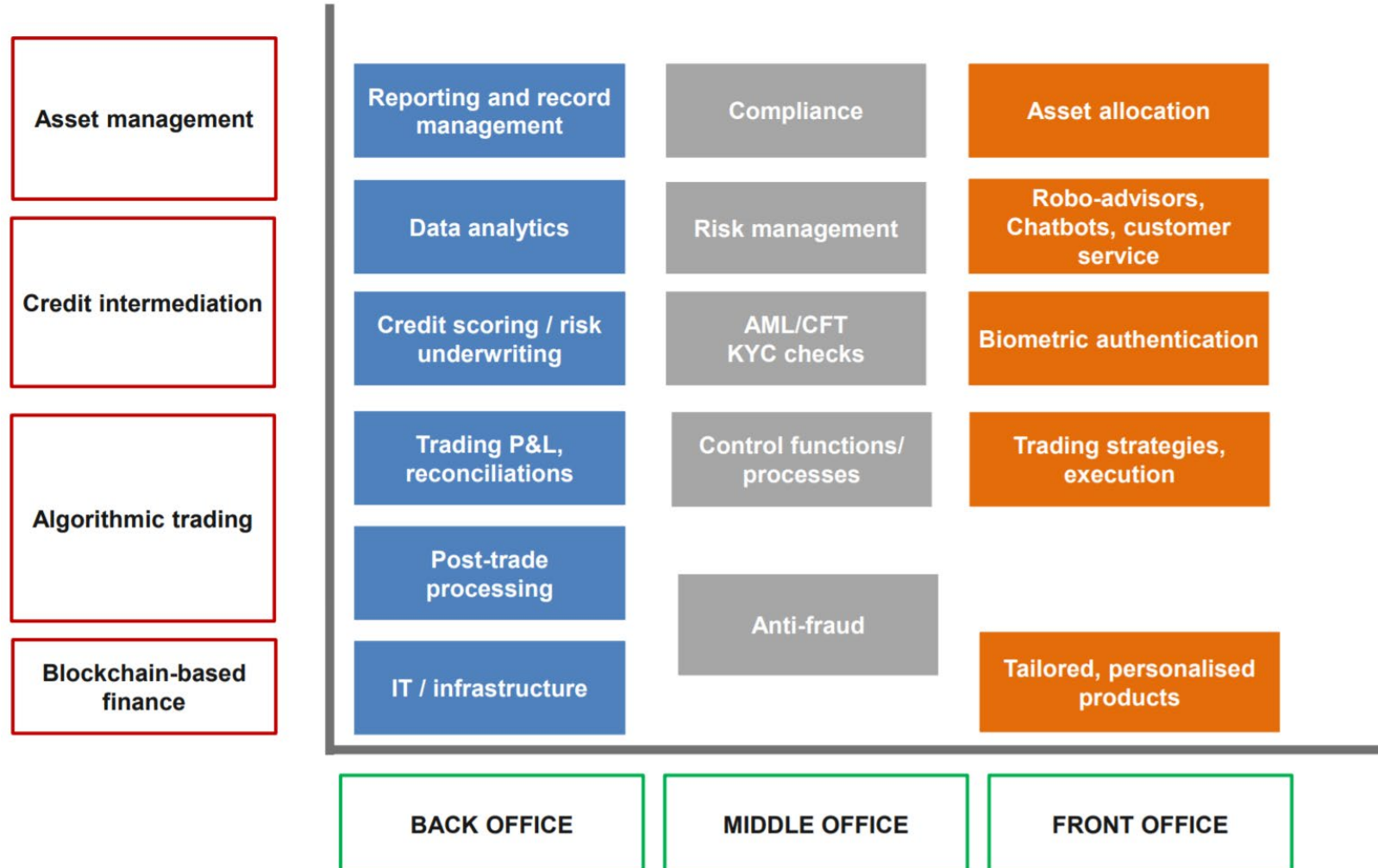
FUTURE DEVELOPMENT

Whoever leads in artificial intelligence in 2030 will rule the world until 2100



A biased tour through AI in Finance Research: Data Science, Fintech and Blockchain Technology

Artificial Intelligence is impacting all business areas



Source: OECD staff illustration.

A biased tour through AI research in Finance

Fintech and Risk Management



100+ researchers from 15 European Universities

- Detecting Fraud in Blockchain payments
- Peer-to-peer lending
- Explainable AI
- Credit Risk Network models

Fintech and AI in Finance



200+ researchers from 38 countries



Reinforcement learning for Finance



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Innosuisse – Swiss Innovation Agency

- Reinforcement learning for trading and forecasting financial markets

frontiers
in Artificial Intelligence

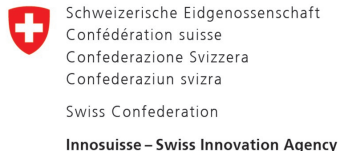
REVIEW article
Front. Artif. Intell., 31 May 2021 | <https://doi.org/10.3389/frai.2021.668465>

The Applicability of Self-Play Algorithms to Trading and Forecasting Financial Markets

Jan-Alexander Posth^{1*}, Piotr Kotlarz^{2,3}, Branka Hadji Misheva², Joerg Osterrieder^{2,4} and Peter Schwendner¹

A biased tour through AI research in Finance II

Credit risk models



- Towards Explainable Artificial Intelligence and Machine Learning in Credit Risk Management

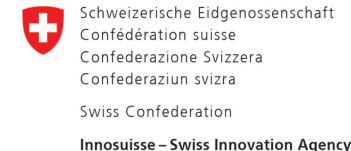
Peer-to-peer lending



Network-based credit risk models

- Network-based feature extraction techniques
- The use of multiple networks in feature extraction
- bagging and hyper-parameter tuning

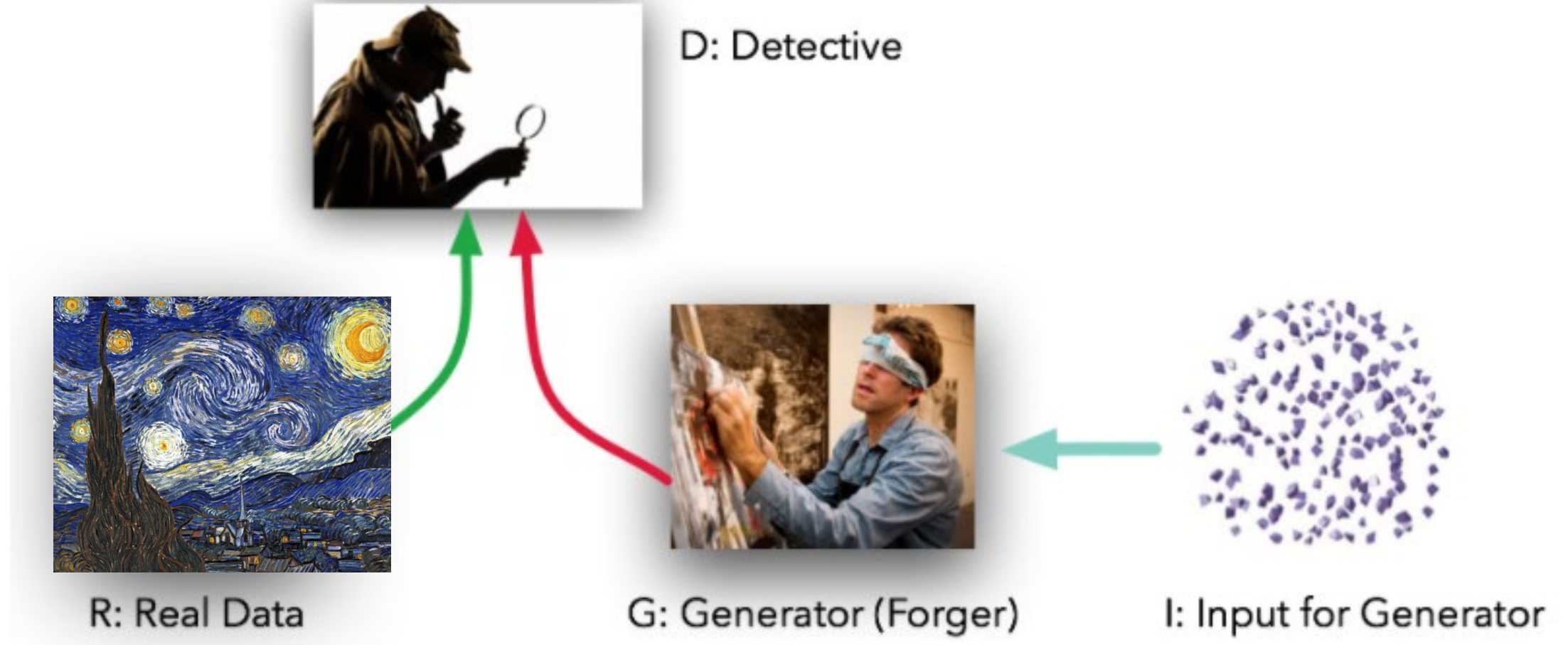
Reinforcement learning for Finance



- Limit order book case study
- Recommender systems
- Factor investing
- Multi-agent Reinforcement Learning

A biased tour through AI research in Finance III

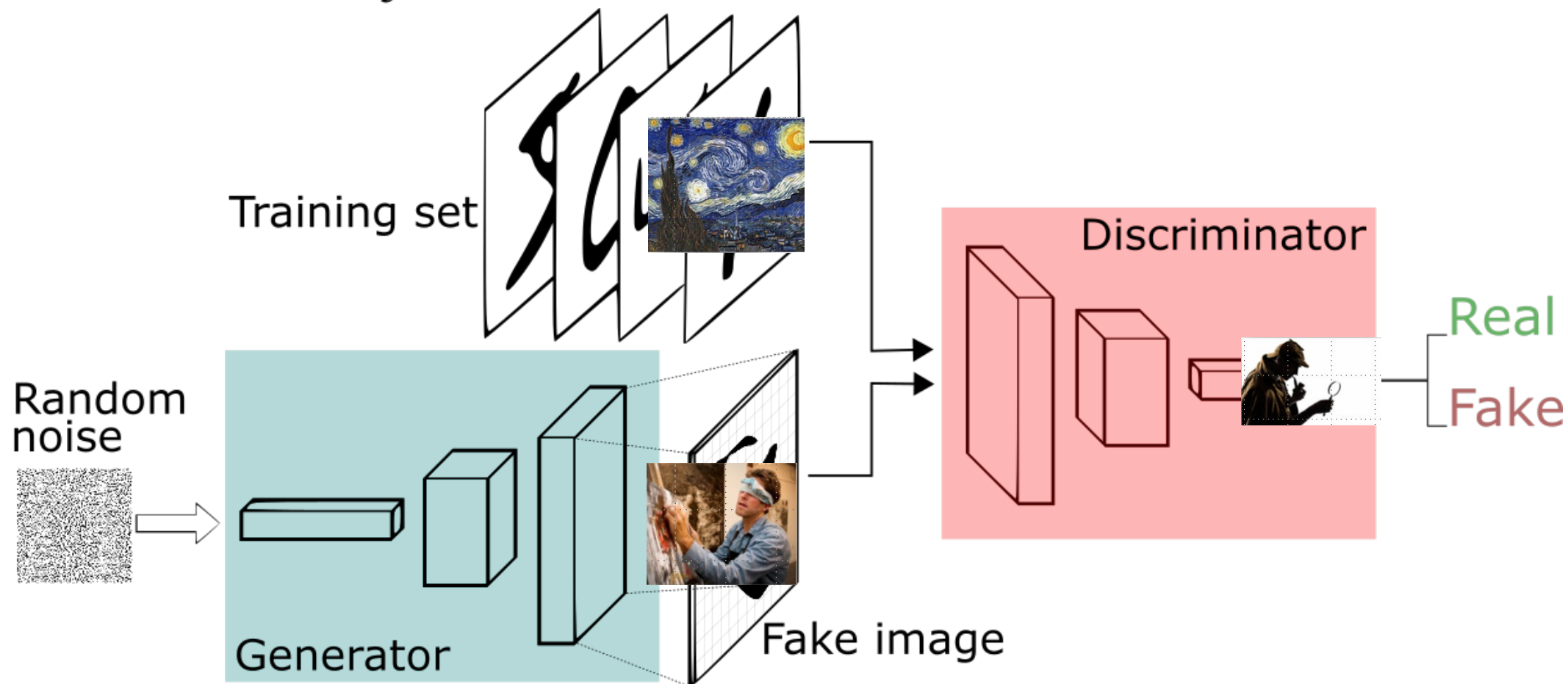
Generative Adversarial Networks



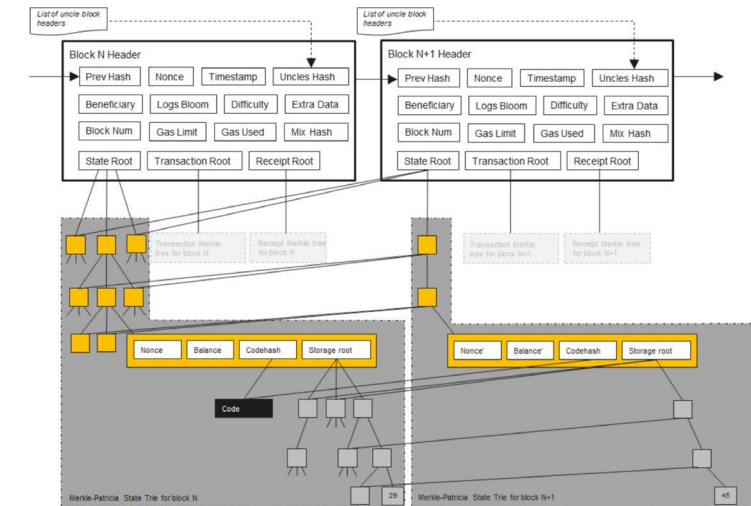
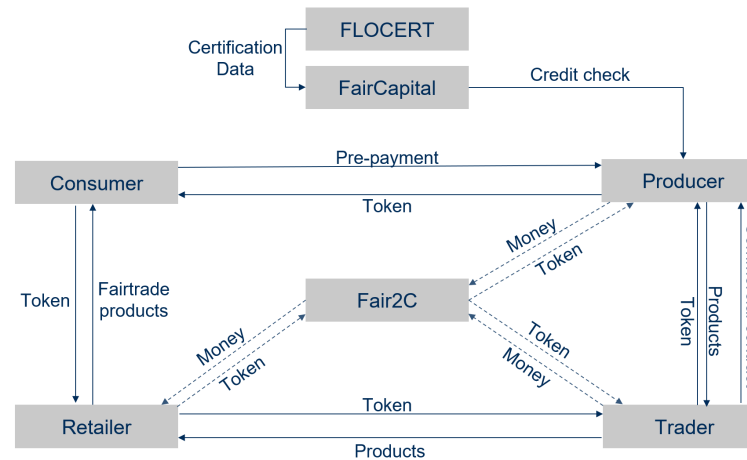
A biased tour through AI research in Finance IV

Generative Adversarial Networks

$$\min_{\mathcal{G}} \max_{\mathcal{D}} E_x [\log(\mathcal{D}(x))] + E_z [\log(1 - \mathcal{D}(\mathcal{G}(z)))]$$



Electronic Signature on the Blockchain



The AI framework

Input Data

Feature Engineering

Feature Creation

Encoding

Feature Selection

Embedding

PCA

Kernel PCA

AutoEncoder

Prediction

Time Series Forecasting

Forecast with Alternative Data

Hyper-parameter Optimization

Neural Architecture Search

AutoML

Ranking

ML-based Ranking

Empirical Ranking

LIME

SHAPLEY

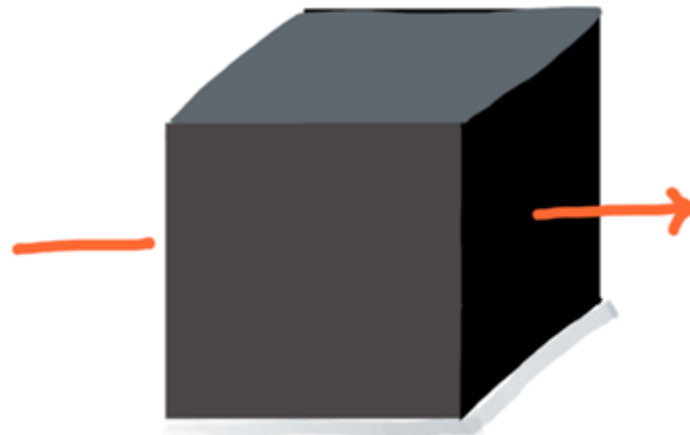
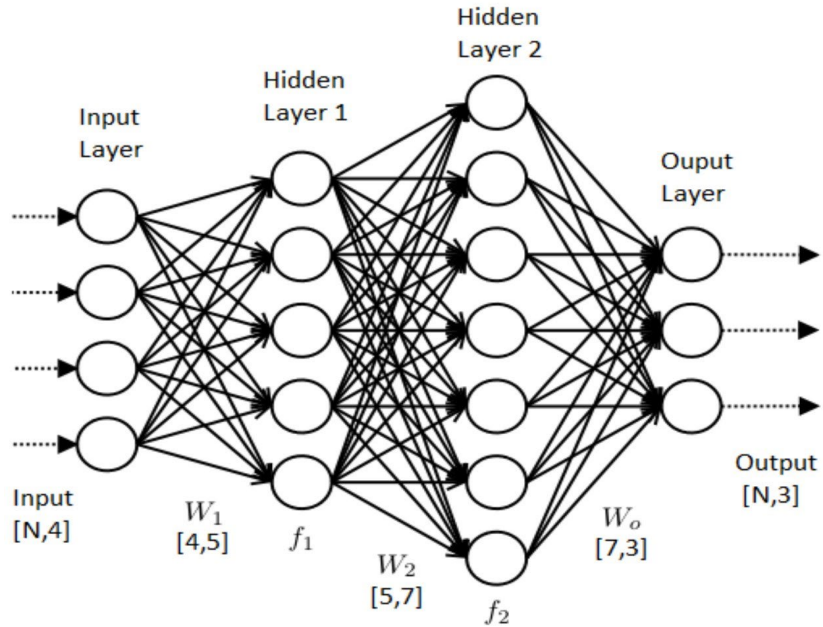
Explanation



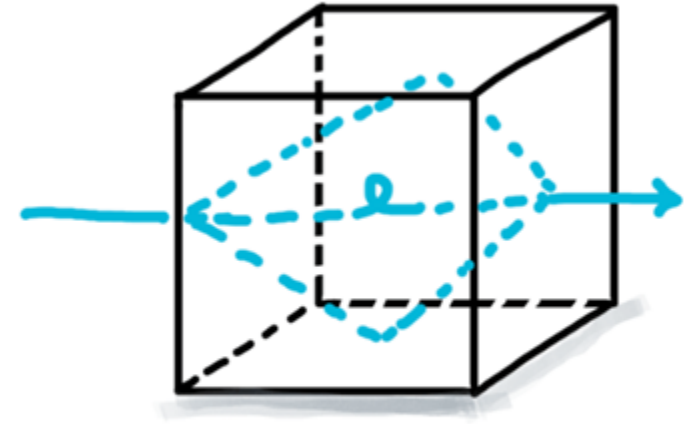
Regulatory aspects

The Need for eXplainable AI

It is not clear how variables are being combined to make predictions!



BLACK BOX AI



EXPLAINABLE AI

The Need for eXplainable AI

It is not clear how variables are being combined to make predictions!



It has found some snow!

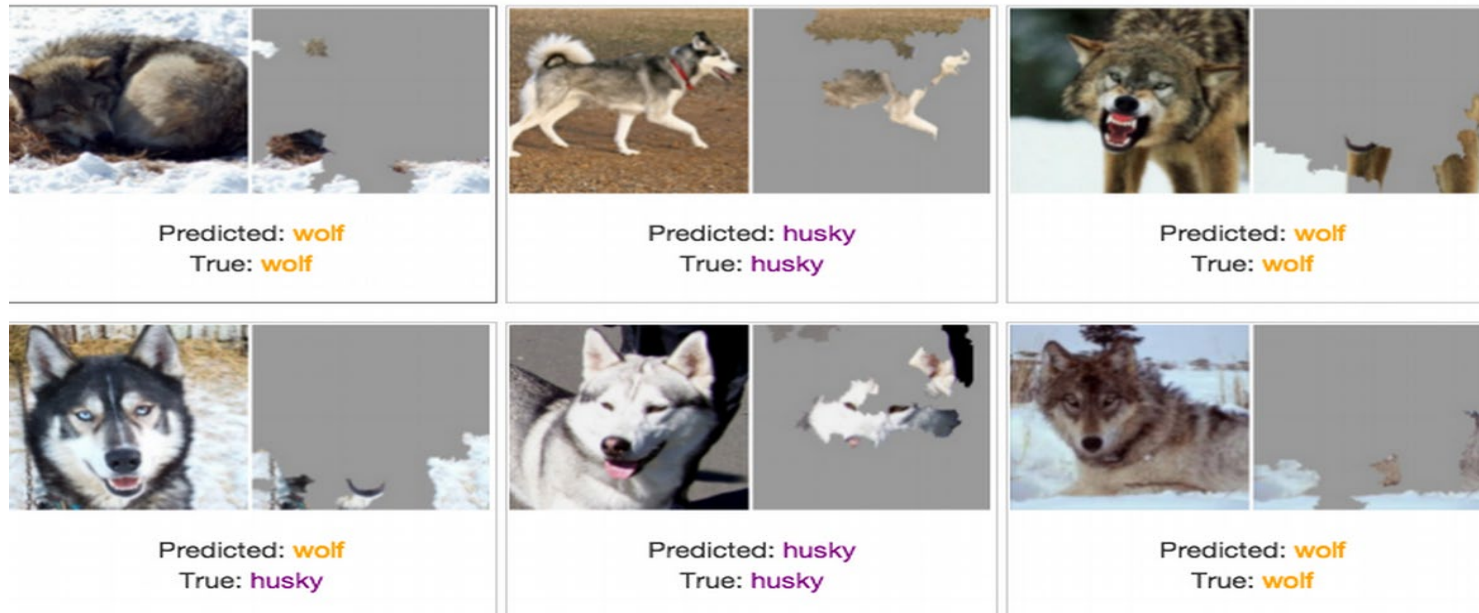


Image source: medium.com

Seven key requirements for AI systems



The European approach to trustworthy AI

Unacceptable risk

- Clear threat to the safety, livelihoods and rights of people
- Systems or applications that manipulate human behaviour to circumvent users' free will
- 'social scoring' by governments

High-risk

- Critical infrastructures
- Educational or vocational training
- Safety components of products
- Employment, workers management and access to self-employment
- Essential private and public services
- Law enforcement
- Migration, asylum and border control management
- Administration of justice and democratic processes

Limited risk

- AI systems with specific transparency obligations
- Users should be aware that they are interacting with a machine so they can take an informed decision to continue or step back

Minimal risk

- Free use of applications such as AI-enabled video games or spam filters
- The draft Regulation does not intervene here, as these AI systems represent only minimal or no risk for citizens' rights or safety

Coordinated Plan on AI

- Funding allocated through the Digital Europe and Horizon Europe programmes, the Recovery and Resilience Facility and Cohesion Policy programmes
- Creation of enabling conditions for AI's development
- Foster AI excellence
- Ensure that AI works for people
- Build strategic leadership

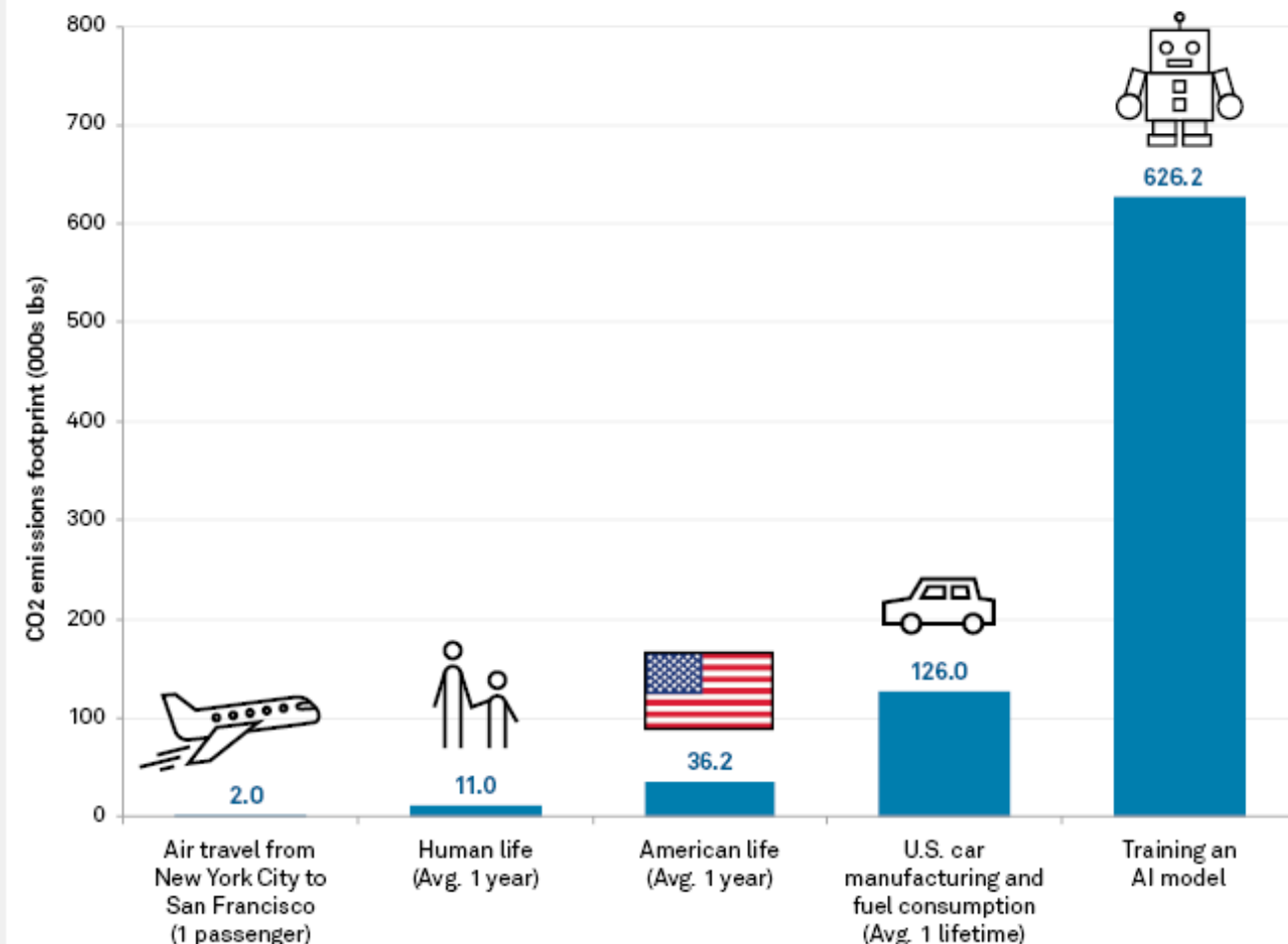




Quantum Computing

AI and sustainability

CO2 emission benchmarks



Data compiled Oct. 9, 2019.
An "American life" has a larger carbon footprint than a "Human life" because the U.S. is widely regarded as one of the top carbon dioxide emitters in the world.
Source: College of Information and Computer Sciences at University of Massachusetts Amherst

In 2030, using AI for climate control could help reduce

2.6 to 5.3 gigatons

of GHG emissions, or 5% to 10% of the total

and could provide

\$1 trillion to \$3 trillion

in value added when applied to corporate sustainability generally

Source: BCG analysis.

How quantum computing could change financial services

In a Historic Milestone, Silicon Quantum Computing Just Exceeded 99% Accuracy

JP Morgan Chase Unleashes Honeywell's Quantum Computer on Tough Fintech Problems

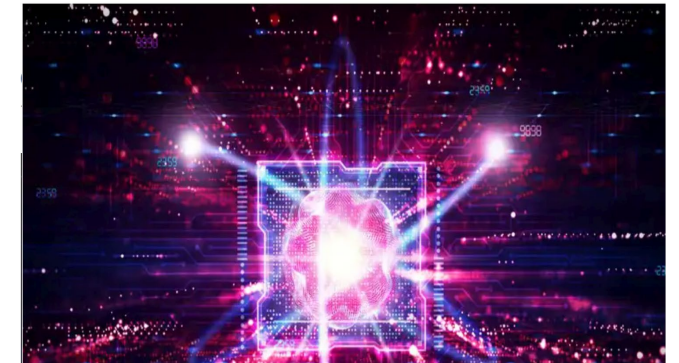
Quantum Computing in Banking and Finance
– Threat or Opportunity?

Quantum Computing Is Coming. What Can It Do?

by Francesco Bova, Avi Goldfarb, and Roger Melko

July 16, 2021

First fully programmable quantum computer based on neutral atoms

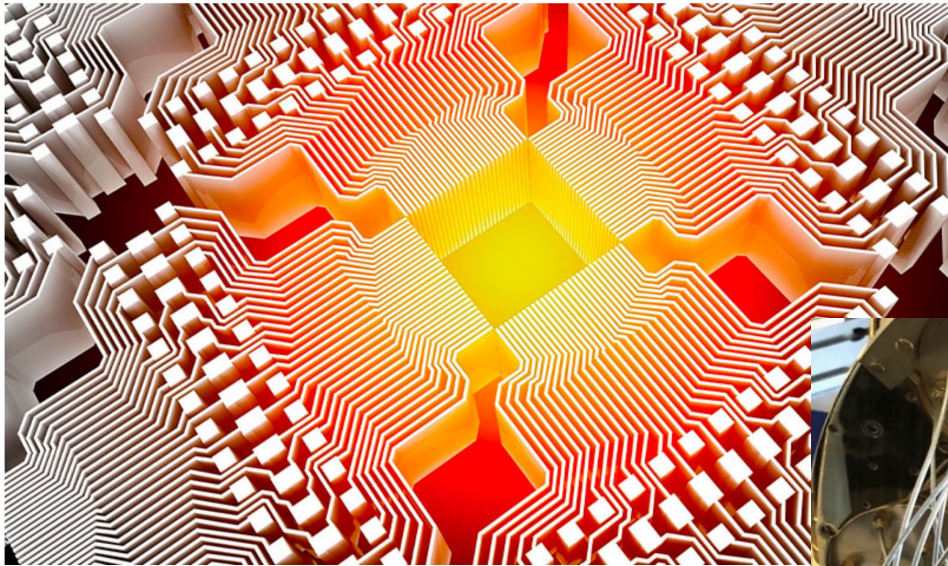


Bloomberg

Markets | Markets Magazine

Quantum Computing Might Be Here Sooner Than You Think

The potential of quantum computing for finance



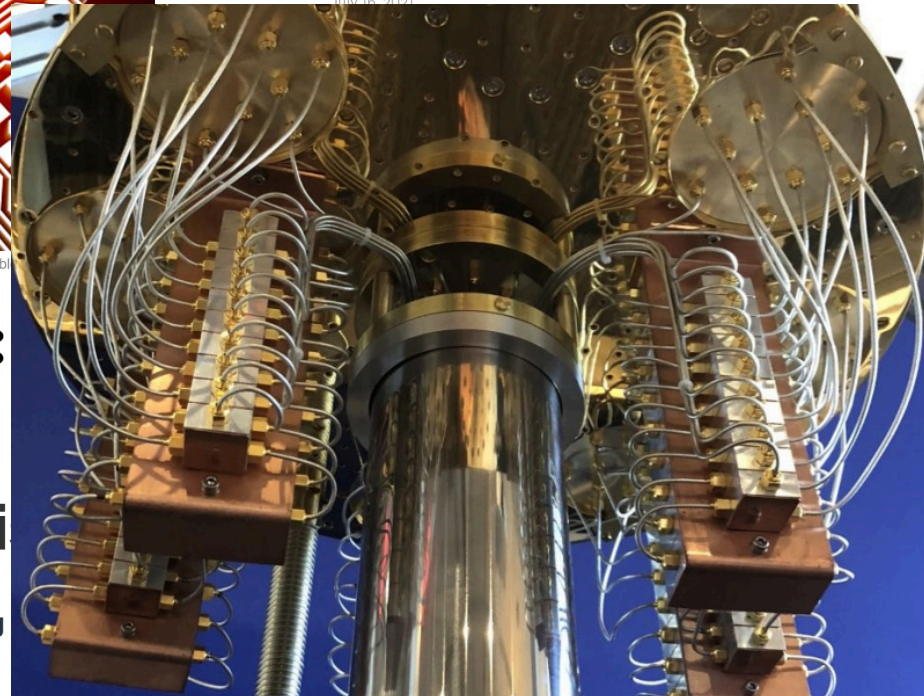
The first, and most important, is confirmation that quantum computers will be able to deliver processing power on an unimaginable scale.

Nuclear quantum computing:

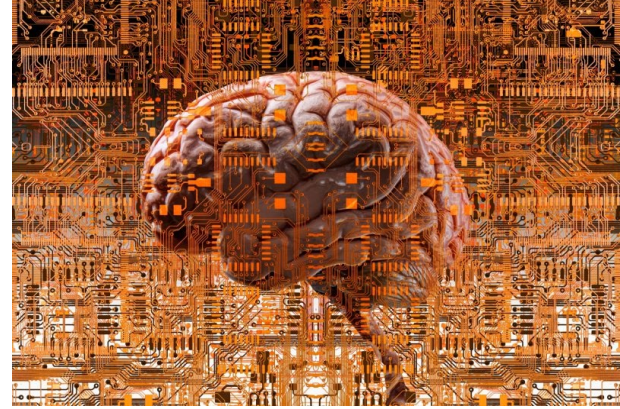
Brought to you by the US Army

Finland moves to industrialize quantum computing

For the Finnish government, now is the time to start preparing quantum computers will have practical value



The Future of Artificial Intelligence



- *“The development of full artificial intelligence could spell the end of the human race....It would take off on its own, and re-design itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded.”— Stephen Hawking*
- *“Artificial intelligence would be the ultimate version of Google. The ultimate search engine that would understand everything on the web. It would understand exactly what you wanted, and it would give you the right thing. We're nowhere near doing that now. However, we can get incrementally closer to that, and that is basically what we work on.” —Larry Page*
- *“Artificial intelligence will reach human levels by around 2029. Follow that out further to, say, 2045, we will have multiplied the intelligence, the human biological machine intelligence of our civilization a billion-fold.” —Ray Kurzweil*

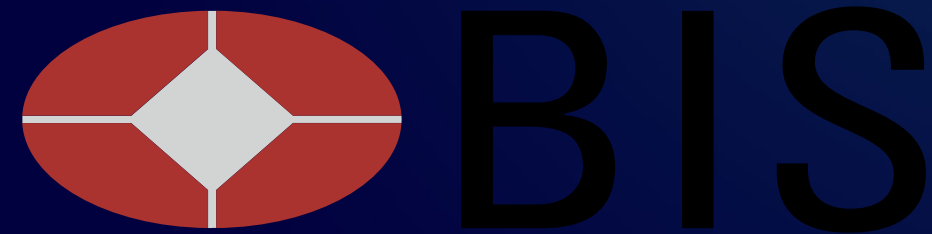
Artificial Intelligence in Finance Quo Vadis?



IFC workshop on Data science in central banking:
Applications and Tools
14 – 17 February, 2022

Irving Fisher Committee on Central Bank Statistics

Prof. Dr. Jörg Osterrieder



IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Swimming in the data lake: an application to NielsenIQ Homescan¹

Minnie H Cui, Gene (Fa Gui) Jiang and Botlhale Mosweu,
Bank of Canada

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Swimming in the Data Lake

An application to NielsenIQ HomeScan

Minnie H. Cui^{Ψ*}, Gene (Fa Gui) Jiang^{*} & Botlhale Mosweu^{*}

Abstract

We demonstrate the functionality of Azure Data Lakes and cloud computing tools for large data set processing through the case of NielsenIQ HomeScan, a newly onboarded data set at the Bank of Canada. With already more than 61 million observations total from over 40 thousand household panelists, we selected a Data Lake storage solution with a final destination in a Microsoft Azure SQL Database to maintain ongoing quarterly deliveries and to access Azure-based cloud-computing resources. We illustrate the capabilities of Data Lakes for securely and cost-efficiently storing and backing-up data from multiple sources of various formats, structured or unstructured. This flexibility of data formats allows users to apply transformations to data only when needed. Users have the choice of Azure Databricks or various ported software, including JupyterLab, Stata, and R, for analytics. Finally, we show that Databricks significantly reduces runtimes by running on clusters which parallelizes certain processes automatically and auto-scales worker nodes as needed. In our experimentation, Databricks ran 40 times faster than Stata on desktop.

Keywords: cloud computing, Data Lake, Databricks, data science, large data sets

JEL classification: C55, C88

^Ψ Corresponding author: minniecui@bankofcanada.ca.

^{*} Bank of Canada, 234, Wellington Ave., Ottawa ON, K1A 0G9, Canada. The views presented are the authors' and do not represent the official views of the Bank of Canada.

Contents

| | |
|--|----|
| Swimming in the Data Lake..... | 1 |
| An application to NielsenIQ HomeScan..... | 1 |
| 1. Introduction..... | 3 |
| 2. Cloud Storage & Computing for Central Banking | 3 |
| Security..... | 3 |
| Memory Limits..... | 4 |
| Computing Limits..... | 4 |
| Analytical Tools | 5 |
| 3. Data Lake Pipeline..... | 5 |
| NielsenIQ HomeScan Pipeline | 5 |
| Notable Features of the Pipeline | 6 |
| Modularity | 6 |
| Performance Gains by Parallel Processing | 7 |
| Data quality enhancements via automated data validation | 8 |
| Maintaining data security during the data validation process | 8 |
| 4. Analysis & Performance..... | 8 |
| Loading data into the Data Lake..... | 8 |
| Analytical performance | 9 |
| 5. Conclusion..... | 10 |

1. Introduction

As the developed world becomes increasingly connected to digital goods and services, the role of technology in the economy becomes increasingly more important. Observational data sets are now available at much higher frequency and larger scale than traditional surveys. This availability has created opportunities for economists and policymakers to examine economic systems with greater accuracy. However, new methods and technologies are required to take full advantage of these data sets *efficiently*.

Cloud-based technology can significantly improve computational capabilities while bringing cost-efficient storage solutions to large organizations. Azure Data Lakes, for example, can securely store and back-up data of multiple formats, structured or unstructured. More importantly, analytical tools like Databricks are integrated with Azure Data Lakes, reducing data movement while delivering significant performance gains.

In this paper, we discuss our implementation of an Azure Data Lake for the storage and analysis of Canadian NielsenIQ HomeScan data. We discuss an automated data pipeline that handles ingestion, validation, processing, and delivery. We also discuss its benefits for security, memory and computational limits. Finally, we show that in testing, Databricks delivered runtimes that are 40 times faster than Stata on desktop for analysis through cluster auto-scaling and parallelization.

2. Cloud Storage & Computing for Central Banking

When working with larger data sets like NielsenIQ HomeScan with security concerns and scheduled deliveries, certain traditional practices have become outdated as storage and security needs become increasingly important. To eliminate some of these concerns, the Bank of Canada is piloting cloud-based solutions in Microsoft Azure.

Security

When dealing with protected data, traditionally, research groups at the Bank secured these data sets on various drives on the Bank's servers with special permissions and/or on the Bank's High Performance Computing (HPC) clusters. Researchers with approved access were given reading permissions to prevent accidental overwriting of raw data. With these permissions, researchers can copy out data to their local drives for analysis.

Without compromising security protocols, Azure provides researchers with single sign-on to access data, compute resources and other services¹. In comparison, prior to the transition to Azure, researchers were required to maintain multiple credentials to be authenticated in multiple systems. In essence, by utilizing the Azure Active Directory (AD) service, Azure provides researchers with a hassle-free experience in accessing the data while protecting against security breaches.

¹ Murray, D. & Omondi, J. (12 May 2021). "[What is single sign-on?](#)" *Microsoft Azure*.

Since NielsenIQ HomeScan has a Protected-B security status, we put in place multiple measures to ensure only authorized personnel have the right access. Azure has a rich set of measures to ensure data security. On a high level, we utilize AD to authorize and authenticate researchers. For storage, we employ the Role-Based Access Control (RBAC) to control the access level to data containers, and the Access Control List (ACL) to folders and files under the containers. For other Azure services used by the pipeline, such as Azure Data Factory (ADF), Databricks, and SQL Databases, we use AD's single sign-on to manage researchers' access to data.

In the Azure Data Lake, we use Azure Gen2 Storage Accounts to store the various formats of NielsenIQ HomeScan data including CSV, Excel Workbooks, parquet, and delta². Simultaneously, we use an Azure SQL Database to store tabular format data. Azure Gen2 Storage firewalls are configured to grant access to known IP addresses, in effect providing network security. The Azure Data Lake is deployed under an Azure Virtual Network (VNet)³, which means the Bank has its own private network. Azure resources behind the VNet such as Azure SQL Database can communicate securely with each other through a virtual network service endpoint.

Memory Limits

When multiple researchers with approved access to protected data sets make copies locally in the pre-Azure environment, the memory required to allow multiple projects to progress simultaneously grows exponentially. This issue is resolved in Azure, since researchers can read from one central data storage location, so long as they have proper permissions to use analytical tools like Databricks. Moreover, within an Azure environment, researchers reduce data movement, which also reduces the rate at which the memory grows.

Computing Limits

Depending on the organization, requirements may exist with the use of compute resources for security reasons. If limits exist, researchers must create custom compute resources (called self-hosted runtime)⁴. A disadvantage of custom resources is the potential of memory limits. If the resources created for the organization are found to encounter memory issues, they can easily be scaled up in Azure without the need to buy physical hardware.

On the other hand, if security wasn't a concern, Azure has pretty much infinite computing power through Auto Resolve Runtime⁵ that automatically scales the memory for users as required to complete processing.

² Brown, L. & McCready, M. (26 January 2022). "[Delta Lake guide](#)." *Azure Databricks*.

³ Dwivedi, K., Berry, D. & Buck, A. (1 Feb 2022). "[What is Azure Virtual Network?](#)" *Microsoft Azure*.

⁴ Li, L. & Burchel, J. (20 Jan 2022). "[Create self-hosted integration runtime](#)." *Microsoft Azure*.

⁵ Li, L., Huff, A. & Burchel, J. (21 December 2021). "[Integration runtime](#)." *Microsoft Azure*.

Analytical Tools

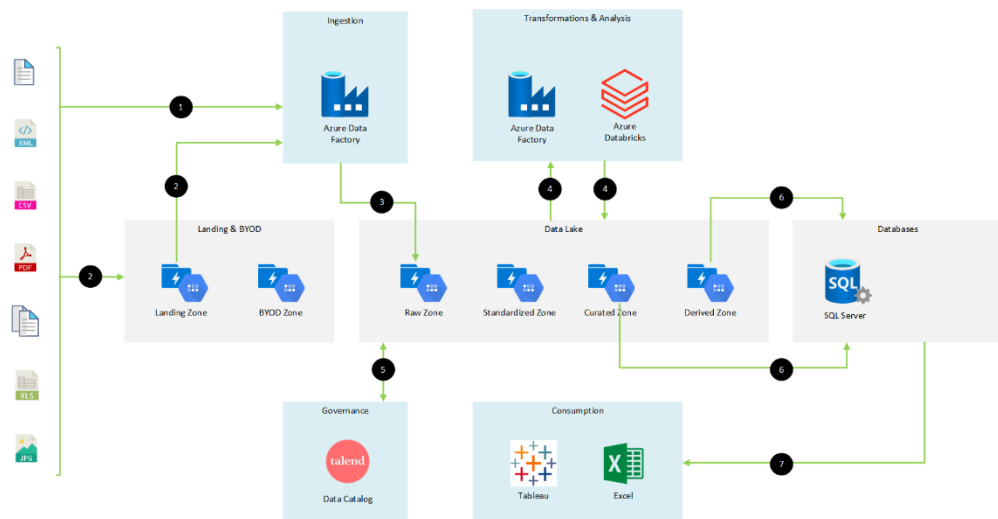
The Data Lake enhances researchers' analytical work by allowing their existing tools such as Stata, Matlab, Tableau and Python to connect to the database using JDBC⁶ or ODBC⁷ connectors. For some tools such as Python and Tableau, the Data Lake also allows researchers direct access to data under Azure Gen2 Storage Accounts which does not require connectors. In addition, researchers are also given access to Azure Databricks⁸ notebooks to perform analysis using the highly scalable Databricks clusters. Section 4 describes the significant performance gains when performing analysis using Databricks notebooks.

3. Data Lake Pipeline

NielsenIQ HomeScan Pipeline

The design and implementation of the pipeline follow the Bank's Azure Data Lake architecture, as illustrated in **Figure 1**. The pipeline reflects the whole process of data flow that begins with ingestion, undergoes transformation, and concludes with loading for analysis. Data for each phase of the process is stored in different zones to provide data availability, traceability, and consistency to the highest level. ADF is used to orchestrate the end-to-end pipeline run.

Figure 1: Bank of Canada Data Lake Architecture



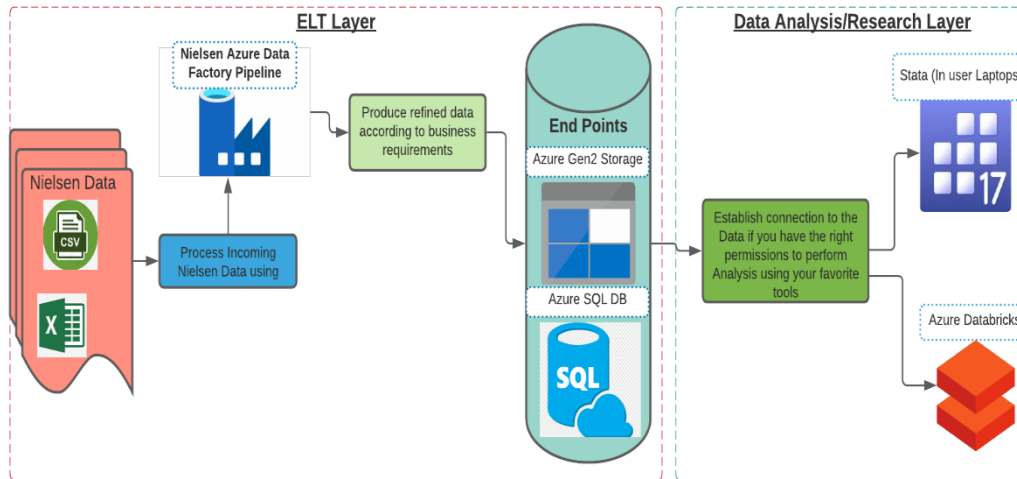
⁶ Engel, D., Roth, J. & Sharkey, K. (8 December 2021). "Download Microsoft JDBC Driver for SQL Server." *Microsoft SQL*.

⁷ Engel, D., Roth, J. & Sharkey, K. (2 November 2021). "Download ODBC Driver for SQL Server." *Microsoft SQL*.

⁸ McCready, M. & Brown, L. (27 May 2021). "What is Azure Databricks?" *Microsoft Azure*.

Figure 2 below is a workflow for NielsenIQ HomeScan. The workflow starts from receiving data from the vendor, then processing and refining the data, and ends with delivery to researchers using various endpoints.

Figure 2: NielsenIQ HomeScan workflow

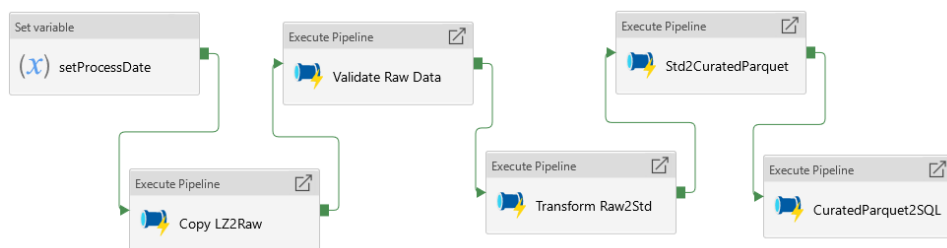


Notable Features of the Pipeline

Modularity

Figure 3 below illustrates the construct of the main pipeline, consisting of five highly modularized sub-pipelines. Each of the sub-pipelines could be run separately or jointly, providing greater flexibility and robustness of handling different use cases.

Figure 3: Nielsen Main pipeline



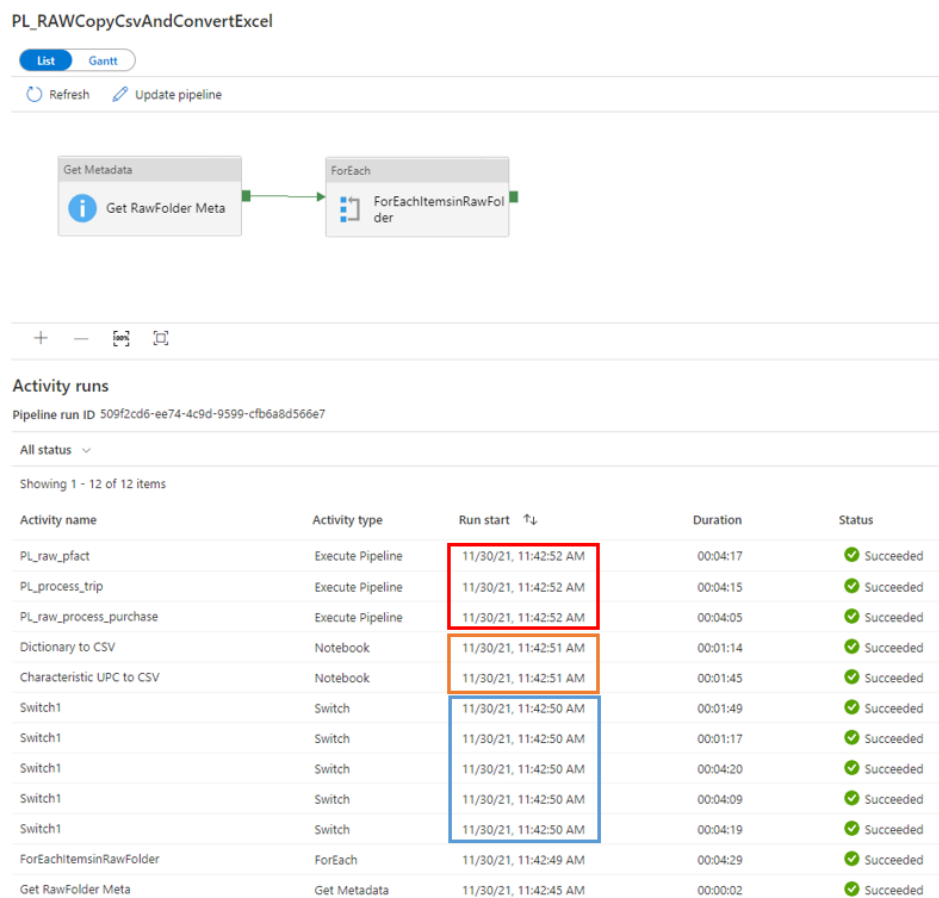
The pipeline aims to automate the process as much as possible, however we are also mindful of providing utilities to researchers to deal with actual cases. For example, in the "SetProcessDate" module, researchers can ask to process files from dates other than the default date by changing the process date parameter. Researchers may also want to pause an in-progress pipeline, then pick up later without losing any previous output. High modularity empowers researchers with such flexibility.

Performance Gains by Parallel Processing

The runtime for this pipeline has been reduced to a fraction of the pre-Azure runtime on a HPC server. These performance gains are obtained thanks to the parallel processing capability built in ADF and Databricks.

Using the “PL_CopyLZ2Raw” sub pipeline as an example in **Figure 4**, the distributed nature of Azure Data Lake storage, coupled with the “ForEach” construct in the ADF, enable the copying activities to be performed in parallel.

Figure 4: Parallel tasks processing



Another major contributor to performance gains is the use of Databricks clusters in handling the file loading and consequent transformations. Databricks builds on Apache Spark⁹, a well-known project dedicated to big data and distributed computing.

⁹ Zaharia, M. (19 Nov 2018). “What is Apache Spark?” *Databricks*.

Data quality enhancements via automated data validation

Data quality assurance is vital to the quality of research produced. In this pipeline, we embed procedures to improve data quality. Embedded procedures have the following benefits:

- Automated data validation for future data deliveries, thus ensuring we do not corrupt existing data with new invalid data
- Reliable and consistent data validation compared to manual intervention
- Fast runtimes, thus providing almost instantaneous feedback to the data vendor

We use a program to validate incoming data according to the vendor's technical specifications. In the past, we relied mainly on manual efforts to spot potential errors.

The program validates business logic, which are difficult to identify manually. During development, this program helped the Data Lake team identify loss of precision in unique numeric identifiers, as well as business logic errors. For example, we found inconsistencies in the translation from the week a shopping trip took place to the corresponding trip month. Business logic errors similar to the previous example created inaccuracies in the analysis of the aggregated data, especially when weeks in December were categorized in the aggregated data as November or January. This shifting of weeks skewed seasonal shopping trends prior to correction.

Maintaining data security during the data validation process

When the data fails any specified logic, a unique error message is logged to Azure Log Analytics and appropriate Bank personnel are alerted. However, Log Analytics is designed to be accessed by everyone at the Bank, which can be a problem if the data is protected.

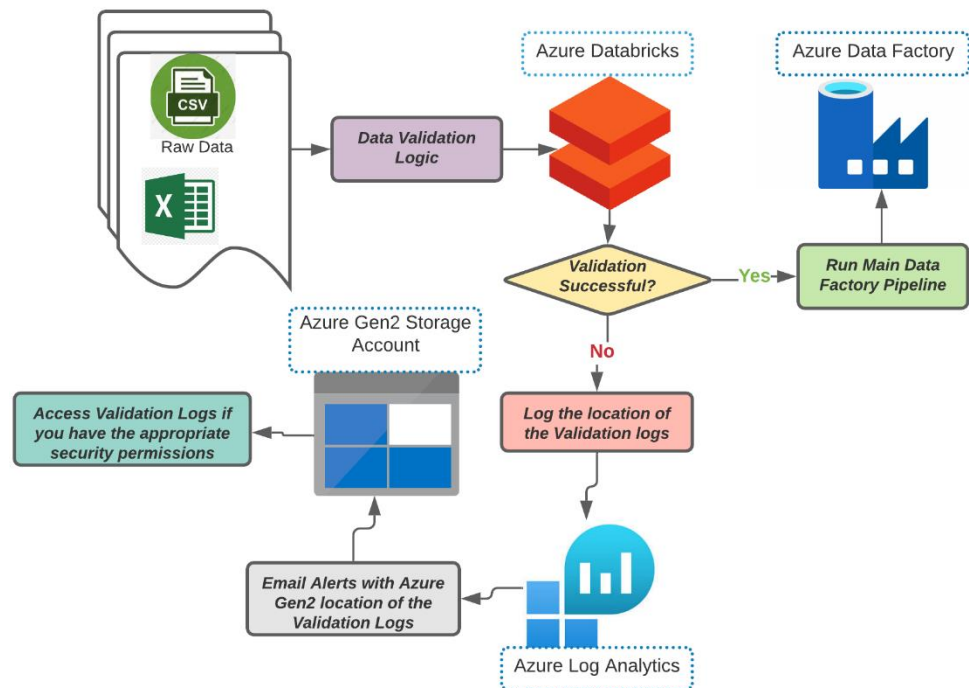
To ensure protected data is not accessed by unauthorized users, the data validation failures are logged to Log Analytics describing the *location* of the detailed logs in Azure Gen2 Storage rather than the *content* of failures. In Azure Gen2 Storage, data is secured using AD groups. Therefore, someone with the correct credentials can login and see the detailed data validation failure and share it with the vendor for resolution. **Figure 5** below describes this process.

4. Analysis & Performance

Loading data into the Data Lake

Prior to the Data Lake pipeline development, new data deliveries required a researcher-run data cleaning procedure. This involved loading raw Excel Workbook data into Stata and merging with household-level sampling weights, data dictionary files, as well as household-level monthly spending summary files. Given the size of the data set, this procedure took upwards of 3-4 hours running on the Bank's HPC clusters at each delivery, even after runtime reduction procedures such as transforming all Workbook files to CSV. The NielsenIQ HomeScan pipeline only requires upwards of 20 minutes to load new data upon delivery.

Figure 5: Maintaining security during data validation



Analytical performance

The benefit of data storage in the cloud is the ease of analysis using analytical tools such as Databricks. In **Table 1**, we summarize the performance improvements of Databricks compared to two alternative methods on a NielsenIQ HomeScan data set with over 61.5 million observations. In the time it requires to run the same analysis *once* in Stata on desktop, Databricks could run it more than 40 times.

Table 1: Comparison of analytical runtimes, seconds (s)

| | Loading data set | Data cleaning | OLS regression | Total runtime |
|------------------------|------------------|---------------|----------------|-----------------|
| Stata (desktop) | 9,800 – 12,897 | 401 | 250 | 10,451 – 13,548 |
| Stata (HPC) | 9,800 – 12,897 | 224 | 110 | 10,134 – 13,231 |
| Databricks* | 13 | 137 | 107 | 257 |

*At peak runtime, 64GB and 16 cores were allocated to this job. Highlighted are fastest runtimes in each category.

Databricks notebooks are run on a cluster that consists of a driver node and different numbers of worker nodes depending on the workload. The driver node and each of the worker nodes is configured with 14GB memory and 4 cores. At peak time, the cluster auto-scaled to 4 worker nodes. **Figure 6** visually illustrates how the analysis described in **Table 1** is completed in Databricks notebooks. We can see that

Figure 1. Reader computing and master data loading in Eudistone

Executors

- Added (Blue box)
- Removed (Red box)

Jobs

- Succeeded (Blue box)
- Failed (Red box)
- Running (Green box)

Timeline:

- 19:00:** Executor driver added. Job `df_c` is added.
- 19:01:** Executor 1 added. Executor 0 added. Job `df_c` is running.
- 19:02:** Job `df_c` is completed (Succeeded).
- 19:03:** Job `df` is added. Job `df_hf_ct = d` is added.
- 19:04:** Job `df_hf_ct =` is added. Job `df_hf_ct` is added.
- 19:05:** Executor 3 added. Executor 2 added. Job `df_hf_ct =` is completed (Succeeded).
- 19:07:** Executor 0 removed.
- 19:10:** Executor 1 removed.

Large organizations like central banks can adopt cloud-based technology to significantly improve computational capabilities while ensuring security and reducing costs. Cloud-based Data Lakes allow for an integrated and automated data pipeline that can securely store and back-up data of multiple formats, structured or unstructured. Analytical tools like Databricks are integrated with Azure Data Lakes, reducing data movement within an organization while delivering significant performance gains.

10



BANK OF CANADA
BANQUE DU CANADA

IFC-BANK OF ITALY DATA SCIENCE IN CENTRAL BANKING WORKSHOP
16 FEBRUARY 2022/16 FÉVRIER 2022

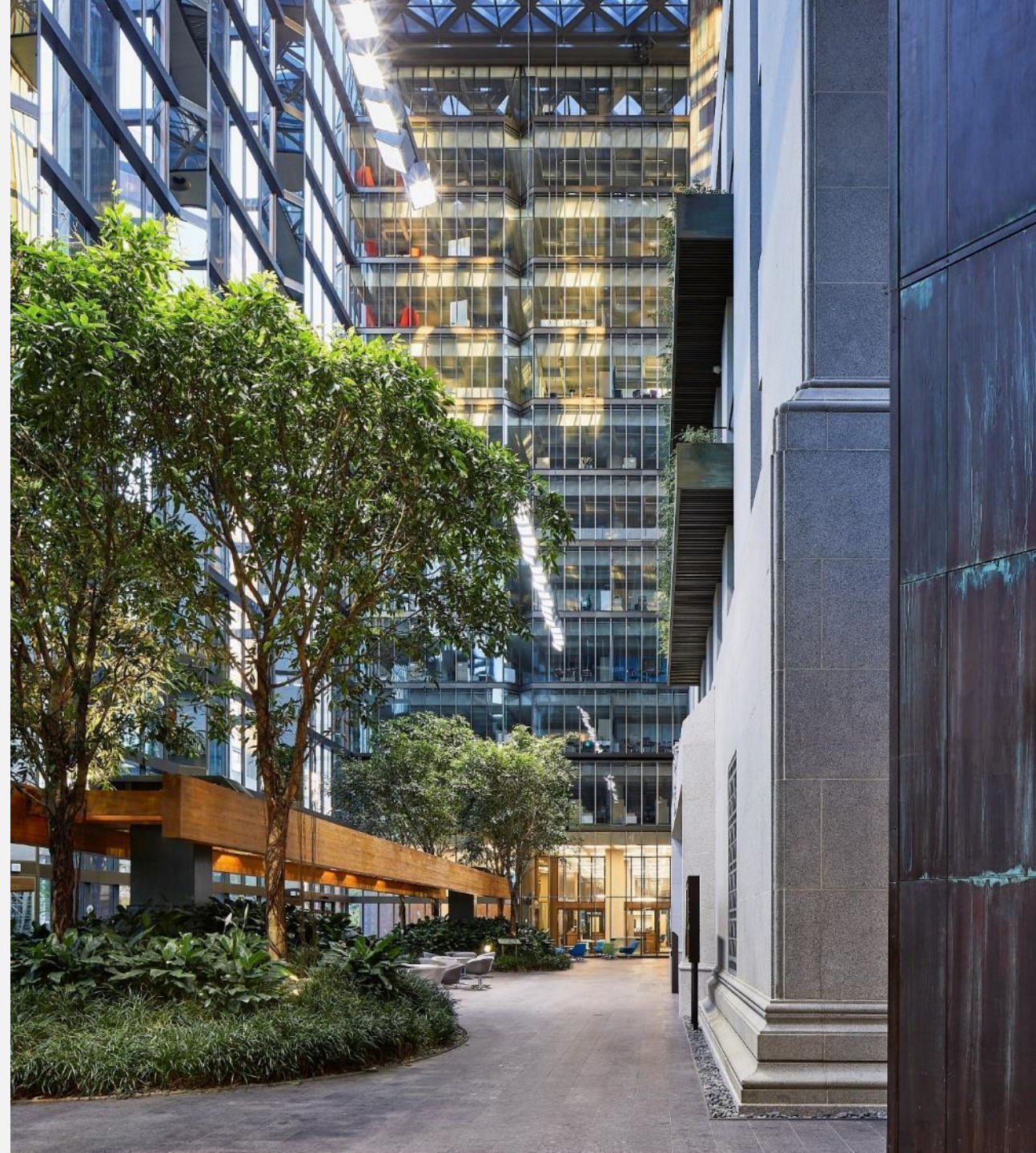
Swimming in the Data Lake

An application to NielsenIQ HomeScan

Minnie H. Cui

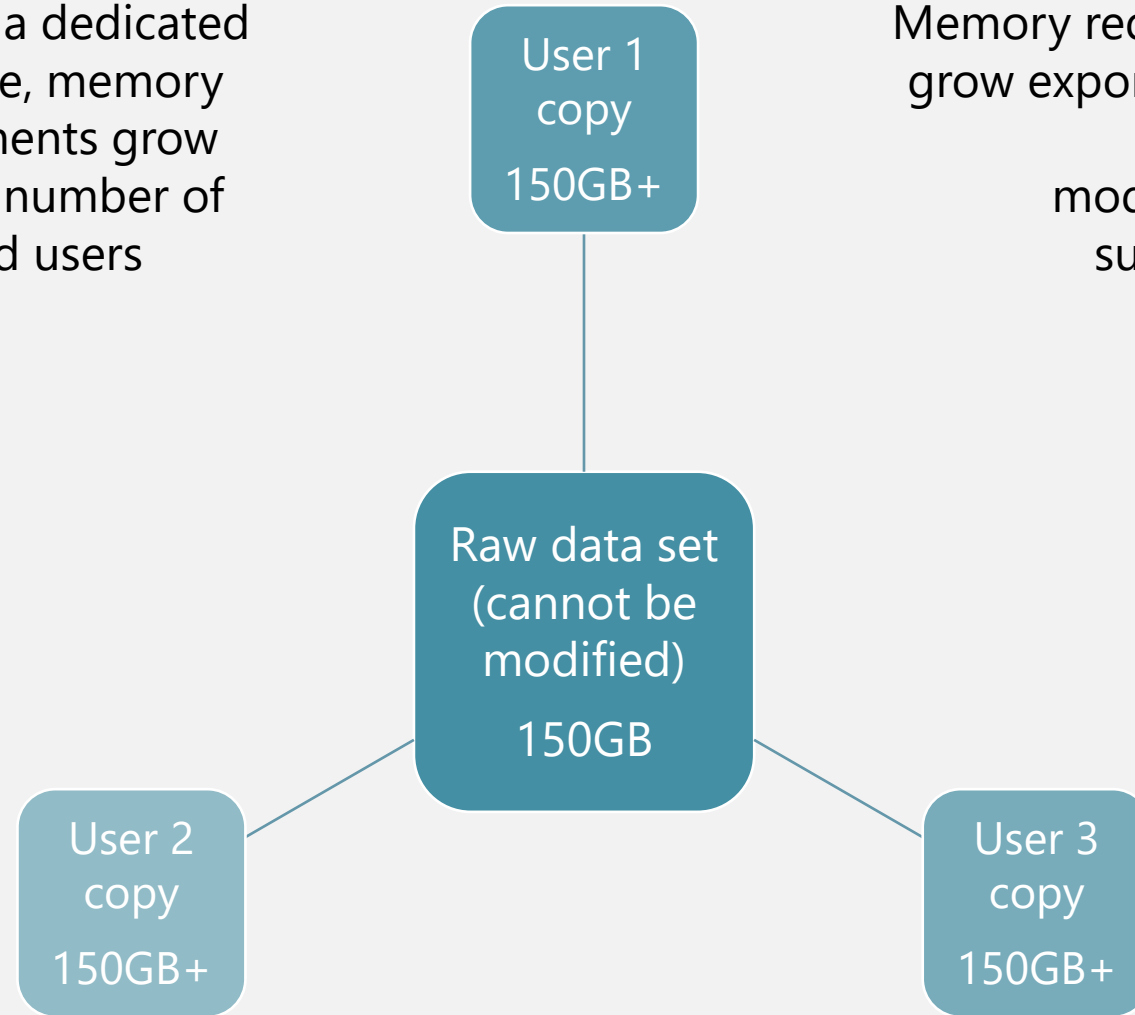
JOINT WITH BOTLHALE MOSWEU & GENE (FA GUI) JIANG

The views represented in this presentation are the authors' and do not represent the official views of the Bank of Canada.



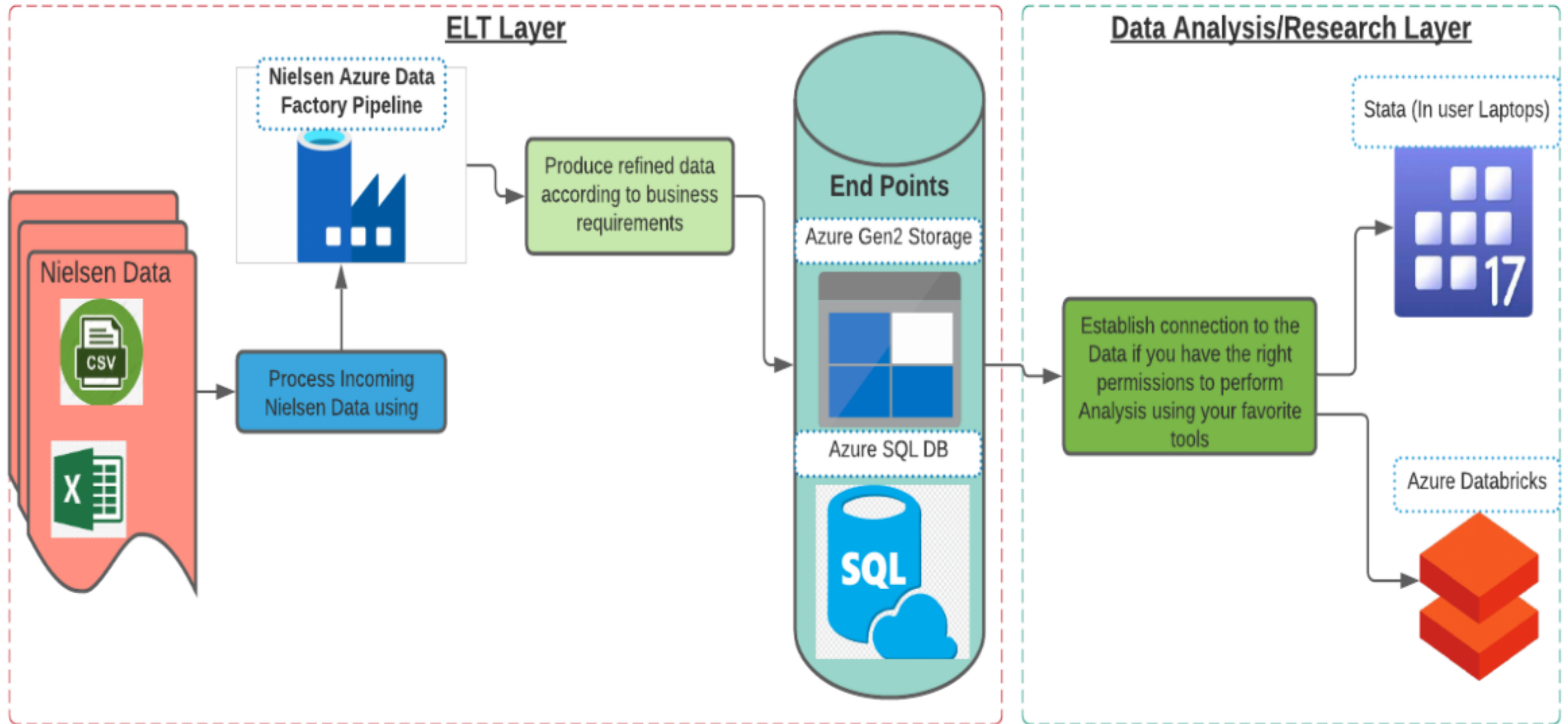
Pre-cloud protected data storage at BoC

Without a dedicated
data base, memory
requirements grow
with the number of
approved users

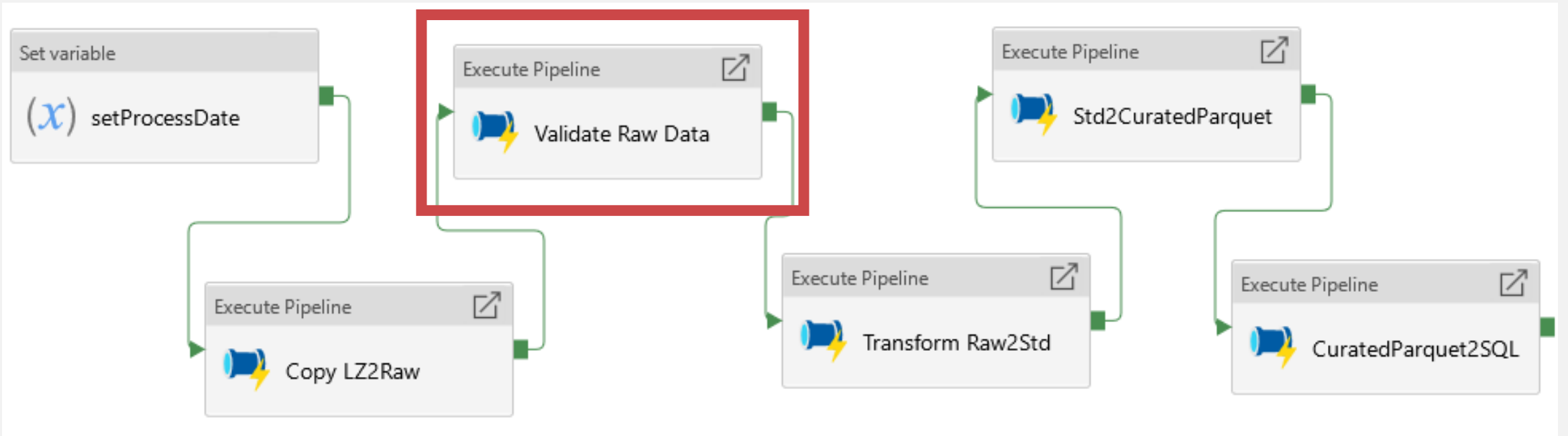


Memory required can
grow exponentially if
users save
modifications,
subsets, etc.

NielsenIQ HomeScan workflow



NielsenIQ HomeScan pipeline



Benefits of automated validation on data quality

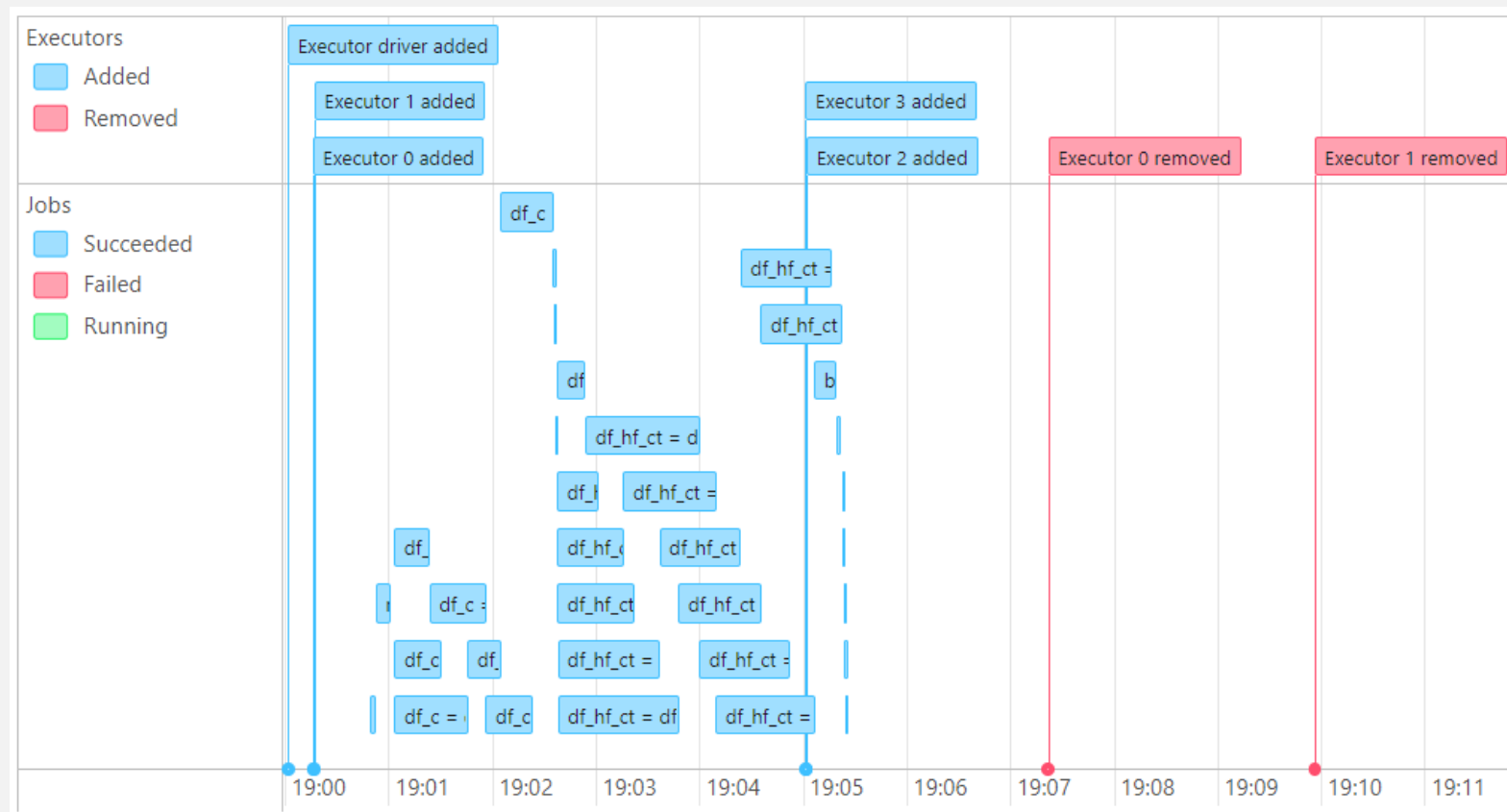
- Automated data validation for future data deliveries thus ensuring we do not corrupt existing data with new invalid data
- Reliable and consistent data validation compared to manual intervention
- Runs fast providing almost instantaneous feedback to the data provider

Comparison of analytical runtimes, in seconds (s)

| | Loading data set | Data cleaning | OLS regression | Total runtime |
|--------------------|---------------------|------------------|-------------------|--------------------|
| Stata (desktop) | 9,800 – 12,897 | 401 | 250 | 10,451 – 13,548 |
| Stata (HPC) | 9,800 – 12,897 | 224 | 110 | 10,134 – 13,231 |
| Databricks* | 13 | 137 | 107 | 257 |

*At peak runtime, 64GB and 16 cores were allocated to this job. Highlighted are fastest runtimes in each category.

Parallel computing and cluster auto-scaling in Databricks



↑
Certain processes are automatically run in parallel to reduce runtime

↑
Cluster automatically removes worker nodes when the analysis is complete

Ensuring security

- Without compromising security protocols, Azure provides users with single sign-on access to data
- Given HomeScan's Protected-B status, we use
 - › Azure Active Directory (AD) to authorize and authenticate users
 - › Role-based access control (RBAC) to control the access level to data containers
 - › Access control list (ACL) to folders and files under the containers

Ongoing questions about security

- BoE has expressed concern with “secretive” nature of cloud computing providers
- Is the data in the cloud secure from unauthorized access, for instance, from the provider?
 - › Azure Data Lake implementation at the BoC is behind an Azure Virtual Network (VNet)
- What is the risk that multiple banks, for instance, can be affected by cyber attacks or service outages?
- These issues should be discussed more before a large scale implementation...



Questions?

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Implementing a multi-tenant big data platform – challenges and approaches taken in the BIS¹

Hiren Jani and Anand Kannan,
BIS

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



Implementing a Multi-tenant Big data platform

Challenges and the approach taken in the BIS

Hiren Jani and Anand Kannan

Why a centralized platform?

Organizational Landscape

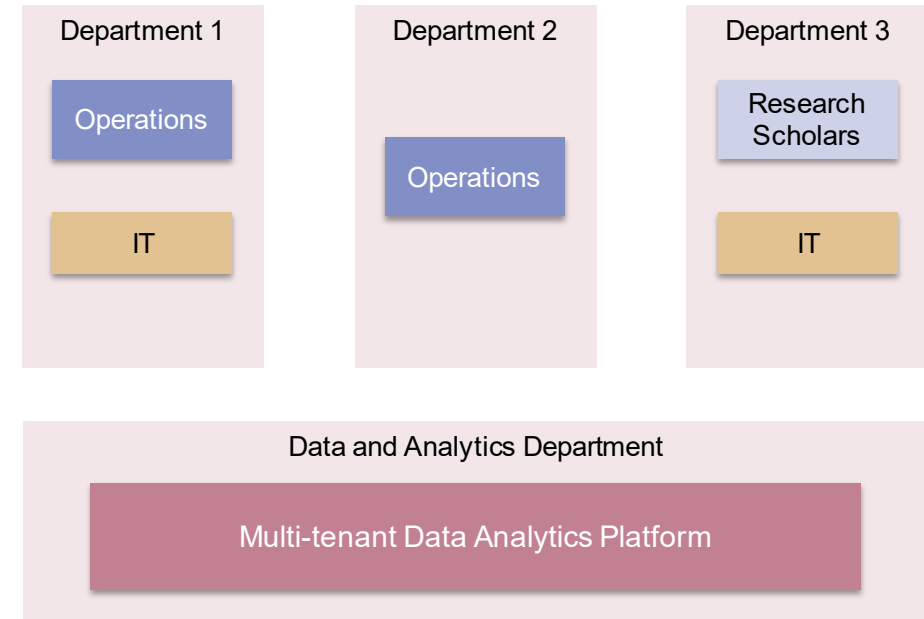
- Various departments in the bank use the platform for their Big Data Analytics needs
- Departments with varying IT capabilities

Our Approach

- Big data platform offered as a shared platform (Horizontal function)

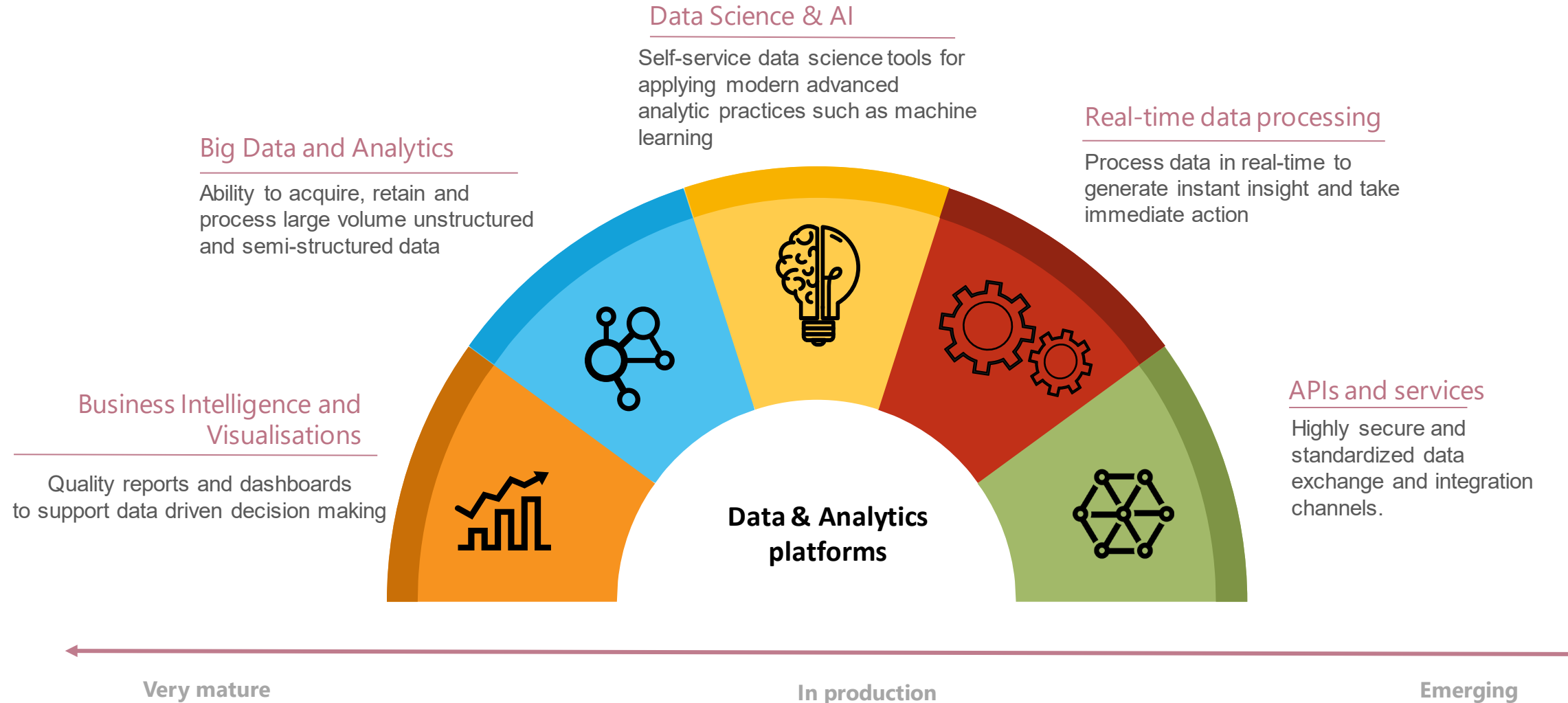
Motivation

- Economies of Scope and Scale
- Ease of Governance
- Enabling self service

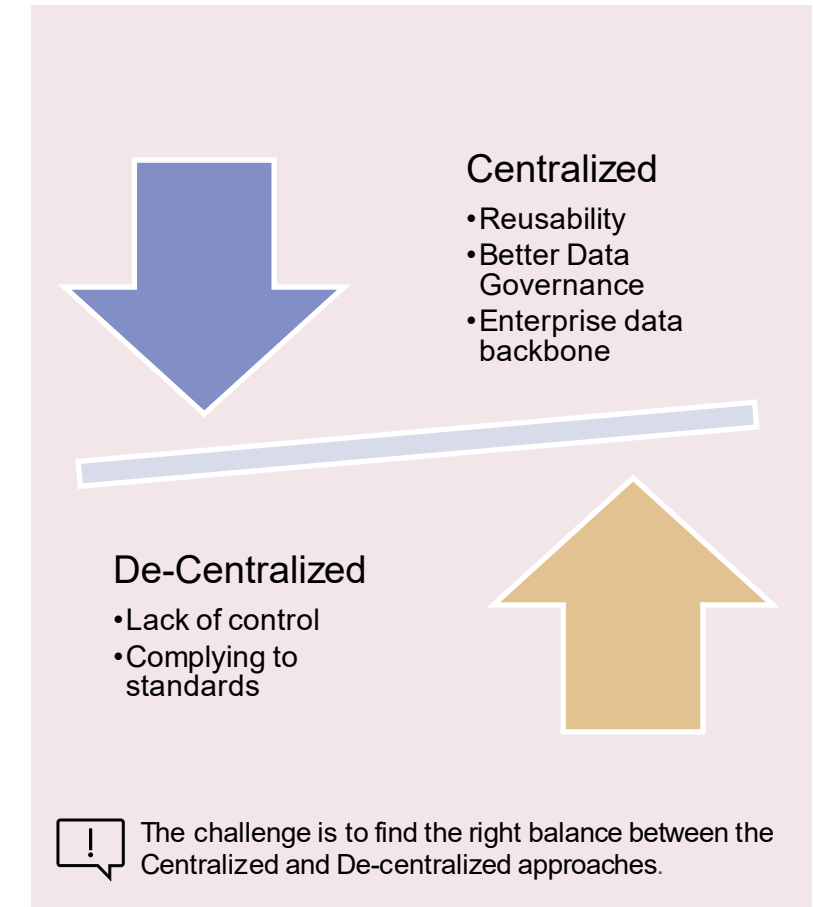
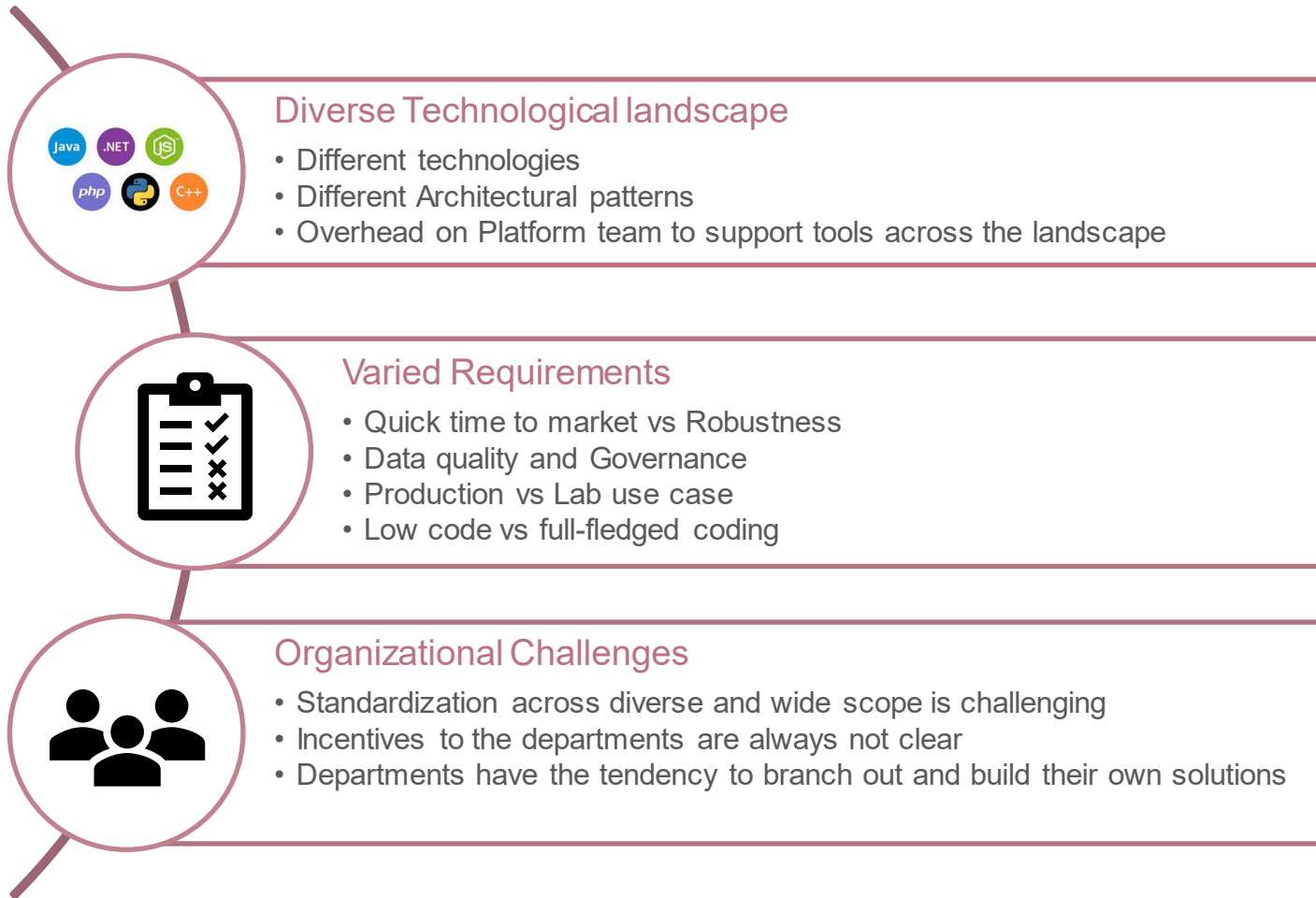


The diverse departments in the BIS and the position of the Data and Analytics as a shared horizontal function.

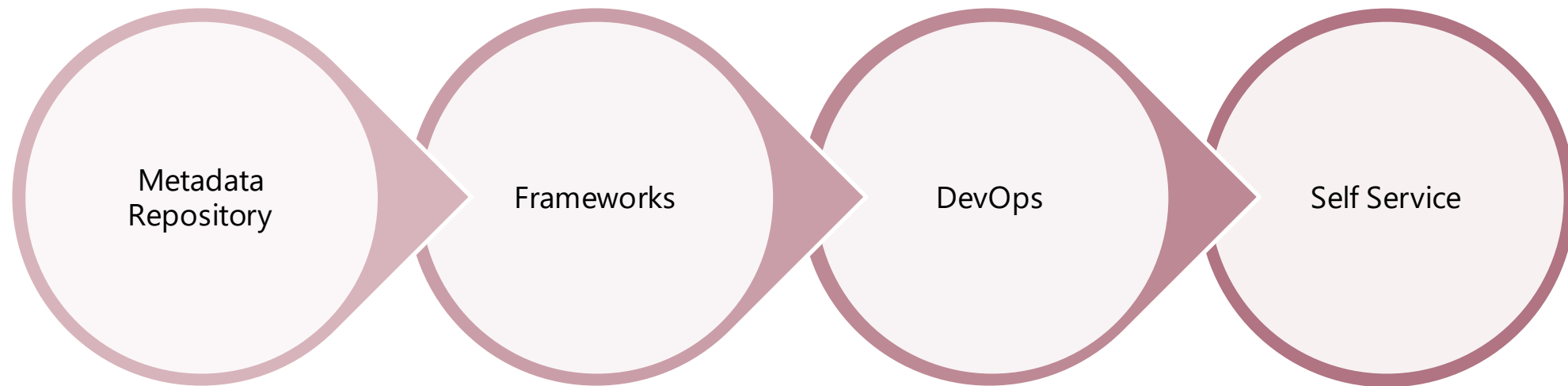
Platform Components



Challenges



Empowerment with Self Service



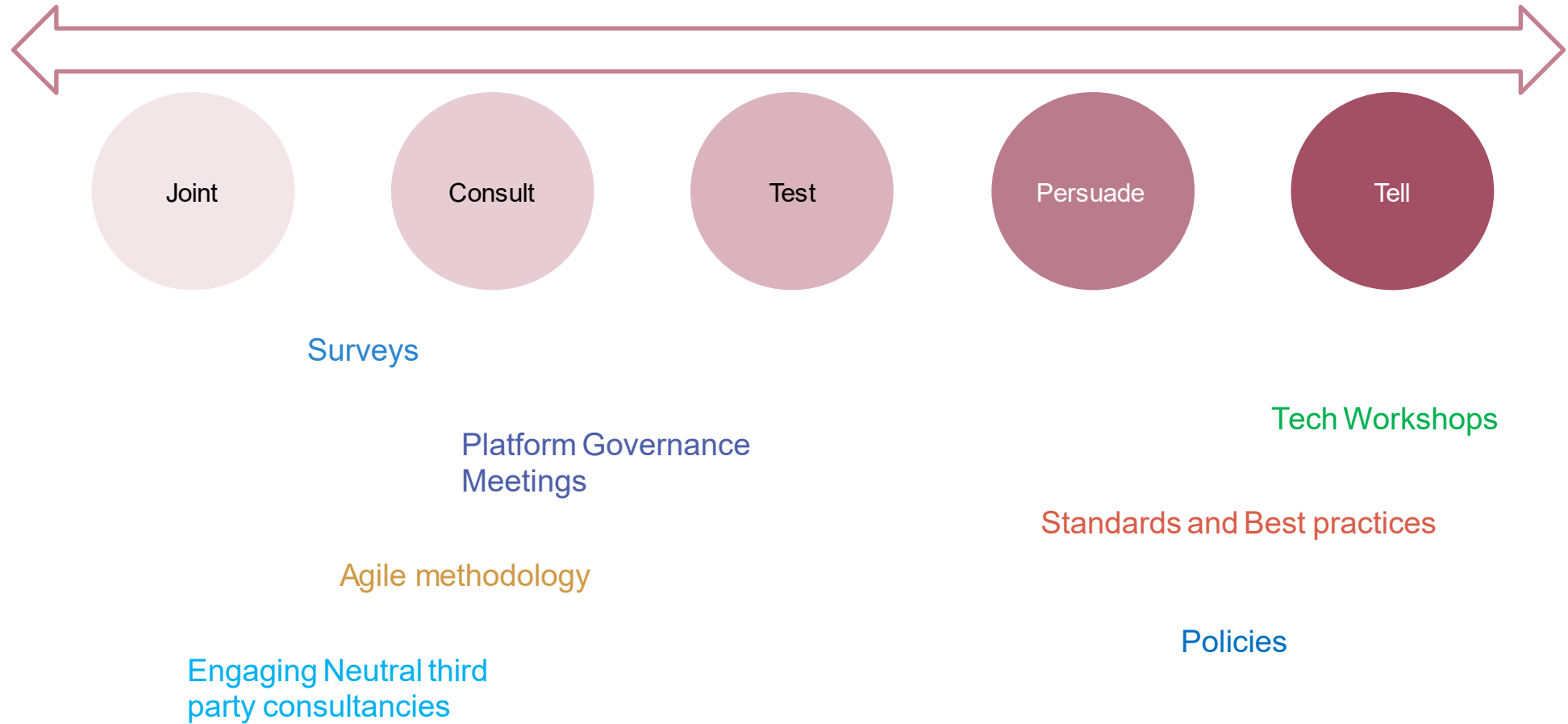
- Data assets generated from metadata (tables, topics, pipelines etc.)
- Define schema for configuration files
- Standard naming convention
- Enforce security

- Data Ingestion (Batch and Real time)
- Data pipeline Monitoring
- Extensible frameworks

- Automated deployment to the platform
- Validation of provided metadata
- Trigger deployment as part of Business Unit release process.

- ✓ Business Units are self sufficient
- ✓ Minimize dependency on platform team to manage and deploy components independently
- ✓ Grant controlled access to Business Units

A Spectrum of Decision Making¹



¹ <https://www.managementcenter.org/resources/modes-decision-making/decision-making-spectrum-3/>

Learnings

- Technology is evolving at a rapid pace – platform team to be cautious when identifying and promoting standard tools.
- Focus on change management from the early stages of the project.
- Operational costs needs to be factored in to planning.



Questions?

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Modern computing platforms as key technology for central banks, financial supervisors, and regulators¹

John Ashley and Jochen Papenbrock,
NVIDIA

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Modern Computing Platforms as Key Technology for Central Banks, Financial Supervisors, and Regulators

Dr. Jochen Papenbrock and Dr. John Ashley¹

Abstract

A team from NVIDIA examined the leading and most challenging use cases for data science and the corresponding IT/software tools that central banks, financial supervisors, and financial regulators are currently discussing and implementing around the world. NVIDIA is a leading global provider of accelerated computing platforms for AI, Generative AI (GenAI), data processing, high-performance computing (HPC), and the industrial metaverse across multiple industries including financial services and public sector. Accelerated computing reduces time to results and cost, and is significantly more energy efficient, allowing existing data centers to be more sustainable.

There are two main observations that can be drawn from the team's analysis. First, central banks and supervisors are embracing Big Data, AI, GenAI, ML (machine learning) and accelerated simulation methods, and the open-source software ecosystem to support them. Second, cloud-native, full stack AI & simulation factories have started to be on the radar of central banks, financial supervisors, and regulators as key technology. However, the common workflows and tools used by many today exhibit computational bottlenecks due to CPU-only, non-accelerated computing. This creates severe roadblocks in scalability, productivity, agility, and time-to-value – unnecessarily reducing return on investment and driving unnecessary levels of cost and energy consumption. Also, the tools that are available for model optimization during training, inferencing and MLOps are rarely used.

Accelerated computing platforms using specialized hardware and software stacks based on massively parallel execution, integrated with industry standard software and open-source frameworks can help address these issues. GPUs – graphics processing units – are the most broadly available, supported, and performant foundations for such platforms. The accelerated hardware layers are just one building block of an accelerated computing platform. The other equally important building block is the complementary software stack with corresponding application frameworks. Those symbiotic hardware and software stacks are the basis for building entire 'factories' for AI and simulation to solve the most challenging problems, to predict the future, and thus are a highly important technology for central banks, financial supervisors, and regulators.

Such modern accelerated computing platforms can operate in any environment. Because GPUs are ubiquitous, these platforms can be deployed in public or private clouds, or in-house/on-prem data centers and colocation facilities. They can even be used in hybrid, cloud-native set-ups. They leverage popular open-source data science and engineering tools and Python packages. These characteristic of modern accelerated computing platforms are very important for broad acceptance in data science and IT departments and in global developer communities.

This paper discusses a wide range of central bank use cases and outlines how to implement these in an effective and efficient way utilizing accelerated computing platforms. These technologies also help implementing AI governance, AI model risk management and building trustworthy, transparent and explainable AI that will further increase the confidence of supervisors, regulated entities, and the public. These same

¹ Respectively, Head of Financial Technology EMEA and Lead Developer Relations Manager Banking Global, NVIDIA, Germany (jpapenbrock@nvidia.com); Chief Architect, AI Nations & Director, NVIDIA AI Technology Centers, NVIDIA Corp., USA (jashley@nvidia.com)

platforms support the technical side of the transition to building and supervising a more sustainable and resilient financial system, taking economic and climate-related change into account.

The paper includes somewhat more technical content to highlight the ready availability of GPU-accelerated tools and application frameworks that enable and accelerate a range of data science operations that are key solutions building blocks for many current and future problems in central banking and macroprudential analysis. This includes areas such as large graph analysis, graph neural networks, conversational/speech AI, deep learning, GenAI and foundation models, NLP and Large Languages Models (LLMs), geospatial AI, quantum computing simulation, simulated digital environments, digital twins, physics-informed AI/ML and climate intelligence. Leveraging full stack accelerated computing platforms delivers the scalability, productivity and time to insight central banks, supervisors, and regulators need to accelerate and enhance their mission, mandate and policies.

Keywords: central banks, data processing, explainable AI, trustworthy AI, RAPIDS, ESG, accelerated computing, climate intelligence, Generative AI, LLM, foundation model, financial supervision, financial regulation, RegTech, SupTech, AI computing

JEL classification: Q58; Q57; Q56; E58; G28; C31; C81

Contents

Accelerated Computing Platform for Enterprise Data Science and AI.....4

Generative AI and Large Language Models in Central Banks..... 16

Candidate Use Cases at Central Banks for Accelerated Computing 22

Appendix: Selected accelerated libraries and frameworks..... 28

Acknowledgements 29

References30

Accelerated Computing Platform for Enterprise Data Science and AI

A team from NVIDIA examined the use cases for data science and the IT/software tools that central banks are currently discussing and implementing around the world.

We drew on several sources such as IFC/BIS publications² and attended central bank conferences such as the "International Conference on Statistics for Sustainable Finance"³ and "Data Science in Central Banking, Part 1: Machine Learning Applications"⁴. We held discussions and were involved in projects with some of the leading central banks. One project we would like to highlight that involved many European central banks and regulators is the EU Horizon2020 project FIN-TECH⁵.

There are two main observations that can be drawn from these sources:

First, central banks and supervisors are embracing both Big Data and AI/ML (artificial intelligence and machine learning) methods and the open-source software ecosystem to support them. The establishment of a transformed supervisory model can be observed that leverages supervisory technology, sustainable finance technology and climate science to

- digest the vast volumes of structured and unstructured data,
- improve timeliness ("real time") of identification, monitoring, and early warning/intervention of (emerging) risks.
- support both the "big picture" and a granular view at different zoom levels

It was possible to identify typical workflows and tools that are very useful and that we have observed and discovered in many other industries dealing with Big Data and AI/ML.

A typical workflow involves 3 steps: data preparation, model training and visualization/explanation. Different types of databases are used, and the analysis sometimes involves graph/network analysis. Many open-source and Python tools are used.

² "Computing platforms for big data analytics and artificial intelligence"

"Big data and machine learning in central banking"

"Big data for central banks"

"The supotech generations"

"The use of big data analytics and artificial intelligence in central banking"

"Central Bank Communications: information extraction and semantic analysis"

³ Conference link: <https://www.banque-france.fr/en/international-conference-statistics-sustainable-finance>; our contribution is titled "Accelerated Data Science, AI and GeoAI for Sustainable Finance in Central Banking and Supervision" and can be found in the second video at 7:13:10 – 7:29:10 [Q&A 8:10:33 – 8:14:50].

⁴ https://www.bis.org/ifc/events/211019_ifc_bdi.htm

⁵ This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 825215 (topic ICT-35-2018, Type of action: CSA). The content reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

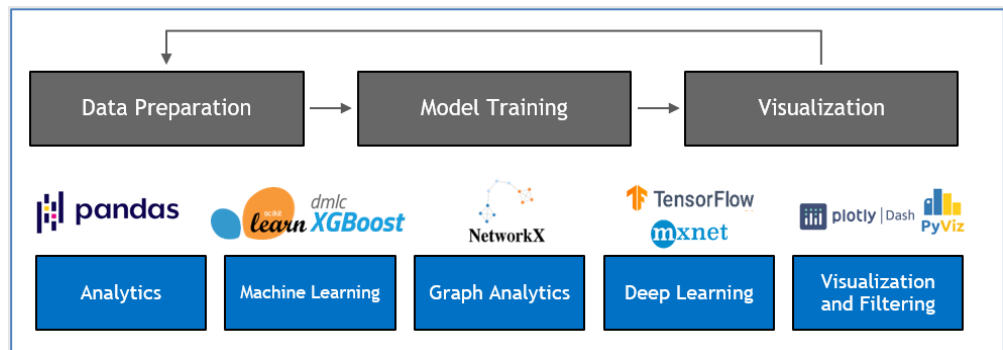


Figure 1: typical workflow and exemplary Python tools in many central bank projects utilizing big data and AI. Open-source projects and Python tools have democratized data science. However, the approach exhibits the typical computational bottlenecks due to CPU-only processing.

At this point in time, less focus is put on industrial tools to deploy the models built and scale them across the organisation in the inferencing (production) step.

Second, in all data science steps mentioned above there are computational bottlenecks with CPU-only processing that are made worse by forcing general purpose infrastructure to serve AI powered workflows.

This has negative impacts:

- inhibits developer productivity growth
- delays time to insight
- limits model scalability
- increases total cost of ownership and carbon emissions
- increases infrastructure complexity.

Also, software tools for model optimization during training and inferencing and for MLOps are rarely used.

In some important and growing areas – such as initiatives to greening the financial system and to develop climate stress tests and scenarios for banks and insurance companies – technologies such as AI, HPC, high-performance data analytics (HPDA) and scientific visualisation still appear rarely to be used.

The Challenge: Enabling the AI Transformation

It all starts with the key business drivers. A bank will be focused on improving fraud detection, enabling virtual assistants, and creating recommenders to produce next best actions. An insurance company will also want to automate claims processing and identify fraudulent claims. Regardless of the type of financial institution surveyed, they all experience traditional AI infrastructure challenges: AI is complex, it is hard to deliver scalability and reliability at the same time and needs to operate within the budget constraints of the bank.

At the same time, fraudsters are already using Generative AI to create “deepfake” calls to bankers to get the bank itself to steal money from customers on the fraudster’s orders. New identities will be backstopped by AI generated histories. Money will be laundered using flows planned and eventually executed by AI to escape detection.

The transition to being an AI enabled bank cannot be leisurely. It is urgent that banks build a native capability for AI – leading the pack is not required for survival but being solidly in the pack is. Leading firms have months of head start with generative AI and in some cases years with pre-ChatGPT AI. No firm can afford to take a “wait and see”

position – ramping up too late, or even too slowly, will expose customers to massive risks and is arguably a failure of fiduciary duty.

The Solution: Accelerated Computing Platform and Enterprise AI Software Stack

Accelerated computing platforms using specialized hardware integrated with industry standard software can help address many of these issues and challenges. GPUs – graphics processing units – are the most universally available, supported, and performant foundations for such accelerated platforms.

According to the IFC-BIS publication "Computing platforms for big data analytics and artificial intelligence" (see Bruno et al. 2020), "Central banks' experience shows that HPC platforms are primarily developed to ensure that computing resources are used in the most efficient way, so that analytical processes can be completed as rapidly as possible. [...] A processor core (or 'core') is a single processing unit. Today's computers – or CPUs – have multiple processing units, with each of these cores able to focus on a different task. Depending on the analytical or statistical problem at hand, clusters of GPUs (graphics processing units, which have a highly parallel structure and were initially designed for efficient image processing) might also be embedded in computers, for instance, to support mass calculations."

Today the superb computing power of GPU (Graphics Processing Unit) clusters are widely used in research where many of the most capable supercomputers are powered with GPUs. Industrial firms and other research organizations have long since adopted GPU computing to address high-performance computing requirements.

Here are some examples of GPU-accelerated computing:

- Cambridge-1 is the fastest supercomputer in the UK, boosting Covid-19 research.
- Large Language Models like GPT-3 with billions of parameters are usually trained on a GPU infrastructure like Selene, one of the fastest supercomputers, built and managed by NVIDIA. NVIDIA technologies power many systems on the Top500 and Green500 lists.⁶
- MLPerf⁷ is a benchmark produced by a consortium of AI leaders from academia, research labs, and industry whose mission is to 'build fair and useful benchmarks' that provide unbiased evaluations of training and inference performance for hardware, software, and services—all conducted under prescribed conditions. To stay on the forefront of industry trends, MLPerf continues to evolve, holding new tests at regular intervals and adding new workloads that represent the state of the art in AI. Systems powered by NVIDIA technology deliver leading performance across all MLPerf tests for training, both per chip and at scale. In inferencing, NVIDIA accelerated systems continued to deliver exceptional performance across the full range of MLPerf tests.
- Meta/Facebook established a Research SuperCluster (RSC) for AI research based on GPUs.⁸

The accelerated hardware layers (GPU servers and high-speed networking) are just one building block of an accelerated computing platform. The other equally important building block is the optimized software stack. It helps to program the hardware, give access to numerous tasks in an accelerated way and to further optimize the compute performance.

Modern AI platforms must work in a similar manner across supercomputers and public cloud all the way to on-prem data centers and edge. They also need to leverage existing

⁶ <https://www.top500.org/>

⁷ <https://mlcommons.org/en/>

⁸ <https://ai.facebook.com/blog/ai-rsc/>

open-source tools and Python packages. These characteristics of modern accelerated computing platforms are important for broad acceptance in data science and IT departments and in global developer communities.

Modern computing platforms support the execution of data science and high-performance computing (HPC) workloads as well as building/deploying AI models at scale. Those platforms are flexible, powerful stacks of hardware and software that are orchestrated to reduce performance bottle necks. These platforms provide access to accelerated data science and complex AI model building for a wide range of users, while addressing a larger number of use cases. Mass calculations and accelerations based on GPUs are crucial as they directly translate into several benefits:

- Accelerated training allows building larger and more accurate models
- Accelerated inferencing allows real-time utilization of trained models, e.g., for detecting fraud and cyber attacks
- Accelerated generation of synthetic data (e.g., Generative Adversarial Networks - GANs) amplifies existing data at large scale and simplifies collaborations of central banks with SupTech and RegTech startups
- Accelerated simulation enables more realistic and complex simulation
- Accelerated computing allows data scientists to better assess, validate, audit, and explain AI models
- People with strong data science skills are a limited and expensive resource globally. Accelerated computing brings leverage to that resource, this reduces time-to-insight and improves productivity.

Another major advantage of modern computing platforms is their leveraging of open-source software. This leverages the innovation potential from global developer and data science communities. Here are some examples:

- Open-source packages like TensorFlow and PyTorch are the de-facto standard frameworks for building and tuning large language models like BERT, ChatGPT, and LLAMA2. GPU support is a standard part of these frameworks.
- There are open-source Python-based projects for enabling end-to-end data science and analytics pipelines entirely on GPUs. This aims to accelerate the entire data science pipeline including data loading, ETL, model training, and inference. In-memory analytics help to scale up and out with accelerated data science. This enables more productive, interactive, and exploratory workflows.
- Open-source systems for automating deployment, scaling, and management of containerized applications as well as virtualization technologies enable business continuity and workload balancing, resource sharing and improved utilization of the existing resources in modern computing platforms.

The Economics and Benefits of Accelerated Computing

GPU accelerated computing moves compute intense and embarrassingly parallel parts of the application to the GPU. This enables dramatic performance improvements, with process that used to take days taking hours, or those that took hours to minutes.

GPU accelerated computing has two main benefits. By accelerating the compute, firms can choose what to optimize for: time to results, more compute intensive (higher fidelity) models, exploring more of the solution space, processing more data, or using a smaller system for the same workload. Because accelerated computing is also more energy efficient, the same results can be had faster and with lower energy consumption, which depending on the energy source, may also mean reduced CO2 emissions.

In addition to the savings in computer systems and energy, improved productivity of data scientists and engineers can have a significant impact on the top and bottom lines.

These performance improvements and energy savings are reflected in industry standard benchmarks like MLPerf and the Green 500 list.

For DL workloads, GPU-based computing platforms have set records for the MLPerf⁹ benchmark (an industry-standard set of benchmarks across a variety of AI modelling tasks), handily surpassing all other commercially available systems (see Mattson et al. (2020)). Comparable results are documented with respect to the STAC-A2™ Benchmark suite¹⁰ which is the industry standard for testing technology stacks used for compute-intensive analytic workloads involved in pricing and risk management.

GPU-accelerated Deep Learning (DL) frameworks¹¹ offer building blocks for designing, training, and validating deep neural networks through a high-level programming interface. Widely used DL frameworks, such as MXNet, PyTorch, TensorFlow, and others rely on GPU-accelerated libraries to deliver high performance, multi-GPU accelerated training. These optimized DL framework containers are performance-tuned for GPUs. This eliminates the need to manage packages and dependencies or build DL frameworks from source. Containerized DL frameworks, with all dependencies included, can be used to develop common applications, such as conversational AI, natural language understanding (NLU), recommenders, and computer vision.

GPU-acceleration translates into faster training and can be used for scaling hyper-parameter optimization and simply training a larger variety of models and framework which is a way to find even better models. Also, GPU acceleration can be used to produce synthetic data to enhance the real data set which can amplify and improve training results.

Enterprise Platform for Building Accelerated Production AI

Accelerated Computing is a key element of Enterprise AI which is a concept including an end-to-end, cloud-native suite of AI and data analytics software that's optimized to enable any organization to use AI and benefits from accelerated computing. It can be deployed from the enterprise data center to the public cloud. It has development tools and frameworks for the AI practitioner and reliable management, orchestration, and virtualization layers for the IT professional to ensure performance, high availability, and security¹².

An accelerated production AI environment is enabled by an enterprise AI system which is characterized by the following capabilities:

- Supporting workflows for Data Science, AI, GenAI/LLM
- SDKs, application frameworks and pre-trained models
- deployable anywhere – from on-prem to cloud
- cloud-native and supporting hybrid set-ups
- supporting open-source software (OSS) frameworks and reducing development complexity
- secure and scalable environments and infrastructure
- certifications and service levels

⁹ <https://mlcommons.org/en/>

¹⁰ <https://stacresearch.com/>

¹¹ <https://developer.nvidia.com/deep-learning-frameworks>

¹² <https://www.nvidia.com/en-us/data-center/products/ai-enterprise/>

Such an enterprise AI stack is the foundation for delivering secure, trustworthy, and scalable AI. It helps addressing the following solution requirements:

- **Acceleration**
 - Faster time to value, high perf and low cost, massive scale
 - putting AI workloads on virtualized solutions preserves the performance while adding the benefits of virtualization, such as ease of management and enterprise-grade security
- **Choice/Flexibility/Sovereignty**
 - Sovereignty of data, operations, and software
 - Range of AI software ecosystem (AI Enterprise, OSS, ISV), vendor-agnostic
 - cloud-native operating model to address multi cloud, hybrid cloud, on-prem environments
 - optimizing your workload placement strategy with cloud-native architecture
 - 'Mobility of compute' to address data gravity (and privacy)
 - balance workload placement based on cost and performance
- **Customization**
 - to improve accuracy, leverage enterprise data, increase trustworthiness (e.g. reduce hallucination and Factual but out-of-context answers in Generative AI)
- **Operations**
 - Easy to deploy, boosts developer productivity, simple to operate and future proof hybrid cloud strategy
 - AI lifecycle management
- **Performance**
 - Satisfying scaling infrastructure demands regarding data and compute (data prep, queries, testing, real-time inferencing)
- **Privacy & Sovereignty**
 - Enterprise data and IP are private and critical – this data needs to be controlled and protected to prevent leakage outside the organizational boundary.
 - securely run your private corporate data to do fine-tuning, run inferencing, and, in some cases, even training in-house.
- **Safety and Security**
 - Intrinsic security at every layer of the stack
 - integrated security and management
 - Guardrails are topical, and for safety and security
- **Cost and Sustainability**
 - easy to deploy, boosts developer productivity, Simple to operate and future proof hybrid cloud strategy
- **Compliance and AI Governance**
 - compliance needs that enterprise solutions, including AI, must meet. Access control and audit readiness

The entire enterprise AI stack has the following components:

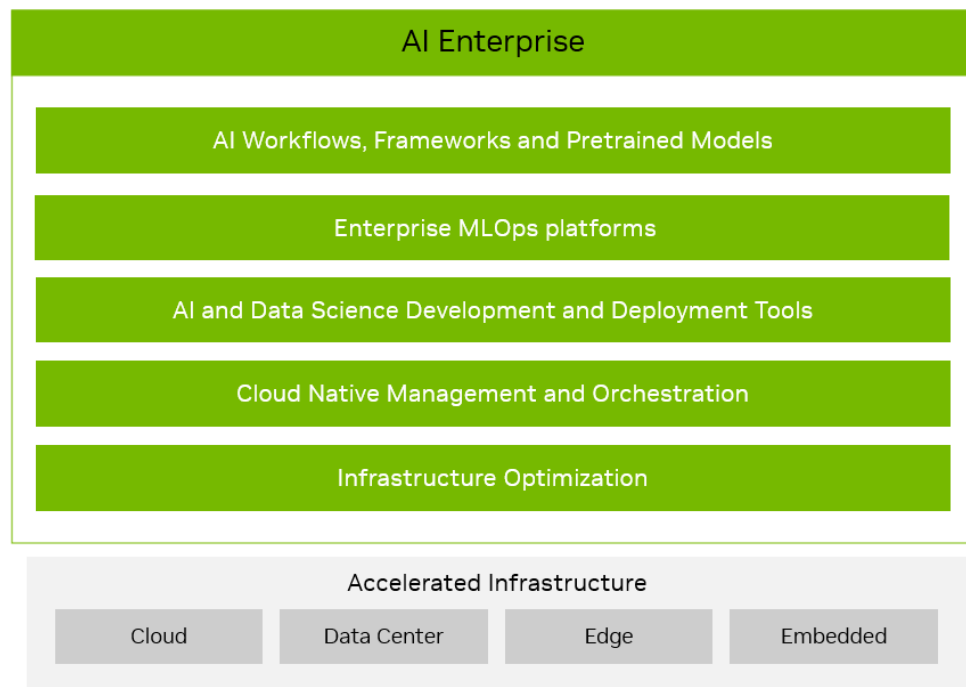


Figure 2: Components of an entire enterprise AI stack.

The technological basis is the optimized, accelerated infrastructure plus cloud-native management and orchestration layers. On top of this layer are tools for development and deployment of AI and Data Science. An enterprise MLOps platform can be added. This end-to-end stack can be used in any environment and established workflows based on this platform are portable across public and private clouds, can be used in a multi-cloud environment and in a hybrid way (like cloud plus on-prem).

An example of such a cloud native stack can be found in a repository of tools, models, and sample workflows called NGC (NVIDIA GPU Cloud)¹³. This repository also contains GPU-activated tools for model optimization and model inference serving. It is a cloud native repository for many of the previously discussed SDKs, applications frameworks, and deployment tools. The content simplifies building, customizing and the integration of GPU-optimized software into workflows, accelerating the time to solutions for users.

- **Containers** package software applications, libraries, dependencies, and run time compilers in a self-contained environment so they can be easily deployed across various compute environments. They enable software portability.
- **Models and Resources:** the NGC Catalog offers pre-trained models for a wide range of common AI tasks. The pre-trained models can be used for inference or fine-tuned with transfer learning, saving data scientists and developers' valuable time. Resources provide reference neural network architectures across all domains and popular frameworks with the state-of-the-art accuracy to enable reproducibility as well as documentation and code samples which make it easy to get started with deep learning.
- **Helm Charts:** Kubernetes is a container orchestrator that facilitates the deployment and management of containerized applications and microservices. A Helm chart is a package manager that allows DevOps to configure, deploy and update applications across Kubernetes environments more easily. The

¹³ <https://www.nvidia.com/en-us/gpu-cloud/>

NGC Catalog provides Helm charts for the deployment of GPU-optimized applications and SDKs.

- **Software Development Kits:** SDKs deliver all the tooling users need to build and deploy AI applications across domains such as recommendation systems, conversational AI or video analytics. They include annotation tools for data labelling, pre-trained models for customization with transfer learning and SDKs that enable deployment across the cloud, the data center, or the edge for low-latency inference.

End-to-end accelerated Data Science and AI

Data science workflows have traditionally been slow and cumbersome, relying on CPUs to load, filter, and manipulate data and train and deploy models. GPUs reduce infrastructure costs and provide superior performance for end-to-end data science workflows using RAPIDS¹⁴ open-source software libraries. GPU-accelerated data science and AI workloads is available regardless of the location where GPUs are deployed, whether in the laptop, the workstation, in the data center, at the edge or in the public cloud.

The NGC repository has containers for RAPIDS which is used for GPU Open Data Science. It is a data science framework that is designed to have a familiar look and feel for data scientists working in Python. It also relies on and connects to many more open-source projects like Apache Arrow. RAPIDS is a suite of open-source software libraries and APIs for executing data science pipelines entirely on GPUs—and can reduce training times from days to minutes, also reducing energy consumption. RAPIDS unites years of development in graphics, machine learning, deep learning, HPC. It can run entire data science workflows with high-speed GPU compute and parallelize data loading, data manipulation, and machine learning for much faster end-to-end data science pipelines.

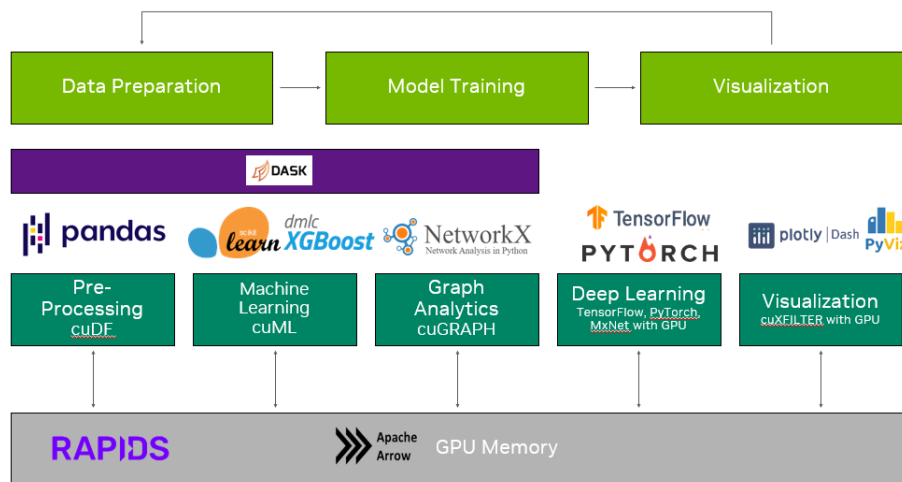


Figure 3: The same workflow as in Figure 1 but with GPU acceleration activated for the entire end-to-end data science workflow including AI, ML, ETL/pre-processing, network/graph analysis and filtering/visualization. There is a minimal change of the involved Python code using RAPIDS components like cuDF, cuML, cuGraph, cuXFilter as well as GPU-accelerated DL Python libraries like PyTorch and TensorFlow. Processing times due to GPU-acceleration can reduce training from days to minutes and there are also MLOPs-related tools to enable real-time inferencing.

RAPIDS can execute end-to-end data science and analytics pipelines entirely on GPUs. Integration into existing workflows normally requires only a few lines of code because

¹⁴ <https://rapids.ai/>

its API was deliberately designed to be consistent with existing data science utilities (e.g., Pandas DataFrame, SciKit Learn). RAPIDS is incubated based on extensive hardware and data science experience from many contributors. It exposes GPU parallelism and high-bandwidth memory speed through user-friendly Python interfaces. RAPIDS focuses on common data preparation tasks for analytics and data science. This includes a familiar data frame that integrates with a variety of machine learning algorithms for end-to-end pipeline accelerations without paying typical serialization costs. It also includes support for multi-node, multi-GPU deployments, enabling vastly accelerated processing and training on much larger dataset sizes.

These libraries democratize the power of GPU accelerated data science with observed accelerations from CPU to GPU that can range from a factor of 10x to 1000x in some cases.

Parallel Data Processing with Spark

Given the parallel nature of many data processing tasks, the massively parallel architecture of a GPU is also able to parallelize and accelerate Apache Spark data processing queries, in the same way that a GPU accelerates deep learning (DL) in artificial intelligence (AI). There has been implemented a GPU acceleration through the release of Spark 3.0 and the open-source RAPIDS Accelerator for Spark¹⁵. The RAPIDS Accelerator for Apache Spark uses GPUs to:

- Accelerate end-to-end data preparation and model training on the same Spark cluster.
- Accelerate Spark SQL and DataFrame operations without requiring any code changes.
- Accelerate data transfer performance across nodes (Spark shuffles).

As ML and DL are increasingly applied to larger datasets, Spark has become a commonly used vehicle for the data pre-processing and feature engineering needed to prepare raw input data for the learning phase. The Apache Spark community has been focused on bringing both phases of this end-to-end pipeline together, so that data scientists can work with a single Spark cluster and avoid the performance penalty of moving data between Spark based systems for data preparation and PyTorch or TensorFlow based systems for Deep Learning. Apache Spark 3.0 represents a key milestone, as Spark can now schedule GPU-accelerated ML and DL applications on Spark clusters with GPUs, removing bottlenecks, increasing performance, and simplifying clusters. In Apache Spark 3.0 there is now a single pipeline, from data ingest to data preparation to model training on a GPU powered cluster.

¹⁵ Further reading:

- <https://nvidia.github.io/spark-rapids/>
- <https://www.nvidia.com/en-us/deep-learning-ai/solutions/data-science/apache-spark-3/>
- <https://developer.nvidia.com/blog/gpus-for-etl-run-faster-less-costly-workloads-with-nvidia-rapids-accelerator-for-apache-spark-and-databricks/>
- <https://developer.nvidia.com/blog/accelerated-data-analytics-machine-learning-with-gpu-accelerated-pandas-and-scikit-learn/>

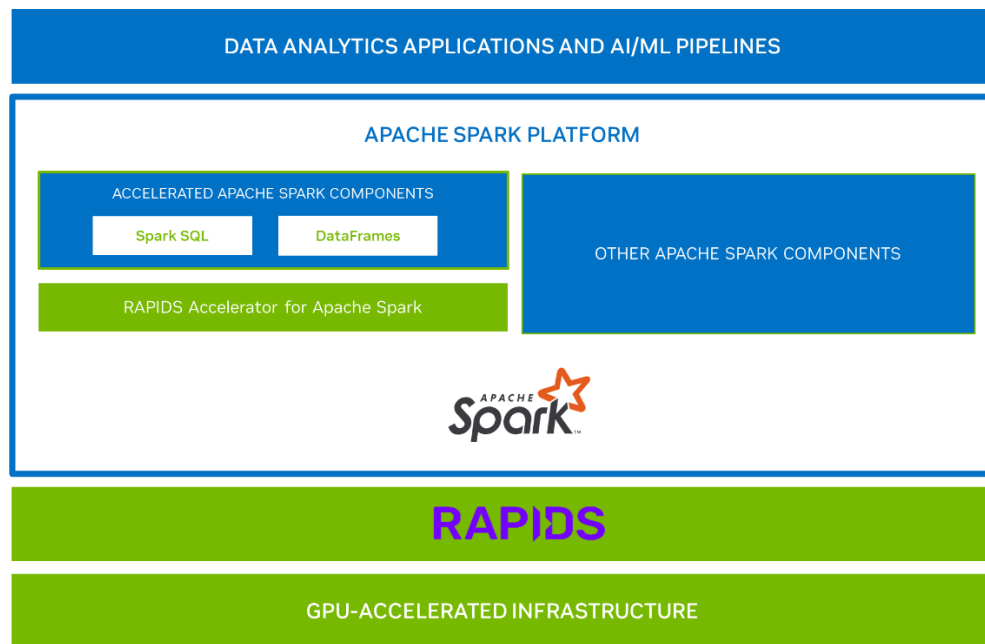


Figure 4: Overview of the Spark architecture leveraging the RAPIDS accelerator for data analytics applications and AI/ML pipelines.

Trustworthy, Explainable AI models

Another role of accelerated computing platforms in creating trustworthy, explainable AI models and in implementing automated AI model audit and evaluation. This goes far beyond MLOps which is a set of best practices for institutions to run AI successfully. Recent regulatory efforts like the draft AI Act released by the European Commission require transparent and trustworthy AI for certain more consumer-facing applications but central banks might want to adopt these principles, too.

There are a number of tasks to be addressed in AI assurance like post-hoc explainability layers (e.g. SHAP values based on cooperative game theory), testing for unwanted bias and adverse model reactions, visualizing large amounts of data and model outcomes, repeatable audits, impact assessments, unit tests, model drift detection, as well as techniques to support more human-centric and data-centric model building. High-performance computing is key in addressing all these dimensions of AI Assurance. As more AI/ML models are deployed around the globe it becomes clearer that the community has strong needs for tools and approaches that help to maintain trust and transparency in the models used. Researchers, developers, engineers, and architects are currently developing approaches and MLOps tools that support the implementation of AI quality, governance, trustworthiness, explainability and other approaches to ensure compliance with upcoming AI regulation and to assure the quality of AI, especially in high-risk AI applications. Modern computing platforms play a key role here as they can support building high quality of AI models and to test, benchmark, validate and certify them on an ongoing basis. Those platforms support internal teams for model risk management and model validation but also the entire TIC (test, inspect, certify) industry and the growing ecosystem of ethical AI startups.

RAPIDS also supports explainability of ML models. Model Interpretability aids developers and other stakeholders to understand model characteristics and the underlying reasons for the decisions, thus making the process more transparent. Being able to interpret models can help data scientists explain the reasons for decisions made

by their models, adding value and trust to the model. There are six main reasons that justify the need for model interoperability in machine learning¹⁶:

- Understanding fairness issues in the model
- Precise understanding of the objectives
- Creating robust models
- Debugging models
- Explaining outcomes
- Enabling auditing

RAPIDS provides GPU-accelerated model explainability through Kernel Explainer and Permutation Explainer. Kernel SHAP is the most versatile and commonly used black box explainer of SHAP. It uses weighted linear regression to estimate the SHAP values, making it a computationally efficient method to approximate the values.

Using a specialized tree based SHAP, a single GPU can provide explanations 20x faster than a 40-core CPU node for moderate-sized tree models, with even further acceleration possible for explanations of feature interactions.

An example in a bank could be to accelerate explainable AI implementation in managing risk in a loan book or credit portfolio. Credit risk management is a critical task for financial institutions, as it involves assessing the likelihood of borrowers defaulting on their loans. Traditional methods of credit risk management rely on statistical models and machine learning algorithms, which can be time-consuming and resource intensive. AI can be used to improve credit risk management by analyzing large amounts of data, identifying patterns, and making predictions about borrower behaviour. GPUs can accelerate data processing and AI workloads, making it possible to train models faster and more efficiently. Trustworthiness in those AI systems is important and can be achieved through techniques such as data validation, model interpretability, and explainability. GPUs can support this implementation e.g., by GPU-accelerated SHAP computations. The entire workflow from data preparation, model training, model inferencing and model explanation is accelerated by GPUs so real-live data sets for real-life portfolio sizes can be processed in minutes or a few hours, which is crucial in production.¹⁷ Visual dashboards and ad-hoc analytics tools for large data sets support the implementation of trust into the data science and AI workflow.¹⁸

¹⁶ <https://developer.nvidia.com/blog/model-interpretability-using-rapids-implementation-of-shap-on-microsoft-azure/>

¹⁷ The entire use case is presented here including code release: <https://developer.nvidia.com/blog/accelerating-trustworthy-ai-for-credit-risk-management/>

¹⁸ See example dashboard here: <https://dash-demo.plotly.host/nvidia-xai/practical-XAI/loan-default-dataset>



Figure 5: Interactive Plotly dashboard with focus in explainable and trustworthy AI for large scale data sets and models.

Graph Analytics and Graph Neural Networks

Learning from graph and relational data plays a major role in many applications. In the last few years, Graph Neural Networks (GNNs) have emerged as a promising new machine learning framework, capable of bringing the power of deep representation learning to graph and relational data. This ever-growing body of research has shown that GNNs achieve state-of-the-art performance for problems such as link prediction, fraud detection, target-ligand binding activity prediction, knowledge-graph completion, and product recommendations. The Deep Graph Library (DGL) is a DL library with efficient implementations for GNNs and demonstrate the speedup for GNN training and inference on GPUs. GPU-accelerated DGL containers will enable developers to work more efficiently in an integrated, GPU-accelerated environment that combines DGL and PyTorch.

cuGraph is paving the way in the graph world with multi-GPU graph analytics, allowing users to scale graphs with billions of nodes and edges. Accelerated algorithms for many common graph analytics tasks exist, across areas like centrality, community, link analysis, link prediction, and other traversal methods.

An example of leveraging this technology is the optimization of fraud detection based on financial data.¹⁹ Another example is detecting fraud with generative models, network analysis and synthetic data.²⁰

¹⁹ A related blog can be found here: <https://developer.nvidia.com/blog/optimizing-fraud-detection-in-financial-services-with-graph-neural-networks-and-nvidia-gpus/>

²⁰ A related blog can be found here: <https://developer.nvidia.com/blog/detecting-financial-fraud-using-gans-at-swedbank-with-hopsworks-and-gpus/>

Generative AI and Large Language Models in Central Banks

The acceleration of deep learning ignited the big bang of AI. ChatGPT, a large language model powered by a DGX AI supercomputer²¹, reached 100 million users in just two months. Its magical capabilities have captured the world's imagination. Generative AI is a new computing platform, like the PC, internet, and mobile cloud. Accelerated computing and AI have fully arrived.

The past months of developments in GenAI/LLM have dramatically pushed the boundaries in this area towards artificial general intelligence. These models are increasingly complex and trained on an increasingly large text corpus. There is empirical evidence for Power Law scaling in multibillion parameter models.²²

These kinds of models can create anything with a text structure, and not just human language text. It can also automatically generate text summarizations and has potential to automate tasks that require language understanding and technical sophistication. Examples show that it can interpret complex documents, launch actions, create alerts, or generate code.

And those models are not only about natural language but also other languages, images, video – even in a multi-modal way.

With massive size comes massive generalization ability: those models are competitive in many benchmarks without even tuning on the target task and the model still scales smoothly in performance instead of plateauing, implying that still-larger models would perform even better.

The generalization capabilities have a very meaningful impact and how we produce models and on the cost of model production. One can use a pre-trained large model and adapt it to new domains and task with the help of a few shots and prompts. The amount of labelled training data can thus be reduced, and this is very beneficial because many labelled data sets are expensive and can also contain labelling errors.

What does that mean Central Banks and Financial Supervisors? Generative AI (GenAI) and Large Language Model (LLM) systems will certainly be used to create better applications, like digital assistants and intelligent chatbots but the true power will come from its ability to ingest a wide variety of unstructured data and then to synthesize answers to natural language queries. GenAI is a smart language-powered interface to complex data, and other analytical and AI capabilities. Theoretically, every employee and staff member can become a researcher, knowledge worker or coder. Companies, organizations, and institutions will build AI factories for intelligence production, leveraging GenAI/LLM frameworks to train, customize, validate and deploy such models.

Here are some examples how Central Banks and Financial Supervisors can leverage those technologies and models, like building digital assistants including powering Avatars²³, enterprise search, summarization, translation, and report generation:

- predicting inflation and analyzing sentiment, using alternative data like text, images, videos, etc.
- document management: summarization and report generation will optimize middle/back office workflows

²¹ <https://www.nvidia.com/en-us/data-center/dgx-platform/>

²² Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. arXiv preprint arXiv:1712.00409.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling Laws for Neural Language Models. arXiv preprint arXiv:2001.08361.

²³ Like <https://developer.nvidia.com/omniverse/ace>

- research and surveillance activities in/of financial systems, markets and institutions based on huge amounts of unstructured data like text
- search and Q&A: optimized information retrieval by evaluating multiple sources, and summarizing results
- Transaction fraud: improves accuracy and generates reports, reducing investigations and compliance risk
- analysing sustainability and climate risk of the financial system but also the Central Banks' own initiatives like greening the financial system
- summarizing news feeds and market sentiment
- understanding the impact of the Central Banks' own activities, programs, projects and policies

Due to the disruptive, innovative, and transformative nature of GenAI/LLM, some of the following topics will be discuss, or are already being discussed in the communities of Central Banks and Financial Supervisors:

- Which use cases in FSI will be addressed and at which impact? How will this change financial service, financial markets, and financial centers? How will different segments of FSI, like banking, insurance, investments adopt?
- How will sustainable finance, ESG, climate risk, physical risk and biodiversity loss be addressed using GenAI and foundation models?
- Which other foundation models will we see in FSI besides natural language, like foundation models built on payments data or geospatial data?
- How will LLMs and GenAI be implemented, customized and how will adoption, flexibility, agility, and model performance be increased and how will cost, efficiency, and latency be decreased?
- Will we see more models in the cloud or rather in hybrid, or on-prem systems? What role will customization play? How will the optimal infrastructure, platforms, stacks, data processing technology and MLOps platforms look like?
- How will Trustworthiness, Security, Safety, Guardrails, Explainability and AI Governance be addressed? How will regulatory sandboxes look like? How will automated GenAI assessments and certification processes be established?
- How will large models be supervised and monitored? How will financial supervisors, regulators and central banks leverage those technologies for themselves and how will they monitor/supervise FSI activities?
- How will Avatars and Digital Assistants look like that are powered by LLMs?
- How will talents and startups develop and how will job profile change?

Challenges Of Developing and Deploying GenAI in Central Banks

There are some specific requirements for leveraging GenAI and LLMs in central banks like:

1. leveraging data that can only reside on-prem due to data privacy reasons
2. need for customization to better fit the the specific task, leverage proprietary data and for application trustworthiness.
3. safety and security reasons

And there are some challenges using foundation models without further customization:

Lack of Domain or Enterprise-Specific Knowledge:

1. Foundation models are trained on general datasets and may not possess the specialized knowledge required for specific domains or enterprises.
2. This limitation restricts their ability to provide accurate and relevant information in context-specific scenarios.
3. Without domain-specific knowledge, the models may struggle to understand and generate content that aligns with specific industry jargon, terminology, or practices.

Limited Adaptability to Changing Requirements:

1. Foundation models are static and do not have the inherent capability to adapt and evolve with evolving requirements.
2. As new trends, technologies, or business needs emerge, the models may become outdated and fail to provide up-to-date information.
3. Without the ability to continuously learn and integrate new knowledge, the models may lose their effectiveness over time.

Generation of Inaccurate or Undesired Information:

1. Foundation models, when used as is, can occasionally generate content that is inaccurate, misleading, or irrelevant to the user's needs.
2. This phenomenon, known as hallucination, occurs when the models generate information that appears coherent but lacks factual accuracy.
3. The generation of undesired or irrelevant information hampers the reliability and usefulness of the models for specific tasks or applications.

Risk of Bias and Toxic Information:

1. Foundation models may inadvertently reflect biases present in the training data, leading to biased outputs.
2. This bias can manifest in various forms, including gender, racial, or cultural biases, which can perpetuate unfair or discriminatory information generation.
3. Additionally, the models may inadvertently generate toxic or harmful content, such as hate speech or misinformation, which can have negative consequences if not addressed.

These challenges highlight the need for customization techniques to address the limitations of foundation models. Users can overcome these challenges and create more reliable, accurate, and tailored large language models that are specific to their domains or enterprises using specific customization methods.

Beyond the aspects of customizing foundation models there are other challenges of building and using them:

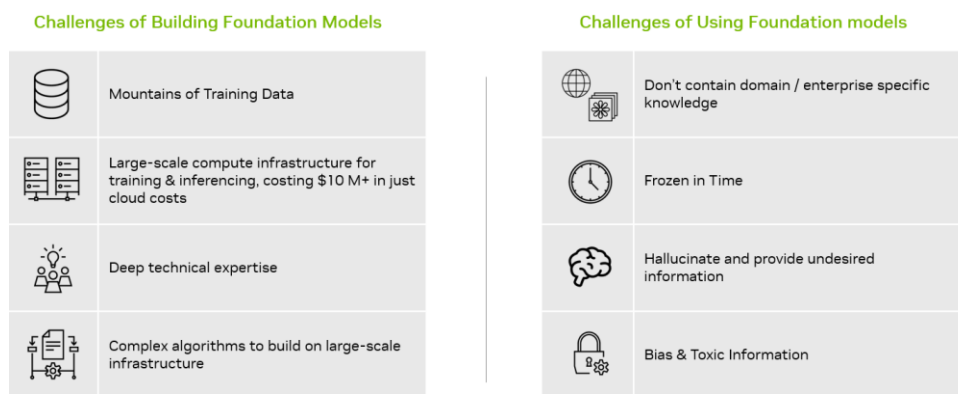


Figure 6: Challenges of building and using foundation models.

The solution is a flexible framework and a flexible, accelerated compute infrastructure plus some specific techniques and tools, e.g., to guardrail the models. ²⁴

Training such large models is an engineering challenge that has been solved by frameworks for efficiently training the world's largest transformer-based language models, based on Pytorch. The framework is for building, training, and fine-tuning GPU-accelerated speech, and natural language understanding (NLU) models with a simple Python interface and can be used to build models for real-time automated speech recognition (ASR), natural language processing (NLP), and text-to-speech (TTS) applications such as video call transcriptions, intelligent video assistants, and automated call center support.

Data curation tools and several engineering approaches to accelerated training like tensor and pipeline parallelism, sequence parallelism and selective activation recomputation are involved:

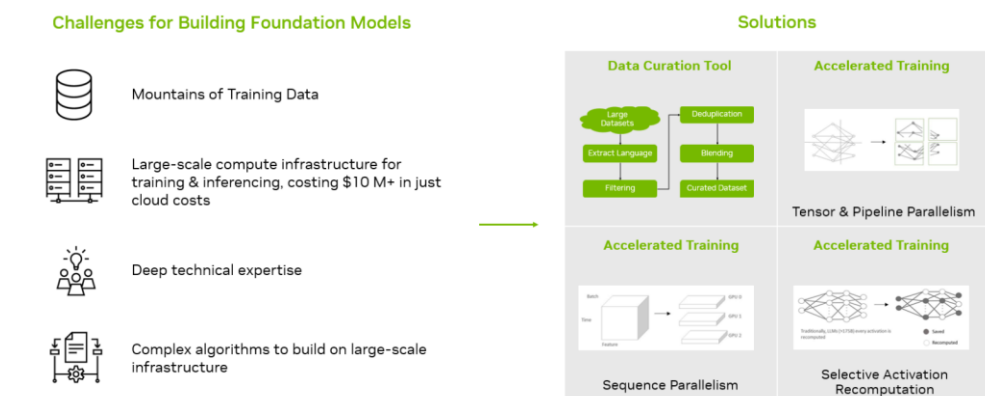


Figure 7: Solutions to meet the challenges of building foundation models.

Several customization techniques help to overcome the challenges using foundation models:

²⁴ <https://www.nvidia.com/en-us/ai-data-science/generative-ai/>

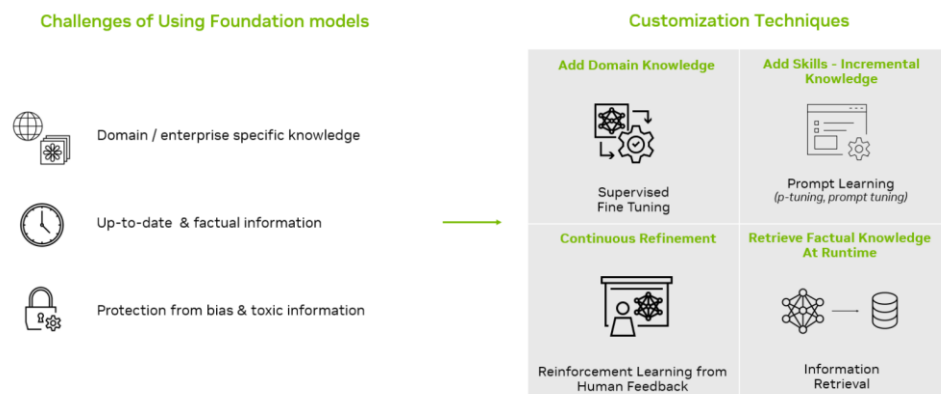


Figure 8: Solutions to meet the challenges of using foundation models.

Increasing the value of generative AI and foundation models in specific business use cases, institutions will increasingly customize pretrained models by fine-tuning them with their own data—unlocking new performance frontiers.

Extensive customization means building models from scratch or performing extensive fine-tuning.

There are a diverse range of customization techniques for generative AI models. These methods range from zero modification to the model's weights, all the way to substantial fine-tuning of every single parameter.

- **Prompting:** The simplest method of customization is by prompting, where no weights are changed within the model. This is a process of finding the right prompt to produce the desired output. As the original model's weights are not altered, the effectiveness of this method relies heavily on the ability to craft an appropriate prompt. It is somewhat of an art and a science, which involves understanding the tendencies of the model and crafting the prompts to leverage these tendencies for a specific task or output.
- **P-Tuning:** This technique stands for Prompt Tuning, where the weights of an additional small model, called the prompt encoder, are fine-tuned. In this case, we keep the generative model's weights fixed and only change the parameters in the external prompt encoder. This allows for better control and adaptability than basic prompting without modifying the primary model. It is considered a form of parameter-efficient fine-tuning, as the tuning happens only on a tiny external component, keeping the resource utilization significantly lower compared to traditional fine-tuning.
- **Parameter-Efficient Fine-Tuning (PEFT):** This technique aims to strike a balance between computational resources and customization. It involves tuning a small fraction, typically less than 1% of the total number of weights. This method includes strategies such as P-Tuning, Adapters, and LoRA (Low rank Adaptation). For example, in the adapter method, small adapter layers are inserted between the pre-trained layers of the model, and only these adapter parameters are trained. These methods aim to deliver high performance with less computational cost, making it an attractive option for many tasks.
- **Fine-tuning:** This is the most comprehensive method for model customization, where all model weights are adjusted. Two instances of fine-tuning are the Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) stages. Here, 100% of the model's weights are tuned to adapt to the specific task at hand. It is often used when the task is substantially different from the original pre-training task or when maximum performance is critical.

The choice of the method for customization depends on the task at hand, the available resources, and the level of performance required. Each of these methods has its trade-

offs and advantages, and a deeper understanding of these can help us make an informed decision. So, generative AI can be thought of as a spectrum, with the needs and solutions varying depending on where an enterprise customer falls on this spectrum. Understanding their specific requirements will help us determine the appropriate type of engagement.

Here is an overview of all customization techniques:

- Using domain knowledge with Supervised Fine Tuning:
 1. Fine-tuning the model with domain-specific data to incorporate relevant knowledge.
 2. Enhancing the model's understanding and generation capabilities within a specific domain.
- Adding incremental knowledge with prompt learning:
 1. Gradually training the model with additional knowledge through prompt-based learning.
 2. Expanding the model's understanding of new concepts and information over time.
- Reinforcement Learning from Human Feedback for continuous knowledge:
 1. Utilizing human feedback to train the model further and improve its performance.
 2. Reinforcing positive behaviors and refining the model's responses based on user input.
- Information retrieval to execute for runtime knowledge to answer proprietary information.
 1. Implementing information retrieval techniques to enable the model to access real-time knowledge.
 2. Allowing the model to dynamically retrieve and incorporate up-to-date information during runtime.

These customization techniques empower users to adapt large language models to their specific needs, enhancing their accuracy, relevance, and reliability by leveraging domain knowledge, incremental learning, reinforcement from human feedback, and runtime information retrieval.

Another important topic is guard railing models for building trustworthy, safe, and secure LLM conversational systems.²⁵ Enterprise use cases require guardrails to exclude everything outside functional domain, eliminate bias and toxicity, and to align to enterprise goals.²⁶

²⁵ <https://developer.nvidia.com/blog/nvidia-enables-trustworthy-safe-and-secure-large-language-model-conversational-systems/>

²⁶ <https://github.com/NVIDIA/NeMo-Guardrails>. NeMo Guardrails is an open-source toolkit for easily adding programmable guardrails to LLM-based conversational systems. It has programmable rails for LLMs with the following functionalities: steering the LLM towards producing outputs that accurately and effectively meet user intent, align the LLMs with the business goals of the enterprise, and prevent the model from generating undesirable, biased, or harmful content. The following guardrails have been implemented to date:

- Topical guardrails: prevent apps from veering off into undesired areas. For example, they keep customer service assistants from answering questions about the weather.
- Safety guardrails – ensure apps respond with accurate, appropriate information. They can filter out unwanted language and enforce that references are made only to credible sources.
- Security guardrails – restrict apps to making connections only to external third party applications known to be safe

Candidate Use Cases at Central Banks for Accelerated Computing

The mandate of central banks and financial supervisors has become challenging. Pandemics, economic instability, cyber risks, climate change, and sustainable investing are just a few examples of complex, emerging risks. Also, the digital, technological disruption of the financial system, as well as the speed of change and the growing complexity of the industry and its unbundling, increase the likelihood of supervisory blind spots.

For this reason, central banks, supervisors, and regulators will need to significantly adapt their operating model over the next decade to include technology and collaboration with (fin)tech firms. Many regulators are beginning to move in this direction with a variety of initiatives. Big Data Analytics and AI/ML will play a critical role, for example, in improving the quality and timeliness of risk identification and monitoring. Central banks already have access to vast amounts of valuable data, drawn from traditional, structured, and unstructured sources; streaming, complex, multi-layered, multi-modal, and alternative data to provide a holistic picture. Using advanced data collection and analytics techniques, certain areas of supervision will be able to use real-time monitoring of emerging risks and generate much earlier warning signals. Central bankers will be able to near cast and even predict the developments and provide a holistic picture of issues but also will be able to drill down into more granular micro developments and trends.

Considering these developments, data science and AI/ML models are becoming important tools used by central banks to fulfil their difficult mandate. Vast amounts of data are becoming available and numerous use cases in central banking are currently being developed. Data Science helps central banks model and analyse the economy and financial system -- AI/ML can support economic forecasting (e.g., for indicators such as inflation, housing prices, unemployment, GDP and industrial production, retail sales, external sector developments) and in business cycle analysis (e.g., compilation of sentiment indicators and use of nowcasting techniques to “forecast” the present).

General Uses of AI and ML

AI/ML can identify risk indicators, assess the behaviour of market participants, identify credit, and market risk, and monitor financial transactions and capital flows, detecting fraud, greenwashing, and assessing climate and economic risk.

AI/ML also supports financial risk assessment and surveillance exercises (functioning of the payment systems). It can detect market abuse with text mining techniques flagging misconduct or insider trading, spot odd patterns in the data, signalling the build-up of possible idiosyncratic vulnerabilities and identify network effects supporting the assessment of system-wide risks.

Using detailed data from the “bottom up” can lead to a substantial improvement in inflation forecasts so central banks are increasingly using more granular data and greater computing power to produce their forecasts²⁷.

Using large-scale accounting data for financial statement audits can help to monitor trustworthiness of financial statements and detect potential misstatements by applying neural networks to learn representations of accounting data that constitute a representative audit sample and using these systems to detect potential anomalies.

²⁷ <https://econpapers.repec.org/paper/boeboeewp/0915.htm>

Neural Networks for Simulating Markets & Analytics

Neural networks for exotic options and risk: complex derivative modelling can be improved by neural nets. The unavailability of analytical solutions, the higher dimensionality for complex interest-rate and foreign-exchange products (volatility surfaces, volatility smiles, multiple curves), and the derivative requirements used to hedge, pose unique challenges. The oracles (traditional modelling) can do complex computations, but they are expensive and slow, and are traditionally done on computer grids overnight. A modern GPU accelerated computing platform can obtain accurate valuations in well under a second.

The application of generative adversarial networks (GANs) and transformers to simulate market data for predictive analytics is another technique to benefit regulators and central banks. Generative adversarial networks are one tool for developing synthetic financial datasets that can be used as training data across many classes of machine learning models. GANs can complement and augment the value of Monte Carlo simulations and replicate regime-specific conditions to better prepare models for more robust predictive analytics. Using a generator and discriminator, with care, the model will transition the simulated data so that it converges to an empirical distribution in a Nash or Quasi-Nash process, preserving more of the real-world temporal characteristics consistent with the targeted market regime.

Systemic Risk Modelling

A special challenge is Systemic Risk Monitoring (SRM) and monitoring 'at scale.' The scaling problem in SRM is the analytical dimension as the number of analyses, indicators, processes, and aggregations increases exponentially:

- Complexity & size: the financial system is inherently complex, interconnected and adaptive
- Datasets exhibit very different formats, granularities and origins but must be cleaned up and combined; several layers of aggregation are needed (micro, meso, macro). There are layers and a taxonomy from areas to countries to sectors to groups to corporations; data quality issues will be prevailing.
- Speed: as markets shift rapidly, analytical outputs must be produced at a faster pace
- Besides economics and financial shocks and risks there are also climate risks (physical and transitional) including stranded assets.

A scalable data model is needed to model the entire financial system at any level as a dynamic multilayer network (multigraph). There need to be filters / aggregations / modifications preserving the data model enabling scalable operations on the data.

This is naturally modelled as a network with a hierarchical taxonomy of nodes representing different nested aggregation layers. The links between the networks are direct transactions or relationships in terms of contracts, ownerships etc. with different weights and intensities.

This graph is a quasi-knowledge graph or multilayered knowledge base that uses a graph-structured data model or topology to integrate data.

Once the graph is established the analytical part can be executed. The graph contains the complex nested relationships and graph/network analysis can be done looking for clusters, communities and spreading nodes, e.g. to understand contagion in such an interconnected system. Shocks can be simulated and links can be predicted.

To allow the use of knowledge graphs in various machine learning tasks, several methods for deriving latent feature representations of entities and relations have been devised. In recent years Graph Neural Networks (GNN)²⁸ have become popular.

²⁸ <https://developer.nvidia.com/gnn-frameworks>

Another attempt to better understand the graph is by visualization in interactive dashboard to understand the large structures but also to deep dive into very detailed local structures.

All mentioned operations in building, analyzing and visualizing the graph require heavy compute workloads, especially when the graph structure is rich and the data set is complex and large. Traditional CPU-only systems quickly encounter bottlenecks making an accelerated infrastructure for storing, analyzing and visualizing large graphs highly desirable.

There are many other network structures in financial supervision like payment/transaction networks that can be analysed and simulated with graph/network technologies. Propagation of shocks, risks, defaults, and fraudulent activity can be better understood with such models and technologies which helps to monitor systems and design policies. NLP (Natural Language Processing) can be used to recognize entities, link structures and types of relationships.

Outlier and Anomaly Detection

A classical application for AI / ML is automatic data validation and outlier control, e.g., of loan and securities microdata. An example is daily transactions in foreign exchange derivatives and interest rates executed in OTC market by financial intermediaries.

The outcomes of the models must be validated and checked for accuracy and plausibility. Results must be interpretable and controllable, and there must be the possibility of automatic selection of informative features. These jobs can be supported by machines as humans would not be able to reliably analyse such large and diverse datasets in reasonable time and quality.

A popular approach is XGBoost which provides a regularizing gradient boosting framework in many programming languages and frameworks. It often achieves higher accuracy than a single decision tree. There are frameworks to tune the hyperparameters and to extract information on feature importance and interaction, e.g., based on Shapley values.

Anomaly detection can also be applied in credit register data to detect reporting gaps and strange patterns. The evolutions of reporting can be overseen, and structural changes and inconsistencies can be identified. The quality deviations can be ranked according to severity. Data assessment processes can thus be much more effective and efficient.

Some modelers use several different AI/ML methods at the same time to build a robust ensemble of models. Typical methods are isolation forests, distance/density based KNN and autoencoders which can all be GPU-accelerated.

Another classical outlier detection application is in time series data, e.g., using unsupervised representation learning and clustering like DBSCAN or network filtering like minimal spanning trees using distance measures like Gower. GPU-accelerated packages for these procedures exist as well as for the compute intensive pre-processing for time series as well as for unstructured data like written reports. Deep learning approaches like LSTM (Long Short Term Memory) are one of the more recent approaches to validation of (financial market) time series.

Outlier and anomaly detection in economic dataset can identify sudden changes in data sets and identify deviations from trends. Deep analytical capabilities and complex data visualization help to gain greater familiarity with large data set and a deeper understanding of data and inherent complex patterns and relationships.

An example for ML for anomaly detection in datasets with categorical variables is the Money Market Statistical Reporting (MMSR) data including additional data sets that are usually added.

Anomalies can be detected in both labelled and unlabelled data, and they can be organized into multiple categories regardless of whether the original data was labelled.

Natural Language Processing (NLP) and Large Languages Models (LLMs)

NLP/LLM is a reliable powerhouse for Central Banks when it comes to processing of natural and non-natural language. Many different tasks can be executed with NLP like sentiment analysis, entity recognition, text summarization, question answering, topic modelling, translation, speech recognition, aspect mining and natural language generation.

NLP is a tool to translate unstructured information like text and voice into structured representations and also generating and responses with text and speech in much the same way as humans do.

An example is using NLP for the identification of topics in FOMC Transcripts from the Federal Open Market Committee Meeting Minutes. It could assist researchers at central banks and institutions to determine topic priorities.

Another example is an NLP tool to understand and assess information provided in text form, such as annual reports and audit reports or capital and liquidity assessments. Findings can be classified, plausibility can be checked, and reference could be provided to corresponding regulatory literature.

Another popular NLP area is analyzing central bank speeches and if they predict financial market turbulence²⁹. NLP can also improve central bank accountability and policy³⁰.

A classical and especially useful NLP task is sentiment analysis for economic forecasting. Increasingly, central banks have been relying on timelier indicators to assess the near term developments of the economy in advance of the release of official statistics. The news sentiment measures move with the fluctuations in economic conditions and can provide information in nowcasting movements in the less timely, quarterly survey-based business sentiment. News sentiment can help to forecast or at least nowcast private investment, CPI, inflation, and employment vulnerability. Data sources can also be alternative data from social media like Twitter. Adding explainability and visualization is important to understand how data and models come to conclusions.

Our chapter 'Generative AI and Large Language Models in Central Banks' described the topic and frameworks in-depth.

Computational Trustworthy AI, AI Governance, Trustworthiness and Explainability

A key issue in financial big data will be interpretable modelling. This is because to make evidence-based policy decisions central banks need to identify specific explanatory causes or factors which they can take action to influence. Furthermore, transparency regarding the information produced by big data analysis is essential to ensuring that its quality can be checked and that public decisions can be made on a sound, clearly communicated basis. Lastly, there are important legal constraints that reduce central banks' leeway when using private and confidential data; interpretable modelling helps address all these issues.

²⁹ <https://www.sciencedirect.com/science/article/pii/S1303070121000329>

³⁰ <https://www.cato.org/cato-journal/spring/summer-2020/how-natural-language-processing-will-improve-central-bank>

The focus of ML/AI models in central banking and supervision cannot be just predictive accuracy. Models must be trustworthy, interpretable, explainable, interactive, fair, robust, accountable, and secure. Proper risk management, data/AI governance, and compliance must be in place.

A bank, commercial bank or supervisor using AI in production needs to overcome the explainability gap to produce transparent, appropriate governance, risk management, and controls over AI. The publication “Financial Risk Management and Explainable, Trustworthy, Responsible AI” (Fritz-Morgenthal et al. 2021) discusses further details.

More specific use cases can be found, moving the discussion from the realm of the general to the specific. One related use case based on SHAP explanation values can be found in Bussmann et al. (2020). The use case had been selected as the best AI case in the EU Horizon2020 project FIN-TECH (www.fintech-ho2020.eu) by the European financial services community including the European supervisors. Other related Explainable AI (XAI) use cases can be found in Jaeger et al. (2021) and Papenbrock et al. (2021). The developed approaches can help to implement Explainable AI using techniques like Shapley values (a local, and global variable importance method with mathematical footings in co-operative game theory) even for large and complex models. For classical datasets, these methods can already substantially improve the transparency of portfolio allocation processes. They also enable the visualization of the variables and their influences of the entire data set in a single analysis. Clustering and network analysis of the variables and their influences are often used to find overall model structure and connections. Real-time monitoring of model drift in continuous learning machines is applied. Simulations and perturbations to test the robustness of the model can be run at large scale. Iterative and evolutionary approaches are now able to create and evaluate millions of models, allowing supervisors to select those that best balance prudential goals. It will also be necessary to meet the upcoming requirements from the European AI Act, especially the technical and auditing requirements for High-Risk AI³¹:

- Creating and maintaining a risk management system for the entire lifecycle of the system.
- Testing the system to identify risks and determine appropriate mitigation measures, and to validate that the system runs consistently for the intended purpose, with tests made against prior metrics and validated against probabilistic thresholds.
- Establishing appropriate data governance controls, including the requirement that all training, validation, and testing datasets be as complete, error-free, and representative as possible.
- Detailed technical documentation, including around system architecture, algorithmic design, and model specifications.
- Automatic logging of events while the system is running, with the recording conforming to recognized standards.
- Designed with sufficient transparency to allow users to interpret the system's output.
- Designed to always maintain human oversight and prevent or minimize risks to health and safety or fundamental rights, including an override or off-switch capability.

In summary, meeting the overall combination of the supervisory, legal, diverse stakeholder, and technical requirements will drive model development, deployment, monitoring, and retirement process that features enhanced auditability, transparency, and explainability. The ever-increasing data volume, velocity, and variety – across structured and unstructured sources – combined with the rapid pace of AI development will drive an overall system architecture that is scalable, flexible, and secure. To be efficient with both people's time and energy, the system must strongly adopt lessons learned in the leading HPC and AI supercomputers of today, leveraging

³¹ See <https://datainnovation.org/2021/05/the-artificial-intelligence-act-a-quick-explainer/>

GPU accelerated compute and networking that can accelerate workloads across the diverse, end-to-end data science use cases of today and tomorrow.³²

³² For additional information, see “Computing Platforms for Big Data Analytics and Artificial Intelligence” (Bruno et al. (2020)), which highlights the experiences of central banks with respect to HPC platforms.

Appendix: Selected accelerated libraries and frameworks

| Topic | Accelerated libraries and frameworks |
|---|---|
| Machine Learning | Rapids ³³ is an open source suite of accelerated Python libraries including XGBoost, which is also available through the open source XGBoost ³⁴ . |
| AI Development and Training | Performance optimized containers and code for PyTorch ³⁵ , TensorFlow ³⁶ and others. Enterprise Support ³⁷ available. |
| Graph Models | Accelerated libraries for graph neural networks (GNN ³⁸) and graph analytics(cuGraph) ³⁹ . |
| Language Models | Support for development (NeMo ⁴⁰) and open source repositories (HuggingFace ⁴¹) plus runtime support via Triton inference serving ⁴² . |
| Speech and Transcription Models | Customizable, multi-cloud speech to text, text to speech, and translation (RIVA ⁴³). |
| AI/ML Models in Production | CPU and GPU support for latency or throughput optimized production deployment of AI and tree based models(Triton). |
| Recommendation Engines | Support a variety of AI powered models for product or next best action recommendations (Merlin ⁴⁴). |
| Quantum Simulation & Integration | Support for quantum simulation (cuQuantum ⁴⁵) and integrating |

³³ <https://rapids.ai/>

³⁴ <https://github.com/dmlc/xgboost>

³⁵ <https://catalog.ngc.nvidia.com/orgs/nvidia/containers/pytorch>

³⁶ <https://catalog.ngc.nvidia.com/orgs/nvidia/containers/tensorflow>

³⁷ <https://docs.nvidia.com/ai-enterprise/index.html>

³⁸ <https://developer.nvidia.com/gnn-frameworks>

³⁹ <https://github.com/rapidsai/cugraph>

⁴⁰ <https://github.com/NVIDIA/NeMo> and <https://catalog.ngc.nvidia.com/orgs/nvidia/containers/nemo>

⁴¹ <https://huggingface.co/nvidia>

⁴² <https://github.com/triton-inference-server/server>

⁴³ <https://developer.nvidia.com/riva>

⁴⁴ <https://github.com/NVIDIA-Merlin/Merlin>

⁴⁵ <https://github.com/NVIDIA/cuQuantum>

| | |
|--|---|
| | traditional and quantum computers(CUDA Quantum ⁴⁶). |
| Spark and ETL / Data Processing | User transparent Spark plug-ins ⁴⁷ and Python libraries (Rapids). |
| Federated Learning | Federated learning using homomorphic encryption. (NVFlare ⁴⁸) |
| Visualization | Plotly Dash, cuxFilter (part of Rapids), and other libraries ⁴⁹ enable accelerated visualization and interaction with large amounts of data. |
| Mathematical & Scientific Computing | Comprehensive support via the HPC-SDK ⁵⁰ , cuNumeric ⁵¹ for open source Python computing. |

Acknowledgements

This work partially relies on the support from the European Union's Horizon 2020 research and innovation program "FIN-TECH: A Financial supervision and Technology compliance training programme" under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA).

The authors are grateful to Kevin Levitt and Marc Staempfli for valuable advice concerning this paper.

⁴⁶ <https://github.com/NVIDIA/cuda-quantum>

⁴⁷ <https://nvidia.github.io/spark-rapids/>

⁴⁸ <https://github.com/NVIDIA/NVFlare>

⁴⁹ <https://docs.rapids.ai/visualization>

⁵⁰ <https://developer.nvidia.com/hpc-sdk>

⁵¹ <https://github.com/nv-legiate/cunumeric>

References

- Ashley, John, Papenbrock, Jochen and Schwendner, Peter, (2022), "Accelerated Data Science, AI and GeoAI for sustainable finance in central banking and supervision" in Settlements, Bank for International eds., Statistics for Sustainable Finance, vol. 56, Bank for International Settlements, <https://EconPapers.repec.org/RePEc:bis:bisifc:56-23>.
- Bruno, Giuseppe, Hiren Jani, Rafael Schmidt, and Bruno Tissot. 2020. "Computing platforms for big data analytics and artificial intelligence." IFC Reports 11. Bank for International Settlements. <https://ideas.repec.org/p/bis/bisifr/11.html>.
- Bussmann, Niklas, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2020. "Explainable Machine Learning in Credit Risk Management." Computational Economics, September. <https://doi.org/10.1007/s10614-020-10042-0>.
- Fritz-Morgenthal, Sebastian and Hein, Bernhard and Papenbrock, Jochen, Financial Risk Management and Explainable Trustworthy Responsible AI (June 25, 2021). Available at <http://dx.doi.org/10.2139/ssrn.3873768>
- Jaeger, Markus, Stephan Krügel, Dimitri Marinelli, Jochen Papenbrock, and Peter Schwendner. 2021. "Interpretable Machine Learning for Diversified Portfolio Construction." The Journal of Financial Data Science 3(3). <https://doi.org/10.3905/jfds.2021.1.066>
- Li, A., S. L. Song, J. Chen, J. Li, X. Liu, N. R. Tallent, and K. J. Barker. 2020. "Evaluating Modern Gpu Interconnect: PCIe, Nvlink, Nv-Sli, Nvswitch and Gpudirect." IEEE Transactions on Parallel and Distributed Systems 31 (1): 94–110. <https://doi.org/10.1109/TPDS.2019.2928289>.
- Mattson, Peter, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, et al. 2020. "MLPerf Training Benchmark." <http://arxiv.org/abs/1910.01500>.
- Papenbrock, Jochen, Peter Schwendner, Markus Jaeger, and Stephan Krügel. 2021. "Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios." The Journal of Financial Data Science, March, jfds.2021.1.056. <https://doi.org/10.3905/jfds.2021.1.056>.
- Radhakrishnan, Ramesh, Yogesh Varma, and Uday Kurkure. 2019. "Evaluating Gpu Performance for Deep Learning Workloads in Virtualized Environment." In 2019 International Conference on High Performance Computing & Simulation (Hpcs), 904–8. IEEE.
- RAPIDS-Spark. 2021. "RAPIDS Accelerator for Apache Spark." GitHub Repository. <https://github.com/NVIDIA/spark-rapids>; GitHub.
- Serena, Jose Maria, Bruno Tissot, Sebastian Doerr, and Leonardo Gambacorta. 2021. "Use of big data sources and applications at central banks." IFC Reports 13. Bank for International Settlements. <https://ideas.repec.org/p/bis/bisifr/13.html>.
- Tissot, Bruno, Timur Hulagu, Per Nymand-Andersen, and Laura Comino Suarez. 2015. "Central Banks' Use of and Interest in "Big Data"." IFC Reports. Bank for International Settlements. <https://EconPapers.repec.org/RePEc:bis:bisifr:3>.
- Tissot, Bruno. 2018. "Big Data for Central Banks." In International Workshop on Big Data for Central Bank Policies–Bali, 23:25.
- Zeranski, Stefan, and Ibrahim Ethem Sancak. 2020. "Digitalisation of Financial Supervision with Supervisory Technology (Suptech)." J. Intl. Banking L. & Reg., no. 8. <https://ssrn.com/abstract=3632053>.



MODERN COMPUTING PLATFORMS AS KEY CENTRAL BANKING TECHNOLOGY

BY DR. JOCHEN PAPENBROCK, FINANCIAL SERVICES AND TECHNOLOGY DEVELOPER RELATIONSHIP LEAD EMEA

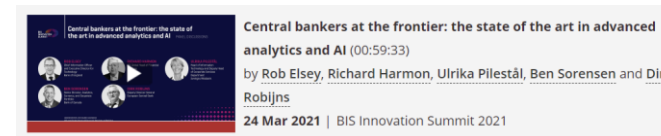
jpapenbrock@nvidia.com

DATA SCIENCE AT CENTRAL BANKS

We have analysed (and contributed to) the data science use cases and IT/software tools the central banks are currently establishing.

Sources:

- IFC/BIS publications:
 - "Computing platforms for big data analytics and artificial intelligence"
 - "Big data and machine learning in central banking"
 - "Big data for central banks"
 - "The supotech generations"
 - "The use of big data analytics and artificial intelligence in central banking"
 - "Central Bank Communications: information extraction and semantic analysis"
- Own projects and interactions with some of the leading central banks globally



Observation

Central banks embrace Big Data and AI with typical tools and workflows but acceleration is not yet realized.

TYPICAL WORKFLOW AND TOOLS

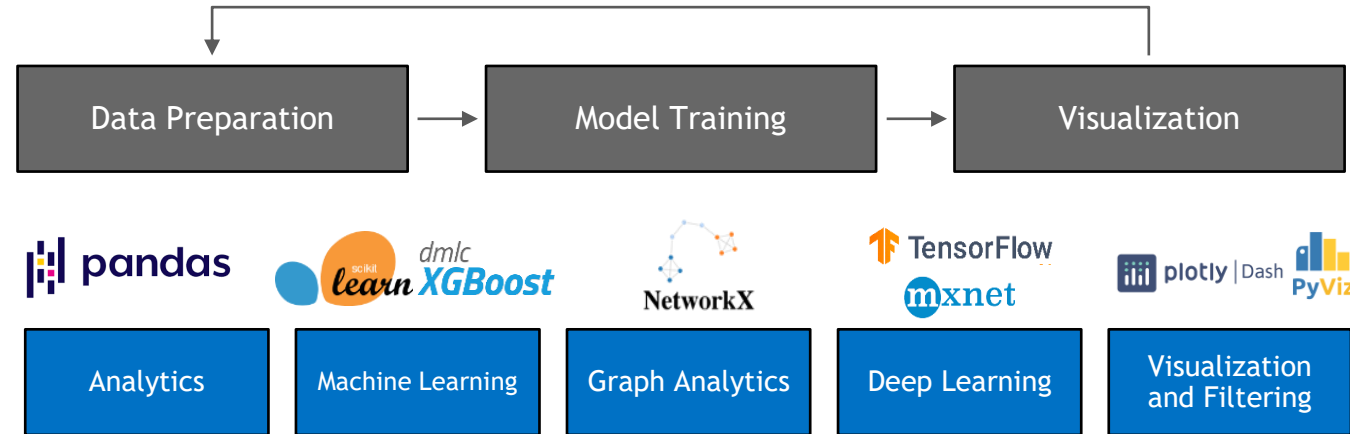
Open-source projects and Python tools have democratized data science


SQL and graph data base





 NumPy



There are computational bottlenecks with CPU-only processing

GPU SUPPORTS PARALLEL MASS CALCULATIONS

IFC Report
No 11

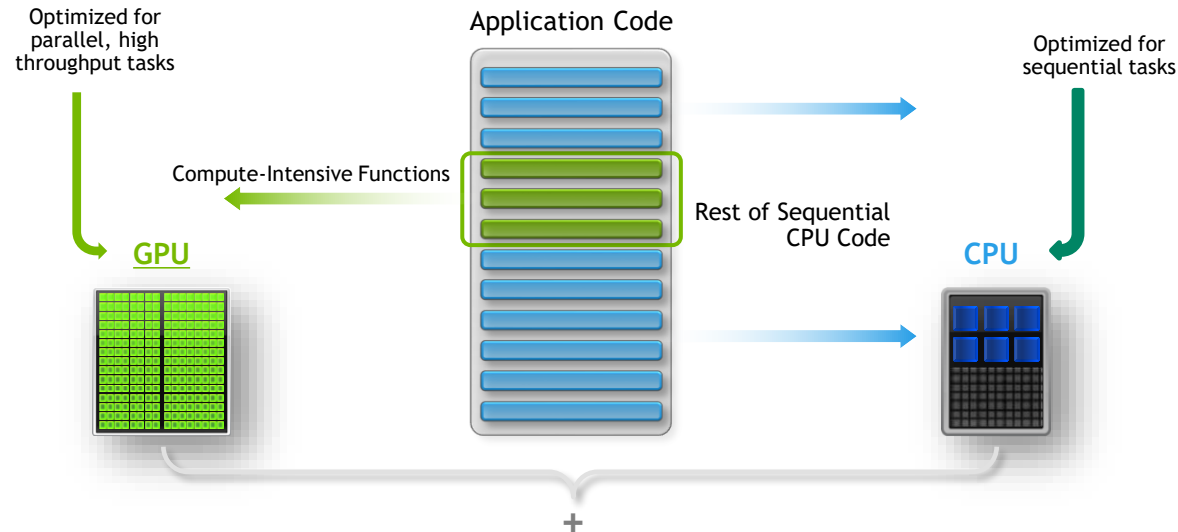
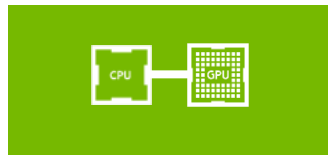
Computing platforms for
big data analytics and
artificial intelligence

April 2020



BANK FOR INTERNATIONAL SETTLEMENTS

“Depending on the analytical or statistical problem at hand, clusters of GPUs (graphics processing units, which have a highly parallel structure and were initially designed for efficient image processing) might also be embedded in computers, for instance, to support mass calculations.”



WHAT HAPPENS WHEN WE EMBED GPU?

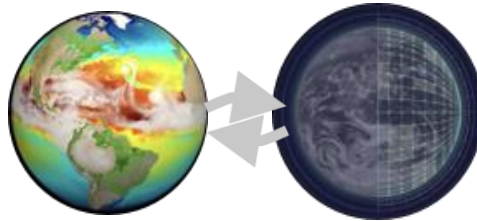
CAMBRIDGE-1

Boosting COVID-19 research



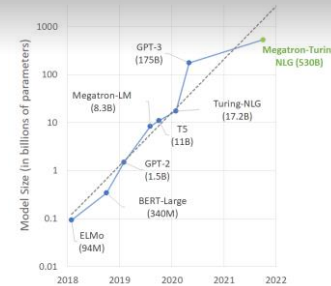
SUSTAINABLE FINANCE

Earth digital twin to forecast climate change



LARGE LANGUAGE MODELS

Of the size of GPT-3 and Megatron-Turing (530B)



AI RESEARCH

Research SuperCluster (RSC) with Meta / Facebook



Accelerated Computing puts Big Data Processing, AI, Simulation, and Visualization to a new level

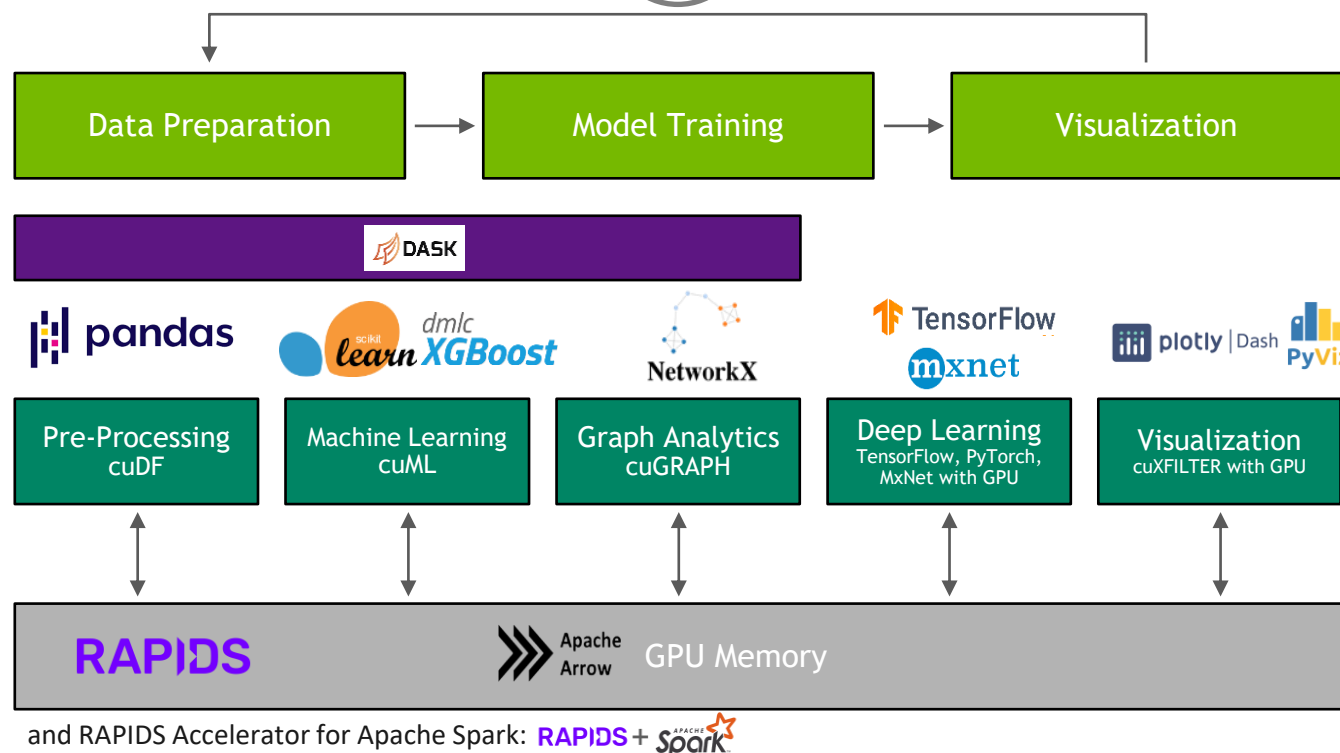
EXAMPLES OF ACCELERATED APPLICATIONS

1

| SDK | NLP Task |
|-----------------|--|
| NeMo / Megatron | Named Entity Recognition Topic Modeling Intent Identification Relation Extraction Sentiment Analysis Language Translation Text Summarization |
| RIVA | Speech to Text |
| TAO | Model Training (Computer Vision & NLP) |
| TensorRT | Model Optimization |
| Triton | Inference Serving |

Accelerated NLP to
address ESG needs

2



Accelerated end-to-end data science

ACCELERATED COMPUTING PLATFORM

1. Chips & Systems
2. Platform Software
3. Application Frameworks



Platform Symbiosis
Hardware & Software

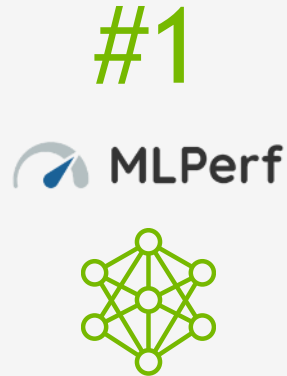
2.5M
Developers

30M
CUDA Downloads

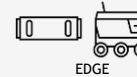
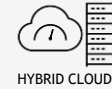
2,500
GPU-Accelerated Applications

9,000
AI Startups

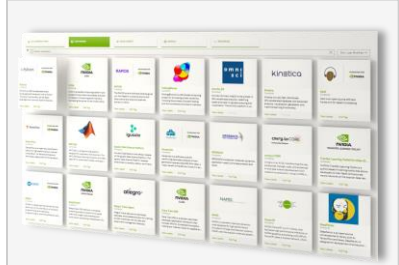
Ecosystem
Developers & Partners



Productivity
World Records



Available Everywhere
Every Cloud and OEM



150 SDKs (often pythonic)
Ready Application Frameworks

OUTCOMES

- increases developer productivity and scalability
- reduces TCO, time to insight and infrastructure complexity

AI TRUSTWORTHINESS AND AI GOVERNANCE

- We are engaged in numerous projects, webinar series and software projects



- Development of new XAI workflows in investment management with Munich Re

Existing infrastructure



DGX Station with
4 V100 32GB GPU

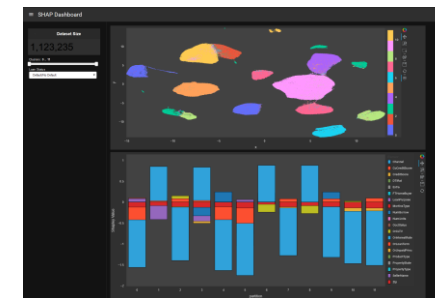


RAPIDS



- FAIC (Financial AI Cluster)

Stimulating an ecosystem around a collaborative technology platform for automating compliance with AI regulation, AI governance, AI assurance



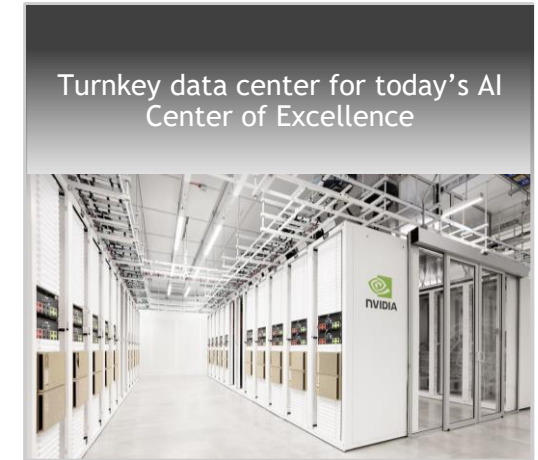
Computation of Explainable AI based on an arbitrary AI black box model
Interactive Exploration of the entire model

SUMMARY AND OUTLOOK

- There is a rise of AI/HPC workload in enterprises and organizations around the globe
- There is a need for a new generation of computing platforms
- Many organizations build their AI Center of Excellence including the appropriate infrastructure
- We help with hardware, software, ecosystem and know-how



With many developers and technology companies presenting, registration is free





CONTACT

Dr. Jochen Papenbrock

jpapenbrock@nvidia.com

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

The data analytics lab – from innovation to products¹

Mona Amer, Hiren Jani and Mathieu Le Cam,
BIS

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



The Data Analytics Lab

From innovation to products

Mission and Vision statements

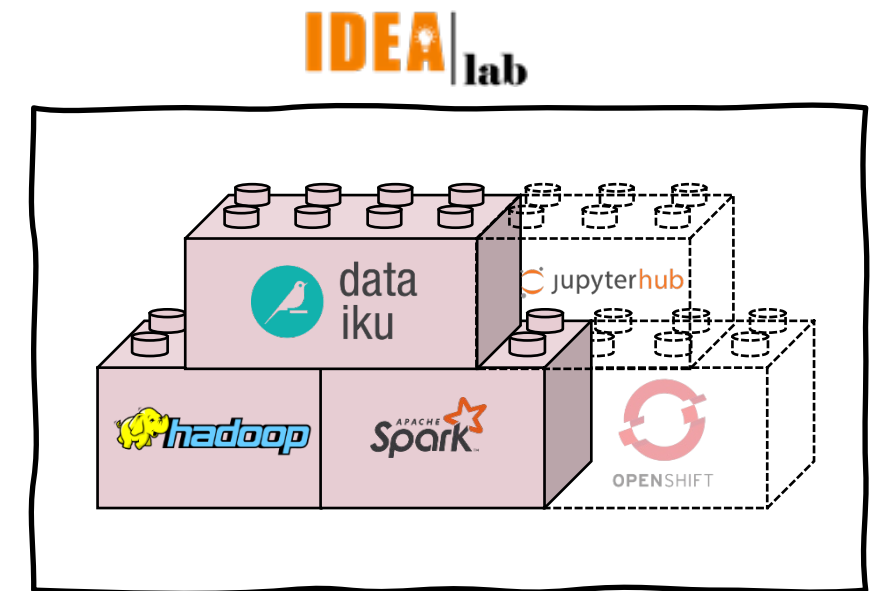
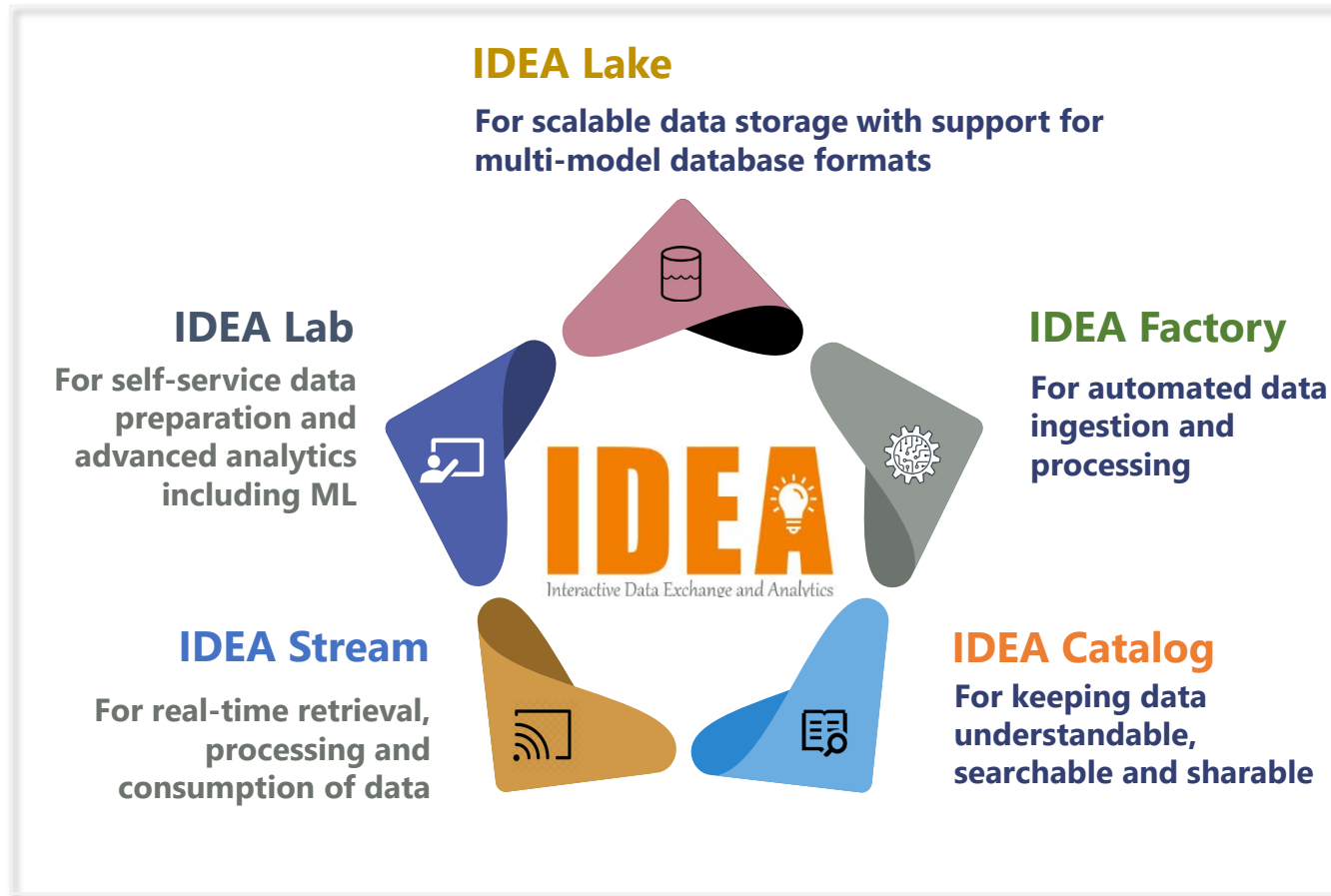
MISSION: provide **self-service platform** that allows users to use data science, AI and flexible data wrangling capabilities to turn data into insight.



VISION: improve efficiency and collaboration across data analytics community in the Bank by providing data lab solutions that are **easy to use, to scale, to govern, to operationalise** and **to integrate** with existing and new data and analytical systems.

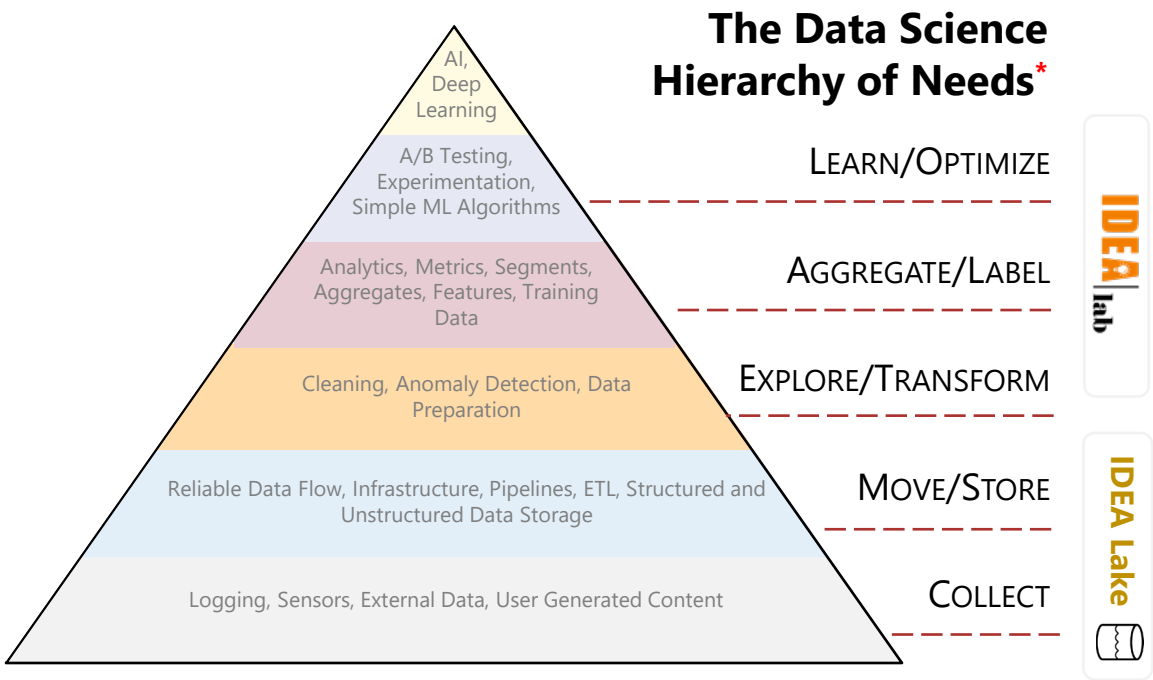


IDEA Lab & the BIS Big Data Platform

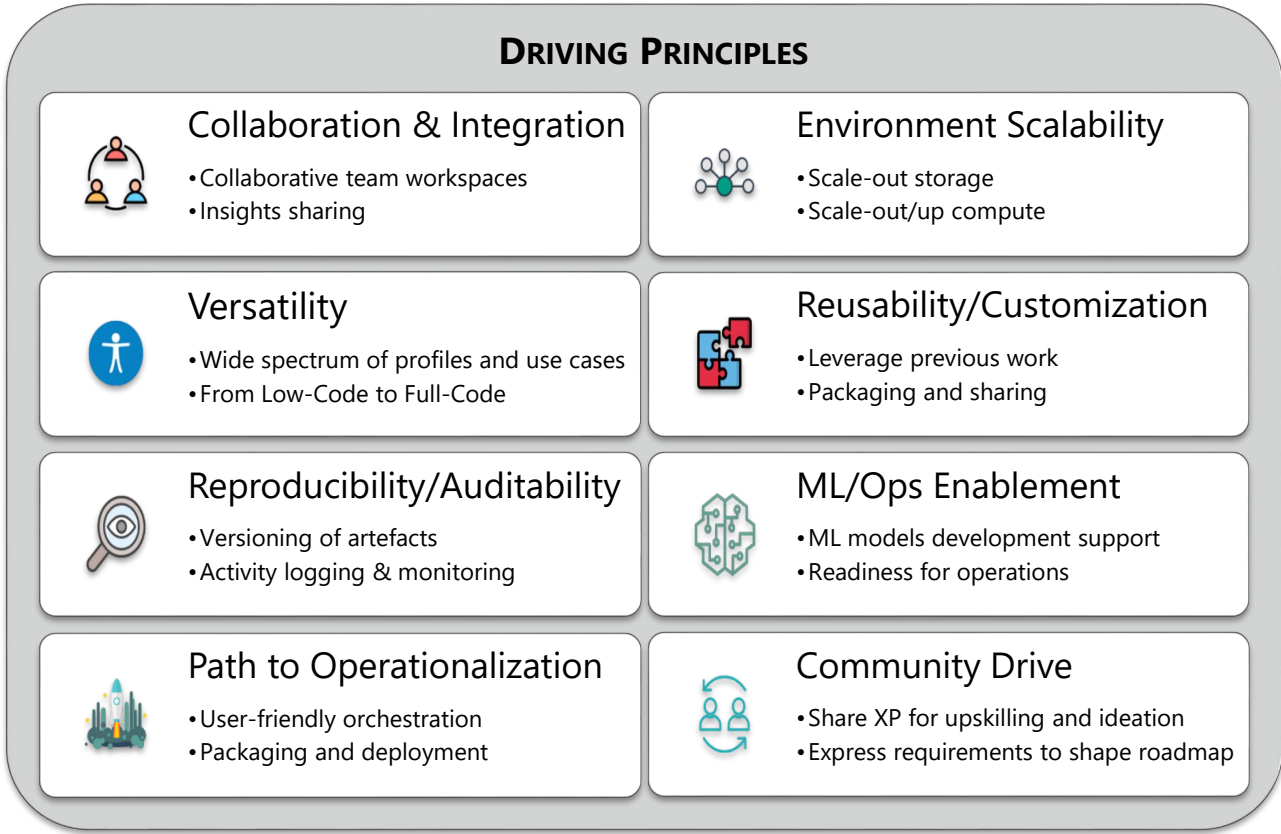


The IDEA Lab design components offer data scientists and analysts greater agility and innovation to maximize value derived from diverse data sources in an efficient, yet, simple manner

How IDEA Lab Delivers Value?







The IDEA Lab goes above and beyond the needs of a successful Data Science and Advanced Analytics Implementation



* Based on <https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007>

Spotlight on products

| | Financial Projection | D2NLP | TweetWatcher |
|---|--|---|--|
|  | <ul style="list-style-type: none"> • Mid-term projection of the Bank's key financial measures • Support discussion regarding dividend policy | <ul style="list-style-type: none"> • NLP to support document management • Help human curate very large document stores | <ul style="list-style-type: none"> • Analyze tweets related to the Central Banking community • Identify trends and topics of interest |
|  | <ul style="list-style-type: none"> • BIS internal financial information & reporting structure • BIS interest rates projection | <ul style="list-style-type: none"> • Documents, archives, correspondence • Millions of text documents in various formats (Word, PDF, etc) | <ul style="list-style-type: none"> • User tweet & mention timelines • User details and metrics • Filtered stream – real time feed of public tweets |
|  | <ul style="list-style-type: none"> • Financial data preparation and consolidation • User-defined projection scenarios & parameters | <ul style="list-style-type: none"> • Apply language models to compute similarity scores between documents • Documents summarization via BERT embeddings • Topic modelling on ad-hoc, user-defined collections | <ul style="list-style-type: none"> • Data denormalization and enrichments • Hashtags and entities extraction • Language detection and offline neural machine translation of non-English tweets • Sentiment analysis of user mention timelines |
|  | <ul style="list-style-type: none"> • Configurable financials projection solution • Automated reporting and dissemination of projection results | <ul style="list-style-type: none"> • Database of potential duplicate with metadata, summaries and analysis results • Web application for exploration and analysis requests | <ul style="list-style-type: none"> • Growing database of enriched Twitter datasets • Filtered stream rules management solution • Executive dashboards to deliver key insights |

Challenges & Enhancements



Challenges of Establishing and Running a Data Lab

- ✓ **Resistance to change** – how to move from “comfort-zone” to “growth-zone”
- ✓ **Keeping up with the needs** – to meet swiftly new connectivity requirements, customizations and process capabilities
- ✓ **Making data accessible** – set up a process to discover, find and get access to datasets, but also share and ensure quality
- ✓ **Finding the right governance** – product criticality varies from initiative to initiative, one size doesn’t fit all



Envisaged Enhancements

- ✓ **Community development** – increase knowledge sharing sessions, schedule dedicated technical workshops
- ✓ **New compute options** – complementary software, scale out/up processes with container orchestration & GPU computing
- ✓ **Cloud-based Innovation Zones** – extensions to the IDEA Lab with higher degree of freedom, to facilitate experimentation with non-sensitive data

Lessons Learned



Self-service is (indeed) a catalyst for innovation



Upskilling is critical, requires time and close collaboration between units



Balance between freedom and governance is essential



The frame of a platform constrains but stimulate delivery



Mindset and culture are keys to adoption and take time to change



THANK YOU!

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Containerization for research collaboration: platform independent economics¹

Venkat Balasubramanian and Kim Huynh, Bank of Canada;
Danielle Handel, Stanford University; Anson Ho, Ted Rogers School of Management
David Jacho-Chávez and Carson Rea, Emory University

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Containerization for Research Collaboration: Platform Independent Economics

Venkat Balasubramanian, Danielle V. Handel, Anson T. Y. Ho, Kim P. Huynh, David T. Jacho-Chávez, Carson H. Rea

Abstract

We present containerization as a tool to facilitate collaboration between researchers across varying institutions and platforms. Docker aids in construction and deployment of containers, and we introduce the relevant concepts and tools for the interested researcher. Advantages of using containers in economic research include enhanced reproducibility, efficiency, portability, and ease of collaboration. Mirroring the empirical use case of Handel et al. (2021a), we use containerization in tandem with cloud computing resources to expedite a computationally intensive spatial analysis mapping access to financial services in Canada.

Keywords: Containerization; Docker; Spark, Azure, AWS.

JEL codes: A11; C87; C88.

Contents

| | |
|--|---|
| Introduction..... | 2 |
| Containers..... | 2 |
| Docker..... | 3 |
| Constructing a Docker container image..... | 4 |
| Container management alternatives and supplements..... | 5 |
| Containers on the cloud..... | 5 |
| Empirical use case..... | 5 |
| Conclusion and considerations..... | 6 |

1. Introduction

Research that supports the functions of central banks increasingly necessitates collaboration across banks, universities, and other stakeholders, requiring more advanced technology to effectively communicate. We have identified a suite of tools that can facilitate rapid, reproducible research across time zones and infrastructures. Our previous work highlights the utility of cloud platforms in research, focusing particularly on Azure Databricks and the automation of resource handling through Spark. We note that cloud computing enables researchers to easily leverage high performance computing for computationally intensive tasks (Handel et al., 2021a). However, there is an obvious practical barrier to implementing this approach: the ability to use platforms like Databricks is conditional on institutional access to cloud platform memberships. This may not be available for all collaborators, or preferred vendors may differ across institutions. In this paper, we present containerization as another collaboration tool with no barriers to access and further demonstrate its compatibility with cloud platforms.

The rest of this paper is organized as follows: Section 2 describes containers and their uses, Section 3 introduces Docker, provides an example to illustrate the construction of Docker containers, and demonstrates the portability of containers by leveraging cloud platforms. Section 4 briefly revisits the empirical example following Handel et al. (2021a) executed with containerization, and Section 5 concludes by highlighting the advantages and considerations associated with containerization for research collaboration.

2. Containers

Containers are standalone, executable packages of software that include everything needed to run an application: code, runtime, system tools, system libraries and settings. With all dependencies included, containers act to create an environment isolated from the user's host operating system, meaning that program functionality and output are not dependent on particular settings and software versions that may differ across collaborators. Containers and virtual machines (see Handel et al., 2021b) have similar functionality. However, in the case of containers, the virtualization occurs at the operating system level, as opposed to the hardware level. Consequently, containers are lighter and faster to use. As noted by Boettiger (2015), this enables researchers to run even 100's of containers on a standard laptop. Figures 1a and 1b illustrate the architecture used by containers (a) as compared with that of virtual machines (b). Note that, for containers, the applications sit on top of the shared container engine (Docker in this case), which in turn uses the host operating system.

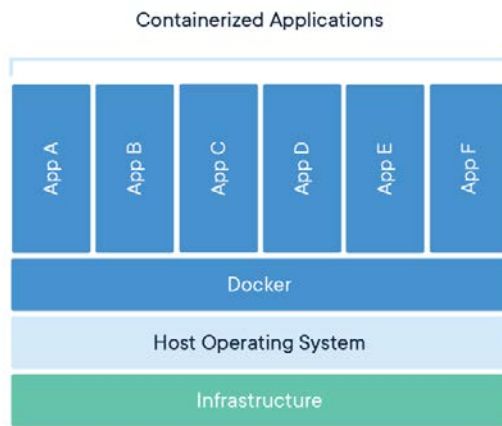


Figure 1a: Container ecosystem

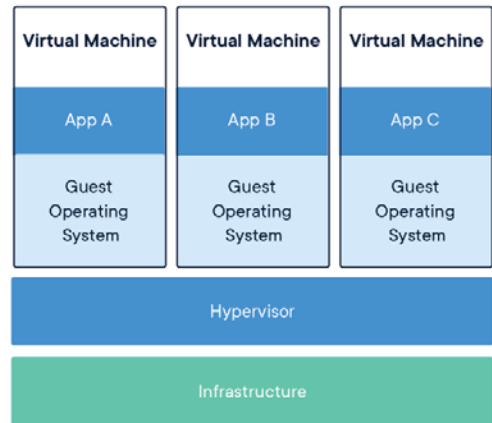


Figure 1b: Virtual machine ecosystem

3. Docker

Docker is an open-source project that streamlines the construction and use of containers. While the official Docker documentation provides a complete and robust set of directions for getting started in implementing containerization, we will briefly introduce the relevant concepts. Researchers can use platforms like Docker for the simplified management and deployment of Windows- and Linux-based container instances using the tools in Table 1.

| Docker tools | | Table 1 |
|--------------|---|---------|
| Docker CLI | A command line interface for managing container instances | |
| Dockerfile | Text-based file with a list of instructions for image assembly | |
| Image | Container template and metadata (how the application is stored) | |
| Container | A running instance of an image (how the application is run) | |
| Build | Command line action to construct an image according to a Dockerfile | |
| Docker Hub | A cloud registry for distributing and storing Docker images | |

Users leverage these tools to create self-contained packages by first creating a text-based script with the instructions to make an image in a Dockerfile, which requires some basic knowledge of shell scripts. These instructions may include loading a pre-built image from Docker Hub and adding additional data and dependencies manually. Using the command line, a user then builds the image from

the Dockerfile specifications using the Docker build command. The Docker up command is used to create a running instance and launch the program in a container. To facilitate exchange, researchers can push their images complete with code, input data, and any dependencies to Docker Hub, where they can then be pulled down by collaborators.

3.1 Constructing and deploying a Docker container image

The easiest way to construct a container is using prebuilt images, which can be found in and pulled from Docker Hub. The use of prebuilt images allows developers to avoid unnecessary building, expediting the creation of the final product. The use of prebuilt images also reduces deployment latency while increasing the success rate.

Popular prebuilt environments include Postgres, Ubuntu, and Python, each having over one billion downloads. By pulling one of these images, say Python, the created container would install everything needed to execute code written in Python, even if the local machine does not have it installed. All future containers that attempt to pull the Python image would no longer require the download, expediting the deployment process. Consequently, a container's initial deployment may take the longest compared to subsequent uses. This holds true for any of the over 8 million prebuilt images found on Docker Hub. Figure 2a displays the Dockerfile that will be utilized in the use case located in Section 4. As outlined, line 1 of the Dockerfile begins "FROM docker.io/bitnami/spark:test" which indicates that the container will be pulling a pre-build image to be used, in this case Spark. Likewise, Figure 2b displays the accompanying .yml file that will be called when the user inputs "docker-compose build" or "docker-compose up."

Another option to consider when building a container is whether to include end-to-end platform capabilities. Containers can be set up in such a way so that a user can take full advantage of an IDE or Jupyter Notebook alongside the working environment. Further, a working file system can be mounted. The setup required is beyond the scope of the paper but demonstrates the extent to which containerization can supplement a workflow.

```
0 lines (6 sloc) | 249 Bytes
1 FROM docker.io/bitnami/spark:test
2
3 USER root
4 RUN pip install pyspark && \
5     mkdir -p /usr/local/src/app
6
7 WORKDIR /usr/local/src/app/program
8 ENTRYPOINT [ "spark-submit", "--verbose", "--master", "local[*]", "--driver-memory", "10", "distance.py"]
```

Figure 2a: Dockerfile

```
16 lines (15 sloc) | 374 Bytes
1 version: '3'
2
3 services:
4   spark:
5     build:
6       context: .
7       dockerfile: Dockerfile
8     image: b1s_container_demo
9     environment:
10      SPARK_MODE: master
11      SPARK_RPC_AUTHENTICATION_ENABLED: "no"
12      SPARK_RPC_ENCRYPTION_ENABLED: "no"
13      SPARK_LOCAL_STORAGE_ENCRYPTION_ENABLED: "no"
14      SPARK_SSL_ENABLED: "no"
15     volumes:
16       - ./usr/local/src/app/
```

Figure 2b: docker-compose.yml

3.2 Container management alternatives and supplements

Singularity is an alternative platform for the construction and deployment of containers. It allows for the same functionality as Docker, specifically optimized for the deployment on high-performance computing (HPC) clusters. The Docker Registry is also accessible by Singularity, enabling the same expedited build process. Docker images can be loaded and implemented in Singularity containers without having to go through the installation process for Docker. Singularity is not compatible with Windows, limiting its functionality to Mac OS and Linux. To demonstrate the full potential of platform independent computing, our use case implements Docker.

Kubernetes supplements containerization by managing and automating services and workloads. Kubernetes is a portable, open-source framework which orchestrates containers with the aim of providing a smooth expedited workflow. Kubernetes distributes network traffic to stabilize app deployment, allows for storage to be automatically mounted, and enables self-healing by restarting, adding, or killing containers as needed. These management services, along with all others Kubernetes provides, are all aimed towards the facilitation of app deployment, so focus can be placed on the research rather than the infrastructure.

3.3 Containers on the cloud

While containers can be built and run locally, the same can be done on virtual machines, independent of platforms. Azure and AWS, products of Microsoft and Amazon respectively. Virtual machines have the same functionality regardless of platform, although Azure requires a Windows operating system. It is worth noting that while entirely possible, the construction and running of containers is not optimized for Windows. In fact, our use case makes use of both Azure and AWS to highlight equal compatibility. If using AWS, selecting Ubuntu as the processor is likely the easiest and most compatible with containerization, as Linux comes out of the box with everything required for virtualization and Git. For instructions on how to create a virtual machine, see Handel et al. (2021b), which uses AWS hosted VMs to support a virtual econometrics laboratory.

4. Empirical use case

We present an empirical use case for containerization by completing an exercise which maps consumer access to financial services in Canada as introduced by Handel et al. (2021a). We use Canadian postal codes to their closest bank branch and examine trends in access. Given consumer reliance on physical bank branches for the purchase of complex financial products (Mintel, 2018) or first-time banking interactions, this provides vital insight into the larger concern of consumer access to financial services. The Bank of Canada is also particularly interested in how consumers are connected to the supply of cash, in which physical bank branches play a large role (Chen & Strathearn, 2020).

The consumer location data we use are shape files of the 24,000 Canadian Postal Codes, which we obtain from Statistics Canada. We use the Statistics Canada Postal Code conversion files to link this data with demographic information from the 2016

Canadian Census. Data on the locations of the 14,000 physical bank branches is obtained from the Financial Institutions File from Payments Canada.

Canadian postal codes are compact regions, often comprising a single block. So, this turns out to be a relatively computationally intensive task, with a high volume of postal codes and bank branches. We complete this task by deploying a Docker container including Python code with PySpark and an optimized SQL nearest-location finding algorithm for managing the large volume of input data, which are also hosted on the container. See Handel et al. (2021a) for a brief discussion of the empirical findings. To address the computational needs, we also deploy the container on the cloud using an Amazon AWS EC2 instance running a Linux operating system and a Microsoft Azure virtual machine running Windows 10.

Before containers can be constructed or deployed on any machine, Docker must first be installed. While the deployment of containers will be consistent between platforms, the installation process for Docker itself may differ because it must be done on the hardware level like any normal operation. As a result, the installation process for Windows may require additional steps to properly structure the backend. Installation guides for Mac, Windows, and Linux can be found on Docker's website. In this demonstrational use case, both virtual machines were completely bare, aside from the installation of Docker and Git.

The process for executing the code was identical between the machines once the machines were properly set up. The appropriate command line interface (CLI) was opened with administrative rights. The necessary GitHub repository was cloned using the "git clone" command¹. The command "docker-compose build" initialized the Docker image. Finally, "docker-compose up" ran the image and exited upon success.

5. Conclusion and Considerations

We are not the first to suggest implementing containerization into economics research. The American Economic Association (AEA) Data Editor highlights the advantages of using containers for academic research, focusing on their implications for reproducibility and urging researchers to make use of containers when they submit data and code to the AEA's set of high impact journals² (AEA Data Editor, 2021). As outlined by (Boettiger, 2015), using containers eliminates the role of dependencies and software versions and imprecise documentation in creating discrepancies in program output across researchers.

Containers have ability to be pulled from a registry like Docker Hub and used on any local machine regardless of host environment, which highlights their portability. As noted earlier, containers may also be used on virtual machines on cloud computing platforms such as Azure and AWS EC2, which can aid in the management of computational resources for especially intensive tasks. Also contributing to their portability is their size, which is an order of magnitude smaller than virtual machines. As discussed earlier, containers require significantly less space and time to operate and start up than virtual machines, motivating their use for collaboration among

¹ We direct readers to TK to find all inputs and scripts necessary for replicating this exercise on their own machine. Once Docker is installed, our pre-built container can be deployed using only 2 commands.

² See the AEA Data Editor GitHub page for more detailed instructions on using Docker for replication

researchers with varying levels of resource constraints. These advantages in reproducibility, portability, and efficiency position containerization as a powerful tool for research collaboration.

References

- AEA Data Editor. (2021, November 21). *Use of Docker for Reproducibility in Economics*. Office of the AEA Data Editor. <https://aeadataeditor.github.io/posts/2021-11-16-docker>
- Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1), 71–79. <https://doi.org/10.1145/2723872.2723882>
- Chen, H., & Strathearn, M. (2020, February 6). *A Spatial Model of Bank Branches in Canada (No. 2020–4)*. Bank of Canada. <https://doi.org/10.34989/swp-2020-4>
- Handel, D., Ho, A., Huynh, K., Jacho-Chavez, D., & Rea, C. (2021, October). Cloud Computing Research Collaboration: An Application to Access to Financial Services. *Data Science in Central Banking: Machine Learning Applications*. IFC and Bank of Italy Workshop on “Data Science in Central Banking,” virtual event hosted by the Bank of Italy.
- . (2021). Econometrics Pedagogy and Cloud Computing: Training the Next Generation of Economists and Data Scientists. *Journal of Econometric Methods*, 10(1), 89–102. <https://doi.org/10.1515/jem-2020-0012>
- Mintel, “The Branch Banking Experience - Canada - February 2018,” Technical Report, Mintel February 2018.



Stanford University

Containerization for Research Collaboration

Platform Independent Economics

Venkat Balasubramanian, Danielle Handel, Anson Ho, Kim Huynh, David Jacho-Chávez, Carson Rea

Previous work

IFC Workshop Part 1: Data Science in Central Banking: Machine learning applications

Cloud Computing Research Collaboration: An Application to Access to Cash and Financial Services

Danielle V. Handel, Anson T. Y. Ho, Kim P. Huynh, David T. Jacho-Chavez, Carson H. Rea

Abstract

We illustrate the utility of cloud computing tools for big data management and analysis serving the functions of the Bank of Canada. These tools provide the opportunity to easily leverage increasingly complex and large-scale data in an interactive coding environment without worrying about backend infrastructure. As an empirical use case to demonstrate these advantages, we use a cloud computing platform to expedite a computationally intensive spatial analysis mapping access to financial services in Canada.

Keywords: High-Performance Computing; Big data; Spark; Jupyter.

Introducing: the “**but it works on my machine**” problem



See “Use of Docker for Reproducibility in Economics,” from AEA data editor [Lars Vilhuber](#)



Containers Offer a Solution

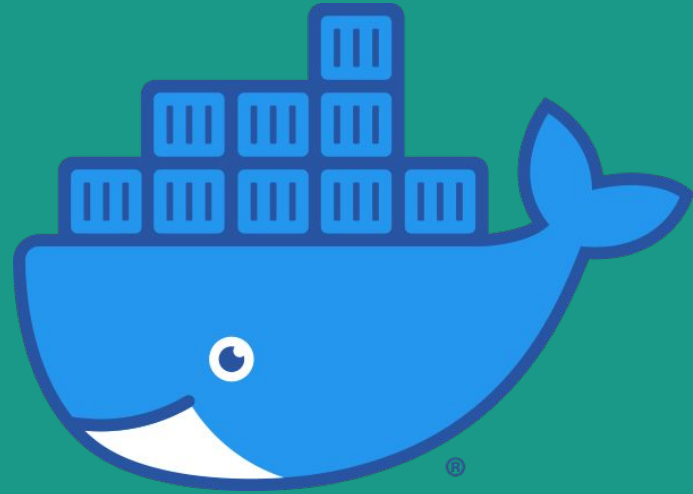
A *container* is an executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings

Solves the “**works
on my machine**”
problem*



REPRODUCIBILITY
PORTABILITY
COLLABORATION

Use Case





Access to Cash and Financial Services

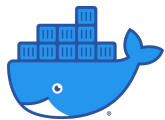
From our [previous work](#): mapping consumers and their nearest bank branch

- Challenge: 24,000+ postal codes
- Computed straight line distances from population centroids to financial institutions
- Run with [PySpark docker image](#)
 - Port to [Microsoft Azure](#) and [AWS ECS](#) to manage computational needs

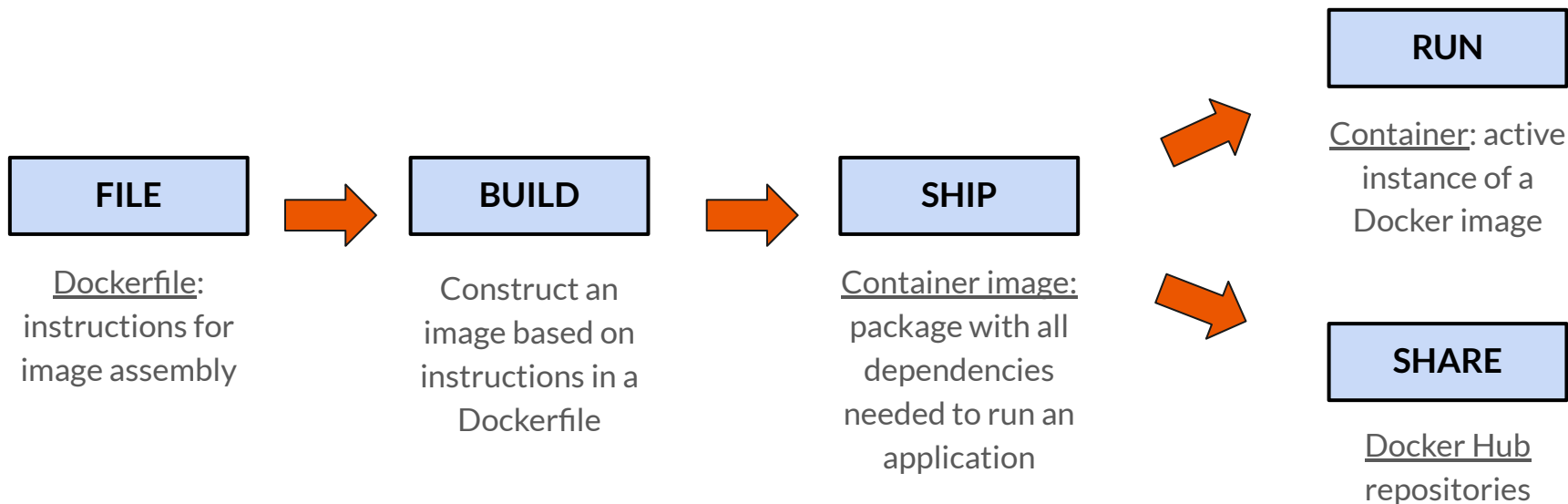


[Repro repo](#) on GitHub with all necessary input files

What is Docker?

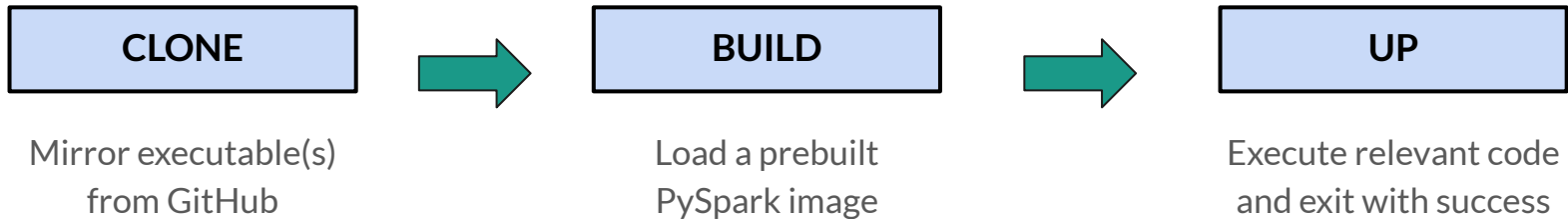


Industry standard tool for the creation and deployment of containers



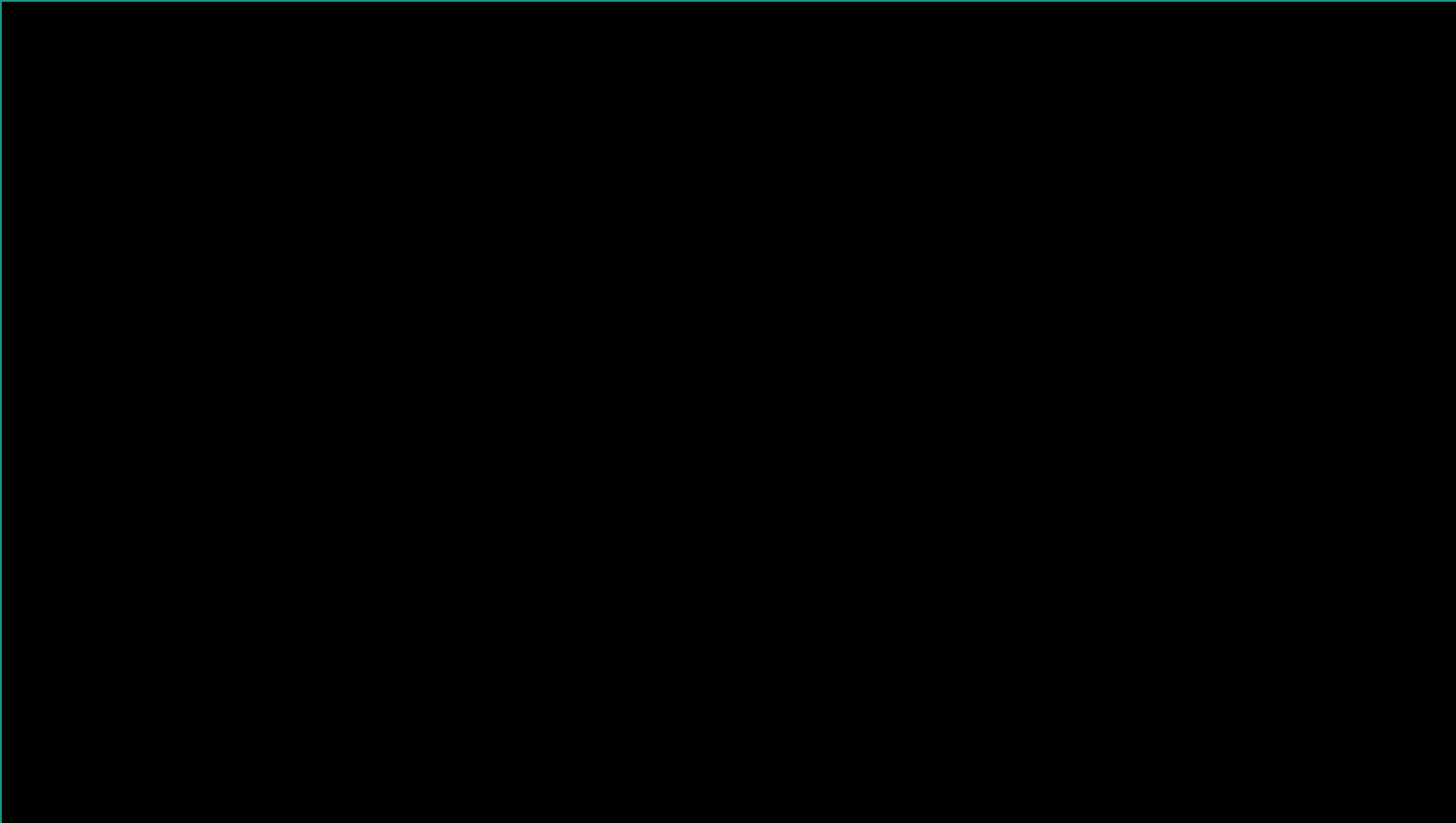


Constructing our Docker Container



Execution on Across Cloud Platforms





Singularity as an Alternative

- [Singularity](#) provides the same functionality as Docker
- Load images from Docker
- Better suited for HPCs
- Only compatible with Mac OS and Linux



Added Value



[AEA Data Editor](#)



[Repro repo](#) on GitHub

COLLABORATION

EFFICIENCY

REPRODUCIBILITY

PORTABILITY



Thanks/Merci

Venkat Balasubramanian
balv@bank-banque-canada.ca

Danielle Handel
dvhandel@stanford.edu

Anson Ho
atyho@ryerson.ca

Kim Huynh
khuynh@bank-banque-canada.ca

David Jacho-Chávez
djachocha@emory.edu

Carson Rea
chrea@emory.edu

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Banco de Portugal data centric-strategy for the adoption of a modern data architecture¹

Caio Costa, Guilherme de Sousa and Hugo Matos,
Banco de Portugal

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Banco de Portugal data centric – strategy for the adoption of a Modern Data Architecture

Caio Costa, Guilherme de Sousa, Hugo Matos¹

Abstract

Banco de Portugal has been on a path towards becoming an increasingly data-driven central bank, in order to take the most advantage of information intelligence. Along with the establishment of an integrated data management program, the challenge of creating a vision and a strategy for the adoption of a Modern Data Architecture has arisen. It is intended to enable the coexistence of corporate analytical environments with new capabilities that enhance fast, agile and flexible access to structured and unstructured data in its most granular state, in order to allow business users to address new advanced analytics or machine learning use cases. This presentation will focus on the approach followed by Banco de Portugal, in a partnership between IT, Statistics and other mission areas, to identify a set of differentiating capabilities, relevant trends and new paradigms in Big Data/Data Science domains, the proof of concepts implemented to validate the business value, as well as the roadmap established for an iterative and governed adoption of a Modern Data Architecture.

Keywords: data strategy, advanced analytics, machine learning framework

1. Introduction

Banco de Portugal is on a path towards becoming an increasingly data driven central bank where the focus is more and more on “data centricity” rather than “application centricity”.

One of the key elements of the Bank’s four years strategic plan 2017-2020 was the creation of an integrated data management programme, aiming at creating the Bank’s data warehouse together with two pillars: the data catalog and the master data system. This programme was a major transformational initiative, jointly coordinated by the IT and the Statistics Departments, in order to strongly contribute to the a better use of the available data in the Bank by means of rationalisation of the processes associated with data collection and processing and to promote its effective sharing throughout the whole organisation (Moreno, 2021).

While running this programme, new challenges emerged, such as: i) business areas were shifting more and more towards the collection of microdata (securities, loans, payments, ...), with higher frequency and more granularity; ii) projects involving big data were becoming more frequent; iii) new data science approaches were necessary to analyze and explore the new data landscape; and iv) security and privacy becoming a major concern.

It became then evident for the IT Department that we needed a specific IT programme to address all these challenges – the creation of our modern data architecture initiative. It kicked-off late in 2019

¹ Caio Costa (ccfcosta@bportugal.pt), Guilherme de Sousa (garanha@bportugal.pt) and Hugo Azevedo Matos (hmatos@bportugal.pt), Information Systems and Technology Department, Banco de Portugal. The views expressed are those of the authors and not those of the Banco de Portugal. We thank to Sara Cândido (Banco de Portugal) and Filipa Lima (Banco de Portugal) for their valuable comments.

and since then we have been developing our data strategy and our artificial intelligence strategy as part of it.

The modern data architecture governance model foresees a straight collaboration between the business units and the IT department. Within the IT department, we have 3 units that are more directly involved in the programme: i) the BI and Information Management Unit; ii) the Systems and Applications Engineering Unit; and iii) the INOV# Inovation Lab.

The paper is organised as follows: after the introduction section, we describe the different dimensions of our modern data architecture in section 2. In section 3 we present the various use cases already in place. We conclude with section 4 with future developments and final remarks.

2. Modern Data Architecture

In our Modern Data Architecture strategy, we identified three clusters of analytical capabilities:

- 1) Corporate BI – analytical capabilities typically centrally guaranteed by IT, with the aim of providing certified information organized by domains.
- 2) Self-Service BI - analytical capabilities that guarantee greater autonomy to users to analyze and interact with analytical information, allowing the creation of interactive and dynamic outputs.
- 3) Advanced Analytics - analytical capabilities typically related to the automatic identification of patterns using self-learning algorithms and exploited autonomously by advanced users.

A key feature of our programme is to go step by step and to gradually promote the business autonomy and flexibility to interact with data, in environments and platforms governed by the IT department. Given that corporate BI and the self-service BI tools were already relatively well covered for the end user needs (through the application portfolio and powerBI dashboards), we focused on advanced analytical capabilities that we were missing and were demanded by the most savvy end users to boost their productivity, which will be the focus of this paper.

In the advanced analytics space, there are lots of new concepts that originates mislead interpretations so the first step was to build a common foundation of concepts.

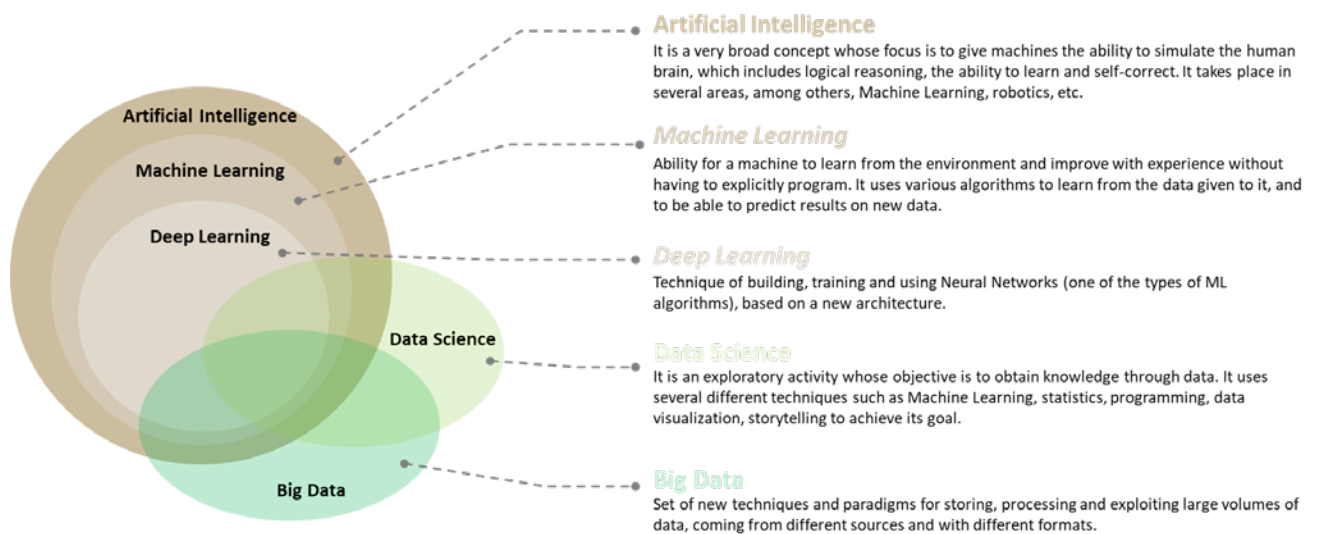


Figure 1 - Advanced Analytics concepts foundations.

Focusing on the transition to a modern data architecture and taking into account our business use cases, we started by defining two parallel paths:

- 1) Path "End users empowerment" - focused on providing new analytical environments with differentiating capabilities to end users take advantage of modern data exploration techniques.
- 2) Path 2 "IT Department capacitation" – focused on empowering the IT Department with the skills, methodologies, frameworks and tools needed to deliver turnkey data science projects.

2.1 End users empowerment - New analytical environments

Based on a straight collaboration between the business units and the IT department, two main needs were identified:

- The first one, aimed to experienced data analysts, was related with having possibility to use SQL language to analyze data in an environment with integrated access to corporate data and with write/persist permissions;
- The second one, aimed to data scientists, was related with providing environments suitable for the exploration of advanced techniques such as machine learning algorithms.

To answer to this needs, Banco de Portugal provided new analytical environments, that enable autonomous exploration by end users, guided by governance and good practices.

2.1.1 SQL Data Labs

SQL data labs to allow users to interact with structured data available on corporate data stores using an SQL interface and persist the analysis results.

User can load adhoc data to join with corporate data and build analysis.

This data labs ensures that the execution of analysis doesn't interfere with scheduled batch corporate processes.

To support this data labs, we leveraged Microsoft SQL servers that we already had in-place and that were most of time idle, due to be used only for secondary replicas.

2.1.2 Data Science Labs

The Data Science Labs - Pitágoras are a new analytical environment for advanced data exploration with modern data science tools, runned on top of a platform with scalable and remote processing capacity, dedicated storage, code versioning (with Github) and a governance model.

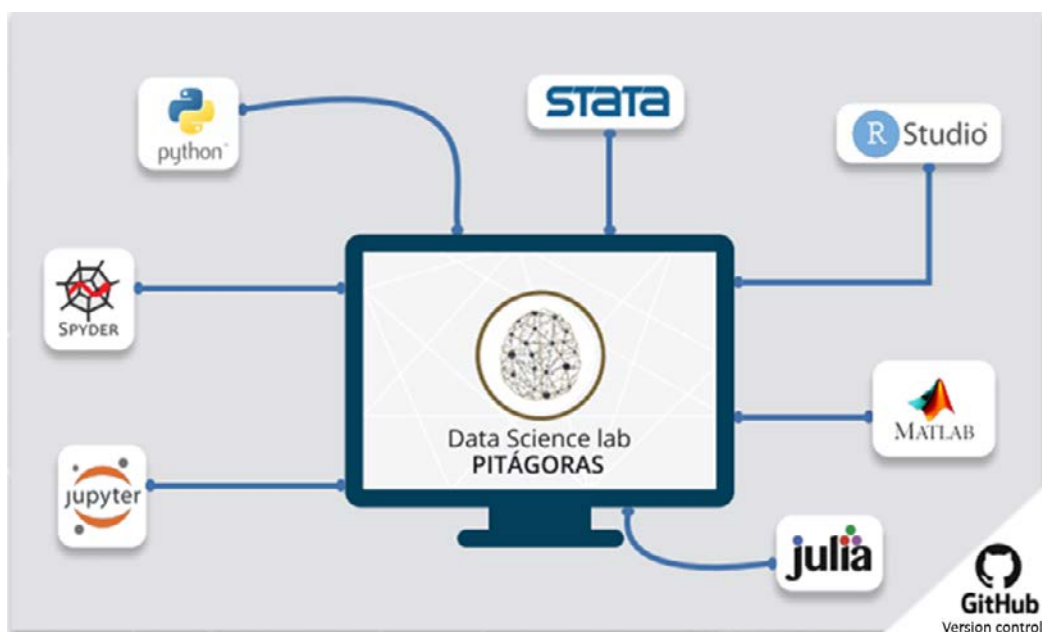


Figure 2 - Data Science Labs flavours.

Tools available for end users work with are Rstudio, Python, Spyder, Julia, Jupiter notebooks, Stata and Mathlab.

In terms of technology, Pitágoras is an in grid computing with containerized and customizable runtimes with singularity, which is a container runtime very similar to docker but designed for this type of use case, and based in CEPH, a Software defined storage that delivers both file and object storage with AWS S3 compatible interface capable of scalling to hundreds of petabytes.

With Pitágoras users are able to execute both batch and interactive jobs - such as submitting a background program, or running their development environment.

Regarding its current size we have 14 servers, totalling 7TB of RAM; 448 CPU Cores; 6 GPUs and 100TBs of storage. It can be scalled to as many servers as needed with ease since all the infrastructure is deployd automatically with Ansible.

We have also deployd a GitHub Enterprise infrastructure for our users source control.

2.2 IT Department capacitation

In order for the IT Department to be able to deliver turnkey data science projects, it was necessary to define project management methodologies suited to the specificities of this type of project, as well as define development and production cycles adapted to ML model, who promote automation, standardization and quality, leveraging good practices.

2.2.1 CRISP-DM

CRISP-DM

To manage our data science projects, we decided to base on CRISP-DM methodology.

The CRISP-DM methodology (Cross Industry Standard Process for Data Mining) is an open standard process model that describes the data science life cycle. It aims to standardize and describe the phases in a data science project, in order to help plan, organize and implement, us-ing common approaches and vocabulary.

It consists of 6 phases, namely:

1. Business Understanding - initial phase that focus on understanding the value-add from a business perspective and translate it to a data science problem definition. Includes definition of business objectives and what success means, and creation of the preliminary plan to achieve those success criteria.
2. Data Understanding - focus on identify, collect and explore initial datasets to know what can be expected and achieved from the data.
3. Data Preparation - prepare the data sets to be used by algorithms, which includes select, clean and create data using data preparation processes.
4. Modelling - phase responsible to build the model that answers to the business needs, as-sessing distinct model techniques against each other.
5. Evaluation - verify if the results meet the success criteria and define next steps. Outcome could be proceed to deploy, iterate again on previous phases to understand why the objec-tives were not achieved and try different approaches, or even to close the project. Also in-cludes reviewing and documenting the processes.
6. Deployment - last phase who is responsible to plan and implement the deployment as well as the monitoring and maintenance of the model after going to production. Depending on the business objectives, the deployment can be integrate the model with an existing software or build a dashboard/output to support decision-making.

The sequence of the phases is not rigid and moving back and forth between phases is common.

This methodology was created many years ago, and despite there are some new project management frameworks used for delivering data science projects (generic project management frameworks like Scrum, Kanban, and also more data science oriented like TDSP, SEMMA), it stills maintains to be one of the most used ones.

2.2.2 Machine learning framework

In order to enforce automation, the use of standards and good pratices, we build a custom machine learning framework, composed by technology and processes.

This framework defines development and production cycles adapted to ML model, leveraging good practices in software development.

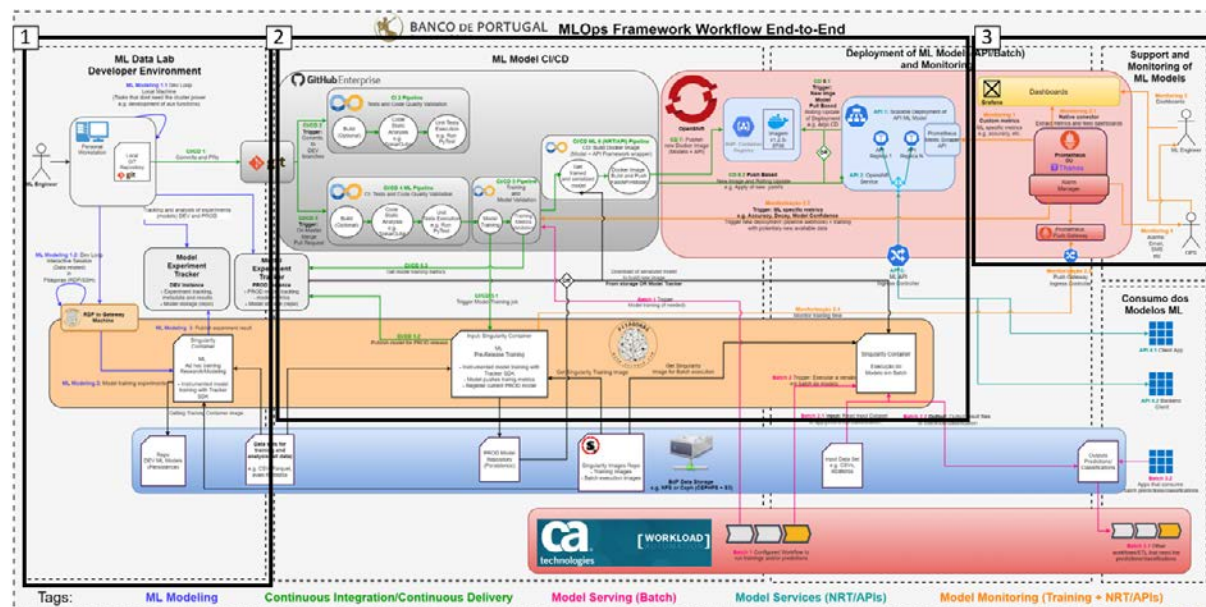


Figure 3 - BdP Machine Learning Framework

The framework is composed of 3 main areas:

- 1) Development: model development
- 2) CI/CD Training Pipeline: code quality validation, automatic tests, (re) training, deploying new versions (batch / api)
- 3) Monitoring & Alerting: collection and monitoring of model metrics, trigger actions based on rules (ex: retrain)

In terms of the development area, our goals were to provide some guidelines regarding good code practices, project structure and model development standardization. To achieve this, we built an example project in which we took all of these aspects into account to demonstrate what has been documented. So far, our top requirements are that models are developed with the kedro framework which brings at almost no cost a lot of good practices, and for the rest APIs to be developed with fast api.

Regarding the CI/CD pipeline, our objective was to have everything automated end to end with the good DevOps practices, plus everything that deploying ML requires: in other words, MLOps. The idea is that this pipeline is generic enough so that everytime a new ML project arrives, it can be reused with just very few changes – we are using github actions, so everything is “infrastructure as code”. In the end, after the tests have passed, the model has been trained in Pitágoras and built into a docker image, it is deployed in our Openshift Cluster.

Last but not least, monitoring and alerting. One of the guidelines of the ML Framework is that the models need to expose a metrics endpoint so that we can monitor it with a Prometheus, Grafana and AlertManager stack. This allows us to monitor the model’s accuracy so that we can be alerted when it goes below the defined threshold and even trigger an automatic training.

3. Use cases

In this section we describe the use cases we have worked so far with various business units.

3.1 SQL Data labs

We start with quality control for individual loans. Experts from the Statistics Department wanted to find outliers in loans information (from our Central Credit Register – CCR). The information has an analytical model defined which is available to the end users. For the first phase of the use case, it was used a SQL data lab, where the experts from Statistics autonomously did the construction / prototyping /and creation of a MVP on Quality Control Processes and at a later stage, after certification, these rules were incorporated into the CCR system (in production environment by the IT Department).

In the second phase, the business wanted an algorithm to detect outliers, so they selected the isolation forest algorithm and created the process in python – using our Data Science Lab - Pitágoras platform. In this case whenever they need they run manually the process, because as we mention we are still defining the best way to incorporate ML code done by end users into corporate systems. (just for a glimpse the Isolation forest (deal with 17 million of new records per month | and in less than 1 hour they are able to train the algorithm).

Another use case worth mentioning is ad hoc analysis for illicit financial activity. Ingestion and analysis of data collected during inspection actions, is done in a completely autonomous way by business users, using skills that are familiar to them (SQL), in an governed environment that has high computing power and that enhances the sharing of information between stakeholders, a SQL data lab.

3.2 Data Science labs

Regarding the machine learning use cases using the framework established and implemented by IT department in project mode we have the following 4 use cases, which were first experimented in our innovation lab and once approved for adoption followed our pipeline for IT project implementation.

1) Automatic validation of credit contracts

The goal is to validate automatically compliance with regulatory standards in the Credit Agreement Drafts, moving from a sampling approach (number of drafts contracts and number of validation rules) to a universal approach (potentially all drafts contracts and all rules). The tool automatically identifies, even if partially, the clauses that do not comply with legal and regulatory requirements, improving the performance of the Conduct Supervision department.

2) Supervisory Board written procedures

The goal is to accelerate the analysis and reply to Supervisory Board written procedures for which a RPA and rules engine were developed. The tool extracts and structures data from a sample of previous written procedure since 2014 and provides NLP capabilities to assist document content and analysis.

3) Automatic classification and response to information requests by banking clients

The goal is to automatically classify information requests and propose the corresponding automatic replies, depending on the classification and content, evaluating thus the feasibility of an automatic answer.

4) Non structured data analysis in acquisition processes

The goal is to have automatic analysis of the acquisition processes, data extraction from acquisition documents (ex.: contract, contract decision, public procurement portal) and audit controls' validation (ex.: mandatory clauses in contracts, amount consistency between documents).

4. Final remarks

For the current 2021-2025 strategic plan we will continue working across the data value chain to fully benefit from new capabilities (data lake and distributed processing engine - Spark) and to promote the use of good practices, such as data lake organization, workload segregation, data ingestion patterns, etc. We will work with the Payments Department towards a new data collection paradigm of individual payments data and with the Statistics Department in order to define a data quality control framework where a set of common consistency and validation rules can be applied irrespectively of the data domain.

We conclude with the key success factors throughout our data journey:

1) Data architecture

Defining the path from traditional to a modern analytics architecture, with business use cases.

Choosing the right technology for the most use cases.

Significant investments in IT infrastructures and software.

2) Data security

Defining very clear rules and principles in terms of information security policy.

Increased use of micro data raises the bar in terms of security issues.
Particularly relevant when considering cloud environments.

3) Data governance

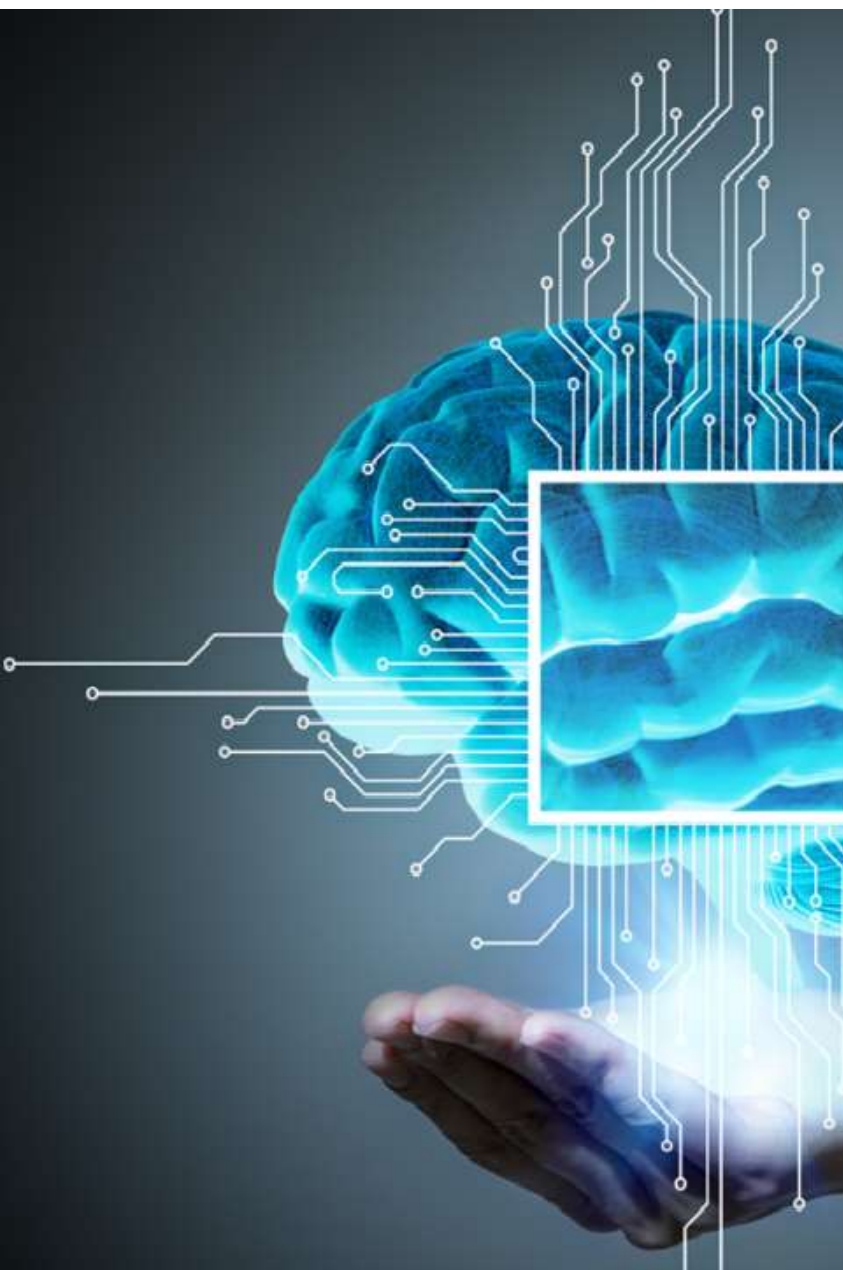
Setting an enterprise-wide roles and responsibilities.
All departments should be involved in the decisions.
Cultural / Organizational change.
Adequate expectations to management is vital.

4) Data skills

New roles are needed (e.g. big data / machine learning engineers, data scientists).
Reinforcing and adapting the skills of employees to the new challenges.
Establishing a training programme is essential.

References

Moreno, M C (2021): "Data Governance: an orchestra of people, processes and technology", IFC Bulletin No 54, Issues in Data Governance.



Banco de Portugal Data Centric

Strategy for the adoption of a Modern Data Architecture

*Caio Costa, Senior Data Engineer
(ccfcosta@bportugal.pt)*

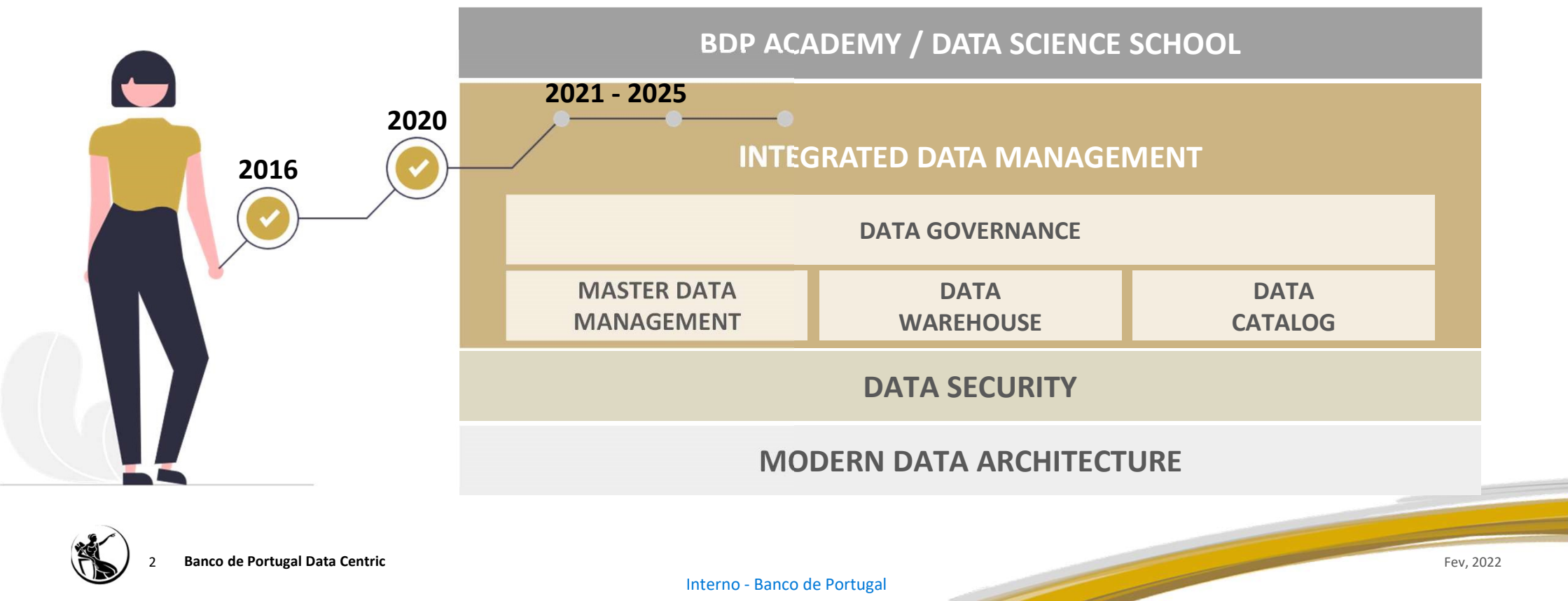
*Guilherme de Sousa, Systems and Applications Engineer
(garanha@bportugal.pt)*



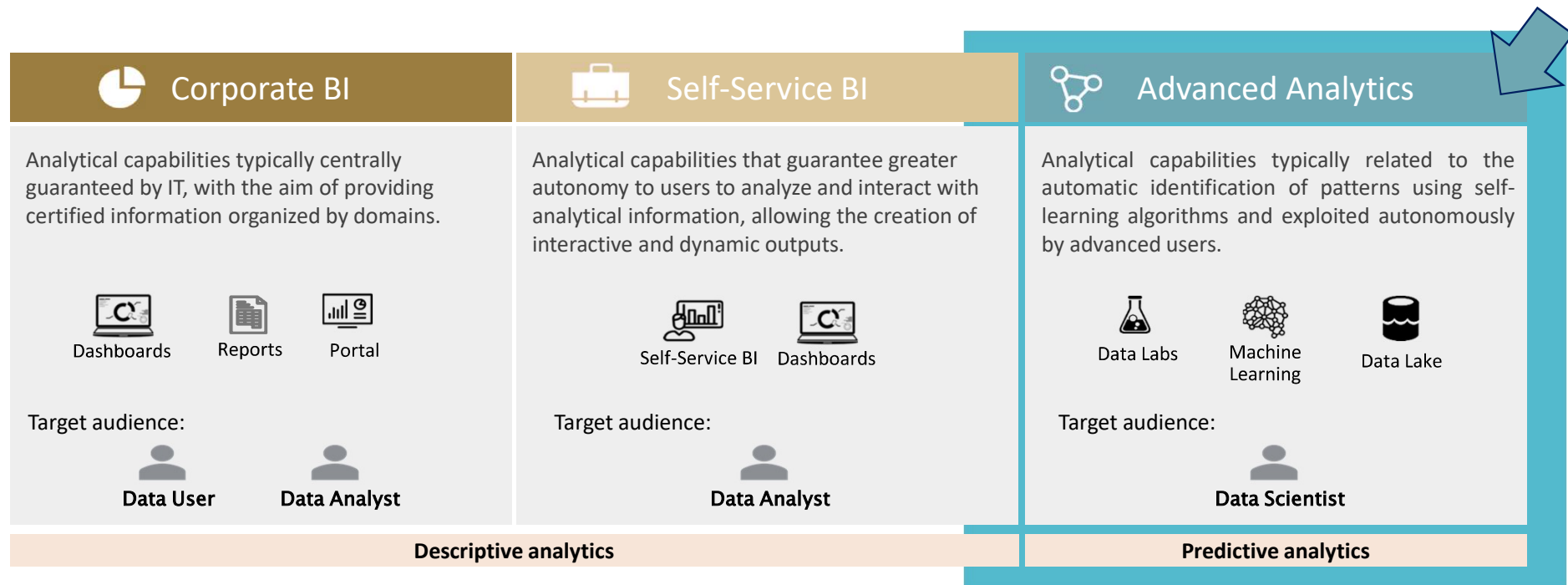
**BANCO DE
PORTUGAL**
EUROSISTEMA

February, 2022

Banco de Portugal on a path towards becoming an increasingly data centric central bank



In our **Data Strategy**, there are three main complementary clusters of analytical capabilities



In the **Advanced Analytics cluster**, with a target audience of advanced users, we provided **new analytical environments**

SQL DATA LAB

Interact with structured data using an **SQL interface** and persist the analysis results, with **high computing capacity** and without interfering with the execution of the scheduled batch corporate processes.

Users have **write permissions on the environment**.

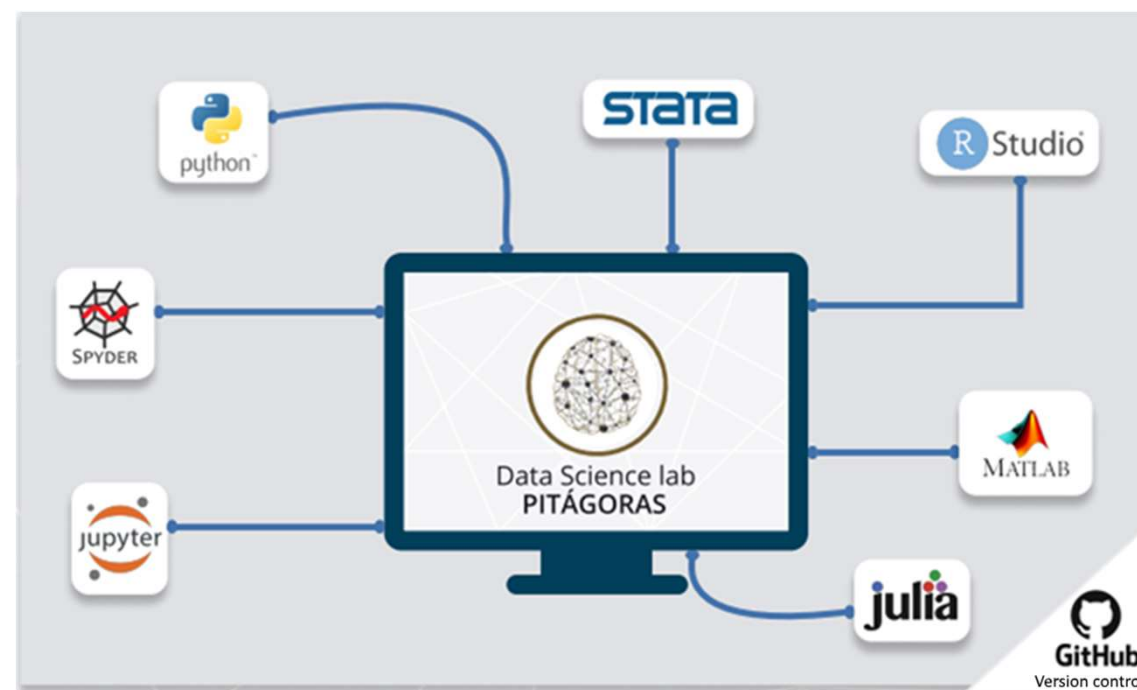
DATA SCIENCE LAB (PITÁGORAS)

Autonomy to use new data exploration techniques such as Machine Learning, Natural Language Processing through multiple tools / languages. With **scalable and remote processing** capacity, **dedicated storage** and **code versioning**.

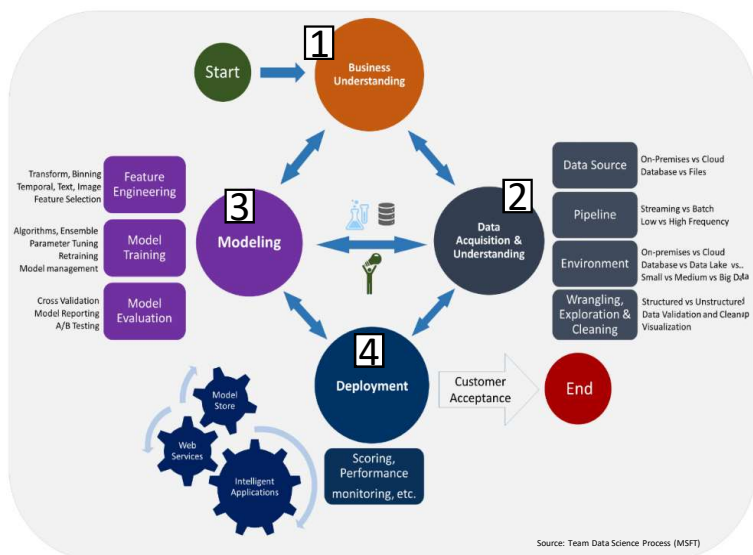


The **Data Science Lab** is a custom built open-source platform aimed for data scientists

- Grid computing solution
- Containerized and customizable runtimes and IDE's (singularity containers)
- Dedicated storage
- Batch executions
- Interactive executions (IDE's, Notebooks, etc.)
- + GitHub Enterprise Server



The IT Department provides **turnkey data science projects** using standard methodology and MLOps pipelines



Skills:

(not exhaustive)

 **Subject-Matter Expert**

 **Data Scientist**

 **Machine Learning Engineer**

- Data Science processes typically run through this life cycle based on **methodology (CRISP-DM)**, with well-defined steps that follow a logical sequence (identified by numbers)
- For each stage, **objectives** are defined, **techniques** to achieve them and **outputs**.
- In order to ensure that this process is **efficient**, the practice of **MLOps** is in place to **increase automation and improve the quality of the ML model** development and production cycle.

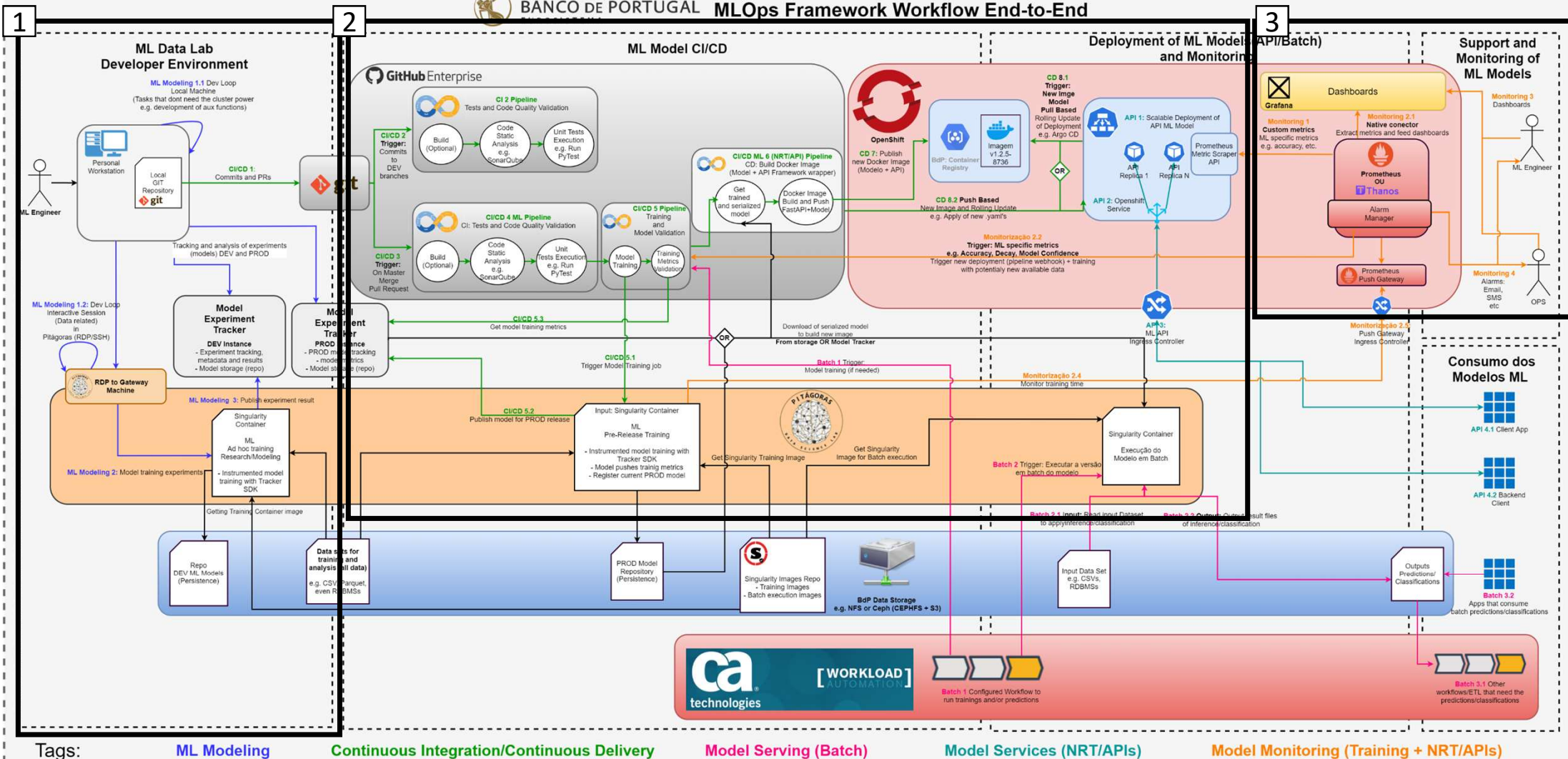
MACHINE LEARNING FRAMEWORK

The objective of this framework is to **increase automation**, promote **standardization** and **improve the quality** of the development and production cycle of ML models, leveraging **good practices** in software development.





BANCO DE PORTUGAL MLOps Framework Workflow End-to-End



Tags:

ML Modeling

Continuous Integration/Continuous Delivery

Model Serving (Batch)

Model Services (NRT/APIs)

Model Monitoring (Training + NRT/APIs)



7

Banco de Portugal Data Centric

Interno - Banco de Portugal

Fev, 2022

Banco de Portugal has deployed use cases that take advantage of the new analytical capabilities

| | | |
|------------------|--|---|
| Self-service | Illicit financial activity (adhoc analysis) | <ul style="list-style-type: none"> Ingestion and analysis of data collected during inspection actions, in a completely autonomous way by business users, using skills that are familiar to them (SQL), in an governed environment that has high computing power and that enhances the sharing of information between stakeholders - SQL Data Lab. |
| | Central Credit Register (quality control processes) | <ul style="list-style-type: none"> Implementation of quality control checks autonomously by business users using both traditional (SQL) and modern (machine learning techniques) approaches – expert users with skills is SQL & Python. After certified by IT, the quality control checks were incorporated into corporate system processes. SQL Data Lab – used for apply quality control rules using SQL queries and to prepare data to apply ML techniques; Data Science Lab – creation and execution of the ML (isolation forest) algorithm for outliers identification. |
| Turnkey projects | Information Requests Classification and Response (Conduct) | <p>Automatic classification of requests for information and generation of response text inov#)</p> <ul style="list-style-type: none"> Evaluate the feasibility of automatically classifying requests for information sent to Banking Conduct Supervision. Depending on the classification and content, evaluate the feasibility of proposing an answer automatically. |
| | Credit Agreement Draft Validation | <p>Validate automatically compliance with regulatory standards in the Credit Agreement Drafts</p> <ul style="list-style-type: none"> In a step-by-step approach from a sampling approach (number of drafts contracts and number of validation rules) to a universal approach (potentially all drafts contracts and all rules). Automatically identify, even if partial, clauses that do not comply with legal and regulatory requirements, improving the performance of the Conduct Supervision department. |



For a success implementation of a Data Strategy, these were the **Key Success Factors**



DATA ARCHITECTURE

Defining the path from traditional to a modern analytics architecture, with business use cases.

Choosing the right technology for the most use cases.

Significant investments in IT infrastructures and software.



DATA SECURITY

Defining very clear rules and principles in terms of information security policy.

Increased use of micro data raises the bar in terms of security issues.

Particularly relevant when considering cloud environments.



DATA GOVERNANCE

Setting an enterprise-wide roles and responsibilities.

All departments should be involved in the decisions.

Cultural / Organizational change.

Adequate expectations to management is vital.



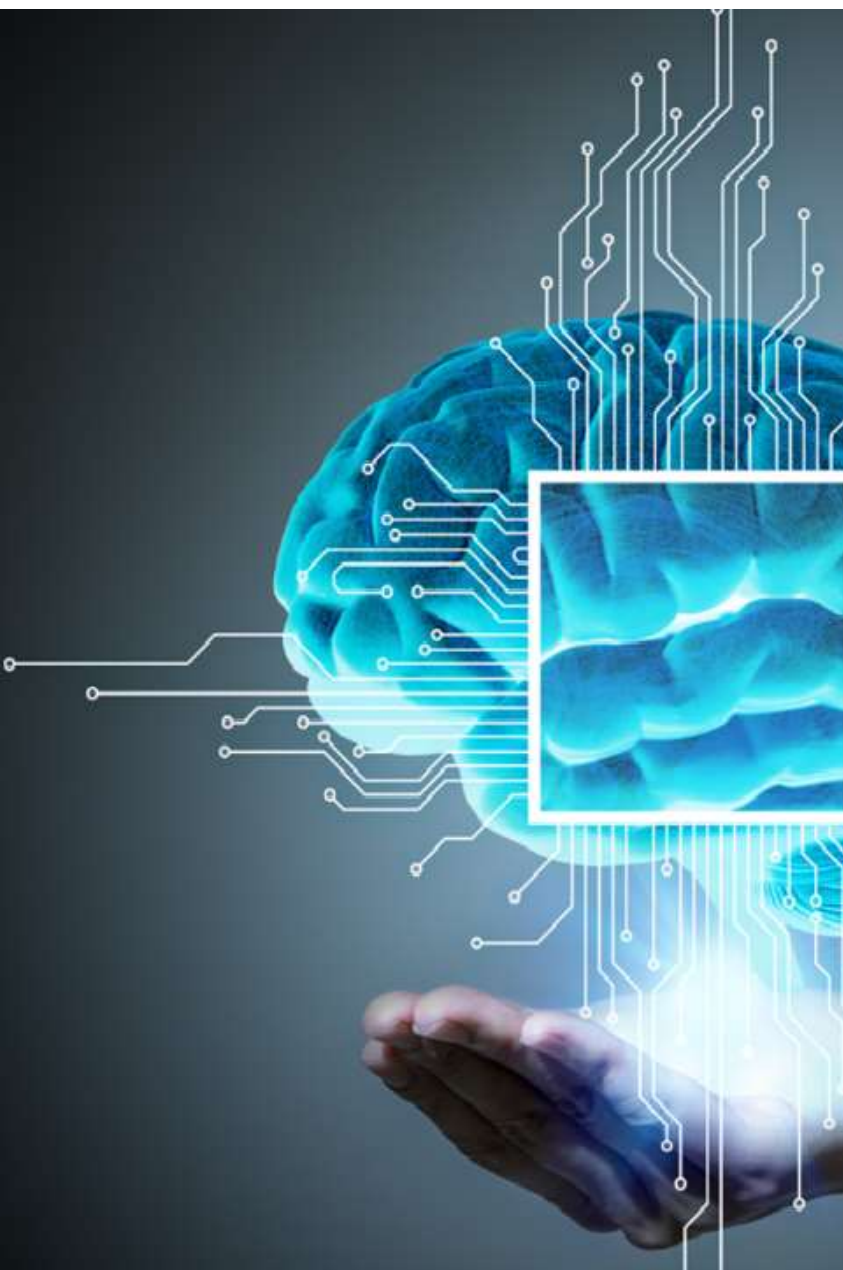
DATA SKILLS

New roles are needed (e.g. big data / machine learning engineers, data scientists).

Reinforcing and adapting the skills of employees to the new challenges.

Establishing a training program is essential.





Banco de Portugal Data Centric

Strategy for the adoption of a Modern Data Architecture

*Caio Costa, Senior Data Engineer
(ccfcosta@bportugal.pt)*

*Guilherme de Sousa, Systems and Applications Engineer
(garanha@bportugal.pt)*



**BANCO DE
PORTUGAL**
EUROSISTEMA

February, 2022

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Microdata utility – data loading with automated data parsing and data structure creation¹

Marcus Jellinghaus,
BIS

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Microdata utility

Data loading with automated data parsing and data structure creation

Marcus Jellinghaus, Bank for International Settlements

Abstract

Each year, the Bank for International Settlements (BIS) processes thousands of heterogeneous data files for economic research and analysis. This poses a challenge for the Information Technology team with the Monetary and Economic Department (MED IT team).

To address this challenge, the MED IT team has developed the 'microdata utility', facilitating the loading of millions of CSV and XML files into a SQL Server database with minimal configuration effort. The database structure is automatically generated based on the structure detected in the data files.

The BIS is using the microdata utility to process data purchased from commercial data providers, used for economic research, and the monitoring of financial markets.

The microdata utility will

- parse the specified data files,
- automatically detect the data structure within these files, and
- create tables in a database that match the structure of the data files.

By using this utility, the effort of setting up and regularly updating datasets can be significantly reduced, and technical resources can be used more efficiently.

Based on feedback, the BIS is open to sharing the utility with partners, and/or making the microdata utility available as open-source software via the BIS Open Tech Hub.

Keywords: Data loading, Automation, Database

JEL classification: C81, C88

1. Introduction

The microdata utility has been developed to load data from different commercial data providers into a relational database and offers a generic method for the loading of CSV files and XML files.

When processing CSV files, the utility will auto-detect the data types of each column and create the appropriate table structure within the database to fit the data. The utility can detect data type changes, new columns / structural changes in data files over time, and allow for the underlying database tables to be automatically altered. The utility allows users to load CSV files without a complex specification, (other than file names and table names).

When processing XML files, which may store the data in a more complex tree-like structure, the structure is automatically analysed and a relational data model is auto-generated.

The utility manages the status of each data file during the loading process and allows for the sequential or parallel processing of data. (To date, the BIS has used the utility to process millions of data files.) The Python- and Pandas-based utility includes options for loading data into SQL-Server using performant 'Bulk Insert'-operations and storing the data using columnstore indices¹. This allows for fast data loading, optimised data storage, and performant query processing.

The microdata utility can process data files available on disk, or via external HTTP or FTP servers. Using a minimal specification, the utility can download data files on a defined update frequency, ensuring that databases are kept up to date.

The utility is already used to process more than 12 commercial datasets within the BIS production environment.

The remainder of this paper is organized as follows:

Section 2 describes the use case and the design objective. Section 3 introduces the data parsing and the structure creation for CSV files, section 4 provides additional information for XML files and their more complex data structures. Section 5 describes how the source files and the destination tables are configured, and section 6 introduces the status tracking for files and the mass processing of files. Section 7 gives an overview of different technical optimisations. Section 8 concludes with the current status and gives an outlook on possible next steps, including the possible sharing of the utility.

2. Microdata use case and design objective

What is microdata?

In MED IT, we define microdata as non-aggregated data at the entity level, eg on individual companies, instruments, investment funds, etc. The BIS is using microdata for economic research and to calculate aggregates for analysing financial markets.

¹ [Columnstore indexes: Overview - SQL Server | Microsoft Docs](#)

This data is purchased from commercial data providers. A dataset may include just a few files, in other cases, it can include more than one million files.

Also, the consumed disk storage space varies, from some megabytes to more than one terabyte.

Storing the data

The microdata is stored in a relational database, making it easily accessible for our end-users. So far, we have worked with Microsoft SQL-Server database servers.

The goal is that the microdata utility loads all data provided by the vendor. While the full data collection may not be required for a current research project, other parts of the data may be required for a future research project, hence the goal to ensure all data is successfully loaded to the database.

Depending on the structure of the data, the data delivered by a provider could be stored in a single table or may need to be spread across several tables within the database.

Data quality and data quality checks

The data is being purchased from commercial data providers. As a general rule, we assume that the data is provided in a logical structure, and the utility will load the data as provided.

The utility performs structural data checks on the data files but does not perform data quality/coherence/validation checks. We assume that data quality issues will be identified by research analysts as they develop research papers, etc. If the data would be used for trading decisions, or to look in more detail at individual institutions, a more rigorous data quality assessment might be required.

Why a generic utility?

In the past, we created one custom application for each dataset. While these applications often had a limited complexity, they were all different and offered different features depending on the dataset to be processed. We have now created a generic data loading utility that allows us to process any dataset.

The microdata utility has been developed in an agile and iterative manner, with each release, adding new features required to process the data files in our collection. Every new dataset provides new challenges, some features are simple to implement, for example, the usual separator in a CSV file is a comma, and we had to make the separator configurable. Others, however, are more complex, for example, the need to implement a new parser for XML files.

The migration away from the custom applications to the microdata utility has challenges. For each dataset, specific features and workflows were implemented, which work sometimes slightly differently in the microdata utility, these need to be reviewed on a case-by-case basis.

Looking at the big picture, we benefit from the generic utility, which allows us to load new and large datasets very quickly. The overall configuration and operations only take a few minutes to define, and the processing of large datasets, eg with more

than 700,000 XML files, are parsed and loaded in less than 100 hours, also based on the degree of data complexity and available resources for parallel data loadings.

Why develop a utility, and not use an existing solution?

While many ETL tools allow analysing the structure of source files, we are not aware of other tools that automatically create and adjust the structure of a database following the structure of the input data files. On the contrary, the analysis often happens at a design stage, and the data structure is then fixed for the runtime. Also, the automatic parsing and translation of XML files to tables (as described in section 3) is not something that we have encountered before. Last but not least, the mass file processing with file status tracking allows for parallelisation, and different technical optimisations allow for efficient usage of resources (see sections 6 and 7). With these advantages in mind, we decided to implement our utility instead of using existing software.

Basic functionality (like reading CSV files, HTTP & FTP downloads, working with data frames, connecting to SQL-Server) was not implemented by us. Instead, we benefit from powerful Python libraries, including Pandas and SQLAlchemy.

Goal: Provide a generic, efficient data processing utility requiring minimal configuration effort

The utility has the goal to minimise the configuration effort, eg we do not want to hardcode the specific column names or their data types. The utility parses the files and detects the columns, their names, and data types. For XML files, the table structure follows the received data structure. This enhances flexibility, reduces the configuration effort, and provides, based on our experience, very high quality of the data structure.

Since we need to process millions of files, the utility needs to be performant and efficient. The files are being received from external providers; their technical quality cannot be controlled. Therefore, the utility needs to be robust and needs to flag possible issues as they are encountered.

3. Processing of CSV files: Data parsing and structure creation, dynamic adjustment of tables

Basic process for parsing data, creating tables and loading the data into a database

The microdata utility can parse CSV files automatically for column names and column types (ie str, float, datetime, etc.). When reading CSV files, the Pandas library automatically detects the different columns. We have added a feature to detect the data type of each column, eg whether it is text, dates, or numbers.

The microdata utility creates tables in a database automatically with corresponding column names and types detected when parsing the data.

After creating the tables in the database, the utility loads the data into the database.

Dynamic adjustment of tables

Tables are being automatically adjusted in SQL-Server if additional files have a different data structure:

- **Creation of additional columns:** Let's say an additional data submission includes additional columns – eg because the data provider added further information over time. In that case, the utility will automatically create a new column.
- **Extension of string columns (to store wider strings):** Let's say a column in the database has a certain width, eg to store a company name with 25 characters. In case a new file has a record with a company name of 30 characters, the database column will be automatically adjusted.
- **Adjustment of data types:** Let's say so far, a column only included integer numbers. If an additional data file includes a number that is not an integer but includes decimals, the data type would be automatically adjusted to store the complete number. Eg if a first file contains a "1", and a second file a "1.1", the column will be initially created as an integer column and will then be adjusted to column type "floating number".

The table definition is expanding. If an additional data file has fewer columns, the NULL values will be inserted for these missing columns.

4. Processing of XML files: Translation in relational table structures

The microdata utility also allows for the processing and storage of XML data files. XML files typically include more complex data structures above CSV data files, eg a tree of data items.

Translation to one table

Let's take a simple example. Imagine an XML file that includes data on several students and for each student several attributes like name, email, street, and city. This can be converted into a table with the different students as records, and the different attributes as columns.

Example: Simple translation of XML file into one table

A simple XML data example can eg look like this:

```
<data>
  <student>
    <name>Alice</name>
    <email>alice@mail.com</email>
  </student>
  <student>
    <name>Bob</name>
    <email>bob@mail.com</email>
  </student>
</data>
```

The resulting table looks like this:

| name | email |
|-------|----------------|
| Alice | alice@mail.com |
| Bob | bob@mail.com |

Translation to several tables

Let's take a more complex example: Again, we have students with attribute email. Now, we also have zero, one, or several subjects for each student and also the grades. In that case, we would create a second table "subject" and store the course name and the grade there. This table would also have a relational link to the table student.

The parsing of XML files has been used to process several datasets from different data providers. The resulting table structure and data model within the database provide a logical data structure for our analysts.

Further enhancements

The processing of XML files can be complex, we have identified further enhancements to the microdata utility which we believe will add business value:

- Trees with higher depth are being handled, more tables are being created. In some cases, the XML structure can be quite deep, therefore several tables are being created.
- Details that do not require extra tables get moved up into the main table.
- Unnecessary middle layers without extra data get removed.

Data relations are being tracked. In the database, these are being stored as primary keys and foreign keys. Relations between tables are also being tracked and are stored in a specific table. Further functionality allows presenting the relations as a dynamic network graph.

Example: Translation of XML file into several tables

XML data example with data on different levels:

```
<data>
  <student>
    <name>Alice</name>
    <email>alice@mail.com</email>
  </student>
  <student>
    <name>Bob</name>
    <email>bob@mail.com</email>
    <subjects>
      <subject name = "Math">
        <grade>C</grade>
      </subject>
      <subject name = "History">
        <grade>B</grade>
      </subject>
    </subjects>
  </student>
</data>
```

The resulting tables 'student' and 'subject' look like this:

| student_uuid | name | Email |
|--------------|-------|----------------|
| 5b765840 | Alice | alice@mail.com |
| a81d2b18 | Bob | bob@mail.com |

| name | grade | student_uuid |
|---------|-------|--------------|
| Math | C | a81d2b18 |
| History | B | a81d2b18 |

The field student_uuid is a "universally unique identifier". For each student record, one value has been generated. This identifier is used also in table 'subject' and allows to link records between both tables.

5. Dataset specification via Excel

The microdata utility requires user input to identify which files to load, and into which table these files need to be stored; to do this an Excel template is used to provide the dataset specification. Only a few columns are required to provide basic information like the path and the filename of the source files as well as the target schema and table name. Each row can be used separately, this allows users to specify several source folders or destination tables.

Where a user wants to load many files into one table, the specification supports wildcards like asterisks (*) and question marks (?). It is also possible to specify whether subfolders should be searched for files.

Tables can be grouped in database schemas. This allows users to separate providers and creates a clean organisation of the data tables in the database.

Box 3

Example: Dataset specification

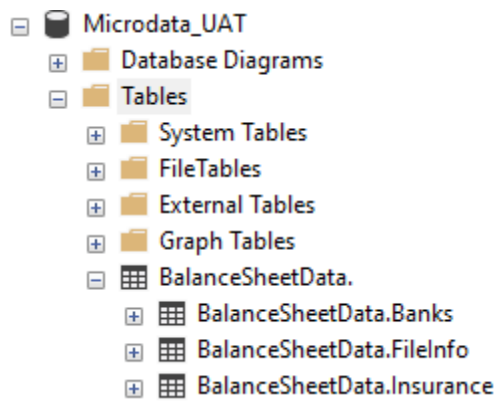
| | A | B | C | D |
|---|------------------|------------|--------------------|----------|
| 1 | DB Schema ▾ | DB Table ▾ | folder ▾ | filter ▾ |
| 2 | BalanceSheetData | Banks | Q:\Data\Banks\ | * |
| 3 | BalanceSheetData | Insurance | Q:\Data\Insurance\ | * |
| 4 | | | | |

Dataset Specification

In this example, the input files are defined in columns C and D. Column C defines from which folder the files should be read, and column D which files should be read.

Columns A and B define where the data should be written to. In Microsoft SQL-Server, the database schema allows to group tables.

The microdata utility automatically creates tables with the name as defined in column B; the column structure is derived from the data files.



Besides the two tables containing the data (in this example: Banks and Insurance), the utility also creates a table FileInfo which contains the names of all data files, and their loading status.

Specifications for updating a dataset

As previously mentioned, the microdata utility can download new data via FTP or HTTP, so that datasets can be kept up to date.

The microdata utility needs to know when to download a file. The dataset specification allows to define different parameters like update frequency (daily, business daily, monthly), point in period (eg end of month) as well as reporting gap and reporting frequency (ie data published after 1 day, or data published after 2 business days).

The utility requires the name of the files to be downloaded from an FTP server. When downloading files from an HTTP server, the name of the new data files must be

specified. For example, users would configure the system with the file naming schema "MyData_*.csv" and the date format "yyyyMMdd". If the date is the 22. September 2021, the file name "MyData_20210922.csv" is calculated.

Wildcards are also supported. This allows downloading several files for a specific date from an FTP server.

Zip files can also be specified, automatic unzipping is supported.

6. Mass file processing and status tracking

The microdata utility allows for the parallel processing of files, which dramatically reduces the time needed to load millions of data files. The status of each file is being tracked within the database. A database table "FileInfo" contains the path and filenames of all files, their processing status, and related information.

Several or even many agents can be used to load data and store it in the database. This allows to speed up the data loading significantly. A potential bottleneck is the database server, which depending on CPU and memory allocation, may only be able to handle a certain amount of data insertions concurrently.

Process for loading data

The standard process for loading a file consists of the following steps: Read Excel specification file, register file, analyse file, load file, mark file as successfully loaded.

In the first step, the dataset specification is read. This specification is used to find files. All found files are being registered in the FileInfo table in the database. In a further step, each file is being analysed for some high-level descriptive information (including the dates covered in the file). In the next step, each file is being read into memory and the data is being parsed. The data structure is being compared to the database. Where required, the database tables are being adjusted. Finally, the data is being inserted into the database table. In the case of very large CSV files, the files are processed in several steps. Once the data has been loaded successfully, the file gets the status "File loaded successfully".

Special cases are also supported, files can be ignored and deleted from the database based on a separate list. Different exceptions when loading files are being tracked inside the table FileInfo, including issues when reading the data, or when the data file is empty. Unexpected issues get logged.

A database view (Status View) allows users to have an overview of all files processed across different schemas. This Status View is also used to create a monitoring dashboard. At the BIS, we use a Tableau dashboard to see the summary statistics of the file loading status by dataset, and a second overview to see the file loading status of the files containing the data of the last 65 days.

7. Technical optimisations: SQL-Utility, Bulk Insert, ColumnStore, and Index Creation

The microdata utility is based on Python and uses libraries like Pandas, SQLAlchemy, and others to store data in SQL-Server.

To enhance the performance of the data loading processes, and to reduce the required resources, several technical optimisations have been implemented, including:

- A small SQL utility wraps different SQL commands into python functions.
- In the case of larger amounts of data, the "bulk insert²" method is used to transfer the data from the client running the microdata utility to the database server. This method speeds up the data inserts into the database.
- To reduce the amount of required storage, columnstore indices are being used. This special table storage method reduces data retrieval time and stores the data compressed on the storage volume. This is transparent to end users / data consumers, and SQL queries for querying the data do not need to be changed.
- Certain database indices are automatically created to increase query performance.

In summary, the microdata utility allows users to load large amounts of data quickly and efficiently.

8. Production status and considerations to share the utility

The microdata utility is used in production since 2020. It has already processed more than 12 datasets. In total, more than two million files have been loaded. All in all, more than 500 GB of compressed (!) data is being stored in the database, including data of various providers:

- different datasets provided by CryptoCompare,
- Fitch fundamental data on banks, data on issues and issuers,
- IHS Markit Iboxx data including constituents, indices, and mappings data
- IHS Markit Credit Indices
- Informa iMoneyNet data on mutual funds,
- iVolatility data
- Lipper Fund data
- Moody's CreditEdge
- MorningStar
- S&P Trucost data

² [BULK INSERT \(Transact-SQL\) - SQL Server | Microsoft Docs](#)

As mentioned above, only a small dataset specification is required to process these datasets quickly and efficiently. There is no need to define detailed table structures, these technical details are automatically generated based on the data. Based on XML files, sets of related database tables might be created. The utility can process millions of files, and dynamically adapt database tables based on developing data structures.

The microdata utility is continuously enhanced as we process additional datasets and migrate legacy dataset processing to use the utility.

Additional features are planned, eg to support hybrid usage on Linux / container stack and on windows servers, and to support a dimensional data model with slowly changing dimensions. Efficiency improvements related to the direct loading of data files from compressed zip files, and to storing GUID data more efficiently, are also envisaged.

Based on feedback, the BIS is open to sharing the utility with partners, and/or making the microdata utility available as open-source software via the BIS Open Tech Hub.

Please let us know whether you would be interested in collaborating in this space. Queries should be directed to Marcus.Jellinghaus@BIS.Org.



Microdata utility - data loading with automated data parsing and data structure creation

IFC and Bank of Italy Workshop on "Data Science in Central Banking", part 2, 14-17 Feb 2022

Marcus Jellinghaus, Principal Data Scientist, MED-IT, BIS

Microdata utility - a generic data loading utility

- What is **Microdata**?
- Why a **generic** utility?
- Next slides:
 - Dataset parsing and database table creation for CSV and XML files
 - Dataset specification
 - Some technical features
 - Current status and possible next steps

Data parsing and structure creation, data loading, dynamic adjustments

- Data parsing
- Data structure creation
- Data loading
- Dynamic adjustments



File parsing

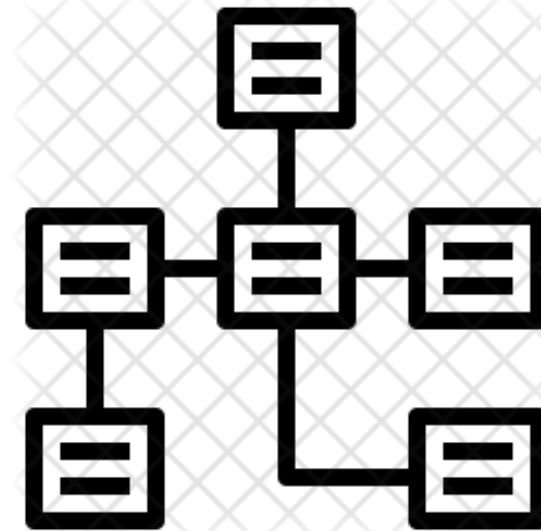
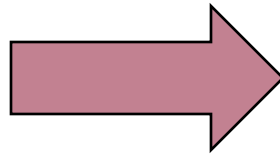


Table creation

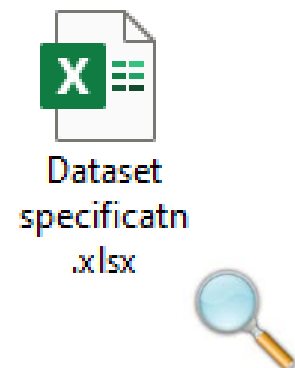


Data loading

Processing of XML files: Translation in relational table structures



Dataset specification with a small excel sheet



| | A | B | C | D |
|-----------------------|------------------|------------|--------------------|----------|
| 1 | DB Schema ▾ | DB Table ▾ | folder ▾ | filter ▾ |
| 2 | BalanceSheetData | Banks | Q:\Data\Banks\ | * |
| 3 | BalanceSheetData | Insurance | Q:\Data\Insurance\ | * |
| 4 | | | | |
| Dataset Specification | | | | + |

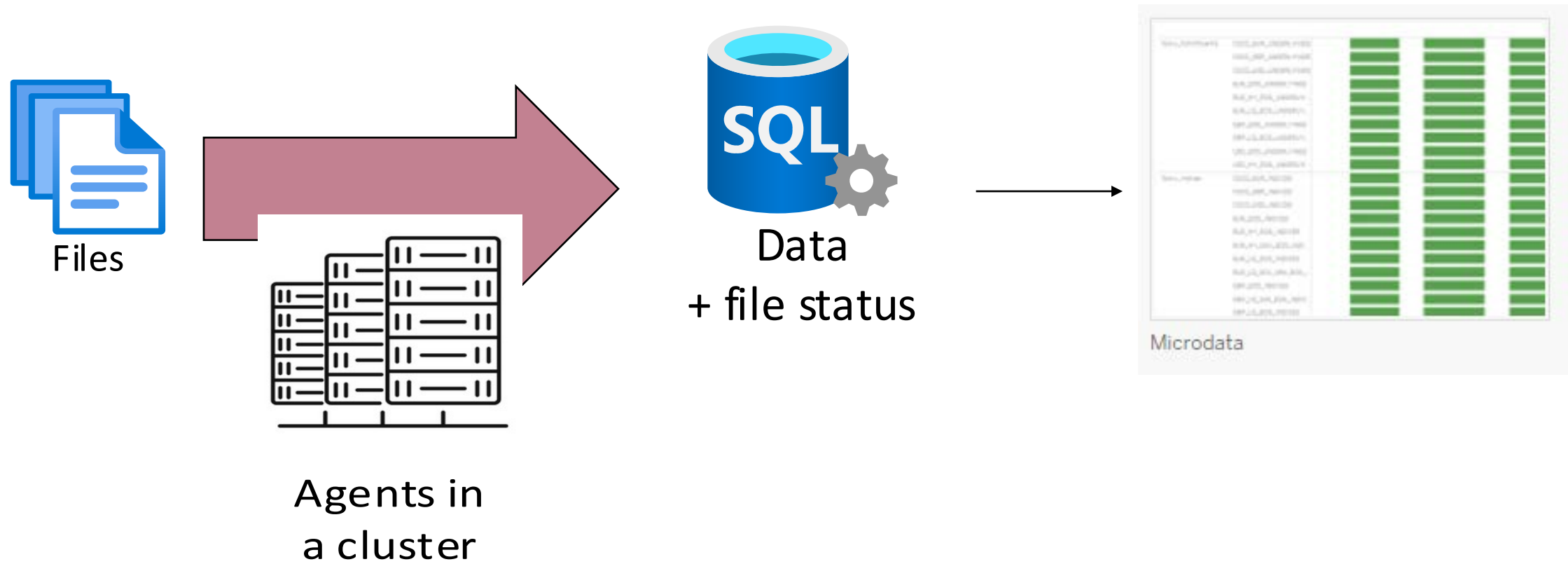


Database



- Microdata_UAT
 - Database Diagrams
 - Tables
 - System Tables
 - FileTables
 - External Tables
 - Graph Tables
 - BalanceSheetData.
 - BalanceSheetData.Banks
 - BalanceSheetData.FileInfo
 - BalanceSheetData.Insurance

Mass file processing and status tracking



Technical platform



Bulk Inserts



Columnstore



Python API
for DB

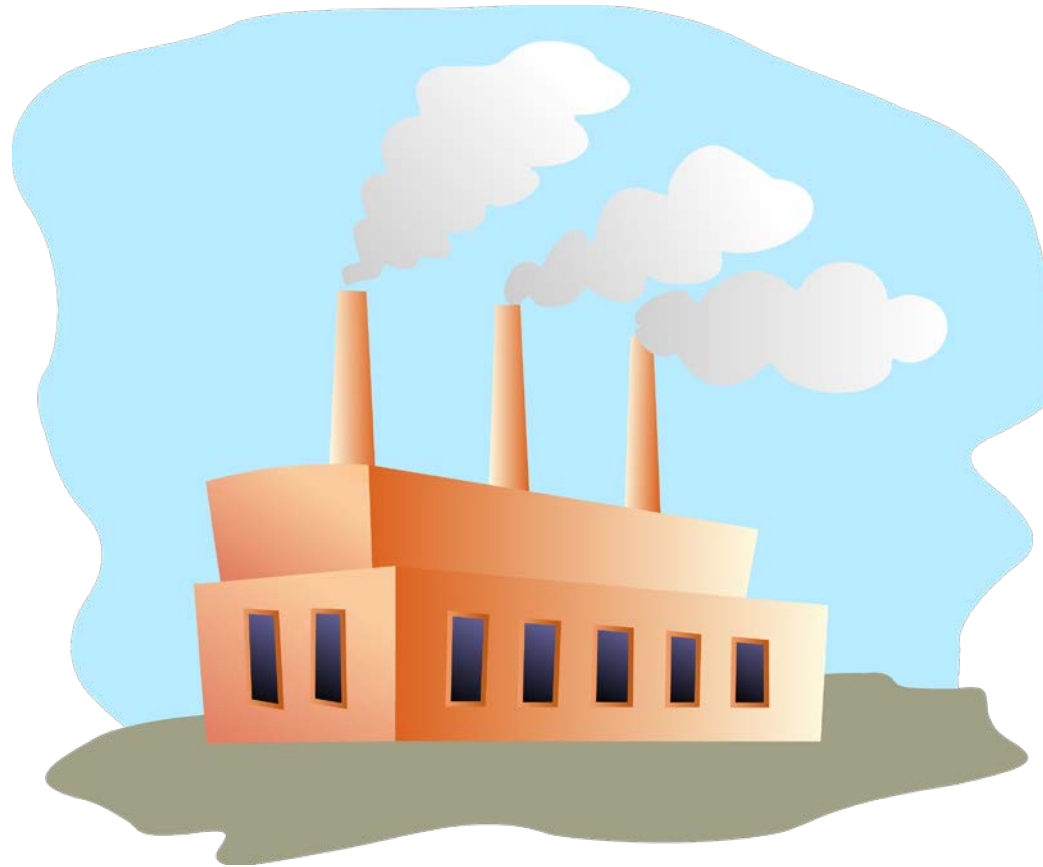


➔ The tool allows to load large amounts of data very fast,
and reduces the amount of required resources

Current status

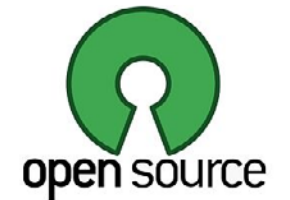


New datasets



Used in production since 2020

- 25 datasets
- 1.4 million files
- 400 GB of compressed data



Contact: Marcus.Jellinghaus@BIS.org

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Using non-traditional point of interest data as merchant survey sample frames¹

Angelika Welte and Joy Wu, Bank of Canada,
Marcel Voia, University of Orléans

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

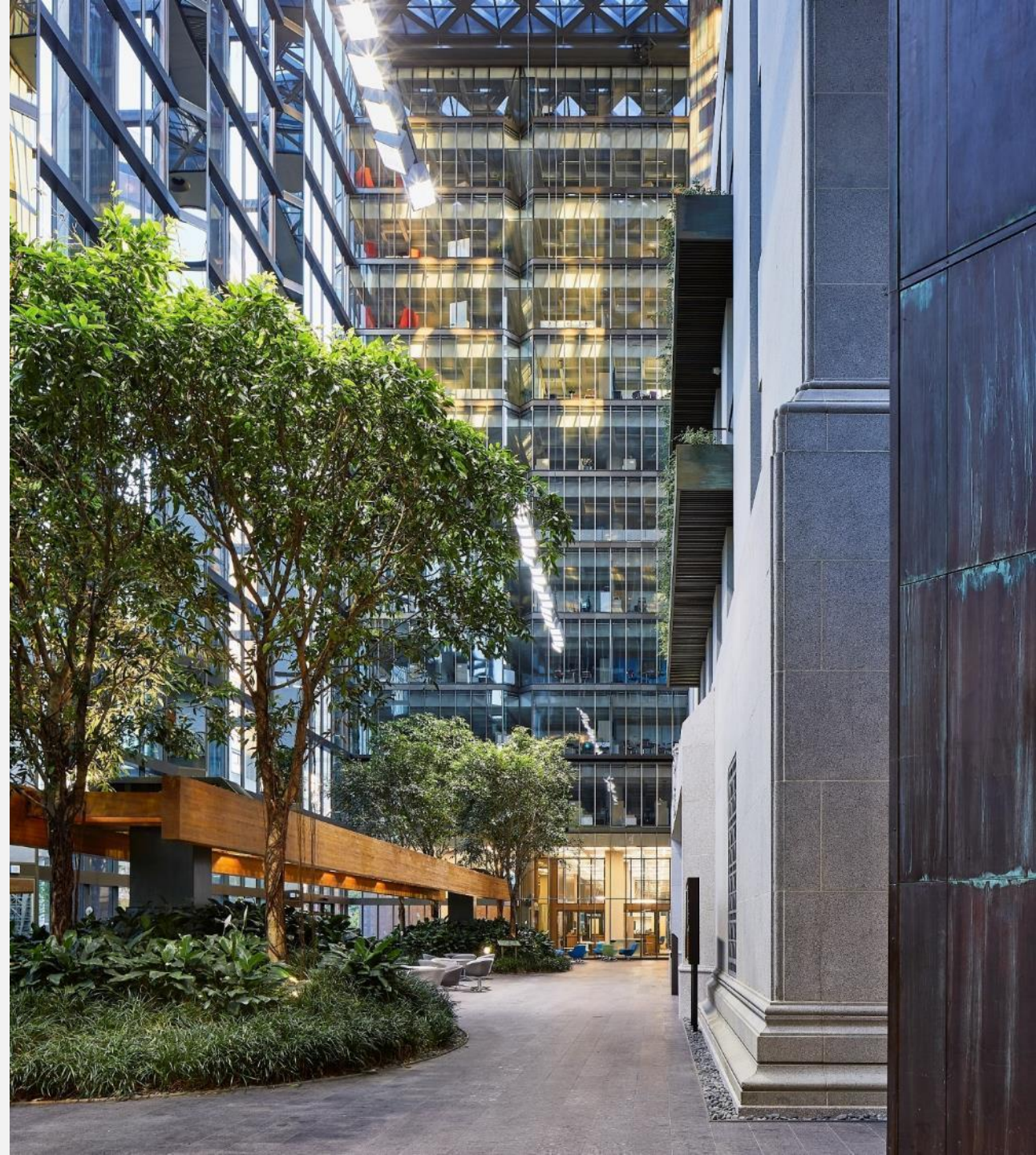
16 FEBRUARY 2022

Using Non-traditional Point of Interest Data as Merchant Survey Sample Frames

BIS-IFC and Bank of Italy workshop on "Data science in central banking"

Presenter: Joy Wu with Marcel Voia and Angelika Welte

THE VIEWS EXPRESSED HERE DO NOT REPRESENT VIEWS OF THE
BANK OF CANADA



Motivation

- The Bank of Canada is conducting a 2021 Merchant Acceptance Survey (MAS)
 - › **Objective:** collect data on how method of payment acceptance at physical point-of-sale has evolved over the COVID-19 Pandemic
- Previous Merchant Surveys faced challenges in constructing the sample frame
 - › **Issues with previous frames:**
 - › Invalid addresses, closed businesses and bad phone numbers
- **Proposed Solution:** Use SafeGraph to improve data quality of record-level business observations in the sample frame

Sources of data and trade-offs

National Statistical agency

- Administrative data
- Probabilistic sampling
- Confidential
- Updates (bi-annual/quarter/months)
- Examples: Statistics Canada ([Business Register \(BR\)](#), [Payroll Deduction \(PD-7\)](#))

Non-traditional data sources

- Convenience/Big Data
- Non-Probabilistic sampling
- Data as a Service (DaaS)
- Updates (weeks/days)
- Examples: Google places, SafeGraph




SAFEGRAPH



Places – 11M+
Global Points of
Interest (POI),
844K+ POI in  Canada


Updated Monthly



Geometry – 7.8
Million Points of
Interest (POI) in
Canada , US, and
GB. 

Updated Monthly



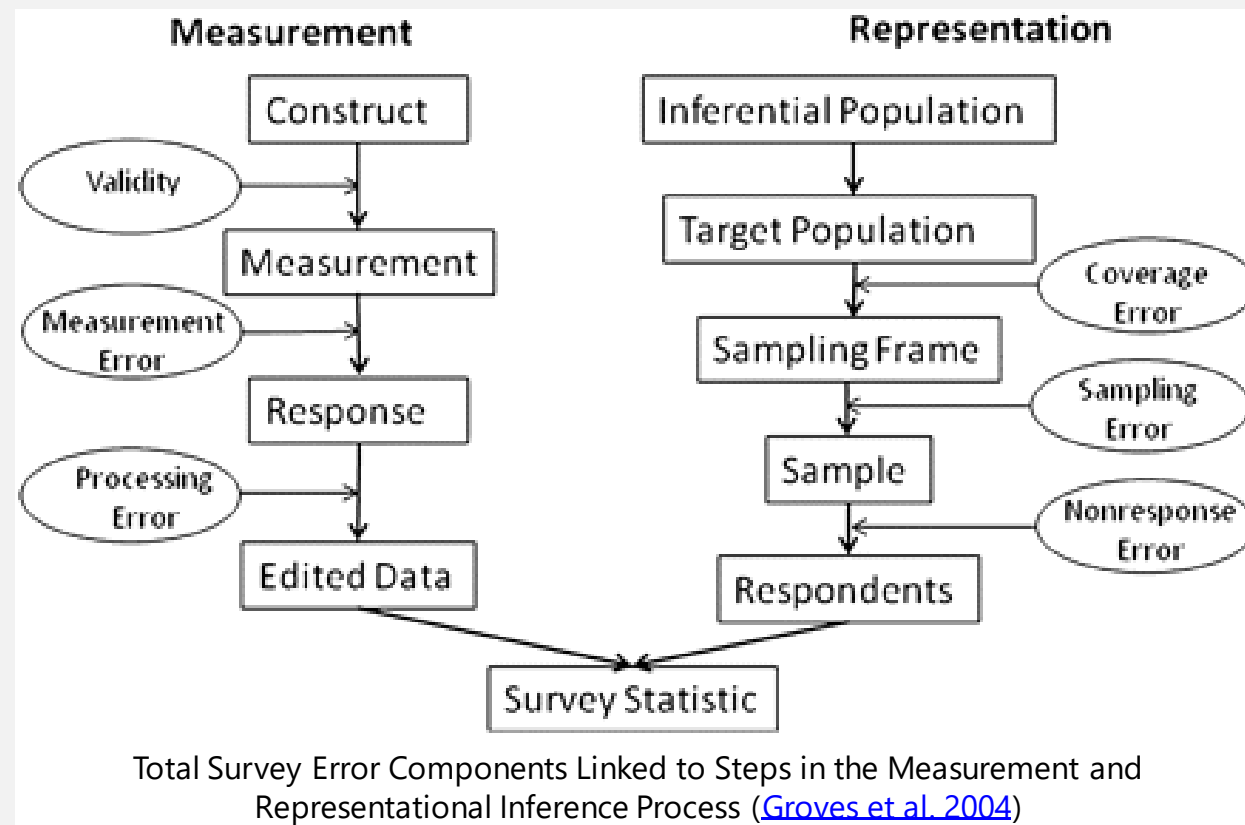
**Canada Weekly
Patterns [BETA]** –
400k POIs and 800
Brands. 

Updated Weekly

SafeGraph data is automatically ingested on a weekly basis using AWS to Azure (Bank of Canada) blob

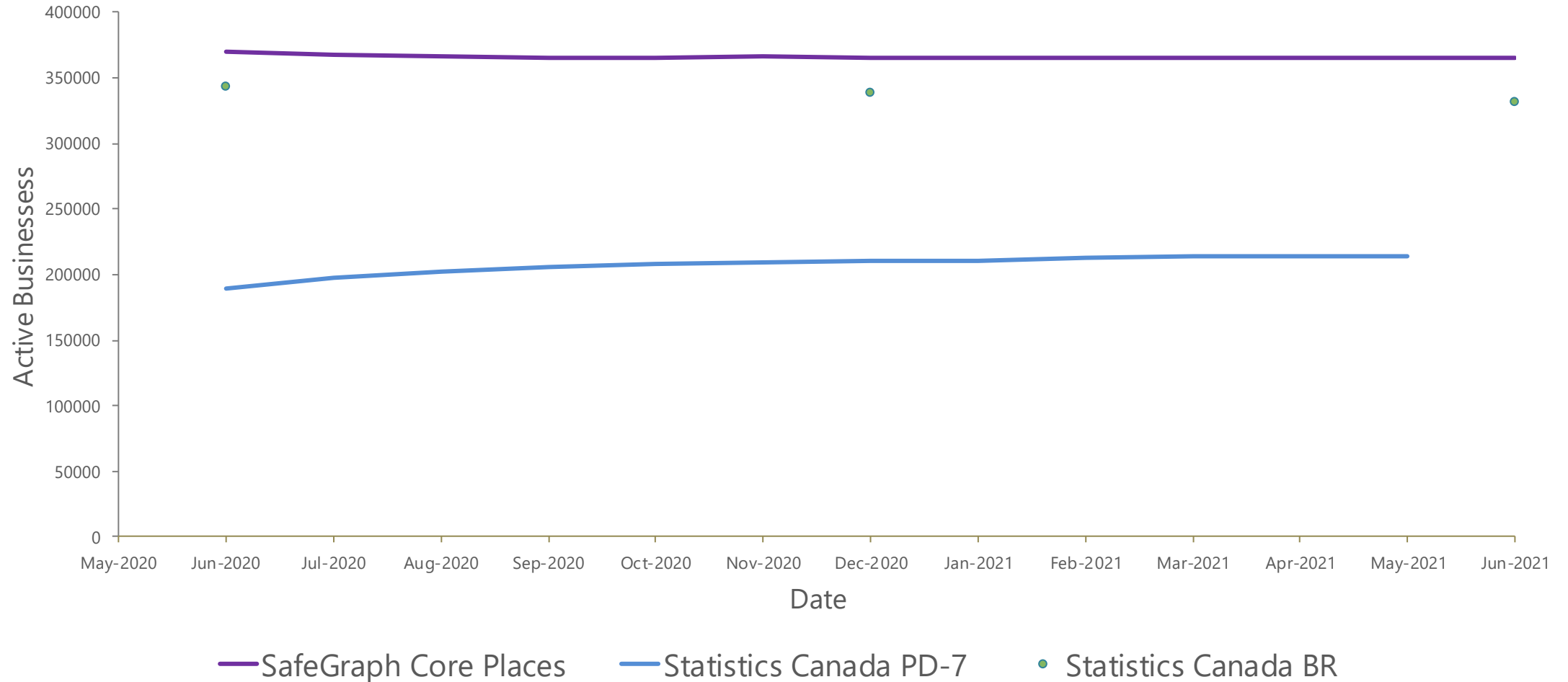
Correspondence with Official Statistics

| | Official Statistics | SafeGraph |
|--|------------------------|---|
| Base Information (Business name, industry, phone number etc.) | Statistics Canada BR | Places |
| Business Activity (Recent and historical) | Statistics Canada PD-7 | Places / Canada Weekly Patterns (Foot traffic) |



Comparison with Official Statistics

Active Businesses (SafeGraph versus Statistics Canada)

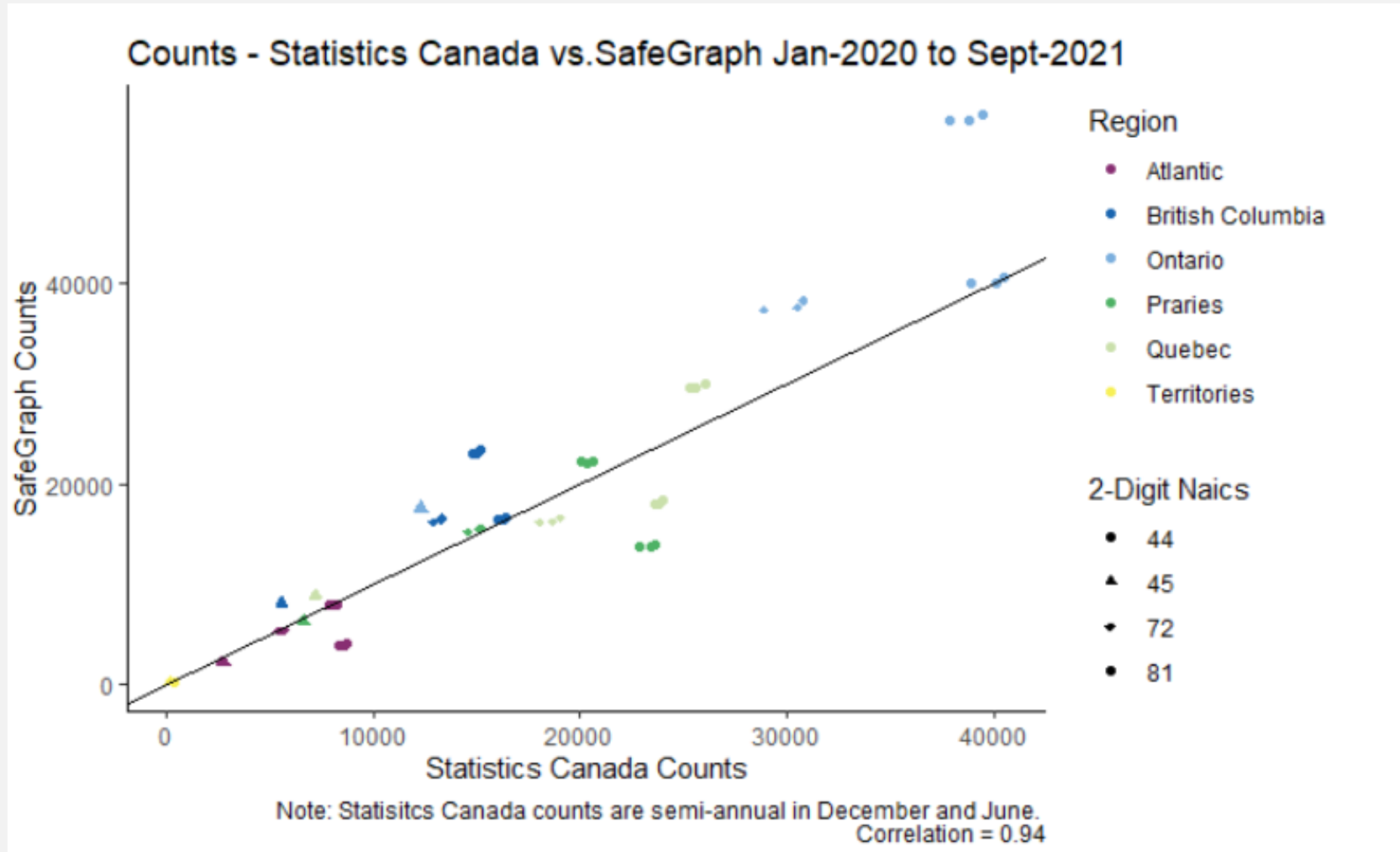


Note: NAICS 44, 45, 72, and 81. All business sizes.

Chart produced in September 2021

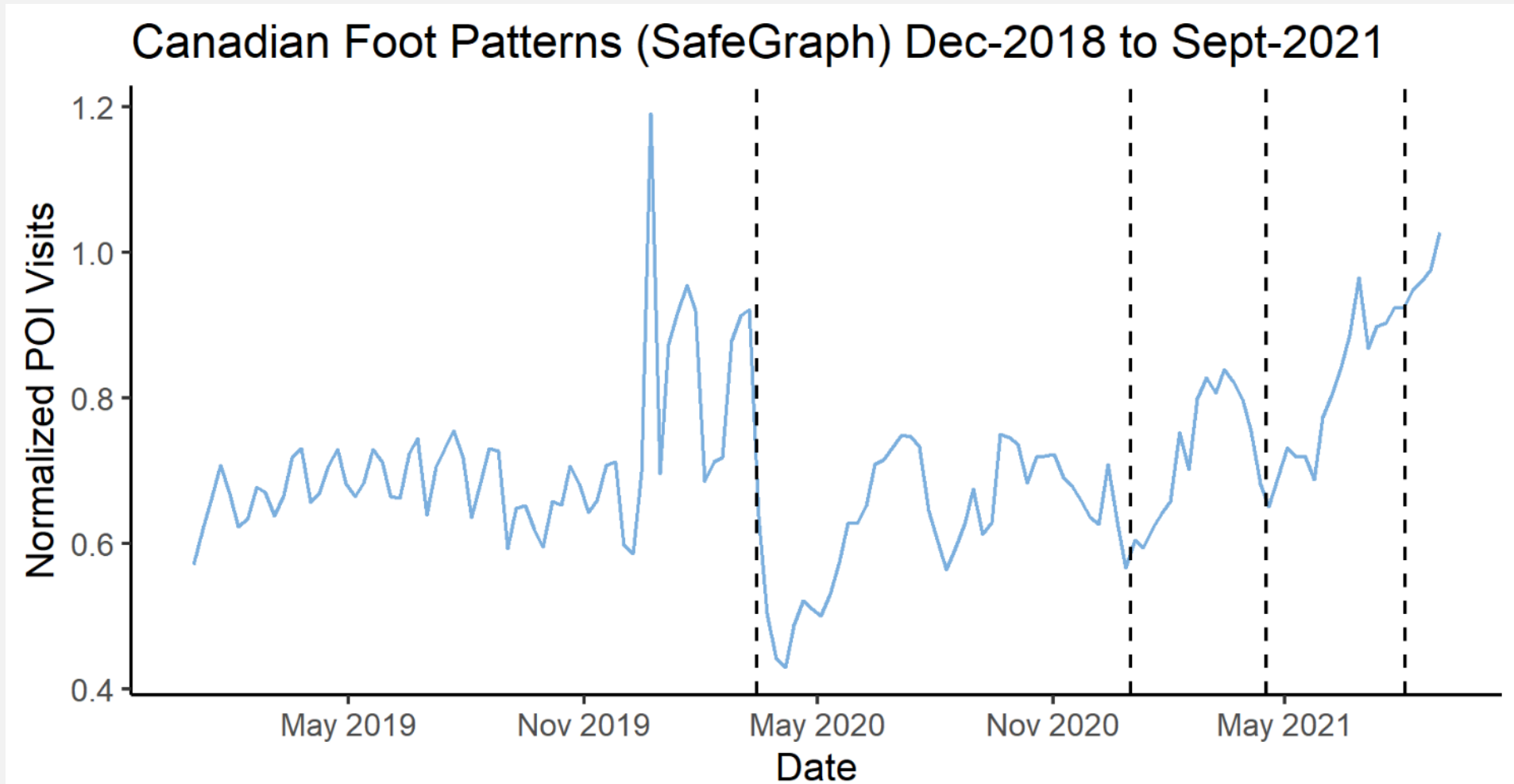
SafeGraph is updated weekly, PD-7 is updated monthly with a three month lag, the BR is updated bi-annually with a lag of two months

SafeGraph Overcount by Region and Industry



- SafeGraph overcounts in distinct region and industry clusters:
 - Ontario 44, 45 and 81
 - British Columbia 44, 81

Foot Patterns

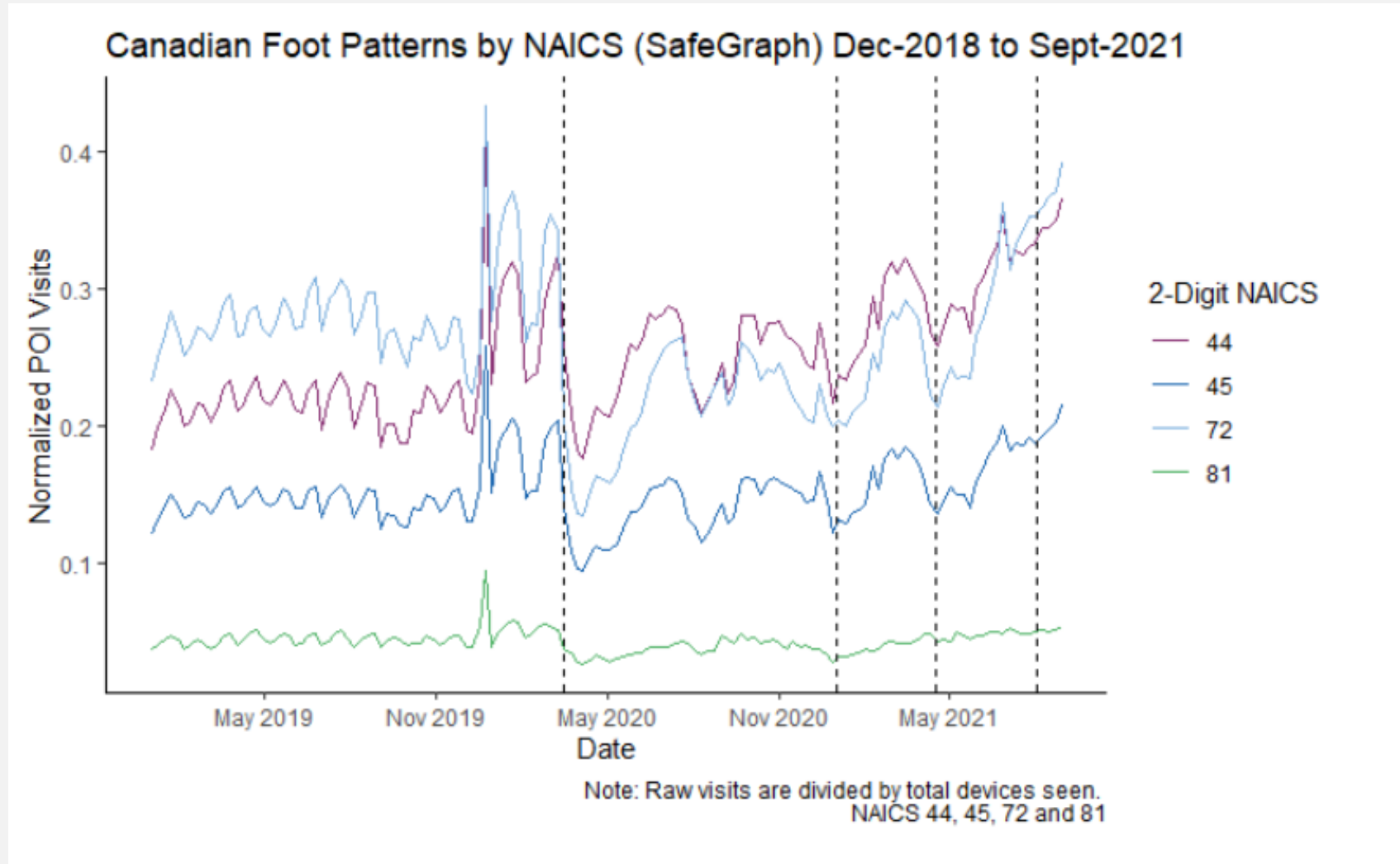


Note: Raw visits are divided by total devices seen.

Note: Approximate Canadian COVID-19 pandemic wave dates indicated by dashed line.

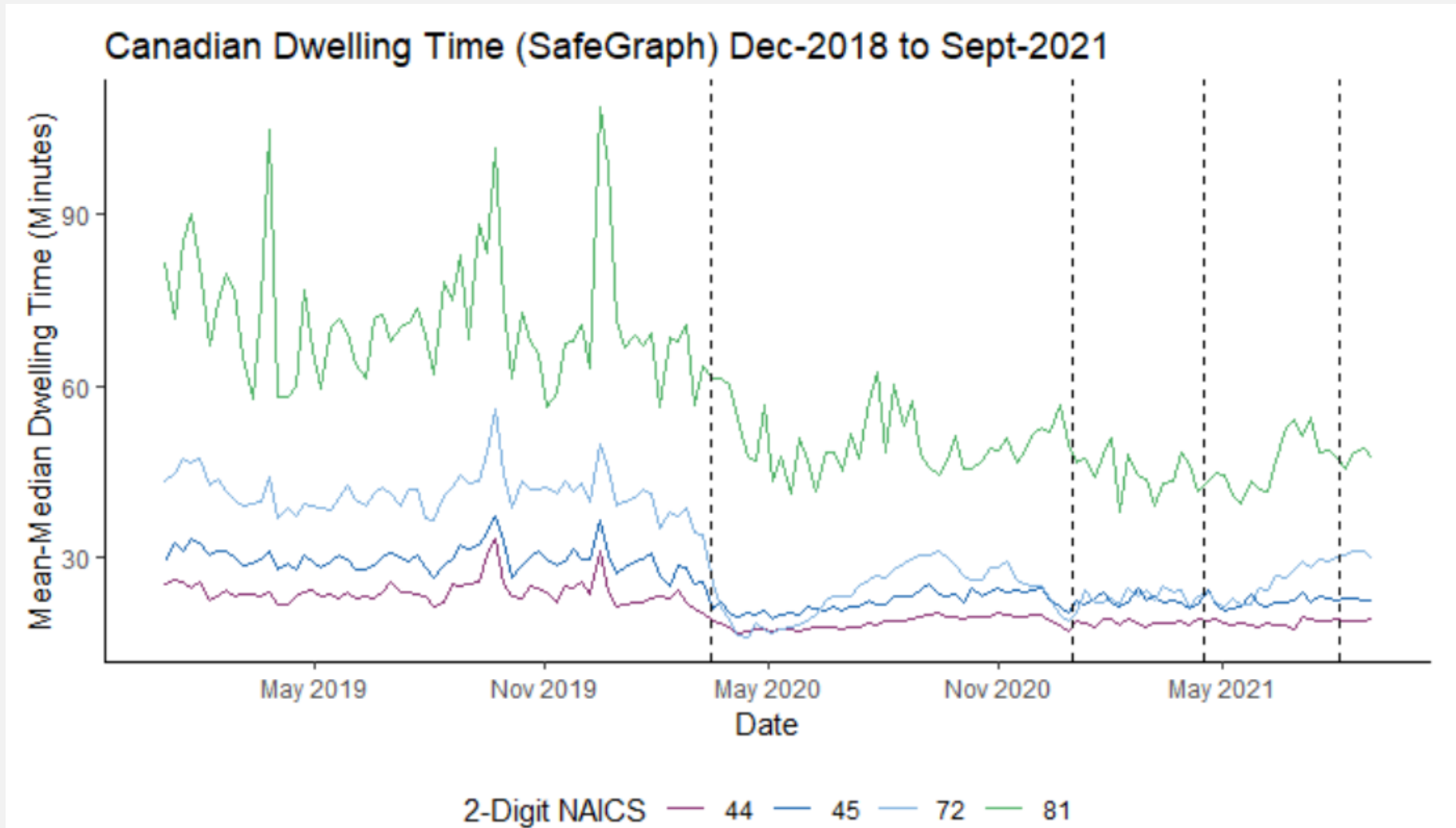
Note: Raw visits are divided by total devices seen.

Foot Patterns by Industry



- Overall increase in foot traffic primarily driven by NAICS 44, 72

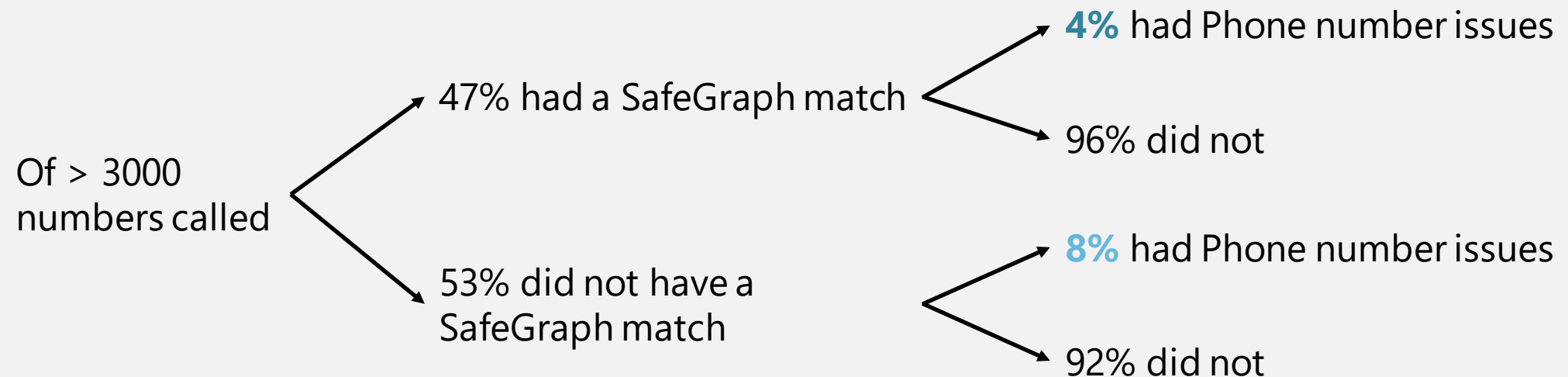
Dwelling Time by Industry



- Increased foot patterns corresponds with decreased dwelling time

2021 Merchant Acceptance Survey Pilot

- MAS Batch 1 was conducted from in 2021
 - › >3000 businesses from an in-house survey frame were contacted



Conclusion

- Complement to national statistical agency data
- SafeGraph provides timely updates on physical business openings/closures
- SafeGraph makes the survey process more efficient
 - › Can help verify addresses, whether businesses are open and closed and bad phone numbers
- More research to automate/integrate processes

Thanks/Merci



IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

gingado: a machine learning library focused on economics and finance¹

Douglas Araujo,
BIS

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

gingado: a machine learning library focused on economics and finance

Douglas K. G. Araujo

Abstract

gingado is an open source Python library that offers a variety of convenience functions and objects to support usage of machine learning in economics research. It is designed to be compatible with widely used machine learning libraries. gingado facilitates augmenting user datasets with relevant data directly obtained from official sources by leveraging the SDMX data and metadata sharing protocol. The library also offers a benchmarking object that creates a random forest with a reasonably good performance out-of-the-box and, if provided with candidate models, retains the one with the best performance. gingado also includes methods to help with machine learning model documentation, including ethical considerations. Further, gingado provides a flexible simulation of panel datasets with a variety of non-linear causal treatment effects, to support causal model prototyping and benchmarking. The library is under active development and new functionalities are periodically added or improved.

Keywords: machine learning, open source, data access, documentation

JEL classification: C87, C14, C82

Contents

| | |
|--|----|
| gingado: a machine learning library focused on economics and finance | 1 |
| 1. Introduction | 3 |
| 2. Data augmentation | 5 |
| 2.1 Basic data augmentation process..... | 6 |
| 2.2 Data augmentation with SDMX..... | 7 |
| 2.3 Is it worth adding more and more data?..... | 8 |
| 3. Automatic benchmark | 10 |
| 3.1 Other use cases for a benchmark model..... | 11 |
| 3.2 Custom automatic benchmarks | 12 |
| 4. Real and synthetic datasets..... | 12 |
| 4.1 Real datasets | 12 |
| 4.2 Simulated datasets | 13 |
| 5. Model documentation..... | 15 |

| | |
|---|----|
| 5.1 Ethical issues in economics and finance for ML models | 18 |
| 6. What to expect next? | 20 |
| 7. Conclusion..... | 20 |
| References | 23 |

1. Introduction

Conceptually, every machine learning (ML) application entails the combination of a specific dataset, algorithm, cost function and optimisation procedure - and each of these components can be replaced more or less independently from the others (Goodfellow, Bengio, and Courville (2016)). The set of possibilities is particularly wide in many economics and finance use cases (Athey and Imbens (2019), Mullainathan and Spiess (2017), Varian (2014), Doerr, Gambacorta, and Garralda (2021) Chakraborty and Joseph (2017)). And there is often no definite answer as to which alternative is best (Mullainathan and Spiess (2017)). Hence, creating a ML application in practice can require multiple iterations and attempted improvements to achieve a satisfactory result. In fact, as of writing this paper, different steps of the process of creating ML models in economics and finance could be more streamlined, ranging from selection and use of the dataset or simulation of a dataset with known generating process, comparison of different ML models, and crucially, also the model's documentation.

gingado is a free, open source library that seeks to facilitate these and other steps in the use of ML in economics and finance in academic and practitioner settings, while promoting good modelling and documentation practices.¹ It offers a number of main contributions. First, gingado facilitates data augmentation of the original user dataset with statistical series from official sources, in a way that is relevant for each case and empirically testable on its ability to improve the model's performance. Second, gingado provides automatic benchmark models that perform well on a broad set of situations and that can train quickly to achieve a reasonable if not the best performance for the dataset at hand; users can also make use of a generic benchmark object to create their own automatic benchmarks. Third, for when the user needs to test a causal inference models, or benchmark existing models, gingado offers functions that flexibly simulate panel data with customised data generating processes that include linear and non-linear interactions and diverse treatment assignment mechanisms and (homogenous or heterogenous) treatment effects. Fourth, gingado promotes documentation of the ML model as part of the development workflow, automating some documentation steps to leave users room for concentrating on more value-added documentation items such as ethical considerations. Also here gingado offers users the possibility for users to create their own model documentation templates that are easy to embed in the modelling practice. And fifth, gingado offers a number of other utilities for helping data science work in economics, in particular in time series of panel data.

¹ gingado's instructions, documentation, practical examples and source code are available at <https://dkgaraujo.github.io/gingado/>. The library is named after the Brazilian concept that is difficult to translate but broadly represents a dance-like non-stop swing of the body that is often associated with flexibility to adjust oneself to adverse situations, eg in life or in a football match. The name is also as an homage to the Afro-Brazilian martial art of Capoeira, where having "gingado" is key. I chose "gingado" as the name for this library focused on using ML in economics and finance because it combines the idea of a constant swing, akin to how economic and financial series also "swing" around a trend through the course of business and financial cycles, with flexibility, similar to how ML models are considerably flexible when fitting functional forms.

gingado is in active development. The library follows three design principles:

1. **Flexibility:** gingado works well out of the box but users can customise its objects in ways that are more suitable for their use cases;
2. **Compatibility:** gingado works seamlessly with other widely used libraries in data science and ML; and
3. **Responsibility:** gingado promotes model documentation and ethical considerations as key steps in the modelling process.

In addition, gingado seeks to be a parsimonious library that complements, rather than redoes, existing functionalities of other widely used libraries. gingado's application programming interface (API) is compatible in particular with scikit-learn (Pedregosa et al. (2011)), which itself is the basis for a variety of other ML libraries, and can be adapted to work with other Python ML libraries with minimal adjustment. In addition, the gingado API can be generally accessed from R (via the integration with Python with the reticulate package, Ushey, Allaire, and Tang (2022)) or from other languages or environments such as Stata and MatLab. The library can be used in a modular way: users might prefer to use gingado only to augment their datasets, create automatic benchmark models, simulate causal datasets or document their models, or any combination thereof.

These characteristics make gingado a potentially useful tool across many domains in economics and finance. ML algorithms are already amply used in economics for prediction problems (in the sense of Kleinberg et al. (2015)), causal inference² (eg, Chernozhukov, Demirer, et al. (2018) and Athey, Tibshirani, and Wager (2019)), model estimation (Maliar, Maliar, and Winant (2021), Fernández-Villaverde, Hurtado, and Nuno (2019) and Duarte (2018)), estimation of models with non-traditional data (Ferreira et al. (2021)) and even being the subject of study themselves (Fuster et al. (2022), Giannone, Lenza, and Primiceri (2021)). gingado's functionalities can be used as appropriate in each instance. Central banks have also been using ML in a variety of applications (Araujo et al. (2023)), and the practitioner use cases in the industry are numerous and diverse.

To showcase practical applications, the online documentation includes two end-to-end examples³ (with more to come over time): the automatic benchmark and model documentation functionalities are illustrated with the cross-country dataset with over 60 variables used to analyse drivers of GDP growth per capita (Barro and Lee (1994)). The dataset was also recently used by Giannone, Lenza, and Primiceri (2021) to study the different predictive abilities of sparse vs dense models. Another example focuses on attempts to forecast exchange rates (Rossi (2013)), illustrating how gingado's utilities can be used to compare different lags of the model, download specific data from SDMX sources, augment the original dataset with other relevant data, quickly create a benchmark model and use it compare different alternatives, and finally how to document the model to promote responsible model maintenance and usage. The examples illustrate ways in which gingado can help economists' workflow.

The next section describes how gingado facilitates data augmentation. Section 3 outlines the automatic benchmark process, followed by a discussion of real and

² A recent compilation of causal inference techniques from various domains is <https://neurips.cc/virtual/2021/workshop/21871>.

³ Available at: <https://dkgaraujo.github.io/gingado/>.

simulated data functions in Section 4. Section 5 describes gingado's model documentation framework. The last section concludes. The online documentation describes how to install the most up-to-date version of gingado. More advanced topics like how to customise automatic benchmark models and model documentation templates can also be found online.

2. Data augmentation

Publicly available data have a long tradition in economics and finance research and practice. For example, the Federal Reserve Bank of St. Louis' Federal Reserve Economic Data (FRED) system has developed in tandem with wider adoption of the internet itself (see Stierholz (2014) for an interesting narrative of FRED's history). Similarly, numerous other central banks, statistical agencies and international financial institutions put datasets in the public domain in one form or the other. A number of statistics organisations created in the early 2000's an initiative to promote the collection, compilation and dissemination of statistical data, the Statistical Data and Metadata Exchange (SDMX), which is now in its version 3.0.⁴ In addition, other data aggregators such as Base dos Dados and DBnomics host an incredible amount of economic and financial series.

Many of these services allow users to access data in a programmatic way, ie setting up a computer programme to download the data instead of the user manually accessing the website, selecting the data, downloading it in a file and incorporating the data in the analyses. Thanks to SDMX and to the broader availability of user-friendly data APIs like the one offered by FRED, querying data from trusted sources to augment the user's original dataset has become much easier than in the past. This allows more research to benefit from the main advantages of programmatic data access as opposed to manual downloads: the data acquisition is reproducible, auditable and scalable. Accessing data programmatically also allows any numeric transformations, consistency checks and data imputation routines that are applied on the dataset to be done in a reproducible and transparent way.

In addition, programmatic access to data can also ensure that any data that are added to the original dataset are done so in a consistent way. For example, SDMX includes the concept of "codelists", which are standardised definitions that apply across dataset domains. One specific codelist contains all possible realisations of the frequency of a dataset (ie, daily, weekly, monthly, etc.), and another codelist encompasses all possible codes for specific currencies. This technology ensures that the user has greater control over the datasets to be incorporated.

The considerations above are even more important when models are in the production stage. In economics and finance, ML models are occasionally designed to be run in production settings instead of a one-off execution; and this is often performed by users that were not involved during development and therefore are not as familiar with the model's internal functioning. For example, a model forecasting stock prices might be used extensively over time - by stock traders or portfolio managers, not the developers. These situations tend to take place in situations where

⁴ A technical description of version 3.0 is found here: <https://sdmx.org/wp-content/uploads/SDMX-3-0-0-SECTION-1-FINAL-1-0.pdf>.

the ML model will require future observations of the datasets used for training. Therefore, automating the process of data augmentation in a way that can ensure consistency between datasets used for training and later for inference helps to insulate users from worrying about these processes related to the use of additional data.

gingado aims to facilitate this process of finding and adding new datasets, leveraging the public availability of reliable data from official sources and automatically ensuring that the augmented dataset is consistent with the original, by having the same frequency and time period. With this, the same publicly available dataset(s) used during model training are downloaded and used each time with the appropriate time period when the model is run in the future, potentially by other people or automatically by systems. The current version of gingado achieves this by means of its data augmentation object `AugmentSDMX`; the idea is to gradually add to the public codebase other "data augmenters" that can scan and download publicly available datasets in a way that is consistent with the original dataset.

2.1 Basic data augmentation process

Similar to other parts of gingado, the API for data augmentation is designed for compatibility with scikit-learn. Specifically, the user decides which specific sources and dataflows will be scanned for data upon instantiation of the data augmener object. Until then, the object does not perform any activity other than to store this and other parameters. It is only when one of the methods `fit`, `transform`, or `fit_transform` are called that the data augmener will (a) read characteristics of the user's data and (b) search in the named sources and dataflows for datasets that correspond to the same frequency and time period. This process is shown in Listing 1, which assumes the user already has a data frame called `X_train` with training data *indexed by time*: gingado uses this time period information behind the curtains to obtain data of the relevant frequency and time periods. In the listing, the user first imports the class `AugmentSDMX`, and then define a dictionary `src` to list the SDMX sources and their relevant dataflows - more on how to find the sources below. The data augmener in this example is set to look for the Composite Indicator of Systemic Stress (CISS) by the European Central Bank (ECB) and the central bank policy rates dataflow from the Bank for International Settlement (BIS). In the next line, an instance of the class `AugmentSDMX` is created using those sources listed in `src`, and the method `fit_transform` learns the frequency and time periods from `X_train` and use that knowledge to fetch all data series from these two sources that comply with these time requirements.

Basic example of the AugmentSDMX application programming interface

Python environment

Listing 1

```
from gingado.augmentation import AugmentSDMX
src = {'ECB': ['CISS'], 'BIS': ['WS_CBPOL_D']}
augmented_X_train = AugmentSDMX(sources=src).fit_transform(X_train)
```

Sources: Author.

The process described above will result in `augmented_X_train`, a dataset that contains the original data, now augmented by potentially numerous other series. The data are indexed by time, and for this reason every combination of permissible codelists in the augmented dataset will create a different variable (ie, a different column). For example, even if the original dataset represents only data from Brazil, the augmentation process shown in Listing 1 I will append the CISS for every country in the list as a different column, as well as the central bank policy rates. `gingado` makes an explicit choice to allow for this, instead of filtering by individual (in this example, retrieving only the dataset about Brazil), because the newly added data can have some level of information on the target variable, and if that is the case the model would probably uncover it. Of course, users that desire to add only data relating to a country, currency, etc might add those as relevant filters, as described in more detail in the documentation.

One important note is that the user remains responsible for ensuring that the data being automatically added is not a covariate that will interfere with the desired statistical properties of the task at hand. In other words, if the economist running the model is interested in anything more than just a predictive exercise without any statistical properties, then greater attention to the potential covariates being added is warranted. For example, ensuring that the covariate is not endogenous or otherwise a bad control (Hünermund, Louw, and Caspi (2023) discuss the sensitiveness of some ML-based inference to inclusion of bad controls.) In some settings, one strategy to ensure only adequate covariates are included is to consider well which data sets will be searched for by `gingado`: ie make sure to include only those broad groups of data that will not negatively interfere with the model inference.

2.2 Data augmentation with SDMX

`gingado` explores and fetches available datasets from official sources to augment user data using SDMX, an initiative by international organisations (BIS, ECB, Organisation for Economic Cooperation and Development - OECD, International Monetary Fund - IMF, World Bank - WB, Eurostat, and United Nations - UN) that develop and publish statistics from these sources. Since its inception in 2001, SDMX has grown to be used by other sources as well - primarily central banks and statistics agencies - as a standard to disseminate their data.

Downloading data from SDMX sources can be advantageous to users because of the variety of sources and the consistency of the concepts describing the datasets. Instead of dealing with the specific data descriptions and download processes of each of the SDMX participant institutions, users can rely on an API based on standardised concepts to fetch the data. For example, using SDMX to look across multiple sources for data at quarterly frequency related to the countries of Argentina, Brazil and South Africa for the period spanning the first quarter of 2015 to the fourth quarter of 2021 would use the codelists (introduced above) for frequency and for reference area. These are the same across the SDMX sources and dataflows, which facilitates the process of finding relevant datasets.

`AugmentSDMX` is currently based on the `pandasdmx` python package. The backend code searches all listed SDMX sources in this package, and retrieves the dataflows for those it is able to get (some sources may time out). Given that downloading all the relevant data from all sources could be a time-consuming operation, users define the SDMX source(s) from which to get the data. For each

source, the user might define the specific dataflows or ask for all dataflows of that source.

As of the time of writing, the data providers available for automatic download of data using AugmentSDMX are:⁵

- Australian Bureau of Statistics
- BIS
- Countdown 2030
- Deutsche Bundesbank (Germany)
- ECB
- Eurostat
- International Labour Organization - ILO
- International Monetary Fund - IMF
- National Institute of Statistics and Geography (Mexico)
- National Institute of Statistics and Economic Studies (France)
- National Institute of Statistics (Italy)
- National Institute of Statistics (Lithuania)
- Norges Bank (Norway)
- National Bank of Belgium
- Organisation for Economic Cooperation and Development -OECD
- Pacific Data Hub
- Statistics Estonia
- United Nations Statistics Division
- UN International Children's Emergency Fund (UNICEF)
- World Bank Group's "World Integrated Trade Solution"
- World Bank Group's "World Development Indicators"

Each of those sources offers a number of dataflows, which are closely related datasets. For example, one dataflow related to foreign exchange rates could include the time series of multiple individual exchange rate pairs, and each pair can be downloaded in their nominal or real exchange rate. The dataflows from all of these sources could be available to train the ML model at hand. In total, the sources listed above result in 9,110 dataflows.

2.3 Is it worth adding more and more data?

While an increasing amount of data can lead to better results by ML models, there are situations where users might want to consider limiting the amount of data being fed to a model. The computation time might increase as the number of SDMX series are added, which is an important consideration as models in production also need to

⁵ Users can get the up-to-date list with the gingado method `'list_SDMX_sources'` in `'gingado.utils'`.

look at the same variables used during training. Depending on how time-consuming downloading this new data is, having bigger datasets could significantly affect computation speed at run time. However, for cases where immediate response is not required, a bigger quantity of data could be considered reasonable, especially if it leads to a marked performance improvement. Therefore, the answer to this subsection's title will depend on each use case, and gingado can help the user answer it.

AugmentSDMX's API is compatible with scikit-learn in a specific way that allows it to be included in a pipeline. Pipelines are objects that apply a specific sequence of transformations on data, and possibly a final step consisting of an estimator (such as a predictor). These pipelines exist to compare sequences of steps as a whole with proper cross-validation, and to allow for a consistent way to estimate different versions of the same model without losing control of the data pipeline.

What this means in practice is that the user can apply standard parameter search algorithms such as grid search to test whether or not the inclusion of a particular dataset will impact the model, eg by improving its performance, changing the importance of regressors, etc. This process is exemplified in Listing 1: the user created with a few lines a model that, when fitted, will estimate two versions of the model: one without augmentation (ie, will "pass through" AugmentSDMX) and another with augmentation. The first four lines import the necessary objects from gingado and scikit-learn. Then, an instance of the data augmenter is created in the variable `sdmx`, in this case using all dataflows made available by the BIS. Note that unlike Listing 1, neither the method `fit` nor `fit_transform` were called yet and therefore no calculation or data download is done at this stage. In the next step, an instance of a Pipeline object is created - but still not fitted. This pipeline chains together two steps: the data augmentation and a random forest. Finally, a dictionary with a grid of parameters to be tested empirically and grid, a variable that performs an ML calibration using grid search with cross-validation, are created. `param_grid` is the key part of this listing, since it tells grid to test the two versions of the ML model: one that bypasses the augmentation step and uses only the user-provided dataset, and the other one that uses the `sdmx` variable to augment automatically the user dataset with all relevant and compatible (ie, with the same frequency and time periods) data from the BIS. When fitted using training data on covariates and dependent variables, grid will select the parameters in `param_grid` that result in models with the best performance. Thus the question of whether or not to use more data can be answered empirically and in a completely data-driven way. Beyond that, the user can even jointly search richer combinations of different parameters governing data augmentation, data transformation and model estimation by combining AugmentSDMX with scikit-learn's Pipeline.

Use of AugmentSDMX in a scikit-learn pipeline

Python environment

Listing 2

```
from gingado.augmentation import AugmentSDMX
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.pipeline import Pipeline
```

```
sdmx = AugmentSDMX(sources={'BIS': 'all'})

pipeline = Pipeline([
    ('augmentation', sdmx),
    ('forest', RandomForestRegressor())
])
param_grid = {'augmentation': ['passthrough', sdmx]}
grid = GridSearchCV(
    estimator=pipeline,
    param_grid=param_grid
)
```

Sources: Author.

3. Automatic benchmark

gingado offers users the possibility of automatically developing a benchmark ML model fine-tuned for their particular case, in a short time and with no input from the user other than the data (although users can also fine tune all aspects of the benchmark). This is achieved by means of an embedded parameter grid search mechanism that evaluates different versions of the underlying algorithm on the user's data, and selects that one that performs better as the benchmark.

The objective is to help the user during the exploratory phase of the development of a ML model. The practice of establishing a baseline model is common in ML practice. Without a goalpost, a baseline model that is relatively simple to understand and benchmark against, it can be difficult in practice to realise if one's model is actually performing well or just improving from a low base. So gingado allows users to quickly create a fully functioning model that can serve as benchmark, as shown in Listing 3. Similar to Listing 1, this listing assumes the user already has available two data frames: `X_train`, containing a panel of covariates, and `y_train`, which is the dependent variable. The first line imports the object `ClassificationBenchmark`. In the second line, an instance called `benchmark` is created and already fitted in the same line. Behind the scenes, this instance of `ClassificationBenchmark` will perform a grid search using a random forest with default parameters that tend to work well in my experience in a variety of datasets that tend to be in line with the size of empirical papers in economics. The fitted object then represents the model with the best performance, and can be used by the user for prediction, etc. Naturally, the user can pass other estimators and also other sets of parameters for the grid search, illustrating gingado's combination of convenience with flexibility.

Creation of an automatic benchmark model

Python environment

Listing 3

```
from gingado.benchmark import ClassificationBenchmark  
benchmark = ClassificationBenchmark().fit(X_train, y_train)
```

Sources: Author.

The off-the-shelf implementations of this automatic benchmark model are based on random forests (Breiman (1996), Breiman (2001)), with one type for regression tasks and another one for classification. Random forests, one of the most widely used ML methods according to industry practitioners (Howard and Gugger (2020a)) present several advantages that make them good candidates for benchmark models. They tend to have a very good out-of-sample fit across a wide variety of data generating processes, and require little data preparation work compared to other ML models. Random forests also provides an intuitive way to measure the individual importance of regressors, although they don't lend themselves to interpreting the channels by which the regressors contribute to the prediction (Varian (2014)). These importance measures are the mean reduction in impurity occurring in the trees that use that particular variable. The values are then typically scaled so that the sum of all feature importance measures is always one. Practitioners use this measurement as one possibility for variable selection (Géron (2019), Kohlscheen (2021)).

Whenever a benchmark model is fitted to the data, the model automatically jumpstarts its documentation from a template and automatically filling some information on the model. In the off-the-shelf implementation, this documentation is done via the ModelCard object, described in more detail in section 5. From then on, the model documentation can be accessed - and most importantly, filled out by the user.

Benchmark models also have a specific method to compare themselves with other candidate models. This allows users to directly compare their own candidate models with the existing benchmark. When this is done, via the module `compare`, gingado also includes amongst the candidates to be tested an ensemble combination of all the candidates and the current benchmark. The inclusion of this candidate ensemble serves to leverage the advantage that ensemble models have over simpler models (Giannone, Lenza, and Primiceri (2021)).

3.1 Other use cases for a benchmark model

In addition to serving as a baseline model during the development of ML, users can simply retain the automatic benchmark as their production model. Beyond benchmark-specific functionalities, these objects behave also as normal scikit-learn models and therefore can be applied as any normal model would.

Benchmarks can be used as a test to see which regressors, if any, differentiate the most between two or more groups, for example to see if one group of samples has out-of-domain data (as proposed by Howard and Gugger (2020a)). The test proceed as follows: a random forest classifier is trained on the dataset, with group identifiers serving as the target variable. The forest's calculated regressor importance

can then uncover what are the variables that differentiate between groups. This test can be generalised to identify which regressors vary substantially between two or more groups.

3.2 Custom automatic benchmarks

In spite of the empirical qualities of random forests, no single algorithm could plausibly cover all potential use cases. For example, random forests could potentially not perform as well as other established algorithms if the economic or financial data at hand contains multimodal data, ie images, videos, texts and other such data. These data types are increasingly relevant in economics research. Some examples of this growing literature on non-traditional data types include Gentzkow, Kelly, and Taddy (2019) and Li et al. (2020) for text and Yeh et al. (2020) and Bajari et al. (2023) for images. In addition, random forests cannot extrapolate beyond the range of the training data that was fed to them. And there are some empirical settings in which gradient boosting trees are more used than random forests and other ML methods (Brotcke (2022)). Similarly, Gorishniy et al. (2021) show that neural networks can achieve similar, and in some cases better, performance than tree-based methods in some cases. And Taylor and Letham (2018) describe Facebook's own tool (now open sourced) for automatic forecasting, without relying on random forests.

So random forests might not be the first choice for all cases, even though they are expected to work well in a wide array of situations. In addition to the off-the-shelf implementations described above, `gingado` offers two possibilities for users to set up their own benchmark models. The most straightforward one is to simply pass as argument a model object to the benchmark's method `set_benchmark`. This will cause the benchmark object to put this new model in place of the previous benchmark.

The second object involves the creation of a new benchmark object class altogether. `gingado` ships with a base class, `gkdBenchmark`, that contains all the necessary features for a benchmark object. Users that wish to create their own benchmark models can sub-class from `gkdBenchmark` and implement the desired functionalities. The user's benchmark will then work in the same way as the original `gingado` benchmark models. This user-created benchmark can include any algorithm or combination of algorithms, as long as the API is maintained.

4. Real and synthetic datasets

`gingado` aims to provide users with an easy way to load real datasets that can be used as benchmarks in research, along with functions that allow users to create synthetic datasets from a wide range of data generating processes.

4.1 Real datasets

Economics and finance research often relies on benchmark datasets used in canonical papers to explore new insights derived from the original work or to evaluation and question the original findings. Prominent (but not exhaustive) examples include the [Angrist Data Archive](#); the data used by Giannone, Lenza and Primiceri (2021) to test ML models; the macro-history datasets of Jordà, Schularick, and Taylor (2017), Jordà

et al. (2019) and Jordà et al. (2021); the technology adoption (CHAT) dataset of Comin and Hobijn (2009); and the datasets used in synthetic control studies by Abadie and Gardeazabal (2003), Abadie, Diamond, and Hainmueller (2010) and Abadie, Diamond, and Hainmueller (2015); and others. Some of these benchmark data may be useful for building and testing ML algorithms.

In many cases, these datasets are shared in a way that makes them convenient to use with almost no manipulation. But even in these cases, these datasets are formatted using different platforms (Stata's DTA file, or in CSV/Excel files, etc). What gingado aims is to make available selected benchmark datasets in a standardised format that can be readily used by economists. At this point, the only dataset loaded this way in gingado is the one from Barro and Lee (1994); Listing 4 illustrates its use. After importing the `load_BarroLee_1994` function in the first line, the user then attributes the result of calling this function to two datasets, `X` and `y`, which can then be used as normal for the training of ML models. The goal is for other datasets to have a similar structure, where their usage can be as simple as importing the appropriate function and calling it once. Naturally, the original source of the all data used should be cited and acknowledged accordingly by the end user.

Importantly, gingado does not aim to restrict this section only to datasets that support well-cited papers. Users that want to propose other datasets, including from their own work, are welcome to do so. In any case, the data must be used in a published academic paper.

Loading data from Barro and Lee (1994)

Python environment

Listing 4

```
from gingado.datasets import load_BarroLee_1994()

X, y = load_BarroLee_1994()
```

Sources: Author.

4.2 Simulated datasets

In many cases, researchers testing new econometric estimators use simulated data, created under "lab-like" conditions with a known data generating process. Such datasets enable the user to test whether proposed estimators really work as intended for a dataset with given characteristics, and can be especially helpful for causal estimands (Imbens and Rubin (2015)). The possibility of simulating a datasets of varying lengths - on the time and the cross-sectional dimensions - also facilitate the testing of asymptotics. gingado's `make_causal_effect` function offers users the possibility to simulate non-trivial data, including with non-linear interactions and a rich set of treatment-related variables.

More specifically, `make_causal_effect` allows users to creates a dataset $\{y_i, X_i, D_i\}$ of outcome variable, covariates and treatment vectors respectively, with the number of samples $i = (1, \dots, n)$ of the dataset, which can be interpreted as peer units or as time periods in a panel data, and $k = (1, \dots, M)$ number of features. Users are able to specify the functional form by adjusting the following characteristics:

- $y_i | X_i$, the pre-treatment outcome. For the untreated units, this corresponds to the observations of y_i ; for the treated units, this is the portion of y_i before adding

the treatment effect. The pre-treatment outcome depends on the covariates X_i and on a constant. It might also have a random component⁶

- $W_i = p(D_i \neq 0 | X_i)$, i 's propensity of being treated, ie having a treatment that is different than zero.⁷ The propensity can be a function that depends on X_i , either deterministically or with a random component. Alternatively, it can also be set with a scalar, in which case it is uniform across all samples, or with a random assignment to each sample according to a parameterised or empirical random distribution; in both cases the propensity simplifies to $p(W_i = 1 | X_i) = p(W_i = 1)$.
- $D_i \neq 0 | W_i$, the actual treatment assignment. The default value is purely a function of the treatment propensity through a binomial distribution: $p(D_i \neq 0 | W_i) = B(1, W_i)$. This is probably the most relevant use case. However, there are instances where researchers might want to explore a more complex treatment assignment relationship. For example, treatment can be rationed and only applied to the ψ samples with the highest W_i , or applied with probability $B(1, W_i)$ but also subject to this rationing. In these cases, the treatment assignment would also be a function of ψ , in addition to W_i . For example, the case where only the ψ most propense units would be treated can be defined as: $1(D_i \neq 0) = B(1, W_i) \times 1(B(1, W_i) \in \text{top}(W_i, \psi))$, where $\text{top}(\cdot, n)$ is the function that orders the set and returns the highest n observations.
- D_i , the treatment value. The treatment can be set to 1 in the most simple of cases, but the treatment magnitude (and sign) can also vary as a function of covariates X_i , for $D_i | X_i$. A random variation in the sample-specific treatment can also be introduced.
- $\tau_i | D_i$, the treatment effects on y_i for each treated sample. This represents the difference between the actual observation y_i and the pre-treatment outcome y_i for treated samples. In the most simple cases, it may be a one-to-one mapping to the treatment value D_i . But, it may also vary by sample according to covariates X_i .

All dependencies on covariates above can include non-linearities such as minimum or maximum comparisons, various types of interactions, exponentials, etc. In short, anything that can be coded using a NumPy array and respects the necessary constraints of each argument (for example, the treatment propensity must always be a number between 0 and 1). It is also important to note that a more complex treatment chain (propensity to assignment to value to effect) can depend on covariates in a different way at each step. Hopefully this flexibility in creating datasets with causal effects can provide researchers with ways to compare different ways to correct for interferences in the estimation of different potential outcomes.

⁶ All random components in gingado code can be made reproduceable with the use of the same random seed.

⁷ The treatment value itself can be set by the user, and therefore is not necessarily a binary dummy.

5. Model documentation

Stakeholders often require some level of documentation of the ML models. From simple descriptions of the model to standardised model reports to fully-fledged evaluation of models, gingado provides a way for models to be more easily documented. The basic idea is that a *documentation template* outlines the specific items to be documented, and various methods allow the user to interact with this template. This template can be seen at any time by the user with the method `show_template`. A gingado documenter object also specifies which of those items are to be filled in automatically (eg, model description can be parsed from the ML algorithm itself), and how exactly this should be done.

After a documenter is instantiated, it can read information from an existing model as well. Listing 5 demonstrates how straightforward it is to automatically read information from a model, and then see what are the questions from the documentation template that were not answered automatically by the `ModelCard` object, even when the ML model itself is not a gingado object.⁸ The first two rows import the necessary objects: a `gingado.ModelCard` class and the keras ML library. The following chunk establishes the structure of a neural network comprising of two dense layers (each with 16 nodes) followed by an end layer that predicts the probability in a classification model. This neural network is then fit using training data frames `X_train` and `y_train` assumed in Listing 5 to already exist, as before. Now, an instance of `ModelCard` object, called `model_doc_keras` is created, and it then reads the information from the keras classification ML model that was created and trained right before. Finally, in the last line the code presents to the user what are the questions, or fields, in the model documentation that are still open because they were not read automatically. This nudges the user to consider answering these questions (and thus documenting their model in an easy and recordable way).

Example of `ModelCard` reading information from an existing neural network model

Python environment

Listing 5

```
from gingado.model_documentation import ModelCard
import keras_core as keras

keras_clf = keras.Sequential()
keras_clf.add(keras.layers.Dense(
    16,
    activation='relu',
    input_shape=(20,)
))
keras_clf.add(keras.layers.Dense(16, activation='relu'))
```

⁸ Currently the base documenter `ggdModelDocumentation`, from which all documenters in this library derive, can read models created using gingado, scikit-learn and keras. Support for automatically reading information from models built with other libraries such as PyTorch is under development.

```
keras_clf.add(keras.layers.Dense(1, activation='sigmoid'))
keras_clf.compile(optimizer='sgd', loss='binary_crossentropy')

keras_clf.fit(X_train, y_train, batch_size=10, epochs=10)

model_doc_keras = ModelCard()
model_doc_keras.read_model(keras_clf)
model_doc_keras.open_questions()
```

Sources: Author.

At any time, the user can view the current state of the document with the method `show_json`, and save or read it to file with `save_json` and `read_json` respectively. Additional information items that were not included automatically are filled with `fill_info` (the user can also override automatic entries). And the remaining items from the template that are not yet filled are shown to the user with the method `open_questions`. These methods (and others, not shown for brevity) aim to provide the user with a more direct, hands-on approach to documenting the model, compared to a more traditional setting where a separate document (ie, an MS Word file) need to be written and maintained. Allowing users to document their models from inside their ML development environment will help embed the documentation process as another step of the model development. Another advantage of gingado's approach is that it is much easier to keep the documentation aligned with the current version of the model. This is particularly important in the settings where the user expects to iterate over different specifications until a suitable model is achieved.

The model documentation is stored as a JSON file, a flexible format that is easy for machines to read, and that can quickly be transformed into objects humans can read more easily, too. gingado uses JSON files because they are a common language to serialise this type of structured information that works across platforms (ie, a JSON file works in Windows machines the same way it works in MacOS, Linux or any other system). Users that want to automatically produce documents in other formats (eg, PDF files) can do so by using these JSON files with other libraries of their choice. And in settings where multiple models are developed and in production, JSON files are a more streamlined format to offer information on each model to comparison scripts.

gingado includes two ready-to-use documenter objects, `ModelCard` and `ForecastCard`. The `ModelCard` documentation template is based on the work of Mitchell et al. (2018), which I believe strikes a good balance between being a general documentation template while also prompting the user to answer questions about the model that are relevant from a broader stakeholder perspective. `ForecastCard` is a version of `ModelCard` with questions that are targeted for forecasting tasks. Similar to how users can create their own class of benchmark models, gingado enables users to create their own custom documenters, from the base class `ggdModelDocumentation`. This allows specific documentation templates to be used in a more automatic way, and can be of particular importance in the context of organisations using ML, since they might have their own documentation preferences. Users can also benefit from the machinery underlying the `ggdModelDocumentation` base class to create dataset documentation (eg, à la Gebru et al. (2021)).

The template for the ForecastCard can provide an example of the type of information a documenter could either acquire automatically from reading the model object and from asking the economist; note that its language is only slightly adapted from Mitchell et al. (2018) with some fields directly reproducing a model card after those authors:

- Model details (basic information about the model)
 - Variable(s) being forecasted or nowcasted
 - Jurisdiction(s) of the variable being forecasted or nowcasted
 - Person or organisation developing the model
 - Model date
 - Model version
 - Model type
 - Description of the pipeline steps being used
 - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
 - Information about the econometric model or technique
 - Paper or other resource for more information
 - Citation details
 - License
 - Where to send questions or comments about the model
- Intended use (use cases that were envisioned during development)
 - Primary intended uses
 - Primary intended users
 - Out-of-scope use cases
- Metrics (metrics should be chosen to reflect potential real world impacts of the model)
 - Model performance measures
 - How are the evaluation metrics calculated? Include information on the cross-validation approach, if used
- Data (details on the dataset(s) used for the training and evaluation of the model)
 - Datasets
 - Preprocessing steps
 - Cut-off date that separates training from evaluation data
- Ethical considerations (considerations that went into model development, surfacing ethical challenges and solutions to stakeholders. Ethical analysis does not always lead to precise solutions, but the process of ethical contemplation is worthwhile to inform on responsible practices and next steps in future work)
 - Does the model use any sensitive data?

- What risks may be present in model usage? Try to identify the potential recipients, likelihood, and magnitude of harms. If these cannot be determined, it can be noted that they were considered but remain unknown
- Are there any known model use cases that are especially fraught?
- If possible, this section should also include any additional ethical considerations that went into model development, for example, review by an external board, or testing with a specific community.
- Any other caveats or recommendations (additional concerns that were not covered in the previous sections)
 - For example, did the results suggest any further testing? Were there any relevant groups that were not represented in the evaluation dataset?
 - Are there additional recommendations for model use? What are the ideal characteristics of an evaluation dataset for this model?

While most economists trained in traditional econometric techniques using time series quantitative data for forecasting might find some of these fields "over-kill", the objective of this documenter is to be a reasonably simple template for models that might, who knows, forecast economically variables by, say, also including textual data or some other form of non-traditional data.

5.1 Ethical issues in economics and finance for ML models

One of the reasons why gingado facilitates model documentation is to promote a greater role for ethical considerations as part of the development of ML models in economics and finance: if the model documentation becomes part of development workflow and certain parts of the documentation are automated, then users are presumably more likely to consider these issues at development time, not *ex post*. The importance of ethical considerations in finance applications of ML is underscored by the ability of ML to drive results in economics and finance: for example, Frost et al. (2019) show evidence that ML models can outperform traditional credit bureau data in predicting loan default rate in Argentina, which illustrates the strong incentives for investment in ML models by lenders. Other evidence of the real-life impact of ML applications might materially impact stakeholder outcomes is studied by Fuster et al. (2022), in the context of US mortgage lending. These authors, who also document evidence of ML outperforming traditional models, find that greater use of ML would lead to relatively higher estimated propensities of default for Black and Hispanic borrowers, even when race is not used as a feature in the models.

More broadly, as discussed by Rambachan et al. (2020), differences in predictions between groups that might be attributable to algorithmic bias can be seen as a combination of basal differences observed in society (and itself possibly the result of a societal bias), and of measurement error and estimation error differences. Of course, only the last two can be addressed by better developed ML models. Still, simply bringing the existence of this bias to light as done during the documentation process might be helpful in driving economists to make models that address this issue at least partially (Cowgill et al. (2020)). Further to that, models that are likely to result in algorithmic bias even from predominantly underlying societal causes should ideally make it clear for users that this could be the case, so that users can exercise due discretion in whether and how to deploy these models. Another strategy that could also be documented with gingado is to choose different samples of the data based

on subpopulations in which the outcome is perceived or known to be less biased, as suggested by Ludwig and Mullainathan (2021). In any case, it is plausible to expect that real-life models where these issues loom large should include documentation describing the strategy taken by model developers.

There seems to be a wide awareness of the implications from biased datasets feeding into complex, black-box ML models in economics and finance.⁹ They are illustrated for example by Brotcke (2022), who present a practitioner view on using ML while seeking to originate loans without bias, by the law enforcement examples of ML gone awry discussed in Ludwig and Mullainathan (2021) and by Doerr, Gambacorta, and Garralda (2021), who describe central banks' concerns with fairness implications from greater use of ML. In the broader language domain, Bender et al. (2021) highlight ethical and societal issues stemming from the incredibly large size of datasets used to train large language models (LLMs).

But the ethical implications of ML do not stem just from the datasets used for training. It is likely that similar issues are important in a variety of uses, in varying degrees. For example, Bender et al. (2021) also discuss model-related aspects: eg, the environmental impact of the large-size architecture of these models, and the risks originating from the higher (apparent) fluency of these models. Mullainathan and Obermeyer (2017) and Thomas and Uminsky (2022) discuss pitfalls and risks from the metrics chosen during ML development. And importantly in the case of economics, Ludwig and Mullainathan (2021) pinpoint the poor experience with ML in the judicial system to a case of faulty development of models, including due statistical consideration to the fact that modeled outcomes are in many cases only a subset of the necessary information, and that decision is taken ultimately by a human decisionmaker. In addition, properly informing the user on algorithmic design choices and recommended and unrecommended use cases might prevent situations where the model is designed to be unbiased, but its deployment is botched due to not considering that the model would work best if delivered equally to all subpopulations of interest, as shown by Lambrecht and Tucker (2019) in the case of an ML-delivered ad for careers in sciences, technology, engineering and maths (STEM) designed to be gender neutral but that was more widely seen by men due to the higher ad costs for women. Therefore, promoting opportunities for economists developing and deploying ML models to think about these implications from the complete model pipeline, from data to application to usage, seems more than warranted, in line with Cowgill and Tucker (2019).

Consideration of these and other ethical issues might be made easier by facilitating model documentation, which nudges users to explicitly document (and thus consider) features of the data and model that go beyond purely model architecture features, which are important but can be read automatically by software. In fact, I hope that the mere availability of model documentation functionalities might act as a reminder that this is an important step (Cowgill et al. (2020)). At the same time, automatically covering model information in the documentation opens up space for the user to reflect on the open human-answerable questions, many of which include issues around ethics. In other words, can model documentation address all issues in ML ethics in economics? Definitely no. But the hope is that enabling and

⁹ A related discussion where ethics and fairness in ML has made strides but needs more progress is related to explainable artificial intelligence (XAI) (for example, Barredo Arrieta et al. (2020)).

facilitating model documentation might help users take steps in that direction, both in academic as well as in practitioner settings.

6. What to expect next?

The vision for gingado is for it to become a collection of tools that can help economists deploy ML in various use cases in research or practitioner work. In this sense, development of new tools to be added in gingado is guided by two considerations. First, what are the pain points of the ML workflow for economists that can be tackled ergonomically? Second, what are the areas of ML that can benefit economics use cases? Using these considerations as backdrop, four areas of active development as of writing of this paper are:

- new canonical datasets to help new users get up to speed with initial models, fomenting the build-up of AI skillsets for beginner-level users or simply being used as benchmarks;
- clustering functionalities that automatically divide a population of entities into clusters, and retain only those that are in the same cluster as a specified entity of interest. This can be used to find (and retain only) related entities to unit(s) of interest in multidimensional settings, providing a useful selection of control units for example, for matching applications (Imbens and Wooldridge (2009)) or in the estimation of causal effects using synthetic controls (Abadie and Gardeazabal (2003), Abadie (2021)) in a more data-driven way, as proposed by Araujo et al. (2023).
- causal ML, implementing Python versions of algorithms that have been designed statistically to allow for causal inference. Examples that might be implemented include the generalised random forest of Athey, Tibshirani, and Wager (2019), the ML-based version of synthetic controls of Araujo et al. (2023), the double/debiased ML of Chernozhukov, Chetverikov, et al. (2018) and other algorithms.
- functionalities to use and adapt LLMs (eg, OpenAI (2023), Touvron et al. (2023)) in settings of economic interest.

7. Conclusion

gingado is a free, open source ML library focused on use cases in economics and finance - in both research and practice. Its compatibility with widely used ML frameworks, most notably scikit-learn, allows it to leverage on the wide familiarity with these tools and complement existing user codebases. And its flexible approach simultaneously affords users the ability to advance ML model development with few steps, while enabling users to tweak all tools to meet their goals, including adjusting models and their outcomes to better suit econometric use cases when necessary (Athey and Imbens (2019)). The toolset provided by gingado was first created for my own use as a practitioner of ML in economics, and I will continue developing it over time to incorporate functionalities that can be of most value-added to researchers and practitioners in economics and finance. Because the code is open, also the broader

public can propose improvements and even new functionalities, either by directly suggesting it or by proposing code that implements these ideas. At the same time, I hope that gingado can contribute to facilitate greater use of ML in economics and finance, while promoting good modelling practices including a greater role for model documentation and ethical considerations.

Of course, both research and empirical applications are already benefiting from more intense usage of ML, including in policy organisations (FSB (2017), Araujo et al. (2023), Frost et al. (2019), Doerr, Gambacorta, and Serena (2021))). And recent breakthroughs in the field (such as GPT-4, OpenAI (2023)) are likely to further promote this, as exemplified by Korinek (2023). In this context, gingado can help new and more experienced economists familiarise themselves with ML practice in an ergonomic way, by removing some of the complexity in getting such models set up, trained, and documented. The particular areas focused by gingado so far are data augmentation, automatic benchmark models, real and simulated datasets and model documentation.

Adding more data to one's dataset can often be cumbersome from an operational perspective, and especially so when multiple sources are involved. This is of course much harder when the model is used in a production setting, instead of a one-off analysis. gingado addresses this by augmenting the user dataset through an object that fits nicely in standard ML pipelines. This also allows the user to test whether or not adding more data actually improves the model (according to any criteria defined by the user). In addition, gingado's data augmentation method focuses on ensuring that the data provided to users is from trusted sources. When more data augmentation objects are added to the gingado codebase, the reliability of data sources will be a key criterion.

Automatic benchmark models are not new. Howard and Gugger (2020b) and Taylor and Letham (2018) for example enable users to quickly set up models with a reasonably good performance for the vast majority of use cases in their respective domains. This is also followed for tabular data by Apple's CreateML framework that underpins numerous ML applications in Apple devices. gingado builds on that insight and provides users with additional functionalities that to my knowledge are novel. Namely, it offers a way to conveniently compare candidate models (and their ensemble) and pick the best one as the new benchmark. It also offers the automatic documentation of the benchmark model, and the ability to create one's own benchmark from a base class that ensures users' customised benchmark models would be compatible with the other functionalities.

These models, of course, need to be trained on data. gingado provides the user with real and simulated datasets. The former currently contains one dataset, used by Barro and Lee (1994), but others will be included over time with the goal of forming a portfolio of academic benchmark data representing various areas of economics and finance studies. And the functionality to simulate data can be used to generate a wide range of datasets, with rich treatment chains and non-linear dependences of the outcome variable on the covariates. These can be especially useful in the context of causal ML using the potential outcomes framework (Imbens and Rubin (2015)).

And finally, there is model documentation. Mullainathan and Spiess (2017) mention the risk that ML models in economics and finance are applied naively or have their outputs misinterpreted. This risk increases with greater deployment of ML in important aspects of daily life, such as banking. But the risk probably also grows as the technical preconditions for ML, such as greater availability of higher-dimensional

datasets and of compute power, facilitate model development by more people. As the popularity of ML models amongst economists grows, my goal with gingado is to contribute a small part to embed model documentation in the process of model development, automating some of the questions to afford the humans in control the opportunity to properly document their models and pay due consideration to ethical issues arising in each particular situation, hopefully facilitating research in the field and leading to better AI models developed and deployed in practice.

References

- Abadie, Alberto (2021): "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects", *Journal of Economic Literature*, v. 59, n. 2, p. 391–425.
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program", *Journal of the American Statistical Association*, vol 105, n. 490, p. 493–505.
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2015): "Comparative Politics and the Synthetic Control Method", *American Journal of Political Science*, vol 59, n. 2, p. 495–510.
- Abadie, Alberto and Javier Gardeazabal (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country", *American Economic Review*, vol 93, n. 1, p. 113–32.
- Araujo, Douglas, Giuseppe Bruno, Juri Marcucci, Rafael Schmidt, and Bruno Tissot (2023): "Machine Learning Applications in Central Banking", *Journal of AI, Robotics and Workplace Automation*, vol 2, n. 3.
- Athey, Susan and Guido W. Imbens (2019): "Machine Learning Methods That Economists Should Know About", *Annual Review of Economics*, vol 11, n. 1, p. 685–725.
- Athey, Susan, Julie Tibshirani, and Stefan Wager (2019): "Generalized Random Forests".
- Bajari, Patrick, Zhihao Cen, Victor Chernozhukov, Manoj Manukonda, Suhas Vijaykumar, Jin Wang, Ramon Huerta, et al. (2023): "Hedonic Prices and Quality Adjusted Price Indices Powered by AI", *arXiv Preprint arXiv:2305.00044*.
- Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. (2020): "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges Toward Responsible AI", *Information Fusion*, vol 58, p. 82–115.
- Barro, Robert J., and Jong-Wha Lee. 1994. "Sources of Economic Growth." *Carnegie-Rochester Conference Series on Public Policy* 40: 1–46. [https://doi.org/https://doi.org/10.1016/0167-2231\(94\)90002-7](https://doi.org/https://doi.org/10.1016/0167-2231(94)90002-7).
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. *FACt '21*. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>.
- FSB, Financial Stability Board. 2017. Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications.
- Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24 (2): 123–40.
- Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.
- Brotcke, Liming. 2022. "Time to Assess Bias in Machine Learning Models for Credit Decisions." *Journal of Risk and Financial Management* 15 (4). <https://doi.org/10.3390/jrfm15040165>.

Chakraborty, Chiranjit, and Andreas Joseph. 2017. "Machine Learning at Central Banks."

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21 (1): C1–68. <https://doi.org/10.1111/ectj.12097>.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018. "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India." *National Bureau of Economic Research*.

Comin, Diego A, and Bart Hobijn. 2009. "The CHAT Dataset." Working Paper 15319. Working Paper Series. National Bureau of Economic Research. <https://doi.org/10.3386/w15319>.

Cowgill, Bo, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. 2020. "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics." In *Proceedings of the 21st ACM Conference on Economics and Computation*, 679–81.

Cowgill, Bo, and Catherine E Tucker. 2019. "Economics, Fairness and Algorithmic Bias." Preparation for: *Journal of Economic Perspectives*.

Doerr, Sebastian, Leonardo Gambacorta, and José Maria Serena Garralda. 2021. "Big Data and Machine Learning in Central Banking." *BIS Working Papers*, no. 930.

Doerr, Sebastian, Leonardo Gambacorta, and Jose Maria Serena. 2021. "How Do Central Banks Use Big Data and Machine Learning." In *The European Money and Finance Forum*, 37:1–6.

Duarte, Victor. 2018. "Machine Learning for Continuous-Time Economics." Available at SSRN 3012602.

Fernández-Villaverde, Jesús, Samuel Hurtado, and Galo Nuno. 2019. "Financial Frictions and the Wealth Distribution." *National Bureau of Economic Research*.

Ferreira, Leonardo N et al. 2021. "Forecasting with VAR-teXt and DFM-teXt Models: Exploring the Predictive Power of Central Bank Communication."

Frost, Jon, Leonardo Gambacorta, Yi Huang, Hyun Song Shin, and Pablo Zbinden. 2019. "BigTech and the Changing Structure of Financial Intermediation." *Economic Policy* 34 (100): 761–99.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." *The Journal of Finance* 77 (1): 5–47. <https://doi.org/https://doi.org/10.1111/jofi.13090>.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–74.

Géron, Aurélien. 2019. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd Edition. Sebastopol, CA: O'Reilly Media.

- Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri. 2021. "Economic Predictions with Big Data: The Illusion of Sparsity." *Econometrica* 89 (5): 2409–37.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
- Gorishniy, Yury, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. "Revisiting Deep Learning Models for Tabular Data." <https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html>.
- Howard, Jeremy, and Sylvain Gugger. 2020a. *Deep Learning for Coders with Fastai and Pytorch: AI Applications Without a PhD*. O'Reilly Media, Incorporated.
- Howard, Jeremy, and Sylvain Gugger. 2020b. "Fastai: A Layered API for Deep Learning." *Information* 11 (2). <https://doi.org/10.3390/info11020108>.
- Hünermund, Paul, Beyers Louw, and Itamar Caspi. 2023. *Journal of Causal Inference* 11 (1): 20220078. <https://doi.org/doi:10.1515/jci-2022-0078>.
- Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86. <https://doi.org/10.1257/jel.47.1.5>.
- Jordà, Òscar, Katharina Knoll, Dmitry Kuvshinov, Moritz Schularick, and Alan M Taylor. 2019. "The Rate of Return on Everything, 1870–2015." *The Quarterly Journal of Economics* 134 (3): 1225–98.
- Jordà, Òscar, Björn Richter, Moritz Schularick, and Alan M Taylor. 2021. "Bank Capital Redux: Solvency, Liquidity, and Crisis." *The Review of Economic Studies* 88 (1): 260–86.
- Jordà, Òscar, Moritz Schularick, and Alan M Taylor. 2017. "Macrofinancial History and the New Business Cycle Facts." *NBER Macroeconomics Annual* 31 (1): 213–63.
- Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95. <https://doi.org/10.1257/aer.p20151023>.
- Kohlscheen, Emanuel. 2021. "What Does Machine Learning Say about the Drivers of Inflation?" Available at SSRN 3949352.
- Korinek, Anton. 2023. "Language Models and Cognitive Automation for Economic Research." Working Paper 30957. Working Paper Series. National Bureau of Economic Research. <https://doi.org/10.3386/w30957>.
- Lambrecht, Anja, and Catherine Tucker. 2019. "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads." *Management Science* 65 (7): 2966–81. <https://doi.org/10.1287/mnsc.2018.3093>.
- Li, Kai, Feng Mai, Rui Shen, and Xinyan Yan. 2020. "Measuring Corporate Culture Using Machine Learning." *The Review of Financial Studies* 34 (7): 3265–315. <https://doi.org/10.1093/rfs/hhaa079>.
- Ludwig, Jens, and Sendhil Mullainathan. 2021. "Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System." *Journal of Economic Perspectives* 35 (4): 71–96. <https://doi.org/10.1257/jep.35.4.71>.

- Maliar, Lilia, Serguei Maliar, and Pablo Winant. 2021. "Deep Learning for Solving Dynamic Economic Models." *Journal of Monetary Economics* 122: 76–101.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. "Model Cards for Model Reporting." CoRR abs/1810.03993. <http://arxiv.org/abs/1810.03993>.
- Mullainathan, Sendhil, and Ziad Obermeyer. 2017. "Does Machine Learning Automate Moral Hazard and Error?" *American Economic Review* 107 (5): 476–80. <https://doi.org/10.1257/aer.p20171084>.
- Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. <https://doi.org/10.1257/jep.31.2.87>.
- OpenAI. 2023. "GPT-4 Technical Report." <https://arxiv.org/abs/2303.08774>.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.
- Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2020. "An Economic Perspective on Algorithmic Fairness." *AEA Papers and Proceedings* 110 (May): 91–95. <https://doi.org/10.1257/pandp.20201036>.
- Rossi, Barbara. 2013. "Exchange Rate Predictability." *Journal of Economic Literature* 51 (4): 1063–1119.
- Stierholz, Katrina. 2014. "FRED®, the St. Louis Fed's force of data." *Review* 96 (2): 195–98. <https://ideas.repec.org/a/fip/fedlrv/00023.html>.
- Taylor, Sean J, and Benjamin Letham. 2018. "Forecasting at Scale." *The American Statistician* 72 (1): 37–45.
- Thomas, Rachel L, and David Uminsky. 2022. "Reliance on Metrics Is a Fundamental Challenge for AI." *Patterns* 3 (5): 100476.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv Preprint arXiv:2307.09288.
- Ushey, Kevin, JJ Allaire, and Yuan Tang. 2022. Reticulate: Interface to 'Python'.
- Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28.
- Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. "Using Publicly Available Satellite Imagery and Deep Learning to Understand Economic Well-Being in Africa." *Nature* 11. <https://doi.org/https://doi.org/10.1038/s41467-020-16185-w>.

gingado: a machine learning 🤖 library
focused on economics and finance
/ʒĩ.'ga.du/

```
if __name__ == "__main__":  
    print("Author: Douglas Araujo")  
    print("Date: 15 Feb 2022")
```


What is *gingado*?

- Open-source machine learning library written in Python (under development 🛠️)
 - Objective: broaden the accessibility of state-of-the-art models to a wide range of practitioners in economics and finance
 - Main features:
 1. Automatic benchmark models
 2. Automatic data augmentation
 3. Automatic documentation
- ... all of that with a simple API

Automatic benchmark models

- Once the model and the dataset are defined, *gingado* automatically trains a benchmark model (unless asked not to train it)
- Benchmark models are useful to compare results from user attempts
- Worse case scenario, if all your attempts fail at beating the benchmark, at least you have a reasonable model

Automatic data augmentation

- “Data augmentation” means to append more data to your dataset. This is known to generally improve the performance of ML models.
- For example, in ML models working with image, data augmentation involves flipping, cutting, zooming in or out, etc.
- In economic/financial datasets, augmentation involves adding other information related to the observations in the dataset
- For example, if the dataset is a country-level panel data, data augmentation would add a lot of publicly-available data for the countries in the dataset
- New data sourced using  **sdmx** and other APIs from official sources

Automatic data augmentation

Ensuring the data augmentation works

1. user defines the variables of interest for augmentation: time + geographies
2. *gingado* fetches data from BIS, IMF, ECB, Eurostat, & other sources
3. a new benchmark model is run with *all* data and it is compared to the original
4. all of the augmented data is kept if performance improves
5. if performance does not improve, only the variables that have low correlation to all of the variables in the origination dataset are kept



Automatic model documentation

- Model card inspired by Mitchell et al (2019)
- “Meta-data” about the model
- Transparency about:
 - envisage use contexts
 - how the model was evaluated, etc
- Important in model development, management, audits
- *gingado* uses JSON to store the raw model card info

Model Card

- **Model Details.** Basic information about the model.
 - Person or organization developing model
 - Model date
 - Model version
 - Model type
 - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
 - Paper or other resource for more information
 - Citation details
 - License
 - Where to send questions or comments about the model
- **Intended Use.** Use cases that were envisioned during development.
 - Primary intended uses
 - Primary intended users
 - Out-of-scope use cases
- **Factors.** Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
 - Relevant factors
 - Evaluation factors
- **Metrics.** Metrics should be chosen to reflect potential real-world impacts of the model.
 - Model performance measures
 - Decision thresholds
 - Variation approaches
- **Evaluation Data.** Details on the dataset(s) used for the quantitative analyses in the card.
 - Datasets
 - Motivation
 - Preprocessing
- **Training Data.** May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
 - Unitary results
 - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Backend

- Backend based on *fast.ai* library 
- *gingado* inherits the following characteristics from *fast.ai*:
 - simple yet hackable
 - flexibility to include non-numeric datasets (texts, etc)
 - production-ready results
 - excellent community support for any *fast.ai*-specific issues
- Extensive use of  as well

If you are interested...

Please use it and share your experience!

- Open up an “issue” on GitHub if you experience any bug
- You may also open up “issues” for feature requests (please be as specific as possible)
- Contributors are welcome! Please get in touch before opening a pull request

<https://github.com/dkgaraujo/gingado/>

Thank you for your attention!

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

A multi-layer dynamic network for significant European banking groups¹

Annalaura Ianiro and Joerg Reddig,
European Central Bank

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

A multi-layer dynamic network for significant European banking groups

Ianiro, Annalaura; Reddig, Jörg¹

Abstract

In the aftermath of the financial crisis there was a clear demand for granular statistical data which would allow user to gain insights into topics ranging from financial stability, monetary policy, banking supervision and diverse research questions which could not be answered before. The statistical function of the Eurosystem answered this demand by collecting new and highly granular data sets on securities, money market transactions, derivative trades, and loans which are accompanied by a collection of master data on institutions allowing interlinking this information. In this work, we first show how we integrate these data to build a multi-layered network of the significant European banking groups. Then, we proceed to analyse the data with a new Python toolkit, NATkit, specifically developed to work with multi-layer dynamic networks.

Keywords: Multi-layer dynamic network, significant banking groups, granular data

JEL classification: G21

The views expressed are those of the authors only and do not necessarily reflect those of the European Central Bank.

¹ Annalaura Ianiro, Annalaura.Ianiro@ecb.europa.eu, European Central Bank

Jörg Reddig, Joerg.Reddig@ecb.europa.eu, European Central Bank

Contents

| | |
|--|----------|
| A multi-layer dynamic network for significant European banking groups | 1 |
| 1. Introduction..... | 3 |
| 2. Related Work – The Approach for Multi-Layer Networks..... | 4 |
| 3. The datasets..... | 6 |
| 3.1. List of significant banking groups (ROSSI list) | 6 |
| 3.2. Group structure of the significant banking groups (RIAD) | 7 |
| 3.3. Securities layer (CSDB / SHSG)..... | 7 |
| 3.4. Loans layer (AnaCredit) | 8 |
| 3.5. Money market transactions data (MMSR) | 8 |
| 3.6. Derivative data layer (EMIR) | 9 |
| 3.7. Other datasets used..... | 9 |
| 4. Building the multi-layer dynamic network..... | 10 |
| 4.1. Retrieving the group structure of the significant institutions..... | 10 |
| 4.2 In depth example: building the securities cross-holdings layer..... | 11 |
| 4.3. The infrastructure..... | 12 |
| 5. Analysing the multi-layer dynamic network..... | 13 |
| 5.1 NATkit: An Overview..... | 13 |
| 5.1.1. Visualization module | 13 |
| 5.1.2. Topology module..... | 13 |
| 5.2. Using NATkit to analyse the multi-layer dynamic network of significant institutions | 13 |
| 5.2.1. The Layers | 13 |
| 5.2.2. Building the multiplex layers | 14 |
| 5.2.3. Visualizing the network | 16 |
| 5.2.4. The multiplex topology..... | 19 |
| 5.2.6. Identifying non-linearities..... | 32 |
| 6. Conclusions..... | 35 |
| References..... | 37 |

1. Introduction

After the global financial crisis, among the many lessons learned, policymakers got reminded that for non-linear events driven by financial instability a timely analysis of interconnected granular data can help to better understand the developments in the financial sector at the level of individual agents and their interactions in various markets, helping to prevent or contain financial crises. Consequently, the European Central Bank (ECB) invested in appropriate modelling tools and started to purposely collect more granular data, in order to support the creation of multipurpose interconnected data sets that allow the insights that have been missing before. Given the way they were designed and conceived, these datasets have the potential to be a game changer for understanding the landscape and interconnectedness of the euro area financial markets. It is, thus, important to analyse interconnected markets in a joint manner. Only then, we might be able to uncover meaningful insights –i.e. exposure of financial institutions to certain economics sector, concentration risk, substitution effects between asset classes, etc. – and the potential propagation channels of shocks.

To do so in a systematic and coherent way, granular datasets need to be integrated with each other and be analysed jointly. Indeed, as former ECB President Mario Draghi stated, “developing the analytical toolkit to adequately monitor interconnectedness and contagion requires granular datasets, and the ability to map and link data across entities and markets” (Draghi, 2019).

The ECB’s Data Committee sponsored a project to develop tools that would allow users from various business areas and policy makers to apply scalable and reproducible advanced analytics on the interconnections of institutions and financial markets via the use of integrated granular data. These would include loan-by-loan and security-by-security (issues and holdings) data, money market transactions, and over-the-counter derivatives transactions. Such data could also be combined with supervisory information on banks like COREP, FINREP and other supervisory data, monetary policy data on MFI balance sheets and interest rates, monetary policy operations as well as commercial datasets like Orbis/Bankfocus on firms and financial institutions. Advanced analytics based on the resulting integrated granular dataset would allow supplementing the standard models based on macroeconomics variables, in particular by being able to better capture non-linear contagion effects that in particular stem from the interconnectedness of institutions in different financial markets.

Advanced, scalable and reproducible analytics not only need common data, but also common analytical tools that can be easy to use, well document and extensible. The integrated granular dataset mentioned above can be seen as a multi-layer dynamic network, in other words a combination of multiple individual networks (the layers), that change over time (dynamic). The available tools for network analysis lack the functionalities needed to deal with this type of networks, forcing analysts either to come up with custom solutions (which are very specific and difficult to scale) or to analyse the network as if not multi-layer nor dynamic (which loses important information). NATkit (Network Analytics Toolkit) is an analytical toolkit that was developed (as a Python library) to fill in this gap and analyse multi-layer dynamic networks with the proper framework. Its purpose is to provide an easy-to-use toolset, which could help to explore and understand complex multi-layer dynamic network data. The toolkit leverages on existing visualisation and network analytics libraries,

building upon them the framework and functions to properly handle this type of networks.

This work is a showcase of the tool's functionalities on a sample of the multi-layer network of significant institutions of the euro area, built combining granular data loans, securities, derivative and money market transactions. It is structured as follows: section 2 offers a brief overview of the literature on networks analytics for multi-layer networks, with specific focus on financial networks. It explains why multi-layer networks need a specific analytical approach and provides a formal definition of what a multi-layer network is. Section 3 describes the dataset that are used for building the individual layers of the network while section 4 describes how exactly the data were integrated and the individual layers were constructed with an example of the securities layer. Information about the toolkit and its empirical application to the multi-layer network are presented in section 5. The final section concludes and gives an outlook on the future work.

2. Related Work – The Approach for Multi-Layer Networks

Many complex real-world phenomena can be represented as a network. Mathematically, a network is a *graph*, in other words a collection of vertices (nodes) connected among each other by lines (edges), making the study of complex networks the domain of graph theory.

In the beginning, complex networks with no apparent structure were described with *random graphs*. Random graphs, also known as *Erdős-Rényi graphs* from the names of the two mathematicians who first theorised them in 1950, have a fixed set of nodes, N , while the existence of an edge between any two nodes is equal to a probability p . The beauty of this model lies in its simplicity and nice properties; however, its ability to properly represent complex network systems has been since reconsidered: the topology of many real-world networks can hardly be described as random. Therefore, many tools and measures have been developed to properly describe the topologies of real-world graphs. Réka and Barabási (2002) provide an interesting overview on the statistical mechanics of complex networks, still a relevant read despite its age.

The specific literature of financial applications of network analysis began with the seminal contributions by Allen and Gale (2000) and Freixas et al. (2000) that recognised the importance of the structure of interconnections between financial institutions in a theoretical framework. From these studies, a rich literature has developed, defining the canons and standards of network analysis for complex financial and interbank networks. Such standards, as argued by Battiston et al. (2014) and Kurant and Thiran (2006), often ignore whether a network is multi-layer or not and can be divided into two types. One type studies the graph resulting from the aggregation of all the edges between a certain set of nodes, regardless of the type of relationship they represent. The other type treats the different layers of the same network as separate entities. Hence, despite their much more complex reality, for a long time interbank and financial networks have either been compressed into a single aggregate layer or different layers have been studied in isolation. Both these approaches overlook important information and only paint an incomplete picture about the structure and functioning of financial relationships among market players.

Representing the network as a *multi-layer graph* or *multiplex* allows to overcome the limitations of the standard approach. When analysing the financial and thus interbank networks, in particular, the literature has highlighted at least three advantages of the multi-layer approach:

- It accounts for the heterogeneity of each layer. Bargigli et al (2015) finds that several topological and metric properties are layer-specific, whereas other properties are of a more universal nature.
- It allows to capture non-linearity, which is important for risk assessment. Poledna et al. (2015) show the existence of non-linearity in the way risks are aggregated: in their application, systemic risk for the aggregated network may be considerably larger than for the sum of the component sub-networks.
- It can provide relevant metrics for banking regulators and supervisors. Aldasoro and Alves (2015) propose a measure of systemic importance that allows to decompose the global systemic importance index for any bank into the contributions of each of the sub-layers. Such a measure can lead to better tailored policy responses.

Having briefly highlighted the advantages of the multi-layer approach, we will explain what a multi-layer network, or multiplex, is, as well as extend the definition to include the dynamic component. While Kivela et al (2014) provide a very extensive formalisation of multi-layer networks, for the purpose of this work we will consider the definition given in Battiston et al (2014), which is more intuitive. According to them, a multiplex is “a network where each node appears in a set of different layers, and each layer describes all the edges of a given type”. A dynamic multiplex is, therefore, the set of the time-specific multiplexes over the given time interval.

More formally, we could think of a network in terms of its *adjacency matrix*. Focusing on a single time period t , an adjacency matrix is a square $n \times n$ matrix, where n is the number of nodes in the network. Therefore, an entry e_{ij} represents an edge connecting node i to node j . When the network is unweighted, such entry is equal to 1 if an edge exists, 0 otherwise. In case of a weighted network, e_{ij} is equal to w_{ij} (the weight value) if there exists an edge, 0 otherwise. Another feature of the network is whether it is directed or undirected². If the network is undirected, the adjacency matrix is symmetric, so that $e_{ij} = e_{ji}$. Each layer of the multi-layer network has its own adjacency matrix. Overlapping them (summing their entries together point by point) allows to obtain the *multiplex adjacency matrices*. Indeed, there are two types (three in case of weighted networks) of multiplex adjacency matrices for each multi-layer network. These are:

- A : aggregated multiplex adjacency matrix. This matrix has entry a_{ij} equal to 1 if there exists an edge in at least one of the layers, 0 otherwise.
- O : overlapping multiplex adjacency matrix. This matrix is built by overlapping the unweighted adjacency matrices of all the layers. Therefore, it has entry

² Financial networks are usually directed. For example, in case of loan exposures the two counterparties have distinct roles: one is the lender, the other is the borrower. An undirected network is, for example, friendship on Facebook: if person A is friend of person B, the opposite is always true. When, instead, we look at the network of Twitter following, which is directed, if person A follows person B, this doesn't imply that person B follows person A.

$o_{ij} = \sum e_{ij}$. In case there is no edge between nodes i and j in any of the layers, the entry is 0.

- O^w : weighted overlapping multiplex adjacency matrix (only for weighted networks). This matrix is built by overlapping the weighted adjacency matrices of all the layers. Therefore, it has entry $o_{ij}^w = \sum w_{ij}$. In case there is no weighted edge between nodes i and j in any of the layers, the entry is 0.

A dynamic multi-layer network is represented by the set of the above matrices for each specific time period t in consideration.

Each of the matrices can be viewed and analysed as its own network. However, in order to properly account for the non-linear relationships that might exist among the different layers, it is important to consider both the multiplex representations and the separate layers in the analysis. Section 5 further elaborate and explain this concept.

3. The datasets

Having defined what a multi-layer dynamic network is, we now turn to describing which granular statistical data are available at the ECB that we used for building the multi-layer network. All these datasets are stored on the ECB's Data Intelligence and Service Center (DISC) platform. This is the ECB's Hadoop-based datalake that allows to query the data via SQL or use them directly on an Apache Spark cluster.

3.1. List of significant banking groups (ROSSI list)

Starting point for the creation of our multi-layer dynamic network for banking groups is to define which entities shall be included in the sample. As almost all available granular datasets comprise information on the largest banking groups in the Euro Area and participating member states which are directly supervised by the Single Supervisory Mechanism (SSM), we decided to use this list as the sample for our network.

The ECB maintains a list of all significant banks under its direct supervision and less significant banks under its indirect supervision, which is publicly available on the SSM homepage.³ As of 1 November 2021, this list of significant institutions comprised a total of 115 banking groups. However, the list is reviewed every year and ad-hoc assessments are carried out during the year whenever necessary to assess the significance status of banking groups based on a size criterion which considers the total value of the supervised entity's or the supervised group's assets, at consolidated level.⁴ Therefore, the number of supervised institutions changes over time.

Internally, this list is made available via the Repository of the SSM Supervised Institutions (ROSSI). From this list of banking groups, we extract the names of the significant head institutions under direct SSM supervision as well as their Legal Entity Identifiers (LEI), which is a unique global identifier for legal entities participating in

³ See also <https://www.bankingsupervision.europa.eu/banking/list/html/index.en.html>.

⁴ A detailed description on the assessment of the significance status can be found here.

financial transactions. These identifiers are used to connect the supervisory entities with the statistical datasets.

3.2. Group structure of the significant banking groups (RIAD)

Once we have identified the list of significant institutions relevant for our network, the next step is to get data on the group structures of these banking groups. This way, we will get a view on all entities that belong to the group, which will be the foundation of the construction of the network.

This information can be retrieved from the Register of institutions and affiliates database (RIAD), the Eurosystem's unique master data repository. It not only contains detailed information on financial institutions but also on their group structures. In fact, four types of group structures are available for different purposes. In the analysis here, we use the RIAD group structure type A, which is based on all *direct and indirect control relationships* in a group. RIAD data are constantly updated and enriched. For the purpose of this analysis, we use monthly snapshots on the end-of-month values for the group structures to match the other datasets which we use to construct the different layers of our network.

3.3. Securities layer (CSDB / SHSG)

For the securities layer, we employ two granular datasets: data on the issuance activities of the banking groups, which we can retrieve from the Centralised Securities Database (CSDB), and the holdings of securities, which are available in the Securities Holdings Statistics Database by Group dataset (SHSG).

The CSDB is a reference database for securities which aims to cover all securities relevant for statistical purposes of the European System of Central Banks. It comprises information on debt securities, equity instruments and investment fund shares which are stored on a security-by-security basis where each instrument is identifiable by the International Securities Identification Number (ISIN). For each of these securities a vast number of attributes are available. For the purposes of building a securities layer in the multilayer network the CSDB gives us important information on the issuers, like issuer name, identification number, and sector classification and on the securities, like instrument type, prices, total issuance, original and remaining maturity, and ratings. The data in the CSDB are available on the ECB's data lake in monthly frequency and cover the periods from starting from 2014.

Sources for this database are commercial data which are bought from several providers as well as input from the Eurosystem central banks. Together, the ECB and the National Central Banks do a constant quality assessment on the data. This guarantees the necessary coverage of securities information used in various statistics of the Eurosystem and a high-quality standard to make the data fit for purpose.⁵

While the CSDB allows us a very good view on the issuance of securities by the significant banking groups, the SHSG dataset provides the reported holdings of securities by the banks. As of 2018-Q3 the significant banking groups must report their holdings of debt securities and equity instruments (including investment funds

⁵ For more information on the CSDB, please see *The "Centralised Securities Database" in brief* (European Central Bank, 2010).

shares) on an entity-by-entity basis. This means, that all securities holdings must be broken down by the entities of the group that are holding these securities on a global level. Thus, we get a complete overview which part of the banking group is holding the securities, inside and outside the euro area.

Like CSDB data, the amounts of securities held by the significant institutions are reported on a securities-by-securities level and can be identified by the ISIN code. This allows us as well to merge the two granular securities datasets and construct the securities layer of the multilayer network, where the so-called cross holdings between the banking groups form the edges of the network.

3.4. Loans layer (AnaCredit)

The layer describing the loan exposures between the significant banking groups is based on the AnaCredit dataset. The dataset contains loan-by-loan information collected from euro area banks extended to corporations. These data are reported at monthly frequency and are available as of September 2018. They comprise more than 80 attributes which give a detailed picture of the nature of these loans. Among others, information on the outstanding amount, maturity, interest rate, collateral/guarantee, and on counterparties are collected for each of the individual loans. The data are collected for each entity of a banking group separately (unconsolidated), which needs to be considered for the calculation of the total loans of a banking group.

This makes AnaCredit an incredibly rich dataset that can be used for various analytical purposes. However, as more than 25 million credit instruments are recorded every month and the large number of attributes, this dataset is also computationally very resource demanding.

3.5. Money market transactions data (MMSR)

The money market statistical reporting (MMSR) data are collected on transaction-by-transaction basis from a sample of euro area reporting agents. This way, it covers the most relevant institutions which are active on the European money market. It provides information on the secured, unsecured, foreign exchange swap and overnight index swap euro money market segments. The ECB uses this information to calculate euro short-term rate (€STR).

For the money market layer, we therefore have two important differences: Firstly, the data collection only comprises a sample of the money market participants. This means that the network does not cover all significant institutions but only a subset of them. An additional complication is that often trades between two market participants are conducted via a central counter party. In those cases, we employ a matching procedure to identify the indirect trading partners. Secondly, the MMSR dataset focuses on daily short-term financial transactions between the money market participant. In contrast, the securities and loan layers comprise end-of-period stock data that are reflected in the balance sheets of the reporting institutions.

The granular MMSR trade data include amongst other attributes the interest rate for the transaction, the volume and counterparty information as well as collateral type information. Moreover, both sides of a transaction (lending and borrowing) are reported by the parties involved. In our analysis of the MMSR data we focus on the secured money market segment as this is the most active part of the market. Analysing these data can bring very valuable information on the structure of the

money market. Moreover, it helps to identify liquidity shortages on the market and thus helps to monitor the risks for market participants.

3.6. Derivative data layer (EMIR)

European Market Infrastructure Regulation (EMIR) dataset contains a vast amount of data on derivatives markets and is collected via authorised trade repositories by the European Securities and Markets Authority (ESMA). The EMIR regulation covers transactional-level derivatives data for all counterparties established in the euro area as well as all contracts where at least one entity is located within the euro area or where the reference obligation is sovereign debt of a euro area member.

The data are reported at daily frequency and cover 129 data attributes which are reported for each contract. On the one hand, these attributes cover information on the counterparties involved and on the other information on the characteristics of the contract (for example the type of derivative, the underlying, outstanding amounts and prices), as well as details about how the contracts were executed. This wealth of information makes it one of the largest datasets currently available at the ECB.

3.7. Other datasets used

Apart from these granular and master data, we also used some bank balance sheet information from the supervisory FINREP data and the statistical Balance sheet items statistics. We mostly used data on the total balance sheets of the banking groups to enrich the information on the nodes in the network.

An overview of the used granular information is shown in table 1, which shows as of when data are available on the ECB's datalake as well as the frequency that these datasets have.

Granular datasets overview

Table 1

| | Time frame | Frequency |
|-----------|------------|----------------|
| Datasets | | |
| AnaCredit | 2018-09 | Monthly data |
| EMIR | 2014-02 | Daily data |
| MMSR | 2018-01 | Daily data |
| CSDB | 2014-01 | Monthly |
| SHSG | 2018-Q3 | Quarterly data |

Sources: ECB

4. Building the multi-layer dynamic network

The integration framework described in this section is the core building block of the multi-layer dynamic network of the significant institutions of the euro area. We would like to highlight that this framework is only one of the possible ways of integrating the datasets; in particular, it is an integration as end-users of each datasets, rather than an integration at the source. Trying to merge datasets that were not conceived for that implied considerable challenges, which we overcame to the best of our abilities. Despite its limitations, we hope that this work can become the basis of future analysis proving the value that integrating granular datasets have for an institution like the ECB, thus paving the way for data collection policies that envisage integration at the source.

One final remark: even if not explicitly written to keep the explanation easier to read, it's important to clarify that each dataset is joined for each reference period considered (i.e. the reference date filed of each dataset is used as a secondary key for each join).

4.1. Retrieving the group structure of the significant institutions

The construction of each layer of the multi-layer dynamic network starts with the integration of all the datasets together. We begin with the list of significant institutions that we got from the ROSSI database. As described in the previous chapter, this list is maintained by the supervisory function of the ECB. It includes two types of identifiers that can help to uniquely identify the parent company of the SSM supervised institutions: the LEI and the ESCB-internal RIAD code. While the LEI is an internationally agree-upon standard, it can happen that some entities do not have such a code or that there are doubts whether the correct code is stored in the database. Therefore, we choose the RIAD code, which is universally used for statistical dataset collected by the ECB, as our entity identifier to connect the datasets with each other.

Having the identifiers of the significant institutions, we proceed to retrieve the list of all the subsidiaries belonging to a specific banking group. To achieve this, we join the ROSSI list with the RIAD database to retrieve the group structure of each significant institution. As previously explained in section 3, RIAD provides four different types of group structures. We choose RIAD group structure type A, which includes all directly and indirectly control entities in a group⁶. Given our focus on the core banking group, we apply a filter to the join, so that only credit institutions in the group are taken into account. The group structure thus obtained will allow us to get an overview of the activities of all group members in each layer of the network. In the following subsection, we provide an in-depth example on how, starting from the group structure, we are able to get the cross-holdings information of each banking group (i.e. who is issuing a specific security and who is holding it). We apply a similar methodology to the other datasets to build the layers of the network for loan data, derivatives data and the money market layer, but for brevity's sake, we will not describe them. It is important to highlight, though, that the different layers are constructed in a consistent way, as we start from the same group structure information.

⁶ Specifically, it considers ownership relationships with an equity share greater than 50% and/or existence of control.

4.2 In depth example: building the securities cross-holdings layer

The first building block of the securities cross-holdings layer is retrieving the total issuance of each credit institution in each banking group. In practice, this implies joining all RIAD codes⁷ identifying the group members with the CSDB, the master database for securities. This gives us a comprehensive list of securities issued by all the members of the banking group, where each security can be uniquely identified by its ISIN code. With it, it is also possible to retrieve attributes on the securities that we need for further analysis of the securities layer. For example, the CSDB contains information on total amount issued for each security, their price, the date of issuance, the original and remaining maturity, and the classification of the securities into debt or equity instruments. For the purpose of this work, we treat the securities layer one as one layer, aggregating all instruments at parent entity level. However, depending on the analysis, it could make sense to define additional sub-layers, e.g. the available attributes let us distinguish between short-term and long-term debt securities layers or between equity and debt layers.

The next step is looking at the holding side and retrieve the total holdings of each credit institution in each banking group. We collect this information by joining SHSG with the group structure, through the RIAD codes. This gives us a comprehensive list of securities held by all the members of the banking group, where each security can be uniquely identified by its ISIN code, as it is the case for the issuance side. Lastly, we need to combine the issuance and holding side, using the ISIN code as the unique identifier and main key for the join.

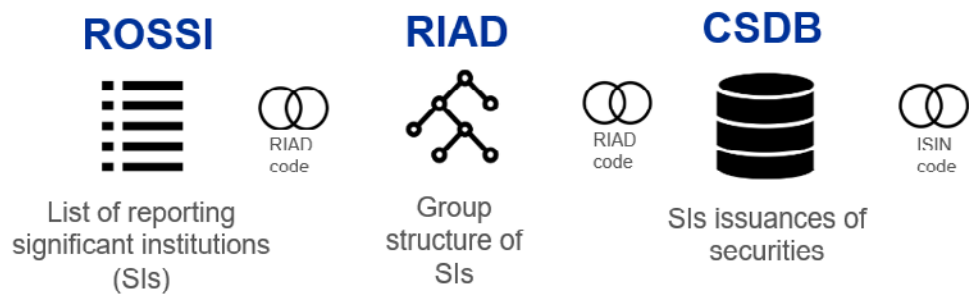


Figure 1: Stylised joins used to create the securities layer of the network

Figure 1 shows a stylised representation of the different joins that are used to create the securities layer for the multi-layer network. As we explained, there are four different datasets involved and two main different identifiers to merge the data. As a result, we get a comprehensive overview of the cross-holdings of securities among the significant institutions supervised by the SSM. Figure 2 shows a representation of this layer through a chord diagram. Thanks to the granular nature of the data used in this exercise and the rich set of attributes available, we are able, among others, to see from which countries the significant institutions are from (represented with the different colours in the diagram). Furthermore, other attributes can be used in a

flexible way depending on the question that researchers want to tackle with the network.

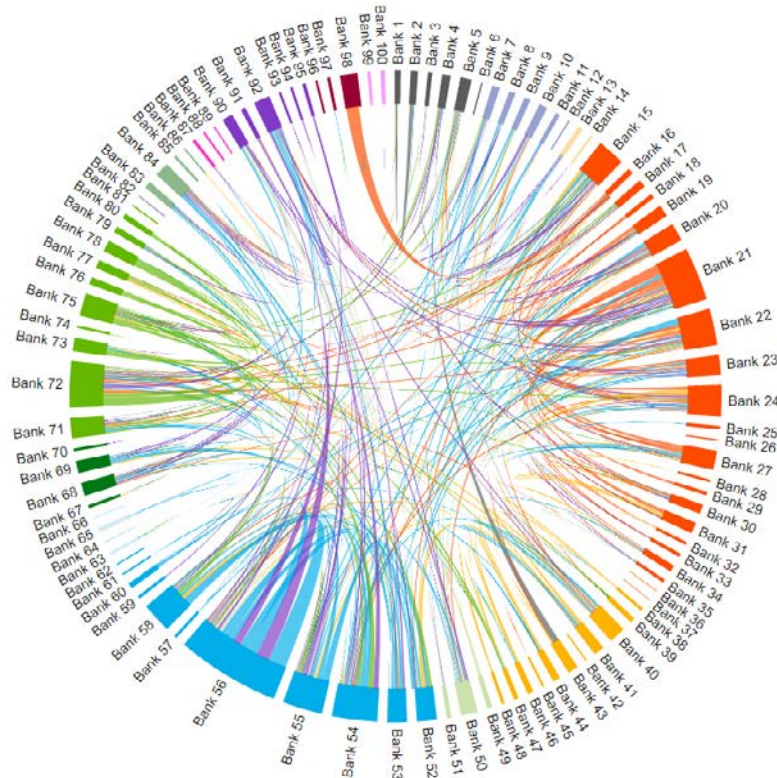


Figure 2: Chord diagram of the securities cross-holdings layer for period 2020-06.

4.3. The infrastructure

The integration framework is a set of SQL and PySpark scripts and their accompanying documentation, available to all colleagues at the ECB. Each script is built in blocks, prioritising flexibility and reusability. This allows us to easily update the code when necessary, as well as quickly adapt it to integrate new layers, as they become available. It is also possible to modify some of its parts to adapt it to different purposes (e.g. one might want to look at all entities belonging to a certain banking group, instead of just at the credit institutions).

Running the scripts allows the users to retrieve the integrated data, which can then be analysed with the preferred language (e.g. R, Python, MATLAB, etc.).

5. Analysing the multi-layer dynamic network

5.1 NATkit: An Overview

NATkit is a Python tool developed to analyse multi-layer dynamic networks. The existing standard Python libraries for network analyses, such as Networkx, lack the functions to deal with a multi-layered graph object; therefore, NATkit aims to fill this gap, providing a tool built specifically for the type of financial data analysed at the ECB. The first prototype of the tool has working versions of the two core modules, visualisation and topology.

5.1.1. Visualization module

This core module allows the users to visualise the network in an interactive way. Combining different Python visualization libraries, it provides easy-to-use functions to plot complex graph objects. The interactivity allows the users to interact with the plot itself, for example to get information on mouse-hoover, as well as to see the evolution of the network over time. As an extra feature, users have the possibility to use an edge bundling algorithm to better disentangle the edges in the graph and obtain a clearer view of the network.

5.1.2. Topology module

This core module enables the users to analyse a multi-layer network in the proper framework, thanks to a combination of visualizations and topology metrics, which have been updated from the traditional ones to account for the multiplex nature of the graph object. In particular, while some topology metrics can be computed on the aggregated and overlapping representations of the network, some others have to rely on a combination of the different individual layers.

5.2. Using NATkit to analyse the multi-layer dynamic network of significant institutions

NATkit can help researchers to deep-dive into complex network structures and explain the interactions between the different layers. The purpose of the remainder of this paper is to demonstrate on real data how to use NATkit in order to analyse a dynamic multi-layer network. Therefore, the following analysis is going to be atypical, in the sense that it is not driven by a specific research question defined by business users, but is rather motivated by its ability to showcase the capabilities of the toolkit. In its development, priority has been given to functionalities that can allow the users to perform an analysis of multi-layer and dynamic networks as comprehensive as possible.

5.2.1. The Layers

Recalling section 4, the integration of the different datasets has resulted into the following layers:

- **Securities cross-holdings layer:** this layer is the combination of ROSSI, RIAD, CSDB and SHSG. It contains information about the securities holdings of the significant institutions of the euro area.

- **Loan exposures layer:** this layer integrates data from ROSSI, RIAD and AnaCredit. It contains information of the loan exposures of euro area significant institutions towards each other.
- **Secured money market transactions layer:** this layer is obtained by joining ROSSI, RIAD and MMSR. It contains information on the secured money market transactions of the significant institutions of the euro area.
- **Interest rate derivative exposures layer:** this layer is built by joining ROSSI, RIAD and EMIR. It contains information on the interest rate derivative exposures of the significant institutions of the euro area towards each other.

All layers have information aggregated at parent company level and have been enriched with total asset information coming from FINREP. For the purpose of this analysis, a sample of the above data has been extracted, covering monthly snapshots⁸ of the sub-layers from December 2019 to May 2020.

5.2.2. Building the multiplex layers

Building a multi-layer network can be a rather complicated endeavour since the very beginning, even in choosing which layers to use. We decided to use the layers aggregated as described in the previous sub-section; however, each of them has its own sub-layers⁹, which could be aggregated into different multiplex layers. Therefore, for future analysis, it could be worth to discuss in-depth which sub-layers to consider and whether to include sub-sub-layers.

Of course, having the data for each single layer is not enough to build the multiplex layers. Some data reshaping is necessary and the toolkit provides a framework and related functions to do so.

After using the integration framework as described in section 4, each layer comes in a table form as an edge-list¹⁰, with different naming conventions and fields. In order to build the multiplex layers, the different layers need to share the same structure and the same attributes. This means identifying:

- **Source nodes and target nodes:** given that each layer is a directed network, this first identification is crucial because it sets the basis for the economic interpretation of the connection. In this specific case, in each layer the source node is identified as the institution providing funds; therefore, each edge represents the exposure of such entity towards the target entity, which receives funds instead.
- **Edge weight:** in our data, this is the amount of the exposure. Specifically, it is the creditor's outstanding nominal amount for AnaCredit, the transaction

⁸ While all other layers have source data at monthly or even daily frequency, the security cross holdings layer has SHSG as one of the sources, which has quarterly frequency for reporting. FINREP data are reported quarterly as well. In order to match the monthly frequency of the other layers, data as of Q4 2019 (December) has been used for January and February 2020; data as of Q1 2020 (March) has been used for April and May 2020.

⁹ For example, the securities cross-holdings layer can be further divided into short-term securities versus long-term ones.

¹⁰ An edge-list is a data structure used to represent a network as a list of its edges.

nominal amount for MMSR, the nominal value for the security cross-holdings and the notional amount for derivatives. All these amounts have been standardised by the total assets of the source node institution and have been rescaled between 0 and 1, in order to be aggregated in the multiplex layers.

- **Nodes attributes:** these are all the attributes of source and target nodes, which can be interesting for the analysis. In this specific case, the attributes are total assets and home country of the institution (home or host approach?).
- **Layer:** a field specifying the layer. The naming convention used is *ana* for AnaCredit, *mmsr* for the money market layer, *ch* for the security cross-holdings layer and *emir* for the derivatives layer.
- **Date:** a field specifying the time period (in this case the month) the data refer to.

Once all the layers follow this structure, they can be joined in the same table and pivoted, so that for each time period, there is the same set of source and target nodes in each layer.

At this point, we have the weighted edge-list for each layer. The next step is computing the unweighted edge-list, simply dividing each weighted edge-list by itself. Finally, it is possible to construct the following multi-layer edge-lists:

- **multi_a:** this is the aggregated multiplex layer, A . It is such that the weight of an edge is equal to 1 if there exists an edge in one of the sub-layers or 0 otherwise.
- **multi_o:** this is the overlapping multiplex layer, O . It is obtained by summing the unweighted edge-lists of each sub-layer. Therefore, given that there are four sub-layers, its values are integer and range between 0 and 4.
- **multi_ow:** this is the weighted overlapping multiplex layer, O^w . It is obtained by summing the weighted edge-lists of each sub-layer. Given the weight rescaling of each sub-layer, its values range between 0 and 4.

The reader might wonder why constructing three different multiplex layers. The weighted overlapping layer, *multi_ow*, characterises the edges from the perspective of the amount of the exposure. The bigger the weight in each layer and the more layers the edge exists in, the bigger the weight in the multiplex. This perspective, however, contains two others: the perspective that there exists a relationship in any of the sub-layers (a pure unweighted perspective) and the one that this relationship exists in one or more specific sub-layers. These other two aspects are captured respectively by the aggregated multiplex layer (*multi_a*) and the overlapping multiplex layer (*multi_o*). These different multiplex layers, as well as the weighted and unweighted versions of the four sub-layers, are equally important to consider in the analysis to better describe and understand the network as a whole.

The last step is transforming the edge-lists into proper graph objects. At this point, we can fully exploit NATkit's modules.

5.2.3. Visualizing the network

“Classic” network visualization is a powerful analytical tool: if done properly it can help gaining useful insights which can guide the analysis. There is one caveat though: when looking at a multi-layer network, it might not be that informative to visualise the layers all together. One layer is likely to be “messy enough” on its own. Therefore, when developing the visualization module, I tried to build a tool that could help users find a pattern in the chaos, focusing on one layer at a time.

Heatmaps are an underestimated way to visualise a network. They provide clear “fingerprints” of how a network looks like: how dense it is, if there are clusters, etc.

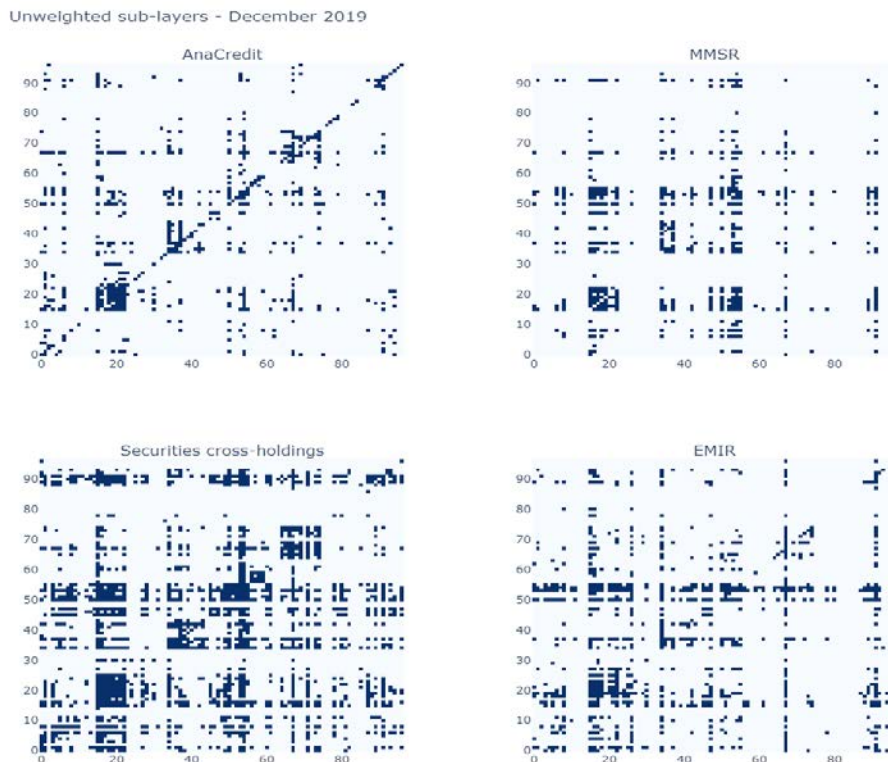


Figure 3

Figure 3 plots the heatmap of the unweighted sub-layers as of December 2019. For confidentiality reasons, the names of the institutions are not displayed, but the nodes have been ordered by country. The differences in the four networks clearly stand out just from a simple visual inspection. It appears that the security cross-holdings layer and the derivative layer are the densest. In the AnaCredit layer, instead, there is a clear pattern of intra-group loans (on the diagonal). Given the way the nodes are ordered, the rectangles we can see around the diagonals represent intra-country clusters, those not on the diagonals are inter-country clusters. One advantage of using heatmaps to visualise a network is that changing the order of the nodes allows to discover and checks for new patterns. For example (not shown here), ordering by total asset size can allow to see whether there are clusters based on size. We could also look at the same picture for the weighted layers, shown in Figure 4. The colour scale here is reversed compared to Figure 3 in order to visualise the low-weight connections: it is easier to see them on a darker background. Compared to the unweighted counterparties, there are only few relevant connections in the layers, highlighted by the lighter colours in the heatmaps.

Weighted sub-layers - December 2019

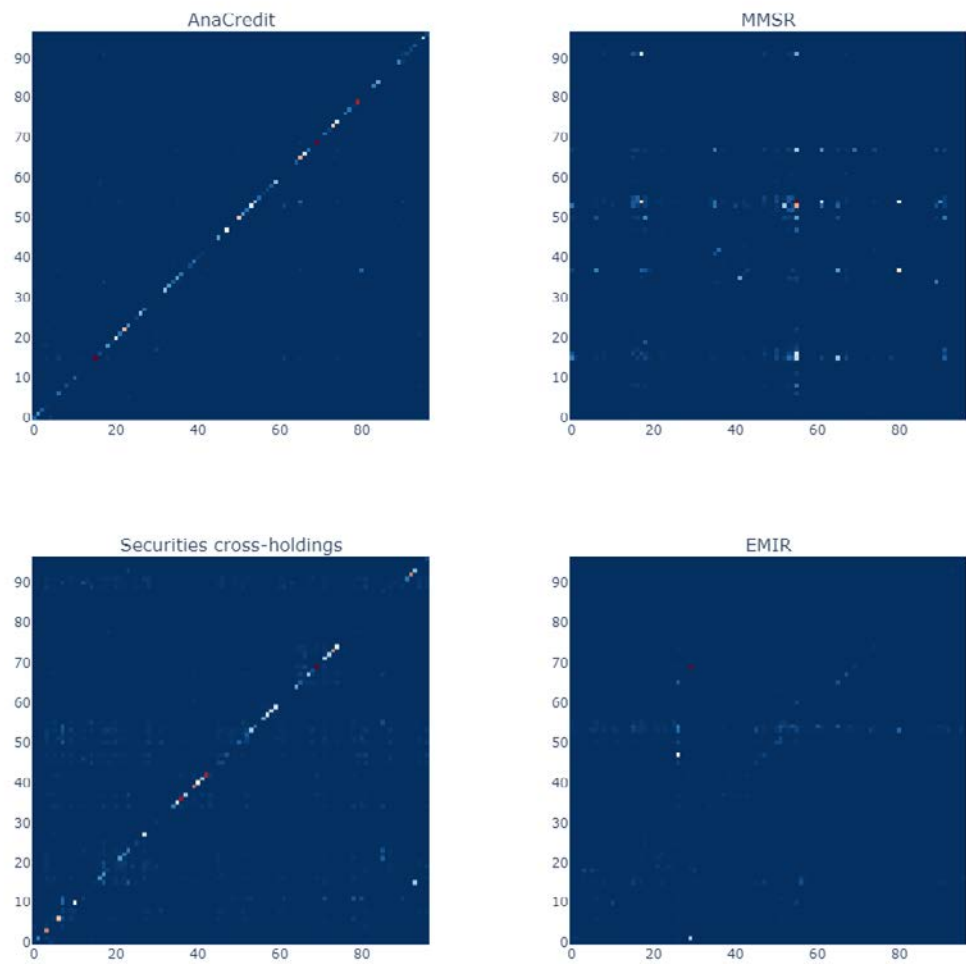


Figure 4

Apart from heatmaps, the tool allows to visualise the network layers in a more standard way. Graphs are usually represented as node-link diagrams, with dots as nodes and lines as the edges among them. The users can use the visualization module to plot the network and see it change over time. Figure 5 has a snapshot of this.

Multiplex overlapping layer

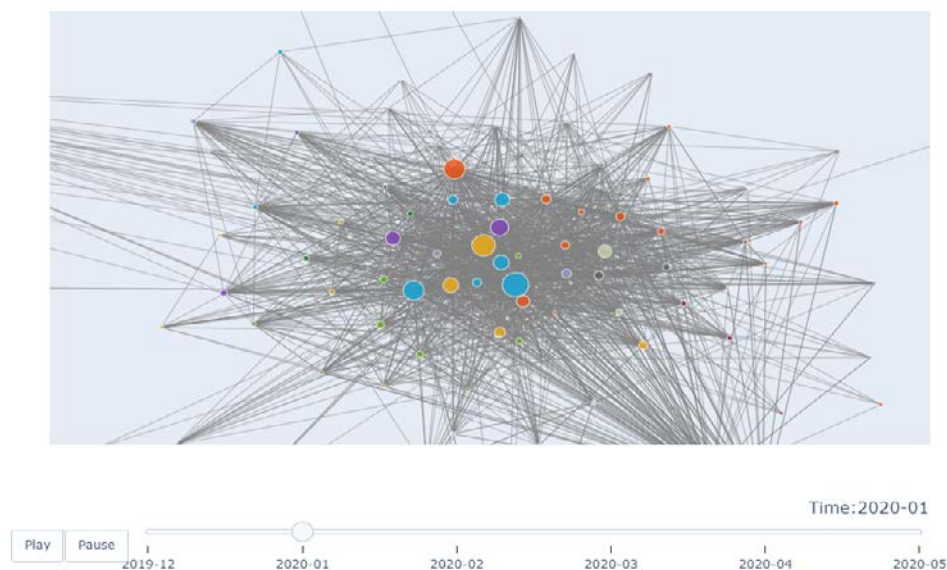


Figure 5

The overlapping layer pictured in Figure 5 is rather dense and, especially in the centre, it is not easy to understand which edge goes where. As stated in Holten and Van Wijk (2009), although node-link diagrams provide an intuitive way to represent graphs, visual clutter quickly becomes a problem in case of graphs comprised of a large number of nodes and edges. One of the most recent trends in network visualisation to address this issue is *edge-bundling*. What edge-bundling does is effectively grouping together edges that go in the same direction. In this way, it is easier to see how the nodes are connected to each other, even when the network is very dense. Figure 6 shows the same network with edge-bundling¹¹ applied to it. Compared to Figure 5, the structure of the network is now more evident, with nodes in the core and others in the periphery.

Visualising a network is only the first step in its analysis; in order to properly understand its characteristic and dynamics, we need to look at its topology.

¹¹ The specific bundling algorithm is *hammer_bundle*, a variant of Hurter, Ersoy and Telea (2012) Kernel Density Estimation (KDE) based edge-bundling algorithm. It is implemented through the Python library Datashader. See [here](#) for more details on Datashader and [here](#) for an intuition on KDE edge-bundling.

Multiplex overlapping layer

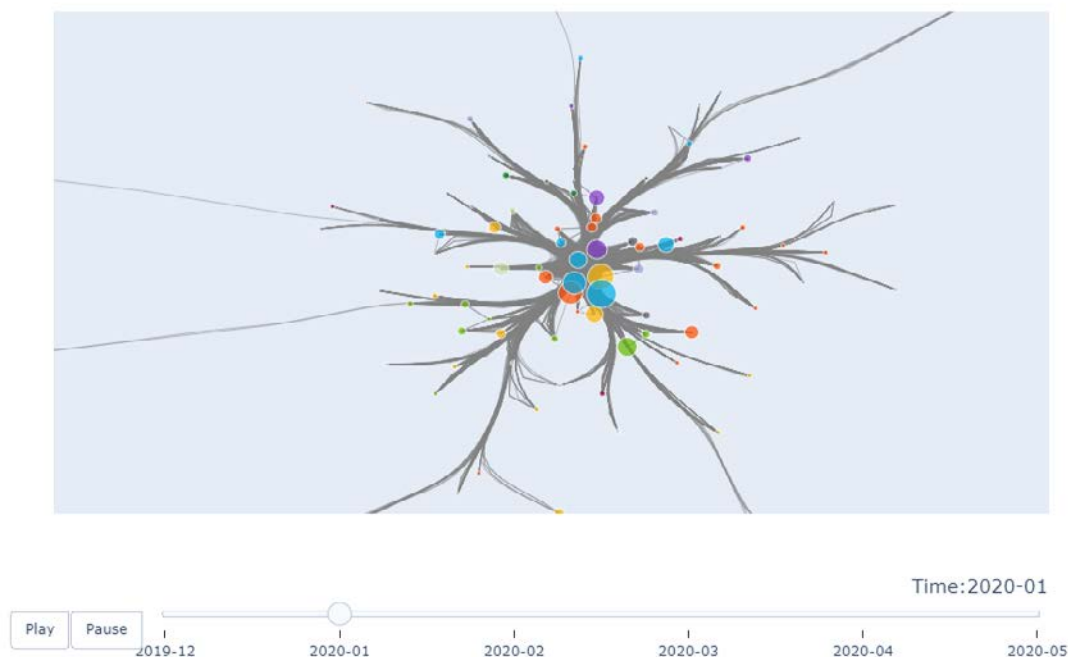


Figure 6

5.2.4. The multiplex topology

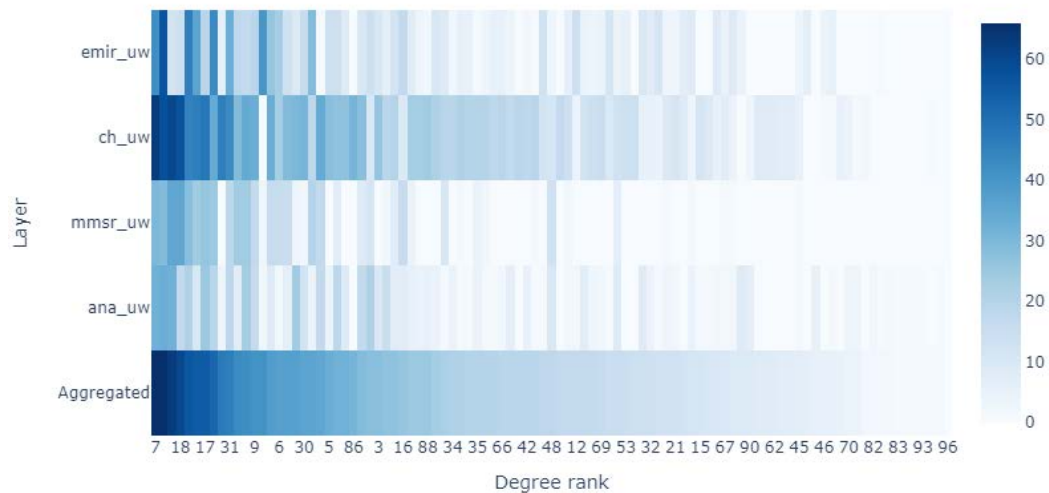
One of the first metrics to look at when analysing the topology of a network is the *degree* of a node. The degree tells how many edges are connected to a specific node; in particular, for directed graph (as the one under analysis) we can talk about *in-degree* (edges that come from other nodes) and *out-degree* (edges that go towards other nodes). The degree is a simple and intuitive centrality measure: a node with high degree (either in or out) is central for the network; therefore, this metric can be used to immediately identify the key nodes.

However, since we are analysing a multi-layer graph, the traditional concept of degree centrality needs to be re-framed. When can be a node defined as central? When is it central in all the layers? Or is being central in one layer enough? There is no straight answer; what matters is to find an analytical approach that can give the most complete picture.

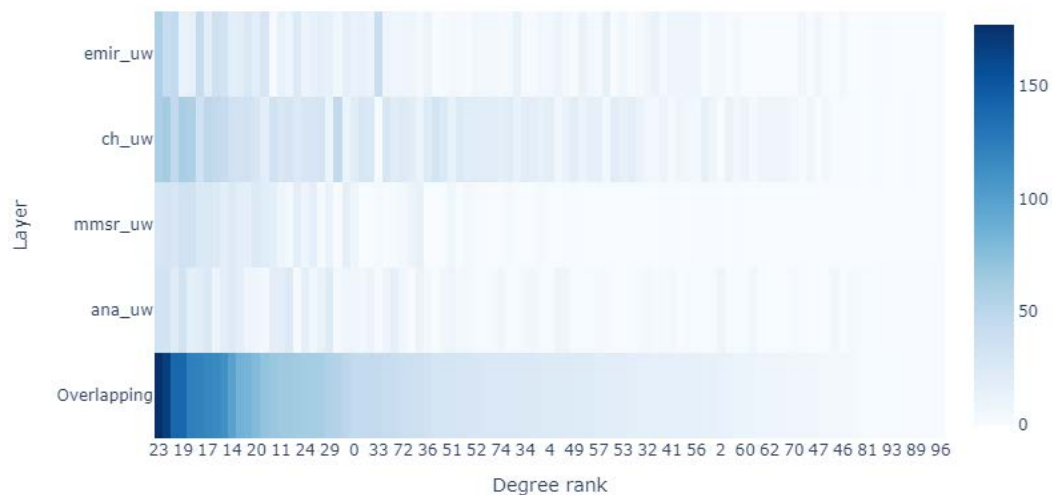
Looking at the relationship between the degree distribution in the multiplex layers and the ones in each of the four sub-layers is a first step in figuring out the centrality profile of the nodes. One might wonder whether nodes with high degree in the multiplex are also high-degree nodes in the sub-layers and whether this behaviour changes when we consider the weighted or the unweighted networks. Figure 7 shows the out-degree distributions in the different layers as of December 2019, ordered by the degree ranking of the multiplex. Panel a) focuses on the unweighted layers. Looking at the different distributions, it appears that the cross-holdings layer (ch_uw) is the one resembling the aggregated multiplex layer the most. In other words, high degree nodes in the aggregated layer tend to remain as such in the cross-holding layer. Another feature emerging from a visual inspection is that the out-degree distributions of the loan exposures and secured money market layers and in part that of the derivative layers have a lower range of values (the colour band is lighter) than

the one of the aggregated layers. This means that, considering the same institution, this is likely to have more connections in the cross-holdings security layers than in the others. Since we are looking at the out-degree and given how we defined the source and target nodes, the above fact implies that an institution tends to have more exposures towards other institutions in the security market.

a) Ordered out-degree - December 2019 - unweighted layers and aggregated multiplex



b) Ordered out-degree - December 2019 - unweighted layers and overlapping multiplex



c) Ordered out-degree - December 2019 - weighted layers and weighted multiplex



Figure 7

Panel b) confirms the results of panel a) and does not offer new insights. Looking at the weighted layers¹² in panel c) instead, we can spot some differences. First of all, in the weighted overlapping layer we can see a sharper contrast in the colour gradient, meaning there are fewer high degree nodes compared to its unweighted counterpart. Secondly, the secured money market layer distribution appears to be more similar to the multiplex one, compared to its unweighted counterpart. In other words, in terms of absolute number of out connections, the money market layer might appear less correlated with the multiplex layer, but when we look at the size of the transaction, the picture changes. In this regard, it might be useful to look at Figure 8. The two heatmaps show the Kendal correlation¹³ as of December 2019 between the different degree distributions.

¹² In a weighted network, the degree is not just the number of edges, but the sum of the weights of the edges.

¹³ The Kendall Correlation is a measure of rank correlation: the similarity of the orderings of the data when ranked by each of the quantities. It is a non-parametric alternative to Pearson's correlation (parametric).

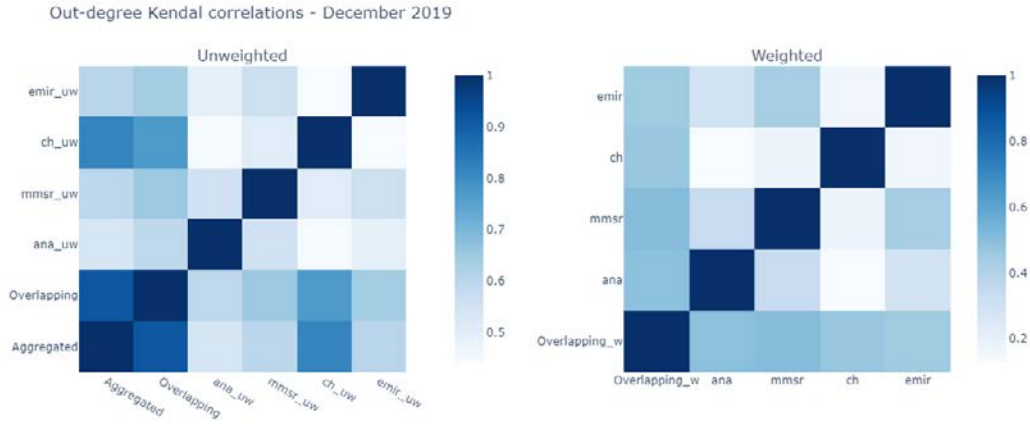


Figure 8

Continuing on the secured money market layer, we can see how, in the unweighted heatmap, mmsr_uw out-degree distribution is not the one with the highest correlation to the degree distributions of the aggregated and overlapping layers: this, instead, is true for its weighted counterpart. Another important difference between the two heatmaps is how the correlation between the multiplex layer and each sub-layer is more diversified in the unweighted networks, compared to the weighted counterparts. This might hint at an inverse correlation between number of transactions in layer and their weight.

Looking at the degree distributions on their own, however, does not tell us clearly whether a high-degree institution in the multiplex layers is so because it equally participates in all sub-layers or instead is a big hub in only one layer. Among the possible metrics to assess this, there is one called the multiplex participation coefficient, defined as:

$$P_i = \frac{M}{(M-1)} \left[1 - \sum_{\alpha=1}^M \left(\frac{k_i^{[\alpha]}}{o_i} \right)^2 \right]$$

where:

- M is the number of layers.
- $k_i^{[\alpha]}$ is the degree of node i in layer α .
- o_i is the degree of node i in the overlapping layer (also known as overlapping degree).

This coefficient measures whether the links of node i are equally distributed among the M layers or are instead mainly concentrated in just one or a few layers. P_i can have values in the interval $[0,1]$; in particular, when it is equal to 0 all the edges of i lie in one layer, while $P_i = 1$ only when node i has exactly the same number of edges on each of the M layers. Figure 9 shows the coefficient distribution as of December 2019, for the unweighted (O) and weighted multiplex layers (O^w).

Multiplex participation coefficient (out-degree) distribution - December 2019

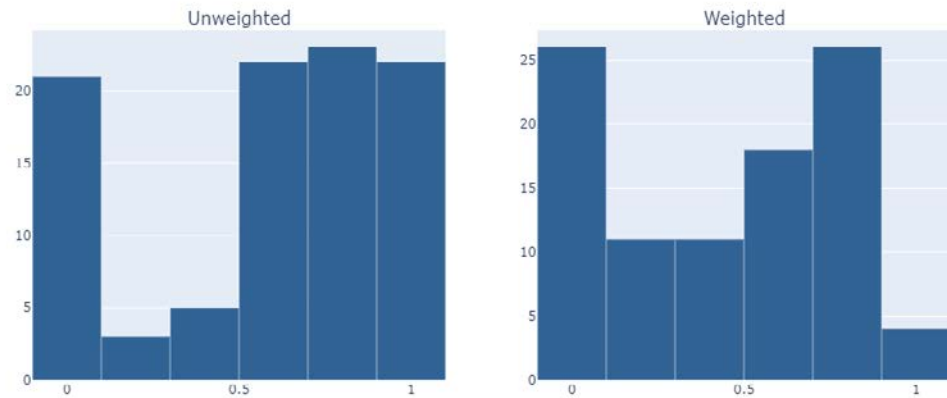


Figure 9

While the unweighted distribution almost looks uniform (if not for a gap between 0.1 and 0.5), the weighted distribution appears to be bimodal. Moreover, while in the unweighted distribution there are over twenty institutions equally participating across all four sub-layers, in the weighted one these institutions decrease to less than five. This means that there are only few institutions with high exposures in all four layers. Another aspect worth understanding is the relationship between the participation coefficient and the degree of a certain node. Figure 10 plots the z-score of the overlapping out degree and the weighted overlapping out-degree over the corresponding multiplex participation coefficients.

Overlapping out-degree (z-score) over multiplex participation coefficient - December 2019

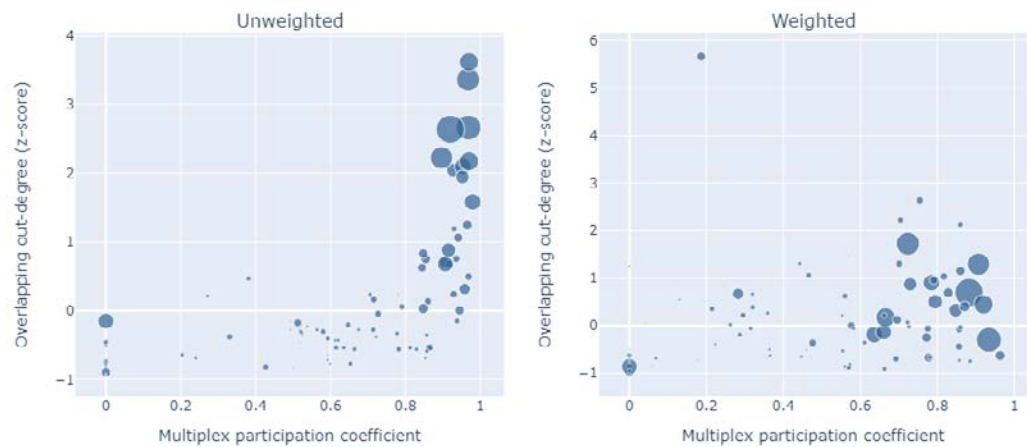


Figure 10 – bubble size represents total assets.

The two scatterplots paint a rather interesting picture. In the unweighted networks, there is a clear positive non-linear correlation between the z-scores and the multiplex participation coefficient. In other words, the more layers an institution has exposures in, the more central it is in the multi-layer network. Moreover, as shown by the size of the bubbles that is increasing moving towards the top-right corner of the graph, there seems to be a positive correlation between z-scores and total assets, as well as

multiplex participation coefficients and total assets. In the weighted networks, instead, it appears that the z-scores are almost uniformly distributed across the different values of the participation coefficient. This means that, from an exposures' perspective, the institutions acting in all or only in few markets are not that different in terms of centrality to the network. From the total assets' perspective, there is still a positive correlation between participation coefficients and total assets, while the one with z-scores is not there anymore.

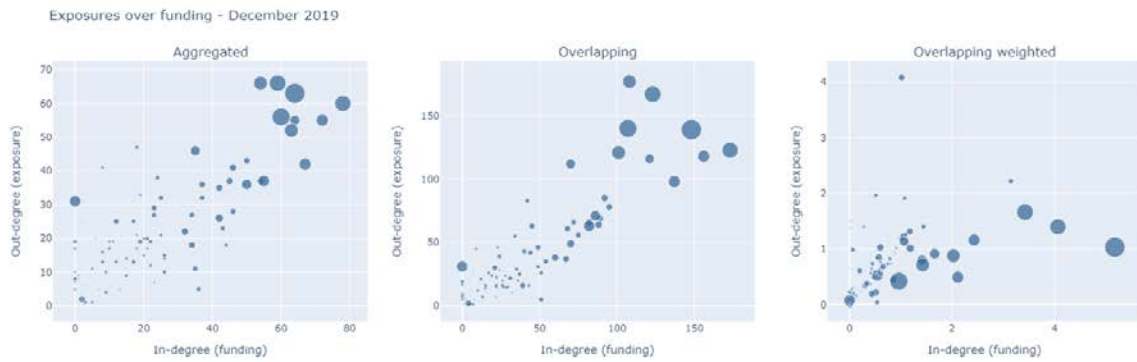


Figure 11 – bubble size represents total assets.

Up to now, we have been looking at the out-degree, thus analysing the network from the exposure side; however, we could also look from the funding perspective with the in-degree. Taking inspiration from Huser and Kok (2019), an interesting relationship to check is the one between exposures and funding, looking at whether institutions with high exposures both in terms of numbers and amount also are the ones receiving more funds, and how this relates to the total assets. Figure 11 provides an interesting insight on the matter. In the aggregated and overlapping multiplex layer there is a positive correlation between funding and exposure: institutions highly connected on the exposure side tend to be so also on the funding side. Moreover, total assets are positively correlated with out and in degree. The overlapping weighted layer tell us a different story: the relationship between exposures and funding is non-monotonic, increasing at first and then decreasing. To better understand this finding, it might be useful to look at each single weighted sub-layer, shown in Figure 12.

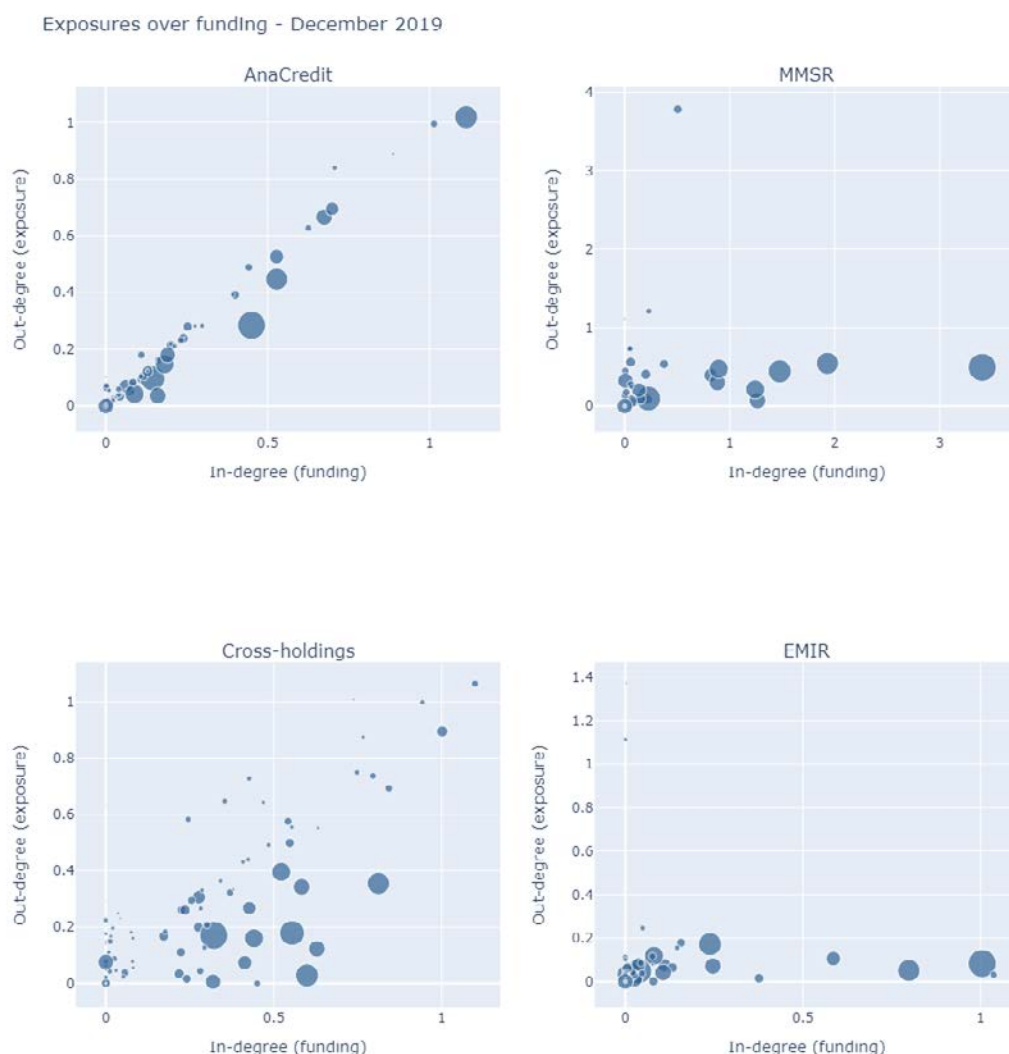


Figure 12 – bubble size represents total assets

While in the AnaCredit and in the security cross-holdings layers there is a positive relationship between funding and exposure, the secured money market layer and the derivative layer do not show such relationship. Instead, it seems that most of the funds received from the network are not redistributed back to it, but likely directed towards other markets. In particular, we can observe nodes with a higher in-degree, compared to the out-degree. One consideration about the cross-holding layer: it appears that the largest (in total asset size) institutions tend to act more as issuers of securities, rather than holders. This could be explained by the fact that large institutions issue more securities compared to smaller ones or maybe securities issued by large banks can be more desirable, because less risky, rather than the ones issued by smaller institutions.

So far, we have shown some topological characteristics of the multiplex layers, and how the sub-layers influence them. As a next step, we look at the influence each sub-layer has over the others. The overlapping degree distributions and the multiplex participation coefficients can give an idea of the existence of inter-layers correlations; however, they are not able to disentangle the effect that each layer has over the others. A metric that can give more insights in this regard is the conditional

probability of finding an edge e_{ij} in layer α' given the existence of the same edge in layer α :

$$P(e_{ij}^{[\alpha']} | e_{ij}^{[\alpha]}) = \frac{\sum_{ij} e_{ij}^{[\alpha']} e_{ij}^{[\alpha]}}{\sum_{ij} e_{ij}^{[\alpha]}}$$

For the weighted layers, the probability is:

$$P_w(e_{ij}^{[\alpha']} | w_{ij}^{[\alpha]}) = \frac{\sum_{ij} e_{ij}^{[\alpha']} w_{ij}^{[\alpha]}}{\sum_{ij} w_{ij}^{[\alpha]}}$$

Figure 13 shows an overview of the different conditional probabilities for the unweighted and weighted version of the four sub-layers.

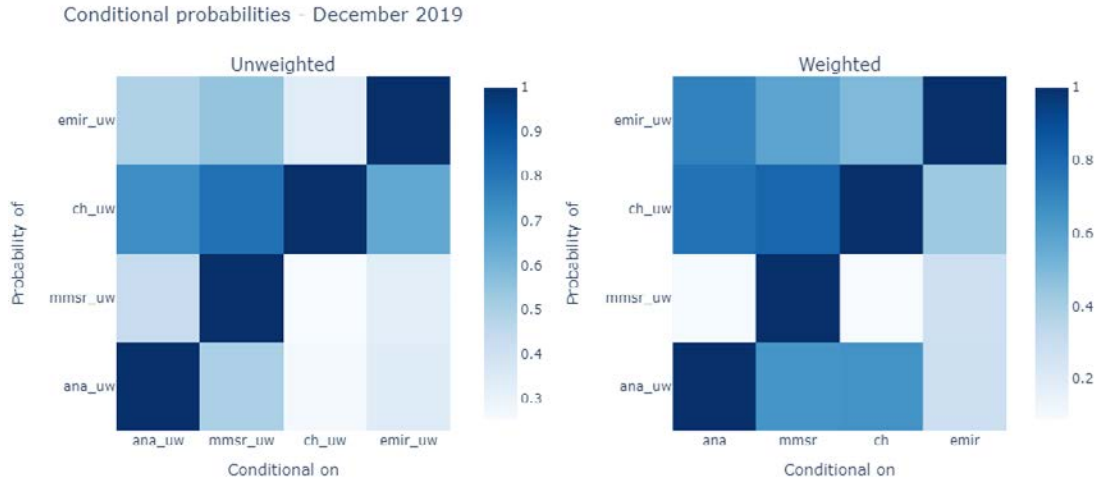


Figure 13

As we can clearly see from the two heatmaps, the probabilities are not symmetric. For example, in the unweighted heatmap, we can see that the probability of an edge being in the cross-holdings layer (ch_uw) conditional on existing in the AnaCredit layer (ana_uw), is much higher than the reverse. Same goes for the probabilities of ch_uw conditional on mmsr_uw. This means that if an institution has a relationship with another on the securities layer, it is very likely that it will also have a relationship on the loan and secured money market layers. However, the inverse is not true. Indeed, this could be explained by the presence of banking groups which are not investment heavy (and thus do not engage on the security layer), but rather focus on more “traditional” banking activity. Overall, the security cross-holdings layer is the one with the highest conditional probabilities, followed by the derivative layer; while AnaCredit and MMSR are the layers doing most of the “conditioning”. This is in line with the idea of traditional banking business. Looking at the weighted conditional probabilities, though, we can observe some differences. The secured money market layer is much less influenced by the loan exposure and cross-holdings layers than in the unweighted scenario, while the AnaCredit layer becomes more influenced. Overall, the weight of the relationship affects the conditional probabilities.

6.5. Time dynamics analysis

Up to now, we have been looking at one specific snapshot of our network; however, the network is dynamic, hence it changes over time. Therefore, it is important to have

a framework to observe and assess such changes, as they might contain insights that would be missed otherwise.

One rather straightforward approach to do so is to repeat the above analysis in different time periods. As an example, Figure 14 plots the Kendall correlations of the out-degree distribution (the one in Figure 6) as of each month in our sample.

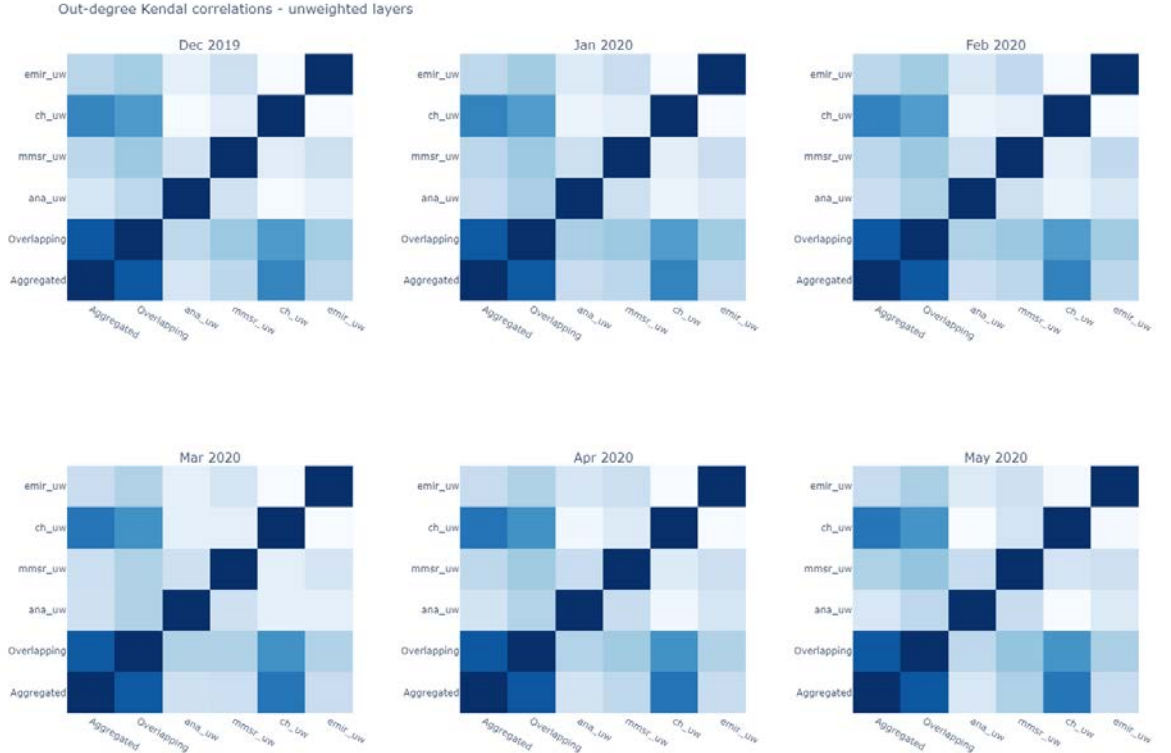


Figure 14

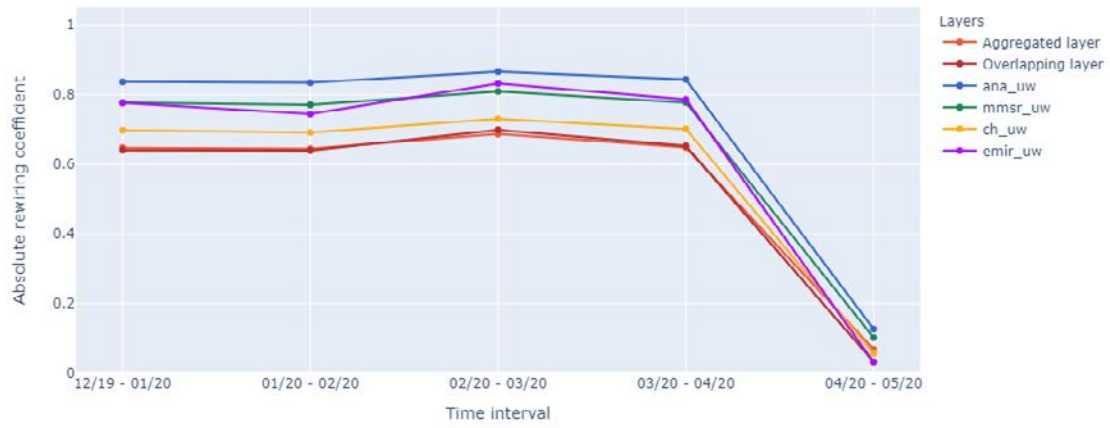
The visualization can be a useful tool to spot relevant differences among time periods. However, looking at this figure in particular, there seem to be none.

A different and maybe more interesting approach is proposed by Goezt and Han (2019). They argue that the majority of measures usually employed to assess network change over time are scalar measures. For example, one can show how the degree of a node changes as time goes by. Instead, they propose to use the cosine similarity to capture not just the change in the node degrees, but its relationship to other nodes. These are vector (or matrix)-based comparisons, rather than scalars, and the authors refer to them as "rewiring" coefficients. With NATkit, it is possible to look at the rewiring coefficients of the multi-layer network over time (Figure 15 and Figure 16). The absolute rewiring coefficient is simply the cosine distance between the same network (represented in a stacked vector form) in two different time periods. The formula is the following:

$$rc^A(N_{t+1}, N_t) = 1 - \cos \theta_{N_{t+1}, N_t}$$

where N is the network under consideration and $\cos \theta_{N_{t+1}, N_t}$ is the cosine similarity between the two snapshots of the network. The closer to 1 is the coefficient, the bigger the rewiring in the network; conversely, a coefficient very close to 0 means little distance between the networks. The advantage of this approach to study network rewiring over time is that, being a vector comparison, it allows to compare the network in its entirety, without being forced to choose the perspective of a specific node or node-related metric.

a) Absolute rewiring coefficients - Unweighted layers



b) Absolute rewiring coefficients - Weighted layers

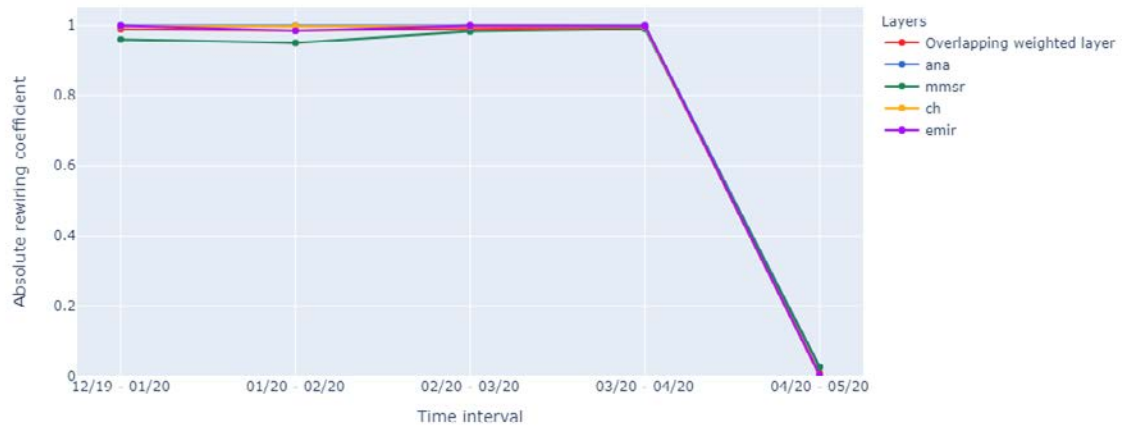
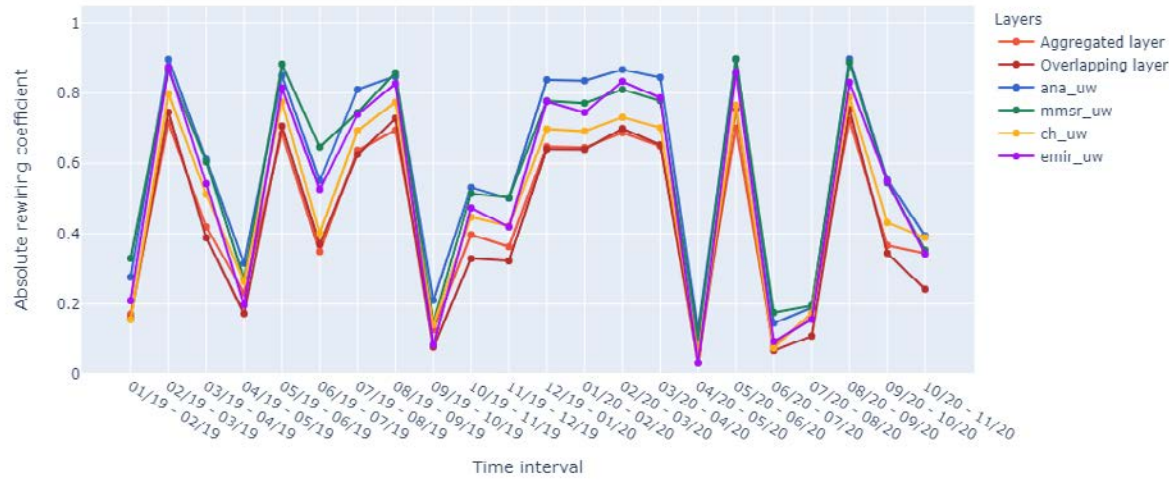


Figure 15

The two plots in Figure 15 show the absolute rewiring coefficients for the different layers in the network. Both unweighted and weighted layers follow the same pattern, with high rewiring observed between December 2019 and April 2020, followed by a sharp drop between April 2020 and May 2020. In order to explain this behaviour, it can be useful to expand the time series of our sample.

a) Absolute rewiring coefficients - Unweighted layers



b) Absolute rewiring coefficients - Weighted layers

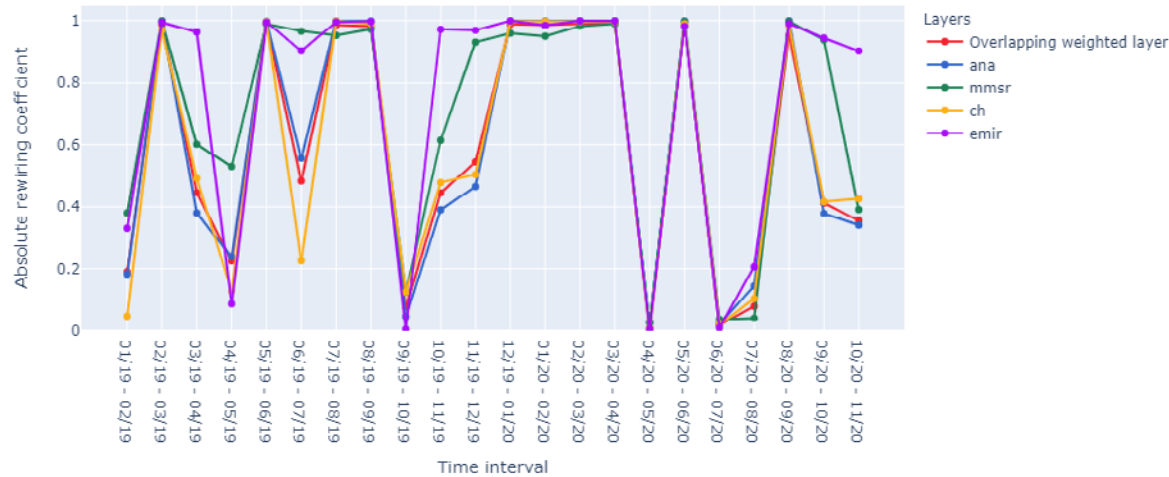


Figure 16

Figure 16 shows the absolute rewiring coefficient over a two-year period. Both the unweighted and weighted time-series show signs of seasonality. Spikes can be observed close to quarter ends: maybe this is due to operations for balance sheet adjustment in view of the reporting obligations. We could check for this hypothesis by simply decomposing the time-series into trend and seasonality¹⁴, using an additive model. Figure 17 shows the results of the decomposition for the aggregated layer. Looking at the seasonality, we can observe a positive spike in the coefficient for the last month of each quarter. This is in line with our hypothesis of rewiring becoming more intense in view of the reporting obligations. More checks should be performed to validate our guess, but this is out the scope of this work. Regardless of the specific economic interpretation of the patterns of the rewiring coefficient, this metric manages to capture something different from metrics such as network density or average degree, which only show an incomplete picture.

¹⁴ The Python library *statsmodel* offers an easy way to do so.

Absolute rewiring coefficients decomposition - Aggregated layer



Figure 17

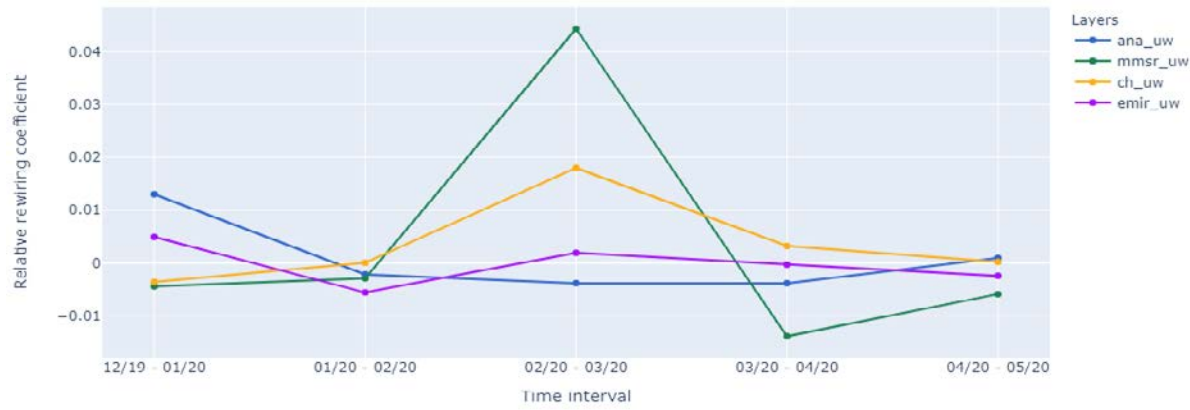
Albeit useful, the absolute rewiring coefficient does not convey the direction of the change. From a vector perspective, the network at $t + 1$ might lie above or below its counterpart at t , pointing in the same or the opposite distance. To identify the direction of change, Goetz and Han (2019) suggest comparing the network change to the equivalent, same-period change in some reference or benchmark network. They call this metrics the relative rewiring coefficient:

$$rc^R(N_{t+1}, N_t, N_{t+1}^B, N_t^B) = \cos \theta_{N_{t+1}, N_{t+1}^B} - \cos \theta_{N_t, N_t^B}$$

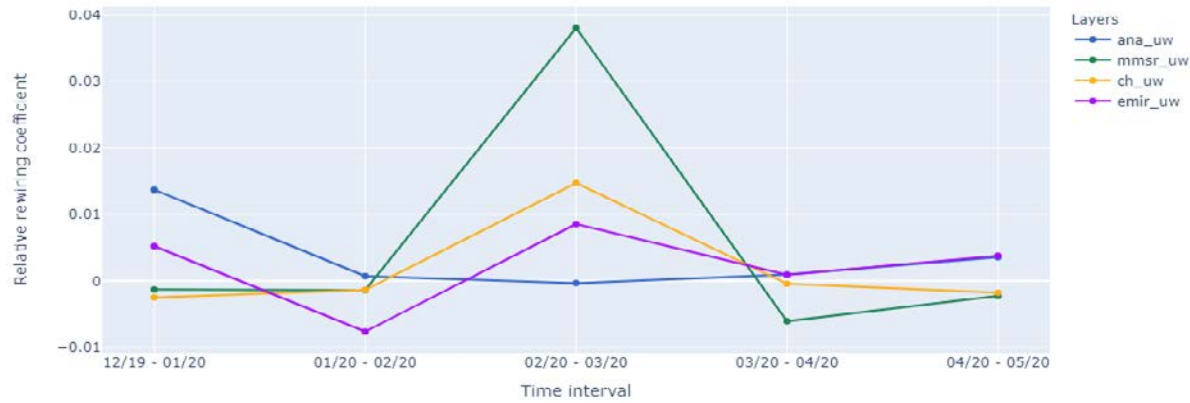
where N^B is the benchmark network. The relative rewiring coefficient can have values between -1 and 1, where the closer to 1, the more similar is the network to the benchmark one; the closer to -1, the more dissimilar from the benchmark.

Going back to our six-month window, we saw in Figure 15 that each sub-layer and the multiplex layers show very similar rewiring pattern, with higher values of the absolute rewiring coefficient until April 2020, followed by a sharp decline in May. We can now look at the relative rewiring coefficients of each sub-layers, taking as the benchmark the different multiplex representations (aggregated, overlapping and overlapping weighted).

a) Relative rewiring coefficients - Unweighted layers - benchmark: Aggregated layer



b) Relative rewiring coefficients - Unweighted layers - benchmark: Overlapping layer



c) Relative rewiring coefficients - Weighted layers - benchmark: Overlapping weighted layer

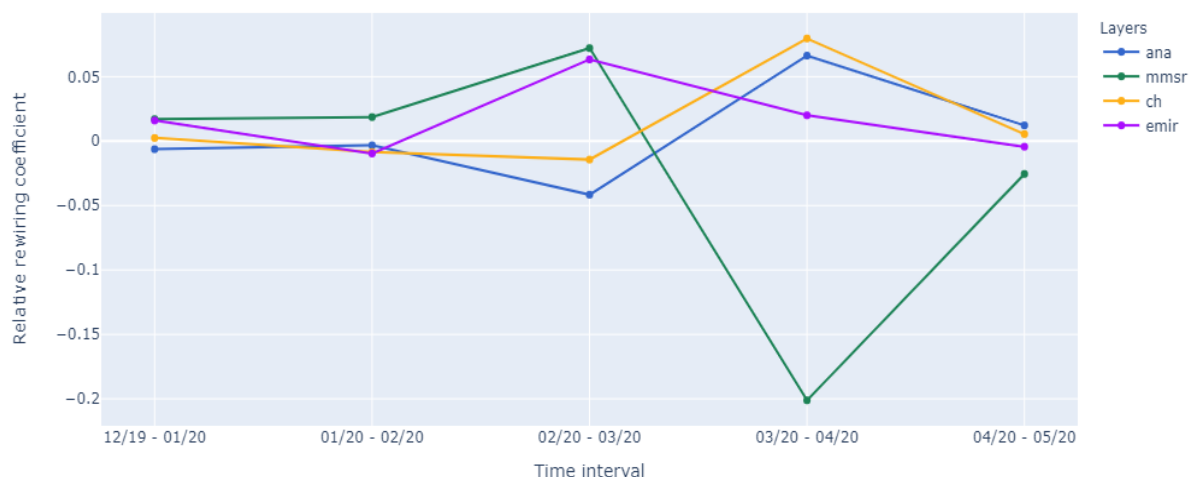


Figure 18

Figure 18 shows the relative rewiring coefficients of the four sub-layers with a different multiplex layer as the benchmark. Panel a) and b) paint a rather similar picture for the unweighted layers. Despite the range of the coefficient being very small and close to 0, we can observe an interesting spike in all the layers, except AnaCredit, between February and March 2020. In that period the all the sub-layers have high absolute rewiring coefficients (Figure 15), but from a relative rewiring perspective, the secured money market layer and the security cross-holdings layer (panel a and b) and the derivative one (panel b only) became more similar to the aggregated and overlapping networks. The loan exposure layer, instead, moved from similarity towards “neutrality”, not truly increasing or decreasing its distance from the benchmark layers. Things look different for the weighted layers. The money market layer is still the one with the highest coefficient, however, in the opposite direction of its unweighted counterpart. Moreover, there seem to be a grouping in the layers, with AnaCredit and security cross-holdings moving very similarly and the derivative and secured money market layer moving together in opposite directions from the other two.

5.2.6. Identifying non-linearities

As explained in this showcase analysis, in order to highlight the properties of the multi-layer network, the multiplex layers and the sub-layers are to be analysed and compared together. Indeed, just looking at the aggregated and overlapping (unweighted and weighted) layers on their own might not give the full picture of the complexity of the connections. The literature on the topic has highlighted that multi-layer networks present non-linearities, which would be lost by simply collapsing all the layers together. This, in turn, would lead to wrong results and decisions. Unfortunately, identifying these effects is not straightforward. Nevertheless, with the following metrics, I hope to give the readers a glimpse of or an intuition about them and prove, why multi-layer networks require their own approach and set of adapted metrics.

One property of real-world networks is the tendency of nodes to form the so-called *triangles*, that is, simple cycles involving three nodes. In a triangle ijm , node i is

connected to node j and node m , and these two are also connected with each other. Another useful concept is the *triad*. One might think of it as an open triangle: in the above example, node i is connected to node j and node m , but these two are not linked. The clustering coefficient, C_i , allows us to quantify how likely it is that two neighbours of node i are also connected among each other, in other words which proportion of triads close into triangles. Each node has its own clustering coefficient, and averaging them give the average clustering coefficient, C . Table 2 shows the values of the average clustering coefficient for the aggregated multiplex layers and the unweighted sub-layers¹⁵ as of December 2019.

Table 1

| Layer | Average Clustering Coefficient |
|------------|--------------------------------|
| Aggregated | 0.601 |
| ana_uw | 0.029 |
| mmsr_uw | 0.033 |
| ch_uw | 0.336 |
| emir_uw | 0.054 |

It is interesting how different the coefficients for the sub-layers are compared to the aggregated layer. The latter has a coefficient almost double the one of the security cross-holdings layer (ch_uw). These results do make more sense when thinking about the definition of the clustering coefficient. Since its the ratio of triangles over triads, in the aggregated layers many of the triads that would not close in a specific layer do close instead. Figure 19 and Figure 20 give some more insight in the distributions of the clustering coefficient.

Ordered clustering coefficients - December 2019

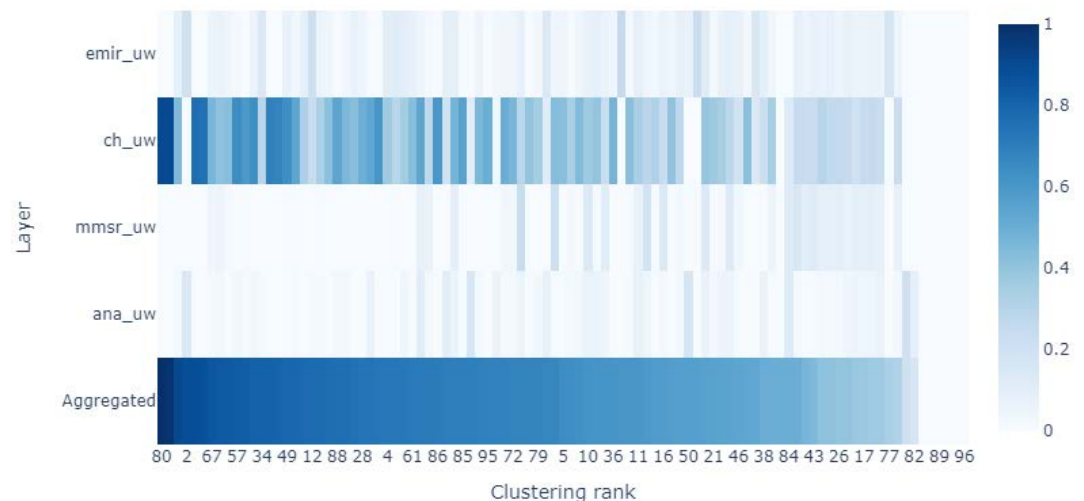


Figure 19

¹⁵ We are focusing on the unweighted layers, because we are interested in the existence of a triangle, rather than its weight.

Clust. coeff. Kendal correlations - December 2019 - unweighted

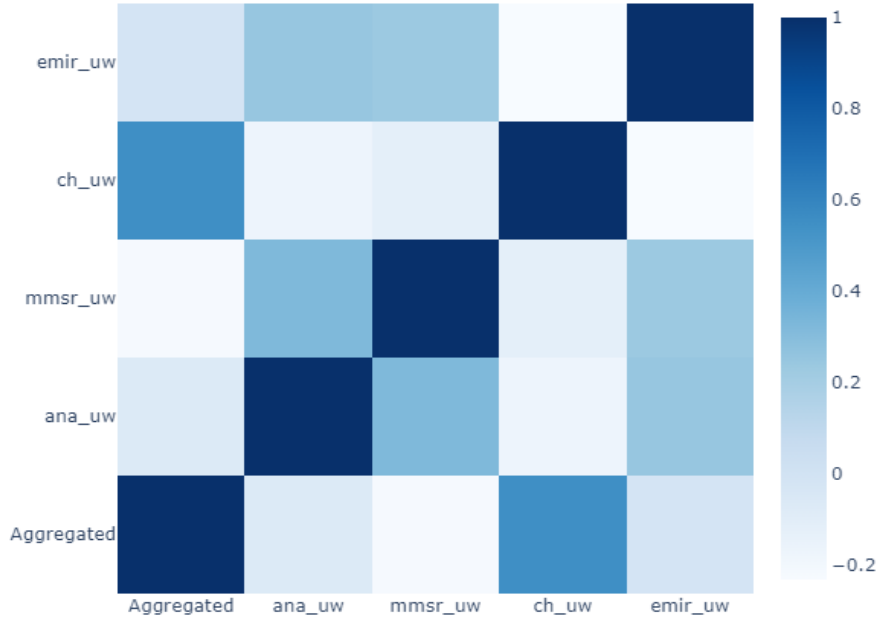


Figure 20

In Figure 19 the contrasts of Table 2 are highlighted. The distribution of the clustering coefficients of the aggregated layer is on a significantly higher range of values, with only the security layer being comparable. Indeed, looking at the Kendall correlations (Figure 20) the aggregated distribution is highly correlated with the one of the security layer.

While these results could be interesting, the traditional clustering coefficient does not consider the multi-layer nature of the network. In terms of computation, the aggregated layer is treated just as any other single layer, and the resulting coefficient tells us very little about the connection across the sub-layers. Therefore, we follow Battiston et al (2014), who computed two modified versions of the clustering coefficient, centred around updated definitions of triangles and triads, more suited for multi-layer systems:

- 2-triangle - a triangle formed by an edge belonging to one layer and two edges belonging to a second layer.
- 3-triangle - a triangle composed by three edges all lying in different layers.
- 2-triad - a triad whose two links belong to two different layers of the systems.

The two new clustering coefficients are defined as:

- C_{1i} : for each node i , the ratio between the number of 2-triangles with a vertex in i and the number of triads centred in i .

- C_{2i} : for each node i , the ratio between the number of 3-triangles with node i as a vertex, and the number of 2-triads centred in i .

The advantage of these two new measures is that they take into account all the sub-layers together, instead of simply looking at the aggregated layer. The new concepts of triangles and triads allow to differentiate between a triangle that is all in one layer and one who is in two or three layers. This is important because the connection between two banks in the loan exposure layer is not the same connection in the derivative layer. Shock propagation can be a practical application where such a difference is important: assuming there is a shock in one specific market, we need to know that it will propagate differently if the next connection is on the same market or on another one.

For the month of December 2019, the average clustering coefficient of the aggregated layer is 0.601. The average C1 and C2 clustering coefficient for the multi-layer unweighted network are 0.136 and 0.244, respectively. This means that simply aggregating the sub-layers together would over-estimate the clustering of the network, because it would overlook the non-linear interactions existing across the layers.

6. Conclusions

After the financial crisis, the ECB accelerated the collection of granular statistical and supervisory datasets. Each of them can bring new insights into the function of the financial markets and offer new ways to assess risks for their stability and thus help policy makers in their decisions. We argue, however, that only looking at these asset classes simultaneously can account for non-linearities that are caused by the interconnectivity of those markets. We therefore constructed a new dataset that integrates these granular data and which we use to build a multi-layer dynamic network which covers loans, securities, derivatives, and money market transactions of the significant banking groups in the euro area.

The newly created network is constructed in a modular and flexible way. This way researchers can use the definition of banking groups they need, or they select only those layers they need for the analysis of the credit market. For example, for some use cases a supervisory definition is needed, while others might want to look at pure ownership structures. In different cases, users might want to focus only on the loan and securities layers in their research and disregard the other layers.

No matter how the multi-layer dynamic network is used, it becomes clear that new tools are needed to deal with the interconnections of the different layers. We therefore describe how we put together useful tools in a Python package NATkit, that supports the analysis of the combined layers. As we have shown in this paper, multi-layer dynamic networks need an ad-hoc analytical approach that can properly capture and describe their topology. A multi-layer network is not just the sum of its sub-layers; therefore, its analysis should always encompass both individual layers and the multiplex layers. Each tells a part of the story, and both sets are needed to get the full picture. With NATkit, we tried to establish the framework to enable this, taking from the relevant literature the best suited approaches and metrics for the type of data under consideration.

We hope that the code and the documentation on the construction of the network as well as the toolkit, spark more research in this topic and accelerates the interest in using the granular datasets. The work on these data, however, is far from over. Our work has shown that unique, reliable, and timely master and meta data are key for the construction of such a network. Without them the granular datasets can not be used to their full potential. Therefore, many initiatives on data integration and further development of master data are going on in the background of the ESCB: working groups are continuously strive to enhance the coverage of identifiers in the datasets, which makes it easier to integrate the data. At the same time, several project as The Integrated Reporting Framework (IReF) and the Banks' Integrated Reporting Dictionary (BIRD)¹⁶ will pave the way to achieve true semantic integration for granular banking data in the ESCB, which will greatly enhance the data quality of such projects as the multi-layer network.

¹⁶ See also the ESCB long-term strategy for banks' data reporting.

References

- Aldasoro, I., & Alves, I. (2018). Multiplex interbank networks and systemic importance: An application to European data. *Journal of Financial Stability*, 35, 17-37.
- Allen, F., & Gale, D. (2000). Financial Contagion. *Journal of Political Economy*.
- Bargigli, L., di Iasio, G., Infante, L., Lillo, F., & Pierobon, F. (2015). The multiplex structure of interbank networks. *Quantitative Finance*, 15(4), 673-691.
- Battiston, F., Nicosia, V., & Latora, V. (2014, March 12). Structural measures for multiplex networks. *Physical Review E*, 89(3).
- Draghi, M. (2019). Macroprudential policy in Europe. *Fourth annual conference of the ESRB*. Frankfurt am Main.
- Erdős, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 290-297.
- Farkas, M. (2020). Network Topology of Target2 Interbank Payments. *ECB Graduate Programme Paper*.
- Freixas, X., Parigi, B. M., & Rochet, J.-C. (2000, August). Systemic Risk, Interbank Relations, and Liquidity Provision by the Central Bank. *Journal of Money, Credit and Banking*, 32(3), 611-638.
- Han, Y., & Goetz, S. J. (2019, July 24). Measuring network rewiring over time. (I. Sendiña-Nadal, Ed.) *PLoS ONE*, 14(7).
- Holten, D., & van Wijk, J. J. (2009). Force-Directed Edge Bundling for Graph Visualization. (H. -C. Hege, I. Hotz, & T. Munzer, Eds.) *Eurographics/ IEEE-VGTC Symposium on Visualization*, 28(3).
- Hurter, C., Ersoy, O., & Telea, A. C. (2012, June). Graph bundling by Kernel Density Estimation. *EUROVIS 2012, Eurographics Conference on Visualization*, 865-874.
- Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. (E. Estrada, Ed.) *Journal of Complex Networks*, 2, 203-271.

- Kok, C., & Hüser, A.-C. (2019). Mapping bank securities across euro area sectors: comparing funding and exposure networks. *ECB Working Paper Series*, 2273.
- Kurant, M., & Thiran, P. (2006, April). Layered Complex Networks. *Physical Review Letters*, 96.
- Poledna, S., Molina-Borboa, J. L., Martínez-Jaramillo, S., van der Leij, M., & Thurner, S. (2015, October). The multi-layer network nature of systemic risk and its implications for the costs of financial crises. *Journal of Financial Stability*, 20, 70-81.
- Réka, A., & Barabási, A.-L. (2002, January 30). Statistical Mechanics of Complex Networks. *Reviews of Modern Physics*, 74(1).
- Spiaggiari, M. (2020). A journey from data management to data analysis with Anacredit. *ECB Graduate Programme Paper*.

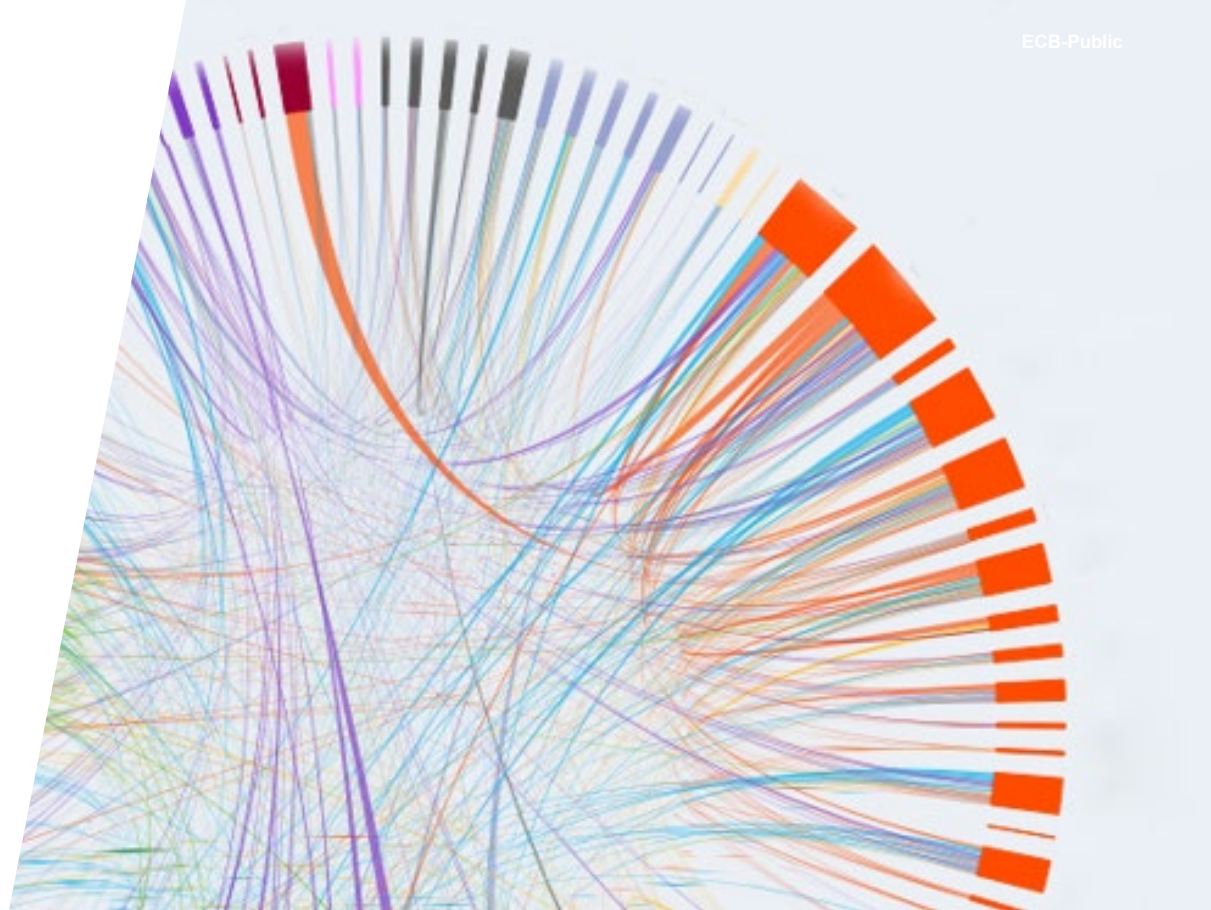


EUROPEAN CENTRAL BANK

EUROSYSTEM

A multi-layer dynamic network

For significant
European banking
groups



14/02/2022

Annalaura Ianiro

Directorate General Macprudential
Policy and Financial Stability

Jörg Reddig

Directorate General Statistics

Presentation: A multi-layer network

- 1 Introduction
- 2 Building a multi-layer network
- 3 Analysing a multi-layer network
- 4 Conclusion and outlook

Granular data at the ECB

The ECB has a rich set of granular data at their disposal

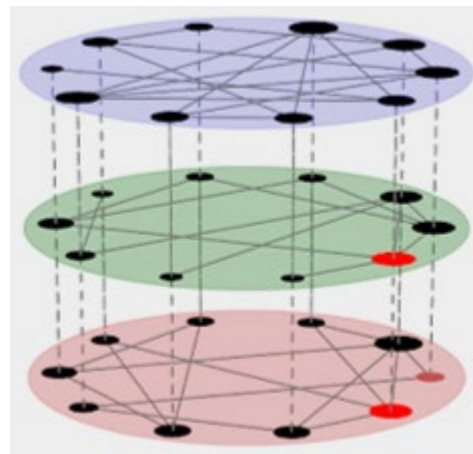
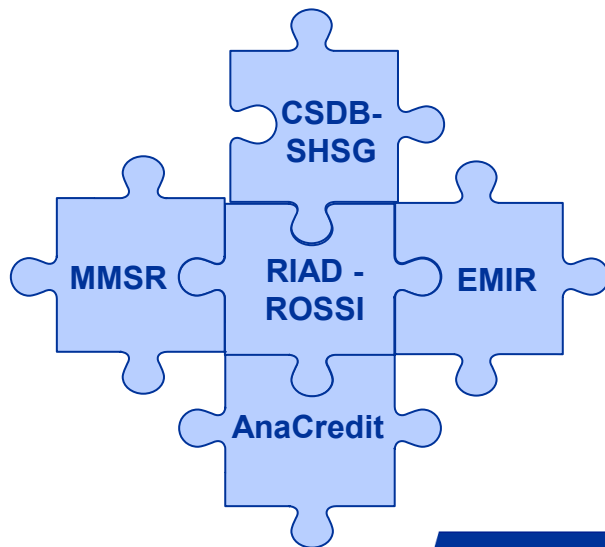
- Data collected by ECB and other European institutions
- Covers around 95% of the asset side of banking groups

The challenge lies in combining these datasets

- Different codelists and dictionaries used
- Different taxonomies (statistical vs supervisory)
- Need for unique, high-quality set of master data

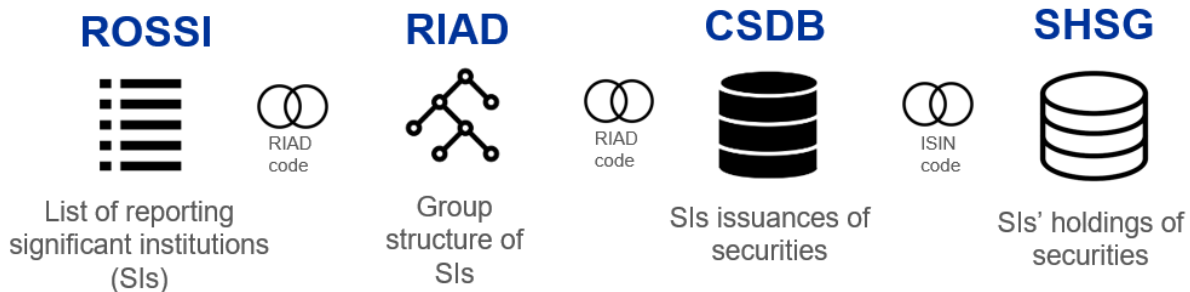
Granular data at the ECB

- For the first time it is possible to integrate and analyse the **interconnections of banking groups** in terms of *securities, loans, money market trades* and *derivatives*.

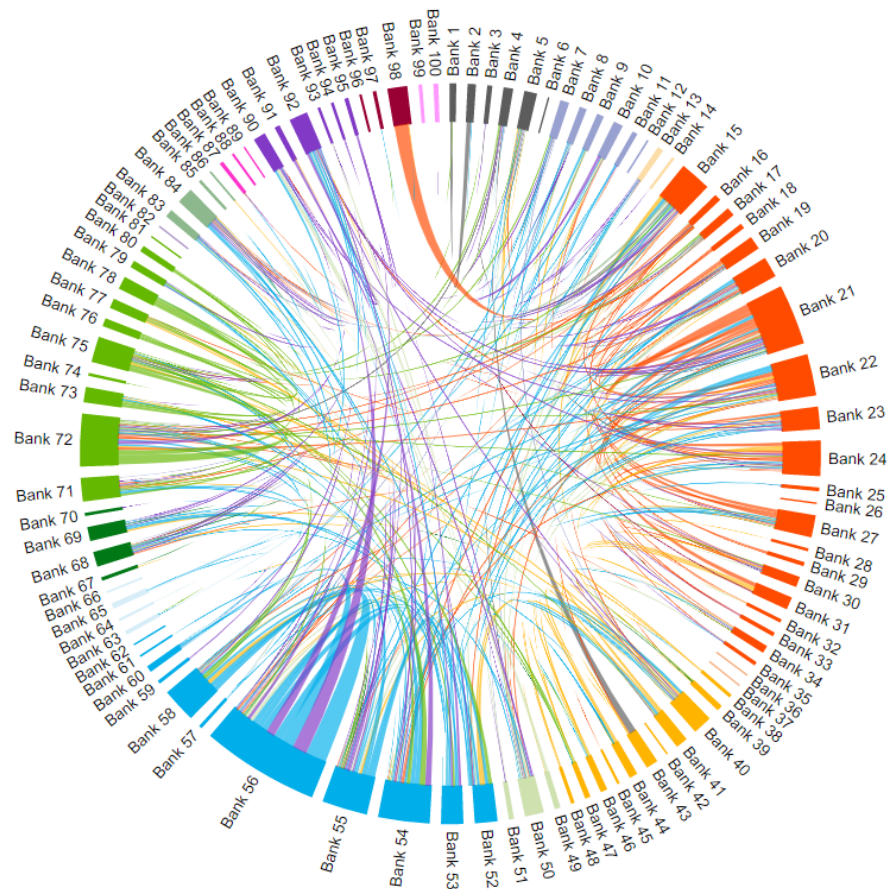


Building a multi-layer network

Integration of data – example of the securities layer



Network



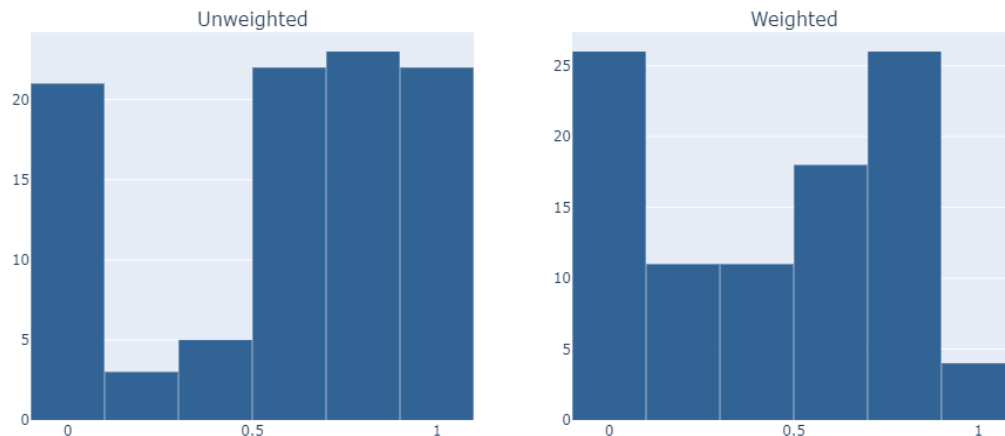
Analysing a multi-layer network

- Network Analytics Toolkit, developed to **analyse multi-layer dynamic networks**.
- Python library with **three modules**:
 - Visualization
 - Topology
 - Shock propagation (early development stage)
- Available as an open source tool within the ECB

Analysing a multi-layer network: multiplex topology

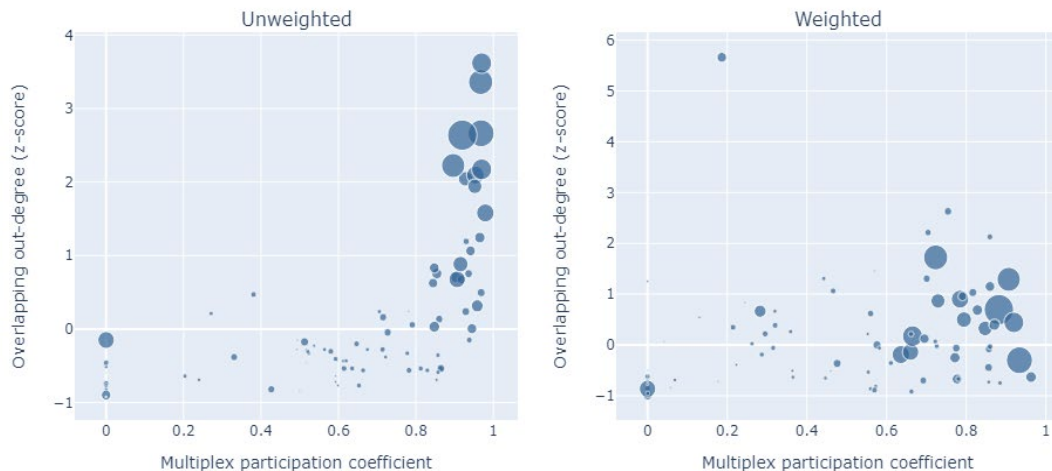
- A high degree node in the multiplex layer can be such because:
 - it **equally participates** in all sub-layers.
 - it is a big hub in **just one layer**.
- The **multiplex participation coefficient** tells if the links of a node are distributed among the sub-layers or are concentrated in just one or few layers.

Multiplex participation coefficient (out-degree) distribution - December 2019



Analysing a multi-layer network: multiplex topology

Overlapping out-degree (z-score) over multiplex participation coefficient - December 2019



- What are the central institutions?
- In the unweighted network, the more layers an institution has exposures in, the more central it is.
- In the weighted network, being active in all or only few markets does not make a big difference in terms of degree centrality.
- Positive correlation between multiplex participation coefficients and total assets.

Conclusion and outlook

- The network opens **new areas of research** not possible before
 - Code and detailed documentation is openly available inside the ECB
 - Can be used as a starting point for data integration projects
- **Important work in the background:**
 - Data quality improvements, harmonising identifiers, building common dictionaries
- Future improvements:
 - Work on **semantic integration** of datasets and **integrated reporting** (IReF project)
 - Adding **new layers** (i.e. SFTDS).



Thank you for your attention!

Contact and Disclaimer

Annalaura Ianiro

Annalaura.Ianiro@ecb.europa.eu



[annalauraianiro](#)

Jörg Reddig

Joerg.Reddig@ecb.europa.eu



[joergreddig](#)

The views expressed are those of the authors only and do not necessarily reflect those of the European Central Bank.

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Single resolution board system for data collection, transformation and analysis¹

Michał Piechocki, Karol Minczyński and Marta Kuczyńska,
Business Reporting - Advisory Group (BR-AG)

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



Single Resolution Board system for data collection, transformation and analysis

IFC-Bank of Italy workshop on "Data
science in central banking"
Applications and tools

Marta Kuczynska

Karol Minczynski

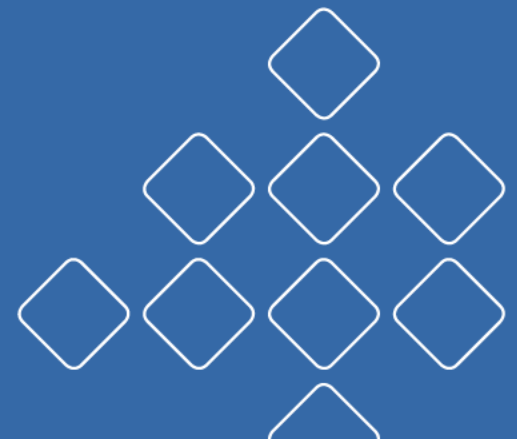
Michal Piechocki



Disclaimer

All views presented are the views of authors only and shall not be considered official views of the Single Resolution Board.

The authors have received permission from the SRB to share the presented information.



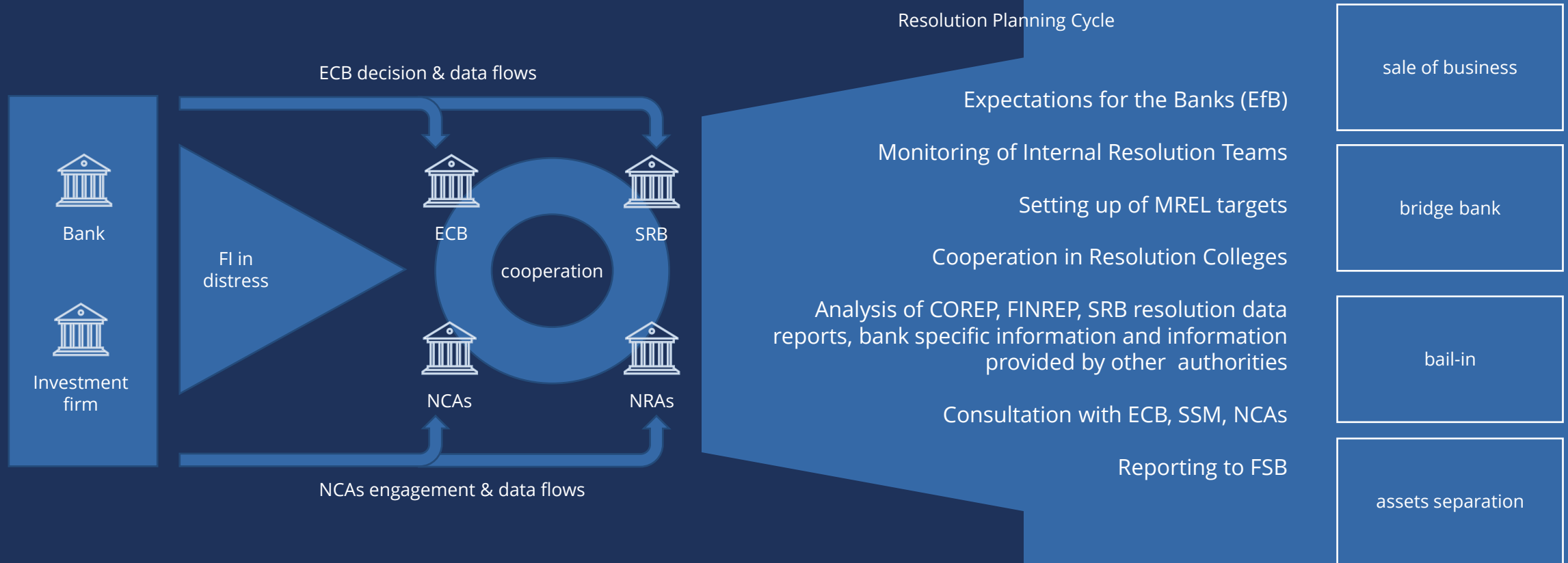
Unveiling the SRB's legal setting

- The SRB was established in and is operational since 2015 by the EU Single Resolution Fund Regulation of 2014
- Primary aim of the SRB is to serve as resolution authority for defaulting financial institutions, as a hub for the Single Resolution Mechanism and as owner of the Single Resolution Fund
- SRB operates in a multi-stakeholder environment including credit institutions, certain investment firms, National Resolution Authorities from 19 eurozone countries plus Croatia and Bulgaria, three European Supervisory Authorities, the ECB, the SSM and the European Commission



The logotype and trademark of the SRB is the intellectual property of the Single Resolution Board@srb.europa.eu

Operational overview



SRB data ecosystem

Structured supervisory data from the ECB/EBA/NCAs:

- Assets Encumbrance
- COREP
- FINREP
- GSII

SRB resolution data reports

- Liability Data Reports, Z and T (part of EBA reporting)
- Additional Liability Data Collection
- Critical Functions Report

Bank specific information

- Structured, semi-structured and unstructured reports developed by Internal Resolution Teams

Market data

- Commercial data providers information

Other data sets that can be taken into account in the future:

- Data provided by other supervisors such as ESMA
- TRACE repositories data based on ISO 20022 standard



Data standards:

- EU Data Point Model used by EBA/ECB/SRB/NCAs for collection and validation of supervisory data
- XBRL used by EBA/ECB/SRB/NCAs for collection and validation of supervisory data
- LEI used by ESAs for identification of counterparties

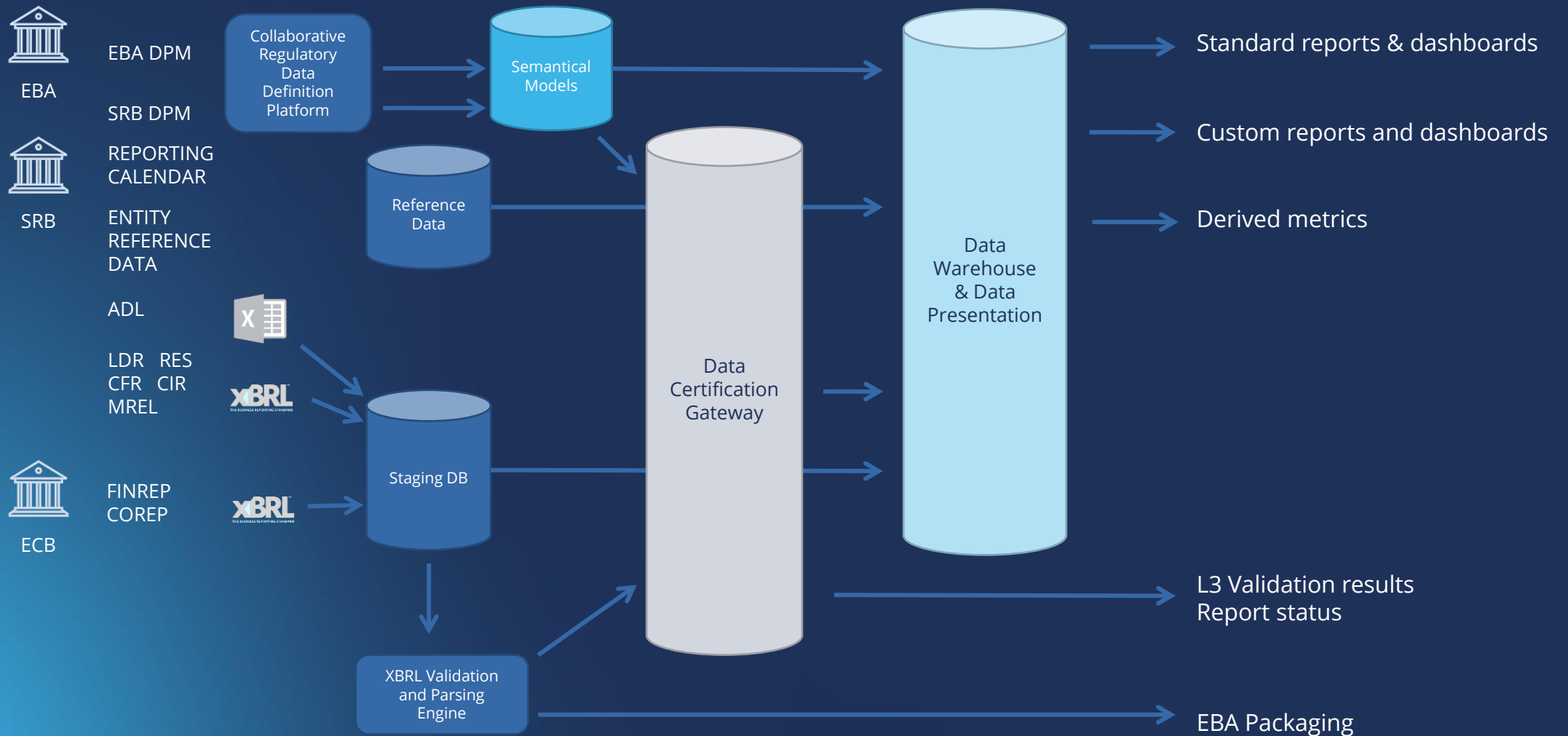
Upcoming and other related developments:

- DPM Refit, a project conducted jointly by the ECB, EBA, EIOPA to prepare DPM for granular data reporting under IREF
- SM-Cube, a methodology combining SDMX-IM and DPM developed and used by the ECB in Statistical Data Dictionary and SSM Wide-data collection database
- ISO 5116, an official ISO Data Point Model standard for which DPM Refit will provide input for update including micro-data definition

Key outputs:

- Resolution plans preparation
- MREL targets definition
- Resolution decisions

Technological architecture



Paving the way for ResTech

- DCG and DWH components are already in production and have successfully processed all FINREP COREP data sets collected by ECB since 2018
- DCG and DWH may in the future be extended to cater for other data sets including market data and to allow analysis of ESAs' granular data sets like TRACE
- According to the SRB Work Programme 2022 the current focus is to allow for automated dashboards and tables in resolution planning, collaboration with NRAs and for exploration of cost-efficient innovative solutions
- SRB will also work on simulations including verification of banks' bail-in simulations
- SRB foresees strengthening of the XBRL-only approach and consideration of more granular data format XBRL-CSV
- SRB is currently conducting its first NLP and ML – based POCs that aim to reduce the time spent by the SBR teams on data processing

“The term ‘ResTech’ refers to the use of innovative technology by resolution authorities to support their work.

The aim is to benefit from innovation by using wise technological solutions, not necessarily involving the purchase of expensive hardware or software.”

SRB Work Programme 2022

Conclusions

- SRB has established a comprehensive metadata model and a **standard-driven platform** for receipt, emphasizing quality assurance, classification and analysis, that is ready to efficiently operate high-volumes of facts
- **Data processed embraces highly-structured** supervisory data that includes both **aggregated and granular data** that may apply to machine-learning models for instance simulating variety of resolution scenarios or impact of specific solutions on creditors, markets and other stakeholders. On the other hand SRB receives **unstructured reports** from banks describing their critical functions that may be subject to NLP algorithms
- The platforms established by the authority are designed with **collaboration principle** including collaboration on new reporting requirements, communication with banks, ESAs, ECB and other stakeholders and finally sharing results of the RPC

1 billion

Number of facts processed by the SRB DCG and DWH platform since production launch



2005-2022 © Business Reporting - Advisory Group

All rights reserved. Any reproduction without written permission from Business Reporting - Advisory Group is prohibited.

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

A network analysis of the JGB repo market¹

Yasufumi Gemma, Takumi Horikawa and Yujiro Matsui,
Bank of Japan

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

A Network Analysis of the JGB Repo Market¹

Takumi Horikawa² Yujiro Matsui³ Yasufumi Gemma⁴

Abstract

In this paper, we examine the characteristics of the Japanese government bond (JGB) repo market by applying network analysis methods to highly granular data on JGB repo transactions. We especially use a measure of "network centrality" which quantitatively identifies financial institutions that play an important role in the transaction network and a "community detection" method which identifies groups of financial institutions that have close transactional relationships with each other. From the results, it was observed that some highly important financial institutions functioned as intermediaries for transactions and that continuous transaction relationships within groups were built around them. These characteristics may contribute to the efficient matching of cash borrowing and lending needs, and to the smooth execution of large-lot transactions. We also conducted some analysis of the behavior of the network structure of the JGB repo market under market stress using the data from March 2020, when the repo rate fluctuated significantly due to the spread of the COVID-19 pandemic. The results of the analysis in this paper indicate the importance of continuously monitoring the functioning of the JGB repo market, and also provide clues for maintaining and improving the functioning and robustness of the market.

JEL Classification: D85, G14, G20, L14

Keywords: Network analysis, Financial markets, Repo transactions, PageRank, Bow-tie decomposition, Community detection

1. Introduction

A repo transaction is a financial transaction in which cash and securities are exchanged for a certain period of time set in a contract. There are two types of repo transactions: general collateral (GC) repo transactions, which do not specify the securities to be traded, and special collateral (SC) repo transactions, which specify the securities to be traded. The former, in general, are used for borrowing or lending cash with securities collateral,⁵ while the latter are used for borrowing or lending specific

¹ The authors are grateful to the conference participants as well as staff members of the Japanese Financial Services Agency and the Bank of Japan for their helpful comments. Any remaining errors are our own. The views expressed in the paper are those of the authors and do not necessarily reflect the official views of the Bank of Japan.

² Financial Markets Department (currently, Personnel and Corporate Affairs Department), Bank of Japan

³ Financial Markets Department (currently, Research and Statistics Department), Bank of Japan

⁴ Financial Markets Department (currently, International Department), Bank of Japan

⁵ "Do not specify the securities to be traded" in GC repo transactions means that the recipient of the securities will accept any securities as long as they are of a somewhat similar quality. Therefore, this is a transaction in which the securities to be delivered are chosen by the security lender.

securities.⁶ Repo transactions are used for a wide range of purposes, including borrowing or lending short-term funds and securities, and these transactions play an important role in the functioning of financial markets.

During the Global Financial Crisis (GFC) in the 2000s, the functioning of repo markets was greatly degraded, which amplified the instability of the financial system. So after the GFC, the G20 and the Financial Stability Board have vigorously pursued efforts to enhance the stability and transparency of repo markets.⁷ As part of these efforts, the Financial Services Agency in Japan (JFSA) and the Bank of Japan (BOJ) jointly started collecting detailed data on individual transaction units for repo markets in Japan from December 2018.⁸ The data are highly granular, namely, include the names of both parties involved in any repo transactions conducted by Japanese financial institutions, those of both the cash borrowing (securities lending) side and the cash lending (securities borrowing) side, as well as such information as the transaction rate and amount of the transaction. This granularity makes it possible to grasp trends in the repo market from a variety of angles.

This paper identifies the network structure of the Japanese government bond (JGB) repo transactions in Japan by taking advantage of these data that include the names of the parties involved in transactions, and then summarizes the characteristics of the network by applying methods of network analysis. Network analysis of financial markets uses a series of analytical methods to understand the structure of networks defined based on the transaction relationships among market participants by visualizing or measuring its characteristics, and to evaluate the robustness and functioning of the entire market. Network analysis of financial markets is widely used, and is especially suitable for research on interbank markets.

The contributions of this paper are as follows. First, for the JGB repo market, we attempted to understand the characteristics of the JGB repo market network by using the "network centrality," which qualitatively identifies the importance of financial institutions in the transaction network and the "community detection" method, which identifies groups of financial institutions that have close transactional relationships with each other. The results show that (i) some highly important financial institutions function as intermediaries for transactions in the market, and (ii) continuous transaction relationships are built around these financial institutions. These characteristics suggest that the JGB repo market is efficient in terms of matching needs of cash borrowing and lending, but may also suggest the need for caution about robustness in the sense that shocks to a few financial institutions that serve as the nexus for many transactions can easily spread throughout the entire repo market.

⁶ SC repo transactions are used, for example, to borrow specific securities with a short position due to short selling to be delivered to the counterparty, or to borrow specific securities to be delivered in bond futures transactions.

⁷ It has been pointed out that repo transactions for some financial products such as securitized products in the U.S. functioned as credit intermediation conducted outside the normal banking system (so-called "shadow banking"), which may have led to the expansion of leverage and excessive risk-taking that were the background factors for the GFC. Against this backdrop, discussions were held on the stability and transparency of the repo market as part of efforts to reduce financial stability risks arising from shadow banking. See Ono *et al.* (2015) for details on the history of the international debate over repo transactions.

⁸ Based on a report by the Financial Stability Board (Financial Stability Board, 2015), efforts are underway to collect similar repo transaction data in each country. See Sasamoto *et al.* (2020) for details on the background of this data collection.

Second, we conducted some analysis of trends of the network structure of the JGB repo market during market stress. Specifically, we examined whether there were any significant changes in the network structure in the period after March 2020, when repo rates fluctuated significantly in response to the spread of the COVID-19 pandemic. As a result, we found from the data that the number of security lenders significantly decreased during the stress period due to certain factors, such as the increased demand for collateral, and that market functioning had deteriorated. On the other hand, the data also suggest that financial institutions were trying to develop new transaction partners during the period.

The analysis in this paper is based on observations of the network structure of the JGB repo market. We did not analyze the factors that contribute to the formation of the network structure, nor did we analyze how the structure of the repo market changes when market stress occurs in detail. However, the results of the analysis in this paper provide some perspectives for monitoring the functioning of the repo market. That is, the results in this paper provide perspectives that should be paid attention to when continuously checking the functioning of the repo market, as well as clues for maintaining and improving the functioning and robustness of the repo market.

The structure of this paper is as follows: Section 2 provides an overview of the data being analyzed and the definition of the network; Section 3 summarizes the characteristics of the network structure of the JGB repo market using network analysis methods; Section 4 provides some analysis of the network structure under market stress. Finally, Section 5 concludes.

2. Data and transaction network

The data used for the analysis in this paper are the transaction information for each repo transaction, which is collected jointly by the JFSA and the BOJ starting as of December 2018.⁹ For each individual transaction with an outstanding balance as of the end of each month, information such as the type of GC/SC repo, the parties to the transaction (both the cash borrowing side and cash lending side), the transaction amount, the transaction rate (interest rate), the type of securities involved in the transaction, and the transaction period is recorded. The top 50 financial institutions in terms of transaction volume report the transactions in which they are either the cash borrowing side or the cash lending side; these constitute more than 90% of the total amount of repo transactions in the market. Although repo transactions relating to equities and other securities are also reported, we analyzed only transactions of JGBs in this paper.¹⁰ Including the transactions in which only one party is among the reporting institutions whereas the counterparty is not, the data capture repo

⁹ There are two types of repo transactions: repurchase agreements (transactions in which parties sell securities with a special agreement to buy them back in the future) and securities lending transactions (transactions in which securities are loaned in exchange for cash or other securities collateral). The data reported jointly by the JFSA and the BOJ cover both. In this paper, we do not distinguish between the two, referring to both as repo transactions.

¹⁰ The analysis covers reported transactions in which the type of securities to be traded is "government-issued bonds," the type of currency is "yen-denominated," and the rating is "investment grade." Although these definitions may include bonds other than JGBs, the majority of transactions are considered to be for JGBs, so we refer to all the transactions under analysis as JGB repo transactions.

transactions conducted by about 170 financial institutions as either the cash borrowing side or cash lending side of the transaction.^{11,12}

By using these data, which include the names of both parties, it is possible to identify the transaction network structure of the JGB repo market. The network structure is represented as a data structure consisting of points or "nodes" and lines or "links" connecting two nodes.¹³ Representations that do not distinguish the direction of links are called "undirected networks" and those that do are called "directed networks." In this paper, we consider a network structure in which financial institutions are nodes and bilateral transaction relationships are directed links: the case where financial institution A's amount of cash lending to financial institution B exceeds its amount of cash borrowing (i.e., A is "net cash lending" to B) is represented by the directed link "A→B."^{14,15} We consider GC repos and SC repos separately because they have different transaction purposes, i.e., whether they are used for the purpose of borrowing/lending cash or for the purpose of borrowing/lending specific securities. For example, Graph 1 shows the network structure of the GC repo market and the SC repo market as of the end of September 2019. This shows that transactions between financial institutions intersect in complex ways, making it difficult to capture the characteristics of the network at first glance. Network analysis can reveal the characteristics of such a complex network structure by visualizing and measuring the number of transaction partners (the number of links to other nodes) of each financial institution and the relative importance of each financial institution in the network.¹⁶

¹¹ Non-reporting institutions are lumped together by type of business and region of residence and treated as a single transaction party because the individual company name of non-reporting institutions cannot be identified in the data. Therefore, the actual number of financial institutions is larger than this figure. The number of non-reporting institutions includes some business corporations (non-financial institutions), but since these make up only a small number of the nodes in the repo market network, all nodes are referred to as "financial institutions" in this paper.

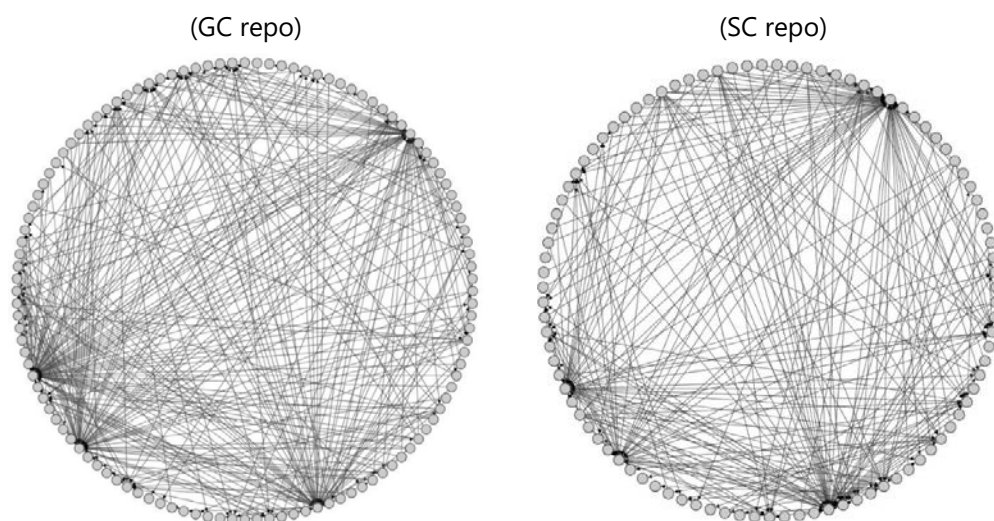
¹² Transactions in which both parties to the transaction are reporting institutions are double reported. In this analysis, we adjust the data for such identical transactions to avoid double counting.

¹³ This network structure is sometimes referred to as a "graph structure," the nodes as "vertexes," and the links as "edges."

¹⁴ There are two transaction schemes for GC repo transactions: the "Subsequent Collateral Allocation Repo Transactions" scheme, which was introduced at the same time that the JGB settlement period was shortened in May 2018, and the traditional "Standard Repo Transactions" scheme. See Fujimoto *et al.* (2019) for details on the differences between the transaction schemes. The former is excluded from the analysis in this paper because the market volume is currently limited compared to the latter, and the risk properties are different from those of the latter, since clearing houses always clear the claims and obligations related to transactions.

¹⁵ In this paper, we use the amount of net cash lending on an aggregate basis, which does not distinguish between different transaction periods, collateral bond issues, and other transaction details. It is conceivable that the network structure may differ across transaction periods and collateral bond issues. Thus, examining the characteristics of these factors remains a matter for future study.

¹⁶ In the analysis, functions of "igraph", a package for the statistical software R, were used as necessary.



¹ Based on transaction relationships as of the end of September 2019.

Source: Data on securities financing transactions in Japan

There have been many previous studies on network analysis of financial markets in various countries. Since it is generally difficult to obtain information on individual financial transactions, most of these studies focused on interbank cash lending and borrowing transactions based on settlement data held by central banks (Inaoka *et al.* (2004), Imakubo and Soejima (2010) [Japan], Furfine (2003), Afonso *et al.* (2013), Soramäki *et al.* (2007) [US], Abbassi *et al.* (2013), Allen *et al.* (2020) [Euro area], Bargigli *et al.* (2015), Iazzetta and Manna (2009), Mistrulli (2011) [Italy], Martínez-Jaramillo *et al.* (2014) [Mexico]). In recent years, however, with the accumulation of transaction data through electronic platforms and other means, there have been some studies covering transactions of corporate bonds (Di Maggio *et al.* (2017) [US]), equities (Di Maggio *et al.* (2019) [US]), municipal bonds (Li and Schürhoff (2019) [US]), securitized products (Hollifield *et al.* (2017) [US]), OTC derivatives (Bardoscia *et al.* (2019) [UK]), and CDS (Markose *et al.* (2012) [US]). In the next and following sections, we attempt to summarize the characteristics of the JGB repo market while referring to the discussions in the previous studies.

3. Characteristics of the Network Structure of the JGB Repo Market

(1) Overview of JGB Repo Market by Network Statistics

Prior studies on network analysis of financial markets have proposed methods to understand the market structure using "network statistics" that measure the characteristics of the network structure. In this section, we summarize the characteristics of the JGB repo market network by focusing on two representative network statistics, "degree" and "shortest distance."

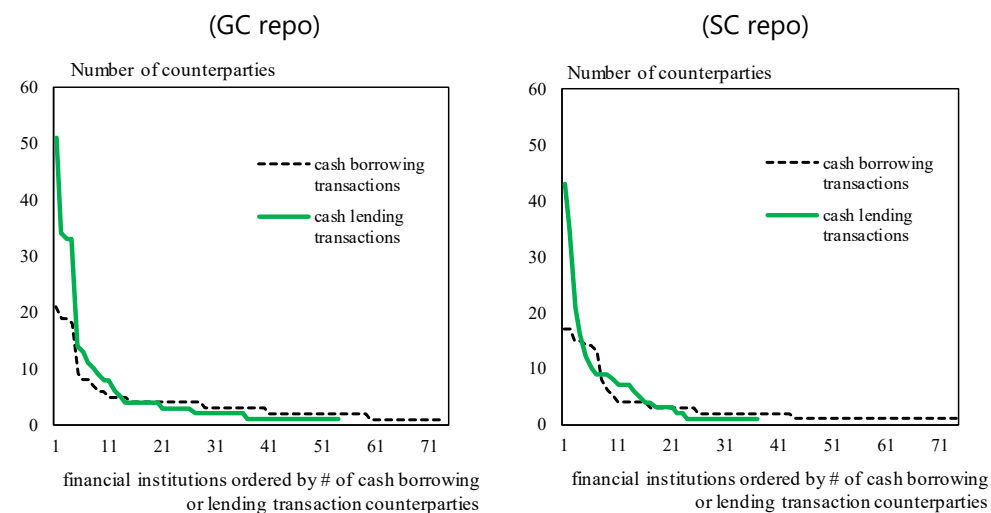
Degree

The "degree" statistic is a measure of the number of counterparties of each financial institution. There are two types of transactions for each financial institution in the repo market: those in which the financial institution borrows cash and those in which it lends cash. Therefore, we distinguish the number of counterparties for cash borrowing transactions as "in-degree" and the number of counterparties for cash lending transactions as "out-degree."

Graph 2 shows the distribution of the number of counterparties for cash borrowing transaction (in-degree) and that of cash lending transaction (out-degree) for each financial institution in the repo market. In Graph 2, the financial institutions are arranged from left to right in the order of the number of counterparties, and the number of counterparties for each financial institution either for cash borrowing or lending transactions is plotted on the vertical axis.

Distribution of the number of counterparties (degree)

Graph 2



¹ Based on the transaction network as of the end of September 2019.

Source: Data on securities financing transactions in Japan

In the left-hand graph for GC repo transactions, some financial institutions have a large number of counterparties in both cash borrowing transactions and cash lending transactions, while other financial institutions have only a small number of counterparties. In the case of cash borrowing transactions, only four financial institutions have more than 10 counterparties, and the majority of the others have an even more limited number, less than five, of counterparties (although it is not clear from this graph, their transactions are concentrated on a few institutions). This tendency is also observed for the SC repo transactions (the right-hand graph). Thus, in the repo market, transactions are concentrated on a small number of financial institutions that act as hubs in the network, whereas a large number of the other entities conduct most of their transactions with those hubs. The characteristic of the network represented by this degree distribution is called the "long-tail characteristic"

and this characteristic of financial market networks has been observed in many previous studies.¹⁷

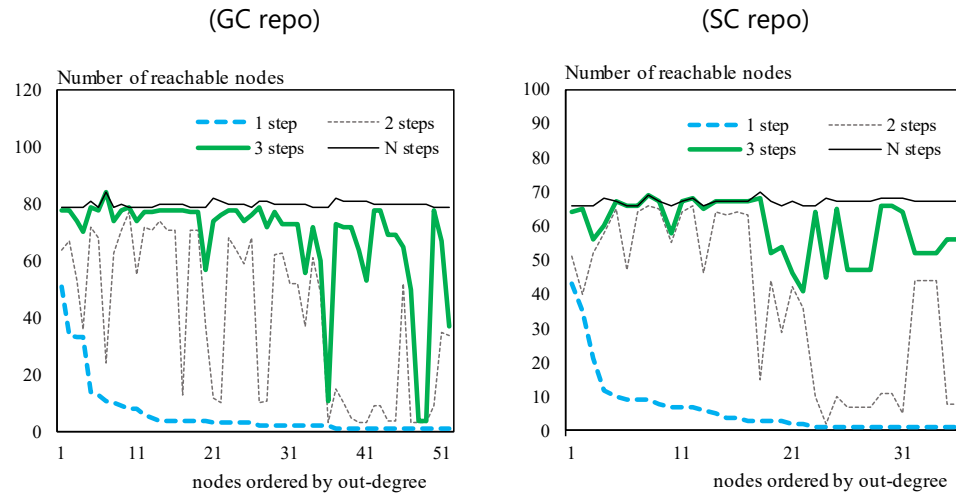
Looking at the top few financial institutions in terms of the number of counterparties, there are more counterparties of cash lending transactions than those of cash borrowing transactions in both the GC repo market and the SC repo market. This indicates that some entities are acting as large cash lenders with a large number of cash borrowing counterparties (asymmetry between cash borrowing and lending transactions).

Shortest distance and reachable region

The length of the shortest path connecting any two financial institutions is called the "shortest distance," and the number of other financial institutions that can be reached from one financial institution in the shortest distance n is called the "n-step reachable region." This indicator shows how close the trade relationships linking financial institutions in a network structure are.

Graph 3 plots the reachable region for each financial institution in each of the GC and SC repo markets using actual transaction data as of the end of September 2019. The results show that for most financial institutions, the region that can be reached in three steps (thick solid green line) is close to the maximum region of N steps (thin solid black line). In other words, most of the financial institutions can reach a significant number of other financial institutions in only three steps. This characteristic has also been shown in many previous studies and is called the "small-world characteristic." It is a characteristic that although many financial institutions in the network are not directly connected to each other, most of the financial institutions are indirectly connected through a few financial institutions. Such a structure may contribute to the efficiency of transactions in the repo market if, for example, an entity with needs of cash borrowing can connect with another counterparty with needs of cash lending through a small number of intermediaries. On the other hand, the fact that the entire network is connected over a short distance suggests that, relatively speaking, shocks that occur in one part of the network are likely to spread throughout the entire network.

¹⁷ Such a distribution is also said to follow a power law, in which the probability distribution is expressed as $p(k) = ak^{-\gamma}$. A distribution that follows a power law is also said to have the "scale-free property" because there is no scale, such as the average, in the number of nodes with an arbitrary number of links (Barabási and Albert, 1999). This characteristic has been reported not only for the degree distribution but also for the distribution of transaction amounts. The scale-free property of the degree distribution can be confirmed by examining whether a linear relationship is observed when the degree and its cumulative distribution are plotted on a two-sided logarithmic graph. In fact, for four cases (out-degree or in-degree for GC repos, and out-degree or in-degree for SC repos), linear relationships indicating distributions that follow a power law have been confirmed.



¹ Based on the transaction network as of the end of September 2019.

Source: Data on securities financing transactions in Japan

So far, we have used the typical network statistics, "degree" and "shortest distance," to examine the characteristics of the repo market.¹⁸ However, these are not enough to explore the efficiency and robustness of the market in depth. Therefore, in the next section, focusing on individual financial institutions and their transaction relationships with each other, we examine which financial institutions play important roles in the entire network, what roles the important financial institutions play in the network structure, and whether there are continuities in the transaction relationships among financial institutions.

(2) Network structure in terms of centrality and community

If financial institutions that deal with a large number of counterparties and play a central role in the market mediate needs of cash borrowing and cash lending in the market, or if continuous transaction relationships are built around such intermediaries, this can lead to more efficient market transactions through prompt matching of cash borrowing and lending needs (Li and Schürhoff, 2019). On the other hand, as was the case in the repo market during the GFC in the 2000s, if financial institutions that deal with many transactions are hit by negative shocks, market participants' actions to stop or reduce their transaction activities, due to concerns about counterparty risk (including failures to deliver securities), may spread through the network, leading to a decline in the market functioning and liquidity. Therefore, in monitoring the market functioning and liquidity, it is useful to identify financial

¹⁸ Prior studies have also examined other characteristics of financial market networks using network statistics such as "clustering coefficients," which measure how closely transaction partners for each financial institution are transacting with each other (see the survey by Iori and Mantegna (2018)). When we calculate the same network statistics for the JGB repo market, although we do not show it graphically here, we can identify features that are common to other financial market networks shown in previous studies.

institutions that play a central role in the repo transaction network and communities that have established close transaction relationships.

In this section, we examine the characteristics of the JGB repo market using a measure of "network centrality," which quantifies the importance of each financial institution -- the extent to which it plays a central role in the network -- based on its relationships with its counterparties, and a "community detection" method, which identifies groups of closely connected financial institutions.

Importance of financial institutions by PageRank centrality

We measure the importance of each financial institution on the repo transaction network by using "PageRank" as a measure of network centrality.¹⁹ PageRank was originally developed as a measure of the importance of web pages on the Internet (Brin and Page, 1998), but has also attracted attention as a measure of systemic risk in financial markets (Allen *et al.*, 2020, Yun *et al.*, 2019). The importance of a web page as measured by PageRank is higher (i) the more web pages are referring to it as well as (ii) the higher the importance of web pages referring to it. Since this importance depends not only on one's own status but also on the importance of the other parties, PageRank is suitable for measuring the degree to which individual financial institutions influence the overall network structure. When this measure is applied to the network of repo transactions, the measure of a financial institution is higher (i) the larger its amount of cash borrowing as well as (ii) the larger the amount of cash borrowing of its cash borrowing transaction counterparties. That is, it is an indicator for measuring how much a financial institution influences the entire transaction network, focusing on cash borrowing transactions.²⁰

It should be noted that we can also consider a measure that focuses on cash lending transactions rather than cash borrowing ones; such a measure of a financial institution is higher (i) the larger its amount of cash lending and (ii) the larger the amount of cash lending of its cash lending transaction counterparties. In this regard, Saltoglu and Yenilmez (2015) and Kaltwasser and Spelta (2019) use two types of PageRank, that is, (i) "borrower PageRank" that measures the importance of each financial institution in cash borrowing transactions and (ii) "lender PageRank," which measures the importance of each financial institution in cash lending transactions.²¹

¹⁹ Other indices such as "eigenvector centrality" have been proposed as measures of network centrality. For the present study, we selected PageRank as the most suitable index to capture the degree of influence of each node on the entire directed network.

²⁰ PageRank for financial institution i is calculated as follows:

$$PageRank_i = (1 - \alpha) + \alpha \frac{\sum_{j \in M} w_{ij} PageRank_j}{\sum_{z \in N} w_{jz}}$$

w_{ij} is the net lending amount of node j to node i , M is all nodes connected to node i , and N is all nodes in the network. α is called the damping factor, which is a device to measure the importance of each node based on the overall transaction relationship even if the network is not fully connected. Because PageRank of each institution also depends on the PageRank values of its transaction partners, it is generally calculated iteratively. The damping factor affects the convergence speed of the solution, and it is considered most efficient to set it to 0.85 by Brin and Page (1998), which many subsequent studies have followed. In the present analysis, it was also set to 0.85.

²¹ Calculated by replacing w_{ji} with w_{ij} and w_{jz} with w_{zj} in the equation of PageRank in the previous footnote. In other words, the PageRank of node i is determined by the directed link from node i to node j (cash borrowing transaction), not by the directed link from node j to node i (cash lending transaction).

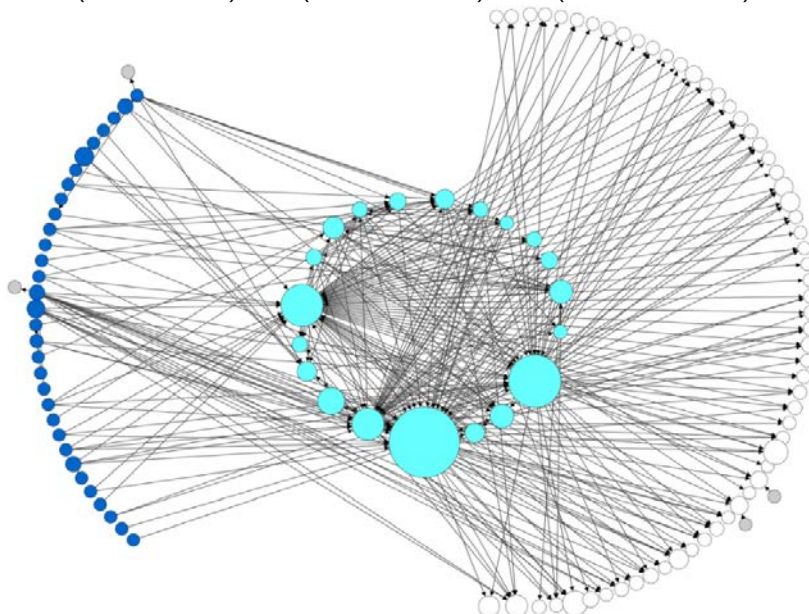
This paper also adopts their approach and calculates two types of PageRank in terms of cash borrowing and cash lending.

In addition, to examine the role of important financial institutions as measured by PageRank in the repo transaction network, we identify where each financial institution is located in the hierarchical structure based on the flow of cash in the repo market (i.e., cash lender, intermediary, and cash borrower tiers) using the bow-tie decomposition algorithm (Yang *et al.*, 2011).²²

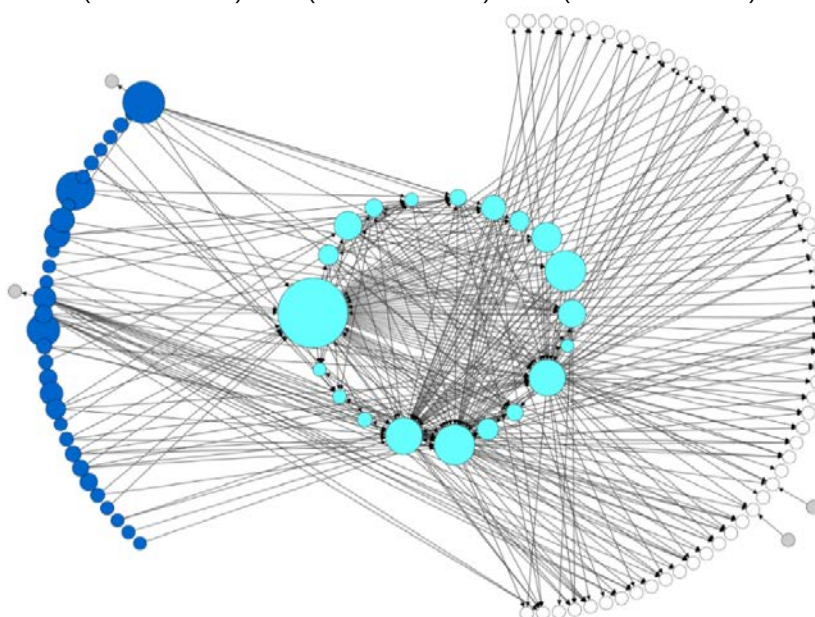
Graph 4 illustrates, for each financial institution, its position on the hierarchical structure in the repo market network and its PageRank. Panels (1) and (2) represent the results of analysis of the GC repo market as sizes of nodes for the borrower PageRank and the lender PageRank, respectively, and panels (3) and (4) represent those for the SC repo market. These results show that, for both cash borrowing transactions and cash lending transactions, the middle tier (light blue) includes highly important financial institutions, which indicates these highly important financial institutions play the role of intermediaries between the final cash borrowers and lenders. It should be noted that, looking at the cash lending transactions in the GC market, the PageRank of the financial institutions in the cash lender tier (dark blue) vary; in particular, some of these institutions are lending large amounts of cash (Graph 4(2)). In addition, looking at the SC market, in terms of cash borrowing (i.e., security lending), not only intermediaries but also those in the cash borrower tier (white) (i.e., security lender tier) have high PageRank (Graph 4(3)). This indicates that some financial institutions play an important role as security suppliers in the SC repo market, which aims to lend and borrow specific securities.

²² The bow-tie decomposition uses the following algorithm to classify financial institutions mainly into three tiers. (i) The set of financial institutions in which every financial institution is reachable through transaction relationships from every other financial institution is defined as the middle tier (intermediary tier). (ii) The set of financial institutions in which the financial institutions are not included in the middle tier but have transactions that allow cash to flow to the middle tier financial institutions is defined as the upstream tier (cash lender tier). (iii) Among the financial institutions not included in the middle tier, the set of financial institutions that engage in transactions through which cash can flow from middle tier financial institutions is defined as the downstream tier (cash borrower tier).

(1) GC repo transaction network structure and borrower PageRank
(cash lenders) \Rightarrow (Intermediaries) \Rightarrow (cash borrowers)



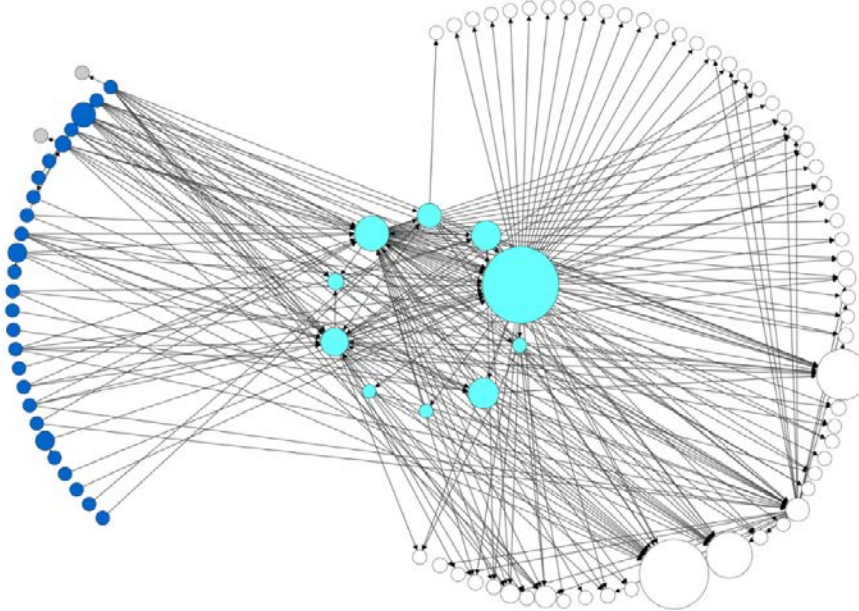
(2) GC repo transaction network structure and lender PageRank
(cash lenders) \Rightarrow (Intermediaries) \Rightarrow (cash borrowers)



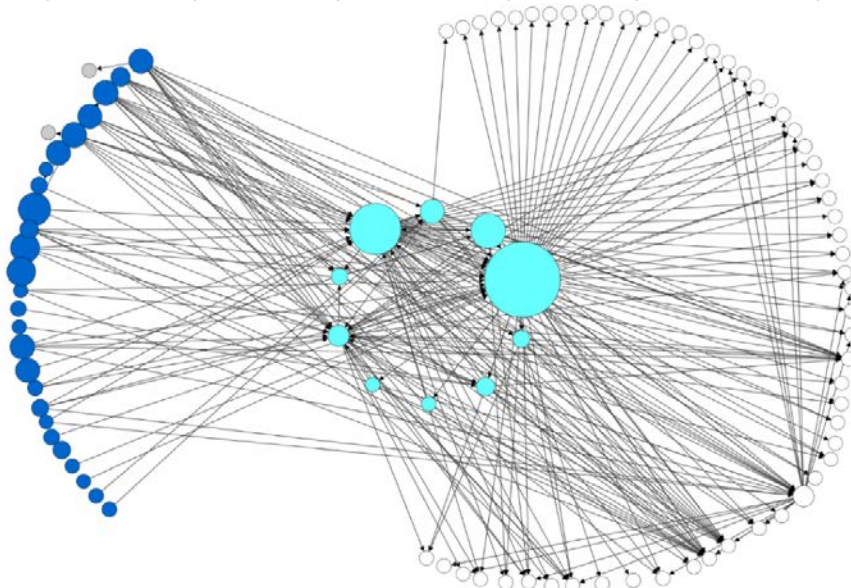
¹ From left to right, financial institutions are classified into three groups: cash lenders (blue), intermediaries (light blue), and cash borrowers (white). The size of the node in (1) corresponds to the borrower PageRank, and the size of the node in (2) corresponds to the lender PageRank. Based on transactions as of the end of September 2019. Financial institutions are arranged in the same positions in (1) and (2).

Source: Data on securities financing transactions in Japan

(3) SC repo transaction network structure and borrower PageRank
 (cash lenders) \Rightarrow (Intermediaries) \Rightarrow (cash borrowers)



(4) SC repo transaction network structure and lender PageRank
 (cash lenders) \Rightarrow (Intermediaries) \Rightarrow (cash borrowers)



¹ From left to right, financial institutions are classified into three groups: cash lenders (blue), intermediaries (light blue), and cash borrowers (white). The size of the node in (3) corresponds to borrower PageRank, and the size of the node in (4) corresponds to the lender PageRank. Based on transactions as of the end of September 2019. Financial institutions are arranged in the same positions in (3) and (4).

Source: Data on securities financing transactions in Japan

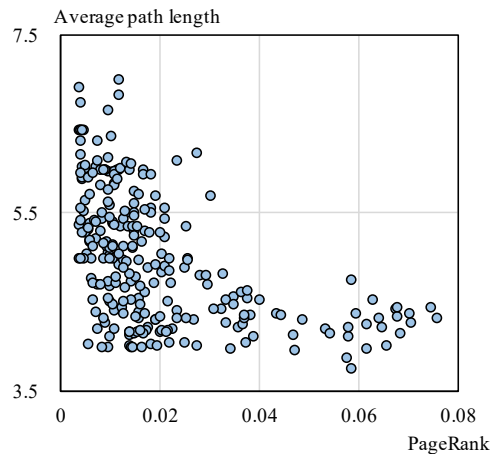
In addition, we perform some analysis of how the circumstance where financial institutions with relatively high importance play the role of intermediaries contributes to the efficiency of the market. If an intermediary connects cash lenders and borrowers through shorter paths, it is considered that the intermediary contributes to the efficient matching of cash lending needs and cash borrowing needs (Li and

Schürhoff, 2019). From the perspective of testing this point for the repo market, we consider the relationship between (i) the importance of the intermediaries and (ii) the average path length of the transaction paths in which they are involved as intermediaries. In detail, for each financial institution in the intermediary tier, (ii) is measured as the average path length of shortest paths connecting pairs of nodes (one each in the cash lender and borrower tiers) that include the intermediary institution. Although the transactions on the shortest path connecting the final cash lenders and cash borrowers are not necessarily the transactions matched by intermediaries, by assuming the transaction path in which it is involved in the intermediation, we treat the indicator in (ii) as a proxy variable to measure the matching power of the intermediary.²³

Graph 5 plots the relationship between (i) and (ii). These results show that intermediaries with high importance tend to connect the final cash lenders and cash borrowers through shorter paths. This suggests that intermediaries with high importance in the network may contribute to the efficient matching of cash borrowing and lending needs in that they connect the final cash lender and borrower through shorter paths.

Relationship between PageRank and the average shortest path length of transactions mediated by the intermediaries

Graph 5



¹ The vertical axis is the average path length of transactions in which each financial institution in the intermediary tier is involved in intermediation. The horizontal axis is the average of the "borrower PageRank" and "lender PageRank" of each financial institution in the intermediary tier, for GC repo transactions only.

Source: Data on securities financing transactions in Japan

Community detection and continuity of transaction relationships

One class of the network analysis methods is "community detection," which extracts groups of closely connected nodes (communities) that have many internal connections and relatively few external connections. In this paper, we apply one of

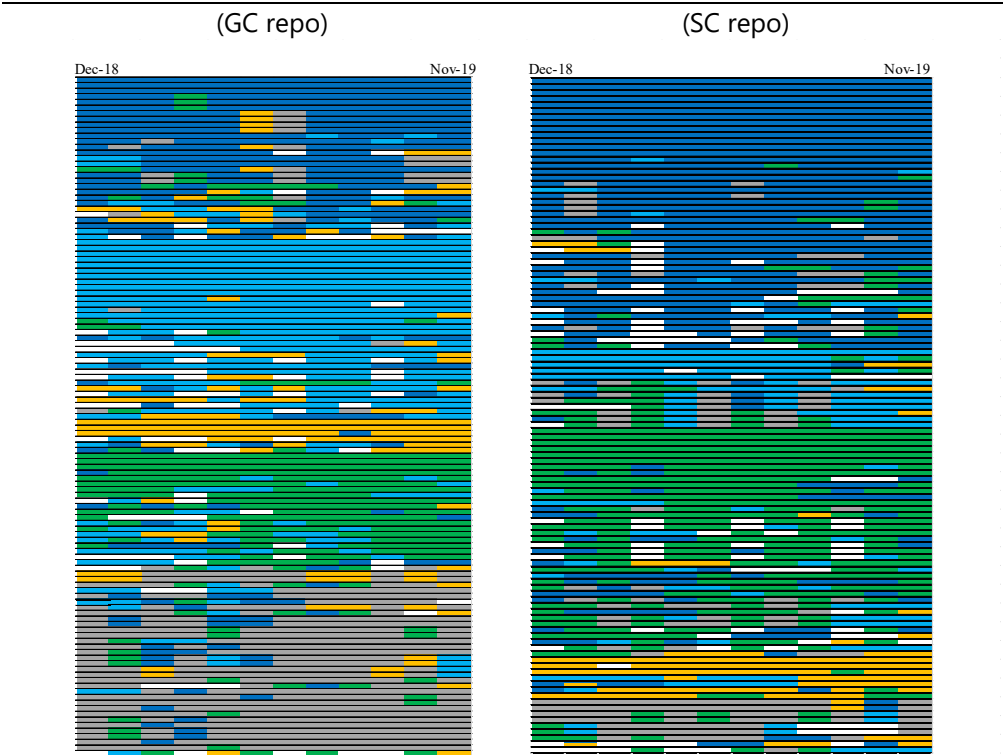
²³ Li and Schürhoff (2019) use daily data and CUSIP codes (unique identification numbers assigned to registered securities) to specify the actual process of specific bonds circulating through intermediaries and analyze the matching power of intermediaries. Since the data in this analysis can only capture transactions that are outstanding at the end of the month and do not fully identify bonds of particular issues, they cannot match the lender and borrower as precisely as Li and Schürhoff, so we adopt the treatment described here instead.

these methods, the spin glass method (Reichardt and Bornholdt, 2006), to analyze the characteristics of the community structure of the repo market.²⁴

The results are shown in Graph 6, where the detected communities are identified by their central intermediary and color-coded. Specifically, financial institutions are lined up vertically and are color-coded according to the community to which they belong. The horizontal direction is the time-series direction. In Graph 6, many of the financial institutions have the same color in the horizontal direction, indicating that many of them have continued to conduct transactions within the same community in terms of community formation in the transaction network. Looking at the frequency with which financial institutions move between communities, we observe that about half of them do not move (Table 1).

Detected communities

Graph 6



¹ GC repo (left) and SC repo (right) transactions. Financial institutions are lined up vertically, and the horizontal direction represents the time series, with color-coded communities to which each financial institution belongs at each point in time. Blue, light blue, green, and orange represent communities formed around major intermediaries, and gray summarizes other smaller communities. White indicates that the financial institution had no transactions. Based on the transaction network at the end of each month from Dec. 2018 to Nov. 2019.

Sources: Data on securities financing transactions in Japan

²⁴ In the spin-glass method, nodes are grouped by community so that there are more links inside the community and fewer links outside the community. Specifically, we consider a score determined as follows. The score is higher if (i) there is a link between any two nodes that belong to the same community or (ii) if there is no link between any two nodes that belong to different communities, while the scores are lower if (iii) there is no link between any two nodes that belong to the same community or (iv) there is a link between any two nodes that belong to different communities. We decompose each node into communities by searching for the grouping of nodes that maximizes the score.

Frequency of community transitions Table 1

| | GC repo | SC repo |
|---------------------|---------|---------|
| from previous month | | |
| to current month: | | |
| Community unchanged | 48% | 54% |
| Community moves | 27% | 26% |
| Other | 25% | 20% |

¹ "Other" includes cases where the transaction status changed from "having no transactions" in the previous month to "having transactions" in current month, and vice versa. Based on the transaction network at the end of each month from Dec. 2018 to Nov. 2019.

Source: Data on securities financing transactions in Japan

In order to find out what role the existence of such a community structure plays as a market function, we attempt a regression analysis of the background factors of the community structure. Specifically, we conduct a probit regression with a dummy variable indicating whether or not each transaction is conducted within the same community (transaction within the same community = 1) as the explained variable.²⁵ The explanatory variables are the transaction rate and transaction amount.^{26, 27}

The results of the probit regression analysis are shown in Table 2. While the coefficient on the transaction rate is not significant, the coefficient on the transaction amount is significantly positive. This indicates that there is a positive relationship between the size of the transaction amount and the probability of the transaction being conducted within the community, indicating that transaction communities may be formed for the purpose of facilitating large-lot transactions.

²⁵ For transactions in which the face value of the underlying bond exceeds 5 billion JPY, there is a market practice to split the transaction up into approximately 5 billion JPY units for execution in order to facilitate settlement. In fact, a histogram of transaction value (face value multiplied by market value) shows that the frequency of transactions around 5 billion JPY to 5.2 billion JPY is notably large, suggesting that split transactions are distributed in this range. Although it is not possible from the data to identify whether a transaction is split from a larger original transaction, we have included multiple transactions whose transaction amounts are in this range and that have the same transaction terms (names of parties, repo rates, and transaction period) as a single transaction with the combined transaction amount.

²⁶ We limited the transaction data used in the estimation to overnight GC repo transactions in order to avoid the effects of differences in transaction period or types of bonds. We also excluded transactions with repo rates of 0% or higher from the data sample for this analysis because they deviate significantly from prevailing short-term interest rates and are considered to be special transactions with nonstandard transaction conditions.

²⁷ For the transaction rate, the deviation from the Tokyo Repo Rate is used. For the Tokyo Repo Rate the tomorrow-next rate with the last day of the month as the execution date is used in order to correspond to the "Standard Repo Transaction." For the transaction rate, we used residuals from the regression analysis with transaction amounts and dummy variables representing the individual financial institutions, considering the possibility of multicollinearity between the transaction rate and the transaction amount. In the results of the regression analysis for the residuals, the coefficient for the transaction amount was significantly negative. This is consistent with the market's view that small transaction amounts tend to increase transaction rates because operation costs are relatively high.

Results of probit regression analysis

Table 2

| Dependent variable: Dummy variable (transactions within the same community=1) | |
|--|--------------------------|
| repo rate | -0.15115 (0.20600) |
| amount (logarithm) | 0.07284*** (0.00389) |
| constant | -1.49711*** (0.08786) |
| time dummies | yes |
| sample size | 20,850 |
| transaction within the same community | 7,989 |
| transaction between different communities | 12,861 |
| Pseudo-R2 (McFadden) | 0.07706 |

¹ *** denotes significance at the 1% level. Figures in parentheses are standard deviations. Robust (heteroskedasticity-adjusted by Huber-White's method) standard errors were used. The sample consists of O/N GC repo transactions with outstanding balances as of the end of each month from Dec. 2018 to Nov. 2019.

Source: Data on securities financing transactions in Japan

In summary, we found that in the repo market transaction network, important financial institutions play the role of intermediaries and continuous transaction relationships are established as communities. In addition, evidence suggested that the repo transaction network may have a structure that contributes to efficient transactions.

On the other hand, some previous studies have pointed out the vulnerability of financial market networks with characteristics similar to those mentioned above, because shocks that occur to highly important financial institutions tend to propagate throughout the network.²⁸ The argument has also been made that the stability of the entire network can be reduced if it is difficult to trade between different communities because of the cost and time required to change transaction partners.²⁹ In light of these arguments, it can be said that the repo market has a structure that is conducive to efficient transactions, but it also needs to be careful about its robustness. Although there is a trade-off between the efficiency and robustness of the network structure and it is difficult to theoretically determine the optimal balance between the two, it is important to understand which aspects of the market structure may lead to a decline in robustness. Therefore, quantitatively "visualizing" the structure of the repo market and understanding its characteristics through the analysis presented in this section

²⁸ Caballero (2015) and Minoiu *et al.* (2015) point out that closely connected networks are more likely to amplify the impact of any negative shocks, increasing the probability of systemic risk; Yun *et al.* (2019) and Bardoscia *et al.* (2019) point out that the presence of important nodes is likely to diffuse shocks.

²⁹ Regarding the community structure of the network, Dong *et al.* (2018) show that when there are few links crossing different communities, the robustness of the overall network is reduced as a result of the tendency of network fragmentation.

are considered to be important for monitoring market functioning. It may also be helpful for individual market participants in considering the efficiency and stability of their transactions.

4. Changes in Network Structure under Market Stress

Prior research has reported that the network structure and related indicators behave in a characteristic manner before and after a stress event in financial markets. For example, it has been reported that the number of links in the transaction network decreases and the exposure per counterparty increases in the interbank market during stress, due to movement to narrow counterparties with an awareness of counterparty risk (Beltran *et al.*, 2015, see also Minoiu and Reyes, 2013). It has also been reported that in the transaction network of the JGB market, when interest rates rise sharply, the financial institutions that form the core of the network actively search for new bond buyers, resulting in a sharp increase in transactions between the core financial institutions and financial institutions that normally have few transaction relationships with them (Sakiyama and Yamada, 2016). Others have reported that network structure changed significantly during stress events (in't Veld and van Lelyveld, 2014, Fricke and Lux, 2015).

In this section, we examine the behavior of the network structure of the JGB repo market during market stress. Although we cannot conduct a precise quantitative analysis of the repo market because the data starting period is just from December 2018 and there is not much data available, we attempt to conduct some verification using data from March 2020 onwards, when the repo rate fluctuated significantly following the spread of the COVID-19 pandemic.

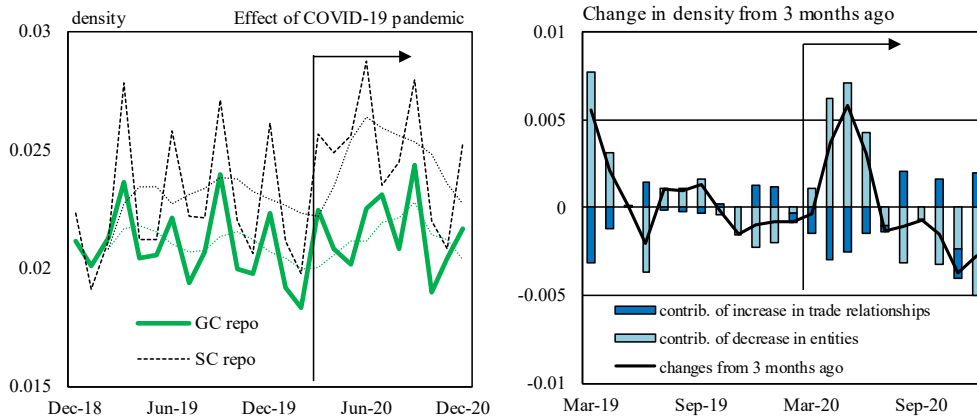
Graph 7 plots the time series of "network density," which represents how transactions are actively conducted in the network, for the transaction networks of GC repos and SC repos. The density of the network is calculated as the ratio of the actual number of transaction relationships to the number of all possible transaction relationships between financial institutions in the network (the number of transaction relationships if all financial institutions transact with each other).³⁰ Considering the seasonality of the spike at the end of the quarter, we can see that the density has remained high since March 2020, when the impact of the spread of the COVID-19 pandemic began (Graph 7, left). Decomposing the three-month change in density (Graph 7, right), we find that while the number of counterparties actually conducting transactions decreased, contributing to a partial reduction in density (dark blue area), the number of financial institutions participating in the repo market decreased (light blue area) and the density has increased in total.³¹

³⁰ Density d of a directed network is calculated by $d = \frac{k}{n(n-1)}$, where the number of transaction relationships is k and the number of financial institutions is n .

³¹ The decomposition was calculated by $\Delta d = \frac{\partial d}{\partial k} \Delta k + \frac{\partial d}{\partial n} \Delta n = \frac{1}{n(n-1)} \Delta k - \frac{k(2n-1)}{n^2(n-1)^2} \Delta n$.

Network density and decomposition of the contribution of changes

Graph 7



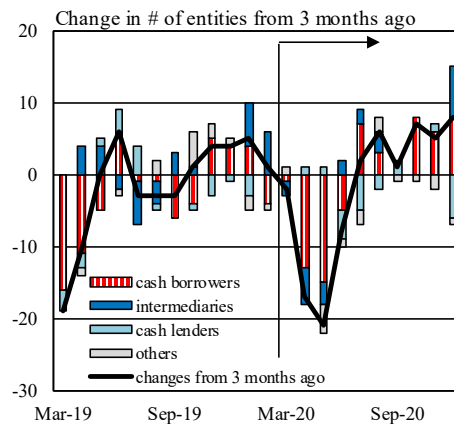
¹ The thin line on the left is the 3-month moving average. The right figure shows SC repo transactions.

Source: Data on securities financing transactions in Japan

Regarding the SC repo transactions, where the density significantly increased, we broke down the decline in the number of financial institutions by category based on the bow-tie decomposition used in the previous section ("cash lenders," "intermediaries," and "cash borrowers"). The decline in the number of cash borrowers (i.e., bond lenders) had the main contribution (Graph 8). Based on this, it can be considered as follows. In the SC repo market, during the stress period after March 2020, the number of bond lenders decreased due to an increase in collateral demand and a decrease in market participants caused by the declaration of a state of emergency and the increase in telecommuting. On the other hand, financial institutions did not reduce the number of counterparties with whom they actually conducted transactions to the extent that the total number of whole market transactions decreased, through development of new counterparties (Sakiyama and Yamada, 2016), which is thought to be behind the increase in density.

Breakdown of changes in the number of financial institutions in the network

Graph 8



¹ SC repo transactions.

Source: Data on securities financing transactions in Japan

Since the data used in this paper cover transactions with outstanding balances at the end of the month, the data fail to capture the short-term trend of transactions executed in mid-to-late March 2020, which is considered to be the most stressed period in the market. The network may have behaved differently during this period than the interpretation given above. To obtain implications for trends of the JGB repo market during times of market stress, it is necessary to reexamine this issue when data accumulation has progressed and data sample size has increased enough that event studies have become feasible.

5. Conclusion

This paper examined the characteristics of the JGB repo market network in Japan. In the JGB repo market, we found that highly important financial institutions in the network act as intermediaries and that continuous transaction relationships were established within groups formed around them. These characteristics suggest that the JGB repo market is efficient but has a network structure in which shocks to a few financial institutions can easily spread throughout the entire market. In order to assess the network structure of the market, it is important to consider whether the market strikes a balance between the aspects that impart transaction efficiency and the aspects that impart robustness. It would be beneficial to continuously monitor the functioning of the JGB repo market in Japan, while keeping in mind these characteristics of the network structure. It is also hoped that the results of this analysis will serve as a reference for market participants when considering the stability and efficiency of transactions through further transparency of the market.

This paper summarizes the basic characteristics of the network structure of the JGB repo market using the network analysis methods. Future research topics include an analysis of the background factors of the network structure and a more in-depth analysis of the behavior of the network structure in response to market shocks as more time-series data are accumulated.

References

- Abbassi, P., Fecht, F. and Weber, P. (2013). How Stressed Are Banks in the Interbank Market? Deutsche Bundesbank Discussion Paper, No. 40/2013.
- Allen, F., Covi, G., Gu, X., Kowalewski, O., and Montagna, M. (2020). The Interbank Market Puzzle. ECB Working Paper Series, No 2374.
- Afonso, G., Kovner, A., and Schoar, A. (2013). Trading Partners in the Interbank Lending Market. FRB of New York Staff Report, 620.
- Barabási, A. L., and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286 (5439), pp. 509–512.
- Bardoscia, M., Bianconi, G., and Ferrara, G. (2019). Multiplex Network Analysis of the UK OTC Derivatives Market. Bank of England Working Papers, No. 726.
- Bargigli, L., Di Iasio, G., Infante, L., Lillo, F., and Pierobon, F. (2015). The Multiplex Structure of Interbank Networks. *Quantitative Finance*, 15 (4), pp. 673–691.

- Beltran, D. O., Bolotnyy, V., and Klee, E. C. (2015). Un-networking: The Evolution of Networks in the Federal Funds Market. Finance and Economics Discussion Series 2015-055, Federal Reserve Board.
- Brin, S., and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30, pp. 107–117.
- Caballero, J. (2015). Banking Crises and Financial Integration: Insights from Networks Science. *Journal of International Financial Markets, Institutions and Money*, 34, pp. 127–146.
- Di Maggio, M., Franzoni, F., Kermani, A., and Sommovilla, C. (2019). The Relevance of Broker Networks for Information Diffusion in the Stock Market. *Journal of Financial Economics*, 134(2), pp. 419–446.
- Di Maggio, M., Kermani, A., and Song, Z. (2017). The Value of Trading Relations in Turbulent Times. *Journal of Financial Economics*, 124(2), pp. 266–284.
- Dong, G., Fan, J., Shekhtman, L. M., Shai, S., Du, R., Tian, L., and Havlin, S. (2018). Resilience of Networks with Community Structure Behaves as if under an External Field. *Proceedings of the National Academy of Sciences of the United States of America*, 115(27), pp. 6911–6915.
- Financial Stability Board (2015). Transforming Shadow Banking into Resilient Market-based Finance: Standards and Processes for Global Securities Financing Data Collection and Aggregation.
- Fricke, D., and Lux, T. (2015). Core-periphery Structure in the Overnight Money Market: Evidence from the e-mid Trading Platform. *Computational Economics*, 45(3), pp. 359–395.
- Fujimoto, F., Kato, T., and Shiozawa, H. (2019). Trends in Market Transactions after the Shortening of JGB Settlement Period -- Focusing on the Repo Market -- (available only in Japanese), Bank of Japan Research Paper.
- Furfine, C. (2003). Interbank Exposures: Quantifying the Risk of Contagion. *Journal of Money, Credit & Banking*, 35(1), pp. 111–128.
- Hollifield, B., Neklyudov, A., and Spatt, C. (2017). Bid-ask Spreads, Trading Networks, and the Pricing of Securitizations. *The Review of Financial Studies*, 30(9), pp. 3048–3085.
- Iazzetta, C., and Manna, M. (2009). The Topology of the Interbank Market: Developments in Italy since 1990. Bank of Italy Temi di Discussione (Working Paper) No, 711.
- Imakubo, K., and Soejima, Y. (2010). The Transaction Network in Japan's Interbank Money Markets. *Monetary and Economic Studies*, 28, pp. 107–150.
- Inaoka, H., Ninomiya, T., Taniguchi, K., Shimizu, T., and Takayasu, H. (2004). Fractal Network Derived From Banking Transaction—An Analysis of Network Structures Formed by Financial Institutions. Bank Japan Working Paper.
- in't Veld, D., and van Lelyveld, I. (2014). Finding the Core: Network Structure in Interbank Markets. *Journal of Banking & Finance*, 49, pp. 27–40.
- Iori, G., and Mantegna, R. N. (2018). Empirical Analyses of Networks in Finance. *Handbook of Computational Economics*, 4, pp. 637–685.

- Kaltwasser, P. R., and Spelta, A. (2019). Identifying Systemically Important Financial Institutions: a Network Approach. *Computational Management Science*, 16(1), pp. 155–185.
- Li, D., and Schürhoff, N. (2019). Dealer Networks. *The Journal of Finance*, 74(1), pp. 91–144.
- Markose, S., Giansante, S., and Shaghaghi, A. R. (2012). ‘Too Interconnected to Fail’ Financial Network of US CDS Market: Topological Fragility and Systemic Risk. *Journal of Economic Behavior & Organization*, 83(3), pp. 627–646.
- Martinez-Jaramillo, S., Alexandrova-Kabadjova, B., Bravo-Benitez, B., and Solórzano-Margain, J. P. (2014). An Empirical Study of the Mexican Banking System’s Network and its Implications for Systemic Risk. *Journal of Economic Dynamics and Control*, 40, pp. 242–265.
- Minoiu, C., and Reyes, J. A. (2013). A Network Analysis of Global Banking: 1978–2010. *Journal of Financial Stability*, 9(2), pp. 168–184.
- Minoiu, C., Kang, C., Subrahmanian, V. S., and Berea, A. (2015). Does Financial Connectedness Predict Crises? *Quantitative Finance*, 15(4), pp. 607–624.
- Mistrulli, P. E. (2011). Assessing Financial Contagion in the Interbank Market: Maximum Entropy versus Observed Interbank Lending Patterns. *Journal of Banking & Finance*, 35(5), pp. 1114–1127.
- Ono, N., Sawada, T., and Tsuchikawa, A. (2015). Towards Further Development of the Repo Market, Bank of Japan Review, 2015-E-4.
- Reichardt, J. and Bornholdt, S. (2006). Statistical Mechanics of Community Detection. *Physical Review*, E 74, 016110.
- Sakiyama, T., and Yamada, T. (2016). Market Liquidity and Systemic Risk in Government Bond Markets: A Network Analysis and Agent-based Model Approach. IMES Discussion Paper Series, Bank of Japan.
- Saltoglu, B., and Yenilmez, T. (2015). When Does Low Interconnectivity Cause Systemic Risk? *Quantitative Finance*, 15(12), pp. 1933–1942.
- Sasamoto, K., Nakamura, A., Fujii, T., Semba, T., Suzuki, K., and Shinozaki, K. (2020). New Initiatives to Improve the Transparency of Securities Financing Markets in Japan: Publication of Statistics on Securities Financing Transactions in Japan, Bank of Japan Review, 2020-E-1.
- Soramäki, K., Bech, M. L., Arnold, J., Glass, R. J., and Beyeler, W. E. (2007). The Topology of Interbank Payment Flows. *Physica A: Statistical Mechanics and its Applications*, 379(1), pp. 317–333.
- Yang, R., Zhuhadar, L., and Nasraoui, O. (2011). Bow-tie Decomposition in Directed Graphs. 14th International Conference on Information Fusion. IEEE.
- Yun, T. S., Jeong, D., and Park, S. (2019). “Too Central to Fail” Systemic Risk Measure Using PageRank Algorithm. *Journal of Economic Behavior & Organization*, 162, pp. 251–272.



A network analysis of the JGB repo market

February 2022

GEMMA Yasufumi

Bank of Japan



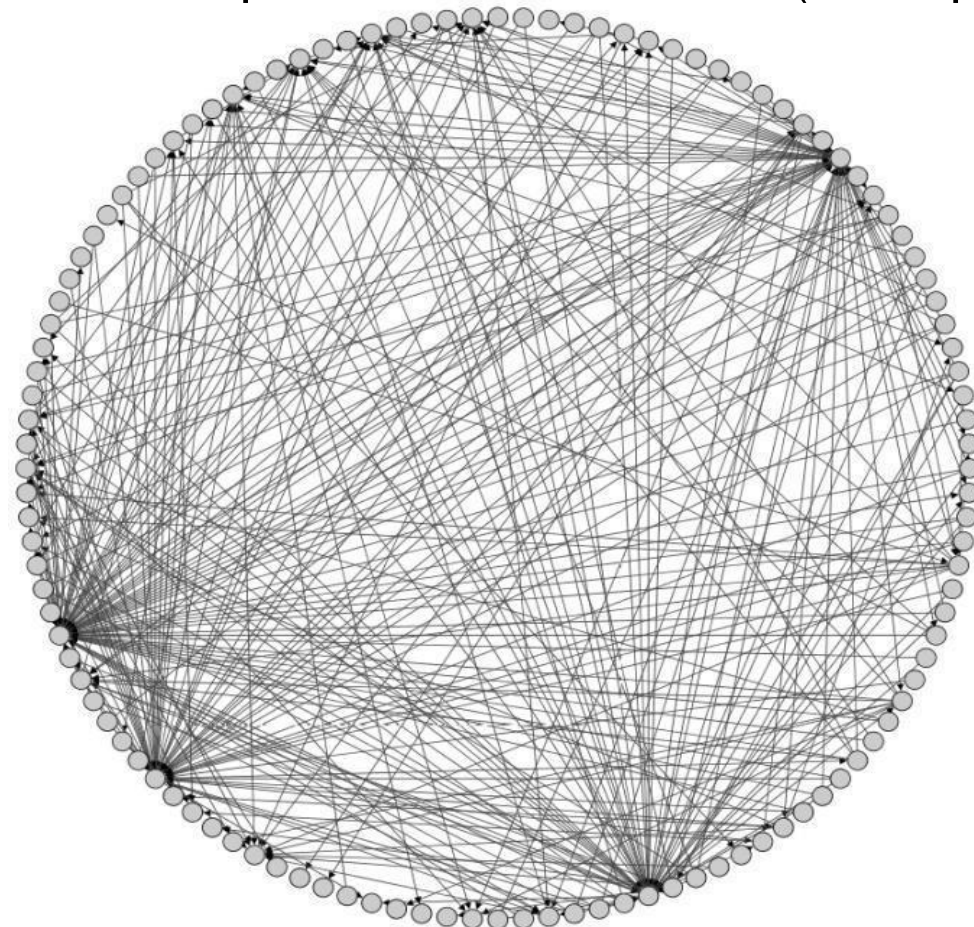
Introduction

- During the GFC, the functioning of repo markets was greatly degraded, which amplified the instability of the financial system. Based on that experience, global efforts to enhance the stability and transparency of repo markets have progressed.
- In Japan, JFSA and BOJ jointly started collecting detailed data on individual transaction units for repo markets from Dec. 2018.
- The **data are highly granular**, including names of both parties involved in any repo transactions in Japan with such information as repo rates and amounts. This makes it possible to grasp trends in the repo market from a variety of angles.
- Taking advantage of the data, we identify **network structure** of the Japanese government bond (JGB) repo transactions in Japan, and apply **network analysis** to examine its characteristics.

Network Analysis

- Network analysis of financial markets is used to understand complex structure of networks based on relationships among market participants by visualizing or measuring its characteristics.
- It can help to evaluate robustness and functioning of the entire financial market.
- Network structure in our work is defined as: FIs are represented as “**nodes**” and net cash lending relationships between two FIs are shown as “**directed links**”

JGB repo transaction network (GC repo)

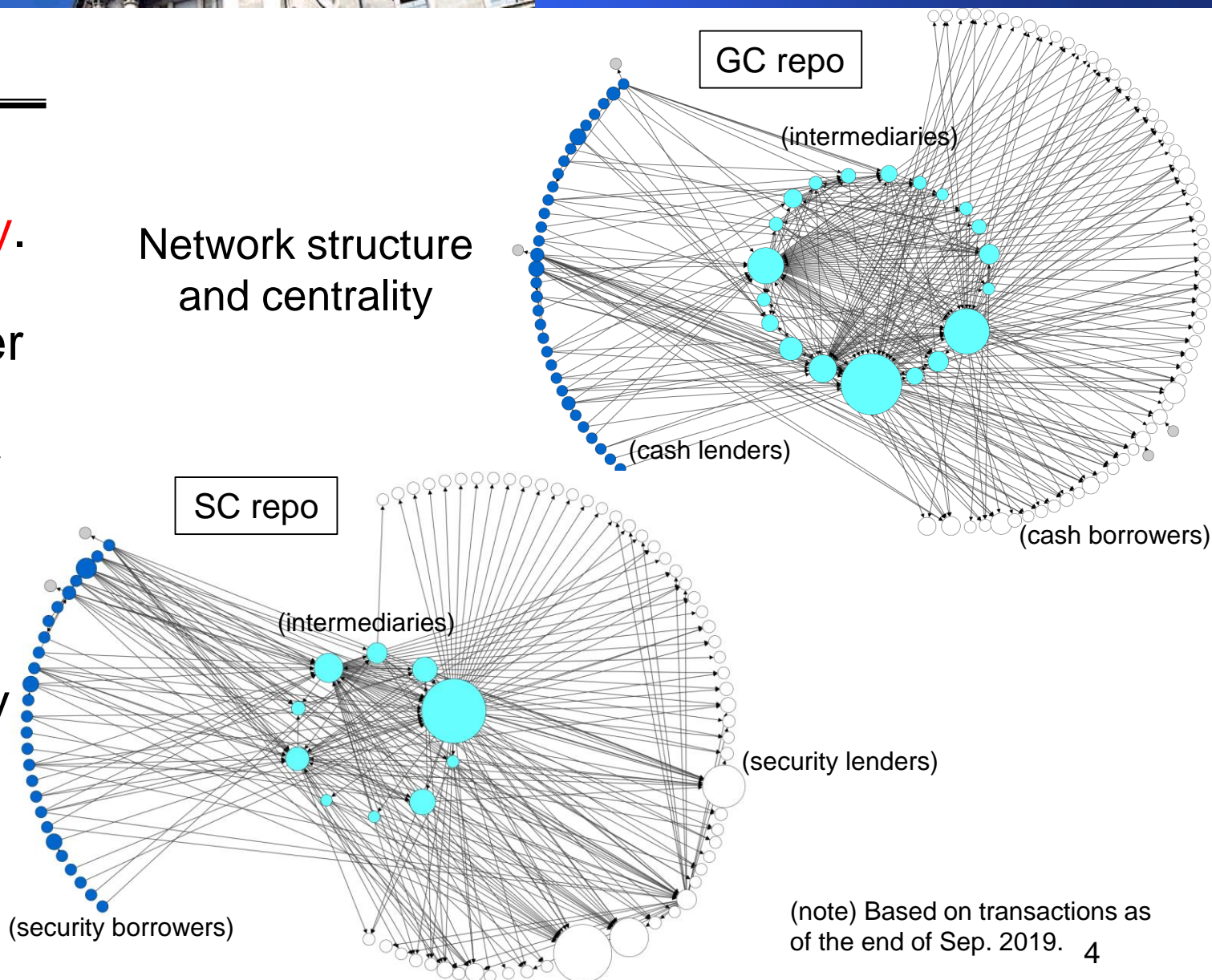


(note) Based on transactions as of the end of Sep. 2019.

Network Centrality

- Each FI's importance is measured by **network centrality**.
- For **GC repo**, intermediaries tier (light blue) includes highly important FIs, (it indicates their central role as intermediaries).
- For **SC repo**, not only intermediaries but also security lenders tier (white) have high centrality (some FIs play an important role as security suppliers).

Network structure and centrality

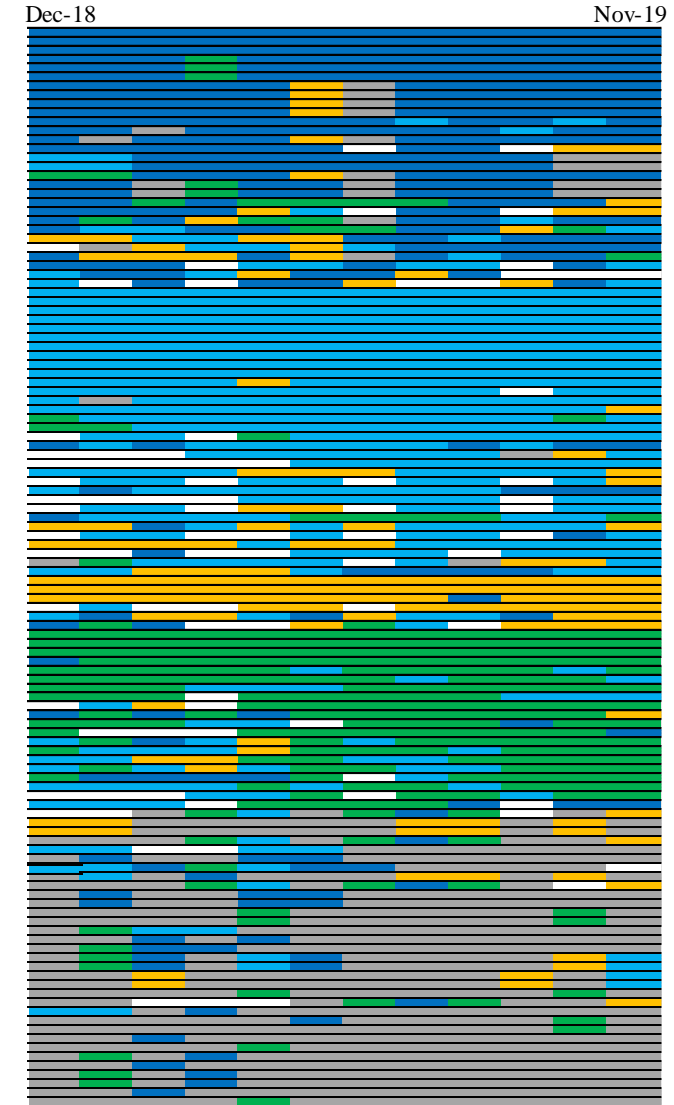


(note) Based on transactions as of the end of Sep. 2019. 4

Community Detection

- We use **community detection** method to analyze the community structure of the JGB repo market.
- Community structure graph for GC repo, where in each data period, FIs are lined up vertically and are color-coded by the formed communities, indicates that many FIs have **continued to conduct transactions within the same community**.
- According to a regression analysis, this community structure is formed for the purpose of facilitating **large-lot transactions**.

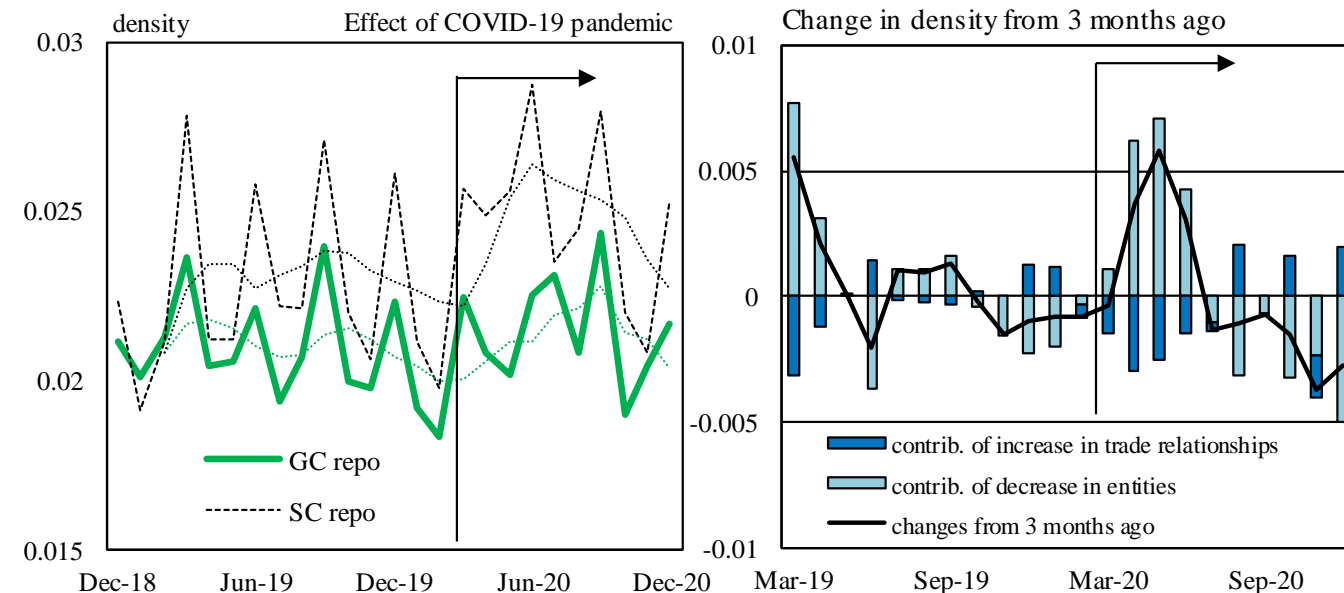
Detected Communities (GC repo)



Network during Market Stress

- We examine behavior of the network structure of JGB repo market **during market stress**, especially during the spread of COVID-19 pandemic early in 2020.
- Network density has remained high since March 2020, especially in SC repo market.
- It is found that decrease in the number of FIs (light blue area) mainly contributed to the higher density.

Network density and decomposition of the contribution of changes

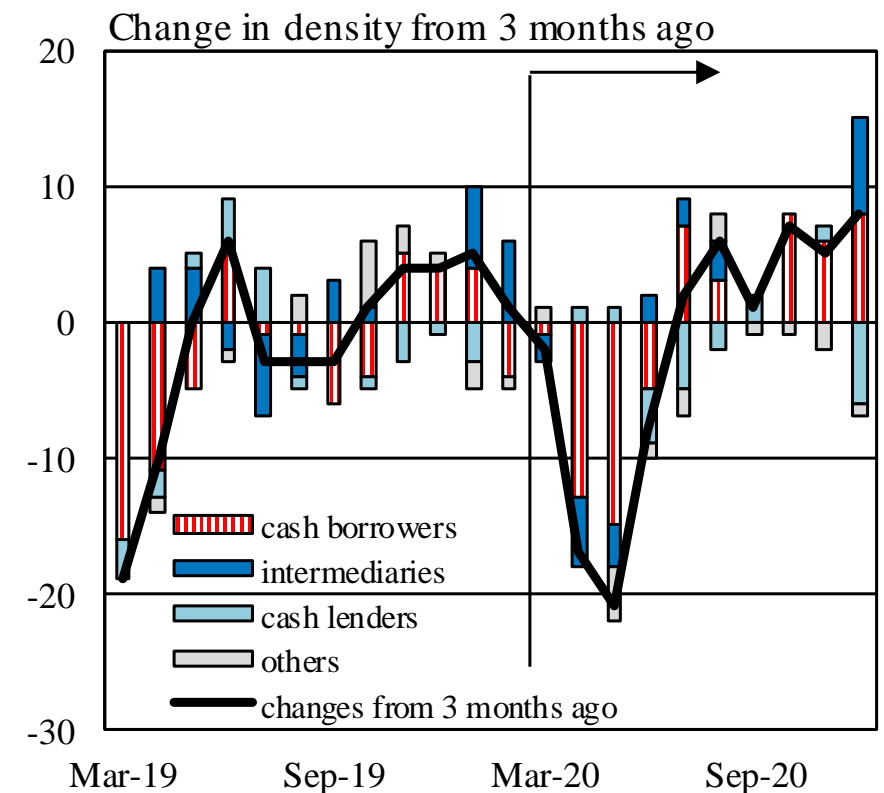


(note) The right graph is for SC repo.

Network during Market Stress

- For SC repo transactions, bond lenders had the main contribution to the decrease of the number of FIs.
- This indicates: In the SC repo market, during the stress period after March 2020, the number of bond lenders decreased due to some reasons (such as collateral demand).
- In the meanwhile FIs did not reduce the number of counterparties to the extent that the total number of whole market transactions decreased, through development of new counterparties.

Breakdown: changes in number of FIs



(note) The graph is for SC repo.



Conclusion

- We examined the characteristics of the JGB repo market network, where we found that highly important FI's in the network act as intermediaries and that continuous transaction relationships were established within groups formed around them.
- These characteristics suggest that the JGB repo market is efficient but has a network structure in which shocks to a few financial institutions can easily spread throughout the entire market.
- In order to assess the network structure of the market, it is important to consider a balance between the aspects that impart transaction efficiency and the aspects that impart robustness.
- It would be beneficial to continuously monitor the functioning of the JGB repo market, while keeping in mind the network structure examined here.

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Data science and statistics: a network analysis to understand the foreign investment¹

João Falcão Silva, Banco de Portugal,
Flávio Pinheiro and Bojan Stavrik, NOVA Information Management School

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Data science and Statistics: a network analysis to understand the foreign investment

Bojan Stavrik, Flávio Pinheiro, João Falcão Silva¹

Abstract

One important feature of the globalization process is the increase in the economic and financial interdependencies across countries. It is crucial to measure the connections between countries in order to identify where the financial centers are located, to characterize the foreign investments, and define the set of world countries that have stronger linkages, among others. In this context, foreign direct and portfolio investments play a crucial role in measuring international investments and understanding these dynamics. Such an interconnected web of foreign investment relationships is difficult to measure due to their complexity, as well as the lack of unified data sources. This article aims to use network analysis to map both the foreign direct investment and portfolio investment global relationships in order to identify patterns, preferential paths for investment, establish trends and describe the relations between countries over time. Secondly, it gathers the results of the network analysis and presents them in an intuitive web application, where the most important findings are highlighted allowing the users to interact with the data and extract insights over all the available years.

Keywords: Foreign direct investment, Portfolio investment, Network analysis, Interactive web application

JEL classification: C02, C63, F21

1. Introduction

With the ever-growing globalization, an increase in trade, and the possibility of easily investing abroad, it is becoming increasingly difficult to track the flow of money between countries. In such an interconnected web of relationships characterized by many players, markets, and investment opportunities, it is complex to map all the linkages between the origin and destination of each investment and address the ultimate investors.

In this context, the external statistics play an important role to analyse the cross-border financial investments between one country and its main investors. The financial account under the Foreign Direct Investment (FDI)² and the Portfolio Investment (PI)³ records the international investment. In the case of the FDI, it aims to establish a long-lasting interest in a foreign business, while the PI is oriented toward

¹ Bojan Stavrik (bojanstavrik@me.com) and Flávio Pinheiro (fpinheiro@novaims.unl.pt), NOVA Information Management School, João Falcão Silva (jmfsilva@bportugal.pt), Statistics Department, Banco de Portugal. The views expressed are those of the authors and not those of the NOVA Information Management School or Banco de Portugal.

² includes the initial investment and all the other financial linkages from residents of one country in an enterprise located in a foreign country, when the investor owns a minimum of 10 percent of the voting power.

³ corresponds to cross-border investments in the form of debt securities and equity that falls below the FDI threshold.

small investors that aim to get short-term returns. In this latter case, most transactions occur in secondary markets, thus not between the original issuer and the final investor.

Traditionally data has been communicated and shared in structured tabular formats. Indeed, many open data platforms these days still rely on such traditional format that makes it difficult for users to explore the underlying complex structures that often characterize the relationships between the elements that are being reported. Moreover, tabular data representations are not only difficult to interpret but also difficult to compare across time and space. In fact, information about size, magnitude, proximity, comparability, and temporal evolution are easier to grasp when data is represented and reported through appropriated visualization formats. A common solution comes in the format of a Business Intelligence dashboard, which offers a layout to quickly present multiple relevant visualizations and key indicators to the user and communicate insights to a specialized audience.

Here, we focus on the analysis of FDI and PI bilateral flows of one country vis-à-vis its main investor partners aiming to analyze the cross-border financial investment linkages in the form of debt instruments and equity in the case of FDI, and in the case of the PI debt securities and equity. To that end we use data provided by the International Monetary Fund (IMF) - Coordinated Direct Investment Survey (CDIS) and Coordinated Portfolio Investment Survey (CPIS), for all the available world countries. Using network analysis to model the linkages between the sources and final destinations of the FDI and PI, we show that the countries with more FDI and PI interconnections usually correspond to advanced economies, financial centres, or countries that offer tax benefits to investments.

Moreover, we explore the development of a web application to report the analytical results to a wider audience. The web application was built in a stack of free and open-source tools comprised of HTML 5, CSS, and javascript (jQuery, D3.js, and d3Plus). The presented web application offers a medium to i) communicate our findings through a personalized and interactive visualization-rich data-driven platform; and ii) allows for users to quickly explore and discovery of relevant partnerships and investment paths, facilitating the comparison between different countries and in different periods in time..

The article is organised as follows: after the introduction, a literature review on the network analysis and its linkage with economic variables is presented in section 2. The data sources and variables are described in section 3. Section 4 presents the methodology and section 5 shows some results. The fi-networks.com portal is described in section 6 and section 7 concludes.

2. Network analysis and economic variables

A network is a system made up of actors (individuals, organizations, countries, etc.) and sets of bilateral ties that represent relationships between them (Wasserman & Faust, 1994). This provides a structure for network analysis, allowing for the identification of central agents in complex local and global networks. Network science offers a unique set of tools and principles for studying complex relationships apparent in nature, technology, and society (Jackson, 2008). They help us understand how diseases spread, patterns in product purchases, languages spoken, voting, and educational decisions, to name a few (Jackson, 2008). Despite evident differences in various network domains, they emerge and evolve based on a set of fundamental mechanisms that govern network science (Barabási, 2013). Ter Wal & Boschma (2008) showcase the huge potential of network analysis to be incorporated in studying the structure and evolution of inter-organizational connection and knowledge sharing. They stress the importance of using high-quality data in building networks, and they identify primary data as the most statistically robust way of building networks (Ter Wal & Boschma, 2008).

Network analysis has a wide range of current literature on the application of network analysis to FDI and portfolio investment relationships, including patterns in banking and cross-country financial investments, both portfolio and direct investment. On the top of some examples, Hafner-Burton, Kahler, & Montgomery (2009) use network analysis on key international outcomes and test network theory in the context of international relations. Focusing on the portfolio investment linkages, Hakeem and Suzuki (2016) took a network approach in foreign portfolio investment of the European Union and its main counterparts. They focus on the relationship between countries' centrality and their economic indicators, showing that the more connections an economy holds the higher the impact on economic growth patterns. Moreover, the literature suggests that the more central a country is, the more embedded it is in a global portfolio investment network, implying greater exposure to foreign financial markets.

On the FDI perspective, Bolívar et al. (2019) find a strong relationship between economy size (measured through GDP) and the centrality of an economy in a global network. Furthermore, a country's commercial openness also has a strong positive relationship with centrality in the network. The more open an economy is the more FDI it attracts and inversely the greater the involvement in the outward global FDI network. On the other hand, political stability and average years of schooling have a moderate effect on inward FDI investment for that economy. As evidenced, developed-to-developed connections represent (66%) of the weighted global FDI networks. More recently, Norgren and Olsson (2021) apply Stochastic Actor-Oriented network models to study the relationship between FDI and institutions. They distinguish between formal and informal institutions, where informal institutions are culture and trends while formal institutions are the laws and rules of society. Additionally, (Lima, Pinheiro, Silva, & Matos, 2020) analysed the use of the network analysis for FDI relationships. The authors highlighted the visualisation capabilities of the network analysis methodology and also its ability to apply metrics that provide useful information about economic relations.

By definition FDI is meant to establish a long-lasting interest and, in many cases, the ultimate investor is difficult to trace, especially when investments involve offshore centers or Special Purpose Entities (SPE's)⁴. Small economies with inexplicably large FDI inflows are one of the clear signs that a country's total FDI value is inflated and its counterparts are not necessarily trying to establish long-lasting relations but instead use it for different financial planning goals (Damgaard & Elkjaer, 2017). Luxembourg and Netherlands are two such economies that host many foreign-owned multinational enterprises or SPE's. Damgaard and Elkjaer (2017) show the difference in those networks by combining CDIS with the OECD data. Using regression analysis, they estimate the amount of "real" FDI each of these relationships holds. Once the transformations are applied there is a 34% decrease in total inward FDI. Representing the "real" and "phantom" FDI in a network shows some differences. Smaller economies, such as Netherlands and Luxembourg weakened their intermediating power in the network although remain one of the most important global intermediators.

3. Variables description and data source

According to the Balance of Payments Manual, in its 6th edition (BPM6), direct investment includes the cross-border investments where there is a control or a significant degree of influence on the management of an enterprise that is resident in another economy⁵. It captures the immediate direct

⁴ SPE's are legal entities that have little or no employment, operations, or physical presence in the jurisdiction in which they are created by their parent enterprises, which are typically located in other jurisdictions (economies) (OECD, 2008).

⁵ The significant degree of influence is determined to exist if the direct investor owns from 10 to 50 percent of the voting power in the direct investment enterprise. Control is determined to exist if the direct investor owns more than 50 percent of the voting power in the direct investment enterprise.

investment relationships, i.e., when a direct investor directly owns equity that entitles it to 10 percent or more of the voting power in the direct investment enterprise. On the contrary, the portfolio investment is defined as cross-border transactions and positions involving debt or equity securities, other than those included in direct investment or reserve assets, meaning that there is no control or significant degree of influence on the non-resident enterprise.

The direct investment is usually presented in two alternative perspectives – following the asset/liability principle (as introduced in BPM6) or directional principle (requested in previous editions), whereas in the case of portfolio investment only the asset and liability principle is presented.

Under the directional principle, direct investment is shown as either direct investment abroad (outward investment⁶) or direct investment in the reporting economy (inward investment⁷). The asset (liability) principle of the portfolio investment, represents the amount invested by resident (non-resident) entities in the form of equity/debt securities, on non-resident (resident) entities.

In this paper, the implementation of the network estimation uses statistical information on the FDI directional principle and asset-liability principle for the portfolio investment. According to the available information, the primary presentation of international accounts shows positions with all non-residents as a total. Although, we follow a network perspective aiming to map the foreign investment linkages between partner economies. In this regard, we obtained information from the Coordinated Direct Investment Survey (CDIS) and Coordinated Portfolio Investment Survey (CPIS) provided by the International Monetary Fund. The selected data contains annual information from 2009 until 2019 on the total inward direct investment (stocks), inward equity direct investment (stocks), portfolio investment assets and liabilities.

Since both CDIS and CPIS correspond to total amounts in US dollars and not proximities, which is the desirable metric for our analysis, it is therefore necessary to transform their values to proximities⁸. In order, to obtain a proxy for proximity we consider the reciprocal of the absolute value of the directional investment. Therefore, the proximity calculation for each investment is as follows:

$$\phi_{ij} = \frac{1}{|f_{ij}|} \quad (1)$$

where $\phi_{ij} \neq \phi_{ji}$. In that sense, we say that the larger the investment amount between two countries, the closer they are to each other. Ultimately a proximity matrix is obtained, which forms the basis for building a directed weighted network that represents global trade flows in foreign investment.

To obtain a proxy of the foreign investment, we combine both Inward from CDIS and the liabilities from CPIS datasets (combined liabilities) and outward from CDIS with assets from CPIS (combined

⁶ Investments by resident direct investors in their direct investment enterprises abroad deducted from the reverse investments by direct investment enterprises abroad in their resident direct investors.

⁷ Investments in resident direct investment enterprises by direct investors abroad minus Reverse investments by resident direct investment enterprises in their direct investors abroad.

⁸ Let us start by defining f_{ij} as the total investment of country i in country j , which can in general terms concern any of the indicators that were used. The investments between two countries can be asymmetrical, that is, $f_{ij} \neq f_{ji}$ implying the directionality of the investments. While, in most cases, the value of f_{ij} is positive, in certain situations it can also be negative. For instance, suppose country i (the parent) invested and is holding a position in country j (the affiliate); the parent can use the affiliate for funding operations back at home. When the total amount of funding that flows back to the parent, exceeds the total amount of investment done in the affiliate it is standard to report it as a negative flow. As such, we shall consider the absolute value of f_{ij} , that is, $|f_{ij}|$. The main reason for taking the absolute value instead of removing such observations is to prevent the loss of information about important investment partners.

assets). It is interesting to understand how the networks change once the main cross-country investment datasets are combined, to map possible international investment linkages across countries⁹.

4. Methodology

A network, G , is composed of two different but complementary elements: a set of N vertices/nodes and a set of links/edges. Edges connect a pair of nodes and identify the existence of a relationship between them. Vertices represent the unit of analysis, while the edges represent the relationship between them. In the context of this work, we shall use nodes of a network as representations of countries, while edges represent the existence of an investment between a pair of countries.

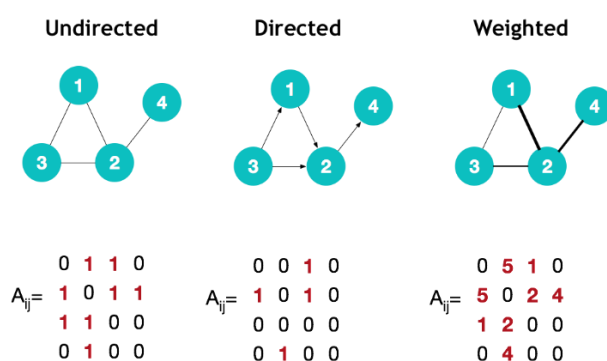


Figure 2 – Graphical (top) and matrix (bottom) representation of networks representing structures that have edges of different nature. Source: (Lima, Pinheiro, Silva, & Matos, 2020).

Depending on the nature of the relationships being modelled, there can be three main types of networks, namely: undirected; directed; and weighted. One way to represent such networks is through the adjacency matrix, see bottom panel of Figure 2. The adjacency matrix, A , of a network informs on the existing relationships between nodes/vertices. In that sense, the entry a_{ij} of A is zero if there is no relationship between nodes i and j , being non-zero if a relationship exists between such a pair of nodes. In a weighted graph, a_{ij} represents the weight (strength) of the relationship between the nodes, where $a_{ij} = a_{ji}$. For the directed network case the matrix is not symmetric along the diagonal and indicates a relationship and its direction, therefore $a_{ij} \neq a_{ji}$. The diagonal entries of each matrix A indicate self-relationships, and as general practice are set to zero.

When building a graphical representation of a network, each edge represents a relationship between two nodes (i.e., person, country, institution). In a directed network, edges are represented with arrows to indicate the direction of the relationship (from source to destination), with the possibility of two links (arrows) between two nodes. Additionally, in a weighted graph, the thickness of the edge represents the strength of the relationship.

Additional attributes can be associated with each relationship and each actor (e.g., we might want to consider the gender or age of individuals). However, these additional attributes do not affect the core structure of the networks, but they add a dimension that allows to classify relationships and profile explanatory factors for the creation of relationships.

⁹ These datasets do not refer to the same statistical concepts because FDI is recorded in the directional principle, whereas the CPIS is recorded on the assets/liabilities perspective. Although there is no available information on the FDI assets/liabilities by counterpart country, therefore, the best proxy that can be used to understand the international investment in the form of debt/equity is to aggregate CDIS Outward with CPIS Assets as Combined Assets, and CDIS Inward with CPIS Liabilities as Combined Liabilities.

Weights can be interpreted either as proximities/similarities or distances. It is important to establish which measure is being used in a network, as they have opposite interpretations. However, the choice hinges on a balance between the available data and the analytical purpose of the network structure under study. It is also common to study a simplified projection of the network, for instance by using an unweighted projection of a weighted network by applying a threshold to edges. For the case of analyzing global FDI and portfolio investment patterns, both directionality and strength matter. Therefore, directed-weighted networks are constructed to define and explain the underlying structure and identify the most central countries and their characteristics.

Often, we want to identify the role of each actor in the overall system through its position in the network. In network analysis, this is done by estimating the centrality of nodes. Several measures exist for that purpose. For instance, one can argue that the most central/important node is the one with the highest degree/connectivity, which is the number of links that are connected/connect to a node. Moreover, in a directed network the measure can be analyzed separately into incoming and outgoing connections. Hence, allowing us to define three measures: degree centrality, in-degree centrality, and out-degree centrality. The standardized formula for degree centrality is:

$$C_D(a) = \frac{v_a}{n-1} \quad (2)$$

Where v_a is the number of nodes a is connected to, and n is the total number of nodes in the network. In terms of in/out-degree centrality, the formula follows the same logic, except taking into consideration only incoming or outgoing edges in the numerator.

However, the number of connections an edge holds can tell little about the role of a node in mediating information between different parts of the network. To that end, betweenness centrality assumes that a node is more important the more shortest paths (paths connecting pairs of nodes in the network) it mediates. The higher the betweenness centrality of a node the more central/relevant it is. The formula for betweenness centrality is:

$$C_B(a) = \sum_{i,j \in A} \frac{\sigma(i,j|a)}{\sigma(i,j)} \quad (3)$$

Where a is the node (country), $\sigma(i,j)$ represents the number of shortest (i,j) -paths, and $\sigma(i,j|a)$ is the number of shortest paths passing through node a , other than i,j . If $i = j$ then $\sigma(i,j) = 1$, and if $a \in i,j$ then $\sigma(i,j|a) = 0$.

Considering the distance between nodes in a network, closeness centrality measures the importance of a node depending on how close it is to the other nodes in the network. The closer it is, on average, the more central it is. Its formula is:

$$C_C(a) = \sum_{v=1}^{n-1} \frac{n-1}{d(v,a)} \quad (4)$$

Where $d(v,a)$ is the shortest path distance between v and a , and n is the number of nodes that can reach a .

Figure 3, visually shows how these different measures of centrality can classify different nodes as the most central, thus highlighting that each one plays a different role on different structural dimensions.

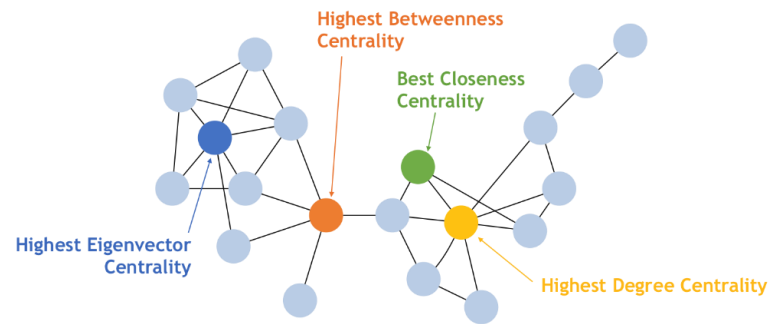


Figure 3 – Simple network with the most important centrality measures Source: (Lima, Pinheiro, Silva, & Matos, 2020)

In this article, we will use nodes to abstract countries and edges to identify financial relationships between pairs of countries. Moreover, edges will represent revealed proximities between countries. We will perform all computations in the entirety of the network (e.g., node centrality and shortest paths), with all its links, however, for visualization purposes (because the networks are very dense) we will represent only the most relevant edges.

To that end we shall follow the following steps: 1) we identify the Minimum Spanning Tree, which is a set of edges that ensures all nodes are interconnected while minimizing the sum of proximities between the selected edges; 2) then we enrich the Spanning Tree with the edges that identify the closest relationships until we reach a minimum average degree of 3.5 links, which we take as a thumb rule for a network density that would allow for interpretable network visualization.

5. Results

In this section, we use network analysis to present and discuss the results that answer the questions formulated in the previous section. Although, there are many different combinations of networks, years, countries, and measurements all results can be found in the web application (see <https://fi-networks.com>).

Correlation

Figures 4A and 4B show the spearman correlation between the betweenness and closeness centrality values of countries in the year 2019. We show that country rankings are highly correlated across networks, thus applying network analysis and using the centrality measures, it is possible to track the position of countries and their importance in a global investment network. Figure 4A shows, for instance, that CDIS Inward has an almost perfect correlation (0.99) with Combined Liabilities, while CPIS Liabilities is substantially lower (0.28). Since CDIS Inward and CPIS Liabilities are the building blocks of Combined Liabilities, their respective correlations indicate that the CDIS Inward is much more influential in the aggregation.

Looking at the Combined Assets, the observed correlations are stronger. In this case, both CDIS Outward and CPIS Assets have the same weight towards the Combined Assets network. Another observation is the strong correlation between both betweenness and closeness centralities in the CDIS Inward and Outward datasets. Hence, it allows us to conclude that the same countries intermediate the most investment paths in the outgoing and incoming investments while being the closest to the other countries in their respective networks. Another conclusion that can be taken out of the correlation analysis is that betweenness and closeness are very similar across the 6 different networks under study.

The closeness rank correlations are slightly stronger. This shows that there are a few most central countries that hold the highest rank positions for these two centralities, and that play a more key role in foreign investment. In other words, a few countries including the United States, Netherlands, Luxembourg, China, Hong Kong, and the United Kingdom are the main global intermediators, which will be justified throughout this section. These countries can usually be characterized by being a global economy such as the U.S. and China or a tax haven such as Luxembourg, the Netherlands, and the Cayman Islands.

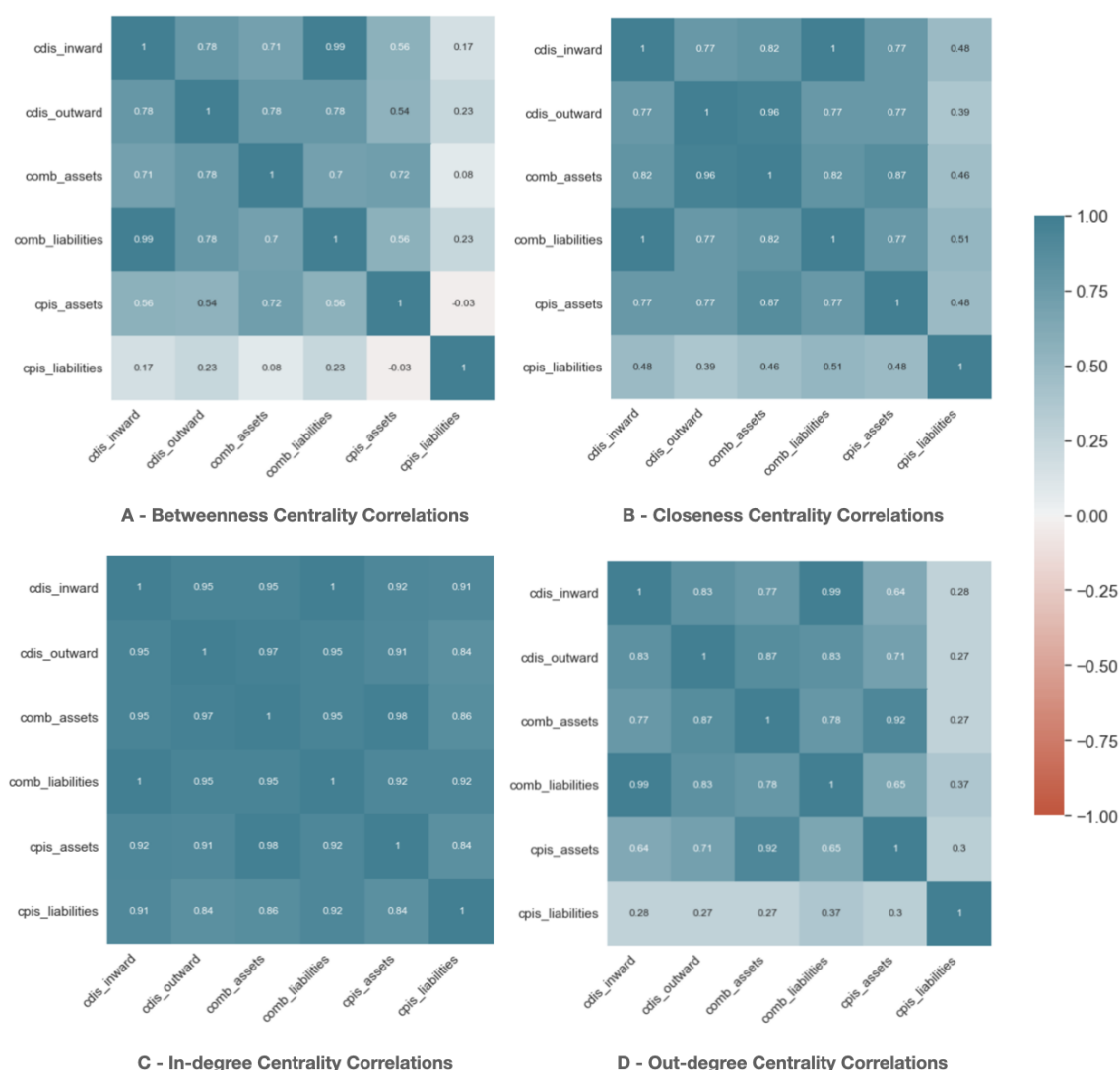


Figure 4 – Spearman Correlation Matrices for Betweenness (A), Closeness (B), In-degree (C), and Out-degree (D) centralities in the Networks estimated for the year of 2019.

In addition to the spearman correlations based on the betweenness and closeness centralities, we also present the spearman correlations based on the in and out-degree centrality values across all networks. Figure 4C highlights the correlations based on in-degree centralities, showing a nearly perfect correlation between all networks. This implies that no matter what network we look at, the ranks based on incoming investments are mostly the same. Figure 4D on the other hand follows a similar pattern to the betweenness centrality correlations across networks.

Foreign Direct Investment-- Inward Network

As previously mentioned, six data sets were considered in our analysis¹⁰. In this section, we illustrate the results for one specific data set: FDI inward. For the remaining data sets, the fi-networks portal will have all the information. Although, the interpretation of the remain data sets follows closely the one that is described below.

FDI Inward is shown in Figure 5. It is noticeable that some of the biggest global economies are present in the top ranks for the centralities throughout the years. In 2009 the top intermediators are the Netherlands, the U.S., the China, the Canada, and the Russia in that order. Similarly, in the 2019 network, the Netherlands and the U.S. hold the top two positions, while third-placed Luxembourg is followed by Hong Kong and China. In the case of the Netherlands, it is interesting to point out that it has substantial inward direct investment flows relative to its GDP, which can be explained by its lenient corporate tax laws. The position of the U.S. on the other hand can be explained by its global economic strength, as well as it being an innovation hub that has attracted and produced some of the most successful multinational corporations. The position of Hong Kong, and its rise as the fourth most intermediary of FDI, implies that it overtook Singapore's position to become an influential player in East and South-East Asia as well as globally. Nonetheless, Singapore continues to play an important role in the region too.

Shifting focus on the highest-ranked countries based on closeness centrality, the results are somewhat different. The highest-ranked countries for 2009 include Bermuda, Netherlands Antilles, Samoa, Cayman Islands, and the U.S. Similarly, the top five ranked countries for 2019 are Samoa, Cayman Islands, Bermuda, Jersey, and the British Virgin Islands. Most of these countries are tax havens. In that sense, considering that closeness centrality highlights nodes that are on average closer to all the other nodes in the network, such countries should offer an efficient route to spread investment from a source investor that is looking to spread its investments to many destination countries.

Nonetheless, the Netherlands and the Luxembourg rank, respectively, 12th and 13th in closeness centrality, while the U.S. is 9th. Additionally, based on the out-degree centrality the same countries tend to stay in the top list for both 2009 and 2019, a ranking that is occupied by Italy in the top position followed by the China, Thailand, and Bulgaria. In contrast, the in-degree centrality ranking lists the U.S. in the top position followed by the U.K., France, Switzerland, and the Netherlands. In both in- and out-degree rankings the top countries are generally not very volatile in their rank.

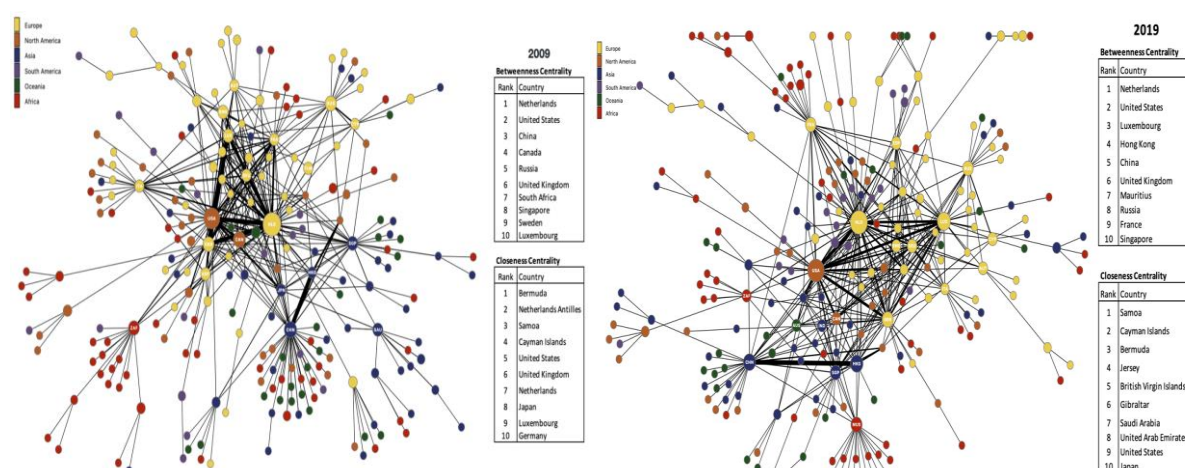


Figure 5 – CDIS Inward Network with node size representing relative amounts of Betweenness Centrality for 2009 (top) and 2019 (bottom). Source: web app

¹⁰ inward/outward of the FDI, assets/liabilities of the portfolio investment and the combined assets and liabilities.

The treemap in Figure 6 further reinforces the finding that several countries are the top facilitators of global trading roots in terms of FDI, by continent. The U.S., Netherlands, China, and Luxembourg combined make up 42.57% of global FDI inward investments for 2019.

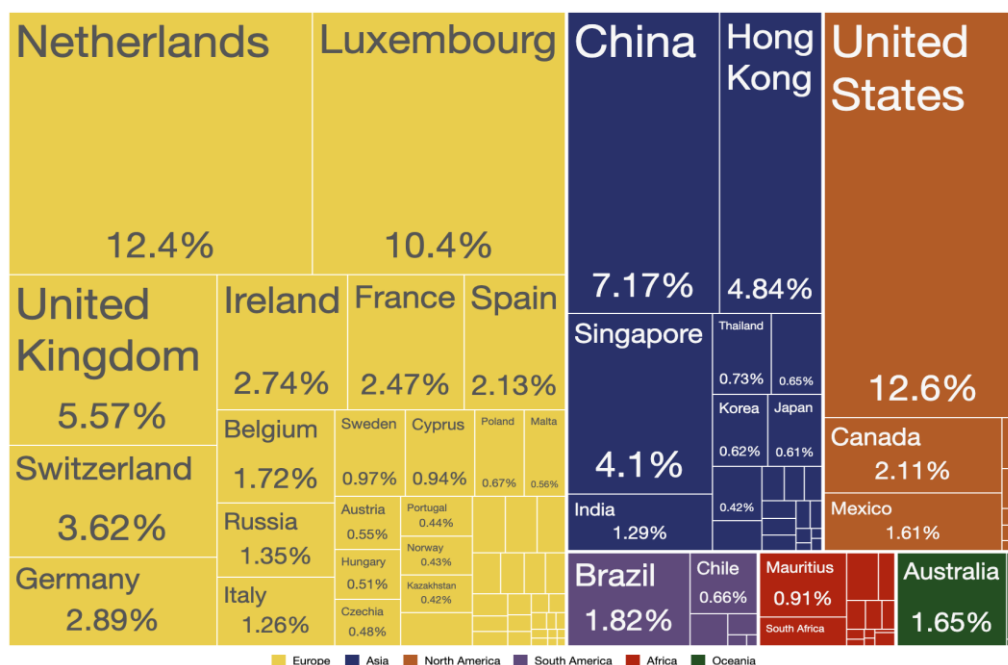


Figure 6 – CDIS Inward Treemap with block size representing the relative amount of FDI investment per country for 2019, with the colours representing the different continents. Source: web app

With China as a global manufacturing powerhouse and the U.S. as an innovation hub and both as one of the biggest world economies, they form one of the most important bilateral relationships in the world. In addition, Figure 7 represents the evolution of this relationship between 2009 and 2019. From 2009 up until 2017 Japan was the only intermediary in this relationship. While from 2018 to 2019 the U.K. and Hong Kong became intermediators between China and the U.S.



Figure 7 – CDIS Inward Shortest Path Visualization, where each node represents a country as part of the shortest paths of investment between the U.S. and China.

6. The fi-networks.com Portal

We chose to develop the web application from scratch using free and open-source tools. The exploration analysis and all computations were done using python and resorting to libraries common in the data science stack, such as Pandas and NetworkX. The early steps of the project consisted of data cleaning and pre-processing, stored in json or csv files. Those documents are the basis for the visualizations displayed in the portal.

The portal, as a web application, was built in a stack of free tools that comprised HTML 5, CSS, and javascript. In particular, several handy javascript libraries — such as jQuery, D3.js, and d3Plus — were used for the development of interactive elements and build the visualizations. Moreover, the bootstrap

CSS framework developed by Twitter (see <https://getbootstrap.com/>) was used for quick prototyping of html elements. Finally, we used Font Awesome icons throughout the website to style dynamic actions (e.g., mouse hovering) elements to buttons.

Concerning javascript libraries — D3.js and d3Plus — they operate as DOM manipulators while taking leverage on the use of Scalable Vector Graphics (SVG's), which allows drawing shapes in a browser window. The main advantage of using SVG's is that these objects do not lose quality when rescaled and have low memory requirements.

Regarding to the data loading and storage, data is loaded directly from csv files due to its lower development complexity and requirements¹¹.

1.1.1 Landing Page

The goal of the landing page is to grab the attention of the user while offering quick access to the different datasets available.

The top of the landing page contains a several relevant statistics cards about the used datasets. Their goal is to give the user an overview of the magnitude of the data being used. Below that, we present five teaser questions are formulated that a user might already be thinking about in the context of this framework. Each teaser corresponds to one visualization on the network page. They are all clickable and linked to their respective visualization, which would ultimately answer the question. Taking the first question as an example, "What is the top direct investment intermediary for 2019?" would lead to the network visualization and highlight the country that has the top rank for betweenness centrality for 2019. The same logic was applied to the rest of the teasers, except they take the user to a different section.

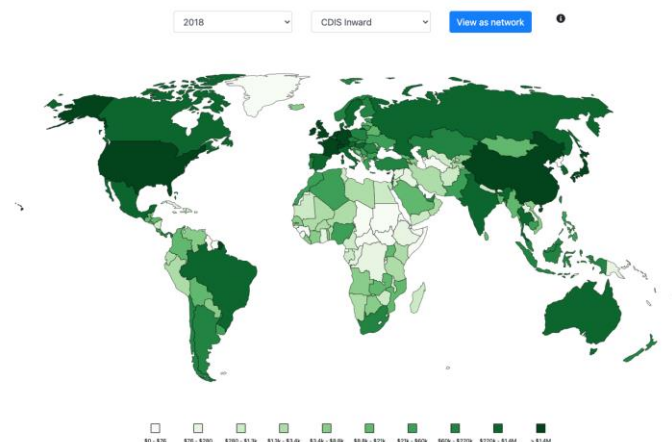


Figure 8 – World map visualization in the landing page shown for 2019 CDIS Inward. Source: web app.

Below the teasers, we present a world map highlighting the total FDI/Portfolio Investment amounts per country using different shades of green. As shown in Figure 8, the goal is to give a breakdown based on the total amount of investment. It shows the relative size of investments per country through the shades of green, the darker the color, the higher these amounts¹².

¹¹ Another option was to build a database that would feed data to the portal through a RESTful API.

¹² The user has the option to switch between all the datasets and years which change the respective values on the map. The "View as network" button takes the user to the network page for the latest selection of year and dataset in the dropdown menus. This adds more interactivity and another direct link to the main visualization of the web application.

1.1.2 Network Page

The network page is the core element of the portal. Each visualization in this page is broken down into two sections. The selection (left bar) side allows the user to select the visualization options, by selecting different networks, years, countries, and dataset. The options change depending on visualization, although all of them present an option to select the dataset and year. There are also information icons, which explain the contents of the visualizations, with a link to the methodology page where everything is explained in more detail. The last element of the selection pane has its own set of teaser questions; these are clickable and highlight the answer in the respective visualization. To explain the storyline of the web app, the United States will be used as an example to show their position, evolution, and conclusions that can be drawn based on the information presented.

The network visualization is the most complex one, putting into evidence the web of trade flows that can be estimated from the different datasets. Each node represents a country, while its size represents the value of the measure that is selected (betweenness, closeness, in-degree or out-degree centrality). The colors represent the continent of each country. The link that connects countries represents an investment relationship between those counterparties, while their thickness is the strength of that investment. Thickness, instead of color scale, was chosen to represent the strength of the relationships because it is a more natural degree of freedom to communicate intensity/magnitude.

The displayed networks represent a subset of the most relevant relationships on top of the Minimum Spanning Tree (MST). By doing this we overcome the issue of representing a very dense network, as most pairwise relationships exist, though they might be irrelevant in most cases. Hence, for visualization we enrich the MST with the most relevant links until a network with average degree of 3.5, which we use as a thumb rule. However, such filters are only done for visualization purposes, all calculations are performed on the fully spanned directed-weighted networks.

The most central countries per selection are highlighted with their ISO3 codes printed inside the nodes. Although, the user can manually click a country to highlight it in the network. Clicking the teaser for the CDIS Inward network in 2019 highlights the U.S.A. as the most central country based on betweenness centrality. In case a country is difficult to identify by looking at the network structure, the search menu helps in finding any country in the network.

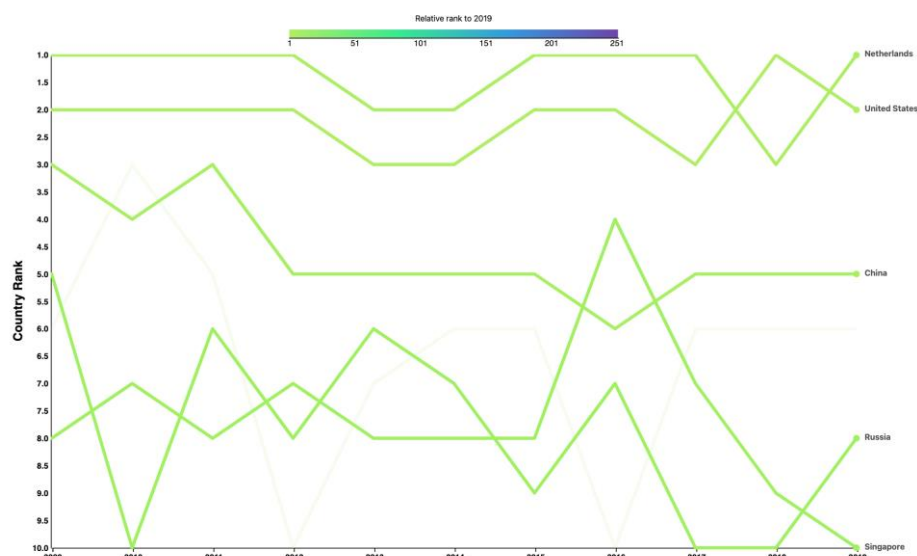


Figure 9 – Rank Evolution Chart between 2009 and 2019 based on Betweenness Centrality, showing the evolution of Netherlands, United States, China, Russia, and Singapore. Source: web app.

The second visualization presented is the rank evolution chart. The purpose of this visualization is to show the evolution of the relative importance of countries through their rank over the available time period per centrality measure. It shows how stable or volatile countries' position is based on the different centrality measures. In Figure 9, the 5 least volatile countries are highlighted in terms of betweenness centrality rank for the CDIS Inward network. The color is fixed at the respective country's rank in 2019, serving as a reference to the 2019 rank throughout the years. There is an option to select the network, centrality, and countries displayed in the chart. In the case of the U.S. Figure 9 shows its low variability in terms of betweenness centrality, never dropping below the third position between 2009 and 2019.

Next we show the time evolution of the shortest path charts. This particular visualization highlights the most likely intermediators if the flow of investments would follow a path of "minimum effort". More importantly, it shows how such paths changed from year to year. Figure 10, shows the evolution of the shortest path between the U.S. and China based on the CDIS Inward network. Between 2009 and 2017 Japan was the sole intermediary, while in the last 2 years it has been the U.K. and Hong Kong. In addition, the platform includes also a treemap chart to show the relative magnitude and breakdown of the total amounts per country and continent dependent on year and network.



Figure 10 – Shortest path chart between the United States and China between 2009 and 2019. Source: web app.

Lastly, the top intermediators chart (Figure 11) aims to identify the countries that more often participate as intermediators in the shortest investment paths globally. It ranks the top 15 countries by the proportion of paths they intermediate, based on two measures: One is the percentage of times a country is part of a shortest path of investment globally; and secondly the other identifies the percentage of times a country is the first intermediary in the shortest path of investment. The U.S. in the selection for CDIS Inward in 2019 is in the second position as top intermediary, being part of 53.73% of shortest investment paths (blue bar). While being the first intermediary 18.17% of the time (yellow bar).

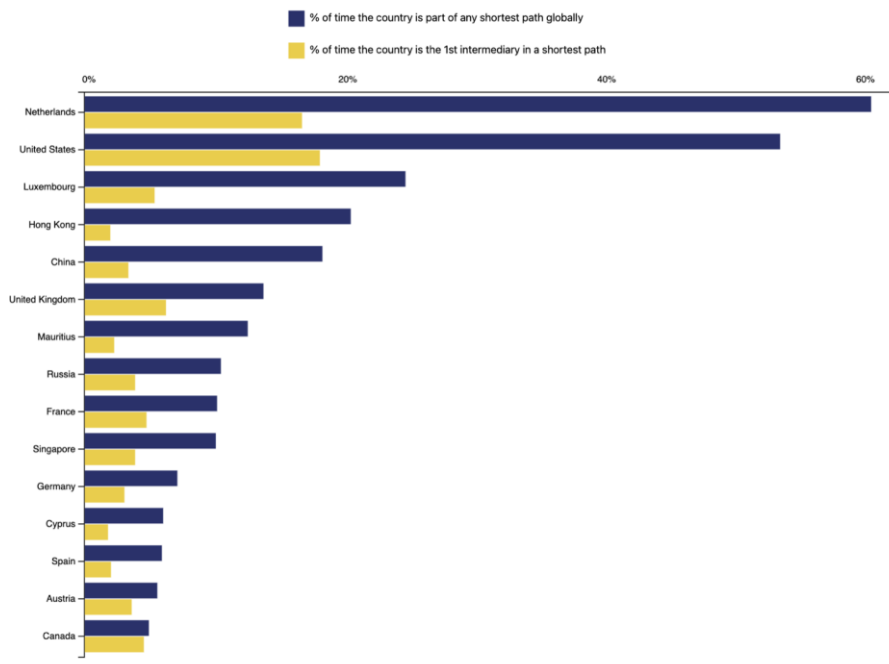


Figure 10 – Horizontal bar-chart chart showing the top intermediators based on the percentage of times they are part of a shortest path (blue bar) and percentage of time they are the first intermediary (yellow bar). Source: web app.

7. Conclusions

This article focuses on the power of network analysis to use immediate counterpart CDIS and CPIS data to trace the ultimate investors, as well as analyze any country's rank, influence, and connectedness in a global network.

Our results show that global economies and strategic partners such as the United States, China, Hong Kong, Netherlands, and Luxembourg rank the highest in the different networks in terms of intermediary power based on betweenness centrality for 2019. Notwithstanding, the variation in these ranks is not the same for all the listed countries. Hong Kong and China, have become some of the most influential intermediators only in the last several years. The United States, Netherlands, and Luxembourg are in the top 20 ranks for most of the years between 2009 and 2019. In terms of closeness centrality, off-shore countries such as the Cayman Islands, Bermuda, Jersey, the and British Virgin Islands are among the top ranks in the last few years, while the in/out-degree centralities rank the biggest global economies the highest without significant volatility between 2009 and 2019.

In addition to the analytical perspective, the web application has proven to ease the tracking of individual countries' positions in global investment networks over time. It gives central bankers, policymakers and all the other users, an easy way to identify underlying investment paths from a global investment network that is built based on CDIS and CPIS. One of the advantages is that the networks and measures are pre-computed, and their results can be quickly rendered on the page. Additionally, the visualization of the results makes it easier to understand and extract valuable information from it.

Even though, this analysis has some limitations. From an analytical point of view, some new insights could be gained by analyzing a larger timeframe. Interesting results could be obtained when looking at the differences in the networks between certain events such as the financial crisis of 2008/2009 or very recently the structure of the networks before and after the Covid-19 pandemic. It is also important to mention that there exists some reporting data gaps as the information regarding to the assets/inwards is not symmetric to the liabilities/outwards.

User experience surveys and interviews have not been performed due to time limitations. They are integral part of building web applications in order to better understand the improvement aspects of the design and communication in the web app. Such information should be gathered from experts in the field, such as central bankers and policymakers.

Finally, the adoption of portals as the case of fi-networks can be a very comprehensive tool for users, and a more visualized way for the organizations to show/highlight the main messages that arise from the data.

References

- Balance of Payments and International Investment Position Manual, 6th edition, International Monetary Fund.
- Barabási, A.-L. (2013). Network science. Philosophical Transactions of the Royal Society A.
- Bolívar, L. M., Casanueva, C., & Castro, I. (2019). Global Foreign Direct Investment: A network perspective. *International Business Review*.
- Damgaard, J and T Elkjaer (2017): "The Global FDI Network: Searching for Ultimate Investors", IMF Working Paper, November 2017.
- Jackson, M (2008): "Social and Economic Networks". Princeton: Princeton University Press.
- Hafner-Burton, E. M., Kahler, M., & Montgomery, A. H. (2009). Network Analysis for International Relations. Cambridge University Press, 559 - 592.
- Hakeem, M. M., & Suzuki, K.-i. (2016). Foreign Portfolio Investment and Economy: The Network Perspective. International Conference on Business, Economics, Management and Marketing.
- Lima, F., Pinheiro, F., Silva, J. F., & Matos, P. (2020). Foreign direct investment – using network analysis to understand the position of Portugal in a global FDI network.
- Norgren, A., & Olsson, M. (2021). Institutional Dynamics in the Global FDI Network. Linköping.
- OECD Benchmark Definition of Foreign Direct Investment - 4th Edition.
- Ter Wal, A. L., & Boschma, R. A. (2008). Applying social network analysis in economic geography: framing some key analytic issues.
- Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Data science and Statistics: a network analysis to understand the foreign investment

Bojan Stavrik | Flávio Pinheiro | João Falcão Silva
bojanstavrik@me.com | fpinheiro@isegi.unl.pt | jmfslva@bportugal.pt |

Data science in central banking

14 February 2022



BANCO DE
PORTUGAL
EUROSISTEMA





China Three Gorges buys EDP stake for 2.7 billion euros

Axel Bugge

LISBON (Reuters) - China Three Gorges won the competition to buy Portugal's stake in utility EDP ([EDP.LS](#)), paying 2.7 billion euros (\$3.5 billion), in a privatization seen key to the indebted euro zone country's ability to sell state assets.

The deal, which also includes Chinese investment in the wider economy, is the brightest news for Portugal since it was forced to seek a 78 billion euro bailout from the European Union and International Monetary Fund in the spring after its financing costs soared.

State holding company Parpublica said on Thursday that China Three Gorges' offer for the 21 percent stake in EDP, Portugal's largest company, was at a 53 percent premium to its share price.

The Chinese energy giant beat Germany's E.ON ([EONGn.DE](#)) and Brazil's Eletrobras ([ELET6.SA](#)) after a tough competition in which Three Gorges had promised to sharply

Budget 2022: Investors brace for another year of high bond supply

The government will need to borrow a minimum Rs 4.5 lakh crore to repay past loans, apart from continued spending to support an economic recovery

APARNA IYER
JANUARY 24, 2022

Find a buyer

Bond supply jumped the most in FY21 due to the pandemic. However, bond yields fell more than 100 basis points during the year. Yields move inversely to bond prices.

The key reason was that the RBI infused a historic amount of liquidity over various tenures through an array of instruments. The central bank also stood in the market as a constant buyer of government bonds.

*In FY22, the central bank has been a net buyer, but a more reluctant one. Its purchases dropped to Rs 1.4 lakh crore this year from Rs 3.1 lakh crore in FY21. **In recent weeks, the central bank has been a big seller of bonds in the secondary market.***

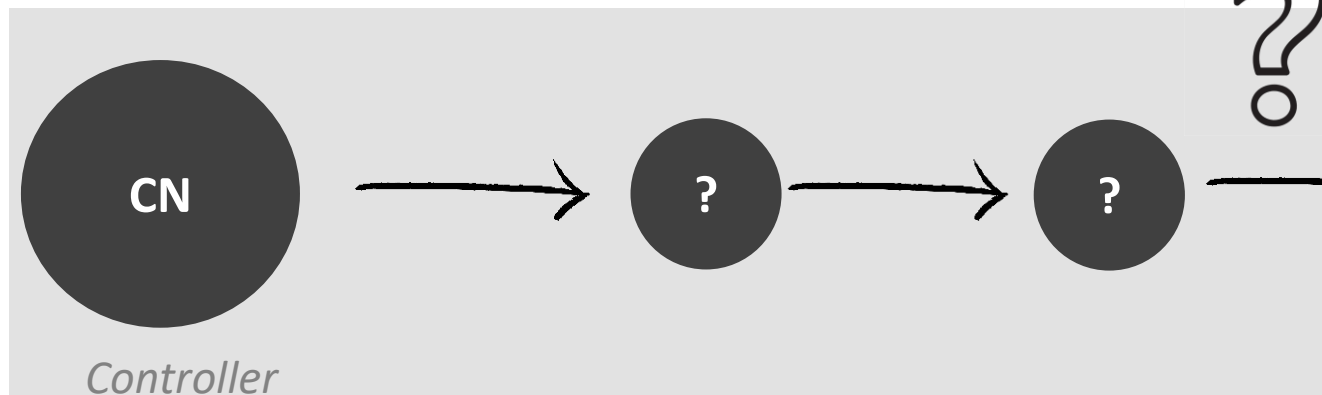


1

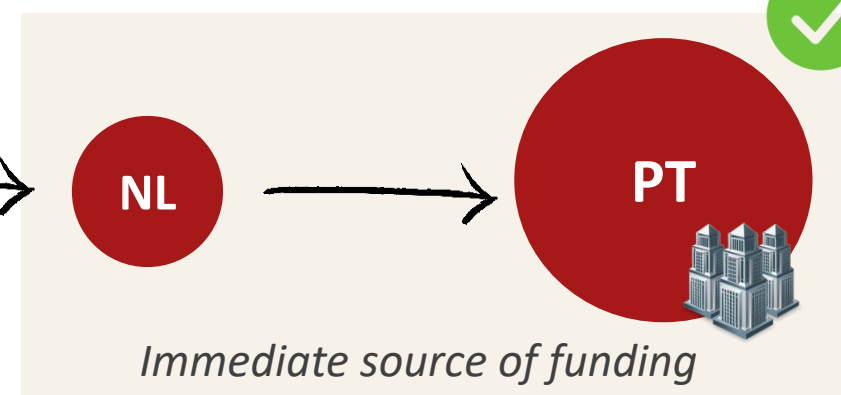
Applying network analysis to the foreign investment - **WHY?**

01

LACK OF DATA

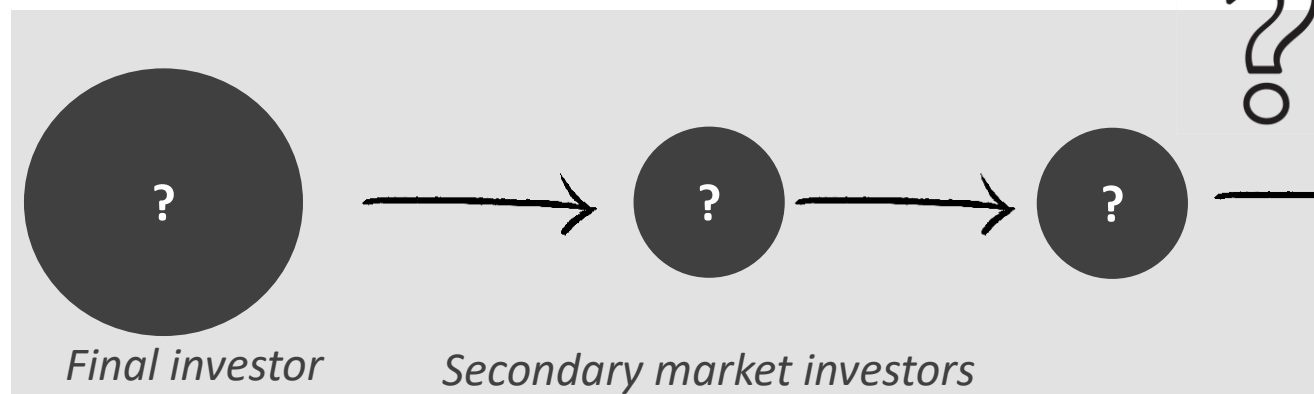


FDI STATISTICS

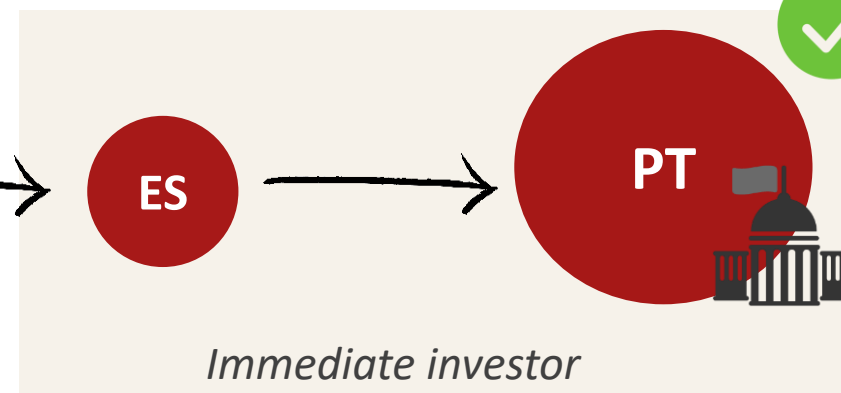


02

LACK OF DATA



PORTFOLIO INV. STATISTICS



2

Applying network analysis to the foreign investment - HOW?

Use the **network analysis** to map foreign investments:

FOREIGN DIRECT
INVESTMENT

AND

PORTFOLIO
INVESTMENT

01

IDENTIFICATION OF PATTERNS

Identify the **patterns** and **shortest paths** between the immediate and ultimate investors

02

ESTABLISHING TRENDS

Establish **trends** and describe the relations between countries over time

03

RESULTS ILLUSTRATION

Illustrate the results of the network in an intuitive **web application**

04

PREDICTIONS

Use the network analysis to **predict** the ultimate **investor** and **intermediaries**



2

Applying network analysis to the foreign investment - **HOW?****DATA
SOURCES**

IMF Coordinated Portfolio Investment Survey - annual data from 2009-2019

IMF Coordinated Direct Investment Survey - annual data from 2009-2019

DISTANCE

Countries with larger bilateral stock are closer, thus are at a shorter distance from each other (the weight of the link is lighter):

$$\phi_{ij} = \frac{1}{|f_{ij}| + |f_{ji}|}$$

STOCKS



**DIRECTIONAL
PRINCIPLE – FDI
INWARD AND
FDI OUTWARD**



**PI ASSETS AND PI
LIABILITIES**



**COMBINED
ASSETS AND
LIABILITIES ***



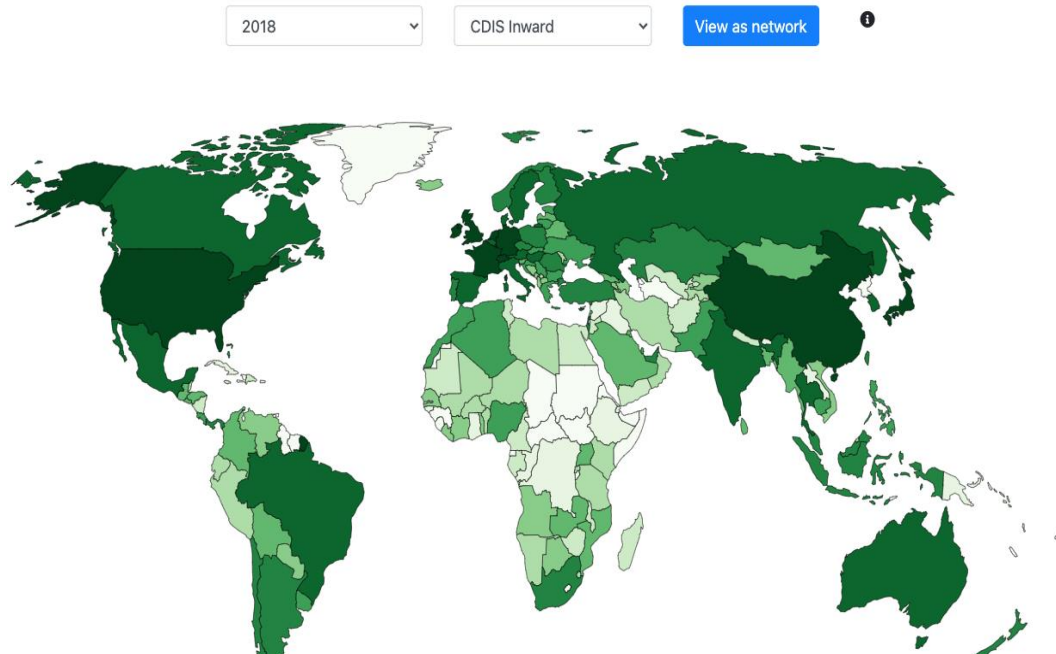
* **Assets** = FDI outward + PI assets

Liabilities = FDI inward + PI liabilities



3

THE PLATFORM: FI-NETWORKS



THE WEBSITE IS HOSTED ON THE **NETLIFY** PLATFORM!

HTML5
AND CSS



D3.JS AND
D3PLUS

✓ For building the **skeleton of the website**

✓ Placing the **selection menus** and **descriptions**

✓ **D3.js** - JavaScript library for developing **dynamic, interactive data visualizations**. It uses Scalable Vector Graphics (**SVG's**)

✓ It builds interactive visualizations, such as the **network visualization, world map, treemap and bar and line charts**



3

THE PLATFORM: FI-NETWORKS

F
I
-
N
E
T
W
O
R
K
S

THE DATA BEHIND
THE
VISUALIZATIONS
IS COMPUTED IN
PYTHON



SVG'S HAVE
LOW MEMORY
REQUIREMENT
AND MAINTAIN
QUALITY WHEN
SCALED



OUTPUTS
STORED IN
CSV/JSON FILES,
ORGANIZED IN
A FOLDER
STRUCTURE



HOSTED ON
NETLIFY, A
SERVERLESS
HOSTING SERVICE
PULLING SOURCE
CODE FROM
GITHUB



VISUALIZATIONS
BUILT WITH
JAVASCRIPT,
LEVERAGING ON
SVG PROPERTIES



5 PAGES:
HOME
NETWORK
METHODOLOGY
ABOUT
DOWNLOADS



THE PLATFORM: Fi-networks:

<https://fi-networks.com>



01

Network science illustrates the enormous analytical power to **predict the ultimate investors for FDI and Portfolio Investment**

02

Few countries including the **US, NL, LU, CN, HK** and **the UK** are **the main global intermediators**

03

FI-NETWORKS, web application to:

- i) communicate these findings in an **interactive data visualization platform**
- ii) quick exploration and discovery of relevant **partnerships/investment paths**



WHAT'S NEXT?

EXPLORE HOW THE WEB APP CAN BE
A FUNDAMENTAL TOOL FOR
REGULATORS AND CENTRAL BANK
OFFICIALS



IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Cross-currency swap market through the lens of OTC derivative transaction data – impact of Covid-19 and subsequent recovery¹

Kazuaki Washimi and Rinto Maruyama,
Bank of Japan

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Cross-Currency Swap Market through the Lens of OTC Derivative Transaction Data

Impact of COVID-19 and Subsequent Recovery

Kazuaki Washimi and Rinto Maruyama

Abstract

Cross-currency swaps are one of the major US dollar funding tools for non-US banks. While their developments have attracted international attention, data for gauging transaction details are limited since these swaps are over-the-counter transactions, not trades on an exchange. This report provides an overview of the Japan's cross-currency swap market with over-the-counter derivative transaction data collected in Japan. Then it briefly reviews the impact of the COVID-19 crisis on these transactions around the spring of 2020. A data analysis indicates that major banks continued transactions as a market maker by breaking trades into smaller blocks and diversifying the counterparties, while smaller banks who do not actively engage in normal times were found to have participated in trading.

Keywords: Cross-currency swaps; Trade repository; Market structure

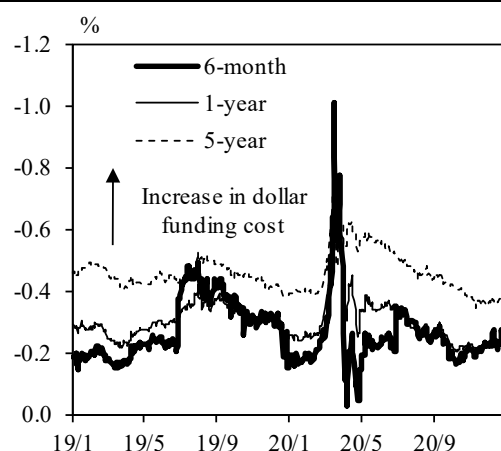
JEL classification: F31; G15

1. Introduction

In recent years, foreign currency funding by banks, especially the US dollar funding, has attracted international attention as cross-border claims have been on an increasing trend. In particular, the spring of 2020, when the market was in turmoil due to the spread of COVID-19, saw a sharp increase in the dollar funding cost (Chart 1). While cross-currency swaps are one of the major US dollar funding tools, concern over its resiliency in the wake of shocks has been pointed out.¹ Increasing understanding of the market structure of the cross-currency swap market will be key for assessing the stability of foreign currency funding going forward.

Cross-Currency Basis (USD/JPY)

Chart 1



¹ As of December 31, 2020.

Sources: Bloomberg

Up until now, there has been little available data for gauging transaction details, since cross-currency swap transactions are over-the-counter (OTC) transactions rather than on-exchange trades. Previous studies have largely relied on the cross-currency basis which represents the dollar funding premium. It has been pointed out that more data would be needed to gauge market liquidity and functioning in the cross-currency swap market.² In relation to this, the Japanese authorities have collected the OTC derivative transaction data as part of a global initiative to expand the data.³

¹ For instance, these points are raised in Chapter 5 of the IMF Global Financial Stability Report (October 2019).

² In BIS (2020) "US dollar funding: an international perspective," CGFS Papers No. 65, data collection and information sharing by country authorities are listed as challenges ahead in US dollar funding. In particular, it is expected that data analysis will be deepened to enhance transparency in OTC derivative transaction data.

³ As discussed later, the Financial Services Agency in Japan started to collect the OTC derivative transaction data in April 2013. The following webpage releases the outstanding amounts (notional amount basis) as of March each year.

<https://www.fsa.go.jp/status/otcreport/index.html>

This report provides stylized facts on the cross-currency swap market in Japan with OTC derivative transaction data. Then it briefly reviews the impact of the COVID-19 crisis on the transactions around the spring of 2020. The focus is placed on cross-currency swaps, given that the OTC derivative transaction data in Japan do not cover FX swaps.⁴

2. What Are Cross-Currency Swaps?

A cross-currency swap is a contract in which one party exchanges one currency for a second currency (e.g., US dollar for Japanese yen) with another party for a certain term typically longer than one year. Though there are various types of contracts in cross-currency swaps, the basis swaps for USD/JPY—known as typical interbank transactions—involve the exchange of principal and interest. In this case, the swap exchanges floating interest in the form of “USD three-month reference rate” and “JPY three-month reference rate plus cross-currency basis (alpha)”.⁵ In the latter, “cross-currency basis” indicates the US dollar funding premium, where a negative value (alpha) means a relatively strong demand for the US dollar from a demand-supply perspective.

According to the BIS triennial survey on the global turnover,⁶ the Japanese yen has the largest cross-currency swap transaction volume against the US dollar, followed by the Euro, UK pound, and Australian dollar (Chart 2). Taking a look at cross-currency basis by currency pair (Chart 3), while those of the Australian dollar and New Zealand dollar hovered in positive territory, those of the Japanese yen and European currencies stayed negative. It has been pointed out that US dollar demand

⁴ With respect to the main difference from cross-currency swap, FX swaps are centered on the short-to-medium term, less than one year. That said, as the paper below pointed out, there is an arbitrage between them since the two can be considered to be transactions which have similar economic impacts.

Yasuaki Amatatsu and Naohiko Baba (2007) “Price Discovery from Cross-Currency and FX Swaps: A Structural Analysis,” Bank of Japan Working Paper Series No. 07-E-12

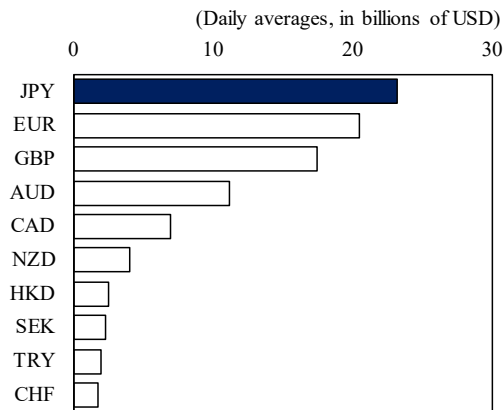
In Japan, OTC derivative transaction data do not cover FX forwards and FX swaps as they are not classified as OTC derivatives in the Financial Instruments and Exchange Act. As mentioned later, the country authorities have collected the OTC derivative transaction data since the global financial crisis as a global initiative. That said, the data coverage varies across countries and regions. For instance, FX forwards and FX swaps are covered in Europe.

⁵ While Libor has been used as a reference rate, risk-free rates are expected to gradually replace it, reflecting the interest rate benchmark reform.

⁶ The BIS, in cooperation with the world’s central banks, conducts the *Triennial Central Bank Survey of Foreign Exchange and Over-the-counter Derivatives Markets*. It collects data from major financial institutions around the world (called “reporting dealers”) covering comprehensive topics consistent with international protocols.

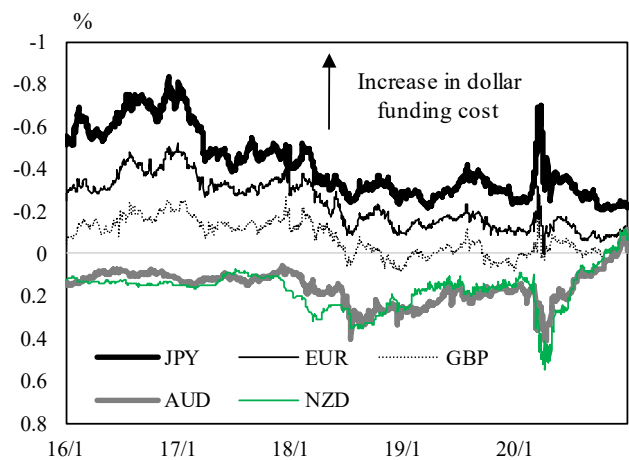
varies across currencies depending on the direction of monetary policy and hedging needs⁷ of banks and institutional investors.⁸

Cross-Currency Swap Turnover (against USD, 2019) Chart 2



¹ As of April 2019. The same applies to charts below.
Sources: BIS "Triennial Central Bank Survey of Foreign Exchange and Over-the-counter (OTC) Derivatives Markets"

Cross-Currency Basis (against USD) Chart 3



¹ A one-year term. As at December 31, 2020.
Sources: Bloomberg

⁷ For instance, Japanese institutional investors such as life insurers use cross-currency swaps and FX swaps to hedge their foreign currency denominated investments.

⁸ The aforementioned BIS report pointed out, for instance, that Australian banks are US dollar providers in the cross-currency swap market on a net basis. The following report discusses the reasons for the rise in the US dollar funding premium in recent years.

Fumihiko Arai, Yoshibumi Makabe, Yasunori Okawara, and Teppei Nagano (2016) "Recent Trends in Cross-currency Basis," Bank of Japan Review Series 2016 -E-7

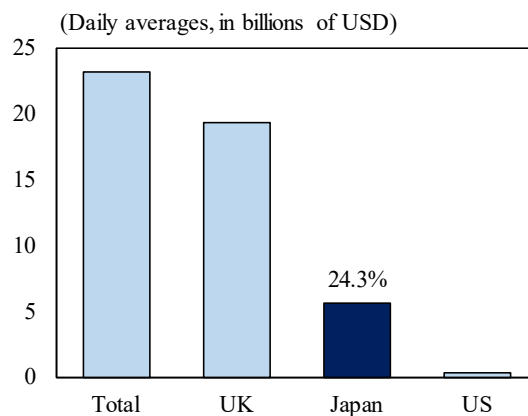
3. Trends in Japan's Cross-Currency Swap Market

Market Structure from BIS Triennial Survey

According to the BIS triennial survey, Japan accounts for about a quarter of trading volume in cross-currency swaps for USD/JPY⁹, being second only to the UK (Chart 4). A breakdown by counterparty shows that financial institutions account for the majority of transactions in Japan, while it also points to a certain presence of institutional investors and hedge funds, etc., globally (Chart 5).

Cross-Currency Swap Turnover by Country (USD/JPY, 2019)

Chart 4

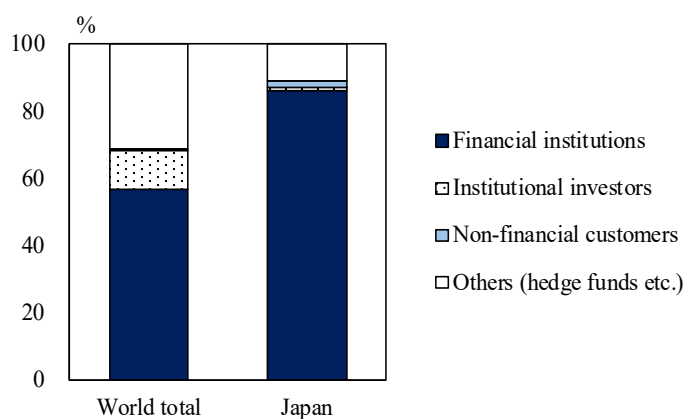


¹ Total does not match the sum of country breakdowns as it excludes the double-counting of transactions between local and cross-border dealers.

Sources: BIS "Triennial Central Bank Survey of Foreign Exchange and Over-the-counter (OTC) Derivatives Markets"

Cross-Currency Swap Turnover by Counterparty (USD/JPY, 2019)

Chart 5



¹ Share by counterparty from the viewpoint of reporting dealers.

Sources: BIS "Triennial Central Bank Survey of Foreign Exchange and Over-the-counter (OTC) Derivatives Markets"

⁹ This report focuses on USD/JPY, since USD/JPY constitutes about 95 percent of cross-currency swap transactions against the yen according to the BIS turnover survey (2019) in Japan.

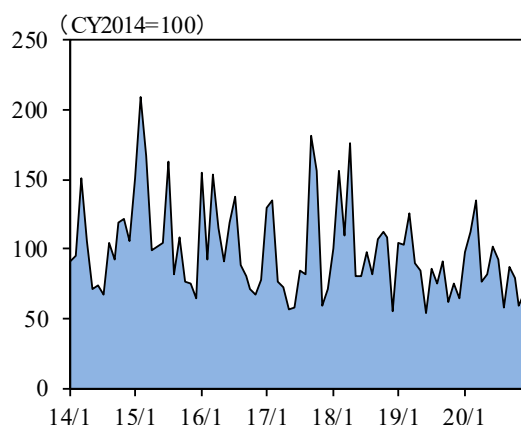
Stylized Facts from Granular Data

This section is aimed at providing more detailed observations using the OTC derivative transaction data. The OTC derivative transaction data in this report are transaction-by-transaction data based on the reporting from trade repository and financial institutions in Japan. The data cover the transactions where at least one of the parties is Japanese financial institutions or foreign financial institutions based in Japan.¹⁰ The country authorities (Financial Services Agency in Japan) have collected transaction data for gauging systemic risk and improving transparency in the OTC derivative market in light of the lessons from the global financial crisis. Nonetheless, there have not been a large number of analyses globally,¹¹ given that the confidentiality of each transaction should be safeguarded and that data cleansing needs substantial time and cost.

Aggregating the new transactions on a monthly basis, the trading volume in cross-currency swap market for USD/JPY has stayed more or less the same albeit with fluctuations, showing no signs of extreme swings in recent months (Chart 6). Trade counts have been mostly in a range of 300 and 600 per month¹² (Chart 7), except for a spike in March 2020 (which will be discussed later).

Cross-Currency Swap Turnover in Japan (USD/JPY)

Chart 6



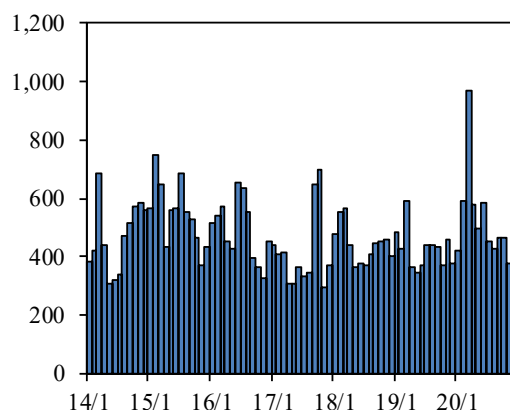
¹ As at December 2020. The same applies to charts below.

Sources: OTC Derivative Transaction Data

¹⁰ The precise scope of reporting entities is defined by Article 6 of the "Cabinet Office Ordinance on the Regulation of Over-the-Counter Derivatives Transactions." Specifically, the data cover the transactions where either or both of the parties are a Financial Instruments Business Operator that conducts Type I Financial Instruments Business, a bank, Shoko Chukin Bank, Ltd., Development Bank of Japan Inc., a federation of Shinkin banks (the district of which is the entire nation) or Norinchukin Bank or an insurance company.

¹¹ In relation to US dollar funding, for instance, the following paper presents breakdowns of FX forwards (and the forward leg of FX swaps) by currency and by maturity using trade repository data in Europe. Cielinska et al. (2017) "Gauging Market Dynamics using Trade Repository Data: The Case of the Swiss Franc De-pegging," Bank of England Financial Stability Papers, 41

¹² In terms of trade amount and counts, no distinct seasonality was observed, whereas transactions for a relatively long tenor tended to increase at the end of the fiscal year (March).

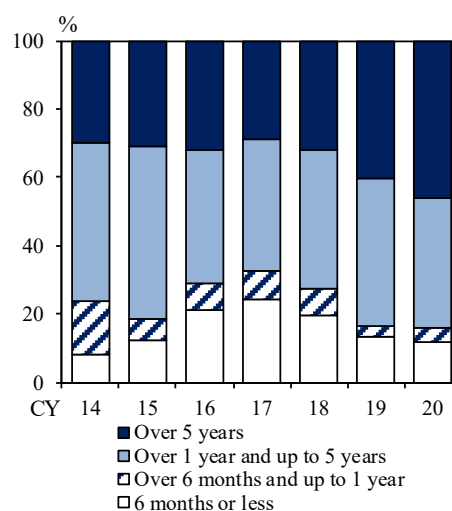


Sources: OTC Derivative Transaction Data

Breaking down the trading amount by maturity, there is no significant change in the term structure except for a slight increase in transactions for longer than five years recently (Chart 8). By type of counterparty¹³, foreign banks and securities companies (including foreign securities companies) are major dollar providers, while major banks¹⁴ are the main dollar takers (Charts 9 and 10).

Share of Trading Amounts by Maturity

Chart 8



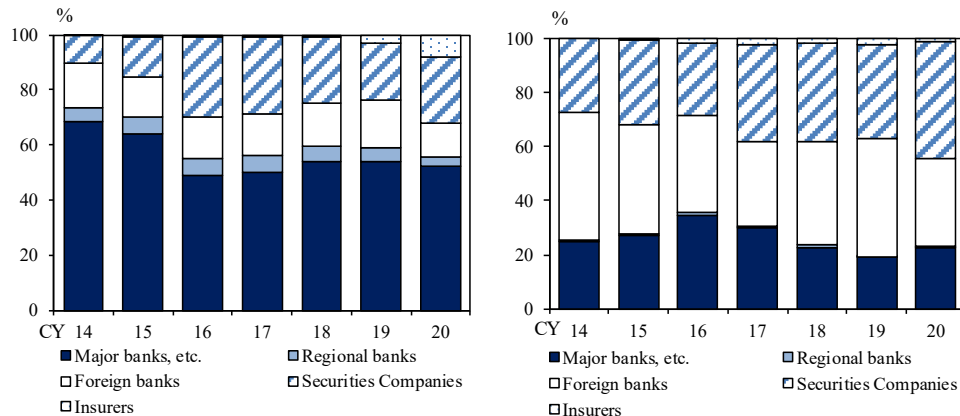
Sources: OTC Derivative Transaction Data

¹³ The transactions by non-financial corporates and others are excluded from the aggregation by type of counterparty.

¹⁴ Major banks. etc. include major banks, Shokochukin Bank, Development Bank of Japan, Shinkin Central Bank and Norinchukin Bank.

US Dollar Taker and Provider by Sector

Chart 9 and 10



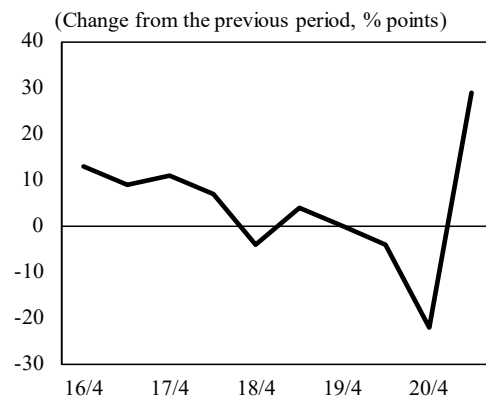
Sources: OTC Derivative Transaction Data

Evolution around the Time of COVID-19 Crisis

This section summarizes the market developments of cross-currency swaps around the spring of 2020, when the dollar funding conditions remarkably deteriorated, using the OTC derivative transaction data. The existing data suggest a sharp increase in the dollar funding cost in mid-March, as mentioned above (Chart 1). Various surveys indicate a notable deterioration in the swap market functioning in the spring of 2020 (Chart 11). Another survey suggests that market participants coped with the strains in the dollar funding market by breaking trades into smaller blocks and diversifying counterparties (Chart 12).

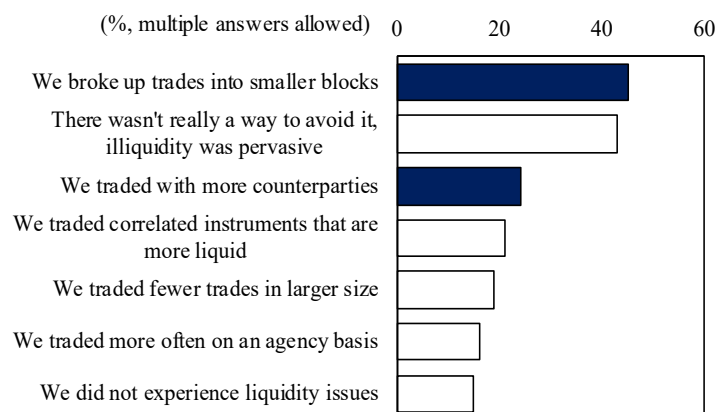
Diffusion Index on Functioning in Swap Market

Chart 11



¹ Covers FX swaps and cross-currency swaps. As of April and October for each year. Diffusion index is calculated as percentage share of those responding "high" minus percentage share of those responding "low".

Sources: Tokyo Foreign Exchange Market Committee "Turnover Survey of Tokyo Foreign Exchange Market"



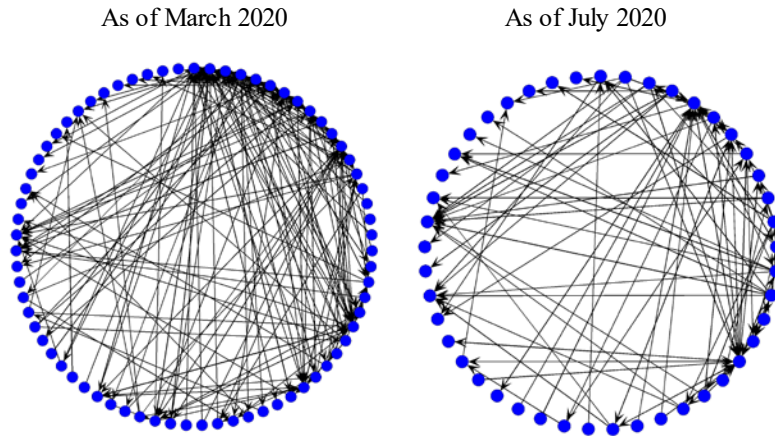
¹ Questionnaire survey on liquidity in overall swap markets. Survey respondents are 172 market participants globally.

Sources: ISDA/Greenwich Associates 2020 COVID Crisis Swaps Liquidity Survey

Nevertheless, the quote prices for currency basis and these surveys do not reveal what entity traded and in what terms (amounts, rates) the transactions were made. The OTC derivative transaction data can be used to assess the behavior of market participants by comparing before and after the spring of 2020 based on the network measures and the spreads between the reference rates and contract rates.

Network analysis is known as a method for visualizing the interconnectedness of financial market transactions. It can check on the changes of trading behavior by quantifying and showing how each entity is connected to other entities through transactions. For instance, the network topology indicates that the transactions were interconnected in a more complicated manner in March 2020 relative to July 2020 when the market had recovered to a certain degree (Chart 13).¹⁵

¹⁵ The attributes such as the name and type of each entity are not shown in order to maintain confidentiality.

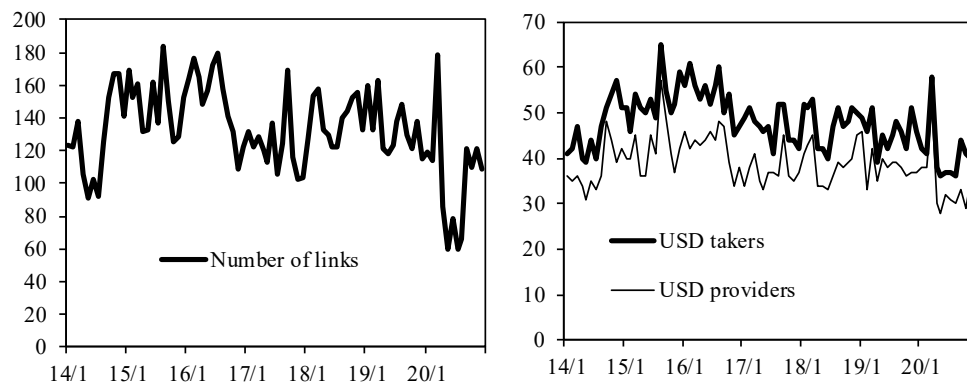


¹ Covers new transactions of cross-currency swaps for USD/JPY (The same applies to the charts below). The circles indicate the players, while the arrows represent the dollar flows.

Sources: OTC Derivative Transaction Data

Number of Transaction Links, US Dollar Taker and Provider

Chart 14 and 15



Sources: OTC Derivative Transaction Data

The number of links in the network can capture how many counterparties each player is connected to in order to see the changes over a relatively long period.¹⁶ They increased sharply in March 2020, and stayed at relatively low levels in April after the first declaration of a state of emergency, and then recovered to some degree in June afterwards with the emergency declaration lifted (Chart 14). Moreover, the

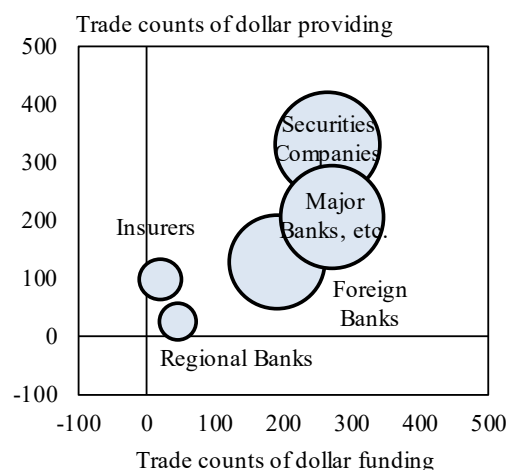
¹⁶ For instance, the following study shows an increase in the number of links in the Japanese Government Bond market in the wake of a sharp increase in interest rates.

Toshiyuki Sakiyama and Tetsuya Yamada (2016) "Market Liquidity and Systemic Risk in Government Bond Markets: A Network Analysis and Agent-Based Model Approach," IMES Discussion Paper Series No. 2016-E-13

numbers of dollar providers and takers¹⁷ show similar trends (Chart 15). Meanwhile, trade counts increased in March 2020, while there was no remarkable change in trading amount, suggesting that the average size of trades became smaller (Charts 6 and 7). By type of financial institution, major banks and securities companies recorded large trade counts on both the funding and supply sides, which indicates that they functioned as market makers (Chart 16).¹⁸

Trade Counts of Dollar Funding and Provision (March 2020)

Chart 16



¹ The size of bubbles represents the centrality of each player within the transaction network based on PageRank measure.

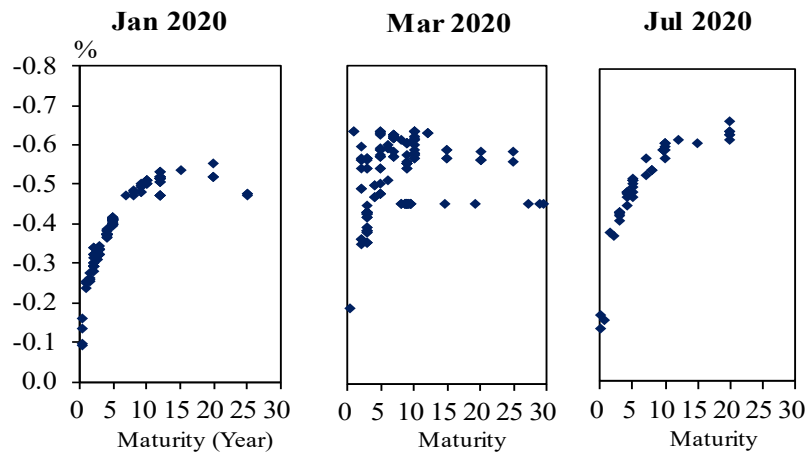
Sources: OTC Derivative Transaction Data

These data suggest that major banks continued transactions as a market maker by breaking trades into smaller blocks and diversifying the counterparties, while smaller banks who do not actively engage in normal times participated in trading amid the deteriorating dollar funding conditions in March 2020. Simultaneously, the distribution of the spreads between the reference rates and contract rates by maturity shows a temporary widening of spreads for the short-to-medium term in March 2020 and a gradual recovery entering the summer as a result of an expansion of dollar swap lines among the central banks (Chart 17).¹⁹

¹⁷ The network analysis in the report uses entity-based aggregation. It should be noted that the numbers of dollar providers, takers and links could vary depending on the aggregation method or data cleansing.

¹⁸ They are considered to play the role of market makers as the trade counts of major banks and securities companies for both dollar funding and supply are high in normal times.

¹⁹ The spreads point to a similar trend even after the counterparty risks (characteristics of the counterparty) are taken into account.



¹ Covers new transactions. It should be noted that there are some missing data and discrepancies in units (some are reported in percent while others are reported in level). Spreads (currency basis) are calculated as reference rates (e.g., LIBOR) minus contracted rates. A negative value indicates an increase in dollar funding cost.

Sources: OTC Derivative Transaction Data

Concluding Remarks

This report provides stylized facts on the cross-currency swap market in Japan with OTC derivative transaction data and reviews the developments around the time of the COVID-19 crisis using network analysis.

The following are the two main implications. First, transaction-level data can be used to visualize the characteristics of the cross-currency swap market in Japan including dollar funding and supply structure by type of market participant that are not necessarily covered in the existing data. Second, the use of timely and high-frequency data could make it possible to examine the market liquidity and the changes in the trading behavior of market participants in a more timely and detailed manner compared to the existing data.²⁰ Continued use of these transaction-level data together with market intelligence would help assess the foreign currency funding markets more carefully.

A potential caveat is that the report captures only a part of the various US dollar funding activities (and is limited to transactions in Japan). Therefore, it is expected that future research would analyze the FX swap market from different angles not covered by Japan's OTC derivative transaction data.

Going forward, accumulating data together with improving data cleansing could help increase transparency in the OTC derivative markets. International discussion including that by the Financial Stability Board (FSB) has expressed an expectation of

²⁰ For the trading volume in the cross-currency swap market, the Tokyo Foreign Exchange Market Committee publishes a turnover survey twice a year, in addition to the aforementioned BIS turnover survey. The Bank of Japan releases Regular Derivatives Market Statistics in Japan twice a year for the outstanding data.

further progress in data development and analyses to gauge the global picture.²¹ In this respect, the Bank of Japan (BOJ) has actively engaged in strict data management and transaction-level data analysis to gauge market liquidity and functioning, as well as setting up the Financial Market Data Planning Group in March 2018. Regarding the OTC derivative markets, the BOJ has released an analysis on FX options and will continue to pursue such initiatives.²²

²¹ These are also mentioned in FSB (2019) "OTC Derivative Market Reforms 2019 Progress Report on Implementation."

²² For instance, see the following.

Kazuaki Washimi (2020) "Revisiting Determinants of Investor Sentiment in the FX Option Market by Machine Learning Approaches," Proceedings of 2020 IEEE Symposium Series on Computational Intelligence (SSCI)

In addition, the BOJ has released the FSB repo data (including cross-currency repo) since 2020. For instance, see the following.

SASAMOTO Kana, NAKAMURA Atsushi, FUJII Takanori, SEMBA Takashi, SUZUKI Kazuya, and SHINOZAKI Kimiaki (2020) "New Initiatives to Improve the Transparency of Securities Financing Markets in Japan: Publication of Statistics on Securities Financing Transactions in Japan," Bank of Japan Review Series 2020-E-1

Cross-Currency Swap Market through the Lens of OTC Derivative Transaction Data: Impact of COVID-19 and Subsequent Recovery

Kazuaki WASHIMI

Bank of Japan, Research and Statistics Department

※ The views expressed in the presentation are those of the author and do not necessarily represent those of the Bank of Japan.

Motivation

- Use of OTC derivatives data at a transaction level
- Gauging US dollar funding activities by banks
 - ✓ Attracting international attention, but data are limited
 - ✓ Conditions deteriorated in the wake of COVID-19

2



- A) Fact finding using transaction-level data**
- B) Network analysis on trading behavior**

Overview of OTC Derivative Data

Aggregate

| Category | Outstanding (Tril. Yen) |
|----------------|----------------------------|
| Interest Swaps | 4,640 |
| FX | 87 |
| Credit | 28 |
| Equity | 20 |

Transaction-Level

| ID | Player 1 | Player 2 | Contract Date | Maturity Date | Notional Amount ... |
|----|----------|----------|------------------|------------------|---------------------|
| 1 | A | B | 2018/1/6 | 2019/1/6 | X bill. Yen |
| 2 | A | C | 2018/3/2 | 2018/9/2 | X bill. Yen |
| 3 | | | | | |
| 4 | | | | | |
| 5 | | | | | |
| 6 | | | | | |
| 7 | | | | | |
| 8 | | | | | |
| 9 | | | | | |
| : | | | | | |

More than millions of cells per month

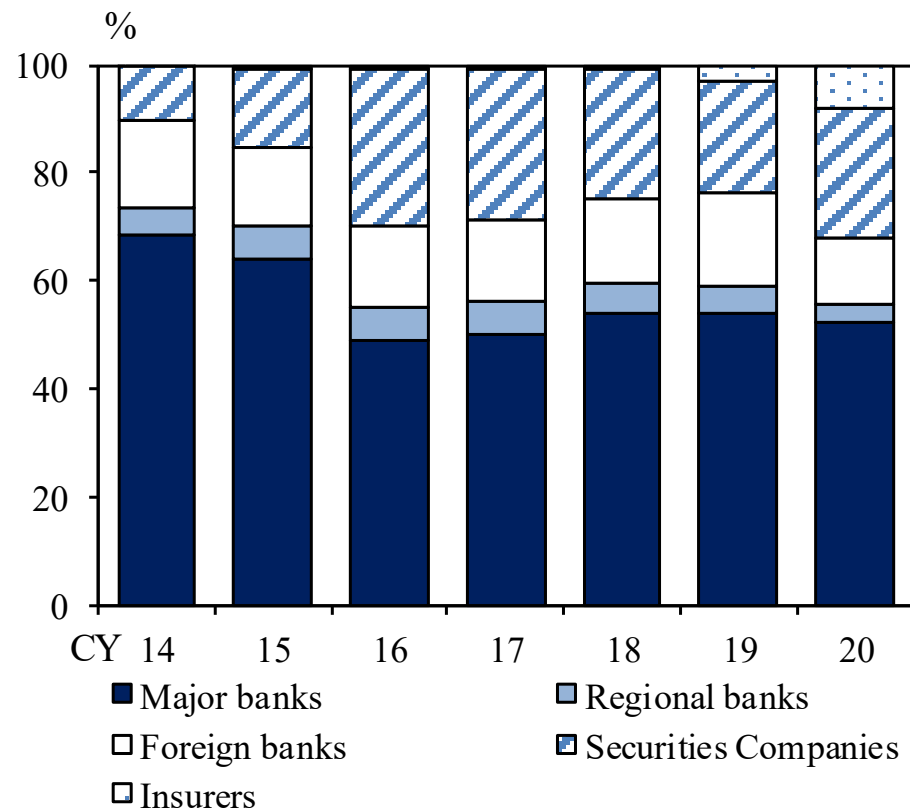
3

Safeguard confidentiality, massive data cleansing needs

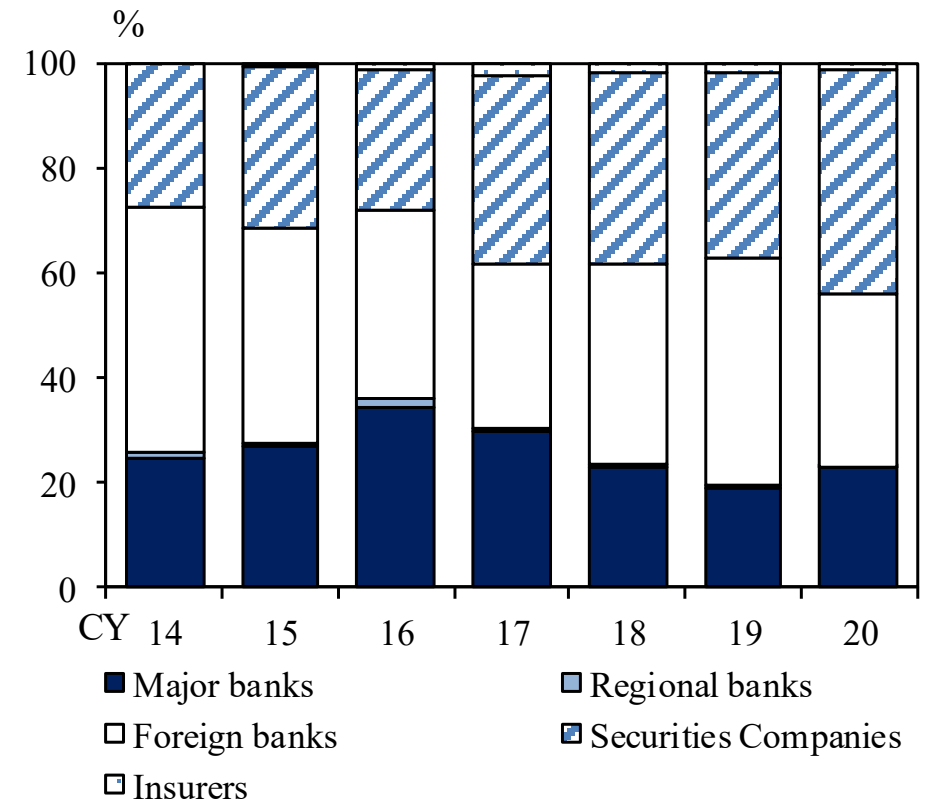
▶ Room for further improvement for transparency

US Dollar Taker/Provider by Sector

US Dollar Taker by Sector



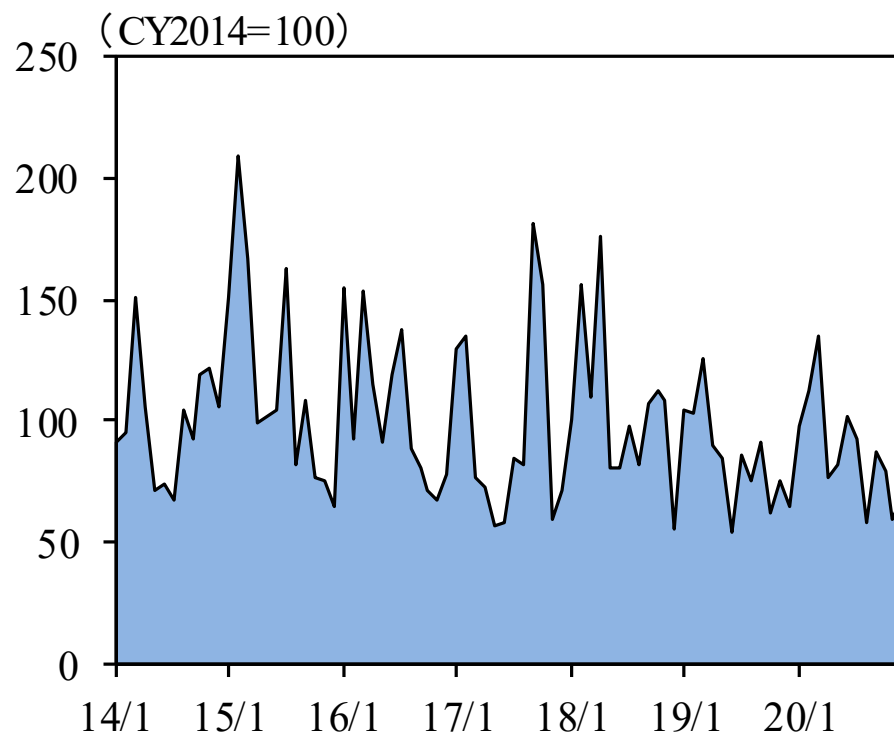
US Dollar Provider by Sector



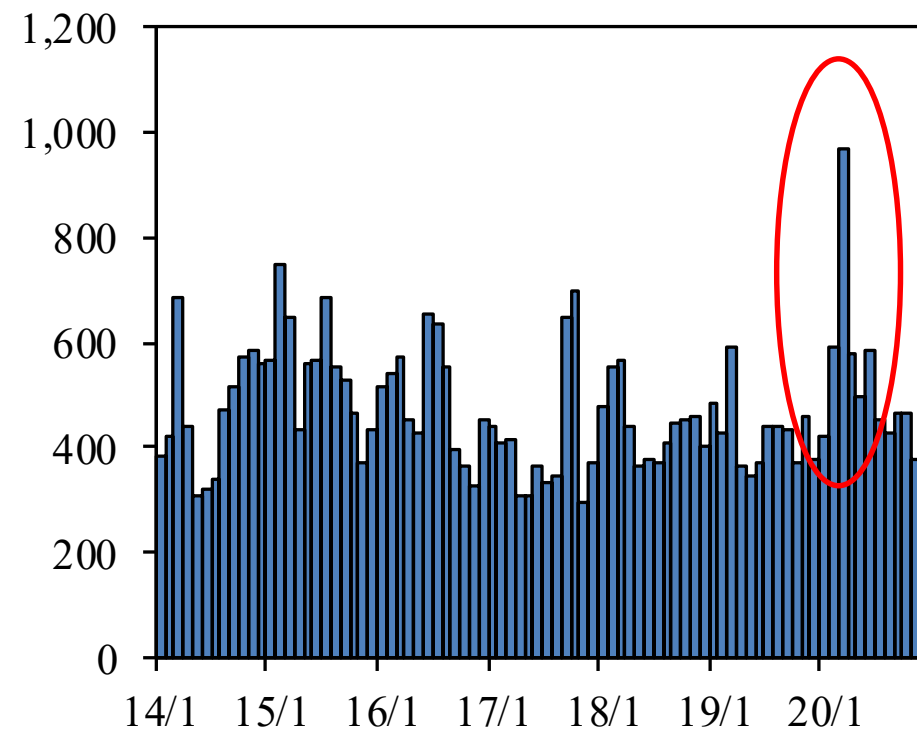
Source: OTC Derivative Transaction Data

Trading Volume and Counts

Trading Volume



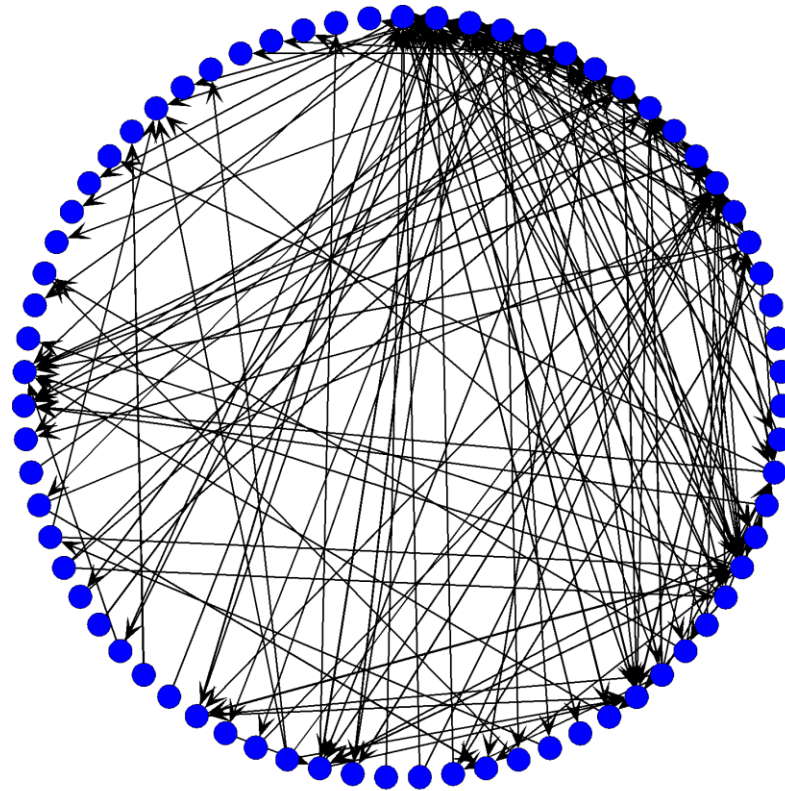
Trading Counts



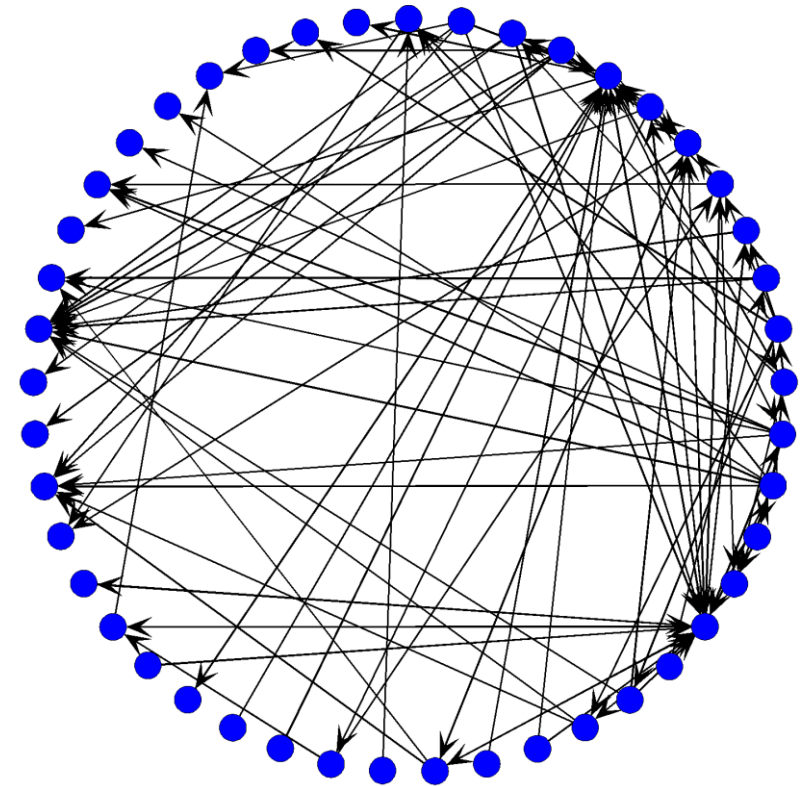
Source: OTC Derivative Transaction Data

Comparison of Network Topology

As of March 2020



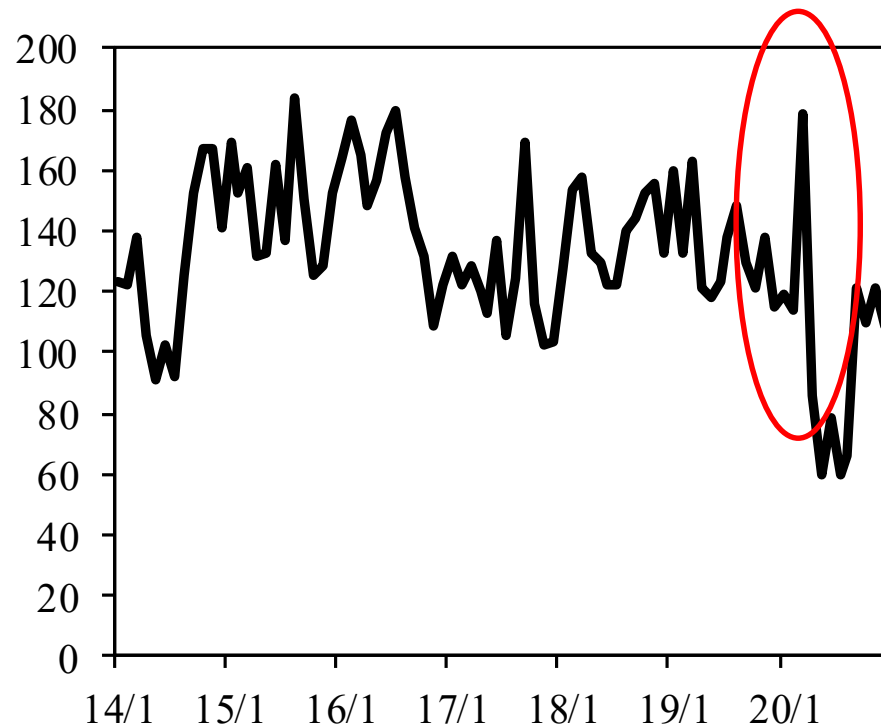
As of July 2020



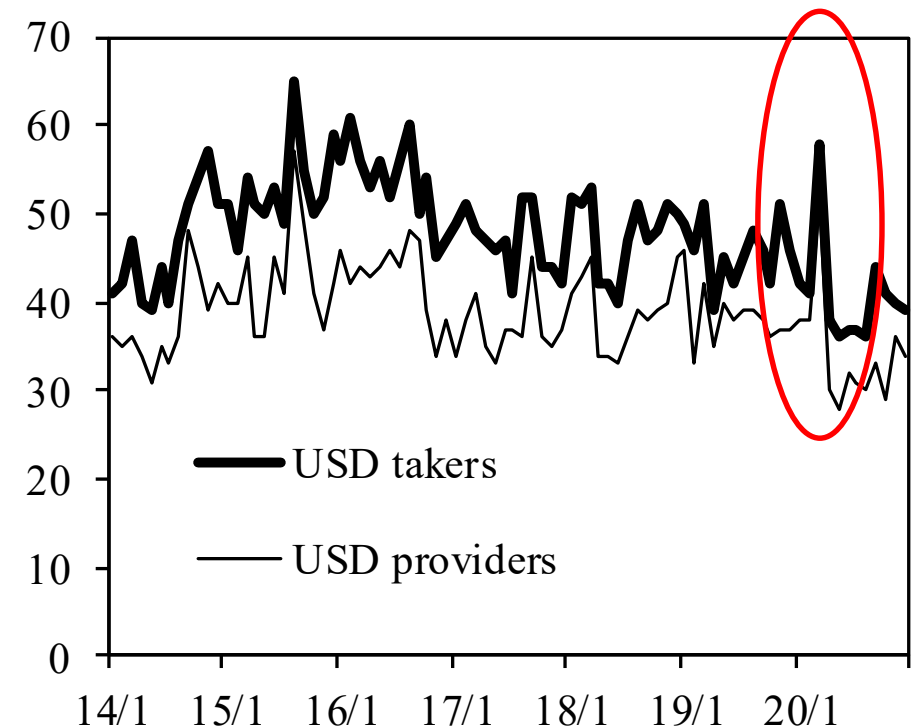
Note: The circles indicate the players, while the arrows represent the dollar flows.
Source: OTC Derivative Transaction Data

Network Measures

Number of Transaction Links



Number of Dollar Taker/Provider



Source: OTC Derivative Transaction Data

Summary

- Transaction-level data can help capture market structure and gauge trading behavior of market participants
- In the wake of COVID-19 crisis, it appears that trades were broken into smaller blocks and traded with more counterparties
- Future work can benefit from assessing other dollar funding tools on a global scale

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Changes in the lending activity of banks in Poland, including the portfolio of non-financial corporate loans¹

Aneta Kosztowniak,
Narodowy Bank Polski

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Changes in the lending activity of banks in Poland, including the portfolio of non-financial corporate loans

Aneta Kosztowniak¹, Economic Analysis and Research Department, Narodowy Bank Polski, SGH Warsaw School of Economics

Abstract

The study aims to identify changes in the lending activity of banks, including the loan portfolio of non-financial corporation's (NFCs) and main determinants of non-performing loans (NPLs) in Poland in the period Q1.2009.Q1.-Q4.2021. This study presents the changes of average interest on corporate loans, capital requirements and mainly regulations concerning the credit policy. The article presents the differences in the NPL rates and debt servicing costs in Poland compared to other countries of the Visegrad Group, as well as Germany and France. The author presents the results of an overview of NPL research on EU countries. In modelling the quality of the portfolio of NPLs granted to NFCs, mainly the following variables are taken into account: market and financial variables of corporations – determine the possibility of servicing loans, and variables of banking conditions – serving as capital hedging against an increase in banking risk. In the analyzed period, banks pursued a liberal policy of interest rates on loans while maintaining adequate capital requirements. The analysis of NPL changes shows that there was a long-term downward trend confirming the improvement in the quality of the portfolio of loans of NFCs in the analyzed period. However, the last quarters (during the COVID pandemic), brought an increase in the NPL ratio, respectively: Q2.2020 (8.7%) and Q4.2020 (9.0%). Results of the impulse function confirmed, that the NPLs showed declining trends in response to impulses from: NPL's own changes, GDP, CPI, WIBOR, ROAC, GFCF, TOFSP and increasing trends in response to changes: CROAC, GTPR, AIRCL, CAR and CRofCR. Results of variance decomposition indicate that the main pillar of the explanation of NPL changes were: GDP, GFCF as well as CPI and WIBOR. Also, the results of the NPLs research confirmed the procyclical nature of lending activity in Poland in the verified years.

Keywords: Banks, credit policy, asset quality, loan portfolio, NPLs, NFCs, credit risk, UE, Poland

JEL classification: E4, E5, G2

¹ The views expressed in this study are the views of the author and do not necessarily reflect those of NBP.

Contents

| | |
|--|----|
| Changes in the lending activity of banks in Poland, including the portfolio of non-financial corporate loans | 1 |
| 1. Introduction..... | 3 |
| 2. Lending activity, capital requirements and regulations of banks in Poland | 4 |
| 3. Non-performing loans (NPLs) – regulations and recommendations | 5 |
| 4. Review of research in the field of NPLs | 7 |
| 5. Differences in the NPLs between EU countries and the structure of the loan portfolio of NFCs in Poland | 9 |
| 6. Research procedure - mainly determinant of NPLs of NFCs in Poland | 12 |
| 7. Empirical results – impulse response functions and variance decompositions.... | 14 |
| 8. Conclusions and way forward..... | 16 |
| Annex | 18 |
| References..... | 19 |

1. Introduction

The bank's credit policy is adjusted to its specificity and changing economic conditions, while maintaining the security of loan portfolio management. In terms of risk assessment in the banking sector, the Basel regulations are of great importance (Basel Committee on Banking Supervision (2014)). An important group of instruments used to mitigate credit risk are exposure concentration limits: internal (created by banks) and external (supervisory), which are defined in legal regulations (including numerous directives).

A significant problem related to the lending activity is the quality of the loan portfolio. The quality of the loan portfolio depends on the exposure to banking risks, especially credit risk. Among the credit risk factors, there are internal factors – endogenous inside business entities and external factors – endogenous in the environment of enterprises and independent of them. The bank's credit policy and credit portfolio management are adjusted to the internal and external factors (independent from them).

Among the many reasons for the quality of the portfolio, the following should be mentioned:

- overall increase in financial risk in the economy,
- sudden changes in the economy, resulting on the one hand in the need for quick adjustments of all business entities, and on the other hand – reducing the possibility of hedging against risk,
- increasing number of bankrupting enterprises, which results from the increase in the number of companies established in order to obtain high ad hoc profits, often on speculative transactions, increasing risk in foreign trade due to the increase in the number of heavily indebted countries and with a high inflation rate,
- growing competition on the banking services market, limiting the banks' ability to choose customers.

When analyzing credit risk, it is important to distinguish between individual risk, portfolio risk and risk of individual client (natural person) from the credit risk of institutional client (enterprise). Individual assessment of the creditworthiness of enterprises is determined on the basis of many financial indicators. In addition, banks for internal needs forecast changes in the quality of loans, for example to enterprises, using: credit exposures, the results of corporate financial statements and numerous macroeconomic indicators. One of the final effects of the assessment of banks' lending policies is changes in the portfolio of non-performing loans (NPLs).

2. Lending activity, capital requirements and regulations of banks in Poland

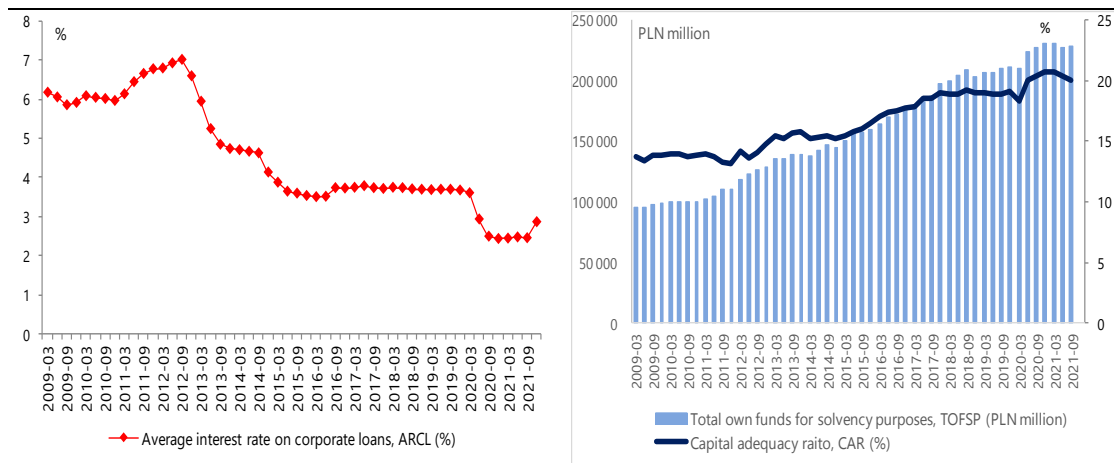
In the years 2009-2021, the credit policy of banks in Poland consisted of:

- on the one hand, the liberalization of interest rates on loans for private and institutional clients, including NFCs – mainly since 2012 – by lowering interest rates on loans,
- on the other hand, increasing the capital requirements of banks, including increasing the capital adequacy ratio and reserves for securing credit risk – which was also served by national recommendations (Polish Financial Supervision Authority, PFSA) and numerous EU directives.

In the period Q3.2012-Q3.2021, the average interest rate on corporate loans fell from 7.0% to 2.4%. Only in Q4.2021 these rates were raised to 2.9% due to the increase in inflation (consumer price index, CPI) (Figure1, left panel). In terms of capital requirements and securing capital adequacy, banks systematically raised them. Total own funds for solvency purposes (TOFSP) grew from PLN 95,692 million to PLN 228,037 million. These capitals made it possible to maintain the capital adequacy systematically growing, from 13.7% in Q1.2009 to 20.0% in Q3.2021 on average in banks in Poland (Figure1, right panel).

Average interest rate on corporate loans (left panel), the total own funds for solvency purposes and the capital adequacy ratio (right panel) in Poland in the period of Q1.2009-Q4.2021 (% , PLN million)

Figure 1



Sources: Author's compilation based on NBP 2022).

In the case of Poland, it is worth paying attention to the numerous recommendations of the PFSA regarding the lending policy of banks, including the quality of assets (KNF, 2022):

- Recommendation M – concerning operational risk management in banks.
- Recommendation P – concerning the management of banks' liquidity risk.

- Recommendation R – concerning the principles of identifying balance sheet credit exposures that are impaired, determining: impairment losses on balance sheet credit exposures and provisions for off-balance sheet credit exposures.
- Recommendation R – concerning the principles of credit exposure classification, estimation and recognition of expected credit losses and credit risk management / Recommendation R comes into force on 1 January 2022.
- Recommendation S – concerning good practices in the management of mortgage-secured credit exposures.
- Recommendation S – concerning good practices in the management of mortgage-secured credit exposures.
- Recommendation T – concerning best practices in managing the risk of retail credit exposures.
- Recommendation U – concerning good practices in the field of bancassurance.
- Recommendation W – concerning model risk management in banks.
- Recommendation W – on model risk management in banks.
- Recommendation Z – concerning the principles of internal governance in banks.

The above-mentioned recommendations and supervision of banks in Poland in terms of their implementation by the PFSA significantly improve the quality of the lending policy, and thus the level of NPLs.

Systematic lowering of interest rates on loans to corporations as well as an increase in capital requirements and securing capital adequacy on the part of banks – to legal regulations (directives, numerous recommendations) – guaranteed an appropriate credit policy and credit risk protection. Due to these actions, there were no bank failures in the commercial banking sector in Poland

3. Non-performing loans (NPLs) – regulations and recommendations

A bank loan is considered non-performing when more than 90 days pass without the borrower paying the agreed installments or interest. According European Central Bank (ECB) – the NPLs are also called “bad debt” (ECB, 2020a).

The ECB requires asset and definition comparability to evaluate risk exposures across euro area central banks. The ECB specifies multiple criteria that can cause an NPL classification when it performs stress tests on participating banks.

The ECB has performed a comprehensive assessment and developed criteria to define loans as nonperforming if they are:

- 90 days past due, even if they are not defaulted or impaired
- Impaired with respect to the accounting specifics for U.S. GAAP and International Financial Reporting Standards (IFRS) banks
- In default according to the Capital Requirements Regulation (CRR).

The increase in the NPL ratio proves the deterioration of the lending policy. If a bank has too many bad loans on its balance sheet, its profitability will suffer because it will no longer earn enough money from its credit business. However, limiting lending on the part of banks means limiting the sources of financing investments among enterprises and, further, increases unemployment in the economy.

Asset quality monitoring is a key area of supervision in banks, along side liquidity and profitability. The asset quality analysis mainly involves calculation:

- NPLs to total loans,
- NPLs less provisions to capital,
- Sectoral distribution of loans to total loans.

NPLs may affect financial stability as they weigh on the viability and profitability of the affected institutions and have an impact, via reduced bank lending, on economic growth. More specifically, high stocks of NPLs can weigh on bank performance through two main channels:

- NPLs generate less income for a bank than performing loans and thus reduce its profitability, and may cause losses that reduce the bank's capital. In the most severe cases, these effects can put in question the viability of a bank, with potential implications for financial stability.
- NPLs tie up significant amounts of a bank's resources, both human and financial. This reduces the bank's capacity to lend, including too small and medium-sized enterprises, which rely on bank lending to a much greater extent than larger companies. In turn, this negative effect in terms of credit supply also reduces the capacity of businesses to invest, affecting economic growth and job creation, hence creating a tangible effect on the real economy (European Commission Services, ECS, 2020).

Due to the multithreaded scope of portfolio quality in banking activity, the scope of legal regulations, including monitoring, is also extensive. Among the regulations of the European Commission (EC) in the field of financial supervision and management, it is worth mentioning (EC, 2022):

- Financial conglomerates - Directive (2002/87/EC)
- Banking prudential requirements - Directive 2013/36/EU
- Banking prudential requirements - Regulation (EU) No 575/2013
- Bank recovery and resolution - Directive 2014/59/EU
- Deposit guarantee schemes - Directive 2014/49/EU
- Credit rating agencies - Regulation (EC) No 1060/2009
- Prudential supervision of investment firms - Directive (EU) 2019/2034
- Prudential supervision of investment firms - Regulation (EU) 2019/2033.

Supervision of the quality of loan portfolio (including NPLs), is one of the key areas of risk reduction in the European banking sector. European Council notes that the financial crisis and ensuing recessions, together with structural factors, accompanied by inadequate loan origination practices, have left the banks in some Member States with high ratios of NPLs.

The Commission and other EU authorities have long highlighted the urgency of taking the necessary measures to address the risks related to NPLs (ECS, (2019). In order to reduce the high NPL stocks, the EU agreed on a comprehensive set of measures outlined in the "Action Plan to Tackle NPLs in Europe" (European Council, 2017), which is currently being implemented. The ongoing decline of NPLs has been and continues to be one of the key areas for reducing risk in the European banking sector. Still, high NPL ratios remain an important challenge, for some (EC, 2019, June 12; EC, 2019, July 11).

Monitoring the quality of the corporate loan portfolio in the banking sector results from prudential regulations. As part of its package of proposals on NPLs put forward in March 2018, the Commission proposed a Regulation amending the CRR (Regulation EU 575/2013, European Parliament, 2013), introducing a 'statutory prudential backstop' in order to prevent the risk of under-provisioning of future NPLs (Regulation EU, 2019/630, European Parliament, 2019). The regulation was adopted in April 2019 and it requires banks to have sufficient loan loss coverage (i.e. common minimum coverage levels) for newly originated loans if these become non-performing exposures (NPEs). In case a bank does not meet the applicable minimum coverage level, it has to deduct the shortfall from its own funds.

In a narrow sense, the monitoring of NPLs concerns the diagnosis of the quality of the banking portfolio in terms of many financial indicators (Annex Table 1, IMF, 2003, May 14, p. 12).

4. Review of research in the field of NPLs

Many researchers analyze changes in NPLs taking into account the impact of many macroeconomic and banking variables. In the group of macroeconomic factors are commonly studied: real GDP growth, value of GDP/GDP per capita, the exchange rate, interest rates and the level of inflation. The results confirm that: real GDP growth usually translates into a higher level of income, improving the financial standing of borrowers and decreasing NPLs. When an economy is below normal conditions or in a recession, NPL levels may rise due to the ensuing rise in unemployment, and borrowers face severe debt repayment difficulties (Salas and Suarina, 2002; Ranjan and Dhal, 2003; Fofack, 2005; Jiménez and Saurina, 2005; Thalassinou et al., 2015). Exchange rate fluctuations may have a negative impact on the quality of assets, especially in countries with a large amount of foreign currency loans. The same applies to interest rate increases, particularly in the case of loans with flexible interest rate (Louzis et al., 2012; Zaman and Meunier, 2017). However, on the one hand, higher inflation may ease debt compensation by affecting the real value of unpaid credit, while on the other hand it may also reduce the real income of unprotected borrowers. In countries where credit rates are flexible, higher inflation may lead to higher rates resulting from monetary policy actions to fight inflation (Nkusu, 2011).

Klein (2013) for NPLs in Central, Eastern and South-Eastern European countries (CESEE) in 1998-2011 confirmed that NPLs responded to macroeconomic conditions, i.e., unemployment, GDP growth and inflation, and that high NPLs in these countries have a negative effect on economic recovery. According to Mazreku et. al. (2018) for 10 transition countries (Central and Eastern Europe, CEE) in 2006 and 2016, dynamic panel estimates show that GDP growth and inflation are both

negatively and significantly correlated with the level of NPLs, while unemployment is positively related to NPLs. Export growth shows largely insignificant results, indicating that NPLs in the sample are mainly influenced by domestic conditions rather than external economic shocks. Vogiazas and Nikolaidou (2011) investigate the determinants of nonperforming creditors in the Romanian banking sector during the Greek crises (2001-2010) and find that inflation and external GDP information influence the credit risks of the banking system in the country. According to Hada et.al.(2020), the exchange rates (mainly EUR, USD and CHF), unemployment rate and inflation rate had a significant impact on NPLs in the Romanian banking system in the period 2009-2019.

Among the banking variables that define NPLs, research focuses on return on assets (ROA), bank efficiency, and bank capital. However, the specificity of each bank and its customers are very important for NPL changes. For example, Godlewski (2008) investigates the association between NPLs and return on assets (ROA) and states that the lower the rate of ROA, the higher the NPLs and vice versa. Boudriga et al. (2010) confirm from their study that there is a negative association between ROA and NPLs. They conclude that when the ROA decreases, then a bank starts to make investments in high-risk projects and as a result the level of NPLs rises. Dimitrios et al. (2016) investigate the various determinants of NPLs in the euro banking system and conclude that ROA has a significant impact upon NPLs. An insufficient control of the loan portfolio (including short-term loans) increases risk and NPLs. Fiordelisi et al. (2011) examine the various factors that increase the risk level in the EU banks and conclude that a declining efficiency hikes the risk level of banks in future. Furthermore, efficiency and performance factors have influence on NPLs in the Greek banking sector (Louzis et al., 2012). Rachman et al. (2018) state that operating efficiency does not influence NPLs.

The effect of bank capital on NPLs works in the opposite direction. For one part, incentivised managers of low capitalized banks tend to get involved in high-risk investments and give loans that are issued without proper credit rating and monitoring (Keeton, 1999). For another part, banks with a high level of capital tend to give loans easily as they know that owing to these loans banks are not going to be bankrupt and fail; therefore, banks are highly engaged with these kinds of risky credit activities suggesting a positive association between capital and NPLs (Rajan,1994). Moreover, the capital adequacy ratio (CAR) shows the ability of an organization to face abnormal losses and to survive that situation. Makri et al. (2014) also state that there is a negative association between CAR and NPLs. Constant and Djiogap et. al. (2012) claim that NPLs and CAR have a positive association with each other. Bank profitability and sustainability can only be provided through a proper flow of interest income generated through the lending function. However, since banks are no longer able to generate enough interest income through classical safe credit and are required to maintain reserves in the form of provisions to cover for eventual loan losses, bank capital decreases together with their health, which is becoming fragile, increasing the trend of NPLs. Therefore, banks are required to take proactive action to deal with the phenomenon of a poor choice of borrowers by identifying and understanding the macroeconomic factors that contribute to the rise of classified credit in the banking system (Anjom and Karim, 2015).

According Paudino et. al. (BIS, 2017) the resolution of NPLs that have reached systemic levels is complex and costly. Bank NPL problems tend to emerge after credit booms or protracted periods of low growth in structurally weak financial

systems. NPLs crowd out new lending, eroding both the profitability and solvency of banks. When high NPL levels affect a sufficiently large number of banks, the financial system stops functioning normally, and banks can no longer provide credit to the economy. A prompt recovery can be obstructed by impaired market functioning and coordination failures among banks. In such circumstances, authorities usually step in to lead the crisis response. To this end, they can deploy a variety of resolution instruments, although these typically require a large amount of resources and take time to deliver results.

Moreover, results of Baudino and Yun show that the resolution toolkit used by the authorities has remained broadly unchanged for several decades in Europe (see also Iwanicz-Drozowska, 2015), and the United States. Success of resolution policies varies from case to case. Important role play structural banking sector conditions, the type of problem assets, the fiscal space for public sector intervention, and legal and judicial frameworks for NPL resolution. These country-specific characteristics determine how far specific resolution options may be applicable and effective in one country but not in another (see also ECB, November 2020b).

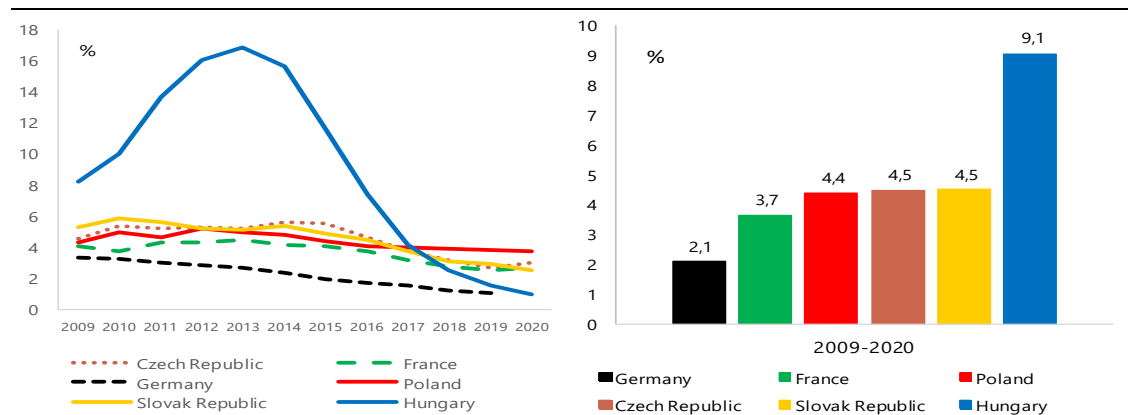
5. Differences in the NPLs between EU countries and the structure of the loan portfolio of NFCs in Poland

The bank non-performing loans to total gross loans according to the World Bank (2021) show significant differences in the banking sectors of EU countries (e.g. 27.0% Greece, 15.0% Cyprus, 5.8 % Bulgaria, 4.9% Portugal, 3.7% Poland, 2.93% Czech Republic, 2.71% France, 2.53% Slovak Republic, 1.1% Germany and 0.93% in Hungary in 2020).

The differences in the average NPL ratio (e.g. for the years 2009-2020) reached several percentage points between the banking sectors. The lowest level of NPL in the presented period was maintained by Germany (2.1%), France (3.7%), Poland, the Czech Republic and Slovakia (4.4% -4.5%) compared to the highest level in Hungary (9.1%) (Figure 2).

Bank non-performing loans to total gross loans in selected countries in 2009-2020 (%)

Figure 2

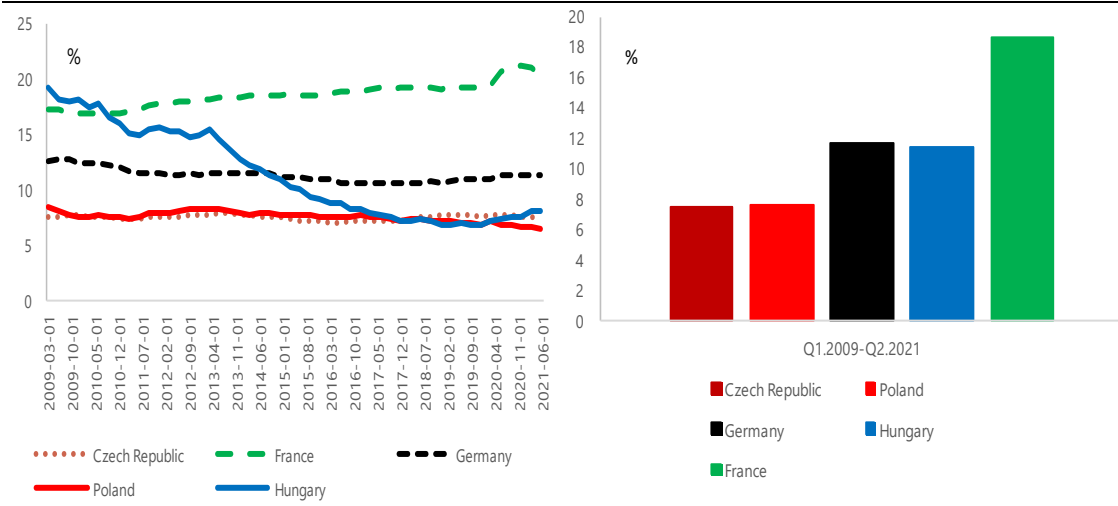


Sources: Author's compilation based on WDI 2022).

The varying quality of bank portfolios in the EU countries is also accompanied by significant differences in debt servicing costs. Lower and relatively stable debt servicing costs are usually accompanied by better portfolio quality and lower NPL values, such as in the Czech Republic, Poland or Germany (Figure 3).

Debt service ratio for the private NFCs in selected countries in Q1.2009.Q1-Q2.2021 (%)

Figure 3



Sources: Author's compilation based on BIS (2022).

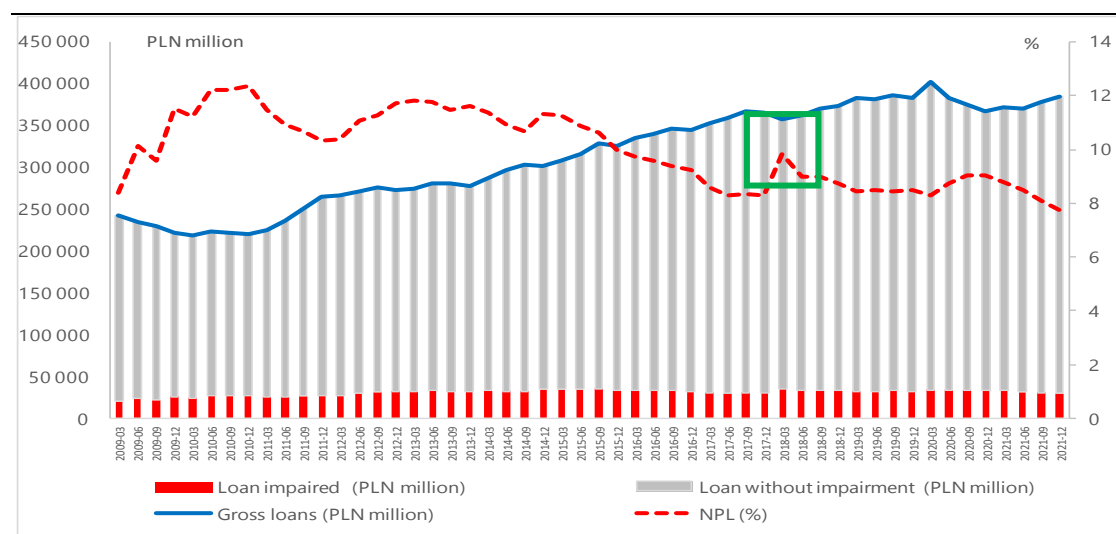
In the case of Poland, a more detailed analysis of the changes in NPLs in 2009.Q1-2021.Q4 in relation to the structure of the NFCs' loan portfolio indicates that with the gross loan increase, the NPL decreased annually.

The total value of corporate loans in the banking sector in Poland showed a general upward trend in the years Q1.2009-Q1.2020 (from PLN 242.9 million to PLN 401.6 million). Only the period 2020.Q2-Q4.2020.Q4 brought a decrease in the total value of loans (PLN 383.5 million and PLN 366.9 million) and re-growth. The NPL ratio showed quarterly fluctuations, however generally it showed a downward trend in the period Q4.2010.-Q1.2020 (from 12.3% to 9.4%), and next quarter decreased in the Q4.2021 (7.7%)² (Figure 4).

² The decline in total loans was also due to a decline in demand for loans from borrowers due to the uncertain macroeconomic situation related to the COVID-19 pandemic.

Changes of loans impaired and without impairment of non-financial corporations in Poland in the period of Q1.2009-Q4.2021 (% , PLN million)

Figure 4



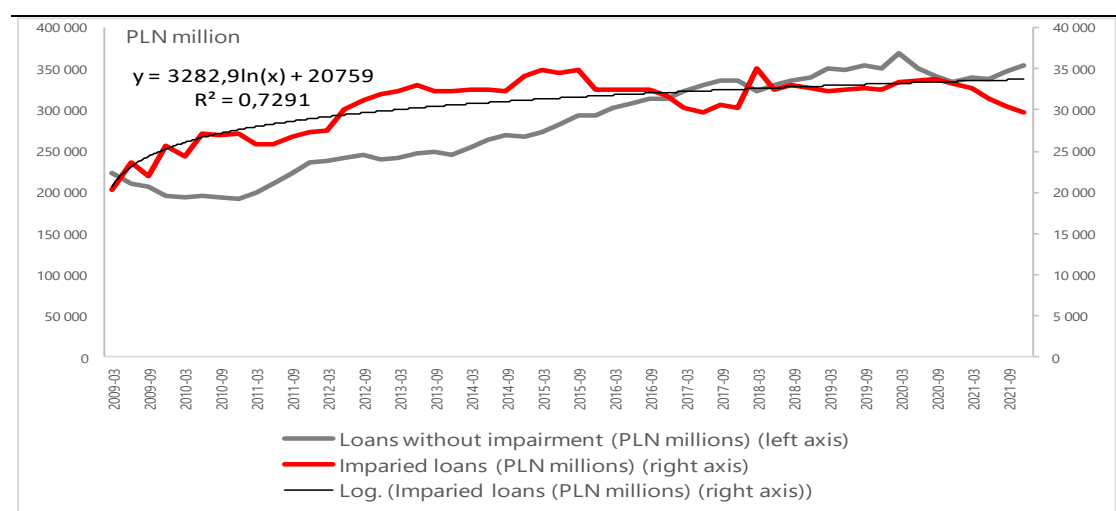
Note: The increase in NPL in the period Q4.2017-Q2.2018 – is the result of changes introduced in the classification of impaired receivables. The anomalies shown in the box in the chart are the result of changes (in the qualification of receivables to phase 3 / impaired) in the mandatory reporting of banks to the NBP (FINREP) and bank adjustments related to the obligation to include in the gross carrying amount also interest on receivables included in the phase 3. After about six months, a significant part of this interest was written off the balance sheet and charged to provisions.

Source: The author's compilation based on NBP (2022).

Despite the fact that the value of loans generally increased, the dynamics of growth of loans without impairment was weaker in the period 2009-2021. The indicated increased in the NPL ratio in the period Q2-Q4.2020 (8.7%-9.0%) was results from quickly decreased in loans without impairment (from PLN 350.0 million to PLN 333.8 million) than impaired loans (PLN 33.5 million to PLN 33.2 million) (Figure 5).

Changes of the impaired loans and loans without impairment of NFCs in Poland in the period of Q1.2009-Q4.2021 (% , PLN millions)

Figure 5



Sources: Author's compilation based on NBP (2022).

The indicated changes in the loan portfolio (Q2.–Q4.2020) and increases in NPL rates were mainly caused by the reduction of economic activity and, consequently, lower income. While in Q4.2019 the value of gross revenues from the total activity of corporations in Poland amounted to PLN 3 235 515.6 million, it decreased to PLN 786 700.6 million in the Q1.2020. The following quarters saw a slow increase in this revenues (Central Statistical Office, CSO, 2022). With the outbreak of the Covid-19 pandemic, additional regulatory requirements were imposed on banks to maintain security in the banking sector (BIS, 2020).

6. Research procedure - mainly determinant of NPLs of NFCs in Poland

The importance of diagnosing changes in NPLs of NFCs in Poland results, apart from the legal obligations of banks, also from the role of this segment of loans. The share of corporate loans in the structure of the gross loan portfolio was 57%. This means that any changes in this portfolio had a significant impact on the entire loan portfolio (NBP, 2022).

The NPL rate is calculated as the ratio of the non-performing loans (impaired loans) and advances of non-financial corporation to the gross value of total loans and advances of these corporations (NBP, 2020).

$$NPL\ ratio = \frac{Non-performing\ loans}{Total\ loans} \quad (1)$$

In this study author made an attempt to assess the quality of the portfolio of loans granted to non-financial corporations, therefore, respectively, impaired loans and total loans granted to these corporations (included in the so-called phase III, portfolio B) were taken into account.

In modelling the quality of the portfolio of NPLs granted to NFCs, mainly the following variables are taken into account: market and financial variables of corporations – determine the possibility of servicing loans, and variables of banking conditions – serving as capital hedging against an increase in banking risk.

In order to analyse the relationship between changes in NPL ratio and chosen variables a final formula for the NPL function was developed:

$$NPL_t = \alpha_0 + \alpha_1 GDP_t + \alpha_2 CPI_t + \alpha_3 WIBOR_t + \alpha_4 ROAC_t + \alpha_5 CROAC_t + \alpha_6 GFCF_t + \alpha_7 GTPR_t + \alpha_8 AIRCL_t + \alpha_9 CAR_t + \alpha_{10} TOFSP_t + \alpha_{11} CRofCR_t + \xi_i \quad (2)$$

The explained variable: NPL_t – The non-performed loan ratio

Explanatory variables: (Table 1) and

ξ_i – random component

t – period

Description of model and data source

Table 1

| Variable | Description | Source | Expected impact on the NPLs |
|---|---|--------|---------------------------------------|
| Macroeconomic variables | | | |
| GDP_t | Gross domestic product | OECD | " – " |
| CPI_t | Consumer price index | CSO | " – " |
| $WIBOR_t$ | Warsaw Interbank Offered Rate | OECD | " + " |
| Variables of the financial standing of corporations | | | |
| $ROAC_t$ | Revenues from the overall activity of corporations | CSO | " – " |
| $CROAC_t$ | Costs of obtaining revenues from the overall activity of corporations | CSO | " + " |
| $GFCF_t$ | Gross fixed capital formation | CSO | " – " |
| $GTPR_t$ | Gross turnover profitability ratio | NBP | " |
| Banking variables | | | |
| $AIRCL_t$ | Average interest rate on corporate loans | NBP | " + " |
| CAR_t | Capital adequacy ratio | NBP | in line with changes in the NPL ratio |
| $TOFSP_t$ | Total own funds for solvency purposes | NBP | |
| $CRofCR_t$ | Capital requirements of credit risk | NBP | |

Sources: The author's compilation based on: NBP (2022), CSO (2022) and OECD Internet databases (2022).

The methodology of changes in the quality of the loan portfolio corresponds to the methodologies used by central banks, e.g. by NBP and IMF (2003), Matthews, Guo & Zhang (2007), Maggi & Guida (2010). The study period includes 52 quarters data for the period Q1.2009–Q4.2021, used the first differences.

In this study, methods are used known from literature on international economics and international finance and econometric methods like the VECM model (Vector Error Correction Method) including the impulse response functions and forecast error variance decomposition analysis.

The data verification procedure and the selection of the analysis method included: ADF test, KPSS stationary test, VAR inverse root, the Engle-Granger and Johansen test and lag order (AIC, BIC, HQC criteria). In order to verify correctness of the VECM model results: two tests were carried out verifying the Autocorrelation Ljung-Box Q' test, and ARCH test. Co-integration was verified by means of the Engle-Granger and Johansen tests which confirmed the occurrence of co-integration and thus justified the use of the VECM model for the lag order 2 and co-integration of order 1.

In accordance with the Granger representation theorem, if variables y_t and x_t are integrated to the order of 1 (1) and are co-integrated, the relationship between them can be represented as a vector error correction model (VECM) (Piłatowska 2003).

The general form of the VECM can be written as:

$$\begin{aligned}\Delta Y_t &= \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} + \dots + \Gamma_{k-1} \Delta Y_{t-k+1} + \pi Y_{t-k} + \varepsilon_t = \\ &= \sum_{i=1}^{k-1} \Gamma_i \Delta Y_{t-i} + \pi Y_{t-k} + \varepsilon_t,\end{aligned}\quad (3)$$

where:

$$\Gamma_i = \sum_{j=1}^i A_j - I, \quad i = 1, 2, \dots, k-1, \quad \Gamma_k = \pi = -\pi(1) = -\left(I - \sum_{i=1}^k A_i\right)$$

and I is a unit matrix.

7. Empirical results - impulse response functions and variance decompositions

Analysis of the NPL response to impulses from the explanatory variables confirmed that the strength of the influence of these impulses increased over time. The impact of explanatory variables increased especially from the 4th-8th quarter, showing changes (positive / negative) in the following quarters.

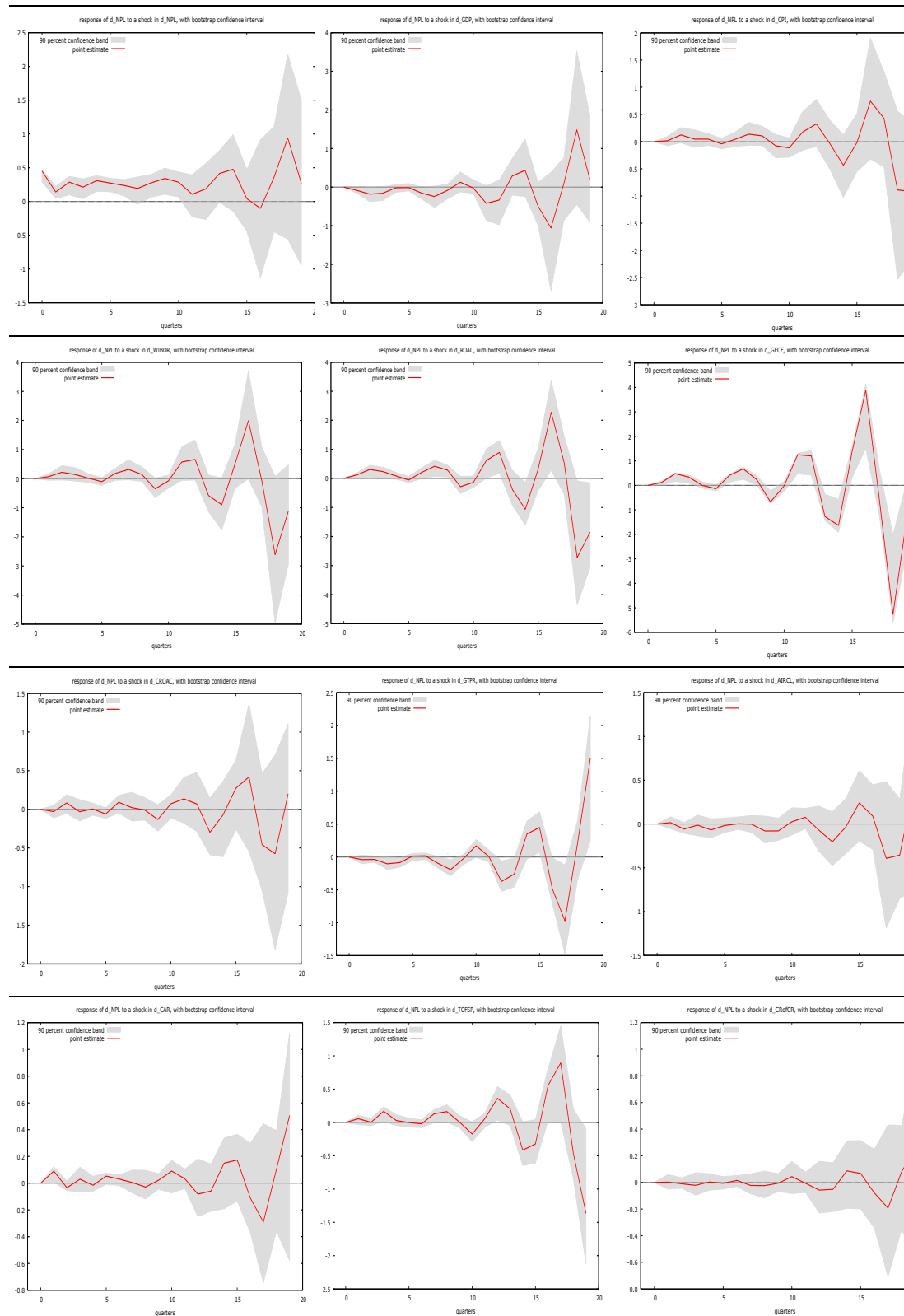
The NPL showed the following reactions (responses) in the end 20th quarter:

- The NPLs showed declining trends in response to impulses from: NPL's own changes, GDP, CPI, WIBOR, ROAC, GFCF, TOFSP.
- The NPLs showed increasing trends in response to changes: CROAC, GTPR, AIRCL, CAR and CRofCR (Figure 6).

The analysis of the decomposition of explanatory variables shows, in turn, that the NPL rate was significant in explaining the changes in the first period: by own changes (100.0%), in the 2nd period own NPL (65.5%), GDP (27.9%), CAR (2.8%), CROAC (1.4%) and TOFSP (1.3%) In the 8th period decreased rate of explanation own NPL (1.5%), increased GDP (94.33) and TOFSP (2.8%). Finally, in the 20th period this rate of explanation was stronger on the side of GDP (95.0%), own NPL (1.5%), GFCF (0.7%), CPI (0.1%) and also WIBOR (0.1%). This means that the main pillar of the explanation of NPL changes were changes in GDP (i.e. changes in the business cycle) and investment expenditure (GFCF) as well as CPI and WIBOR (Figure 7).

The impulse response functions (summary statement), forecast horizon 20q, include bootstrap confidence interval 1- α =0.90 (shaded area)

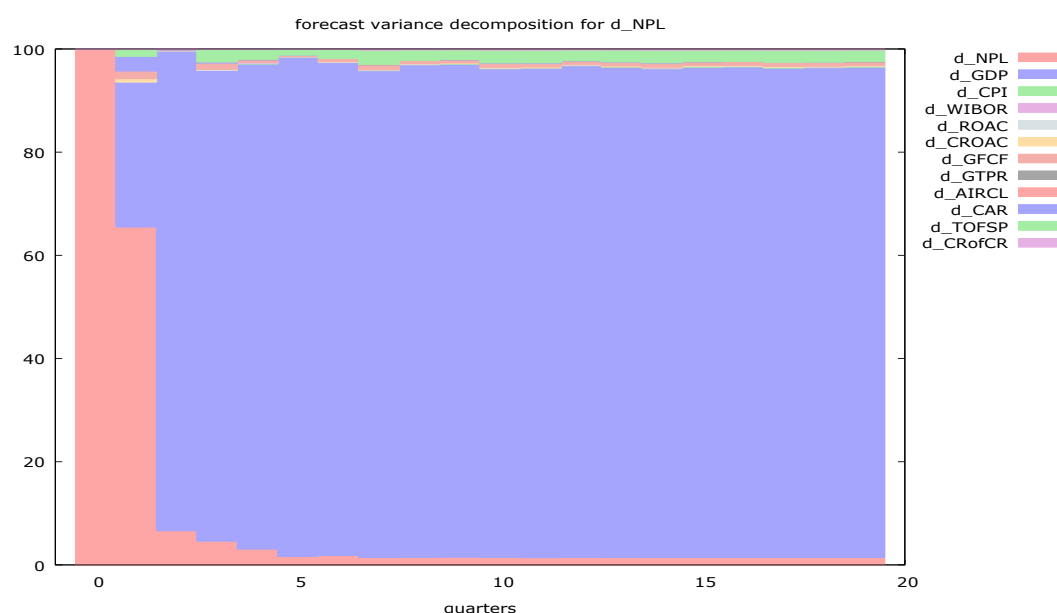
Figure 6



Sources: Author's compilation calculations.

Variance decomposition for the NPL variable

Figure 7



Sources: Author's calculations.

8. Conclusions and way forward

In the analyzed period, banks pursued a liberal policy of interest rates on loans while maintaining adequate capital requirements. The analysis of NPL changes shows that there was a long-term downward trend confirming the improvement in the quality of the portfolio of loans of non-financial companies in the period Q1.2009–Q4.2021. However, the last analysis quarters (during the COVID pandemic), brought an increase in the NPL ratio, respectively: Q2.2020 (8.7%) and Q4.2020 (9.0%). Whereas, in the entire period Q1.2009–Q4.2021, the structure of the loan portfolio in the Polish banking sector showed stable levels.

The results of the response function indicate negative NPL responses (in the end 20th quarter) to impulses (earlier) own NPL fluctuations, GDP, CPI, WIBOR, ROAC, GFCF and TOFSP. The NPLs showed increasing trends in response to changes: CROAC, GTPR, AIRCL, CAR and CRoFCR. The results of variance decomposition indicate that the main pillar of the explanation of NPL changes were: GDP (i.e. changes in the business cycle) and investment expenditure (GFCF) as well as CPI and WIBOR. Also, the results of the NPLs research confirmed the pro-cyclical nature of lending activity in Poland in the verified years.

Resuming, in the period 2009–2021 there was a long-term trend of improving the quality of the loan portfolio of NFCs, which was mainly explained by market (macroeconomic factors). Taking into account the implementation of prudential standards by banks in Poland, resulting from EU directives and numerous recommendations of the Polish Financial Supervision Authority, it should be stated that banks conducted a proper credit policy (they took care of the quality of assets). Moreover, the relatively liberal monetary policy of the NBP (in terms of the basic

interest rates) in the last decade and the maintained GDP growth rate also contributed to lowering the NPLs.

The main problems of banking in Poland (which are mostly common problems of EU countries) are:

- Adaptation to new customer expectations
- The need for new financing
- Macroeconomic situation
- Possible rebound in banking sector profits?
- Consequences of the Court of Justice of the European Union (CJEU) judgment on loans in Swiss francs
- Cybersecurity and efficiency of systems.

Annex

Role of Core and Corporate FSIs

Table 1

| | | | |
|-----------------------------------|---|---|---|
| Banking Sector Financial Strength | <i>Capital Adequacy</i> | Tier 1 capital ratio | Assesses adequacy of highest quality capital, such as shareholder equity and retained earnings, relative to risk weighted assets |
| | | Regulatory capital ratio | A broader measure of capital including items giving less protection against losses, such as subordinated debt, tax credits and unrealized capital gains |
| | <i>Earnings and profitability</i> | Return on equity | Assesses scope for earnings to offset losses relative to capital or loan and asset portfolio |
| | | Return on assets | |
| | | Interest margin to Gross income | Indicates the importance of net interest income to earnings and scope to absorb losses |
| | | Non-interest expenses to income | Indicates extent to which high non-interest expenses weakens earnings |
| | <i>Asset quality</i> | NPLs to total loans | Indicates the credit quality of banks' loans |
| | | NPLs less provisions to capital | Shows NPLs net of provisions taken against them relative to capital |
| | | Sectoral distribution of loans to total loans | Identifies credit exposures concentrations to particular sectors by the whole banking sector |
| Banking Sector vulnerabilities | <i>Liquidity</i> | Liquid assets ratio | Assesses the vulnerability of the banking sector to loss of access to market sources of funding or a run on deposits |
| | | Liquid assets to shortterm liabilities | |
| | <i>Sensitivity to market risk</i> | Duration of assets and liabilities | Measures maturity mismatch to assess interest rate risk |
| | | Net open foreign exchange position to capital | Measures foreign currency mismatch to assess exchange rate risk |
| Corporate sector | Leverage ratio | Leverage ratio | Gives an indication of the credit risk as a highly leveraged corporate sector is more vulnerable to shocks that could impair its capacity to repay loan |
| | Return on Equity | Return on Equity | Indicates the extent to which earnings are available to cover losses |
| | Earnings to interest and principle payments | Earnings to interest and principle payments | Reveals to what extent earnings available to cover losses are reduced by interest and principle |
| | Net FX exposure to equity | Net FX exposure to equity | The vulnerability of the corporate sector to exchange rate changes |
| | Number of bankruptcies | Number of bankruptcies | Serves as an indicator of corporate sector distress |

Sources: IMF (2003, May 14, p. 12). .

References

- Anjom, W., Karim, A.M. (2016), "Relationship Between Non-Performing Loans and Macroeconomic Factors (With Specific Factors: A Case Study on Loan Portfolios- SAARC Countries Prospective", *Asia Pacific Journals of Finance and Risk Management*, 15(3), pp. 84-103.
- Basel Committee on Banking Supervision (2014, October), "Basel III: the net stable funding ratio".
- Baudino P., Yun H. (2017), "FSI Insights on policy implementation No 3 Resolution of non- performing loans – policy options", *Financial Stability Institute, BIS*, October, pp. 1-39.
- BIS (2020), "Governors and Heads of Supervision announce deferral of Basel III implementation to increase operational capacity of banks and supervisors to respond to Covid-19", (Press Release). Retrieved 19 February 2021.
- BIS (2022), "Debt service ratios for the private non-financial sector", https://www.bis.org/statistics/dsr.htm?m=6_380_671
- Boudriga, A., Taktak, N.B., Jellouli, S. (2010), "Bank Specific, Business and Institutional Environment Determinants of Banks Nonperforming Loans: Evidence from MENA Countries", *Economic Research Forum, Working Paper*, pp. 1-28.
- CSO (2022), "Macroeconomics indicators", <https://stat.gov.pl/wskazniki-makroekonomiczne/>.
- Dimitrios, A., Helen, L., Mike, T., (2016), "Determinants of Non-Performing Loans: Evidence from Euro-Area Countries", *Finance Research Letters, Elsevier*, Vol. 18, pp. 116-119, doi: 10.1016/j.frl.2016.04.008.
- Djiogap, F.C., Ngomsi A., (2012), "Determinants of bank long-term lending behavior in the central African economic and monetary community (CEMAC)", *Review of Economics & Finance*, 2: 107-114.
- ECB (2020a), "What are non-performing loans (NPLs)?", <https://www.ecb.europa.eu/explainers/tell-me/html/npl.en.html>.
- ECB (2020b), "Consolidated Banking Data. Calculations by Commission services (DG FISMA)", https://www.ecb.europa.eu/stats/supervisory_prudential_statistics/consolidated_banking_data/html/index.en.html.
- EC (2019 July 11), "Council conclusions on Action plan to tackle non-performing loans in Europe", <https://www.consilium.europa.eu/en/press/press-releases/2017/07/11/conclusions-non-performing-loans/>.
- EC (2019, June 12), "Fourth Progress Report on the reduction of non-performing loans and further risk reduction in the Banking Union, Brussels", COM(2019)278 final.
- EC (2022), "EU banking and financial services law", https://ec.europa.eu/info/law/law-topic/eu-banking-and-financial-services-law_en
- ECS (2019 November 15), "Monitoring Report on Risk Reduction Indicators", European Working Group meeting, <https://www.consilium.europa.eu/media/>

37029/joint-risk-reduction-monitoring-report-to-eg_november-2018.pdf, pp. 20-45.

- ECS (2019), "Monitoring Report on Risk Reduction Indicators", European Working Group meeting., https://www.consilium.europa.eu/media/37029/joint-risk-reduction-monitoring-report-to-eg_november-2018.pdf, pp. 20-45.
- ECS (2020), "Report of the FSC Subgroup on Non-Performing Loans. European Working Group meeting", November 20, <http://data.consilium.europa.eu/doc/document/ST-9854-2017-INIT/en/pdf>.
- European Council (2017), "Council conclusions on Action plan to tackle non-performing loans in Europe", 11 July, <https://www.consilium.europa.eu/en/press/press-releases/2017/07/11/conclusions-non-performing-loans>.
- Fiordelisi, F., Marques-Ibanez, D., Molyneux, P., (2011), "Efficiency and risk in European banking, *Journal of Banking and Finance*", Elsevier, Vol. 35, No. 5, pp. 1315-1326.
- Fofack, H. (2005), "Non-performing Loans in Sub-Saharan Africa: Causal Analysis and Macroeconomic Implications", World Bank Policy Research Working Paper No. 3769.
- Godlewski, C.J. (2008), "Bank capital and credit risk taking in emerging market economies", *Journal of Banking Regulation*, Vol. 6 No. 2, pp. 128-145, doi: 10.1057/palgrave.jbr.2340187.
- Hada T., Bărbuță-Mișu N., Iuga I.C., Wainberg D. (2020), "Macroeconomic Determinants of Nonperforming Loans of Romanian Banks", September, *Sustainability* 12(18):7533, pp. 1-19, DOI: 10.3390/su12187533.
- IMF (2003 May 14), "Financial Soundness Indicators – Background Paper", Prepared by the Staff of the Monetary and Financial Systems and Statistics Departments. Approved by C.S. Carson and S. Ingves, <https://www.imf.org/external/np/sta/fsi/eng/2003/051403bp.pdf>, p. 12.
- Iwanicz-Drozdowska M., (2015), "Restrukturyzacja banków w Unii Europejskiej w czasie globalnego kryzysu", (Restructuring of banks in the European Union during the global crisis), Oficyna Wydawnicza SGH, Warsaw.
- Iwanicz-Drozdowska M., (2017), „Zarządzanie ryzykiem bankowym” (Bank risk management), Poltext, Warsaw.
- Jiménez, G., Saurina, J. (2005), "Credit Cycles, Credit Risk, And Prudential Regulation", Working Paper No. 0531, Banco de España. *International Journal of Central Banking* 2(2), pp. 1-34.
- Keeton, W. R. (1999), "Does faster loan growth lead to higher loan losses", *Economic Review Federal Reserve Bank of Kansas City*, pp. 57-75.
- Komisja Nadzoru Finansowego (KNF, 2022), "Rekomendacje dla banków", https://www.knf.gov.pl/dla_rynku/regulacje_i_praktyka/rekomendacje_i_wytyczne/rekomendacje_dla_bankow (10.01.2022).
- Louzis, D.P., Vouldis, A.T., Metaxas, V.L. (2012), "Macroeconomic and Bank-Specific Determinants of Non-Performing Loans in Greece: A Comparative Study of Mortgage, Business And Consumer Loan Portfolios", *Journal of Banking and Finance*, Elsevier B.V., Vol. 36 No. 4, pp. 1012-1027, doi: 10.1016/j.jbankfin.2011.10.012.

- Makri V., Tsagkanos A., Bellas A. (2014), "Determinants of Non-Performing Loans: The Case of Eurozone, Panoeconomicus", Vol. 61, No. 2, pp. 193-206.
- Mazreku I., Morina F., Misiri V., Spiteri J.V., Grima S. (2018), "Determinants of the Level of Non-Performing Loans in Commercial Banks of Transition Countries", European Research Studies Journal Volume XXI, Issue 3, pp. 3-13.
- NBP (2022), "Monetary and financial statistics" <https://www.nbp.pl/homen.aspx?f=/en/statystyka/monetary-and-financial-statistics.html>.
- Nkusu, M. (2011), "Nonperforming loans and macro financial vulnerabilities in advanced economies". IMF Working Papers, 161.
- OECDStat. (2022), <https://stats.oecd.org/>.
- Piłatowska, M. (2003), "Modelowanie niestacjonarnych procesów ekonomicznych. Studium metodologiczne", (Modelling of Non-Stationary Economic Processes. A Methodological Study), Uniwersytet M. Kopernika, Toruń.
- Rachman R.A., Kadarusman Y.B., Anggriono K., Setiadi R. (2018), "Bank-Specific Factors Affecting Non-Performing Loans in Developing Countries: Case Study of Indonesia", The Journal of Asian Finance, Economics and Business (JAFEB), Vol. 5, No. 2, pp. 35-42.
- Rajan R. (1994), "Why bank credit policies fluctuate", The Quarterly Journal of Economics, Vol. 2, No. 109, 1994, pp. 399-441.
- Ranjan, R., Dhal, S.Ch. (2003), "Non-Performing Loans and Terms of Credit of Public Sector Banks in India: An Empirical Assessment", Reserve Bank of India Occasional Papers Vol. 24, No. 3, pp. 1-41.
- Thalassinos, I.E., Stamatopoulos, D.T. and Thalassinos, E.P., (2015), "The European Sovereign Debt Crisis and the Role of Credit Swaps", chapter book in The WSPC Handbook of Futures Markets (eds) W. T. Ziemba and A.G. Malliaris, in memory of Late Milton Miller (Nobel 1990), World Scientific Handbook in Financial Economic Series Vol. 5, Chapter 20, 605-639, doi: 10.1142/9789814566926_0020.
- Salas V., Saurina J. (2002), "Credit Risk in Two Institutional Regimes: Spanish Commercial and Savings Banks", Journal of Financial Services Research, No. 22, pp. 203-224.
- Vogiazas, S., Nikolaidou E., (2011), "Investigating the Determinants of Nonperforming Loans in the Romanian Banking System: An Empirical Study with Reference to the Greek Crisis", Econ. Res. Int. pp. 1-13.
- WDI (2002), World Development Indicators, <https://data.worldbank.org/indicator/FB.AST.NPER.ZS>
- Zaman C., Meunier B. (2017), "A Decade of EU Membership: Evolution of Competitiveness in Romania". European Research Studies Journal, No. 20(2A), pp. 224-236.



NARODOWY
BANK POLSKI

Irving Fisher Committee on
Central Bank Statistics



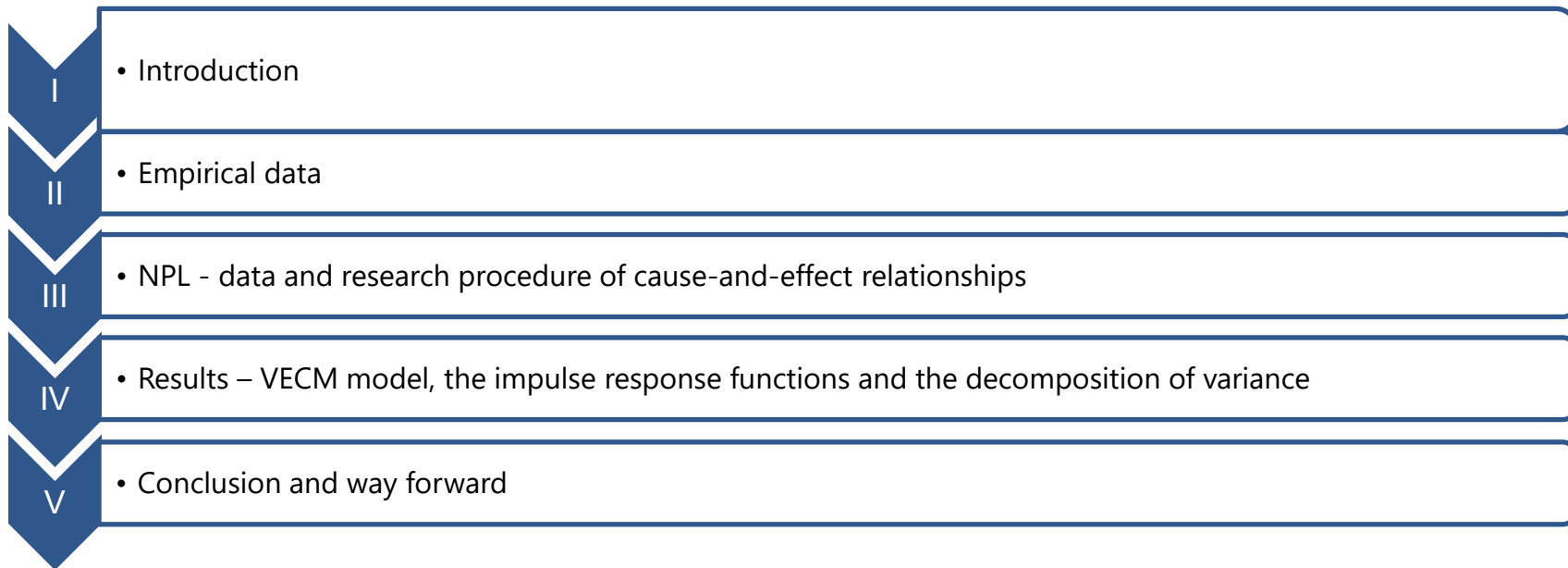
Changes in the lending activity of banks in Poland, including the portfolio of corporate loans

Aneta Kosztowniak, NBP, Economic Analysis and Research Department

- This presentation should not be reported as representing the views of the National Bank of Poland.
- The views expressed are those of the authors and do not necessarily reflect those of the NBP.



Overview



Introduction

A bank loan is considered non-performing when more than 90 days pass without the borrower paying the agreed installments or interest. Non-performing loans (NPLs) are also called "bad debt" (ECB, 2020).

The ECB requires asset and definition comparability to evaluate risk exposures across euro area central banks. The ECB specifies multiple criteria that can cause an NPL classification when it performs stress tests on participating banks.

The ECB has performed a comprehensive assessment and developed criteria to define loans as nonperforming if they are:

- 90 days past due, even if they are not defaulted or impaired
- Impaired with respect to the accounting specifics for [U.S. GAAP](#) and [International Financial Reporting Standards](#) (IFRS) banks
- In default according to the Capital Requirements Regulation.

The NPL rate is calculated as the ratio of the non-performing loans (impaired loans) and advances to the gross value of total loans and advances (NBP, 2020).

$$NPL\ ratio = \frac{Non-performing\ loans}{Total\ loans} \quad (1)$$

Asset quality monitoring is a key area of supervision in banks, along side liquidity and profitability. The asset quality analysis mainly involves calculation:

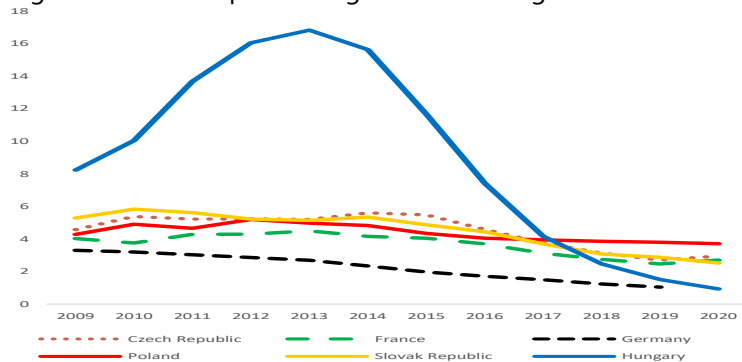
- NPLs to total loans,
 - NPLs less provisions to capital,
 - Sectoral distribution of loans to total loans.
- This monitoring is facilitated by the current assessment of the level and volatility of the indicators listed in Table 1. These statistics are published, among others by The International Monetary (IMF) under the so-called financial soundness (FSIs).

Table 1. Role of Core and Corporate FSIs

| Types of FSI | Specific FSIs | Role of FSIs in Monitoring the Financial Sector |
|--------------------------------------|-----------------------------------|--|
| Banking Sector Financial Strength | Capital Adequacy | Tier 1 capital ratio |
| | | Regulatory capital ratio |
| | Earnings and profitability | Return on equity Return on assets |
| | | Interest margin to gross income |
| | | Non-interest expenses to income |
| Banking Sector vulnerabilities | Asset quality | NPLs to total loans NPLs less provisions to capital |
| | | Sectoral distribution of loans to total loans |
| | Liquidity | Liquid assets ratio |
| | | Liquid assets to shortterm liabilities |
| | Sensitivity to market risk | Duration of assets and liabilities |
| | | Net open foreign exchange position to capital |
| Corporate sector | | Leverage ratio |
| | | Return on Equity |
| | | earnings to interest and principle payments |
| | | Net FX exposure to equity |
| | | Number of bankruptcies |

Empirical data

Figure 2. Bank nonperforming loans to total gross loans in selected countries in 2008-2020 (%)



Source: WDI (2022).

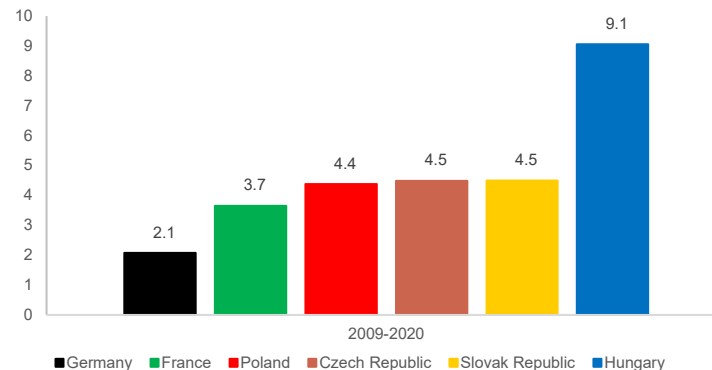
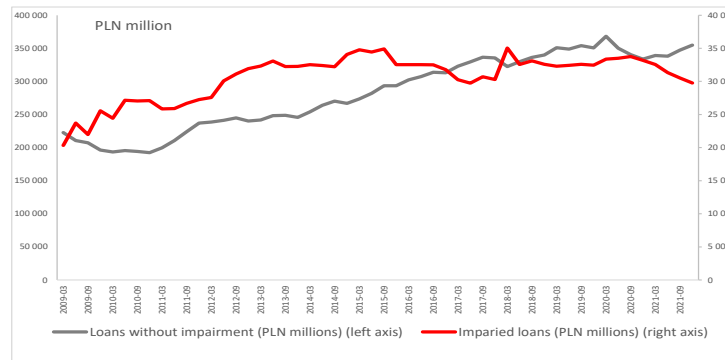
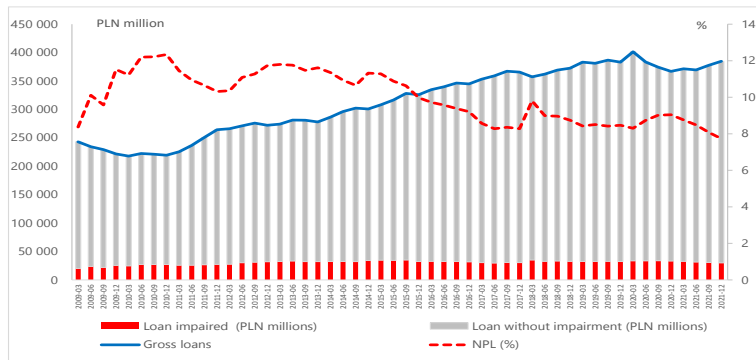


Figure 3. Changes of loans impaired and without impairment of non-financial sector in Poland in the period of Q1.2009-Q4.2021 (% , PLN million)



Source: Author's compilation based on NBP (2022), CSO (2022), OECD (2022) and Eurostat (2022).

Data and research procedure of NPL

Data: the quarters time-series data covering the period 2009:Q1-2021:Q4 (52 quarters; used the the first differences);

Sources: NBP (2022), CSO (2022), and OECD Internet databases (2022).

The data verification procedure and the selection of the analysis method included: ADF test, KPSS stationary test, VAR inverse root, the Engle-Granger and Johanson test and lag order (AIC, BIC, HQC criteria).

In order to verify correctness of the VECM model results: two tests were carried out verifying the occurrence of autocorrelation, i.e.: Autocorrelation Ljung-Box Q' test, and ARCH test.

The final formula for the NPL function:

$$NPL_t = \alpha_0 + \alpha_1 GDP_t + \alpha_2 CPI_t + \alpha_3 WIBOR_t + \alpha_4 ROAC_t + \alpha_5 CROAC_t + \alpha_6 GFCF_t + \alpha_7 GTPR_t + \alpha_8 CAR_t + \alpha_9 TOFSP_t + \alpha_{10} CRofCR_t + \xi_i$$

The explained variable:

NPL_t – The non-performed loan ratio

Explanatory variables:

GDP_t – Gross domestic product

CPI_t – Consumer price index

$WIBOR_t$ – Warsaw Interbank Offered Rate

$ROAC_t$ – Revenues from the overall activity of corporations

$CROAC_t$ – Costs of obtaining revenues from the overall activity of corporations

$GFCF_t$ – Gross fixed capital formation

CAR_t – Capital adequacy ratio

$GTPR_t$ – Gross turnover profitability ratio

$TOFSP_t$ – Total own funds for solvency purposes

$CRofCR_t$ – Capital requirements of credit risk

ξ_i – random component

t – period

Narodowy Bank Polski

The general form of the VECM model:

$$\begin{aligned}\Delta Y_t &= \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} + \dots + \Gamma_{k-1} \Delta Y_{t-k+1} + \pi Y_{t-k} + \varepsilon_t = \\ &= \sum_{i=1}^{k-1} \Gamma_i \Delta Y_{t-i} + \pi Y_{t-k} + \varepsilon_t, \quad (2)\end{aligned}$$

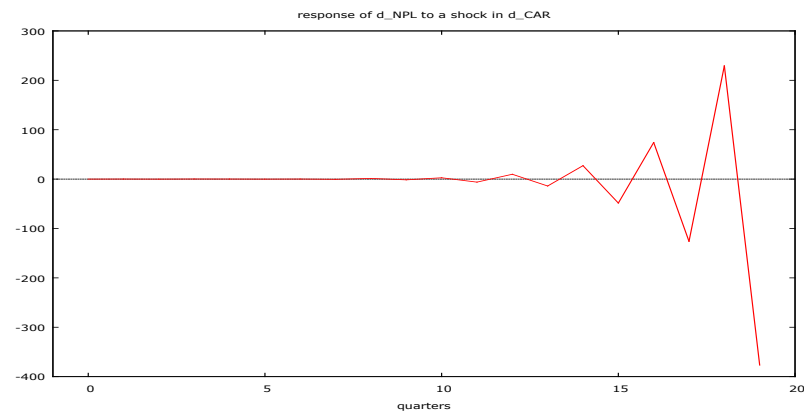
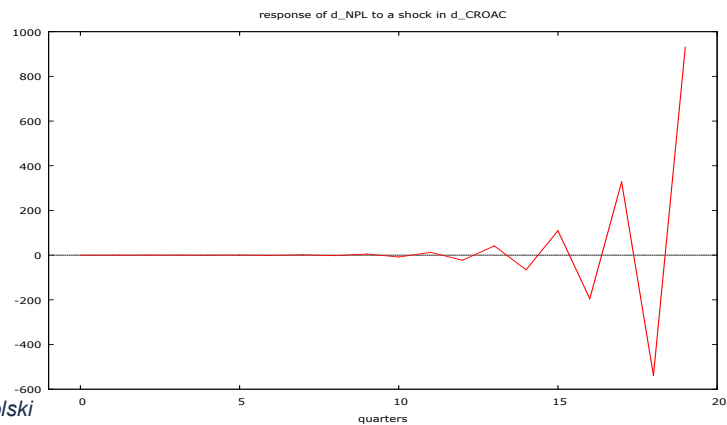
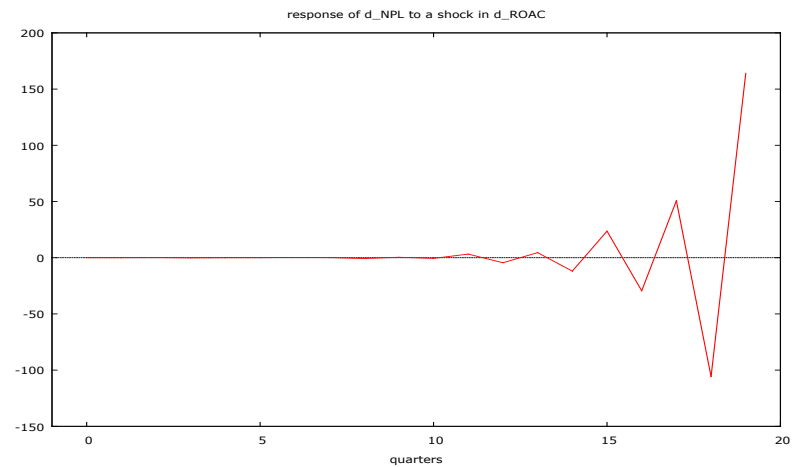
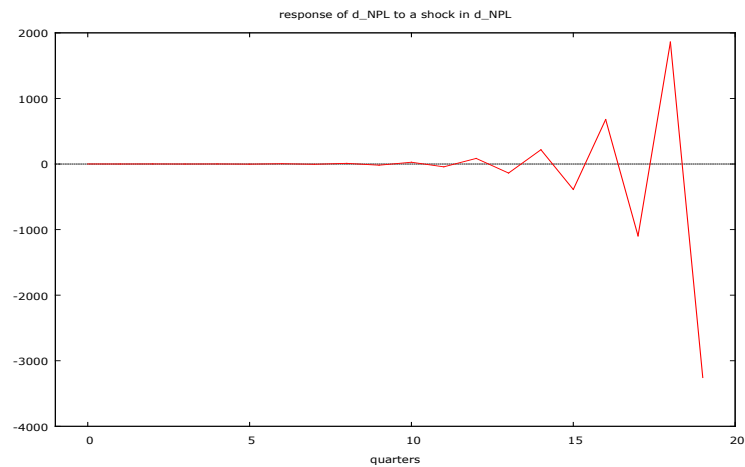
where:

$$\Gamma_i = \sum_{j=1}^i A_j - I, \quad i = 1, 2, \dots, k-1, \quad \Gamma_k = \pi = -\pi(1) = -\left(I - \sum_{i=1}^k A_i\right)$$

and I is a unit matrix.

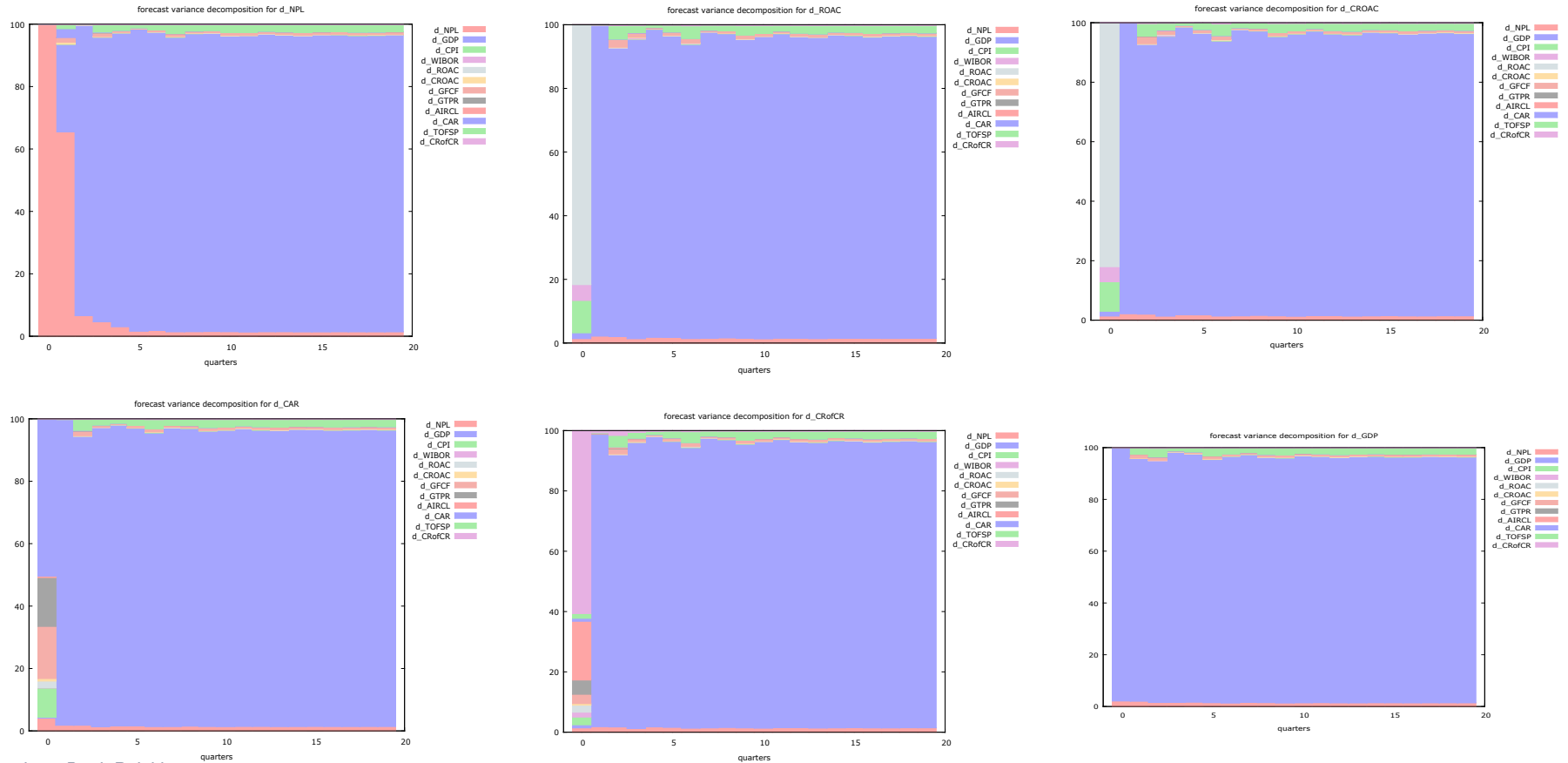
The impulse response functions

Figure 4. Forecast horizon 20q



The variance decompositions

Figure 5. Forecast variance decomposition (quarters)



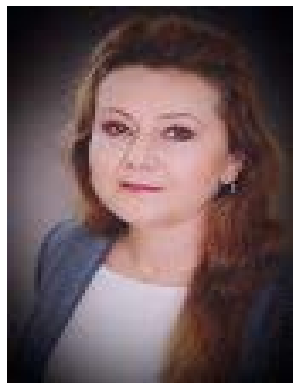
Conclusion and way forward

- I • The analysis of NPL changes shows that there was a long-term downward trend confirming the improvement in the quality of the portfolio of loans of non-financial companies in the period Q1.2009–Q4.2021. However, the last analysis quarters (COVID-19 time), brought an increase in the NPL ratio.
- II • The results of the VECM model confirmed the importance of revenues, economic situation (GDP), indicators of investments, costs of obtaining revenues on the part of corporations and total own funds on the part of banks.
- III • The results of the impulse response were confirmed by the results of the variance decomposition, indicating the importance of market and financial factors both in the volatility and the degree of explanation of NPL in the Polish banking sector. Also, the results of the NPLs research confirmed the pro-cyclical nature of lending activity in Poland in the verified years.
- IV • The results of the research confirm the importance of pursuing an investment policy focused on attracting new investments (new equity), including the so-called greenfield and on maintaining the existing ones (reinvestment of earnings).
- V • Further research should focus on the diagnosis of financial situation of corporation, the adequacy of macroprudential regulations, the credit policy on the part of commercial banks and the interest rate policy on the part of the central bank.



NARODOWY
BANK POLSKI

Thank for your attention



Aneta Kosztowniak

Economic Expert

Economic Analysis and Research Department

phone: +48 22 185 15 07

mobile: +48 691 034 170

e-mail: Aneta.Kosztowniak@nbp.pl; aneta.kosztowniak@wp.pl

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Is mobile money part of money? Understanding the trends and measurement / Evaluating mobile money access and use with non-traditional data sources¹

Kazuko Shirono, Bidisha Das, Yingjie Fan, Esha Chhabra and Hector Carcel-Villanova,
International Monetary Fund

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Is Mobile Money Part of Money? Understanding the Trends and Measurement

Kazuko Shirono, Bidisha Das, Yingjie Fan, Esha Chhabra and Hector Carcel-Villanova¹

Abstract

The rapid uptake of mobile money in recent years has generated new data needs and growing interest in understanding its impact on broad money. This paper reviews mobile money trends using mobile money data from the Financial Access Survey (FAS) and examines the statistical treatment of mobile money under the IMF's Monetary and Financial Statistics (MFS) framework. MFS guidance is straightforward in most cases, as many jurisdictions have adopted regulations which ensure that mobile money is captured in the banking system and thus in the calculation of broad money. However, in cases where mobile network operators (MNOs) act as niche financial intermediaries outside the banking regulatory perimeter and are allowed to invest their customer funds in sovereign securities and other permitted assets, mobile money liabilities may remain outside the banking system as well as monetary statistics. In that case, information on mobile money liabilities need to be collected directly from MNOs to account for mobile money as part of broad money.

Keywords: mobile money, financial inclusion, fintech, economic development, monetary and financial statistics, Financial Access Survey.

JEL classification: E42, E50, G28, M48, K20.

Contents

| | |
|--|----|
| Is Mobile Money Part of Money? Understanding the Trends and Measurement..... | 1 |
| 1. Introduction..... | 3 |
| 2. Mobile Money: Stylized Facts | 4 |
| A. What is Mobile Money? | 4 |
| B. Trends and Developments..... | 5 |
| Mobile Money Usage Patterns: Evidence from M-Pesa Transactional Data..... | 11 |
| 3. Mobile Money Ecosystem..... | 12 |
| A. Mobile Money Value Chain and Business Models..... | 13 |

¹ All authors are from the International Monetary Fund (IMF), except Yingjie Fan, who was Project Officer at the IMF when this paper was prepared. This paper was published as [IMF Working Paper No. 2021/177](#) and is being reproduced by permission of the IMF. The views expressed in this paper are those of the authors and do not necessarily represent the views of the IMF, its Executive Board or IMF management.

| | |
|--|----|
| Narrow Bank Model—Payment Banks of India..... | 15 |
| B. Regulations to Safeguard Mobile Money Customer Funds..... | 16 |
| 4. Mobile Money and Monetary Aggregates..... | 18 |
| Institutional Units and Sectors in Macroeconomic Statistics | 18 |
| Standardized Report Forms (SRFs): 1SR, 2SR, 4SR, and 5SR..... | 19 |
| A. Treatment of Mobile Money in Monetary Statistics..... | 20 |
| B. Compilation of Mobile Money Data for Monetary Statistics..... | 21 |
| Additional Treatments of Mobile Money in MFS Reporting..... | 23 |
| C. The impact of mobile money on broad money composition..... | 25 |
| 5. Conclusion..... | 27 |
| References..... | 29 |
| Annex I. Data Sources for Mobile Money..... | 32 |
| Annex II. Analysis of Transaction-level M-Pesa Data..... | 34 |
| Annex III. SRFs and Compilation of Monetary Aggregates under the <i>MFSMCG</i> | 37 |
| Annex IV. IMF's Monetary and Financial Statistics Database | 38 |

1. Introduction

The fast pace of economic digitalization in the financial sector is changing the way people access and use financial services, with new digital financial products and platforms emerging rapidly. These changes have prompted analysts, policymakers, and statisticians alike to seek new data sources and explore various approaches to systematically classify, measure, and record fintech related activities to evaluate their trends and support policy analysis (Cornelli et al., 2020; Adrian and Mancini-Griffoli, 2019; Claessens et al., 2018).

Mobile money is an early front-runner of such fintech innovations. It is a financial service using mobile money accounts typically offered by a mobile network operator (MNO) or another entity in partnership with an MNO. Unlike mobile banking, which is the use of an application on a mobile device to execute banking services, a bank account is not needed to use mobile money services—the only device required is a basic mobile phone.

Mobile money has had a profound impact on the financial sector landscape in low- and middle income economies, providing traditionally unbanked populations with secure and convenient means to carry out financial transactions and furthering financial inclusion (IMF, 2019a; Espinosa-Vega et al., 2020). While Africa is often considered as the epicenter of mobile money, the usage of mobile money has also grown significantly in other parts of the world, including Asia and Latin America. As of 2019, there are more than a billion registered mobile money accounts over which close to USD 2 billion transactions take place daily (GSMA, 2020a).

The rapid uptake of mobile money has generated new data needs to track its trends and developments for policy purposes. Cross-country comparable data on mobile money can serve as useful inputs for policymakers in formulating and designing targeted financial inclusion policies as well as assessing and benchmarking their impacts in a broader macroeconomic context. The COVID-19 pandemic has created even greater needs in this regard as mobile money can potentially facilitate financial transactions with minimal physical contact to support economic activity (Bazarbash et al., 2020).

The growing presence of mobile money has also raised questions of whether and how mobile money is counted as part of money in the economy, and what data may be needed to ensure mobile money to be properly captured in calculation of monetary aggregates such as broad money. These questions are fundamental in understanding the underlying data used for empirical and policy analysis, particularly given the fact that monetary aggregates are among key macroeconomic variables monitored by policymakers.² Clarifying the treatment of mobile money in monetary statistics, specifically using the IMF's Monetary and Financial Statistics Manual and Compilation Guide (MFSMCG) as a methodological framework, can offer a useful insight in this regard.

Against this background, this paper analyzes recent developments of mobile money, including trends in adoption and usage as well as business models and regulatory requirements. It then examines, from a statistical perspective, the implications of these developments for the measurement of monetary aggregates such as broad money. In doing so, it contributes to the growing literature on measuring digitalization in macroeconomic statistics more broadly.³

The review of the data on mobile money from the International Monetary Fund's (IMF's) Financial Access Survey (FAS), a supply-side database on financial access and use, points to the fact that mobile money now

² For example, some studies have examined empirical relationships between mobile money and monetary policy related variables such as money multipliers and broad money (e.g., Kipkemboi and Bahia, 2019; Mawejje and Lakuma, 2019; Aron et al., 2015).

³ See IMF (2018) and OECD (2020), for example.

offers more access points than traditional banking services in many low- and middle-income economies, with a larger number of mobile money agents available than ATMs and bank branches. Mobile money is also widely adopted in developing economies, with the number of registered mobile money accounts being greater than that of bank accounts in some cases. The FAS data also suggest that mobile money usage measured by transaction values and volumes has increased significantly in many economies over the past years.

Mobile money services are also starting to expand to include new and enhanced products such as credit and interest-bearing savings in some countries. Transaction-level data analysis of mobile money accounts in Kenya confirms this trend and reveals the degree of penetration of these new services among mobile money users included in the study.

The methodological guidance provided by the MFSMCG on the treatment of mobile money is relatively clear-cut—mobile money liabilities, namely outstanding balances in mobile money accounts, need to be included as part of broad money in monetary statistics. However, how mobile money affects the measurement of monetary aggregates, in practice, depends on the mobile money business model and/or the regulatory framework. In most cases, the application of the guidance is straightforward as many jurisdictions have adopted regulations which ensure mobile money to be included in the calculation of monetary aggregates. However, some cases—including when regulations allow mobile money liabilities to be invested in sovereign securities or other permitted assets—require more involved steps and additional data collection to account for mobile money in compilation of monetary statistics. In these cases, compilers of monetary statistics will need to review the prevailing situation in the country and make necessary adjustments in the calculation of broad money.

The rest of the paper is organized as follows. Section II provides an overview of the latest mobile money trends in access, adoption, and usage dimensions, drawing on the FAS database which contains country-level annual data on mobile money. Section III examines mobile money business models and key aspects of mobile money regulations—important factors in understanding the treatment of mobile money in macroeconomic statistics and its impact on the measurement of money. Section IV discusses the guidance from the MFSMCG framework on how to record mobile money and clarifies its implications for calculation of money under different business models and regulatory arrangements. Section V concludes.

2. Mobile Money: Stylized Facts

Mobile money has become a preferred mode of accessing financial services particularly in countries without deep banking penetration and with limited infrastructure. It has played a significant role in facilitating financial inclusion in some of these economies. The novelty of mobile money lies in its ease of access and use. This section starts with defining mobile money and provides an overview of mobile money access, adoption and usage trends across the globe.

A. What is Mobile Money?

Mobile money is a form of mobile payment service typically offered by an MNO or another entity in partnership with an MNO using mobile money accounts. To use mobile money services, customers usually

need to register with a mobile money agent—typically small, local retail stores—of the mobile money service provider and obtain an individual virtual account linked to their mobile phone number, accessible through a SIM card. A bank account is not required to use these services—the services may be accessed only with a basic mobile phone. Mobile money customers can credit funds into the mobile money account by giving cash to a mobile money agent, and in return, receive “mobile money” of equivalent amount via their mobile phone.⁴ They use this electronically stored mobile money to pay their bills, transfer money to their peers, etc.

Mobile money users can also withdraw the money received as salary, deposits, or payments on their mobile money accounts using a mobile money agent. More recently, some MNOs have started to expand financial services offered via mobile money, such as interest earning savings and loans. All these activities are conducted without opening a bank account or visiting a bank branch or an ATM.

The benefits of mobile money have been relatively well documented. Studies have noted that mobile money is a safe, affordable store of value and means of funds transfer for sections of population with no access to traditional financial services (Dupas et al., 2018). It has significantly cut transaction costs of money transfers and remittances (Jack and Suri, 2014). Other studies have focused on how mobile money can facilitate efficient informal risks sharing—by enabling timely transfers of money among family/community members in times of financial distress. This in turn allows households to smooth their consumption and make more efficient investment decisions (Jack and Suri, 2011; Munyegera and Matsumoto, 2016; Suri and Jack, 2016; Blumenstock et al., 2016; Jack and Suri, 2014; Riley, 2018).

While these studies based on micro-level household surveys provide useful insights at the individual country level, it is useful to examine mobile money developments in a global context to gain a broader perspective. The next subsection examines cross-country data to highlight key mobile money trends in recent years.

B. Trends and Developments

The analysis of this subsection draws on mobile money data from the IMF’s FAS, a unique supply-side database which contains, as part of its data collection, annual country-level data on three key aspects of mobile money—access, adoption and usage.

Financial Access Survey

The FAS, launched in 2009, has been collecting annual data on access to and use of financial services covering 189 jurisdictions with data spanning over 15 years. In 2014, the FAS introduced a mobile money module in its annual data collection exercise to gather mobile money data on seven different series with historical data dating back to 2007—the time when mobile money came into prominence after its launch in Kenya. Using these series, the FAS produces ten indicators of mobile money (Table 1). As of 2020, 79 countries (about 90 percent of countries that have mobile money services) report mobile money data to

⁴ If the mobile money user has a bank account, they can transfer funds to the mobile money accounts from the bank account. However, people without bank accounts can also access mobile money services.

the FAS on an annual basis. The source of mobile money data in the FAS is administrative data obtained from the regulators of mobile money service providers.⁵

Mobile Money Data Collected in the FAS

FAS Mobile Money Indicators

Table 1

| | |
|-----------------|--|
| Access | Registered mobile money agent outlets per 1,000 km ² |
| | Registered mobile money agent outlets per 100,000 adults |
| | Active mobile money agent outlets per 1,000 km ² |
| | Active mobile money agent outlets per 100,000 adults |
| Adoption | Registered mobile money accounts per 1,000 adults |
| | Active mobile money accounts per 1,000 adults |
| Usage | Value of mobile money transactions, percentage of GDP |
| | Number of mobile money transactions per 1,000 adults |
| | Outstanding mobile money balance on active accounts, percentage of GDP |
| | Average number of transactions per active mobile money account |

Source: IMF staff and IMF Financial Access Survey.

Insights from the FAS Data

The FAS mobile money data offer a cross-country perspective on mobile money trends and developments. Specifically, these data point to the following three themes on access, adoption, and usage:

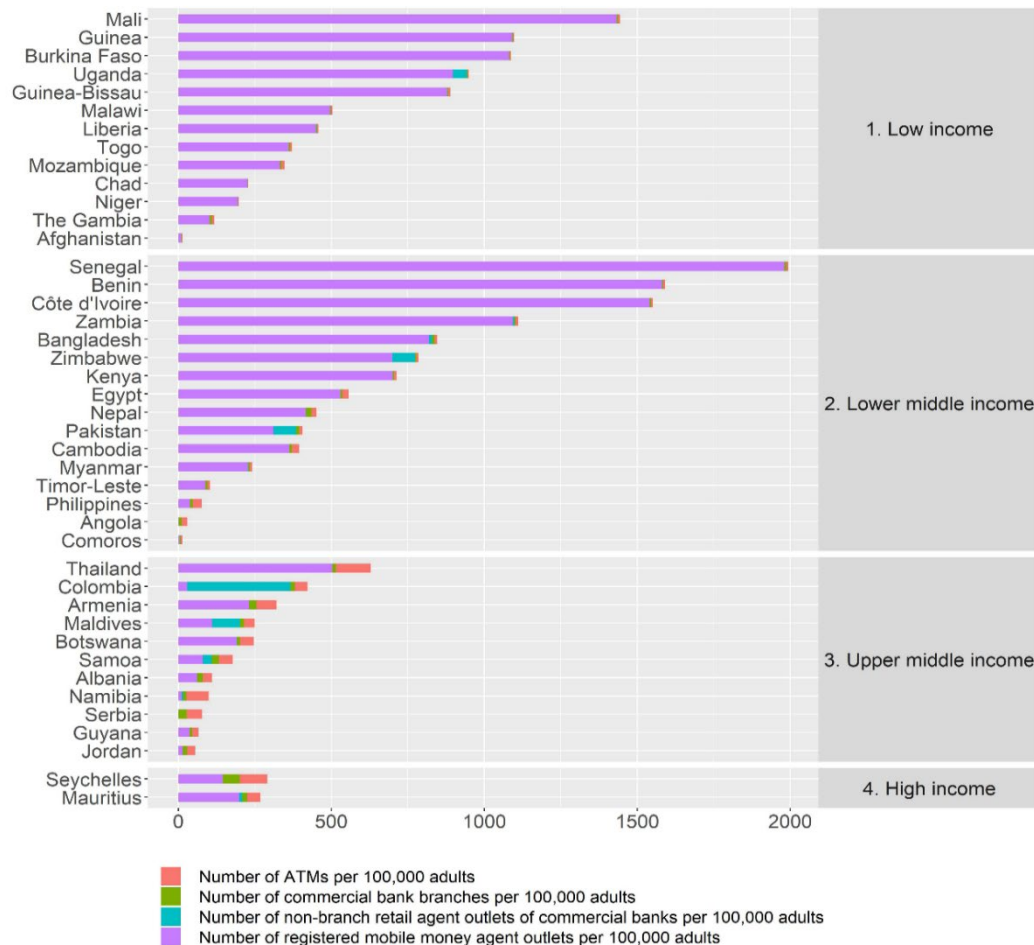
First, mobile money is now more accessible than traditional banking in several low- and middle-income countries. Mobile money is a service brought to its customers by a network of mobile money agents while banks in developing economies typically rely on physical branches and ATMs. In the last decade or so, banks have also offered “agent banking” to broaden financial service provision in geographical areas not reached by bank branch networks.⁶ Despite such innovations by banks, in many low- and middle-income countries, the presence of mobile money agents is higher than ATMs, commercial bank branches and non-branch retail agents combined (Chart 1). For example, in Guinea, for each commercial bank branch, there are 174 mobile money agents as of 2019. Similarly, in Mali, for every commercial bank branch, there are close to 100 mobile money agents.

⁵ While the FAS provides supply-side annual country level data on mobile money, there are also other databases which cover a range of aspects of mobile money. See Annex I for select data sources.

⁶ Agent banking is based on non-branch retail agents, which typically include retail stores, post offices and small businesses acting on behalf of banks to carry out financial transactions. Non-branch retail agents are different from mobile money agents. The range of financial services provided by non-branch retail agents is usually limited, often including account opening, and cash-in/cash-out transactions. These retail agent outlets are also known as “business correspondents.”

Mobile Money Agents are Dominant Access Points in Many Developing Economies

Chart 1



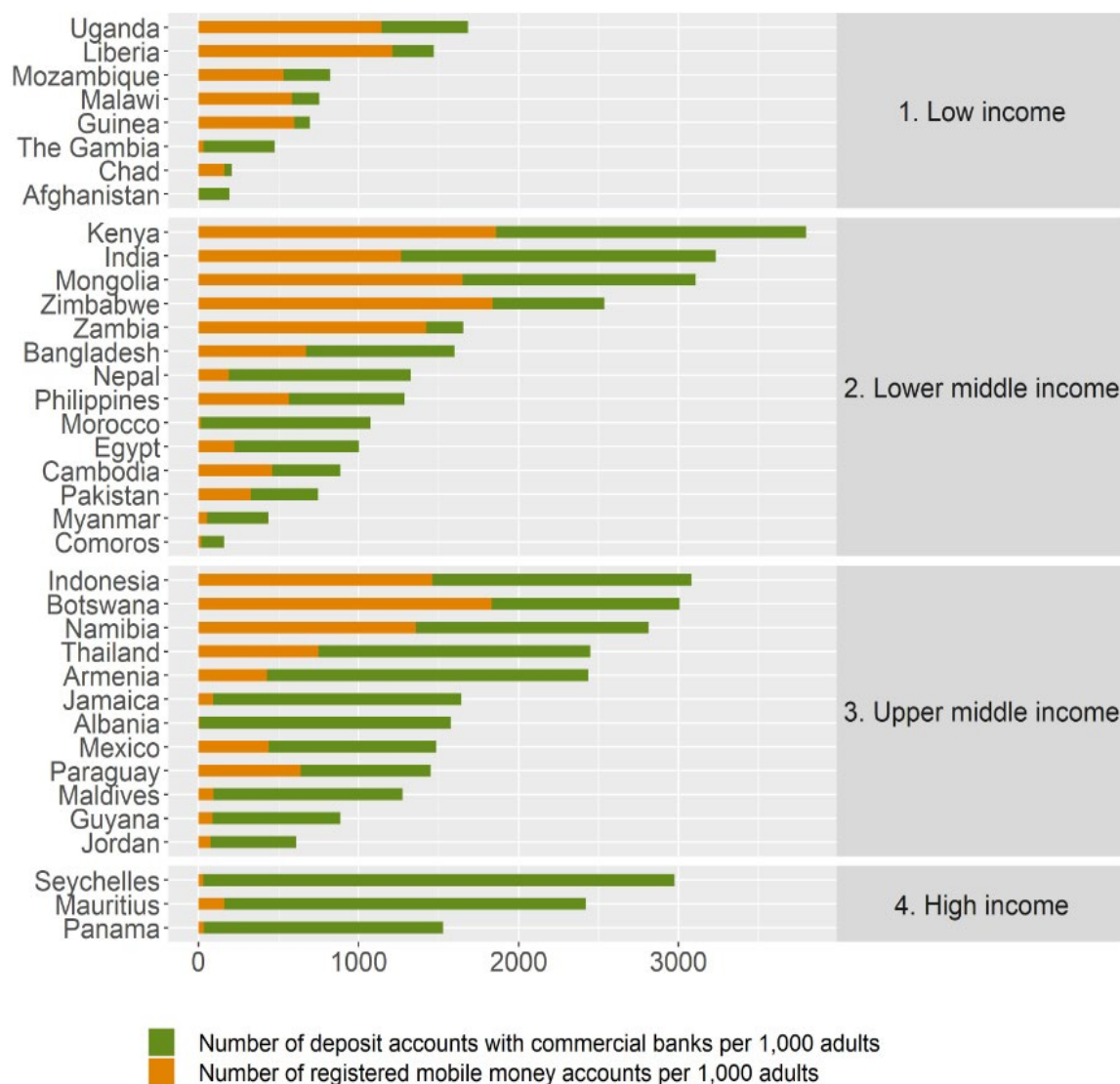
Source: IMF staff calculations and IMF Financial Access Survey.

Note: This figure shows the number of access points for mobile money (registered mobile money agent outlets) as well as for traditional banking, which includes ATMs, commercial bank branches, and non-branch retail agent outlets. Data cover 2019 or the latest year available.

Second, the adoption of mobile money as a popular mode of financial access is evident in the fact that there are more registered mobile money accounts than bank accounts in several low- and middle- income countries (Chart 2). This trend is more pronounced among low-income countries. In the middle-income countries, mobile money seems to play a complementary role to traditional financial services offered by commercial banks.

Mobile Money Accounts are Widely Adopted in Many Economies

Chart 2



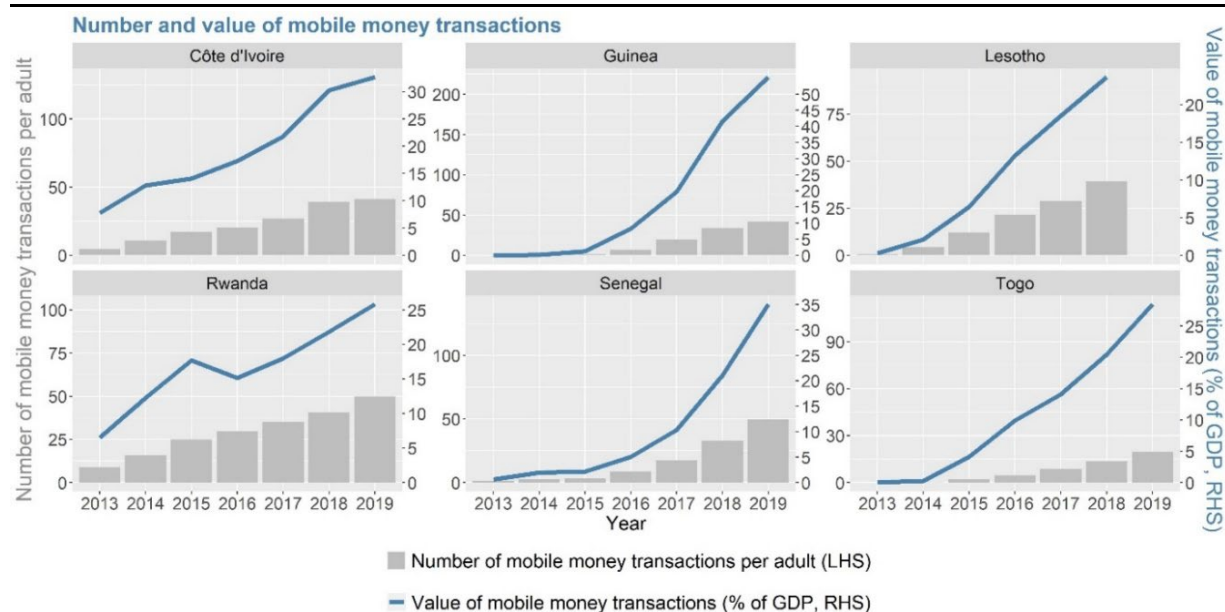
Source: IMF staff calculations and IMF Financial Access Survey.

Note: This figure shows the number of registered mobile money accounts per 1,000 adults and deposit accounts with commercial banks in 2019 or the latest year available.

Third, mobile money usage measured by transaction values and volumes has increased significantly over time, especially among early adopters of mobile money. In line with the growing access and adoption indicators, FAS mobile money usage indicators show fast growth over the last decade for many economies (Chart 3).

Mobile Money Usage Has Grown Over Time

Chart 3



Source: IMF staff calculations and IMF Financial Access Survey.

Note: This figure shows the average number of mobile money transactions per adult (gray bars on the left y-axis) and the total value of mobile money transactions as percentage of GDP (blue lines on the right y-axis) during 2013-19 for select countries.

New Roles of Mobile Money

The FAS data also suggest that mobile money users have started maintaining higher balances in their mobile money accounts. Outstanding balances on active mobile money accounts can be seen as something similar to bank deposits a sum converted from cash to mobile money but not yet used to transfer to peers or pay bills. Chart 4 shows the outstanding balances on active mobile money accounts as a percentage of GDP for select economies. The outstanding balances have risen over time, suggesting that mobile money accounts may be increasingly used as a way to store money, beyond being a mere payment platform.⁷

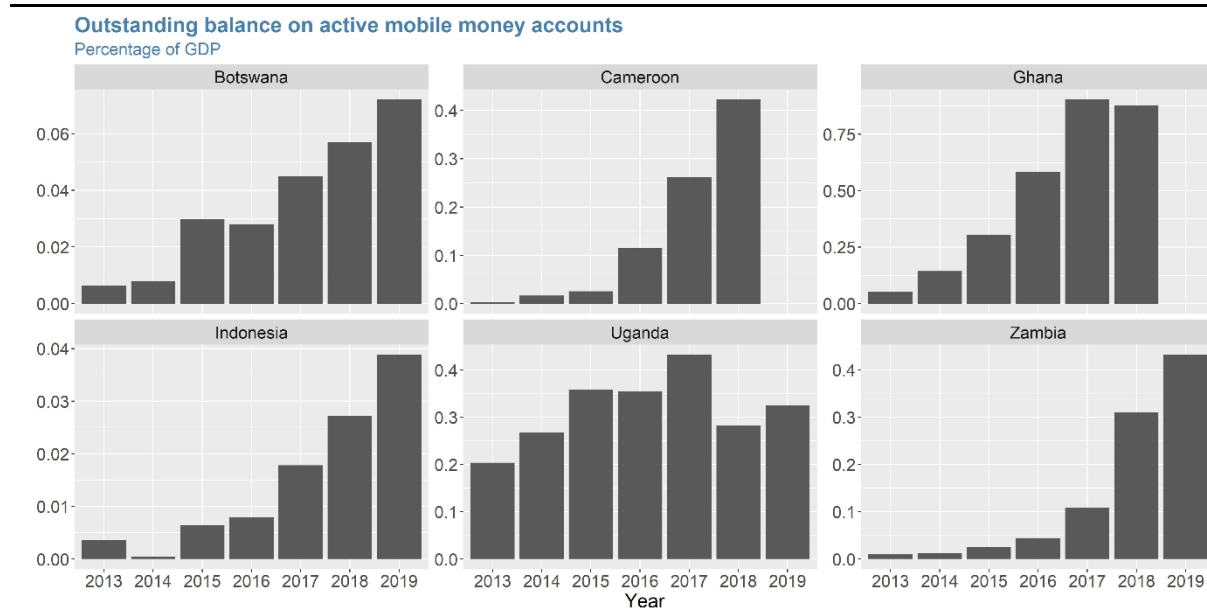
As mobile money matures, new and enhanced services including credit and insurance may be offered through the mobile money channel on a larger scale in collaboration with traditional financial service providers including insurance firms, banks, and microfinance institutions (GSMA, 2018). This trend is already seen in some of the early adopters of mobile money including Kenya and Tanzania, where mobile money account holders can apply for loans/microcredit through their mobile money service provider (Safaricom/Vodafone).⁸

⁷ Outstanding balance over the number of users (average balance, not shown) also shows a similar trend for most countries available in the FAS.

⁸ Other examples include Commercial Bank of Africa, a banking partner of Safaricom/Vodafone, which uses the usage history of the mobile money account to develop a credit score for the applicant, assigns individual credit limits, and decides on the application. The only way to withdraw or repay these loans is via the mobile money account. In Kenya, Safaricom's microcredit service is known as M Shwari and in Tanzania, a similar service offered by Vodafone is known as M-Pawa. Kenya Commercial Bank Group (KCB), a financial

Mobile Money Balances are Growing Over time

Chart 4



Source: IMF staff calculations and IMF Financial Access Survey.

Note: This figure presents the outstanding balance on active money accounts as percent of GDP during 2013-19 for select countries with all available data in the FAS. The 2019 data are not available for Cameroon and Ghana.

While these developments are important to monitor, data are limited and the degree of actual usage of these new services is yet to be well understood. One way to shed light on this is to draw on transaction level data. Consultative Group to Assist the Poor (CGAP) has recently made available transaction level mobile money data on M-Pesa accounts for Kenya, collected as part of a survey conducted by the Busara Center for Behavioral Economics, during the period of July 2017-August 2018 (see Box 1). While the sample size is limited, the data reveal interesting usage patterns. Specifically, the data show that mobile money users are leveraging additional financial services offered through this channel, with roughly half of the M Pesa accounts in the study having been used for new services such as loans and interest bearing savings.

services provider headquartered in Nairobi, also provides a savings service that enables M-Pesa customers to earn interest from their savings balance.

Mobile Money Usage Patterns: Evidence from M-Pesa Transactional Data

Given the lack of transaction level data in the FAS or GSMA, relatively little has been studied about mobile money usage patterns. Transaction level M-Pesa data made publicly available by the Consultative Group to Assist the Poor⁹ (CGAP) help shed light in this regard. The data consist of anonymized 418 M-Pesa accounts belonging to low-income users in Nairobi between July 2017-August 2018 (see Annex II for details). While the data may not be nationally representative, it provides important insights into the use of mobile money in Kenya, especially the usage patterns and transaction volumes and values among low-income users.

Using entity recognition and keywords extraction techniques, the transactions during the sample period were classified into eight main categories as listed in the summary table.

| Category | Number of transactions | Share of total transactions (percent) | Usage: Number of accounts | Usage: Share of total accounts (percent) | Average value (USD) |
|---------------------------------|------------------------|---------------------------------------|---------------------------|--|---------------------|
| P2P Transactions | 41433 | 28.07 | 416 | 99.52 | 6.6753827 |
| B2P / P2B Transactions | 33311 | 22.56 | 363 | 86.84 | 3.7215588 |
| Airtime Purchase | 27992 | 18.96 | 410 | 98.09 | 0.3009765 |
| Deposit and Withdrawal of Funds | 26599 | 18.02 | 418 | 100.00 | 12.4832918 |
| Transaction Fees | 12645 | 8.57 | 416 | 99.52 | 0.3012844 |
| Financial Derivatives | 5084 | 3.44 | 202 | 48.33 | 12.2738078 |
| Others | 532 | 0.36 | 187 | 44.74 | 14.9481398 |
| International Transfers | 36 | 0.02 | 10 | 2.39 | 84.9181282 |

Source: IMF staff calculations from CGAP.

Note: Usage percent of total accounts shows how many of the total accounts are being used for the type of transactions.

Key findings include the following:

- Mobile money is mostly used to conduct person-to-person (P2P) and business-to-business (B2B)/person-to-business (P2B) transactions, accounting for about 28 percent and 23 percent of all mobile money transactions respectively. Unsurprisingly, all mobile money accounts carried out at least one or more P2P transactions during the sample period. The average value of such transactions is USD 6.7. More than 85 percent of accounts in the sample engaged in B2B/B2P transactions. Transaction details also reveal that M-Pesa is being used to pay salaries, which corroborates that both the public and private sectors are increasingly relying on mobile payment services given its low-cost nature (Wasunna and Frydrych, 2017).

⁹ CGAP is a global partnership of more than 30 leading development organizations that works to advance the lives of poor people through financial inclusion.

- New services including credit and insurance are being offered through mobile money. Close to 50 percent of the accounts have been used for these additional financial services such as M-Shwari loans (reported under “Derivatives”—see Annex I).
- The share of international remittances is small, but the average value of such transactions is much higher, standing at USD 85, more than ten times an average domestic P2P transfer. In addition, accounts that receive international remittances carry average balances 50 percent higher than the average of those that do not receive them.
- New services including credit and insurance are being offered through mobile money. Close to 50 percent of the accounts have been used for these additional financial services such as M-Shwari loans (reported under “Derivatives”—see Annex I).
- The share of international remittances is small, but the average value of such transactions is much higher, standing at USD 85, more than ten times an average domestic P2P transfer. In addition, accounts that receive international remittances carry average balances 50 percent higher than the average of those that do not receive them.

Transaction level data have great potential to unveil mobile money usage patterns and the degree of penetration of new products. Partnership with private sector entities including MNOs is key to gain further insights into mobile money usage.

In sum, mobile money has gained broader traction in many low- and middle-income economies for the past decade. At the same time, increasing usage of mobile money has prompted growing interest in how mobile money—financial services provided by MNOs—is treated in monetary statistics and how it is captured in measuring money. The following sections cover these measurement issues in detail.

3. Mobile Money Ecosystem

To examine mobile money measurement issues, it is important to understand the mobile money ecosystem, including its business models and regulatory environments—key factors determining the treatment of mobile money in monetary statistics.¹⁰

A distinct feature of mobile money is the role of non-financial institutions like telecom companies in offering basic financial services to customers who would otherwise be excluded or underserved. For example, M-PESA, launched in Kenya in 2007, is operated by Safaricom, Kenya’s largest telecommunication provider—a non-bank. Vodacom in Tanzania and Globe Telecom in Philippines are other examples of non-bank mobile money service providers. This mode of mobile money service provision is characterized as a “MNO-led model.”

In some countries, however, banks have started partnering with third parties such as mobile network operators to offer financial services to the unbanked populations via mobile phones. This has led to the growth of a “bank-led model” of provision of mobile money services. Despite bank involvement, users do

¹⁰ This paper mainly focuses on the mobile money business model, but other players such as merchants and regulators are also an important part of the mobile money ecosystem.

not need to have bank accounts to use these mobile money services, just as with MNO led mobile money services.

While both MNO-led and bank-led mobile money services offer similar user experience, the mechanics of service provision under these two models are structured somewhat differently. In addition, these two models may be subject to different regulations, which have important implications for the treatment of mobile money in calculating monetary aggregates. The following subsections provide details on mobile money business models and corresponding regulatory frameworks.

A. Mobile Money Value Chain and Business Models

Determining the type of mobile money business model starts with defining the value chain of the mobile money model. This involves identifying five basic roles to be fulfilled for mobile money service provision (Bill and Melinda Gates Foundation, 2015):

- **Telecom channel provider:** This is the institution which provides network access to users. This role is carried out by an MNO, irrespective of business model type.
- **Agent network manager:** This role is important to the success of mobile money service provision. Mobile money agents are the interface between the mobile money customers and the mobile money service provider. An extensive network of mobile money agents is considered a key driver for the penetration of mobile money among often unserved last mile customers.
- **Payment service provider:** The institution carrying out this role provides the front-end interface including the phone interface for agents and customers; the back-end processing; and is responsible for clearing and settlement.
- **Mobile-money issuer:** This is the institution responsible for issuing mobile money. In other words, the mobile money issuer takes on the corresponding liability for issuance of mobile money if a mobile money user wants to convert his mobile money to cash, the mobile money issuer is legally bound to provide the required funds.¹¹
- **Deposit holder:** This role, irrespective of the business model, is carried out by a financial institution, typically a bank, which is responsible for safekeeping the funds deposited by mobile money customers.

What determines whether a mobile money service is MNO-led or bank-led is the type of the entity which takes on the mobile money issuance role. If a bank (or MNO) books the corresponding liability for deposits of mobile money, the mobile money service is classified as the bank-led (or MNO-led) model.

The MNO-led model can be also generalized as a “non-bank led model” as the role of mobile money issuer can be carried out by a third party (e.g., a fintech company)—in this case, it is called a “third-party led model.” For simplicity, this paper focuses on the MNO-led and bank-led models and treats the third-party led model as a variant of the MNO-led model. The same statistical treatments applied to the MNO-led model also hold for the third-party led model below.

It is important to note that regardless of the business model, both banks and MNOs are essential to offer these services. While MNOs offer the telecom channel, banks hold the funds that can be converted into

¹¹ A liability is established when one unit (the debtor) is obliged, under specific circumstances, to provide funds or other resources to another unit (the creditor) (see MFSMCG 4.6).

mobile money. MNOs and banks may also partner with other third parties to offer some of these functions (Table 2).

Mobile Money Value Chain: Key Components

Table 2

| | <u>MNO-led Model</u> | <u>Bank-led Model</u> |
|---------------------------------|----------------------|-----------------------|
| Telecom channel provider | MNO | MNO |
| Agent Network manager | MNO/Third party | Bank/MNO/Third party |
| Payment service provider | MNO/Third party | Bank/MNO/Third party |
| Mobile money issuer | MNO | Bank |
| Deposit holder | Bank | Bank |

Source: IMF staff and Bill and Melinda Gates Foundation (2015).

Most mobile money service providers are MNO-led (Chart 5), but MNO-led and bank-led models can also co-exist in a country. In some cases, however, regulations prohibit non-banks from operating as mobile money providers (e.g., India, Bangladesh), and some variants of these models are also possible depending on mobile money regulations of the country. Examples of mobile money business models, including variants, are as follows:

- **MNO-led model:** MNOs are often big telecom service providers uniquely positioned to provide mobile money services as they typically have a network infrastructure in place and brand recognition among customers. Examples include M-PESA in Kenya, M-Pitisan in Myanmar, and MTN mobile money in Uganda. A variant of the MNO-led model, although not as common, also exists wherein a third party provides the mobile money service in partnership with an MNO ("third-party led model"). Examples include OPay and Palm Pay in Nigeria, founded by a Chinese-owned internet company, Opera.
- **Bank-led model:** Given banks are at the forefront of offering these services as registered financial institutions, the bank-led model is sometimes considered more secure than the MNO-led model. Examples include bKash in Bangladesh, Hello Paisa in Nepal, and Eazy Money in Nigeria. A variant of the bank-led model is a narrow-bank model, wherein a new type of institution—such as payment banks in India—is created under existing banking laws and can offer a limited set of financial services (see Box 2).

Different business models are more dominant than others in different regions. Close to two-thirds of the live mobile money services in Sub Saharan Africa, and Middle East and North Africa are MNO-led. On the other hand, in East Asia and Pacific and Europe and Central Asia, that number is only about one-third. These differences can be in part attributed to the legal and institutional frameworks in the region. Pelletier et al. (2020) show that MNOs are more likely to launch mobile money services in countries where legal rights are weaker and credit information is less prevalent

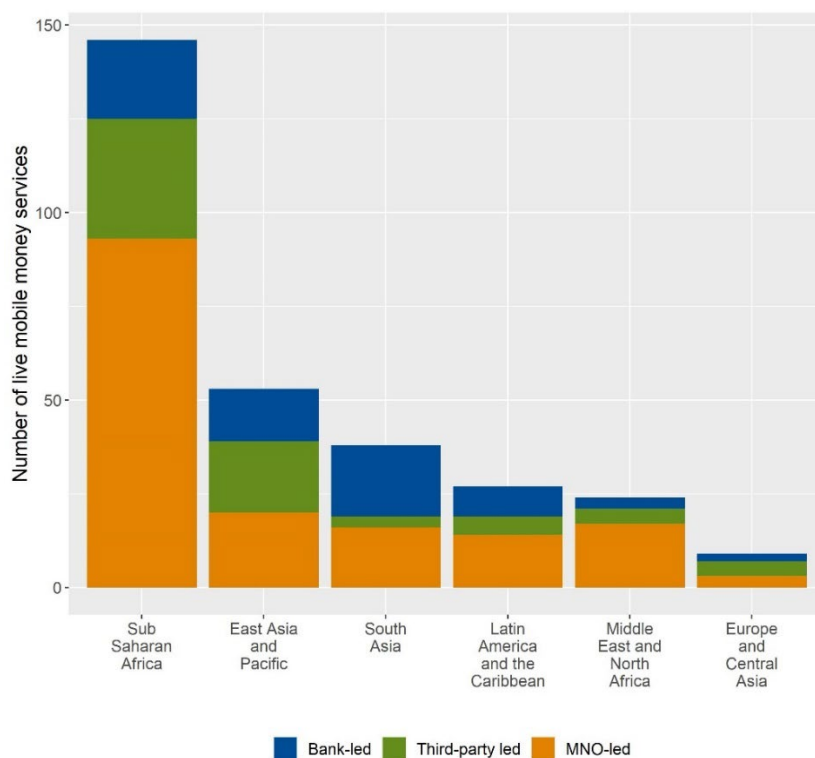
Narrow Bank Model—Payment Banks of India

In India, non-bank e-money issuers including MNOs need to register themselves as payment banks with the Reserve Bank of India. These MNOs which offer mobile money services obtain a type of banking license which allows them to accept deposits, issue ATM/debit cards, and offer other financial services except lending (Reserve Bank of India, 2014).

The payment banks are required to invest a minimum 75 percent of their "demand deposit balances" in government securities/treasury bills with maturity up to one year and hold a maximum 25 percent in current and time/fixed deposits with other scheduled commercial banks¹² for operational and liquidity management purposes.

MNO-Led Mobile Money Services Dominate in Most Regions

Chart 5



Source: IMF staff calculations and GSMA mobile money deployment tracker.

¹² A bank listed in the second schedule of the Reserve Bank of India (RBI) Act, 1934 is considered a scheduled bank. Scheduled banks are required to follow guidelines formulated by the RBI and maintain reserves with the RBI. Scheduled banks are eligible for loans from the RBI at bank rate and are given membership to clearing houses.

B. Regulations to Safeguard Mobile Money Customer Funds

Mobile money regulations typically cover areas such as licensing, AML/CFT, and customer protection (Pelletier et al., 2019). This subsection focuses on mobile money regulations related to customer protection—specifically those put in place to safeguard customer funds stored in the mobile money system. Depending on the mobile money model in question, different regulations apply, and these regulations have direct implications for the treatment of mobile money in monetary statistics.

One of the most significant risks associated with mobile money usage, especially from the perspective of mobile money users, is loss of customer funds, which can arise due to insufficient liquidity or insolvency of the mobile money service provider. This concern is more pronounced for MNO-led mobile money services as these are operated by non-banks, which lie outside the scope of banking regulations.

To mitigate these risks, countries have introduced mobile money specific regulations over time, but it has been an evolving process. In 2007, when Safaricom began operating M-Pesa in Kenya, there was no law or regulation in the country for mobile money service providers. At the time, they started operating on the basis of a letter agreement with the Central Bank of Kenya. It was only in 2014 when, with the adoption of the National Payment Systems (NPS) Regulations, a formal legal framework for mobile money was put in place in the country. The NPS regulations have implemented several regulatory practices, including the safeguarding of customer funds. Following this example, other countries also adopted similar regulations, such as the Bank of Tanzania's e-Money Regulations in 2015 and the Bank of Ghana's Guidelines for e-Money Issuers in 2015 (Greenacre, 2018). All countries now mandate all mobile money service providers to keep 100 percent of their mobile money liabilities in liquid assets (GSMA, 2020b). Details are as follows:

MNO-led model

To date, the most common form of regulations for the MNO-led mobile money services is requiring MNOs to hold their mobile money liabilities as deposits at regulated financial institutions. However, some countries also use a combination of frameworks including allowing MNOs to invest in low risk securities. Depending on the jurisdiction, MNOs have different regulatory options on maintaining the liquid assets:

- **Holding customer funds with banks:** The most common approach is to require MNOs to deposit their entire outstanding mobile money liabilities in one or more banks or other regulated deposit taking institutions. Depending on a country's legal system, this type of account may be called a trust, escrow, fiduciary or custodial account (Kerse and Staschen, 2018). In Eswatini for instance, an MNO needs to ensure that balances in the trust account should always be equal to the total outstanding mobile money liabilities (Central Bank of Eswatini, 2019). In Seychelles, MNOs need to split the liabilities into trust accounts with a minimum of two banks (Central Bank of Seychelles, 2014). Other countries which follow a similar approach include Kenya, Ghana, Paraguay and Uganda.¹³
- **Depositing customer funds with the central bank:** Though not a common approach, some regulators require that MNOs hold at least a portion of their outstanding mobile money liabilities in the central bank while holding the rest in other regulated financial institutions like commercial banks. In a few limited cases, the regulations may even require the MNOs to hold their entire mobile money issuance as deposits with the central bank. The risk of loss of funds is minimal with the central bank being the lender of last resort. For example, Article 10 of El Salvador's Financial Inclusion Bill requires that mobile

¹³ Central Bank of Kenya E-money Regulations 2013, Ghana e-Money Guidelines, Art. 16, Resolución No. 6 de 2014 – Reglamento de Medios de Pagos Electrónicos, Art. 15, Uganda e-Money Guidelines, Art. 16.

money service providers store customer funds in the central bank (Asamblea Legislativa de El Salvador, 2015).

- **Investing funds in other liquid assets:** Some countries give the option to MNOs to either maintain funds at regulated financial institutions or invest the customer funds in low-risk securities, such as government securities. This is the case for the West African Economic and Monetary Union (WAEMU), where customer funds can be invested in bank deposits but also in securities issued by central governments, regional financial institutions, and companies listed on the West African Regional Securities Exchange (BCEAO, 2015). In Malaysia, funds may be invested in high quality liquid assets in the form of deposits placed with licensed institutions or debt securities issued or guaranteed by the Federal Government and the Central Bank (Bank Negara Malaysia, 2003). In the Philippines, customer funds may be invested in bank deposits, government securities or other permitted liquid assets (Bangko Sentral ng Pilipinas, 2009). Similarly, in Bolivia, outstanding mobile money liabilities must be held in either cash by MNOs or invested in securities issued by the Bolivian Government and Central Bank or other permitted assets by the Central Bank (Autoridad de Supervisión del Sistema Financiero de Bolivia, 2018) (see Table 3).

Select Examples Allowing MNOs to Invest in Other Liquid Assets

Table 3

| Country | Regulation | Year | Regulatory Requirement | Source |
|-------------|--|------|--|--|
| Bolivia | "Modificaciones al reglamento para empresas de pago móvil, al reglamento de fidecoismo, y al manual de cuentas para entidades financieras, Circular ASFI 548." | 2018 | Outstanding mobile money liabilities must be held in either cash by MNOs or invested in securities issued by the Bolivian Government and Central Bank, or other permitted assets by the Central Bank. | Autoridad de Supervisión del Sistema Financiero de Bolivia (page 26, article 8) |
| Malaysia | "Guideline on Electronic Money (E-Money), Act 627." | 2003 | Funds may be invested in high quality liquid assets in the form of deposits placed with licensed institutions or debt securities issued or guaranteed by the Federal Government and the Central Bank. Also, other instruments as may be specified by the Bank. | Bank Negara Malaysia (page 10, section C) |
| Philippines | "Guidelines governing the Issuance of Electronic Money (e-money) and the Operations of Electronic Money Issuers (EMI), Circular No. 649." | 2009 | Customer funds may be invested in bank deposits, government securities or other permitted liquid assets. | Bangko Sentral ng Pilipinas (page 4, section D) |
| WAEMU | "Instruction N°008-05-2015 regissant les Conditions et Modalités d'Exercice des Activités des Emetteurs de Monnaie Électronique dans les États Membres de l'Union Monétaire Ouest Africaine (UMOA)." | 2015 | Customer funds can be invested in bank deposits but also in securities issued by central governments, regional financial institutions, and companies listed on the West African Regional Securities Exchange. | BCEAO (page 30) |

Source: IMF Staff based on relevant documents listed in the source column.

Bank-led model

Mobile money issued by banks may be classified in their books either as deposit liabilities or as a distinct mobile money liability within deposit liabilities. Mobile money customer funds are typically pooled in a single account rather than individual accounts (Grossman, 2016). These funds are monitored under the overall prudential supervision of the bank (Izaguirre et al., 2019).

The next section examines how these mobile money business model types and regulations affect the treatment of mobile money in monetary statistics and the calculation of monetary aggregates.

4. Mobile Money and Monetary Aggregates

Methodological questions pertaining to the compilation of monetary statistics are best discussed in the context of the IMF's [2016 Monetary and Financial Statistics Manual and Compilation Guide \(MFSMCG\)](#) which offers guidance for the compilation of monetary statistics to promote the production of cross-country comparable monetary data on the central bank, other depository corporations (ODCs), and other financial corporations (OFCs) (see Box 3).

Box 3

Institutional Units and Sectors in Macroeconomic Statistics

For macroeconomic statistics purposes, all corporations—namely, institutional units that produce goods and services for the market, are legally separated from their owners, and are legally liable for their actions—are divided into financial corporations and nonfinancial corporations. The 2008 System of National Accounts (SNA) divides the Financial Corporations sector into nine subsectors: (1) central bank, (2) deposit taking corporations except the central bank, (3) money market funds (MMFs), (4) non-MMF investment funds, (5) other financial intermediaries except insurance corporations and pension funds, (6) financial auxiliaries, (7) captive financial institutions and money lenders, (8) insurance corporations, and (9) pension funds.

For monetary statistics, the MFSMCG combines SNA subsectors (2) deposit taking corporations except central bank and (3) MMFs into one subsector called **Other Depository Corporations (ODCs)**. All other sub-sectors except the central bank and the deposit taking corporations are collectively known as the **Other Financial Corporations (OFCs)**.

Financial corporations (FCs), for monetary and financial statistics purposes consist of the central bank, ODCs and OFCs:

- **ODCs** are all financial corporations that issue liabilities or in other words accept deposits included in broad money. In most countries, commercial banks are the bulk of ODCs, but other institutions that issue monetary liabilities such as credit unions, savings banks, microfinance institutions and money market funds may also be included in ODCs.
- **Depository Corporations (DCs)** consist of the central bank and ODCs.

- **OFCs** include insurance corporations, pension funds, non-money market investment funds, and other financial intermediaries which incur liabilities that are excluded from broad money.

Non-financial corporations are corporations or quasi corporations whose principal activity is production of goods and nonfinancial services.

DCs are the only money issuers in most countries—thus the balance sheet data of central bank and ODCs sectoral balance sheets are especially important from the perspective of broad money calculation. DCs are thus considered as **broad money issuers**. On the other hand, OFCs, NFCs, households and governments are considered as **broad money holders**.

Source: *MFSMCG*.

More than 170 countries and territories follow the recommendations of the *MFSMCG* in compiling monetary and financial statistics and report them to the IMF using the standardized report forms (SRFs), a reporting framework prescribed in the *MFSMCG*. Specifically, the balance sheet information related to the central bank is reported in the SRF-1SR form, the aggregated balance sheet of the banking system or ODCs in the SRF-2SR form and the aggregated balance sheet of non-banks or the OFCs in the SRF-4SR (Box 4). Information in the SRFs are then used to compile and disseminate key monetary aggregates including monetary base and broad money (see Annex III). These data are reported to the IMF and disseminated as the monetary and financial statistics (MFS) through the [MFS database](#) (Annex IV).

Box 4

Standardized Report Forms (SRFs): 1SR, 2SR, 4SR, and 5SR

The MFSMCG prescribes standardized report forms (SRFs) to report monetary data to the IMF, thus providing countries with a tool to compile and report harmonized data for the central bank, ODCs, and OFCs. There are three SRFs corresponding to the three sub-sectors of the financial corporations sectors:

- **SRF-1SR:** sectoral balance sheet of the central bank
- **SRF-2SR:** sectoral balance sheet of ODCs
- **SRF-4SR:** sectoral balance sheet OFCs

These SRFs use a harmonized accounting presentation of assets and liabilities (stocks only) of the FCs with primary breakdowns by financial instrument (presented in order of their relative liquidity, including nonfinancial assets), then disaggregated by currency of denomination (domestic and foreign), and finally by counterpart sector (corresponding to the main sectors of the 2008 SNA).

Based on the SRFs, analytical surveys or presentations are prepared by reorganizing the sectoral balance sheet data to present the intermediation role of the relevant sector—the two important analytical surveys for monetary policy purposes are the central bank survey (CBS) and the depository corporations survey (DCS) (see Annex III).

In the CBS, the liability side is structured to show the components of **monetary base** and the asset side focusses on the financing extended to the non-residents and the various domestic sectors.

Similarly, in the DCS the liability side shows the components of **broad money** and the asset side shows the DC's claims on non-residents and the domestic sectors, thus providing a link between broad money supply and the net foreign assets and net domestic assets.

In addition to the SRFs-1SR, 2SR and 4SR, the **SRF-5SR**, based on the DCS, contains additional line items for components of broad money by institutional units other than FCs. In SRF-5SR, the data reporting is not standardized across countries, but the MFSMCG provides guidance on determining financial instruments to be included in money aggregates in accordance with the structure and other features of the financial system. These include currency in circulation issued by the central government, holdings of foreign currency that is widely accepted as a medium of exchange in domestic sectors, and electronic deposits issued by other nonfinancial corporations.

This section examines the statistical treatment of mobile money in computing monetary statistics and how mobile money impacts the measurement of money based on the *MFSMCG* framework and the data collected in the SRFs.

A. Treatment of Mobile Money in Monetary Statistics

According to the *MFSMCG*, money has the following properties: a medium of exchange, unit of account and store of value.¹⁴ The most common and widely used measure of money aggregates is "broad money." The *MFSMCG* describes broad money as the sum of all liquid financial instruments held by households, businesses etc. that are widely accepted in an economy as a medium of exchange, and/or that can be converted into a medium of exchange at short notice at, or close to, their full nominal value. Currency in circulation and transferable deposits¹⁵—the most liquid financial instruments—meet the definition of broad money. Nontransferable deposits such as savings deposits, sight deposits, and fixed deposits of short-term maturity are also included in broad money as they are redeemable at full value upon request without penalty and fee and satisfy the features of money, i.e., high degrees of liquidity and stability as a store of nominal value.^{16,17} Put another way, broad money consists of the liquid liabilities of the central bank and the banking system (collectively known as the Depository Corporations (DCs) sector in the *MFSMCG*—see Box 3) to the rest of the domestic economy, except the central government.

A key question in this paper is whether mobile money is part of broad money according to the *MFSMCG*. The treatment of mobile money in monetary statistics is addressed under the broad category of electronic money (e-money). E-money is defined as a payment instrument whereby monetary value is electronically stored on a physical device or remotely at a server which represents a claim on the issuer (*MFSMCG* 4.38).

¹⁴ See *MFSMCG* 6.8-6.10.

¹⁵ Transferable deposits comprise all deposits that are (1) exchangeable for banknotes and coins on demand at par and without penalty or restriction; and (2) directly usable for making payments to third parties by check, draft etc. (*MFSMCG* 4.30). Non-transferable deposits comprise all claims, other than transferable deposits, that are represented by evidence of deposit including fixed-term deposits, sight deposits, etc. (*MFSMCG* 4.43).

¹⁶ See *MFSMCG* 6.12.

¹⁷ Money-market fund (MMF) shares held by the money-holding sectors and some short-term debt securities may also be included in broad money (*MFSMCG* 6.47-6.48).

Examples of e-money include prepaid cards, mobile wallets or web-based e money (such as PayPal, if monetary value is electronically stored), and mobile money (IMF, 2018; MFSMCG; Kireyev, 2017).¹⁸

In the *MFSMCG*, e-money is classified as deposits rather than currency.¹⁹ As e-money including mobile money can be used for direct payments to third parties, it qualifies as a transferable deposit. Since transferrable deposits are typically included in broad money, mobile money is included in broad money according to the *MFSMCG*.

B. Compilation of Mobile Money Data for Monetary Statistics

While the guidance is clear that mobile money liabilities are to be treated as transferrable deposits and included in broad money, this may not directly translate into how mobile money data are collected for the purposes of monetary statistics in practice. The treatment of mobile money in monetary data compilation varies depending on the mobile money business model and/or the regulatory framework adopted in the country. This subsection elaborates these details.

MNO-led Model

The MNO-led model of mobile money consists of MNOs or entities that partner with MNOs to offer mobile money services. As noted earlier, all countries around the world where mobile money services exist currently have regulations which require MNOs to keep 100 percent of their mobile money liabilities in liquid assets—which in most cases are in the form of deposits at the central bank and other regulated financial institutions and in limited cases, in the form of investments in sovereign securities or other permitted assets²⁰ (GSMA, 2020b). The MFS treatment of mobile money in each of these cases is discussed below.

Case 1: Mobile money liabilities of MNOs held as deposits at a regulated financial institution

In countries where MNOs are required to keep their mobile money liabilities in the form of deposits at regulated deposit-taking financial institutions like banks, accounting for mobile money in monetary statistics may be relatively straightforward. Since the MNOs by law are required to maintain deposits with banks equivalent to the amount of mobile money issuance, the data on outstanding mobile money balances in the country are reflected on the balance sheet of banks (or ODCs using the *MFSMCG* terminology). These deposits by design are highly liquid so that the MNOs can make available these funds to the mobile money account holders on demand. Thus, these are recorded under transferrable deposits on the bank's balance sheet.

In terms of MFS data collected by the IMF, the aggregated balance sheet of the banking system including all assets and liabilities organized by instruments are reported in the SRF-2SR. Since all the liquid liabilities of the banks need to be included in the calculation of broad money in the country, the transferrable deposits

¹⁸ Many digital instruments or services used for payments do not qualify as e-money. Bitcoins are not e money (they are classified as nonfinancial assets—see IMF, 2018), and neither are credit and debit cards, as no monetary value is stored on them. Store cards are also not e-money as their use is limited to only the issuing stores.

¹⁹ According to *MFSMSG*, currency consists of notes and coins that are of fixed nominal values and are issued or authorized by central banks or governments while deposits are nonnegotiable contracts that represent the placement of funds available for later withdrawals (*MFSMCG* 4.25 and 4.29). E-money falls under deposits and specifically under transferrable deposits.

²⁰ Depending on the jurisdiction, other permitted assets may include debt securities guaranteed by the federal government or central bank or issued by regional financial institutions.

of MNOs at the banks—which are equivalent to the value of their mobile money issuance—are also accounted for in broad money.

In a few countries where MNOs are required to deposit customer funds at the central bank, equal to the total amount of mobile-money liabilities, the process of accounting for mobile money in the calculation of broad money is quite similar. The demand deposits of MNOs at the central bank reflecting their mobile money customer funds become part of the liabilities of the central bank balance sheet. In the MFS, the balance sheet of the central bank is reported to the IMF in the SRF-1SR. Given the MNO's demand deposits at the central bank are highly liquid and transferable, they are also included in the calculation of broad money.

In countries where MNOs are given the option to keep their mobile money customer funds received from customers either at the central bank or at the ODCs, the treatment is no different. These deposits are reflected in part in the central bank liabilities and in part in the banking system liabilities which ultimately feeds into the calculation of broad money. Care must be taken by compilers to avoid duplication—i.e., the mobile money liabilities of MNOs are not separately added in broad money calculation as they have already been accounted for.

The challenge, however, arises when the deposits maintained by the MNOs in the central bank or at the banks are restricted in nature—i.e., the deposits that cannot be accessed by the depositor (MNOs) until the appropriate conditions and obligations have been fulfilled, for example the escrow account can be used by the MNOs only to meet mobile money customer demand.²¹ In those cases, these deposits are considered as “restricted deposits” (a type of nontransferable deposits in the MFSMCG) and excluded from the calculation of broad money. In other words, the amount of mobile money liabilities in the country will not be included in broad money automatically. In these cases, compilers need to ensure that mobile money liabilities of these MNOs are covered for the purposes of MFS compilation separately so that they can be included in the computation of broad money. How this can be done is discussed in detail in Box 5.

In many countries, MNOs earn interest on their deposits (corresponding to the customer balances) kept with commercial banks. There has been an ongoing debate of whether MNOs should pass on the interest to the customers or not. The practice has varied country to country depending on the regulations (Suri et al., 2021). The MNO in Tanzania, Tigo Pesa was the first to pass on the interest to the customers (McKay, 2016). Bank of Ghana also required the MNOs to pass on at least 80 percent of their interest earned to the customers (Bank of Ghana, 2015). In Kenya, interest on MNOs' trust accounts are applied to corporate social responsibility (CSR) activities (Ahmad et al., 2020).

Irrespective, the statistical treatment of mobile money in MFS does not change. The MFSMCG requires that the accrued interest should be included in the outstanding amounts of the underlying financial assets or liabilities. In other words, given that the deposits of MNOs with banks (equivalent to customer funds) are in the banking system and recorded in monetary statistics, the interest amounts are also already included.

Case 2: Mobile money liabilities of MNOs invested in sovereign securities or other permitted assets

In a few select cases such as Malaysia, Philippines, Bolivia or the WAEMU member countries, MNOs have the additional option of investing in sovereign securities or other permitted assets in order to fulfil the requirement of maintaining 100 percent of their mobile-money liabilities in liquid assets (see Table 3). In such cases, the process of accounting for mobile money in compilation of broad money is more involved. While the proportion of mobile money liabilities of MNOs maintained as deposits at banks or at the central

²¹ Note that MNOs are usually required to segregate customer funds and keep them in trust or custodial accounts to safeguard them and prevent comingling with the companies' funds. The restrictions, if any, are to the customer funds only and not to the other deposits made by the MNOs at the banks.

bank are already included in monetary data reported to the IMF (in the balance sheet data of the banks and central bank), the proportion invested in securities are not. In such cases, where the regulations are layered, compilers may consider collecting balance sheet information of the MNOs and including them in the aggregated balance sheet of the ODCs (see Box 5 for details).

Furthermore, in case the regulations requiring MNOs to maintain their mobile money liabilities in liquid assets are not fully implemented or enforced, these countries are encouraged to collect data from MNOs to be included in the broad money calculation as detailed in Box 5. In other words, compilers of monetary statistics will need to review the prevailing situation in the country and decide how mobile money are to be accounted for in compilation of monetary statistics and make necessary adjustments in the calculation of broad money with additional information which may be separately collected from MNOs.

Box 5

Additional Treatments of Mobile Money in MFS Reporting

In some countries, the regulatory framework may be such that mobile money liabilities of MNOs are not reflected in the ODC balance sheet or the central bank balance sheet. In those cases, compilers need to take additional steps to account for the mobile money liabilities in the country by collecting balance sheet data from the MNOs directly. There are two ways to account for this information in the monetary data reported in SRFs.

Assess if MNOs can be considered ODCs for MFS and include their balance sheet in the sectoral balance sheet (SRF-2SR)

MNOs are usually telecom companies in the business of offering telecom and internet services which for the purposes of macroeconomic statistics are classified in the non-financial corporations' sectors. For monetary statistics, however, this may not always be the case. The sectoring of institutional units is based on the economic objectives, functions, and behavior. More generally, the key to classifying a unit in macroeconomic statistics is not its legal status but rather the economic nature of the entity.²²

The *MFSMCG* indeed suggests that electronic money institution should be classified as an ODC when they are in the business of financial intermediation and accept deposits which are included in broad money (Para 3.137 *MFSMCG*). MNOs which offer mobile money services thus can be treated as ODCs for MFS purposes and their balance sheet included in the sectoral balance sheets for ODCs.

Most jurisdictions require the mobile money issuers to be established as a separate legal entity from the MNO. Their balance sheet which consists of their mobile money activities alone must be included while compiling the SRF-2SR.

If the MNOs do not organize the business of mobile money as a separate unit but one of the activities is issuing e-money amongst other things, then their treatment as ODCs for monetary statistical purposes depends on whether their mobile money business can be treated as a quasi-corporation. A "quasi-corporation" in the *MFSMCG* terminology is defined as an unincorporated enterprise that functions in almost all respects as if it were incorporated. For purposes of sectoring, they are treated as institutional units (corporations) separate from the units that own them. In practice, for a quasi-corporation to exist, it must have a complete set of accounts including a balance sheet of assets and liabilities or can produce such a set of accounts as needed; its assets and liabilities must be separate from its owners; and are self-contained and independent. If the mobile money business of the MNO can be considered a quasi-corporation for statistical purposes, then it is to be classified as part of the ODCs.

²² *MFSMCG* 3.13.

In sum, when an MNO is classified as ODC, then balance sheet are included in the aggregated balance sheet for the ODC sector (SRF-2SR) like in the case of commercial banks and credit unions, etc. The customer funds are then treated as transferrable deposits which will ultimately feed into the calculation of broad money.

Report mobile money as additional components of broad money (SRF-5SR):

While the *MFSMCG* provides some discussion on the characteristics of the financial instruments that should be included as part of broad money, the definition is only intended to help monetary statistics compilers determine the scope of broad money, taking into account the structure and other features of the financial system in their own economies. Thus, in the SRFs, there is an additional report form (SRF-5SR) which provides the option to countries to include components of broad money that are not captured in the report forms for the central bank (SRF-1SR), the other depository corporations (SRF-2SR), and the other financial corporation (SRF-4SR) (see Box 4).

The SRF-5SR form has the option of including deposits of nonfinancial corporations like MNOs in the broad money calculation. Therefore, in countries where MNO deposits at the central bank or at banks are restricted in nature and are not captured by a default broad money calculation, the mobile money customer funds with MNOs may be included in the 5SR form to be reflected as part of broad money.

Note that these additional data collection methods are needed only when mobile money liabilities are not counted in the banking system.

Bank-led Model

In the MFS, the treatment of bank-led models of mobile money is much simpler. As discussed above, in a typical bank-led model, the practice is for banks to pool and store funds from its mobile money customers in a single account rather than opening an individual deposit account for each customer (Grossman, 2016).

For such bank-led mobile money service providers, as long as the mobile money customer balances are accounted for in the banks' books either as customer deposits or e-money accounts, these are classified as transferrable deposits with the banks which will ultimately feed into the calculation of broad money under the framework of the *MFSMCG*.

In sum, mobile money appears to be largely captured in the calculation of broad money in most countries that have adopted the *MFSMCG* guidance (Chart 6). This is mainly because most jurisdictions around the world have regulations that require the issuers to mirror the value of the outstanding mobile money in an account at a regulated financial institution or at the central bank. This means that mobile money balances are already being accounted in the existing monetary and financial statistics data reported in the SRFs to the IMF (IMF, 2018).

Treatment of Mobile Money in MFS

Chart 6

| | A: Regulation | B: Impact of the regulation on data | C: Additional treatments needed? | D: How will mobile money liabilities be reflected in broad money calculation? |
|-----------------------|---|---|--|--|
| MNO-led model | Mobile money liabilities of MNOs held as deposits at a central bank or commercial banks. | Data collected from the central bank and commercial banks on the amount of transferrable deposits include data on mobile money liabilities of MNOs. | No separate data collection needed for mobile money liabilities. | Transferrable deposits with the central bank and commercial banks are part of broad money calculation. |
| | Mobile money liabilities of MNOs held as bank deposits or invested in sovereign securities or other permitted assets. | Data collected from the central bank and commercial banks on the amount of transferrable deposits may not reflect the mobile money liabilities. | Assess if MNOs can be considered ODCs and include MNOs in the balance sheet of the ODC sector. Mobile money liabilities will be considered transferrable deposits. | Liquid liabilities of the ODCs are part of broad money calculation. |
| | Regulations not yet implemented. | | <u>OR</u> Collect data on mobile money liabilities of MNOs directly and include in the monetary aggregates (SRF 5SR). | The SRF-5SR provides the options to include components of broad money that are not captured in the SRFs for central bank, ODCs and OFCs. |
| Bank-led model | Mobile money liabilities held as deposits at a bank by definition and subject to prudential regulations. | Data collected from commercial banks on the amount of transferrable deposits include data on mobile money liabilities of MNOs. | No separate data collection needed for mobile money liabilities. | Transferrable deposits commercial banks are part of broad money calculation. |

Source: IMF staff.

Note: This figure summarizes the treatment of mobile money in the MFS. Column D shows how mobile money liabilities are reflected in the calculation of broad money assuming that the treatments described in column C are implemented.

C. The impact of mobile money on broad money composition

More generally, mobile money has the potential to change the composition of broad money without impacting its size or volume. As mobile money is increasingly used, currency in circulation will likely decrease while deposits with the banking system will increase since mobile money balances outstanding with the mobile money service providers will in most cases be accounted for in monetary statistics as part of transferrable deposits with the banks.

Kipkemboi and Bahia (2019) have studied whether mobile money has led to a decline to the use of cash, by examining the trend of the ratio of the currency outside the banking sector to broad money in select

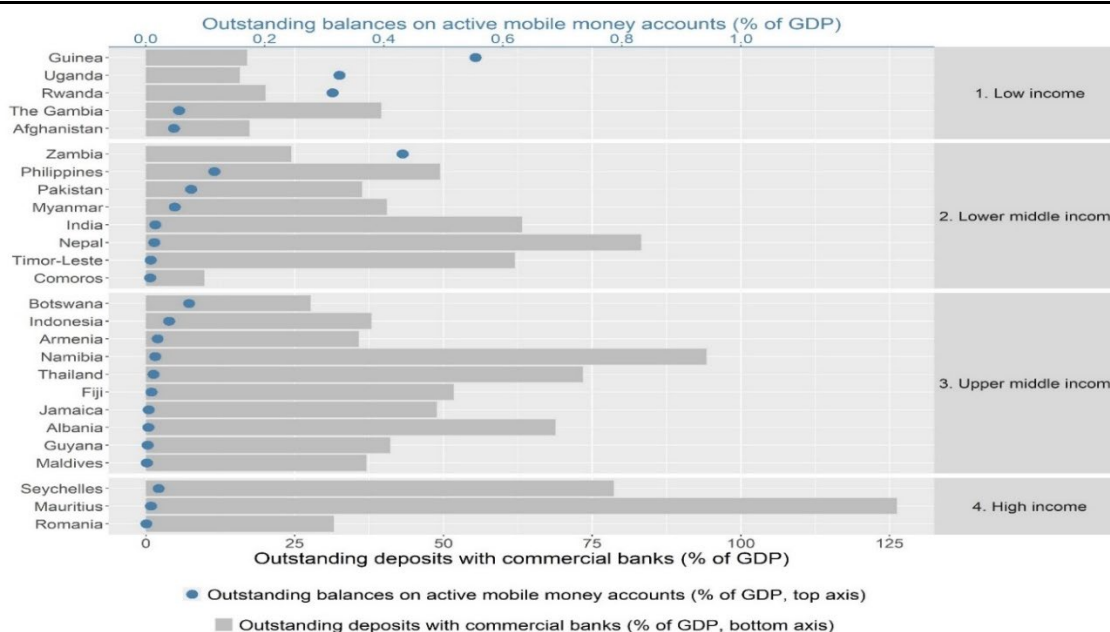
countries of Kenya, Uganda, Ghana and Rwanda. They find a negative growth of the ratio, suggesting that the use of cash has declined while bank deposits have increased.²³

However, mobile money balances are still small relative to the size of transferable deposits and currency in most countries to see this trend prominently across countries. For example, the outstanding amount of mobile money balance is a mere fraction of that of deposits at commercial banks (Chart 7).

While mobile money is only a small part of broad money currently, as mobile money gains more popularity, its impact on calculation of monetary aggregates may become more pronounced. It is thus crucial that statistical compilers ensure that mobile money is being accounted for in the calculation of monetary aggregates. In this regard, there may be some merit in distinguishing between deposits of customer funds (from mobile money) in banks and other types of bank deposits included in broad money in the future. More generally, given mobile money has implications for various policy aspects ranging from its impact on financial inclusion to monetary aggregates, it is important to continue to monitor mobile money trends and developments.

Mobile Money Account Balances and Bank Deposits

Chart 7



Source: IMF staff calculations and IMF Financial Access Survey.

Note: This figure presents the outstanding deposits with commercial banks (gray bars on the bottom x-axis) and outstanding balance on active mobile money accounts (blue dots on the top x-axis) as percentage of GDP in 2019 or the latest year available for select countries.

²³ More broadly, the cash component of broad money ("cash in circulation") is often used as an indicator to track cash usage in an economy. Given the growing interest in digital money including central bank digital currencies (CBDCs), understanding the costs and drivers of demand for cash is becoming important (see for example, Khiaonarong and Humphrey, 2019). Comprehensive data on broad money components such as cash usage and e-payments are key in this regard (Bech et al., 2018).

5. Conclusion

This paper analyzed recent developments of mobile money from a statistical perspective, examining measurement and data collection issues. Specifically, the paper reviewed stylized facts about mobile money using available data and also clarified the treatment of mobile money in monetary statistics drawing on the methodological framework provided by the *MFSMCG*.

The FAS data provide useful insights into how mobile money has helped unbanked or underbanked populations to gain access to financial services in many developing economies. The data show that mobile money is a much more accessible means of using basic financial services than traditional banking as there are more mobile money agents than ATMs and bank branches combined in these economies. This increased access can be also seen in the number of mobile money accounts—there are more registered mobile money accounts than traditional bank accounts in many low- and middle-income economies. A similar trend also holds in the usage of mobile money, as evidenced by the increased transaction values and volumes.

Mobile money service providers have recently started to expand their suite of financial services for customers, including savings and loans products via mobile money accounts. However, relatively little is known about the degree of the usage of such services. The analysis based on transaction level data from Kenya confirms that mobile money customers are indeed starting to use new financial services, with about half of the mobile money accounts in the study being used for such transactions. These findings also point to the importance of proper measurement of mobile money activity, especially for the central bank policymaking, as mobile money continues to expand into other streams like credit.

On measurement issues related to mobile money, the paper clarified the treatment of mobile money in monetary statistics and its implications for calculation of monetary aggregates to address the question of whether mobile money is counted as part of broad money. Answering this question warrants an understanding of the mobile money value chain, the ensuing business models, and the regulations in place to safeguard customer funds. Mobile money business models can be either MNO-led or bank-led, depending on the type of the entity responsible for the issuance of mobile money. The type of business model then determines the regulations to be applied, which in turn have direct implications for the treatment of mobile money in monetary statistics.

The methodological guidance provided by the *MFSMCG* on the treatment of mobile money is relatively clear-cut—mobile money liabilities need to be included as part of broad money in monetary statistics. However, how mobile money affects the measurement of monetary aggregates depends on the mobile money business model and/or the regulatory framework. For an MNO-led model, mobile money is reflected in broad money as long as regulations require mobile money liabilities are secured in the form of deposits at a regulated financial institution, which is the case in many countries. If regulations allow mobile money liabilities to be invested in sovereign securities or other permitted assets, more involved steps may be needed, including collecting additional information from the MNO. For a bank-led model, since mobile money balances are already recorded as part of the bank balance sheet, they feed directly into the calculation of broad money.

While this paper focused on mobile money, there are other fintech products and services whose data needs and potential measurement issues are yet to be fully explored. These financial innovations include mobile payment applications known as mobile wallets and neobanks (online only banks), which offer internet-based financial services without physical presence. More recently, policymakers are also increasingly focusing on the potential digital form of money such as stablecoins and central bank digital currencies (CBDCs).²⁴ These are important topics that require further research, and understanding the statistical treatment of these products in monetary statistics is fundamental in monitoring their developments as well as supporting policy analysis.

²⁴ See IMF (2019b) which briefly touches upon CBDCs. More work is under way on the statistical treatment of digital money including CBDCs.

References

- Adrian, T. and T. Mancini-Griffoli. 2019. "The Rise of Digital Money". IMF Fintech Notes No 19/01, International Monetary Fund, Washington, DC.
- Ahmad, A. H., C. Green and F. Jiang. 2020. "Mobile Money, Financial Inclusion and Development: A Review with Reference to African Experience." *Journal of Economic Surveys* 34 (4): 753-792.
- Aron, J., J. Muellbauer and R. Sebudde. 2015. "Inflation forecasting Models for Uganda: is Mobile Money relevant?" Centre for Policy Research Discussion Paper Series 10739: 1-66, London.
- Asamblea Legislativa de El Salvador. 2015. "Reformas a la ley para facilitar la inclusión financiera, Decreto No. 464." San Salvador.
- Autoridad de Supervisión del Sistema Financiero de Bolivia. 2018. "Modificaciones al reglamento para empresas de pago móvil, al reglamento de fidecoismo, y al manual de cuentas para entidades financieras, Circular ASFI 548." La Paz.
- Bangko Sentral ng Pilipinas. 2009. "Guidelines governing the Issuance of Electronic Money (e-money) and the Operations of Electronic Money Issuers (EMI), Circular No. 649." Manila.
- Bank of Ghana. 2015. "Guidelines for E-Money Issuers in Ghana." Accra.
- Bank Negara Malaysia. 2003. "Guideline on Electronic Money (E-Money), Act 627." Kuala Lumpur.
- Bazarbash, M., H. Carcel-Villanova, E. Chhabra, Y. Fan, N. Griffin, J. Moeller and K. Shirono. 2020. "Mobile Money in the COVID-19 Pandemic." IMF Special Series on COVID-19, International Monetary Fund, Washington, DC.
- Bech M. L., U. Faruqui, F. Ougaard, and C. Picillo. 2018. "Payments are a-changin' but Cash still rules," *BIS Quarterly Review* (March).
- BCEAO. 2015. "Instruction N°008-05-2015 regissant les Conditions et Modalités d'Exercice des Activités des Emetteurs de Monnaie Électronique dans les États Membres de l'Union Monétaire Ouest Africaine (UMOA)." Dakar.
- Bill and Melinda Gates Foundation. 2015. "Assessing Risk in Digital Payments." Special Report, Financial Services for the Poor, Seattle.
- Blumenstock, J. E., N. Eagle and M. Fafchamps. 2016. "Airtime Transfers and Mobile Communications: Evidence in the Aftermath of Natural Disasters." *Journal of Development Economics* 120: 157-181.
- Central Bank of Eswatini. 2019. "Practice Note for Mobile Money Service Providers." Mbabane.
- Central Bank of Seychelles. 2014. "Mobile payment services – Approval to commence Pilot Program." Victoria.
- Claessens, S., J. Frost, G. Turner and F. Zhu. 2018. "Fintech Credit Markets around the World: Size, Drivers and Policy Issues." *BIS Quarterly Review* (September).

Cornelli, G., J. Frost, L. Gambacorta, R. Rau, R. Wardrop and T. Ziegler. 2020. "Fintech and Big tech Credit: a new Database." BIS Working Papers No 887, Bank for International Settlements, Basel.

Dupas, P., D. Karlan, J. Robinson and D. Ubfal. 2018. "Banking the Unbanked? Evidence from Three Countries." *American Economic Journal: Applied Economics* 10 (2): 257-297.

Espinosa-Vega, M., K. Shirono, H. Carcel-Villanova, E. Chhabra, B. Das, and Y. Fan. 2020. "Measuring Financial Access: 10 Years of the IMF Financial Access Survey." IMF Departmental Paper No. 20/08, International Monetary Fund, Washington, DC.

Global System for Mobile Communications (GSMA). 2018. "Mobile Money Policy and Regulatory Handbook." London.

Global System for Mobile Communications (GSMA). 2020a. "State of the Industry Report on Mobile Money 2019." London.

Global System for Mobile Communications (GSMA). 2020b. "The Mobile Money Regulatory Index 2019." London.

Greenacre, J. 2018. "Regulating Mobile Money: a functional Approach." Pathways for Prosperity Commission Background Paper Series no. 4, Oxford.

Grossman, J. 2016. "Safeguarding Mobile Money: How Providers and Regulators can ensure that Customer Funds are protected." GSMA Report, Global System for Mobile Communications, London.

Hammer, C. L., D. C. Kostroch, G. Quirós, and STA Internal Group. 2017. "Big Data: Potential, Challenges, and Statistical Implications." IMF Staff Discussion Note 17/06, International Monetary Fund, Washington, DC.

International Monetary Fund (IMF). 2018. "Measuring Digital Economy." IMF Staff Report, Washington, DC.

International Monetary Fund (IMF). 2019a. "Mobile Money Note 2019." Washington, DC.

International Monetary Fund (IMF). 2019b. "Treatment of Crypto Assets in Macroeconomic Statistics." Washington, DC.

Izaguirre, J. C., D. Dias and M. Kerse. 2019. "Deposit Insurance Treatment of E-money – an Analysis of Policy Choices." CGAP Technical Note, Washington, DC.

Jack, W. and T. Suri. 2011. "Mobile Money: The Economics of M-Pesa." NBER Working Paper No. 16721, National Bureau of Economic Research, Cambridge, MA.

Jack, W. and T. Suri. 2014. "Risk Sharing and Transactions Costs: Evidence from Kenya's Mobile Money Revolution." *American Economic Review* 104 (1): 183–223.

Kerse, M. and S. Staschen. 2018. "Safeguarding Rules for Customer Funds held by EMIs." CGAP Technical Note, Washington, DC.

Khiaonarong, T. and D. Humphrey. 2019. "Cash Use across Countries and the Demand for Central Bank Digital Currency," IMF Working Paper No 19/46, International Monetary Fund, Washington, DC.

Kipkemboi, K. and K. Bahia. 2019. "The Impact of Mobile Money on Monetary and Financial Stability in Sub-Saharan Africa." GSMA Report, Global System for Mobile Communications, London.

Kireyev, A. P. 2017. "The Macroeconomics of De-cashing." IMF Working Paper No 17/71, International Monetary Fund, Washington, DC.

Mawejje, J. and P. Lakuma. 2019. "Macroeconomic Effects of Mobile Money: Evidence from Uganda." *Financial Innovation* 5 (23).

McKay, C. 2016. "Interest Payments on Mobile Wallets: Bank of Tanzania's Approach." CGAP Blog Series: Emerging Regulatory Enablers in Digital Financial Services, Washington, DC.

Monetary and Financial Statistics Manual and Compilation Guide. 2016. International Monetary Fund, Washington, DC.

Munyegera, G. K. and T. Matsumoto. 2016. "Mobile Money, Remittances, and Household Welfare: Panel Evidence from Rural Uganda." *World Development* 79: 127–137.

OECD. 2020. "Digital Economy Outlook 2020". Organisation for Economic Co-operation and Development, Paris.

Pelletier, A., S. Khavul and S. Estrin. 2019. "Regulating Mobile Money, what's at Stake." LSE Business Review, London School of Economics, London.

Pelletier, A., S. Khavul and S. Estrin. 2020. "Innovations in Emerging Markets: the Case of Mobile Money." *Industrial and Corporate Change* 29 (2): 395-421.

Reserve Bank of India. 2014. "Guidelines for Licensing of Payments Banks". Mumbai.

Riley, E. 2018. "Mobile Money and Risk Sharing against Aggregate Shocks." *Journal of Development Economics* 135: 43–58.

Suri, T. and W. Jack. 2016. "The Long-Run Poverty and Gender Impacts of Mobile Money." *Science* 354 (6317): 1288–1292.

Suri, T., J. Aker, C. Batista, M. Callen, T. Ghani, W. Jack, L. Klapper, E. Riley, S. Schaner, and S. Sukhtankar. 2021. "Mobile Money." *VoxDevLit* 2 (1).

Wasunna, N. and J. Frydrych. 2017. "Person-to-Government (P2G) Payment Digitization: Lessons from Kenya." GSMA Report, Global System for Mobile Communications, London.

Annex I. Data Sources for Mobile Money

While the IMF's Financial Access Survey (FAS) provides supply-side annual country level data on mobile money, there are also other databases which cover a range of aspects including qualitative information on policies and higher frequency data at the individual country level. This annex summarizes such select data sources.

GSMA: The GSMA's annual Global Adoption Survey collects data from mobile money service providers and presents a set of mobile money metrics for six geographical regions, covering up to 90 countries. The GSMA does not however release data at an individual country level. This database includes annual data by region between 2001-19 on the number of mobile money services worldwide. It also contains quarterly data by region on the number of registered and active mobile money agents, the number of registered and active mobile money accounts, and the volume and value (in USD) of transactions processed by the industry across different products.

Country level data: Some have started to publish higher frequency country level data on mobile money. For example, the [Bank of Zambia](#) reports monthly data on mobile payment values and volumes since 2012. The Central Banks of [Kenya](#), [Uganda](#) and [Bangladesh](#) also publish monthly data on a variety of mobile money series such as the number of active mobile money agents and the number of mobile money accounts. The [Bank of Ghana](#) also provides information on these mobile money features on an annual frequency.

Demand-side data: The World Bank Global Findex is a demand-side data source on financial inclusion. It was first released in 2012 and is updated triennially, covering 140 economies with data collected through sample surveys of roughly 1,000 people in each country. It covers 12 mobile money series, including gender disaggregated data on mobile money accounts.

Qualitative data: Some databases collect qualitative information on mobile money regulations and policies, providing useful information to supplement quantitative data in monitoring and understanding mobile money trends and developments:

- The [GSMA Mobile Money Regulatory Index](#) database provides information on the mobile money regulatory framework to help facilitate mobile money adoption for 90 countries where mobile money services are available. It contains 26 indicators to produce scores for six dimensions of mobile money regulation, which are aggregated into the overall index score. The six dimensions covered are (i) authorization; (ii) consumer protection; (iii) know your customer (KYC); (iv) agent network; (v) transaction limits; and (vi) infrastructure and investment environment. The index and indicators are scored ranging from 0 to 100, with a higher score associated with a more enabling regulatory framework.
- The [GSMA Mobile Money Deployment Tracker](#) monitors the number of live mobile money services globally using both primary and secondary sources on a monthly basis. It contains information on products offered by mobile money service providers as well as their partners in mobile money service provision.

COVID-19 Policy Responses: The COVID-19 pandemic has created new data needs, including information on policies adopted in response to the pandemic. Digital financial services have been recognized as a useful tool in the pandemic context, and many countries have implemented policies to facilitate the use of digital financial services to support financial transactions while social distancing.

- The IMF's [Financial Access COVID-19 Policy Tracker](#) documents policy measures adopted in response to the pandemic across the globe, to support SME financing and usage of digital financial services, including mobile money. The information is collected from publicly available sources as well as feedback received from country authorities. The policy tracker categorizes policy responses aimed at promoting mobile money usage into three types: (i) temporary waiver of transaction fees; (ii) temporary increase of transaction and balance limits; and (iii) easing KYC onboarding requirements.²⁵
- The GSMA's [Mobile Money COVID-19 Regulatory Response Tracker](#) contains information on mobile money specific interventions implemented to support mobile money customers and service providers during the pandemic. It covers a broader set of policy measures, including cash transfer programs using mobile money.

²⁵ See Bazarbash et al. (2020) for the details on the COVID-19 response measures related to mobile money and associated risks.

Annex II. Analysis of Transaction-level M-Pesa Data

The FAS and GSMA data can offer insights into mobile money's development and trends at the country or regional level. However, due to lack of transaction-level data, little has been studied about individual mobile money usage patterns, which can provide more nuanced information about mobile money usage and help policymakers to design better targeted financial inclusion strategies. This annex reports the analysis of the transaction-level M-Pesa data which have been made publicly available by the Consultative Group to Assist the Poor (CGAP).

Data and Methodology

The CGAP has recently published transaction-level M-Pesa data for 418 low-income mobile money users living in Nairobi who gave consent to share anonymized mobile money transaction data and participated in a survey conducted by the Busara Center for Behavioral Economics. This open-source dataset contains the entire transaction history (147,632 transactions) of these 418 low-income mobile money users from July 2017 to August 2018.²⁶ While the data may not be nationally representative, it provides useful insights into the use of mobile money in Kenya, especially the usage patterns and transaction volumes among low-income users.

The dataset includes seven different variables, including mobile money account ID numbers, encrypted phone numbers, receipt numbers, transaction details, transaction time, types, and values. It also provides information on transaction's nature, mobile money agent's location, and a detailed description of the transaction for each transaction entry. Using each account ID number and the encrypted phone number, which are unique to each account holder, the transaction history for each of the account holders in the dataset were recreated. Topic modeling—a machine learning technique—was applied to categorize transactions into eight categories. Then entity recognition, keyword extraction, and random-sampled human coding were used to refine and validate the categorization.

Key Findings

The analysis found that 95 percent of the mobile money accounts were active during the sample period.²⁷ The average balance on mobile money accounts is USD12.8, close to 9 percent of monthly income in Kenya.²⁸ Transactions are categorized into the following eight categories (Table A.II.1):

- **Person-to-Person (P2P) transactions:** This category includes deposit and withdrawal of funds between users and the most popular function measured by transaction share (28 percent of all transactions). The average value of P2P transactions ranges from USD 5 to USD 7.

²⁶ The data are available at: <https://www.cgap.org/blog/how-do-kenyans-really-use-m-pesa>.

²⁷ Active mobile money accounts are defined as those used to conduct a mobile money transfer or cash-in/cash-out transaction over the past 90 days, in line with the FAS guidelines.

²⁸ Monthly income is approximated by GNI per capita per month. The GNI per capita in Kenya was USD 1,750 in 2019 according to the [World Bank](#).

- **Business-to-Person/Person-to-Business (B2P/P2B) transactions:** This covers transactions from and to business enterprises, including vendors, merchants, and companies, accounting for the second-highest share in total transaction. However, the degree of usage varies across subcategories. About 79 percent of the users in the sample have used mobile money to pay bills using M-Pesa while only 20 percent of these users received salary payments via M-Pesa. The average value of salary payments is higher than other categories, at about USD 24.
- **Airtime purchase:** This refers to the purchase made for phone calls, SMS, or cellular data usage. Airtime cannot be used as money to purchase other items or pay bills.²⁹
- **Deposit and withdrawal of funds:** This category is the most basic function of mobile money together with the P2P function, accounting for 18 percent of all transactions and used by almost all users in the sample. The average values of the deposit and withdrawal of funds are around USD 12.
- **Transaction fees:** These are charges collected by the telecommunication operators for mobile money transactions. Whereas all deposits are free, M-Pesa charges for sending and withdrawing money through a Safaricom agent, ATMs, and transferring cash to registered and non-registered M-Pesa accounts. Specific charges depend on the amount of withdrawal and transfers.³⁰ Nearly all account holders were once charged for transaction fees. The average transaction fees across all charges is about USD 0.3.
- **Financial Derivatives:** This category includes any additional financial services provided by the partnering commercial banks in Kenya. Commercial Bank of Africa (CBA) and KCB Bank Kenya Limited allow M-Pesa users to open deposit accounts via M-Pesa and transfer mobile money balance to earn interests. Not only the basic saving accounts, but both KCB and M-Shwari also provide micro-credit loan products, which allows M-Pesa subscribers to borrow money in times of need and to fund a project or enterprise. Borrowing via M-Pesa derivatives can help users build their credit history to facilitate future financing. Noticeably, around 40 percent of the users in the sample have utilized these basic savings accounts with the average value of M-Shwari/KCB M-Pesa deposits standing at USD17, slightly higher than the average value of regular deposits in M-Pesa. These findings show that M-Pesa users now have increased options to save and borrow funds using mobile money.
- **International Transfers:** This refers to international remittances received via an international money transfer service provider. Although the share of international transfers in all transactions is small at 0.02 percent, the average value of such transfers is USD 85—about 60 percent of average monthly income in Kenya—substantially higher than any other category.
- **Others:** This includes transactions not categorized into the seven categories aforementioned.

²⁹ <https://www.worldremittance.com/en/faq/mobile-money#13873>

³⁰ Details on M-Pesa rates are available at <https://www.safaricom.co.ke/personal/m-pesa/getting-started/m-pesa-rates>.

Mobile Money Transaction-Level Data

Table A.II.1

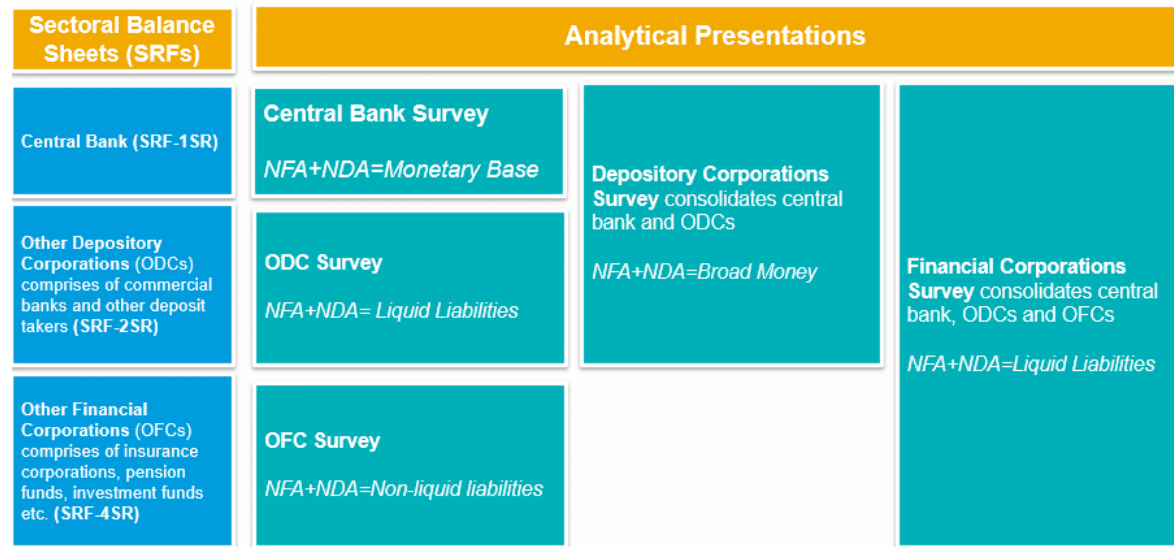
| Category | Category Share in all Transactions (percent) | Transaction Types | Number of Transactions | Share of all Transactions (percent) | Usage: Number of Accounts | Usage: Share of Total Accounts (percent) | Average Value (USD) |
|--|--|-------------------------|------------------------|-------------------------------------|---------------------------|--|---------------------|
| P2P Transactions | 28.07 | Customer transfer | 23501 | 15.92 | 401 | 95.93 | 5.11 |
| | | Funds received | 17932 | 12.15 | 411 | 98.33 | 8.71 |
| B2P/P2B Transactions | 22.56 | Pay Bill | 29320 | 19.86 | 327 | 78.23 | 1.33 |
| | | Other Payment | 3679 | 2.49 | 287 | 68.66 | 9.4 |
| | | Salary Payment | 312 | 0.21 | 85 | 20.33 | 24.36 |
| Airtime Purchase | 18.96 | Airtime purchase | 27992 | 18.96 | 410 | 98.09 | 0.29 |
| Deposit and Withdrawal of Funds | 18.02 | Customer withdrawal | 13622 | 9.23 | 416 | 99.52 | 13.24 |
| | | Deposit of funds | 12977 | 8.79 | 406 | 97.13 | 11.69 |
| Transaction Fees | 8.57 | Withdrawal charge | 12645 | 8.57 | 416 | 99.52 | 0.29 |
| Financial Derivatives | 3.44 | M-Shwari/KCB Withdraw | 2736 | 1.85 | 177 | 42.34 | 8.11 |
| | | M-Shwari/KCB Deposit | 1230 | 0.83 | 164 | 39.23 | 17.25 |
| | | M-Shwari/KCB Loan | 954 | 0.65 | 102 | 24.40 | 12.55 |
| | | M-Shwari Lock Deposit | 164 | 0.11 | 19 | 4.55 | 6.22 |
| Others | 0.36 | Others | 532 | 0.36 | 187 | 44.74 | 1.24 |
| International Transfers | 0.02 | International Transfers | 36 | 0.02 | 10 | 2.39 | 84.92 |

Source: IMF staff calculations from CGAP data.

Note: "Usage: Share of total accounts" shows the share of accounts used for a specific type of transactions in total accounts. For example, if it is 100 percent for a certain category of transactions, then it means that all accounts in the sample had at least one transaction under the category.

Annex III. SRFs and Compilation of Monetary Aggregates under the *MFSMCG*

The chart below complements the discussion in Box 4, summarizing the use of SRFs and the compilation of monetary aggregates under the *MFSMCG* framework.



Source: IMF staff and the *MFSMCG*.

NFA=Net Foreign Assets and NDA=Net Domestic Assets.

Annex IV. IMF's Monetary and Financial Statistics Database

The monetary and financial statistics (MFS) database contains in addition to the monetary aggregates, key credit aggregates like credit to the private sector by the banking system or the claims of the central bank on the central government. The visualization tools embedded in the database allow quick analysis of these credit aggregates for over 170 countries. An interactive map allows users to compare broad-money growth for countries across the globe over the years. Many of these countries are also using the SRFs as the platform to generate the monetary data disseminated through their national publications.

In fact, the MFS database has several presentations of monthly SRF based monetary data dating back to 2001 that provide a framework for analyzing the relationship between the FCs sector and other institutional sectors, including broad money, credit aggregates, and liquidity measures for example, with net domestic assets as a correspondent to broad money. These analytical presentations of monetary data are presented separately for the central bank, the ODCs, and the OFCs. The database also presents consolidated data for the depository corporations (also known as monetary survey) and the entire financial corporations' sector.

More than 40 countries have agreed to disseminate [detailed monetary data based on the SRFs](#) through the MFS database. These SRF data provide the asset and liability positions (and the corresponding flows) that are presented in a balance-sheet-like form by category of financial instrument, by currency (domestic and foreign), and by counterpart institutional sector.

Most countries compile the central bank and ODC data on a monthly basis and report it within one or two months after the end of the reference period. OFC data are sometimes submitted quarterly.

The MFS database also contains monetary data on non-SRF countries, namely those that have not adopted the MFSMCG methodology in compiling monetary data reported to the IMF.



STATISTICS

Evaluating Mobile Money Access and Use with Non-traditional Data Sources

**IFC-BANK OF ITALY WORKSHOP PART 2: “DATA SCIENCE IN
CENTRAL BANKING: APPLICATIONS AND TOOLS”**

FEBRUARY 14-17, 2022

Kazuko Shirono

Deputy Division Chief

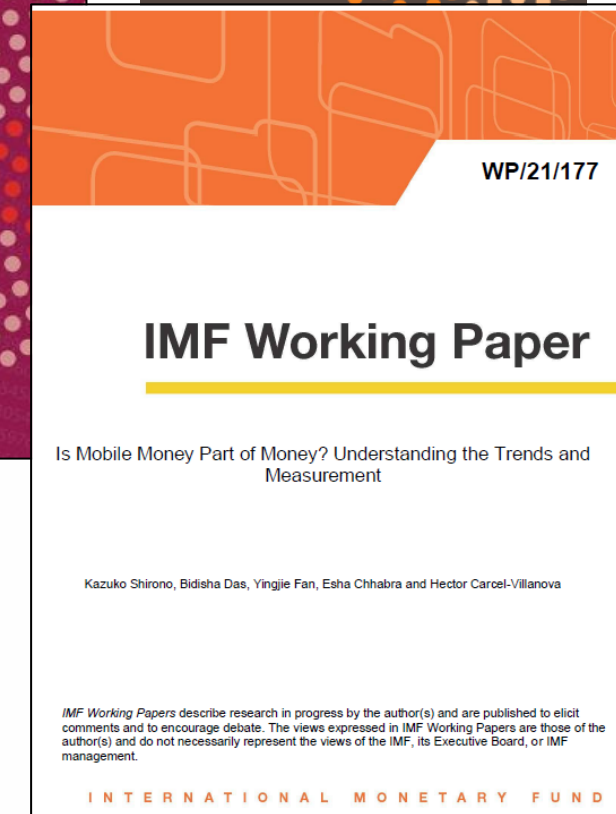
Financial Institutions Division, Statistics Department

International Monetary Fund

The views expressed herein are those of the author and should not be attributed to the IMF, its executive board or its management.

Context

- IMF's Financial Access Survey (FAS)
 - ▶ Supply-side data on access to and use of financial services
 - ▶ Includes data on mobile money
- Exploring non-traditional data sources
 1. Mobile money transaction level data
IMF Working Paper: “Is Mobile Money Part of Money? Understanding the Trends and Measurement”
 2. Geolocational data from Google Map Places API
“Evaluating Financial Access with Big Data Approach”
- *Two use cases of big data approach to examine mobile money access and use.*



What is mobile money?

Mobile money:

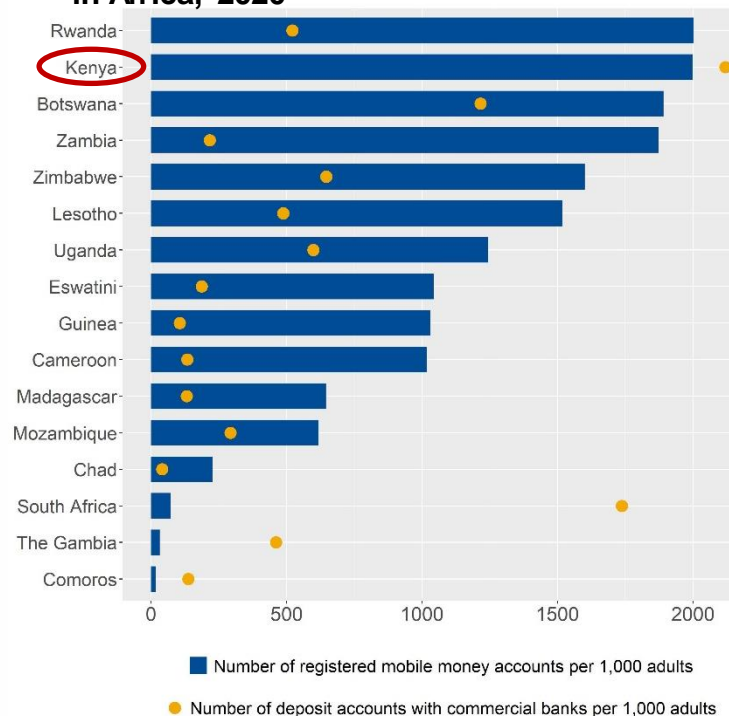
- Financial service offered typically by a mobile network operator, independent of the traditional banking network.
- The only prerequisite is a basic feature cell phone.
- Mobile money agents carrying out cash-in/out transactions.

E.g., Kenya's M-Pesa, operated by Safaricom

Why do we care:

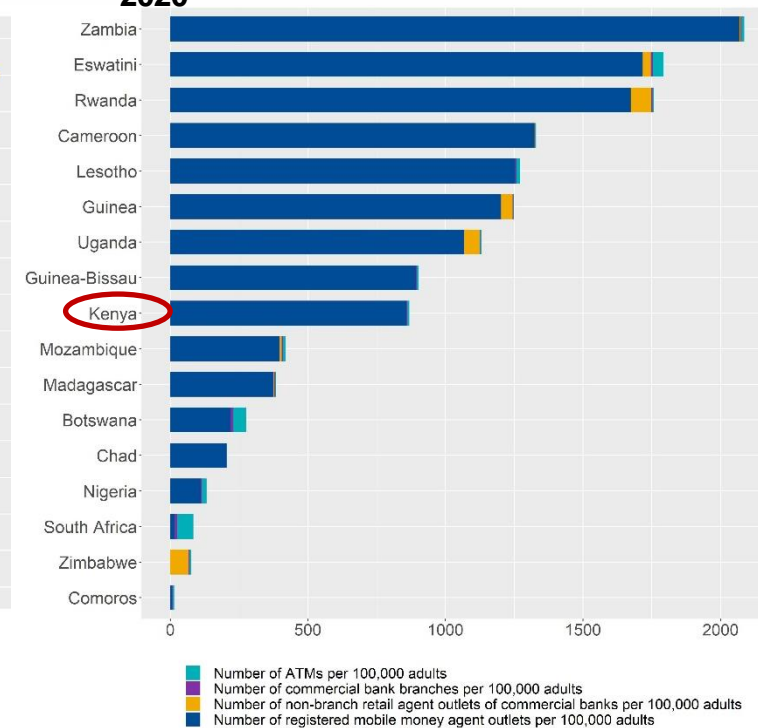
- Mobile money has been a game changer in many low- and middle-income countries.
 - Advancing financial inclusion

Figure 1: Mobile money and bank accounts in Africa, 2020



Source: Financial Access Survey

Figure 2: Financial access points in Africa, 2020



Source: Financial Access Survey

M-Pesa transaction level data

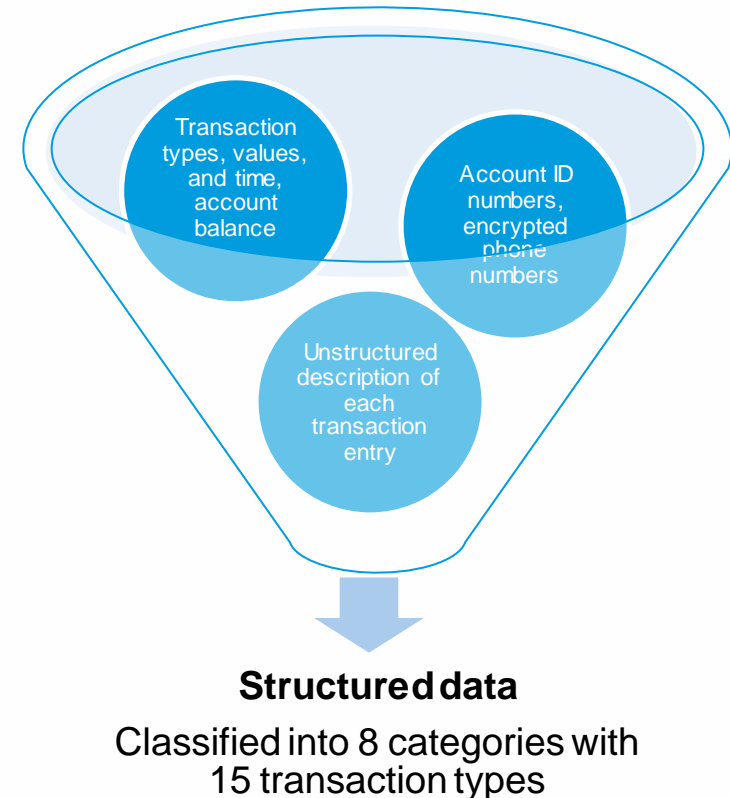
Data:

- Open-source data on M-Pesa transactions
- Anonymized 418 low-income M-Pesa users in Nairobi
- Contain 147,632 transactions during July 2017-August 2018
- Data collected by the Busara Center for Behavioral Economics and made publicly available by Consultative Group to Assist the Poor (CGAP)

➤ *Offer insight into usage patterns of M-Pesa*

Methodology:

- Topic modeling applied to categorize transactions into eight categories.
- Categorization refined and validated using entity recognition and keyword extraction.



M-Pesa transaction level data: Key findings

- **Average account balance:** USD13, about 9 percent of monthly income in Kenya
- **P2P:** Most used transaction category
- **B2P/P2B:** Bill pay was the most used transaction type (20 percent of transactions). Salary payment used by 20 percent of the accounts.
- **Financial derivatives:** 40 percent of accounts deposited in M-Shwari/KCB with average value of deposit being USD 17 (12 percent of monthly income).
- **International remittances:** Higher value than other categories, USD 85.

| Category (Types of transactions) | Share of total transactions (%) | Usage share of total accounts (%) | Average value (USD) |
|--|---------------------------------|-----------------------------------|---------------------|
| P2P (customer transfer, funds received) | 28.07 | 99.52 | 6.6 |
| B2P/P2B (bills pay, salary, other payment) | 22.56 | 86.84 | 3.7 |
| Airtime purchase | 18.96 | 98.09 | 0.3 |
| Deposit/Withdrawal of funds | 18.02 | 100 | 12.4 |
| Transaction fees | 8.57 | 99.52 | 0.3 |
| Financial derivatives (M-Shwari/KCB withdrawal, deposit, term deposit, loan) | 3.44 | 48.33 | 7.63 |
| International transfers | 0.02 | 2.39 | 84.9 |
| Others | 0.36 | 44.74 | 14.9 |

Google Map Places API data

Data:

- Geolocational data on ATMs and mobile money agents in Kenya
- Collected from Google Map Places API using the R googleway package during November-December 2019.
- *Offer insight into geographical distribution of financial access points*

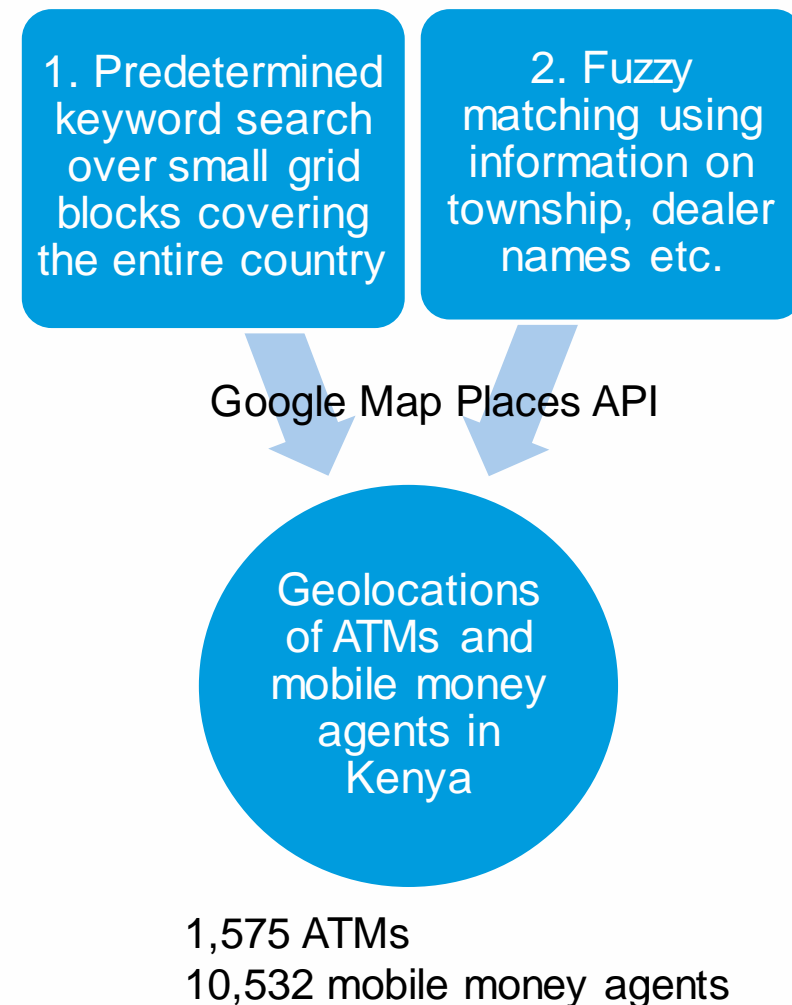
Methodology:

▪ Method 1

- ▶ Divide the map of Kenya into smaller areas with a grid system.
- ▶ Search queries with predetermined keywords (e.g., “ATM, Kenya”) passed to the API within each grid block until covering the entire country.
- ▶ Predetermined keyword search more challenging for mobile money agents which are often small shops/other retail stores.

▪ Method 2

- ▶ Web-scraping additional information on mobile money agents from publicly available lists of M-Pesa mobile money agents.
- ▶ Fuzzy matching to obtain additional geolocations of mobile money agents.
- ▶ Limitations due to lack of latest information



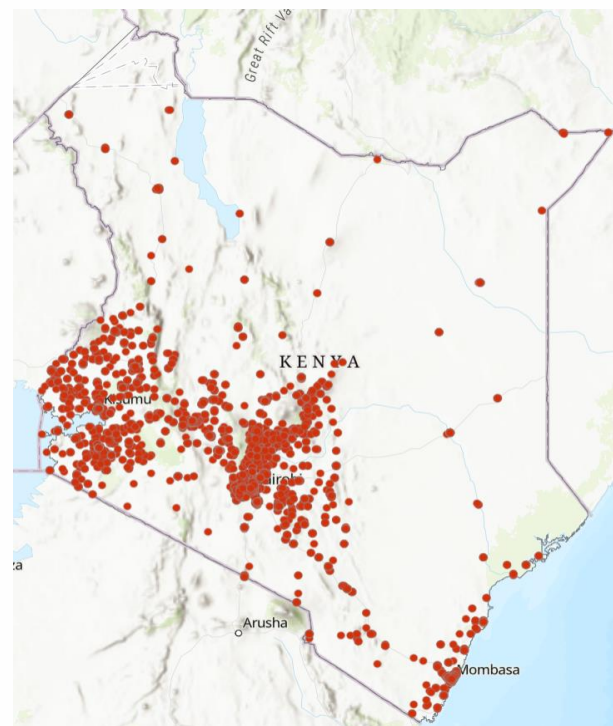
Assessing the Data Coverage

Comparison with other sources:

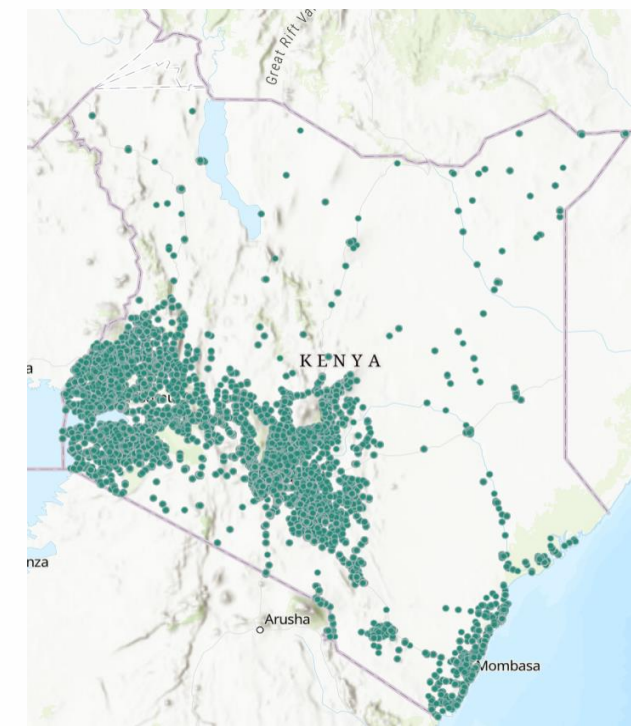
- IMF's Financial Access Survey (FAS) data
 - ▶ The number of ATMs and registered mobile money agents
 - ▶ But mobile money agent tills reported for Kenya so not directly comparable
- FinAccess Geospatial Mapping
 - ▶ Geolocational data on financial access points of various types in 2015
 - ▶ Survey undertaken by a “walk-the-street exercise”
 - ▶ Around 66,000 mobile money agents

| | Google Map Places API | FAS coverage (2019) | FinAccess (2015) | FAS coverage (2015) |
|----------------|-----------------------|---------------------|------------------|---------------------|
| Number of ATMs | 1,575 | 64 percent | 624 | 23 percent |

Mobile money agents



Google Map Places API data



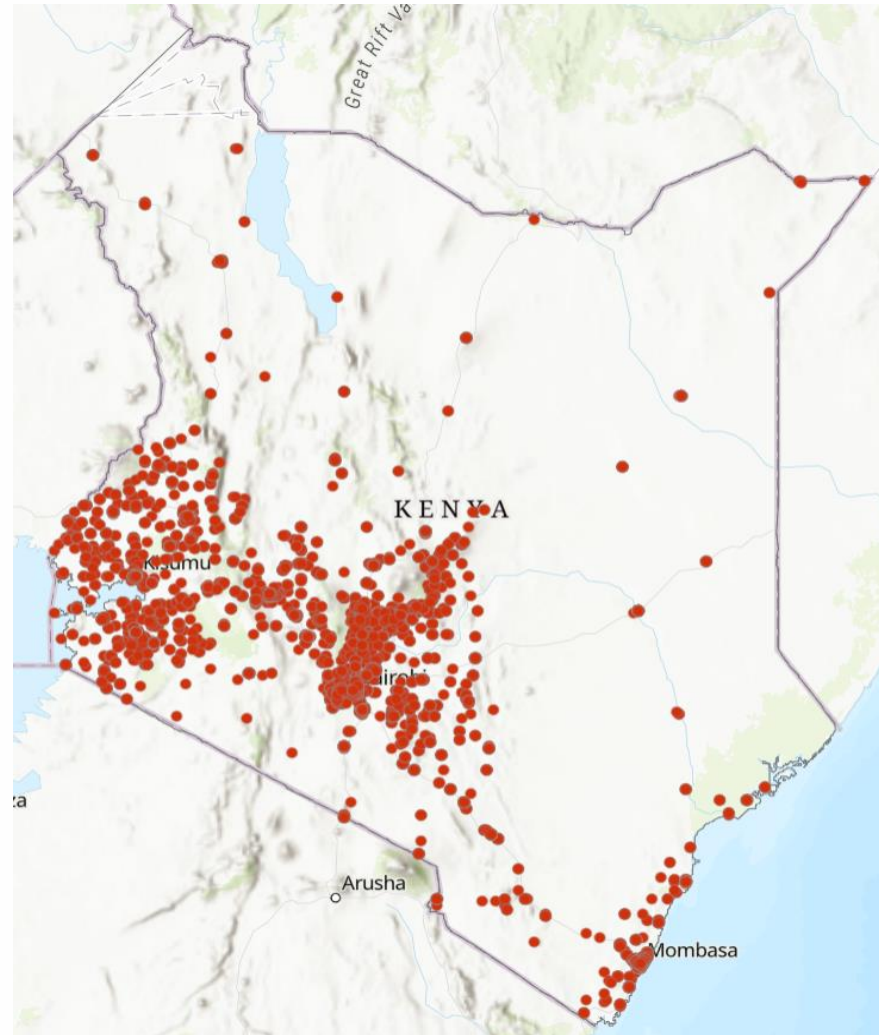
FinAccess data

Similar distribution as mobile money agents tend to be clustered.

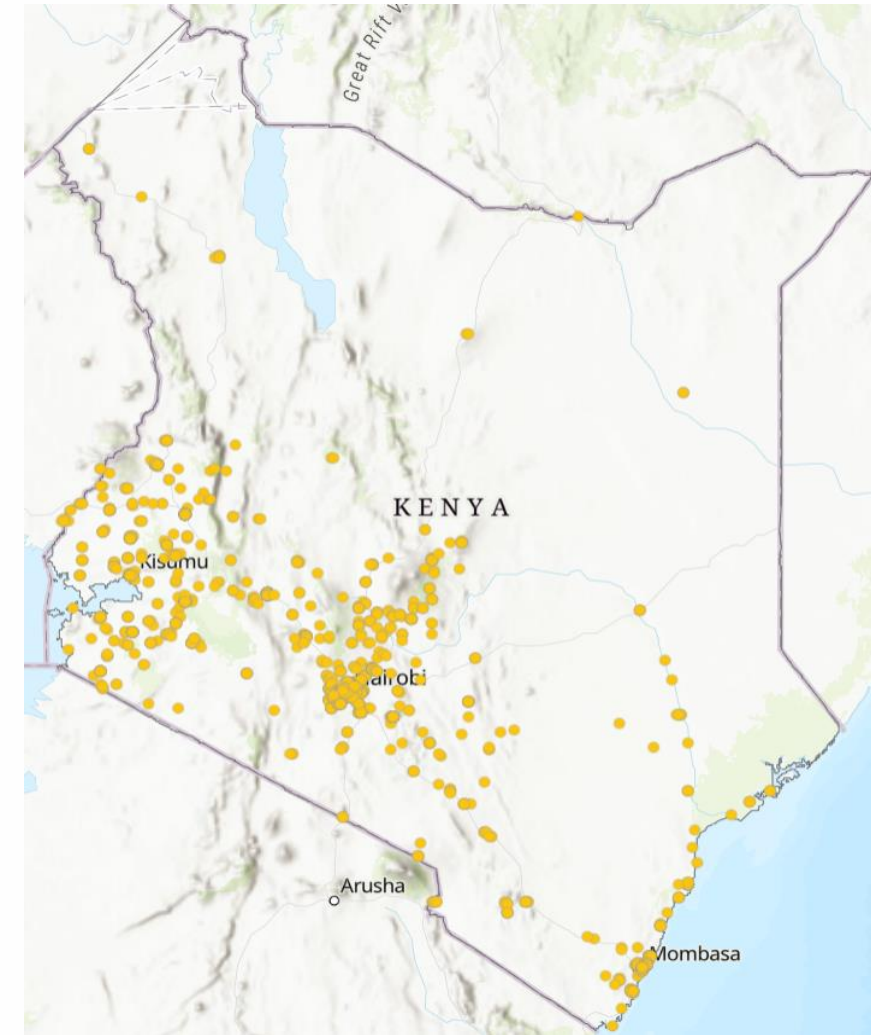
Google Map Places API data: Highlights

Mobile money agents

- Both ATMs and mobile money agents are concentrated in major cities (e.g., Nairobi).
- More mobile money agents than ATMs.
- More mobile money agents in the northern region.



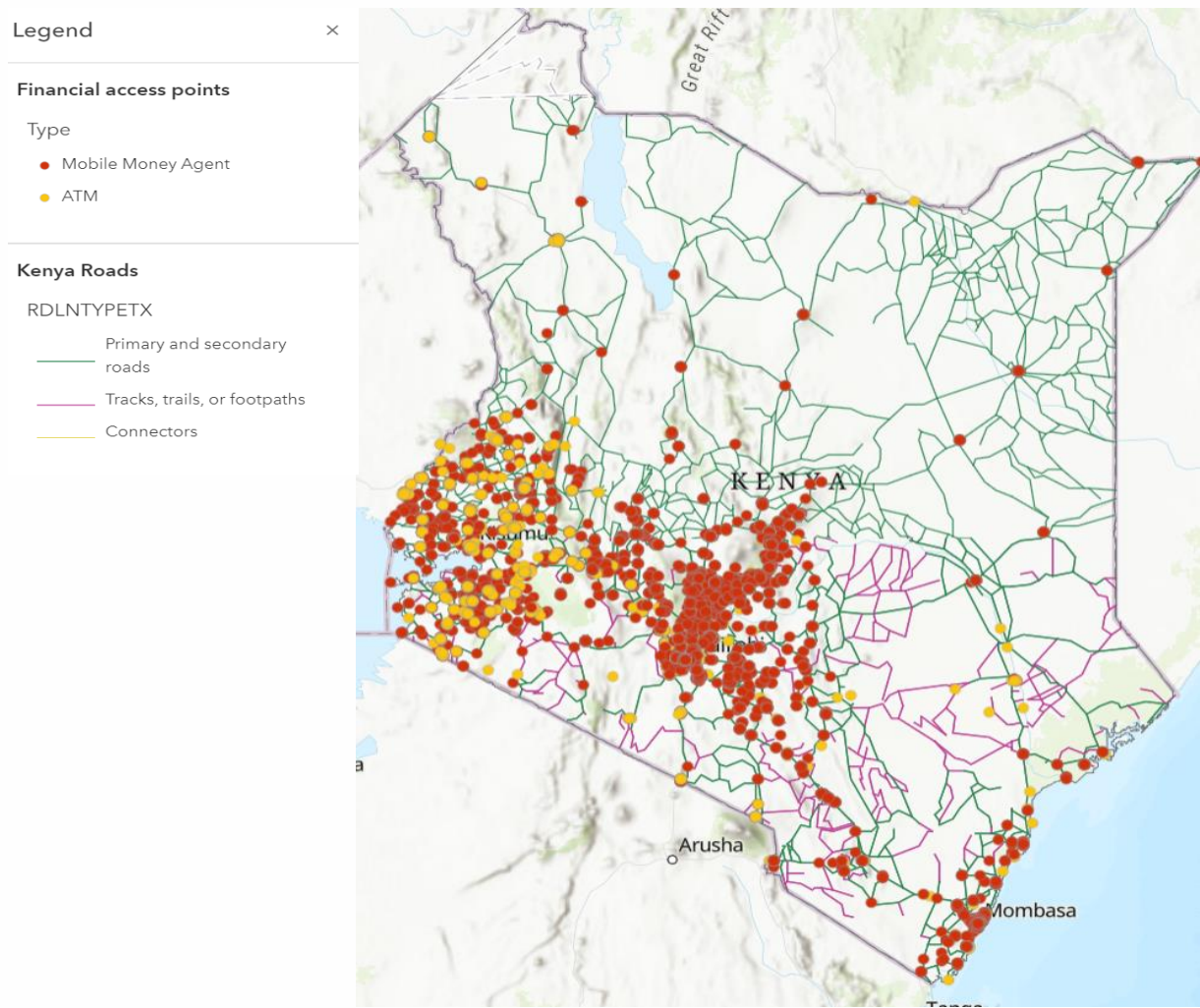
ATMs



Source: ArcGIS and IMF staff.

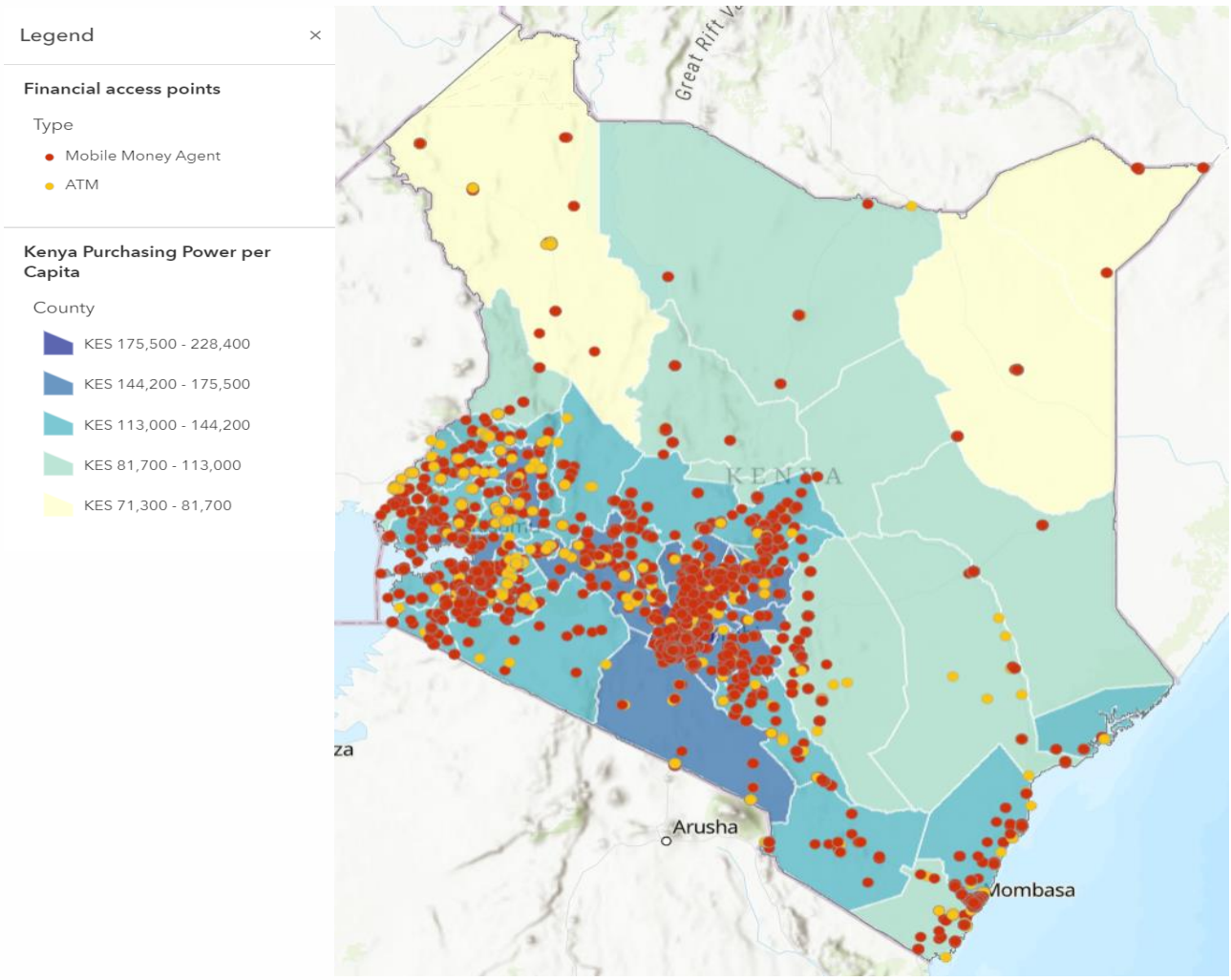
Google Map Places API data: Highlights

Traffic



Source: ArcGIS and IMF staff.

Income levels



Source: ArcGIS and IMF staff.

Takeaways

- Experimental studies to explore big data approach, using non-traditional data sources.
 - ▶ Mobile money usage patterns
 - ▶ Geographical distribution of financial access points
- Non-traditional data can offer new insight into financial access and use, possibly in a cost-effective way.
- Partnership with the private sector may be useful.

Example:

[Development Data Partnership](#): A partnership between international organizations including the IMF and technology companies, to facilitate the use of third-party data in international development.

Thank you!

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Sentiment analysis of tourist reviews from online travel forum for improving Indonesia tourism sector¹

Muhammad Abdul Jabbar, Arinda Dwi Okfania,
Anggraini Widjanarti and Alvin Andhika Zulen,
Bank Indonesia

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Sentiment Analysis of Tourist Reviews from Online Travel Forum for Improving Indonesia Tourism Sector

Muhammad Abdul Jabbar¹, Anggraini Widjanarti², Alvin Andhika Zulen³, Arinda Dwi Okfantia

Abstract

Tourism is one of strategic sector in Indonesia to increase Gross Domestic Product (GDP). In the long run, tourism sector can support exchange rate management and reduce current account deficit. Therefore, it is pivotal for optimizing tourism sector by improving aspect that is important for tourist. Text sentiment analysis, has become a flourishing frontier in the text mining community. This project studies tourist reviews of tourist destinations on the online travel forum, TripAdvisor, to produce 5 (five) tourism indicators, namely: the composition of foreign and domestic tourists, the composition of the country of origin of tourists, the movement of visits from month to month, the rating of tourist attractions, and positive and negative sentiments of tourists towards tourist destinations in the 3A aspect (attractions, amenities, and accessibilities). These five indicators can be used to help build a promotional strategy that is targeted and effective. The results of this positive review can be used to strengthen aspects of tourist destinations that can be promoted. Meanwhile, the negative review results are used as recommendations for future improvements in tourist attractions.

Keywords: text mining, sentiment analysis, tourism.

JEL classification: L83, L86

¹ Statistics Department – Bank Indonesia; E-mail: Muhammad_abdul@bi.go.id

² Statistics Department – Bank Indonesia; E-mail: Anggraini_widjanarti@bi.go.id

³ Statistics Department – Bank Indonesia; E-mail: Alvin_az@bi.go.id

Contents

| | |
|---|----|
| 1. Background..... | 3 |
| 2. Literature Review | 4 |
| 2.1 Tourism Quality Aspects | 4 |
| 2.2 TripAdvisor Review Reliability | 5 |
| 2.3 Text Mining on Tourism | 6 |
| 3. Methodology..... | 6 |
| 3.1 Data | 6 |
| 3.2 Framework | 8 |
| Data Pre-processing | 8 |
| Sentiment Analysis | 9 |
| 4. Result & Analysis | 10 |
| 4.1 Tourists Country of Origins Composition | 10 |
| 4.2 Tourism Destinations Monthly Visits | 12 |
| 4.3 Tourism Destinations Rating | 13 |
| 4.4 Sentiment Analysis Results..... | 14 |
| 5. Conclusion & Future Works | 19 |
| 5.1 Conclusion | 19 |
| 5.2 Future Works | 20 |
| References..... | 21 |

1. Background

Since 2011, Indonesian Government has launched tourism promotion program that has been using "Wonderful Indonesia" as main tourism branding tagline. As one of the strategic economic sector with big multiplier effect, tourism has been one of the sector that highly supported from the government. It has shown with many initiatives from the government to accelerate the development. In 2019 the government has set a high target of 20 Million foreign tourists and foreign exchange earnings of 17.6 Million USD with tourism shares of GDP around 5%. The improvement and development of major tourism destinations and events has been done steadily to realize Indonesia as world top tourism destinations.

Continuous and rigorous tourism sector development is necessary to support the growth and resilience of Indonesia external economy. The tourism sector supports Indonesia foreign exchange earnings and current account balance. To achieve Bank Indonesia vision to create a tangible national economic contribution, Bank Indonesia is closely coordinating with the government to support policy strategy formulation of tourism sector development in Indonesia.

The fact is, in Indonesia, Tourism sector data is relatively hard to find. Other than number of tourists that obtained from immigration, most of the data is acquired via survey. One of useful data that can be used in tourism sector development is the fulfilment of 3A aspects (attraction, amenities, and accessibilities) in Indonesia tourism destinations. This information can be used by regional government to provide recommendation and feedback regarding the development of each of tourism destinations on their region. However, this information is difficult to obtain in a bigger scale and usually done via survey that may need a lot of time and high cost in its collection.

Along with the growth of internet and smartphone technology, there are more online travel platform such as TripAdvisor, Booking.com, Agoda, & Traveloka. With the microblogging trends where users voluntarily and happily provide reviews of their travels, this drive the growth of user generated data in these platforms. These online travel platforms gathered large amount of user generated data that can be useful for many kinds of analysis including sentiment analysis. The Big Data that available on the platforms has big potential not only for the platforms, but also for government, authorities and tourism destination providers.

A study by Pan in 2007 shows that online reviews of products and services including tourism are deemed more trustworthy and credible. Recommendation based on previous user experience is one of the most influential information for the reader travelling decision. With the advancement of computer technology, we can extract information from large amount of text using Big Data Analytics and text mining methodology. These technology can be used to develop new indicator in the tourism sector that is credible and timely.

2. Literature Review

2.1 Tourism Quality Aspects

Tourism development involves components regarding tourism planning approaches, tourism attractions, accommodation, tourism facility and other tourism services such as transportation, infrastructure, and tourism management institutions (Inskeep, 1991). Tourism attractions can be categorized into 3 categories:

1. *Natural attraction* is attraction based on natural environment formation. Examples of natural attractions are climate, scenery, flora and fauna and also other form of nature uniqueness.
2. *Cultural attraction* is attraction based on people cultural activities. Examples of cultural attractions are archeological attractions, religious attractions, and traditional lifestyle.
3. *Special types of attraction* are man-made attraction such as theme park, circus, shopping mall, etc.

From (Yoeti, 1997) studies we understand that several aspects that correlate with the success of a tourism destination are 3As, the destination attractiveness (attraction), how the ease of access to the destination (accessibility) and the facility destination (amenities). Similarly, (Middleton, 2001) and (Mason, 2003) defines tourism product is the combination of 3 main components of attractions, the available facilities, and the destination accessibilities.

First, attraction is the elements in a tourism destination that broadly determine consumer or tourist's choice of what kind of attraction that they want to see or visit. The categories of attractions are natural tourism (involving nature, beach, climate, and other geographical and natural resources), man-made attractions (involving building and infrastructure such as historical and modern architecture, monument, garden, and thematic area), cultural attractions (involving historical and folklore, art and religion, musical theatre, museum, and other special events), and social attractions (involving the life local habitants within an area, language, and other social activities).

Second, amenities or the facility within a tourism destination is the elements that enables the tourists to participate and stay to enjoy the attraction of a tourism attraction. These elements are accommodation (hotel, apartment, villa, caravan, hostel, and guest house), restaurant, transportation (taxi, bus), other facilities (toilet and religious spaces), retail outlet (shops, travel agent, souvenir), and other services.

Third, accessibilities is the ease of access to the tourism destination that involves the cost, smoothness and comfort of travel through the infrastructure, road, airport, railway, seaport, operational factors such as operational route, transportation service frequency and cost, and government regulation regarding transportation safety.

Indonesia ministry of tourism also states that tourism development involves 3 aspects:

- 1) **Attractions**, site attractions (historical sites, places with comfortable or delightful climate, and places with beautiful sceneries), and event

attractions (eventful occasion such as congress, exhibition, or other kind of events);

- 2) **Amenities**, the availability and quality of the facilities, such as: Inn, restaurant, local transportation, and communication services;
- 3) **Accessibility**, the duration, cost, comfort, security, and availability of transportation to the location.

2.2 TripAdvisor Review Reliability

As the participatory nature of the Internet permits easy contribution of user-generated content without any editorial control, sites such as TripAdvisor allow users to share post-trip travel experiences have acquired immense popularity (Lo, 2011). TripAdvisor users can post review without any verification of whether the information that they submitted is factual or not and mostly based on their own experienced that can be very subjective, therefore raise some concerns of its reviews veracity.

Chua's study uses Social Network Analysis (SNA) to form clusters of highly-interlocked hotels that have been evaluated by a common pool of reviewers. The results of SNA indicated 58 isolated hotels that were not evaluated by any reviewer who had also commented on some other hotel in Singapore. These included 39 two-star hotels, 17 three-star hotels and two four-star hotels. The remaining 191 hotels (249 - 58) were interlocked cumulatively accounting for 4,192 interlocks with one another. The highly-interlocked hotels with five or more interlocks were used for further analysis. There were 46 such hotels accounting for 310 interlocks with one another. The hotels included three three-star hotels, 23 four-star hotels and 20 five-star hotels. Reviews posted by the top 100 reviewers in terms of volume of entries for the selected 46 interlocked hotels were used for further investigation. The top 100 reviewers contributed 424 reviews with an average of 8.83 reviews per hotel.

On establishing review baselines, the inter-reviewer and intra-reviewer reliability of contributions was studied using reviewer-centric and review-centric approaches respectively. Results suggest that reviews in TripAdvisor could be largely reliable (Chua, 2013).

SNA Result Illustration (Chua, 2013)

Figure 1

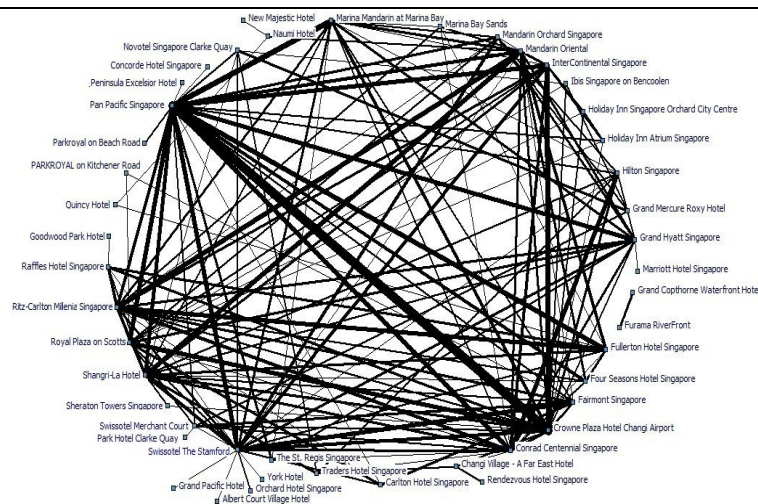


Fig. 1. Interlocking patterns of the 46 hotels with more than five connections.

2.3 Text Mining on Tourism

Text mining as defined by Hearst (2003) is the process of deriving high-quality information from text. It involves the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. In text mining, the goal is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down.

Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people's opinions, attitudes and emotions toward an entity (Medhat, 2014). Sentiment Analysis usually done by deriving information from large amount of data to studies positive and negative perceptions regarding an object. Sentiment Analysis usually done by creating classification models that can accurately predict whether a sentences expresses positive or negative sentiment toward the object that want to be analyzed.

Text mining is widely used in a lot of field, including tourism. Park utilizes text mining and twitter data to help formulate strategies of yacht cruise promotion (S.B Park, 2016). Text mining also utilized in hospitality industry to classify customer satisfaction from customer reviews (Berezina, 2016). A study by Pan states that online reviews of products and services including tourism are more accurate and timely (Pan, 2007). Recommendation based on experience of other users is one of the most influential source of information for a traveler's decision.

Thaha did a study that extract information from tourist reviews of Tangkuban Perahu and Kawah Putih in Indonesia (Thaha, 2020). The study extract the positive and negative reviews on the two tourism destinations from the visitors of the destinations. The result shows the rating of the tourism destinations and the positive sentiments for the rating, also negative reviews with words such as "expensive". The study uses Ekman emotion classification which is a classification of cross-cultural basic emotions proposed by Ekman et.al in 1992. These emotions are anger, disgust, fear, happiness, sadness, and surprise. The study shows that the review is dominated by "joy" emotion which indicates that most tourist enjoy the destination. From the negative sentiments, they extract "sadness" and "anger" emotion, and the details of the reviews can be used for recommendation by the tourism destination management.

3. Methodology

3.1 Data

Indonesia tourism destinations has similar attraction values to those in Thailand and Vietnam. The three peer countries tourism destinations strengths are its natural attractions and the sense of South East Asian spirituality and culture. Indonesia tourism destinations are competing with Thailand and Vietnam destinations since both countries have the same category of attractions and tourists market shares, especially for foreign tourists from Japan, China and Australia. Therefore, to analyze Indonesia tourism destinations competitive value, we compare it with similar tourism destinations in Thailand and Vietnam.

This study uses tourist reviews data that are extracted from TripAdvisor.com using its API. TripAdvisor API is a powerful API that enable its user to extract large amount of user reviews with all of its information in its database with specifiable parameters of tourism destinations or locations. This study used TripAdvisor Content API with specified parameters of tourism destinations in the regions we want to observe. The reviews are extracted from "things to do" pages for tourist destinations Indonesia, Thailand and Vietnam. The pages are crawled to extract its reviews automatically. This methodology extracted 604,395 reviews from 1330 tourist places within 26 major tourism destinations in Indonesia, Thailand and Vietnam from the year 2004 to 2021.

The information that extracted from the reviews are:

1. Tourists Destination Name
2. Category (Nature parks, museum, etc.)
3. Address
4. Rating information (Aggregates and number of reviews per rating scales(excellent, very good, average, poor, terrible)
5. Review information (Reviewer name, country of origin (not mandatory), review date, review content, visit date, rating, language)

The list of tourists destinations are as follows:

List of Tourism Destination

Table 1

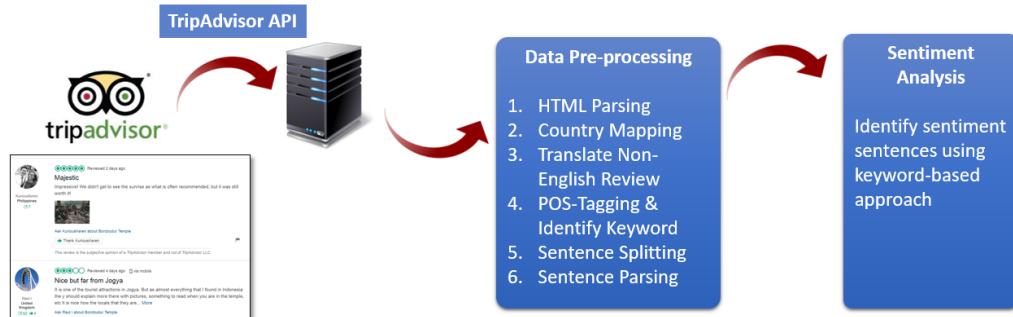
| Indonesia | | Thailand | Vietnam |
|------------------|------------------|------------|------------------|
| Borobudur | Wakatobi | Chiang Mai | Ho Chi Minh City |
| Bromo | Morotai | Bangkok | Halong Bay |
| Banyuwangi | Danau Toba | Phuket | Nha Trang |
| Kepulauan Seribu | Tanjung Kelayang | Ayutthaya | Hoi An |
| Kota Tua | Bali | | |
| Yogyakarta | Solo | | |
| Tanjung Lesung | Semarang | | |
| Mandalika | Batam | | |
| Labuan Bajo | Likupang | | |

3.2 Framework

The text mining framework of the reviews extraction is as follows:

Sentiment Analysis Framework

Figure 2



Data Pre-processing

Pre-processing is done to convert unstructured html data into structured data and enrich the information of each reviews prior to the sentiment analysis of reviews. The pre-processing steps in this paper are as follows:

1. HTML Parsing

The HTML that extracted are separated using HTML tags to parse the required information.

2. Country Mapping

The location information in the reviews are mapped using its geocode to map the reviewer information to country level information.

3. Review translation

Tourists reviews that are not in English are translated to English using google translate API.

4. Keyword identification

Using POS Tag and word frequency, this process identify the keyword that relevant to each review. To identify the keywords, we make a lists of all nouns and verbs and count the occurrence of each word in the reviews. The top 1000 words then filtered manually to extract keywords that are relevant to 3 categories of tourism services: attraction, accessibilities and amenities. This process identify 68 keywords that are relevant to the aspects, 83 negative sentiment keywords, and 39 positive sentiment keywords. These 3 aspects have sub-aspects which are as follows:

- a. Attraction : Point of interests, Hospitality, and Cleanliness.
- b. Accessibilities : Infrastructures and Transportation.
- c. Amenities : Accommodation and Facilities.

Keyword Example

Table 2

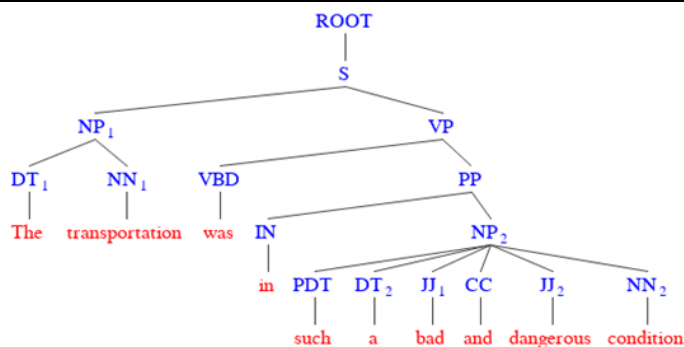
| Keywords | Aspects | Sub aspects |
|-----------|-----------------|-------------------|
| Road | Accessibilities | Infrastructure |
| Street | Accessibilities | Infrastructure |
| Transport | Accessibilities | Transportation |
| Train | Accessibilities | Transportation |
| Coral | Attractions | Point of Interest |
| Island | Attractions | Point of Interest |
| Service | Attractions | Hospitality |
| Toilet | Amenities | Facilities |
| Hotel | Amenities | Accommodation |

5. Sentence Parsing

The review text content are split into sentences and clauses. Then each sentences are parsed and decomposed into sentence parse tree structures based on English grammatical rule. For example, parse tree for the sentence *"The transportation was in such as bad and dangerous condition"* is as follows:

Sentence Parse Tree Example

Figure 3



Sentiment Analysis

This step identify the sentiments of each tourist review sentences using rule-based method based on the keyword that has been identified before. The rules for sentiment analysis are as follows:

1. Keywords for one of the aspect (attractions, accessibilities and amenities) has to be in the same clause as the sentiment keyword. For example in the sentence *"There is a lot of trash on the island, the corals are dead, and the water is dirty"* there are 3 clauses, S1, S2, and S3. In the sentence there are 3 negative attractions aspect sentiments which is identified by the keyword *"island"*, *"coral"*, and *"water"*. In S1, *"island"* keyword is next to *"trash"* keyword

which is a negative sentiment keyword. In S2, "coral" keyword is next to "dead" keyword which is a negative sentiment keyword. In S3, "water" keyword is next to "dirty" keyword.

2. If there is a negation word "not/no" in a clause, then the sentiment is reversed. For example in the sentence "The water is clean and there is no trash on the island" there are 2 positive sentiments on attraction aspect that is identified by "water" and "island" keyword. In the clause "there is no trash on the island" there is a negative sentiment keyword "trash" but there is a "no" as a negation keyword in the same clause. Therefore the clause is identified as a positive sentiment sentence on the attraction aspect.

Using the information collected from the data and sentiment sentences extracted from this process, this study create indicators of tourism destinations that can be used to analyze tourism qualities within different perspectives.

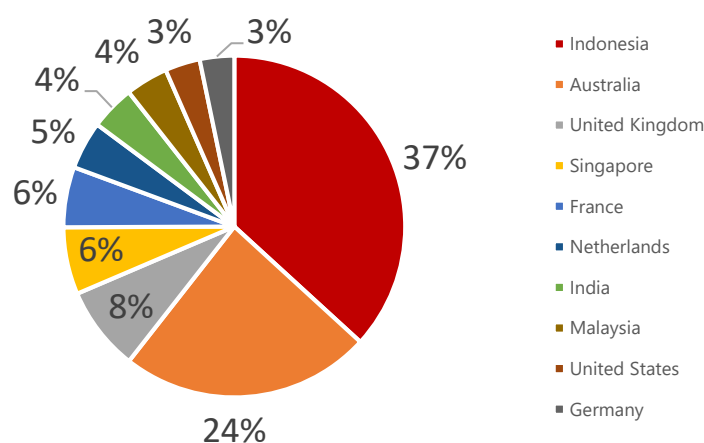
4. Result & Analysis

4.1 Tourists Country of Origins Composition

First, this study identify the composition of tourist's countries of origin from the location information that are obtained from the reviews of Indonesian tourism destination. The result is as follows:

Tourist Review Country of Origins on Indonesian Tourism Destinations (2017-2021)

Figure 4



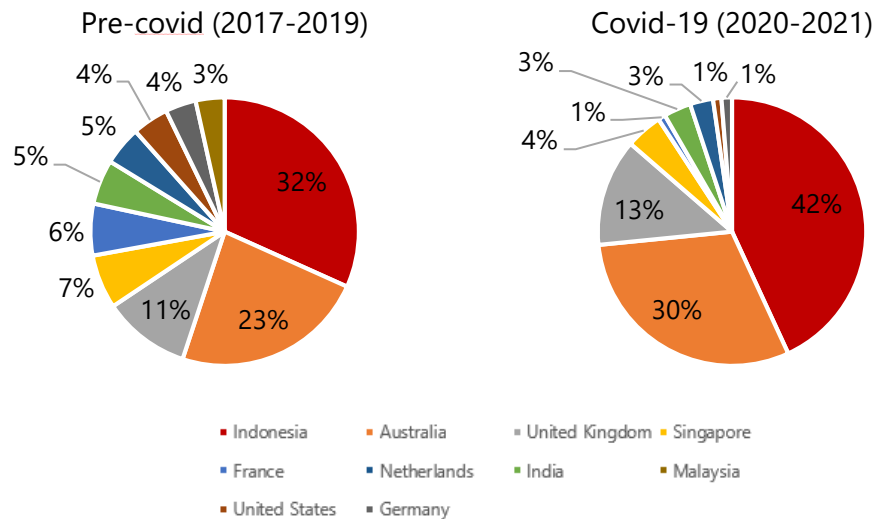
From this figure, we can see the origin of tourists that write reviews of their experience in Indonesia tourism destinations the most. Most of the reviews are from Indonesian local tourists with tourists from Australia, United Kingdom, Singapore and France from second to fifth. We can conclude that domestic tourists dominates the reviews of tourism destination the most, and Australia, UK and Singapore are the three countries that wrote reviews to the tourism destinations the most.

As we know in 2020 to 2021, international tourists are hindered to travel to Indonesia due to covid-19 pandemic that hit Indonesia and travel restrictions both

from the origin countries and Indonesian government. The reviews can be dissected by the year of visits and analyzed by pre-covid-19 (2017-2019) and covid-19 (2020-2021) years. As can be observed, there is a shift of tourist country of origins with much more of the reviews coming from local tourists (10% increase) in covid-19 era.

Tourist Review Country of Origins Pre-covid-19 and Covid-19 era on Indonesian Tourist Destinations (2017-2021)

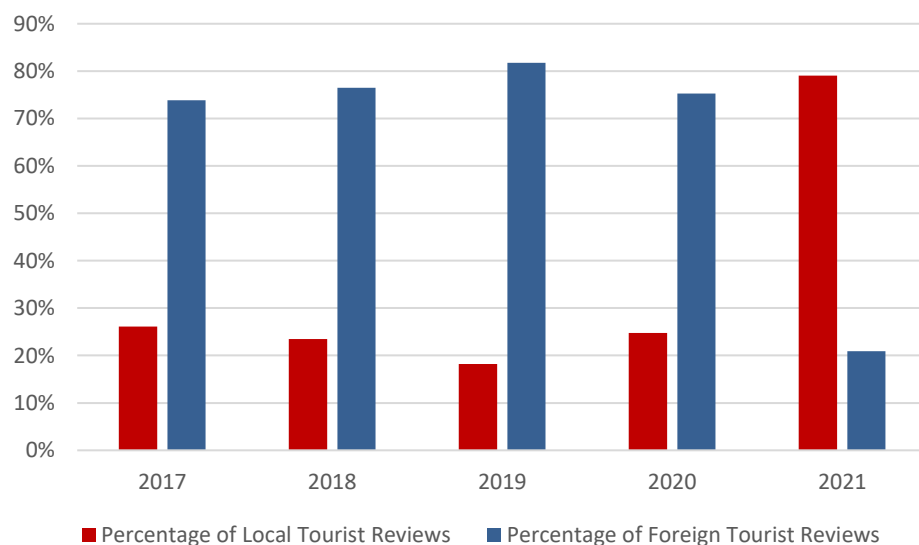
Figure 5



From the percentages of local vs foreign tourist reviews per year, in 2020 to 2021 the percentage rises significantly since there is virtually no international tourism and travel during covid-19 period especially in 2021. The percentages of local tourist reviews are declining in 2017-2019, but rises significantly in 2020 and in 2021 local tourist dominates so much more of the reviews although with low number of reviews in 2021.

Percentages of Local vs Foreign Tourists Reviews on Indonesian Tourism Destinations per Year (2017-2021)

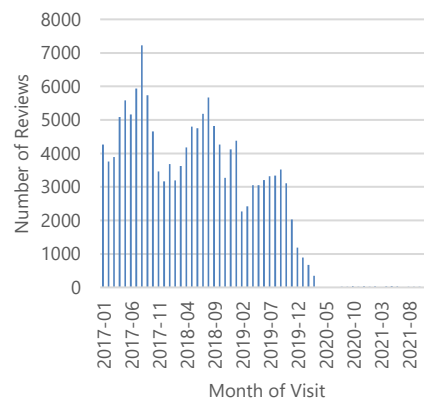
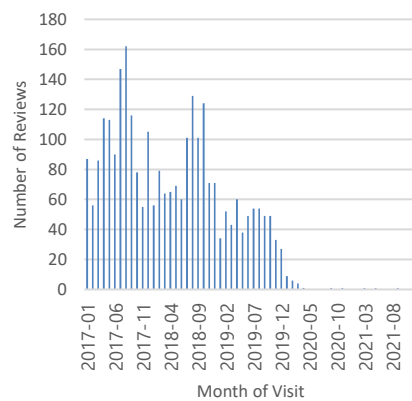
Figure 6



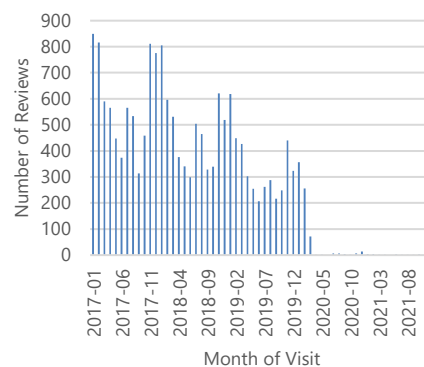
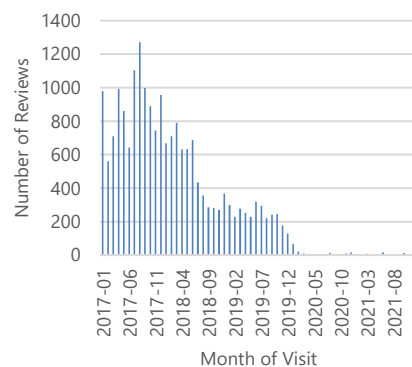
4.2 Tourism Destinations Monthly Visits

Using the visit date information that extracted from the reviews, this figures shows how the monthly visits for each tourism destinations are and when the tourism destination is more popular and have more or less visits. The effect of covid-19 pandemic to the visits of each tourism destination can be seen from the figures. The pandemic affect the visits of tourism destination as seen from the reviews with reviews reaching zero in most months in 2020-2021. Here are examples of 4 tourism destination monthly visits.

| | | | |
|---------------------------------|----------|----------------------------|----------|
| Borobudur Monthly Visits | Figure 7 | Bali Monthly Visits | Figure 8 |
|---------------------------------|----------|----------------------------|----------|



| | | | |
|----------------------------------|----------|----------------------------------|-----------|
| Yogyakarta Monthly Visits | Figure 9 | Chiang Mai Monthly Visits | Figure 10 |
|----------------------------------|----------|----------------------------------|-----------|



4.3 Tourism Destinations Rating

This study extracted the rating of tourism destinations from TripAdvisor and calculate the percentage of each of the rating from all the reviews. There are 5 level of rating (Excellent, Very Good, Average, Poor, and Terrible). Then this study ranks the tourism destination by the TripAdvisor rating to compare Indonesian tourism destinations with peer countries tourism destinations.

Tourism Destination Ranking by TripAdvisor Rating (2017-2021)

Table 3

| Tourism Destinations | Rating | | | | |
|----------------------|-----------|-----------|---------|------|----------|
| | Excellent | Very good | Average | Poor | Terrible |
| Halong Bay | 90% | 7% | 2% | 1% | 1% |
| Wakatobi | 76% | 12% | 7% | 3% | 2% |
| Banyuwangi | 74% | 20% | 4% | 2% | 1% |
| Borobudur | 69% | 23% | 5% | 1% | 1% |
| Bromo | 68% | 24% | 5% | 2% | 1% |
| Labuan Bajo | 55% | 27% | 12% | 3% | 2% |
| Bangkok | 51% | 34% | 12% | 2% | 1% |
| Ho Chi Minh City | 50% | 34% | 12% | 2% | 1% |
| Nha Trang | 50% | 29% | 15% | 4% | 3% |
| Danau Toba | 50% | 36% | 10% | 3% | 1% |
| Likupang | 49% | 29% | 20% | 0% | 1% |
| Ayutthaya | 48% | 36% | 12% | 2% | 2% |
| Chiang Mai | 46% | 35% | 14% | 3% | 2% |
| Yogyakarta | 45% | 33% | 15% | 4% | 3% |
| Tanjung Kelayang | 45% | 42% | 12% | 1% | 0% |
| Hoi An | 41% | 32% | 19% | 5% | 2% |
| Phuket | 40% | 32% | 18% | 5% | 5% |
| Bali | 35% | 41% | 20% | 3% | 2% |
| Kepulauan Seribu | 31% | 44% | 20% | 3% | 2% |
| Solo | 30% | 43% | 22% | 3% | 1% |
| Mandalika | 29% | 34% | 24% | 7% | 6% |
| Semarang | 28% | 46% | 22% | 3% | 1% |
| Morotai | 28% | 25% | 35% | 10% | 2% |

Note: Blue-shaded cells denotes peer countries tourism destinations

By rating, the best major tourism destinations in our list is Halong Bay in Vietnam with 90% of the reviews are in excellent rating. Then in the top 10 ranking we have 5 tourism destination in Indonesia in 2nd to 6th place with excellent rating around 55% to 76%. Lastly from 7th to 9th we have 3 popular tourism destinations in Thailand and Vietnam, Bangkok, Ho Chi Minh City, Nha Trang, followed by Danau Toba¹ in 10th

¹ Danau Toba or Lake Toba is the largest lake in South East Asia and one of the deepest in the world. Located in North Sumatera, Danau Toba is one of the most promising natural attraction in Indonesia

place. Indonesian tourism destination ratings are quite competitive with tourism destinations in Thailand and Vietnam with 6 Indonesian tourism destinations in the top 10 compared to 4 from Thailand and Vietnam.

The 3A tourism quality aspects can't be observed from the rating only, since there is no additional and detailed information comes from the rating. The review sentences are rich with detailed sentiment information about the experience the reviewer has in the tourism destinations. In the next subsection this study presents the sentiment analysis results that extracted more detailed information regarding tourist experience in the destinations.

4.4 Sentiment Analysis Results

From the number of sentiment review sentences, this study shows how each tourism destinations fare in each aspects of tourism quality and how they compare with each other.

Aspect Sentiments Review Sentences (2017-2021)

Table 4

| Tourism Destinations | Aspects Sentiment | | | | | |
|----------------------|-------------------|-----------|-----------------|-------------|-----------|-----------------|
| | Positive | | | Negative | | |
| | Attractions | Amenities | Accessibilities | Attractions | Amenities | Accessibilities |
| Borobudur | 1240 | 170 | 255 | 336 | 30 | 48 |
| Bromo | 698 | 96 | 153 | 296 | 31 | 94 |
| Banyuwangi | 231 | 8 | 15 | 54 | 4 | 74 |
| Kepulauan Seribu | 128 | 39 | 59 | 48 | 7 | 9 |
| Kota Tua | 183 | 47 | 44 | 87 | 14 | 25 |
| Bali | 48585 | 14738 | 11731 | 14616 | 3638 | 3428 |
| Tanjung Lesung | 22 | 9 | 9 | 8 | 1 | 0 |
| Mandalika | 209 | 51 | 33 | 76 | 17 | 11 |
| Labuan Bajo | 1506 | 305 | 324 | 329 | 92 | 85 |
| Wakatobi | 49 | 15 | 13 | 8 | 2 | 1 |
| Morotai | 6 | 0 | 4 | 8 | 2 | 0 |
| Danau Toba | 116 | 39 | 35 | 13 | 1 | 6 |
| Tanjung Kelayang | 44 | 12 | 10 | 8 | 2 | 1 |
| Yogyakarta | 3412 | 640 | 784 | 1090 | 154 | 367 |
| Solo | 336 | 98 | 91 | 63 | 26 | 23 |
| Semarang | 432 | 104 | 109 | 130 | 37 | 47 |
| Batam | 887 | 329 | 194 | 269 | 80 | 69 |
| Likupang | 20 | 10 | 8 | 13 | 4 | 1 |

| Aspect Sentiments Review Sentences (2017-2021) | | | | | | Table 4 |
|--|------|-----|-----|------|-----|---------|
| Chiang Mai | 1939 | 624 | 336 | 1179 | 392 | 386 |
| Bangkok | 2612 | 629 | 831 | 2406 | 860 | 1204 |
| Phuket | 2124 | 699 | 589 | 2109 | 822 | 925 |
| Ayutthaya | 368 | 90 | 91 | 306 | 97 | 100 |
| Ho Chi Minh City | 1724 | 362 | 307 | 1263 | 377 | 347 |
| Halong Bay | 1190 | 635 | 585 | 797 | 410 | 291 |
| Nha Trang | 1085 | 218 | 179 | 867 | 154 | 172 |
| Hoi An | 1947 | 444 | 286 | 1346 | 343 | 325 |

Note: Blue-shaded cells denotes peer countries tourism destinations

Indonesian popular tourism destinations such as Bali and Yogyakarta has more reviews compared to Thailand and Vietnam tourism destinations, while Indonesian less popular destinations have significantly little number of reviews. Generally there are more positive than negative review sentences across all tourism destinations. TripAdvisor reviews tend to be more positive than negative, suggesting that reviewers like to share about their travelling experiences and they not only use the platform to complaint about the bad experiences they may have.

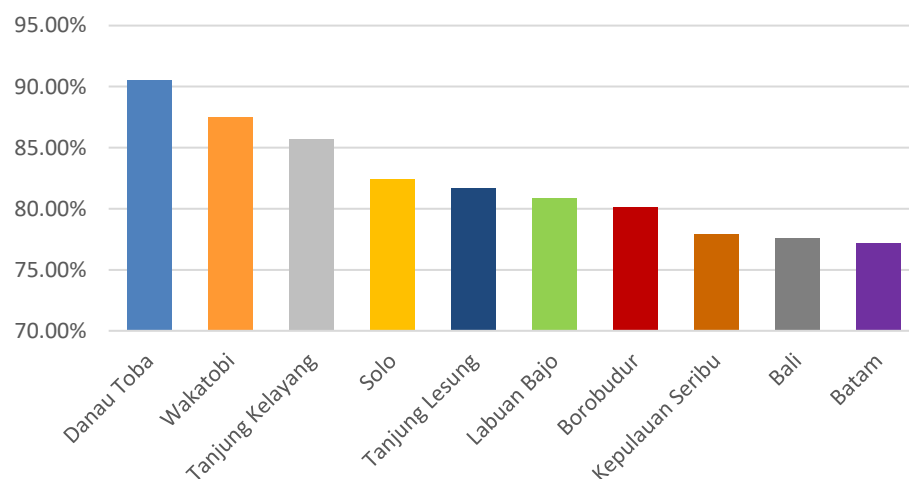
From the number of review sentences, the tourism destinations are ranked by the shares of its positive sentiments from all sentiment review sentences. Tourism Destinations like Danau Toba, Wakatobi², and Tanjung Kelayang³ has the highest shares of positive sentiments compared to other tourism destinations. From this perspective, Indonesia's tourism destinations ranks in all the Top 10 destination from the 26 tourism destinations.

² Wakatobi Regency is a group of ca. 150 islands forming an administrative regency located in Southeast Sulawesi, Indonesia. The four largest islands are Wangi-wangi, Kaledupa, Tomia and Binongko. The Wakatobi Islands are a part of the Coral Triangle, which contains one of the richest marine biodiversity on earth.

³ Tanjung Kelayang is a Special Economic Area that is developed for its beach tourism. It's located in Belitung Island and famous for blue and clear seawater and also the soft, white, and beautiful sand in the shorelines.

Top 10 Tourism Destinations by Shares of Positive Sentiment in All Aspects (2017-2021)

Figure 11

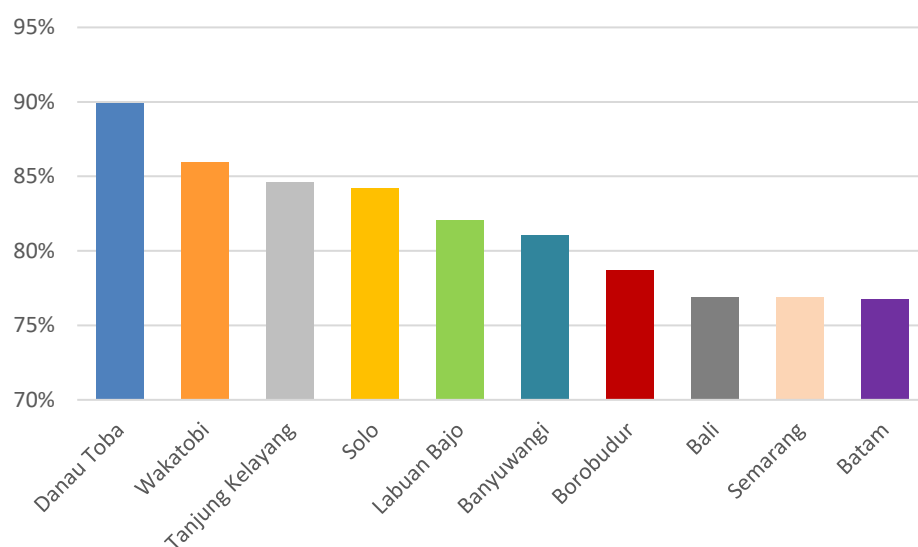


This ranking can be used to give recommendation to the government of which tourism destinations that are better priority to promote more intensely. Danau Toba, Wakatobi and Tanjung Kelayang are three tourism destinations that doesn't have that many review sentiment sentences and can be considered underrated. With better promotional strategies, we can expect improvement of tourist visits to these three tourism destinations after the covid-19 pandemic ends.

The ranking can also be decomposed for each qualities aspects.

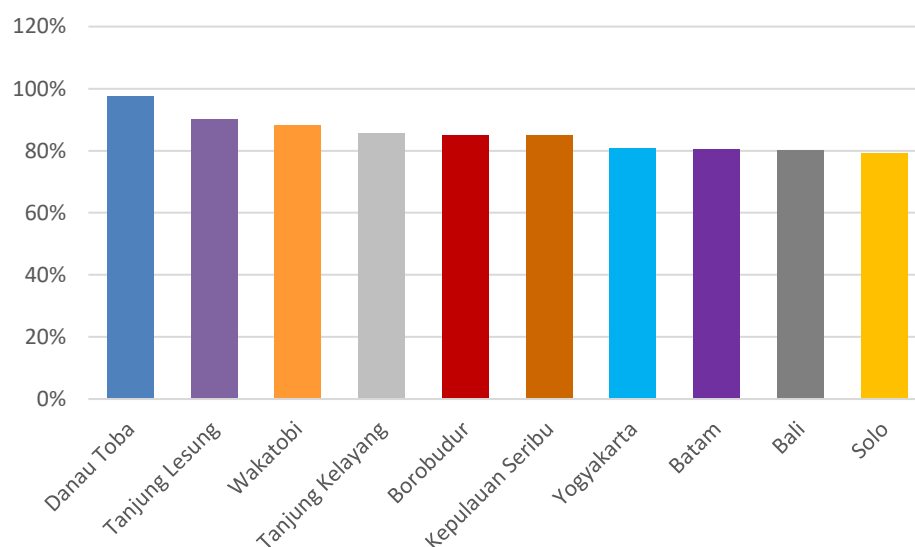
Top 10 Tourism Destinations by Shares of Positive Sentiment in Attraction Aspects (2017-2021)

Figure 12



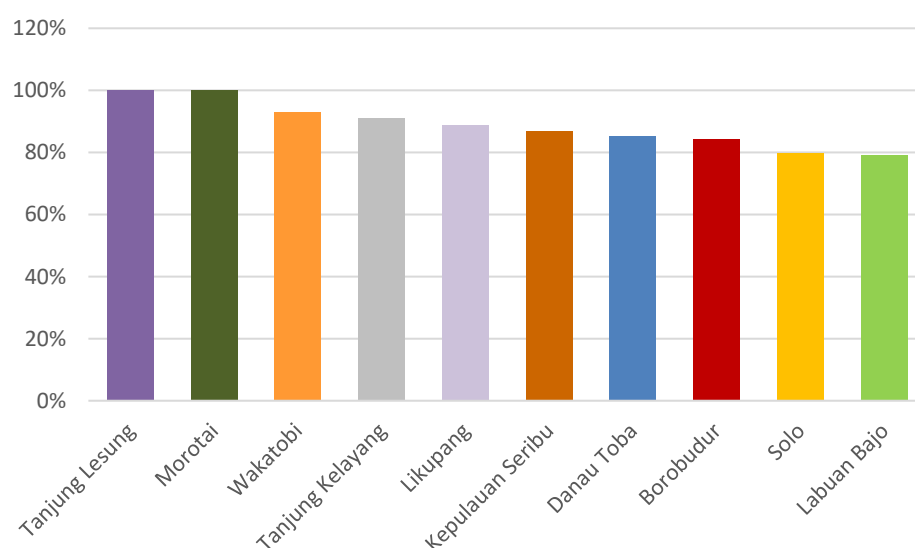
Top 10 Tourism Destinations by Shares of Positive Sentiment in Amenities Aspect (2017-2021)

Figure 13



Top 10 Tourism Destinations by Shares of Positive Sentiment in Accessibilities Aspect (2017-2021)

Figure 14



This table shows the granular details of review sentences that we have obtained using this methodology. We are able to extract positive and negative sentiments sentences that we can use to understand more about the main aspects for promotion and improvement for each of the tourism destinations. The positive sentiments can be used to formulate policy to strengthen promotional aspects of tourism destinations, while the negative sentiments can be used to recommend aspects of tourism quality that needs to be improved.

Tourism Destinations Sentiment Examples Top 10 Tourism
Destination by Positive Review Shares

Table 5

| Tourism Destinations | Sentiment Reviews |
|----------------------|---|
| Danau Toba | The lake and its surrounding hills are beautiful just to watch. |
| | What is lacking is the infrastructure to make this place a real tourist spot. |
| Wakatobi | The care was always excellent, diving places could be gotten in the short term. |
| | Ferry schedule/service is a bit tough due to changing weather and local conditions. |
| Tanjung Kelayang | This beach is quite managed neatly, because the beach is clean, on the beach there are also many boats that can be rented for Hopping Island |
| | Unfortunately the boat rental is quite expensive 500rb / boat with the contents of max 7 people |
| Solo | Very interesting collection of batik textile as a piece of Indonesian history. Excellent helpful very kind guide tour we had only for our family. |
| | Closed for unadvertised lunch between 11.30am and 1pm so we couldn't even get into the museum. Staff were pretty unfriendly so we left. |
| Tanjung Lesung | The beach club provides well-maintained bathrooms and prayer facilities. The beach is also quite clean and well-maintained |
| | After getting very disappointed, i kept my faith and saw the board which listed the prices. i was shocked. An ATV rental for 20 min (max 2 ppl) for 300,000 Rp. and Jet-ski for 120,000 Rp. |
| Labuan Bajo | Very very beautiful beach. We also had a very nice snorkeling. Would really love to be back here again! |
| | If I knew road was this steep and half paved bumpy, I wouldn't go. Not recommend to older adult. |
| Borobudur | Went in before sunrise with a guide who helped find the picturesque spots. The temple is amazing, historic and beautiful. |
| | Ticket prices are ridiculous high for foreigners. Locals paid only IDR 40,000 whereby for foreigners the price is 10 times higher. It is ridiculous. Nothing much to view. |
| Kepulauan Seribu | This was a visit to the volunteer of Earth's Day with LIPI on the Thousand Islands Pari Island...The sunset is pink and beautiful ... |
| | Visited island with my wife. After 10 minutes we arrived my wife started to cry from shock. The island is full of rubbish, dirty, water is dirty too. |
| Bali | We had super good time. The staffs were nice and helpful, they arranged 3 round beach benches for our group. The beach clean, water nice and warm, music nice. Foods also good in here. Everything. |

Tourism Destinations Sentiment Examples Top 10 Tourism
Destination by Positive Review Shares

Table 5

| | |
|-------|--|
| | Boat was overcrowded and life jackets not in place. Sure the boat got there but if there was a problem unnecessary casualties would have occurred. |
| Batam | It's a Chinese Buddhist temple very near to city centre. It is a very silence place & a suitable place for worship. |
| | Very bad experience arriving at this Terminal. Queue was long, custom was rude and clearance was slow. |

Notes: Green shaded cells are positive sentiment reviews. Red shaded cells are negative sentiment reviews

From review sentiment sentences, we can conclude that Indonesia tourism destinations qualities in each aspects are as follows:

1. Attractions : While Indonesia has strong natural and cultural attractions that praised by most of the review sentiments, the cleanliness of its attractions hasn't maintained well in some of the tourism destinations.
2. Amenities : Most of the reviews express good review of clean facilities, but there are still complaints in some of the places with dirty toilets and facilities are the most frequent complaints. Other frequent complaints are also regarding the cost and comfort of transportation within the tourism destinations.
3. Accessibilities: The access and cost of travel to the tourism destination is one of the most frequent complaints in many of the less popular destinations. The accessibilities has to be improved for many of the less popular tourism destinations for more tourists to become recurring visitors of the tourism destinations.

Generally our recommendations to improve the quality of most Indonesia tourism destinations from the result of this study are as follows:

1. Promote its natural and cultural attraction aspect more intensely.
2. Develop a framework to maintain consistent and thorough cleanliness in the tourism destinations locations.
3. Carefully take attention to the transportation services to and within the tourism destinations, especially regarding the cost and quality.

5. Conclusion & Future Works

5.1 Conclusion

TripAdvisor API is a powerful tools to extract large amount of reviews that can be utilized to discover strengths and weaknesses of tourism destinations. Using TripAdvisor API we extracted 604,395 reviews of 26 tourism destinations in Indonesia, Thailand and Vietnam from 2004-2021 and analyze it using sentiment analysis to gather information about how the quality of each tourism destination in 3A (Attractions, Amenities, Accessibilities) tourism quality aspects. We also compare

Indonesia tourism destinations with peer countries (Thailand and Vietnam) tourism destinations using the rating that we obtained from TripAdvisor and the sentiment analysis results.

From the TripAdvisor rating collected, Indonesia have 6 tourism destinations in the top 10 with shares of excellent rating ranges from 50% to 76% compared to 4 peer countries tourism destinations ranks in the top 10 with shares of excellent rating from ranges from 50% to 90%. From sentiment perspectives, Indonesian tourism destination generally have better positive sentiment shares compared to peer countries tourism destination.

The top 3 tourism destinations by positive sentiment shares are Danau Toba, Wakatobi, and Tanjung Kelayang. From the review sentences extracted we can formulate better strategies promotion to improve the 3A aspect of each tourism destination qualities.

Our recommendations for most Indonesia tourism destinations from the result of this study are as follows:

1. Promote its natural and cultural attraction aspect more intensely.
2. Develop a framework to maintain consistent and thorough cleanliness in the tourism destinations locations and the facilities around it.
3. Carefully take attention to the transportation services to and within the tourism destinations, especially regarding the cost, quality, and comfort.

5.2 Future Works

These are some room for improvements and future works that may be done.

There are other travel platform reviews e.g Yelp, Google Places, Booking.com, etc that may have useful similar data. Although TripAdvisor is the most popular tourist review platform, there may be other interesting findings if we perform similar methodology to data from other travel platforms.

More sentiment reviews can be extracted with expansion of aspects and sentiment keywords list. The study has used 1000 of the most frequent English nouns and verbs, but there may have some nouns and verbs that can be used to expand the scope of the aspects and sentiment keywords.

Similar sentiment analysis methodology can be done for accommodation reviews data especially those around tourism destinations. This study will be helpful to understand the qualities of accommodation which is part of the second tourism quality aspect (Amenities) in a more detailed manner.

Sharia tourism is also one of the interesting perspective of tourism that can be looked more closely. As the biggest Muslim country in the world, Indonesia has the potential to capture the majority of Sharia tourist market share. Similar study can be done using sharia tourism specific aspects keyword such as praying room, mosque, halal restaurant, etc.

References

- Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. *Journal of Hospitality Marketing & Management*, 25(1), 1-24.
- Chua, A. Y., & Banerjee, S. (2013). Reliability of reviews on the Internet: The case of TripAdvisor. In *World Congress on Engineering & Computer Science: International Conference on Internet and Multimedia Technologies* (pp. 453-457). York.
- Ekman, P. (1992). Facial expressions of emotion: New findings, new questions.
- Hearst, Marti. "What is text mining." *SIMS, UC Berkeley* 5 (2003).
- Inskip, E. (1991). *Tourism planning: An integrated and sustainable development approach*. John Wiley & Sons.
- Lo, I. S., McKercher, B., Lo, A., Cheung, C., & Law, R. (2011). Tourism and online photography. *Tourism management*, 32(4), 725-731.
- Mason, P. (2020). *Tourism impacts, planning and management*. Routledge.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4), 1093-1113.
- Middleton, V. T., & Clarke, J. R. (2012). *Marketing in travel and tourism*. Routledge.
- Pan, B., MacLaurin, T., & Crotts, J. C. (2007). Travel blogs and the implications for destination marketing. *Journal of travel research*, 46(1), 35-45.
- Park, S. B., Ok, C. M., & Chae, B. K. (2016). Using Twitter data for cruise tourism marketing and research. *Journal of Travel & Tourism Marketing*, 33(6), 885-898.
- Thaha, A. R., & Aziz, F. (2020). Text Mining on Tourism Destinations in Bandung Raya (Case Study: Tangkuban Perahu and Kawah Putih). *Secretary and Business Administration Journal*, 4(2), 146-156.
- Yoeti, Oka A. (1997). *Planning and Development of Tourism*, PT. Pradnya Paramita: Jakarta.

Sentiment Analysis

of Tourist Reviews from Online Travel Forum

for Improving Indonesia Tourism Sector

**Anggraini Widjanarti, Muhammad Abdul Jabbar, Arinda
Dwi Okfantia, Alvin Andhika Zulen.**

IFC-Bank of Italy: Data science in central banking, 14-17 February 2022

*The views expressed here are those of the authors and do not necessarily reflect the views of Bank Indonesia



OUTLINE

2



BACKGROUND AND GOALS



FRAMEWORK



METHODOLOGY



RESULTS AND ANALYSIS



CONCLUSION AND FUTURE WORKS



BACKGROUND AND GOALS

3

Background

1. Tourism is one of strategic priority sector in Indonesia with big multiplier effects to the economy.
2. Discovering big picture and granular details of tourism reviews in popular tourism destinations
3. Untapped Big Data sources (TripAdvisor reviews) that can be analyzed using text mining

Goals

- We extract tourist reviews in Indonesia's and peer countries major tourism destinations to analyze its 3A's tourism qualities aspects (Accessibilities, Accommodation, Amenities).
- Collect positive reviews to strengthen the promotional aspects of tourisms, and negative reviews to identify recommendation for improvement in tourism destinations.

Literature Study

1. Tourism Development Aspects:

Inskeep, E. (1991). Tourism planning: An integrated and sustainable development approach: The study mainly talks about the aspects of tourism development.

"Tourism development involves components regarding tourism planning approaches, tourism attractions, accommodation, tourism facility and other tourism services such as transportation, infrastructure, and tourism management institutions."

2. TripAdvisor Reviews Credibility:

Chua, A. Y., & Banerjee, S. (2013). Reliability of reviews on the Internet: The case of TripAdvisor: The paper studies the reliability of TripAdvisor review by using Social Network Analysis of hotel reviewers in Singapore.

"On establishing review baselines, the inter-reviewer and intra-reviewer reliability of contributions was studied using reviewer-centric and review-centric approaches respectively. Results suggest that reviews in TripAdvisor could be largely reliable."

3. Text Mining on Tourism Data

Thaha, A. R., & Aziz, F. (2020). Text Mining on Tourism Destinations in Bandung Raya (Case Study: Tangkuban Perahu and Kawah Putih). Similar sentiment analysis study using TripAdvisor review data on Indonesia tourism destinations.

"Research findings illustrate that text mining can be applied to travel reviews for tourism destinations in Indonesia."



FRAMEWORK & SCOPE

4

To understand Indonesia's tourism competitive advantages we analyze sentiments to 3 aspects of tourism qualities (Accessibilities, attractions, and amenities) from 18 popular tourism destinations from Indonesia and 8 popular tourism destinations from Vietnam and Thailand.

Framework

1



Review extraction using TripAdvisor API

2



Data pre-processing to structurize the data and enrich the information in the data

3



Sentiment analysis using text mining to extract sentiment sentences from the reviews

Scope

Tourism Quality Aspects

Accessibilities

1. Infrastructure
2. Transportation

Attractions:

1. Point of interests
2. Hospitality
3. Cleanliness

Ammenities:

1. Accommodation
2. Facilities

Tourism Destinations

Indonesia:

1. Borobudur
2. Bromo
3. Banyuwangi
4. Kepulauan Seribu
5. Kota Tua
6. Yogyakarta
7. Tanjung Lesung
8. Mandalika
9. Labuan Bajo
10. Wakatobi
11. Morotai
12. Danau Toba
13. Tanjung Kelayang
14. Bali
15. Solo
16. Semarang
17. Batam
18. Likupang

Peer Countries:

1. Chiang Mai
2. Bangkok
3. Phuket
4. Ayutthaya
5. Ho Chi Minh City
6. Halong Bay
7. Nha Trang
8. Hoi An



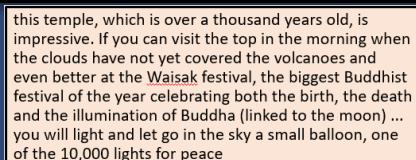
Country Mapping



| | |
|------------------------|---------|
| Name | Dina K |
| Tourist Origins | Izhevsk |

| | |
|---------------------------|---------|
| Name | Dina K |
| Tourist Origins | Izhevsk |
| Country of Origins | Rusia |

Sentence Split & Parsing



But/CC what/WP we/PRP found/VBD was/VBD sad/JJ sad/JJ
sad/JJ . There/EX is/VBZ a/DT lot/NN of/IN trash/NN on/IN
the/DT island/NN , the/DT corals/NN are/VBP dead/JJ , and/CC
the/DT water/NN is/VBZ dirty/JJ .

| Keyword | Aspek | Sub Aspek | Keyword |
|---------------|-----------------|-------------------|--------------|
| Road | Accessibilities | Infrastructure | Dirty |
| Street | Accessibilities | Infrastructure | Bad |
| Transport | Accessibilities | Transportation | Horrible |
| Train | Accessibilities | Transportation | Poor |
| Coral | Attractions | Point of Interest | Pricey |
| Island | Attractions | Point of Interest | Smell |
| Service | Attractions | Hospitality | Impolite |
| Fee | Attractions | Price | Lack |
| Trash | Attractions | Cleanlines | Rip off, etc |

We booked a local tour and departed to nowhere, expecting to find a paradise looking island in the middle of nowhere. But what we found was **sad sad sad**. There is a lot of trash on the island, the corals are dead, and the water is dirty. Still, the people were very welcoming and our tour guide tried hard to make our stay enjoyable. We did some snorkelling, but it just reminded us about how much harm we have caused to the environment.

Sentence Splitting

1. We booked a local tour and departed to nowhere, expecting to find a paradise looking island in the middle of nowhere.
2. But what we found was sad sad sad.
3. There is a lot of trash on the island, the corals are dead, and the water is dirty.
4. Still, the people were very welcoming and our tour guide tried hard to make our stay enjoyable.
5. We did some snorkelling, but it just reminded us about how much harm we have caused to the environment.



METHODOLOGY – SENTIMENT EXTRACTION

6

Sentiments are extracted from the review sentences data using rule-based methodology. The rules for the extraction are as follows:

- 1 Keywords for one of the attractions, accessibilities and amenities aspect has to be in the same clause as the sentiment keyword.
- 2 If there is a no/not word in the clause, the sentiment is reversed (positive to negative, vice versa)

Example

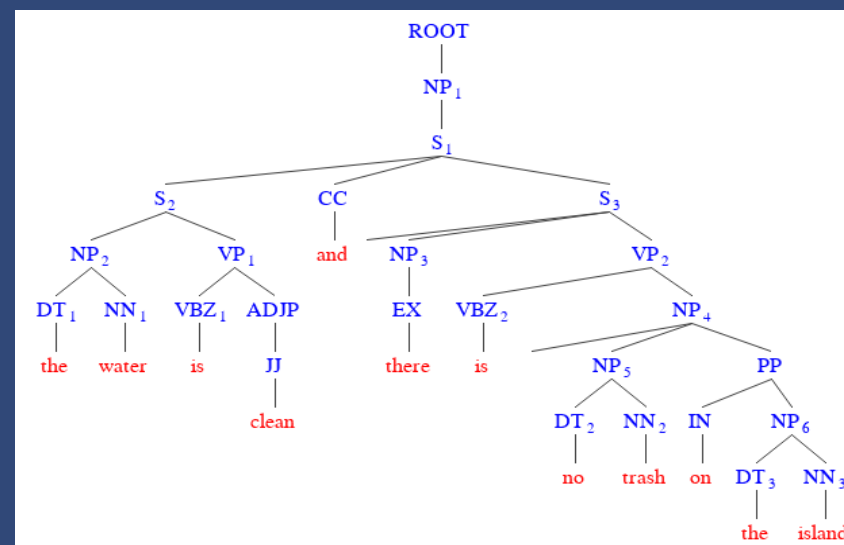
“The water is clean and there is no trash on the island”

There are 2 clauses in the review sentence:

S1: The water is clean

S2: There is no trash on the island

1. S1 is a positive sentiment sentence because there is positive sentiment keyword 'clean' and attraction keyword 'water'
2. S2 is a positive sentiment sentence because there is negation word 'no' with negative sentiment keyword 'trash', and attraction keyword 'island'



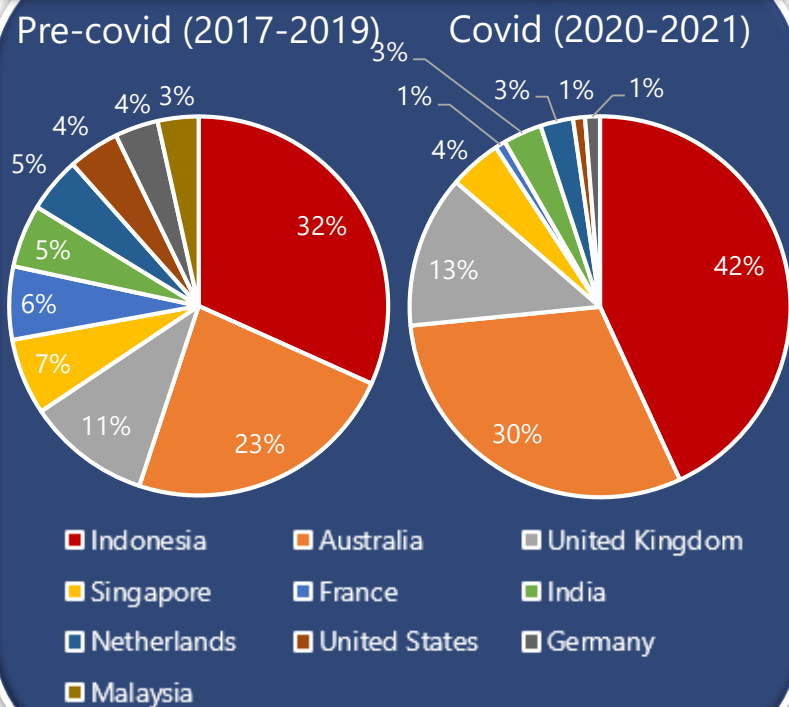


RESULT – GENERAL OVERVIEW

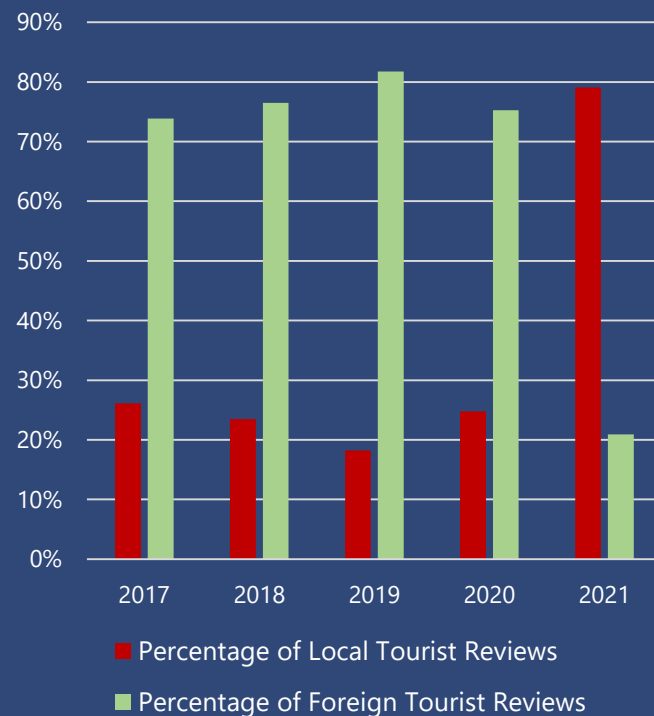
7

- As the covid-19 pandemic causes travel restrictions around the world, we can see the shift of tourists composition in Indonesia travel destinations. It's also shown in the number of reviews in TripAdvisor. On Covid era, local tourists dominates the review compared to foreign tourists.
- Halong Bay is the top tourists destination by TripAdvisor rating followed by 5 tourism destination from Indonesia.

Top 10 Tourist Reviews Countries of Origin*



Percentages of Local vs Foreign Tourist Reviews*



Top 10 Tourist Destinations by Rating 2017-2021

| Tourist Destination | Rating | | | | |
|---------------------|-----------|-----------|---------|------|----------|
| | Excellent | Very Good | Average | Poor | Terrible |
| Halong Bay | 90% | 7% | 2% | 1% | 1% |
| Wakatobi | 76% | 12% | 7% | 3% | 2% |
| Banyuwangi | 74% | 20% | 4% | 2% | 1% |
| Borobudur | 69% | 23% | 5% | 1% | 1% |
| Bromo | 68% | 24% | 5% | 2% | 1% |
| Labuan Bajo | 55% | 27% | 12% | 3% | 2% |
| Bangkok | 51% | 34% | 12% | 2% | 1% |
| Ho Chi Minh City | 50% | 34% | 12% | 2% | 1% |
| Nha Trang | 50% | 29% | 15% | 4% | 3% |
| Danau Toba | 50% | 36% | 10% | 3% | 1% |

*Indonesia tourism destinations only, 2017-2021 period

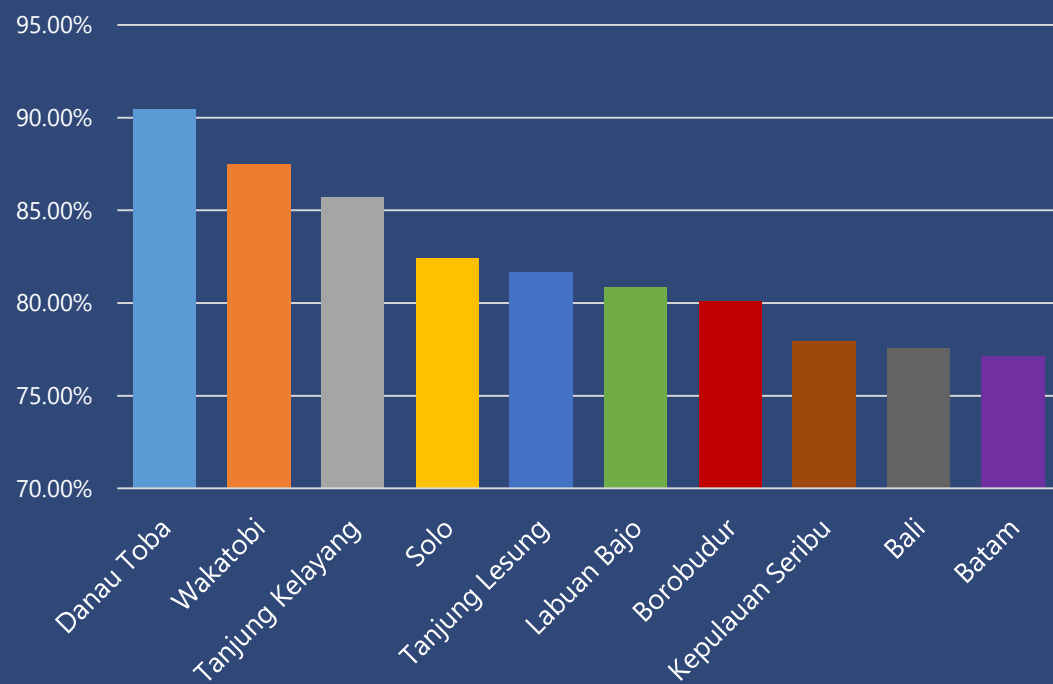


RESULT – SENTIMENT ANALYSIS

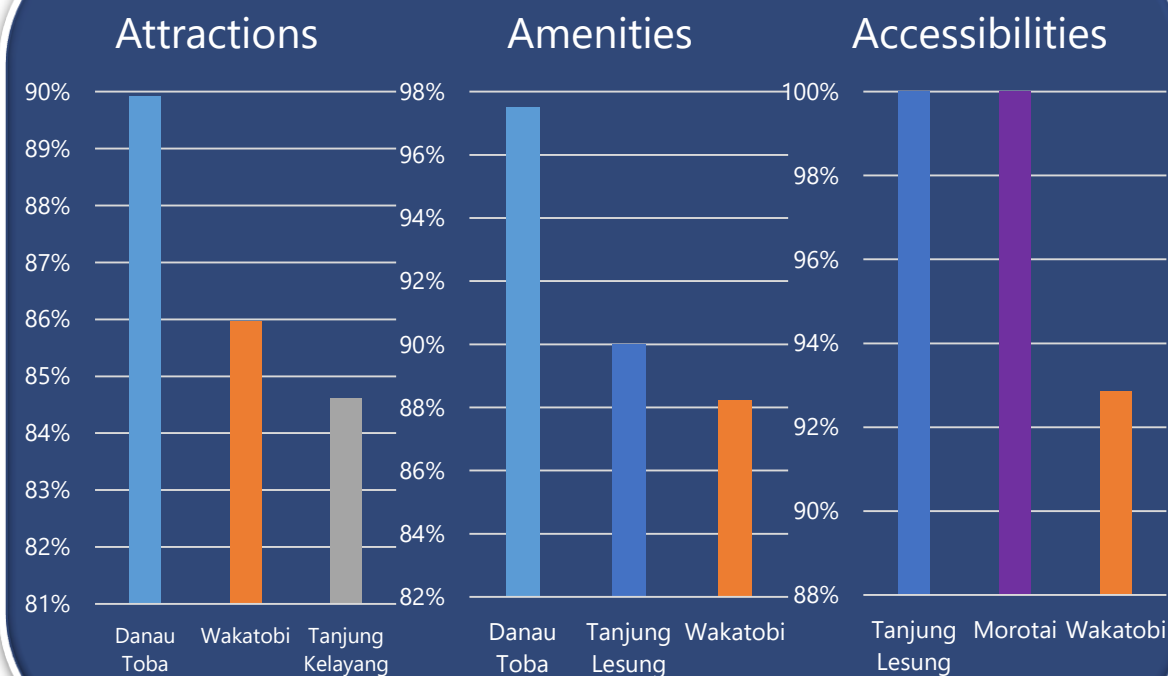
8

From the extracted sentiments we can rank the tourism destinations by its shares of positive vs negative sentiments. We can also further scrutinize the data and rank by its 3A aspects. From sentiments perspectives, Indonesia's tourism destination ranks in all the top 10.

Top 10 Tourism Destination by Shares of Positive Sentiments 2017-2021



Top 3 Tourism Destination in 3A Aspects by Shares of Positive Sentiments 2017-2021





RESULT – SENTIMENT SENTENCES

9

From the review sentences we can see the strengths and weaknesses of each tourism destination and use them to analyze aspects to promote and to improve about for each of the tourism destinations.

Danau Toba Sentiment Sentences



- 1. this area is an absolute beauty.
- 2. The lake and its surrounding hills are beautiful just to watch.

- 1. What is lacking is the infrastructure to make this place a real tourist spot.
- 2. Also, cleanliness must be improved.

Wakatobi Sentiment Sentences



The care was always excellent, diving places could be gotten in the short term.

Ferry schedule/service is a bit tough due to changing weather and local conditions.

Tanjung Kelayang Sentiment Sentences



This beach is quite managed neatly, because the beach is clean, on the beach there are also many boats that can be rented for Hopping Island

unfortunately the boat rental is quite expensive 500rb / boat with the contents of max 7 people



Conclusion

1. TripAdvisor API is a powerful tools to extract large amount of reviews that can be utilized to discover strengths and weaknesses of tourism destinations.
2. Indonesian tourism destination generally have better positive sentiment shares compared to peer countries tourism destination.
3. The top 3 tourism destinations by positive sentiment shares are Danau Toba, Wakatobi, and Tanjung Kelayang. From the reviews extracted we can formulate better strategies to promote and improve the 3A aspect of each tourism destination qualities.

Future Works

1. Exploration of other travel platform reviews e.g Yelp, Google Places, Booking.com, etc.
2. Expansion of aspects and sentiment keywords.
3. Accommodation around tourism destinations review extractions.
4. Sharia tourism specific review extractions.

A world map is shown in the background, colored with a gradient from light blue at the top to dark red at the bottom. The text "Thank you!" is centered over the map.

Thank you!

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Mailbot – optimising the process of answering statistical queries¹

Daphne Aurouet, Nina Blatnik, Samo Boh, Andrea Colombo, Almir Delic,
Jordi Gutiérrez, Gavril Petrov and Kristine Rikova,
European Central Bank

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



EUROPEAN CENTRAL BANK

EUROSYSTEM

MailBot

Optimizing the process
of answering statistical
queries

15/02/2022, IFS workshop

Daphne Aurouet, Nina Blatnik, Samo Boh, **Andrea Colombo**,
Almir Delic, Jordi Gutiérrez, **Gavril Petrov**, Kristine Rikova



Current process for handling statistical queries



Statistical Information Request

Manually assign responsible team

| From | Subject | Received | Categ... |
|--|---------|----------------------|----------|
| Categories: Media (1 item) | | | |
| Rikova... RE: URGENT: Statistical query from the media | | Mon 19/10/2020 20:07 | Me... |
| Categories: Follow up (5 items) | | | |
| Statisti... FW: [EXT] AW: Input needed: RESC - Real Estate Statistics - Com... | | Mon 19/10/2020 21:28 | Foll... |
| Statisti... FW: Input needed: IDCM contributions to growth rate (#4 - 1163... | | Mon 19/10/2020 16:40 | Foll... |

Aspects to improve

1. Slow identification; sometimes wrong BA identified



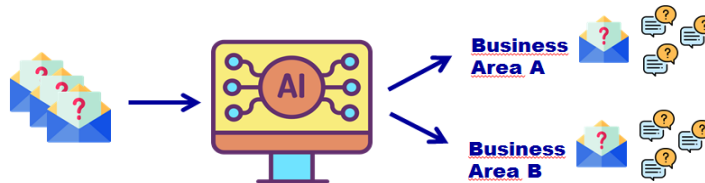
2. Similar queries answered multiple times



3. Manual handling of internal procedures

Solution

1. Classification of BAs
2. Identifying similar queries
3. Automatisation of procedures



Picture sources:

<https://www.ranzey.com/>

<https://www.synthesizerguide.com/synthesizers/how-to-choose-a-synthesizer/>

<https://www.flaticon.com>

Knowledge database



Simple structure → Question + Answer + Responsible BA

Text pre-processing and feature extraction

Data preparation

- Assigning IDs
- Parsing of emails – picking up original question and final reply
- Remove confidential information – greetings, signatures
- General string cleaning

NLP preprocessing

- Strip punctuation
- Tokenize
- Remove stopwords / other email vocabulary
- Lemmatizing
- Keep only nouns, pronouns, verb

Classification of BAs

Subject: Average exchange rates

I would like to know the annual average exchange rate for the year 2019, for Canadian dollar to Euros and American dollar to Euros.



92%

ESSA Exchange rates statistics

4%

Monetary statistics

1%

Market Data Provision

XGBoost

Continuous Bag of Words (CBOW) Model

Stochastic Gradient Descent

Extremely randomized trees

BERT

C or R NN

SVM

Enhanced Naïve Bayes



Extremely Randomised trees



H
E
U
R
I
S
T
I
C
S

Similarity of queries

Subject: Average exchange rates



I would like to know the annual average exchange rate for the year 2019, for Canadian dollar to Euros and American dollar to Euros.



QUERY:

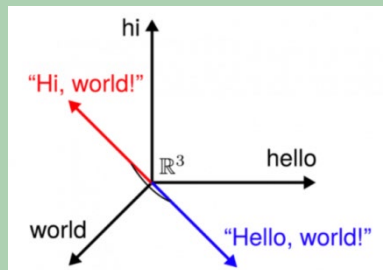
I would like to know the annual average exchange rate for the year 2018, for Canadian \$ to Euros and American \$ to Euros

ANSWER:

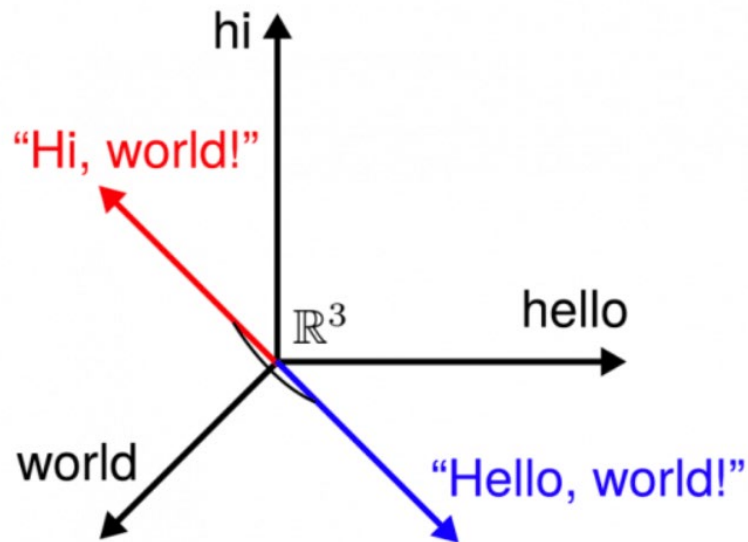
Please find exchange rates data at the following [link](#).



Cosine similarity



Similarity of queries: base model



Cosine Similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Source: machinelearningplus.com,
Wikipedia

Application demo

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Central bank communication: what can a machine tell us about the art of communication? One size does not fit all¹

Joan Huang and John Simon,
Reserve Bank of Australia

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Central Bank Communication: One Size Does Not Fit All

Joan Huang and John Simon

Abstract

High-quality central bank communication can improve the effectiveness of monetary policy and is an essential element in providing greater central bank transparency. There is, however, no agreement on what high-quality communication looks like. To shed light on this, we investigate 3 important aspects of central bank communication. We focus on how different audiences perceive the readability and degree of reasoning within various economic publications; providing the reasons for decisions is a critical element of transparency. We find that there is little correlation between perceived readability and reasoning in the economic communications we analyse, which highlights that commonly used measures of readability can miss important aspects of communication. We also find that perceptions of communication quality can vary significantly between audiences; one size does not fit all. To dig deeper we use machine learning techniques and develop a model that predicts the way different audiences rate the readability of and reasoning within texts. The model highlights that simpler writing is not necessarily more readable nor more revealing of the author's reasoning. The results also show how readability and reasoning vary within and across documents; good communication requires a variety of styles within a document, each serving a different purpose, and different audiences need different styles. Greater central bank transparency and more effective communication require an emphasis not just on greater readability of a single document, but also on setting out the reasoning behind conclusions in a variety of documents that each meet the needs of different audiences.

Keywords: central bank communications, machine learning, natural language processing, readability, central bank transparency

JEL classification: C61, C83, D83, E58, Z13

Contents

| | |
|--|----|
| Central Bank Communication: One Size Does Not Fit All | 1 |
| Introduction | 4 |
| What is Effective Central Bank Communication? | 5 |
| Readability | 6 |
| Reasoning | 7 |
| Who is the audience? | 7 |
| Data | 8 |
| Survey design | 9 |
| Limitation of the survey | 11 |
| Response bias | 12 |
| Survey Results | 13 |
| Summary | 13 |
| Survey respondents | 13 |
| Survey responses | 14 |
| Correlation between readability and reasoning | 17 |
| Correlation with readability formula | 17 |
| Economists versus non-economists | 18 |
| Methodology and Model Data | 19 |
| Introduction to ML | 20 |
| Label paragraphs | 20 |
| Extracting text features using natural language processing | 21 |
| The Models | 22 |
| ML algorithms | 22 |
| Model training | 23 |
| Feature importance | 24 |
| Results | 28 |
| Evaluating a document over time: <i>SMP</i> overviews | 28 |
| Comparing documents with each other | 29 |
| The variation of readability and reasoning within a document | 31 |
| Conclusion | 32 |
| References | 33 |
| Appendix | 36 |

| | |
|---------------------------------------|----|
| Model Tuning Process..... | 39 |
| Feature selection process | 39 |
| Tuning parameters process..... | 39 |
| Top ten features for four models..... | 40 |
| Model Validation Results | 41 |
| Confusion matrix | 41 |
| ROC-AUC | 42 |

Introduction

Central banking used to be a rather secretive business. As Janet Yellen (2012) noted in a speech 'In 1977, when I started my first job at the Federal Reserve Board ... it was an article of faith in central banking that secrecy about monetary policy decisions was the best policy'. But times have changed. There has been a gradual evolution towards greater transparency in central banking practice and an increasing emphasis on the quality of communication (Eijffinger and Geraats 2006; Filardo and Guinigundo 2008; Dincer and Eichengreen 2009).

This evolution has been driven by two primary forces. The first is a literature that has highlighted how expectations are central to the efficacy of monetary policy. If central banks communicate effectively, markets are able to anticipate the policy implications, and should respond to information contained in new economic data when it occurs rather than disruptively when central banks make policy announcements (Hawkesby 2019). Reflecting this, countries that are most effective in this practice tend to experience less interest rate volatility and smaller reactions to monetary policy changes (Blinder *et al* 2008).

The second force is the transition towards independent central banks and the associated need for transparency in support of democratic accountability (Blinder *et al* 2001). As independent public institutions, central banks are ultimately accountable to the public. To support this accountability they need to reveal enough about their analysis, actions and internal deliberations so that interested observers can understand each monetary policy decision as part of a logical chain of decisions leading to some objective and, thereby, assess their performance (Woodford 2005; Bernanke 2010; Preston 2020).

Central banks do, however, confront a trade-off. Monetary policy is not simple. A comprehensive explanation may not be simple or easily understood, but a simple explanation may not be accurate. In practice, central banks have tended to provide more explanation over recent years. For example, the length of the RBA's *Statement on Monetary Policy (SMP)* has increased from just over 10,000 words, in its first iteration in 1997, to over 30,000 words in the latest issues. A review of the Federal Reserve Bank's communication also reveals a general trend towards longer and more complex documents, to the point where the reader requires a university-level education to understand the content properly (Davis and Wynne 2016; Haldane 2017). A related challenge is that it can sometimes be difficult to decide who a central bank's audience is. Is it, for example, market economists who spend their time interpreting central bank actions for their financial market clients? Or is it the general public who are deciding whether to take out a mortgage or to invest in some new equipment for their business? Or companies and unions deciding how they will approach wage negotiations?

Despite the increased emphasis on communication and the many questions in the area, there has been relatively little study of the communication quality of central bank documents and even fewer answers on what makes for effective communication in this area.¹ To fill this gap we analyse central bank communications using surveys and a novel application of machine learning techniques. We focus on how different *audiences*, in particular economists and non-economists, perceive the *readability* of

1 There is, of course, a huge literature on effective communication in general.

and the degree of *reasoning* in various economic communications. Ours is the first work that attempts to measure the degree of reasoning in central bank communication and, consequently, also the first that considers the relationship between readability, reasoning and audience. We discuss the reasons for this particular delineation, and related work, in more detail in the next section.

We have three main results. First, we find that simple readability indices, such as the Flesch–Kincaid (FK) grade level, are barely correlated with individuals’ survey ratings of text readability. This suggests that a focus on simple readability metrics, as is common, may fail to increase a broad measure of readability. Furthermore, we find that the readability of a text is generally uncorrelated with the degree of reasoning in the text. Thus, a focus on readability alone runs the risk of undermining the achievement of transparency to the extent that it de-emphasises the importance of the content of a document.

Second, the way people comprehend a document depends on their knowledge of economics. We find that there is no correlation between the way the economists and non-economists in our sample perceive the readability and reasoning of the same piece of text. Our machine learning results reflect this observation: the textual elements associated with more readable paragraphs vary between economists and non-economists. Put another way, one needs to emphasise different techniques when writing for economists and non-economists; one size does not fit all.

Finally, there seems to be a trade-off between readability and reasoning – at least within a given paragraph. We find, for example, that the introductions of documents tend to be more readable but contain less reasoning while conclusions tend to be less readable but contain more reasoning. This highlights that different parts of a document have different objectives and it would be difficult to achieve these multiple objectives with a single style of writing within a single paragraph. Importantly, the application of any single metric to an entire document will fail to capture the need for different emphases at different points. Consequently, we emphasise that text quality metrics – including our own – should be used to inform rather than prescribe.

While this is the first study that we are aware of that systematically assesses these 3 aspects of central bank communication, our results are individually unsurprising. In other respects, however, they are new. In particular, we have not seen any previous work that considers the various aspects of effective central bank communication simultaneously and, in particular, that acknowledges the trade-offs that exist between the various individual objectives. No one document or style of writing is best for every paragraph, audience or communication objective.

What is Effective Central Bank Communication?

As mentioned in the above, we investigate 3 main aspects, or qualities, of central bank communication: the ease of *reading* and the degree of *reasoning* as assessed by different *audiences*. We discuss the reasons for our choice of these particular lenses in more detail here, along with a brief discussion of the existing literature.

Questions about readability and audience are natural ones when thinking about communication and there is a large literature on these topics. For example, the much-

used FK grade level (Kincaid *et al* 1975)² embodies the concepts of readability and audience to the extent that it highlights how certain texts are more or less appropriate for different audiences based on their level of education.³ The topic of reasoning is also common in the central banking literature, although it is not labelled as such. In particular, the literature on central bank transparency argues that not only does central bank communication have to be understood; it needs to reveal a central bank's analysis, reasoning and thinking. For the purposes of this paper, we label this concept 'reasoning'. That is, we see transparency as combining the concepts of readability and reasoning. Finally, while issues of readability and reasoning are common in the central bank literature, there is surprisingly little explicitly on the topic of audiences; much of the literature implicitly approaches these issues from either the perspective of a trained economist or assumes that simpler communication is universally better. We discuss some relevant literature on these 3 topics in more detail below.

Readability

As implied above, there is a degree of fuzziness in the central banking literature. Some papers implicitly equate readability with transparency, while others treat the existence of information on policy objectives as there being transparency about the objectives. We would suggest that the existence of information is an example of providing the reasons while the manner of its expression is an example of readability – which could be either clear or incomprehensible.

Notwithstanding the fuzziness in the literature, there are a number of papers that explicitly consider readability (e.g. Haldane 2017). Among these studies, simple readability formulae, most notably the FK grade level, are commonly used. The principle of the FK grade level, and alternatives are very similar, is that longer sentences or words with many syllables make a paragraph more difficult to read and comprehend. While many researchers (Jansen 2011; Bulíř, Čihák and Jansen 2012; Luangaram and Wongwachara 2017) adopted this metric for its simplicity and objectivity, others (Redish 2000; Janan and Wray 2012) criticise it for its ignorance of communication content, of common stylistic elements and of format and text structure. Those elements are commonly considered more important to comprehension than the number of syllables in a word or the word count in a sentence (Janan and Wray 2012).

Importantly, easy reading is not an end in itself. More readable communications should lead to a better understanding of monetary policy decisions and less market shocks. Reflecting this ultimate objective, Fracasso, Genberg and Wyplosz (2003) investigated 19 central banks and found that more readable central bank monetary reports are indeed associated with smaller policy surprises. Similarly, Jansen (2011) and Davis and Wynne (2016) found that more readable central bank communication

2 The FK grade level is calculated as:

$$0.39 \left(\frac{\text{number of words}}{\text{number of sentences}} \right) + 11.8 \left(\frac{\text{number of syllables}}{\text{number of words}} \right) - 15.59.$$

3 Although an implicit assumption in much of the literature, and certainly practice, is that a lower grade level is better regardless of the topic or audience. We touch on this implicit assumption later where we note that communication is multifaceted and more of one quality in writing can lead to less of another.

helped to reduce financial market volatility. So, despite debate about the most appropriate measures of readability, it seems that some central banks have managed to develop more effective communication practices – at least as measured by financial market volatility. But, the fact that readability and financial market volatility still varies across central banks and countries suggests that this is not a simple task. In part, this is because communication quality depends on more than just readability.

Reasoning

Effective communication also depends on conveying meaningful and useful information.⁴ For independent central banks, accountability relies on a central bank being transparent about why it thinks what it does. Blinder *et al* (2001) argue that the relevant content should include central banks' economic analyses, actions and internal deliberations, so that the public is clear about what it is trying to achieve, how it goes about doing so, and its probable reactions to the contingencies that are likely to occur. This allows people to form accurate expectations about how the bank will act in the future, which can increase the effectiveness of monetary policy.

Few studies have assessed the nature of the content in central bank communication. Of those that do, they mainly focus on the existence or quantity of certain public information rather than assessing the quality of that information. For example, Fry *et al* (2000) developed a transparency index for 94 central banks by calculating the average of 3 elements: whether the central bank provides prompt public explanation of its policy decisions, the frequency and form of forward-looking analysis provided to the public, and the frequency of bulletins, speeches and research papers. Eijffinger and Geraats (2006) adapted this index to 5 dimensions: political transparency (openness about policy objectives), economic transparency (openness about data, models and forecasts), procedural transparency (openness about the way decisions are taken), policy transparency (openness about the policy decisions) and operational transparency (openness about the implementation of policy actions). Similar studies include Bini-Smaghi and Gros (2001), de Haan, Amtenbrink and Waller (2004) and Dincer and Eichengreen (2014). Implicit in these studies is that the idea that, for example, simply holding a press conference or publishing a model forecast is a demonstration of transparency. But, just because someone asks you a question, it doesn't mean you have to answer it and giving a press conference does not necessarily mean you are being transparent. For this reason, we want to probe the content of communications more deeply to see if we can identify the degree to which they explain the reasons behind decisions. We believe ours is the first paper to attempt something like this and is one of the novel contributions of this work.

Who is the audience?

Central banks communicate with a wide variety of audiences including economists, financial market participants, politicians, the media, and the broader public. Each has different needs for information. Economists, who understand the economic data and

4 A simple example may help illustrate the point. 'The cat sat on the mat' is a very clear sentence. It is, however, purely factual and leaves a lot of information out. In contrast, 'The cat sat on the mat because it was warm' is still clear but now provides information on why the cat sat on the mat. This information could, for example, allow a reader to form expectations about what the cat might do in the future.

models better, are more likely to be interested in technical details about forecasts, while journalists and politicians may like to know more about the bottom line.

Reflecting this diversity of audiences, central banks have adopted a variety of communication strategies. One common strategy is to communicate via different channels. For example, the RBA's quarterly *SMP* provides a comprehensive economic summary that helps particular audiences, especially economists and financial market participants, understand the economic forecasts. RBA speeches tend to have a broader and more varied audience than monetary policy statements, but frequently contain similar information. A slightly different approach adopted by the Bank of England has been to add a visual summary to its *Inflation Report*. The visual summary includes the same key information as the traditional *Inflation Report*, but is written in less-technical language and contains a much heavier emphasis on visuals as a means of conveying information. The visual summary has been found to improve the comprehension of messages delivered in the *Inflation Report* for both members of the general public and economics students (Haldane and McMahon 2018).⁵ Other publications, such as bulletins and research papers, may provide indirect insights into central bank thinking to more academic audiences. There are also an increasing number of central banks that use social media (such as Twitter, YouTube, LinkedIn, etc) to provide information and target their audiences in more accessible ways (Bjelobaba, Savic and Stefanovic 2017).

Nevertheless, despite a variety of approaches that reflect implicit views about audiences and needs, the audience of communication is the least studied aspect of the 3. Born, Ehrmann and Fratzscher (2011) found that financial stability reports and speeches and interviews have different effects on financial stability. But it was not clear how the various audiences of those products, and how well the products targeted their audiences, affected the results. The existing literature tends to take the audience as given. For example, when assessing the transparency of communications, Fracasso *et al* (2003) used economics PhD students to rate central bank reports. This choice implicitly defined the audience they were considering. We show in Section 4.4 that the choice of audience can affect how effective a particular communication is judged to be. What works for some audiences may not work for other audiences.

Data

As noted above, there are 3 main dimensions to central bank communication that we are interested in: what is being communicated, how clearly it is being communicated and who it is being communicated to. The most obvious way to gather this data is the one we choose here: we ask a variety of people to rate the ease of reading and degree of reasoning of economic communication. More specifically, our data consists of 1,000 paragraphs of economic communication that survey respondents rated for their ease of reading and their degree of reasoning.⁶ We collect this data using an online survey that was completed by staff at the RBA with varying levels of economic

5 Haldane and McMahon (2018) also found that the visual summary of the *Inflation Report* improved economics students' reported perceptions of the Bank, but this is not the case for the general public. Furthermore, Bholat *et al* (2019) found that the increase in public comprehension is mainly due to the reduction of complexity of language rather than the inclusion of icons in the summary.

6 Due to the randomisation setting in the online survey there were some paragraphs that were not selected, thus only 833 paragraphs were actually rated.

training.⁷ For simplicity, we divide the audience into 2 broad groups: economists and non-economists. While there will undoubtedly be a range of understanding within those groups, and we do gather more fine-grained estimates of people's economic literacy, the largest differences in perception are likely to exist between these groups, so we focus on that in this study.

We discuss the reasons for various choices we made in the survey below. You can find a sample survey in this link (https://www.surveymonkey.com/r/EC_G1).⁸

Survey design

Sample paragraphs

Our survey asks respondents to rate the ease of reading and degree of reasoning in 10 paragraphs that are randomly selected from a set of 1,000. We chose to focus on paragraphs as these are a natural unit of written communication that are meant to present a single thought or idea that is also not too long and not too short. It was felt that single sentences would strip too much context from the writing and make evaluation of the readability and reasoning more difficult.⁹ On the other hand, asking people to read longer bodies of text would increase the response burden – consequently reducing the size of our dataset – and magnify the difficulties associated with converting the text into structured data.

The corpus of 1,000 paragraphs was selected randomly from a large number of publications from different sources including both central bank and non-central bank documents. This was done to ensure that our sample paragraphs include a variety of writing styles and economic topics and, thus, could provide a good amount of variation in the data. Given our focus on RBA communication, half of the sample paragraphs are from RBA publications, which include the *SMP* (2006–19), speeches (2018 and 2019), and *Bulletin* articles (2017–19). Another 20 per cent of the sample is from Bank of England (BoE) publications, including the *Inflation Report* (2014–19) and speeches (2019). We chose to include writing from another central bank as a way of including a different style of writing in our sample while keeping the underlying content relatively similar. The remaining 30 per cent is from non-central bank documents, including a number of reports published by the Grattan Institute, an economic policy think tank, and various articles from *The Economist*. These

- 7 The sample of RBA staff was a sample of convenience. Notwithstanding the non-representative nature of the sample, it had some useful aspects. The first was that, given we were trusted by the recipients, we obtained a higher response rate than would be the case with a survey sent to the general public. Indeed, we achieved a response rate of almost 70 per cent that would be unheard of in a survey sent more broadly. Also, because the issue was of particular interest to the respondents, they should devote more effort to providing accurate responses – leading to a higher quality dataset. Finally, the fact that the sample is non-representative is, for our purposes, not a particular problem. The primary requirement is that our sample include a range of different 'audiences' – which, because of the diversity of staff at the RBA, it does.
- 8 You may note that the concepts in the survey are referred to as 'clarity' and 'content'. This reflects the fact that earlier versions of this work used the label 'content' rather than 'reasoning' to refer to the extent to which particular text revealed the author's thinking or point of view. On the basis of feedback received on an earlier draft, we decided that the label 'reasoning' better captured the particular aspect of a text's content that we were focusing on and 'readability' better captured the ease of reading.
- 9 As it is, a tendency for some people to write very short or even one sentence paragraphs did create some problems.

documents allowed us to include a wider variety of economic topics as well as writing styles in our training sample. See Table 1 for more details.

| Sample Paragraphs by Source | | | Table 1 |
|---|-------------------------------|----------------------------|----------------------|
| | Number of paragraphs selected | Percentage of whole sample | External or internal |
| RBA publications | 500 | 50 | Internal |
| <i>Bulletin</i> articles | 100 | 10 | |
| Speeches | 100 | 10 | |
| <i>Financial Stability Report</i> | 50 | 5 | |
| <i>SMP</i> | 100 | 10 | |
| Overview/Introduction | | | |
| <i>SMP</i> main body | 50 | 5 | |
| <i>SMP</i> boxes | 100 | 10 | |
| BoE publications | 200 | 20 | External |
| <i>Inflation Report</i> introduction | 50 | 5 | |
| <i>Inflation Report</i> main body | 50 | 5 | |
| Speeches | 100 | 10 | |
| Other economic publications | 300 | 30 | External |
| <i>The Economist</i> ^(a) | 200 | 20 | |
| Grattan Institute ^(b) | 100 | 10 | |
| Notes: A full list of these paragraphs are available in the online supplementary information | | | |
| (a) Paragraphs are extracted from articles published in the 'Finance & economics' section | | | |
| (b) Reports are downloaded from its website (https://grattan.edu.au) and only include those related to economic growth | | | |

Survey design

For logistical and sampling reasons, we divided the 1,000 paragraphs across 5 online surveys. Each survey presented a random selection of 10 paragraphs for the respondent to rate.¹⁰ Asking each respondent to rate 10 paragraphs helped to keep the response burden low while also allowing us to control for a degree of inter-rater variability – different people tended to have different default ratings. Respondents were asked to rate each paragraph on a scale from 1 to 5 on 2 aspects:

- Readability (ease of reading): how easy it was to read. Where 1 is a very hard to read paragraph and 5 is a very easy to read paragraph.
- Reasoning (what versus why): the extent to which the paragraph reveals the thinking, position or point of view of the author. Where 1 indicates a statement of facts (what) and 5 indicates that there is an obvious position being taken or explanation being given (why).

We measure readability using survey ratings rather than existing metrics of readability or reading time. This is because we want to capture a holistic measure of readability (that we can then analyse to see if it is correlated with existing metrics)

10 The corpus of 1,000 paragraphs was divided into 5 randomly selected sets of 200 paragraphs. Each respondent would receive 10 randomly selected paragraphs from the subset associated with the particular survey they were sent.

rather than automatically assuming that shorter sentences, shorter words, or sentences that are read more quickly are necessarily 'clearer'.

The concept of reasoning is harder to capture. The definition used reflects the result of a number of pilots where we refined the question to best reflect the concepts discussed in the theoretical literature on central bank transparency. All of these emphasise the need for a central bank to explain its reasoning and framework to allow informed observers to predict future behaviour and test past behaviour against the central bank's stated framework. Our definition also reflects some overlap with a wider literature that focuses on analysing persuasive texts (Cohen 1984; Olsen and Johnson 1989; Azar 1999; Ferretti and Graham 2019), from which we drew a number of ideas.

Survey participants

Survey participants for this study are all working at the RBA, but in different areas including both economic policy-related areas and non-policy areas.¹¹ To assess their economic knowledge and working background we asked 3 simple questions:

1. How would you rate your overall level of economic literacy? (5 point scale from 'below average' to 'above average')
2. What level of formal economics education do you have? (scale from 'none' to 'post-graduate qualification')
3. Do you currently work in a job that involves economics in some way? (Yes/No)

Using these questions, we can test for the effect of economic knowledge on reader's judgements about the readability and reasoning of a given paragraph. Therefore, we sent the same survey (that is, a survey drawing from the same sub-sample of 200 paragraphs) to both economic policy and non-policy areas of the Bank in an effort to gather views from both economists and non-economists. In practice, the randomisation process meant that not every paragraph was rated by people from each area or by an economist and non-economist. In particular, some paragraphs were rated multiple times while some were not rated at all. We discuss the insights this duplication delivers and how we analyse these responses in Section 4. Other factors, such as age, gender, working experience (years) in economics, may also affect survey ratings. These were not included in our research for both privacy reasons and because we wanted to focus on high-level distinctions in our initial work. Notwithstanding this, the effect of these factors on the ratings would be a fruitful avenue of exploration for future work.

Limitation of the survey

While using a survey is an effective way to collect data in this study, we do face a number of limitations. Two particular ones we focus on here are selection bias and response bias.

11 Economic policy area generally refers to departments in Economic Group, Financial System Group and Financial Markets Group. Non-policy generally refers to the Information Technology Department (IT) and Business Services Group. Some departments, such as Note Issue, have both economists and non-economists working in them. As discussed below, we use the answers to questions 2 and 3 to distinguish between economists and non-economists.

Selection bias

The main selection issue is that the survey participants may not be representative of the general public or average central bank audiences. Indeed, this is undoubtedly the case. As such, the results should not be interpreted as indicating what a representative sample of Australians think about particular documents. Notwithstanding this, our primary objective is to obtain samples from different audiences with different levels of economics training. In this respect, the sample meets our needs. While all participants in our survey are currently working at the RBA, the degree of familiarity with monetary policy among non-economists at the RBA is very limited. Many respondents had relatively short tenures at the RBA, do not work in the policy-related areas, and do not have any economics training. As such, they are generally unfamiliar with economic policy issues. Conversely, among the economists surveyed, we would expect that they would be much more familiar with the ideas associated with central banking and represent a particularly specialised audience. To the extent that our primary purpose is to identify differences between the way specialist and non-specialist audiences understand various communications, this bias is beneficial in highlighting such differences more clearly than a more 'representative' sample might.

A related observation about the sample is that the economist sample may, in fact, be reasonably useful for understanding the way financial market economists perceive RBA communications. It is common for financial market economists to have spent some time working at a central bank or treasury. As such, we think the differences between the way economists at a central bank and economists in the private sector would understand particular communications are likely to be limited. Notwithstanding this, differences in the way the Bank of England publications were rated, discussed further below, suggest that the results reflect the *Australian* financial market economists might view the communications. This may reflect a learned familiarity with the RBA 'house style'. So, while Australian market economists are a relevant audience for RBA documents, UK market economists may perceive things differently and would be a more relevant audience to the Bank of England.

A second, less important, selection issue relates to the text samples chosen. Text selection bias may occur if sample paragraphs are not representative of the documents from which they are drawn. To the extent that this is an issue, it would limit the conclusions we could draw about the readability of or reasoning contained in the overall documents from the survey results alone. In practice, our main objective is to have a wide variety of paragraphs to train our machine learning algorithm rather than a representative sample of paragraphs. Nonetheless, given our selection was random, the survey averages should be a reasonable representation of the average characteristics of the various documents we sampled. In any case, while we present some summary statistics from our training sample, this is not the focus of our study and we do not draw particular conclusions about individual sources from these results alone.

Response bias

People's judgement about a given document can have subjective as well as objective elements. The subjective elements may vary based on people's personality, mood or opinion about the subject. For example, some survey respondents may be more generous or harsh than others and, thus, tend to give relatively higher or lower scores to the paragraphs they read. To control for this bias an effective (but not perfect) approach is to standardise the scores given by each person. That is, we calculate the

mean rating a person gives to each of the 10 paragraphs they rate and the standard deviation of their ratings, and convert their raw scores into normalised scores by subtracting the mean and dividing by the standard deviation. Implicit in this approach is the assumption that the average objective quality of the 10 paragraphs assigned to each respondent is the same. While this is unlikely to be precisely true *ex post*, it is certainly true in expectation because of the random assignment we use. More practically, we found that the additional noise that resulted from not making this normalisation made it very difficult for our models to fit the data well. That is, we believe the random variation in average paragraph quality between questionnaires was substantially less than the random variation in respondent's default or average ratings.

An additional question related to inter-rater variability and response bias is what to do with paragraphs rated by multiple respondents where the normalised (or un-normalised) rating differs. One way to manage this variability would be to use the average scores for the paragraph to measure text quality. An alternative would be to include each response in the dataset so that the same paragraph is associated with 2 (or more) different ratings. We discuss these 2 alternative below and make our choice – to use the average – based on the distribution of the observed survey responses.

Survey Results

Summary

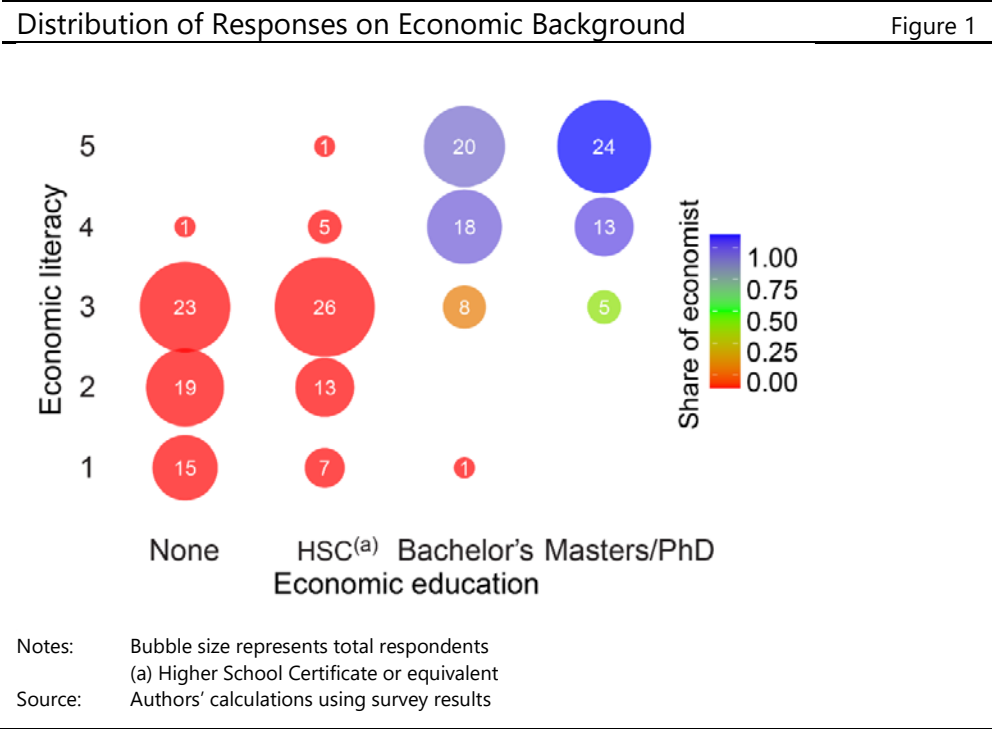
Survey respondents

The survey was sent to approximately 300 RBA staff and complete responses were received from 199. These respondents work in a range of areas including: IT, facilities management, the library, and various economic policy areas of the RBA. In terms of formal education in economics, about 45 per cent of our survey respondents had a bachelor's degree or above, 26 per cent had taken a high school course and about 30 per cent had not received any formal economic education. More details are in Table 2.

| Survey Respondents Background Description | | Table 2 |
|--|-------|---------|
| | Count | Share |
| By economic literacy (Q1) | | |
| 1 | 23 | 12 |
| 2 | 32 | 16 |
| 3 | 62 | 31 |
| 4 | 37 | 19 |
| 5 | 45 | 23 |
| By education background (Q2) | | |
| Bachelor's degree in economics or a related discipline | 47 | 24 |
| Masters or PhD in economics or a related discipline | 42 | 21 |
| High school economics course | 52 | 26 |
| None | 58 | 29 |
| By economic-related job (Q3) | | |
| No | 111 | 56 |
| Yes | 88 | 44 |
| Source: Authors' calculations using survey results | | |

Based on the answers to questions 2 and 3 we divide the 199 respondents into 2 broad groups: economists and non-economists. We define economists as those who have a university-level education in economics and whose work involves economics. Non-economists are those who are either working in a role that is not economics related or have no university-level education in economics. Using this division, 71 respondents are defined as economists, accounting for 36 per cent of the respondents, and 128 are non-economists (64 per cent).

We also asked each person to self-assess their economic literacy using a scale ranging from 1 to 5. Almost every economist assessed their economic literacy as somewhat above average (4) or above average (5). This is in line with their reported education background in economics as having a bachelor's degree or above. By contrast, most non-economists assessed their economic literacy as average or lower. More details on the distribution of economic knowledge in our sample are shown in Figure 1. The proportion of economists, as defined above, in each category of economic literacy and education background are shown in Figure 1 through the colour of the bubbles.¹²

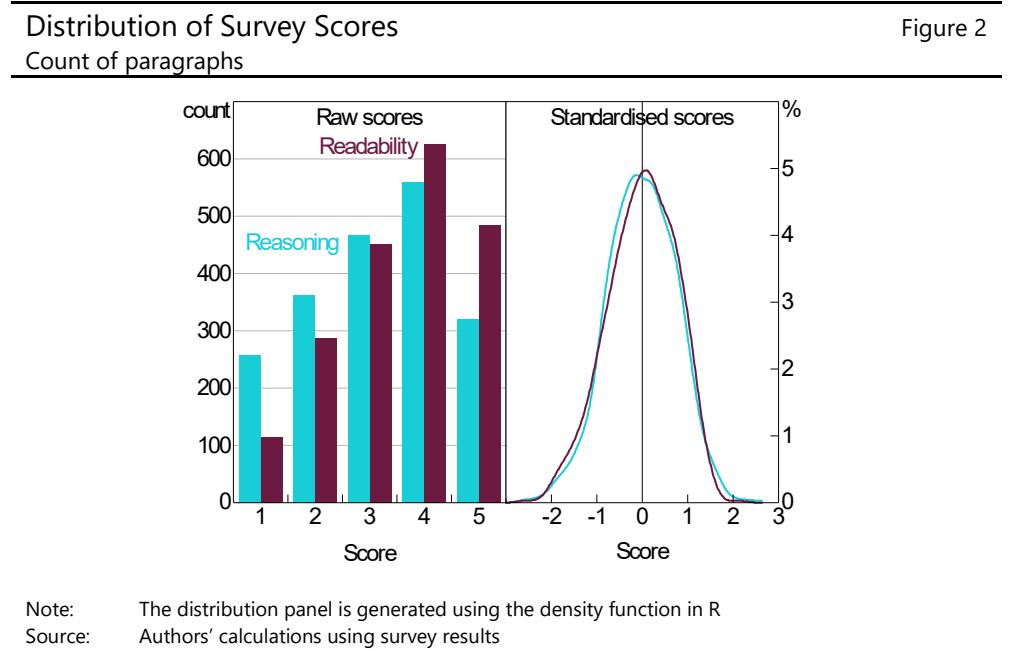


Survey responses

We received 1,695 valid responses covering 833 unique paragraphs. The left panel of Figure 2 shows the distribution of those scores for reasoning and readability. As can be seen, the modal score for both is 4 although the mean rating for readability is

12 Some responses appear anomalous (which is one of the reasons we adopted the definition we use). For example, one respondent reported holding a bachelor's degree in economics but having a very low economic literacy. We think this response (and a couple of others who rated their economic literacy as average despite also reporting holding a post-graduate degree in economics) points to the possibility that some respondents may have overlooked the term 'in economics' when answering the question on education background. That is, they hold a bachelor's or masters degree, but not in economics.

higher than for reasoning. As discussed above, however, different respondents appear to have different default scores – for example, some default to a score of 4 while others default to 3 – so we decided to standardise the scores. The distribution of standardised scores is shown in the right panel of Figure 2.



The fact that we have more valid responses than paragraphs partly reflects the design objective of getting both economist and non-economist ratings for the same paragraph. Of the 833 unique paragraphs, 465 were rated by both an economist and a non-economist, 53 by an economist only and 315 by a non-economist only. A second reason for the higher number of responses is that, due to the randomisation settings in the online survey, some paragraphs were rated by more than one person in each group. As shown in Figure 3, there were over 400 such paragraphs.

The overlapping ratings for a given paragraph, on the one hand, give us an opportunity to investigate the way people's ratings for a given paragraph vary. On the other hand, as mentioned above, they present a challenge in deciding the appropriate score for a given paragraph.

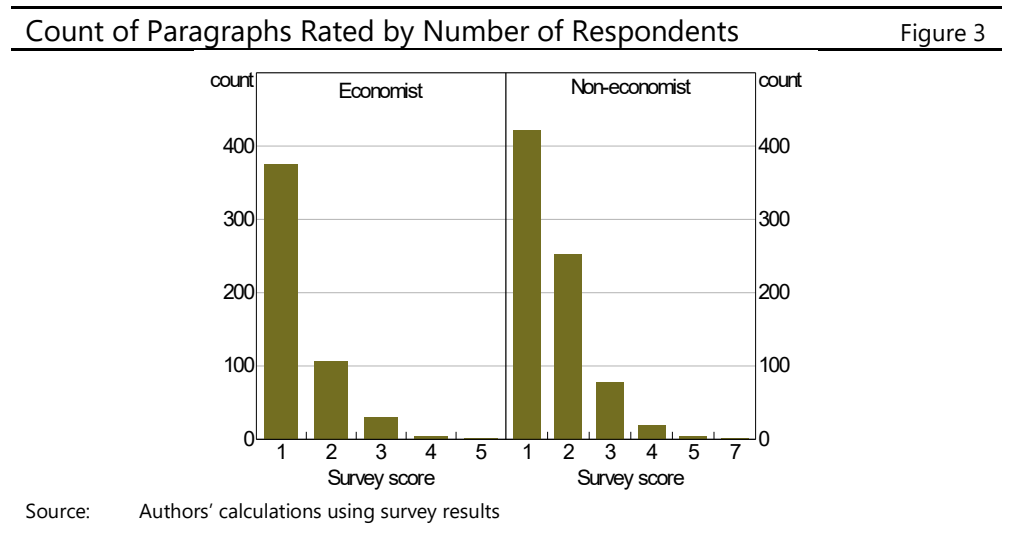


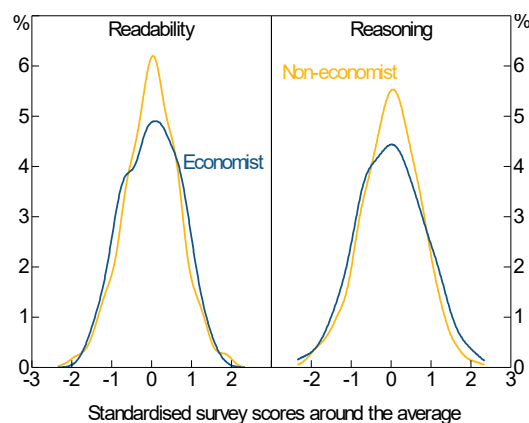
Figure 4 shows the distribution of scores around the average score for a given paragraph that is rated by 2 or more people ($S_i - \sum_{i=1}^n S_i / n, n \geq 2$). We can see that the scores generally cluster around the average – indicating that there is a degree of agreement across respondents about the quality of a given paragraph.¹³ The results suggest somewhat more disagreement among economists than non-economists and more dispersion in the ratings for reasoning than readability. The wider dispersion of reasoning scores is unsurprising given that reasoning is a harder concept to define and possibly more subjective in its evaluation. We leave the reader to make their own judgement about the reason for the greater disagreement among economists than non-economists.

As discussed above, one interpretation of the divergence is that each paragraph has one ‘true’ rating and each observation we have is a noisy signal about that true quality. Under this interpretation, taking the average rating would give the best signal about the true paragraph quality. An alternative interpretation is that different people – perhaps reflecting different backgrounds, knowledge or attitudes – have different interpretations of any given paragraph. Under this interpretation, divergence of ratings about a given paragraph is a signal that the paragraph is inherently ambiguous. That would suggest that each observation should be included in our dataset but, absent information about the reader that might explain the divergence in ratings across multiple readers, it would not be possible to correctly classify all of these paragraphs.

While exploration of the dispersion of ratings for a given paragraph could possibly reveal some subtle insights about effective writing for different audiences, it would also make our machine learning task considerably harder and require significantly more data than we have. Additionally, the single-peaked distribution of ratings for a given paragraph (Figure 4) suggest that assuming there is a true rating for any given paragraph is a reasonable assumption. Thus, we use the simple average of survey scores from multiple respondents as the final score for those repeatedly rated paragraphs.

Distribution of Scores on a Given Paragraph

Figure 4



Source: Authors' calculations using survey results

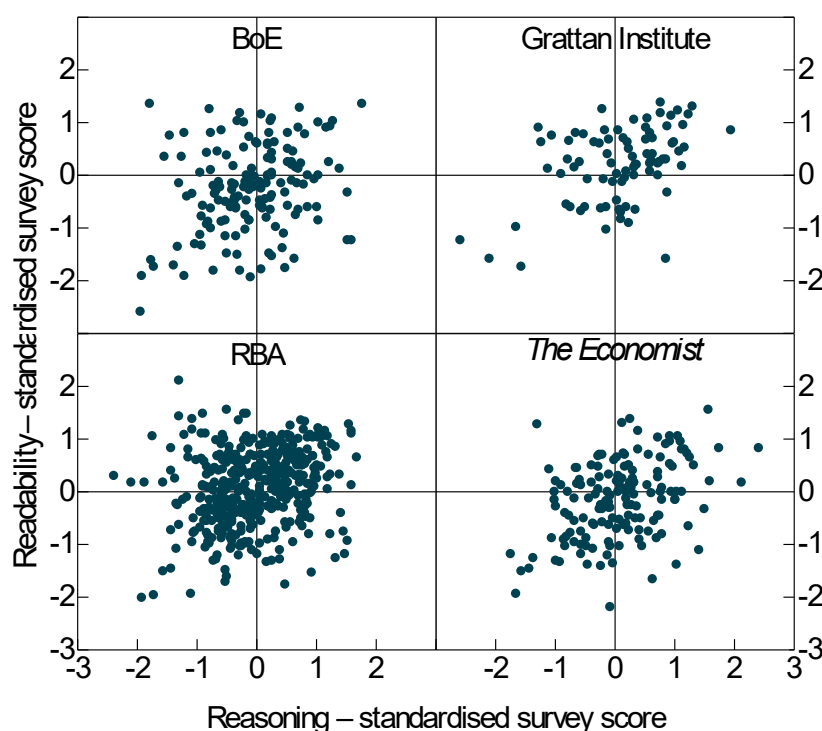
13 We tested normality using the Shapiro-Wilk test, and the results suggested that the distributions, except for the economist-readability data, are not significantly different from a normal distribution. For the full results, please refer to the online supplementary information.

Correlation between readability and reasoning

Figure 5 plots the distribution of reasoning and readability scores across sample paragraphs by text source. In general, all sources contain both high and low readability paragraphs as well as high and low reasoning ones and all combinations thereof. This wide distribution will be useful for training the machine-learning algorithm as it means we have examples of all possible types of paragraphs to help predict the quality of out-of-sample text.

Overall, what is most striking is the lack of correlation between readability and reasoning. There is only a slight positive correlation between reasoning and readability. The lack of close correlation between the scores emphasises how multidimensional writing is. This leads to one of our key observations: Trying to summarise the quality of a paragraph or document with any one metric must inevitably miss many important features of writing.

Correlation between Readability and Reasoning by Text Source Figure 5



Source: Authors' calculations using survey results

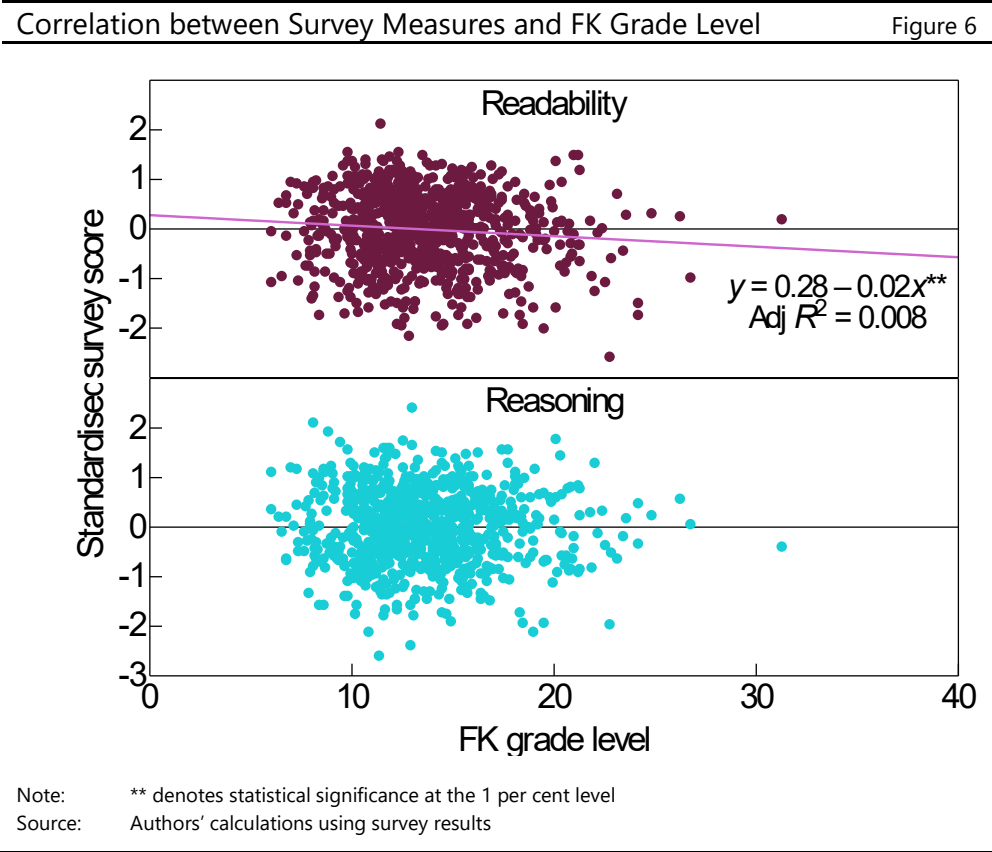
Correlation with readability formula

Simple readability scores, such as the FK grade level, have been widely used in the literature as a measurement of text quality. However, as noted in Section 2, there are a number of criticisms of their accuracy. Given that we have a direct evaluation of readability from our survey, it is interesting to look at the correlation between one of these measures, the FK grade level, and our survey responses. Figure 6 shows the correlation between our 2 measures of text quality and the FK grade level.

We can see a significant, but weak, correlation between the FK grade level and readability scores from the survey. The coefficient is of the expected sign and the

value of -0.021 indicates that an increase of 10 in the FK grade level is associated with a readability rating that is 0.21 standard deviations lower. However, the value of R^2 is only 0.008. This is a very low value, indicating that the FK score may be a poor indicator of the readability of any given sample paragraph. There is no significant correlation between the FK grade level and the reasoning scores.

These findings lend some support to the criticisms of, at least, the FK grade level but are likely much more widely applicable. In addition to the fact seen above that single metrics can miss important aspects of communication, widely used readability metrics may not even measure readability well.



Economists versus non-economists

A key focus of this project is to look at how different audiences understand the same piece of text. This question of audience is another dimension that is missing from simple readability metrics – people with similar education levels but different backgrounds will understand communication differently. Given our focus, we look at the difference between economists and non-economists.

Figure 7 shows the correlation between economist and non-economist ratings for a given paragraph – each dot represents a paragraph that was rated by both an economist and a non-economist. As can be seen, there is very little correlation.

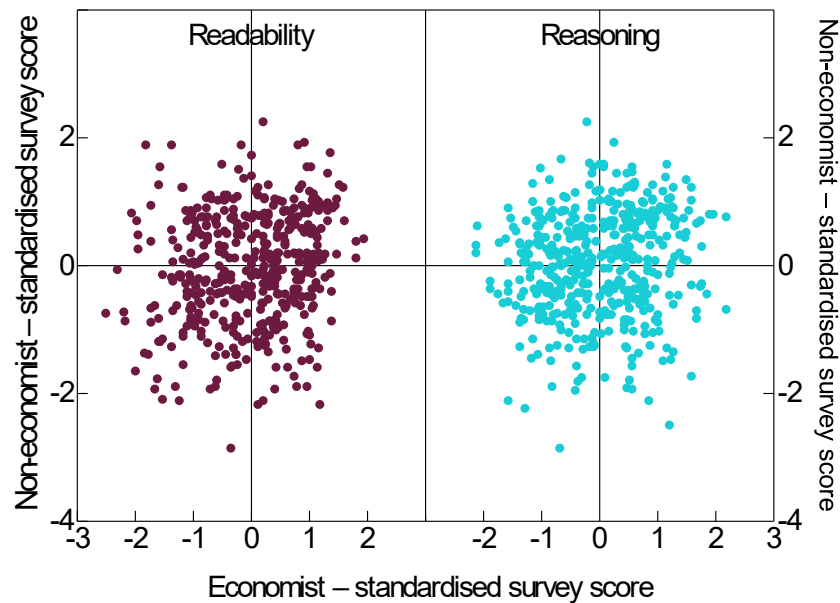
One possible explanation is that non-economists find the language used unfamiliar. As noted by Andy Haldane (2017): “‘Inflation and employment’ leaves the majority of [non-economists] cold. ‘Prices and jobs’ warms them up. ‘Annuity’ deep freezes [non-economists], whereas ‘investment’ thaws’. Nonetheless, while jargon and word choice may explain some of the difference, the variation is more likely to

arise due to the different ways people comprehend a paragraph based on their background knowledge. As noted by Goldman and Rakestraw:

Generally, in situations of high content knowledge, readers will be less reliant on structural aspects of the text than in low content knowledge situations because they can draw on preexisting information to create accurate and coherent mental representations. In low content knowledge situations, processing may be more text driven, with readers relying on cues in the text to organize and relate the information and achieve the intended meanings (Goldman and Rakestraw 2000, p 313).

Correlation between Non-economist and Economist Scores

Figure 7



Source: Authors' calculations using survey results

In other words, economists have sufficient background to understand the significance of pieces of information in a text without needing explicit pointers to their relationships. Conversely, non-economists may need the relationships between pieces of information spelt out explicitly through the structure of the text. A surprising implication is that non-economists might prefer longer sentences (with correspondingly higher FK grade levels) that provide the necessary structure for their understanding. They might find it harder to understand shorter sentences if these just stick to the facts and assume the reader can fill in the linkages. Alternatively, short sentences with sufficient explicit contextual information and a lot of attention to coherence between sentences might also achieve the same goal. More generally, this leads to a second key insight: one size does not fit all.

Methodology and Model Data

While the descriptive analysis above has highlighted a number of interesting features of economic communication, more insights can be gained through the application of machine learning (ML) algorithms. In particular, training an ML model to classify

paragraphs will allow us to consider a much larger range of paragraphs – and gain insights from them – than we could through the survey alone. A second benefit is that, by observing which features the ML algorithm uses to predict paragraph scores, we can better understand some of the features that make for a higher quality paragraph of economic communication.

Introduction to ML

While machine learning has become quite popular in recent years, there is a lot of overlap between ML techniques and traditional statistical and econometric techniques. For example, one of the most fundamental ML techniques is regression analysis, particularly logistic regression, which has long been used in more traditional statistical and econometric areas. In its basic form, logistic regression is used to classify data, based on a range of observable variables, into one of two categories. An example might be predicting whether someone will buy a house in a given year based on attributes such as their age, income, job, sex, relationship status and so on. Machine learning, however, generally approaches problems in different ways and, consequently, asks slightly different questions than those commonly tackled by econometrics.

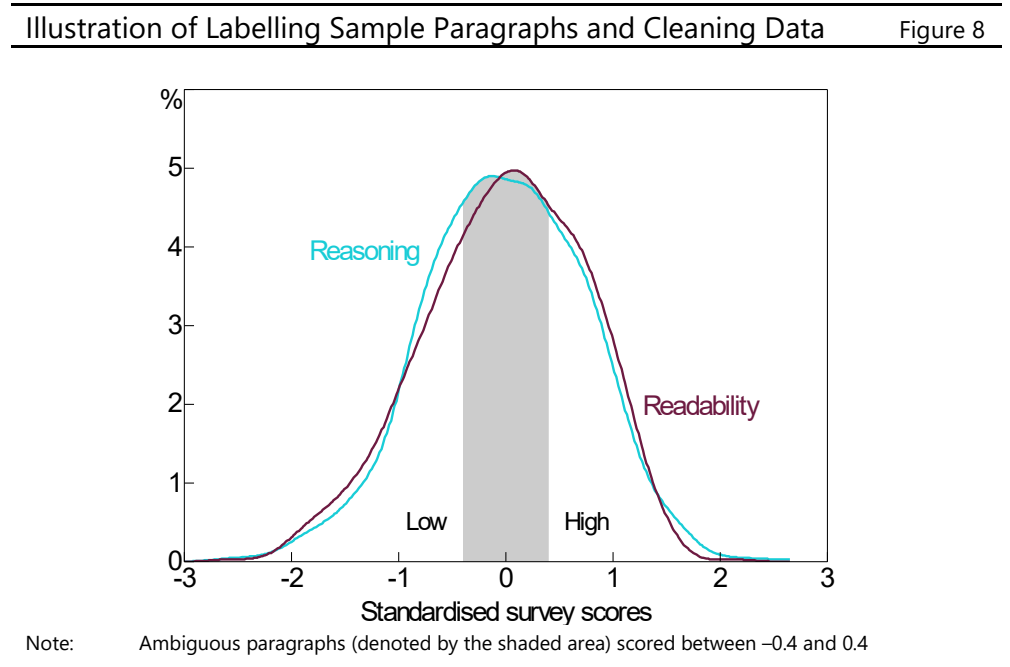
A common, though not universal, quality of ML problems is that there may be limited theory to guide the selection of appropriate explanatory variables, or features as they are called in ML models. Thus, they rely on big data and the associated techniques to learn underlying patterns that can then be used to predict other data rather than relying on theory in the way that more traditional statistics or econometrics tends to. As we have very limited existing theory that can guide us in selecting the set of features that will predict communication quality, we rely on ML techniques in this paper.

Within the field of ML there are a wide variety of techniques. At a high level, these techniques can be divided between supervised and unsupervised ML. In supervised ML the analysis starts with data that has previously been classified and labelled by experts and uses that data to ‘learn’ the basis for that classification. Unsupervised ML, such as cluster analysis, starts with unlabelled data and attempts to infer the underlying structure by identifying patterns. This study uses supervised ML techniques to build models that predict text quality based on the classifications provided by our survey respondents. Given our choice of this technique, there are 2 key elements to our approach that we discuss next: how we choose the labels for paragraphs, and how we convert the text into numerical data amenable to analysis.

Label paragraphs

While our survey asked people to classify paragraphs on a 5-point scale, we collapse these labels into 2 categories – ‘high’ and ‘low’. We do this because using this binary variable generates results that are more reliable. In practice, we also used a third implicit label – ‘ambiguous’ – that applied to paragraphs in the middle that we excluded from the training data. We found that paragraphs scored in the middle were very difficult for the algorithms to classify and the noise this introduced tended to degrade overall performance. Label noise is a well-known problem in machine learning and there are various techniques that have been proposed to deal with it (e.g. Karimi *et al* 2020). We adopt the simple technique of filtering out these

ambiguous labels. More precisely, we exclude paragraphs with a normalised magnitude between -0.4 and 0.4 . Excluding those ‘ambiguous’ paragraphs will reduce our sample size, but provide us with a higher quality dataset.¹⁴ This is illustrated in Figure 8.



Extracting text features using natural language processing

In addition to labelling our paragraphs, we need to convert the unstructured text into numerical data that can be analysed by the ML algorithms. That is, we need to compile a set of variables that numerically describe the individual paragraphs. A common approach to converting text into numeric data is using a dictionary mapping, also known as a ‘bag of words’ approach. This approach counts the frequency of particular words used in a sample of text but disregards grammar and word order. Text sentiment analysis, where the count of positive and negative words is calculated, is an example of this sort of approach.

The results using these approaches were, however, disappointing.¹⁵ Consequently, we investigated and ultimately included more syntactic approaches. The syntactic features of a sentence, rather than the particular words themselves, can have a large effect on readability. As noted by Haldane (2017), paraphrasing Strunk and White (1959): ‘In general, the readability of text is improved the larger the number of nouns and verbs and the fewer the adverbs and adjectives’.

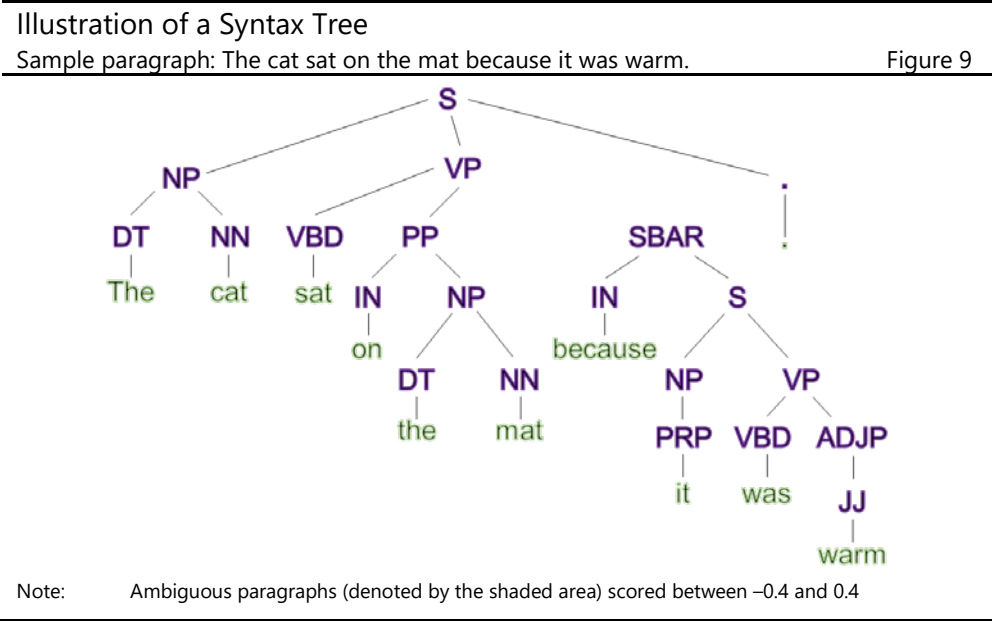
Therefore, we turn to a more advanced natural language processing approach that uses artificial intelligence to decompose text into its grammatical components.

¹⁴ For the readability model, 320 sample paragraphs are removed, 264 paragraphs are labelled as high, and 241 as low; for the reasoning model, 326 sample paragraphs are removed, 248 paragraphs are labelled as high, and 251 as low.

¹⁵ We tested model performance using a number of approaches, such as counting words (after removing stop words and lemmatisation) and mapping words to a clue words list, but found the model accuracy was not good enough to make any reliable predictions for out-of-sample data.

More specifically, we map each word in a sentence into a part of speech (PoS) using a PoS tagger and label each phrase using a parse tree.^{16,17}

As an example, we can decompose the sentence ‘The cat sat on the mat because it was warm.’ into a syntax tree as shown in Figure 9. It identifies ‘The’ as a determiner (DT), ‘The cat’ as a noun phrase (NP), and ‘because it was warm’ as a subordinate clause (SBAR) introduced by the subordinating conjunction (IN) ‘because’. We then use counts of the various parts of speech as our variables of interest. You can see the full list in Appendix B, along with an example of how a particular sentence is converted to numerical data.



The Models

We have 4 different datasets to model: readability for economists, reasoning for economists, readability for non-economists and reasoning for economists. Consequently we develop 4 separate models.

ML algorithms

There are a number of popular ML algorithms, each with their own strengths and weaknesses. To choose our preferred algorithm we first tested a number of popular ML algorithms on the sub-sample of economist data. These algorithms included the generalised linear model (GLM), the elastic net generalised linear model (GLMNET), the support vector machine (SVM), the gradient boost machine (GBM), and the random forest (RF). We chose to use the RF algorithm because it performed the best in our sub-sample testing and because it is relatively robust to overfitting.

16 We use the *openNLP* package in R (Hornik 2019) for this exercise.

17 Bholat *et al* (2017) deployed a similar approach to analyse central bank communication.

RF is a tree-based algorithm that predicts the classification of data by combining the results from a large number of decision trees (the forest part of its name). A decision tree is a flowchart-like structure that separates samples into 2 categories based on a sequence of yes/no decisions. To construct an individual decision tree, the algorithm first searches over all available variables and selects the variable that provides the best separation of the 2 categories as the top node. It then moves to the next layer and repeats the process to find the variables that give the best separation. The splitting stops when no further improvement can be made (Quinlan 1986). The RF algorithm builds its individual trees independently using a random sub-sample of the data and variables (the random part of its name).

Model training

In this project, to protect against overfitting, we randomly choose 75 per cent of our data as the training dataset to build the models and use the remaining 25 per cent as the validation dataset for testing model performance. A few approaches were used to improve model performance. First, we adopt an automatic feature selection method that selects the most relevant features for our model; including too many features may lead to overfitting. Second, the RF algorithm has many hyperparameters¹⁸ that affect model performance and we tune these parameters using a grid search approach. Please refer to Appendix C for details about the feature selection and parameter tuning processes.

As our models return a probability prediction (p_i)¹⁹, we convert p_i to a predicted class label using a threshold. We use the default value of 0.5,²⁰ so the prediction label for a paragraph is *high* if $p_i \geq 0.5$ and *low* otherwise.

$$class(paragraph) = \begin{cases} high, & \text{if prediction probability} \geq 0.5 \\ low, & \text{otherwise} \end{cases}$$

We evaluate model performance using 2 standard evaluation metrics: the confusion matrix and the ROC-AUC curve. A confusion matrix is a 2-by-2 table that is calculated by comparing the predicted labels with the actual labels from the validation dataset. The ROC-AUC curve yields a measure of how well the model separates the two classes of data.

For our models, the accuracy (calculated from the confusion matrix as the proportion of labels that are correctly predicted) is around 70 per cent. The AUC ranges from 0.55 to 0.6 for our models. We report the full test results in Appendix D. Overall these results are modest, our model has reasonable accuracy in predicting

18 For instance, *ntree* and *mtry* are 2 important parameters for the RF algorithm. *ntree* represents the number of trees that will be built and *mtry* is the number of variables that will be randomly sampled for each node in a tree.

19 In tree-based algorithms, the probability is calculated as the proportion of trees assigning a label of high to a given paragraph. For example, if there are 500 trees in an RF model and 300 of them rate an observation as 'high,' it returns a probability of 0.6.

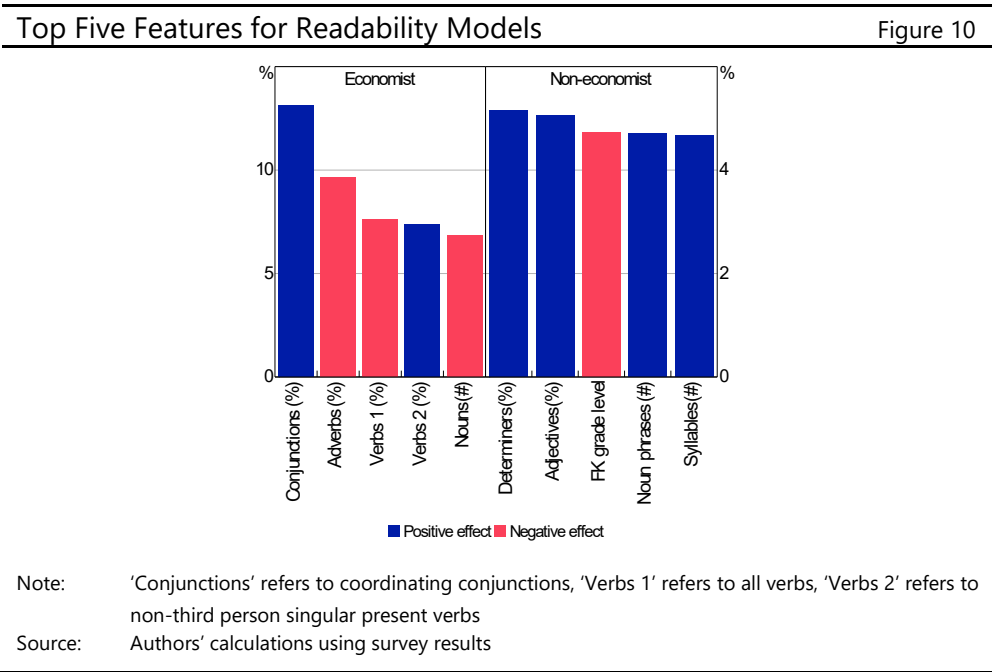
20 There are other ways to set the threshold and, if the data set is unbalanced – with more of one label than the other – the default 0.5 may not be a good threshold. For this study, 0.5 is a reasonable threshold as our datasets are roughly balanced (for the readability model, 264 paragraphs are labelled as high and 241 as low; for the reasoning model, 248 paragraphs are labelled as high while 251 are labelled as low).

whether a paragraph is more likely to be high quality than not, but does not yield definitive predictions about paragraph quality. Given that there is an inherent fuzziness to paragraph quality, we think it unsurprising that our algorithm can not cleanly separate high-quality paragraphs from low-quality paragraphs – we suspect humans would struggle to do so as well.

Feature importance

ML models are often considered to be ‘black boxes’ for their complex inner workings and plethora of opaque parameters. Our dataset has hundreds of features and it is often difficult to understand which features are driving the prediction accuracy of our models. One benefit of the RF algorithm, however, is that it has a built-in function that calculates the contribution of each feature.²¹ This helps to discern some of the inner workings of the black box. However, we must emphasise that the underlying models are nonlinear and complex, so one should not over-interpret the results presented here – they are meant to give a heuristic impression about the models. They are not a precise linear representation of the workings of the model in the manner of linear regression coefficients.

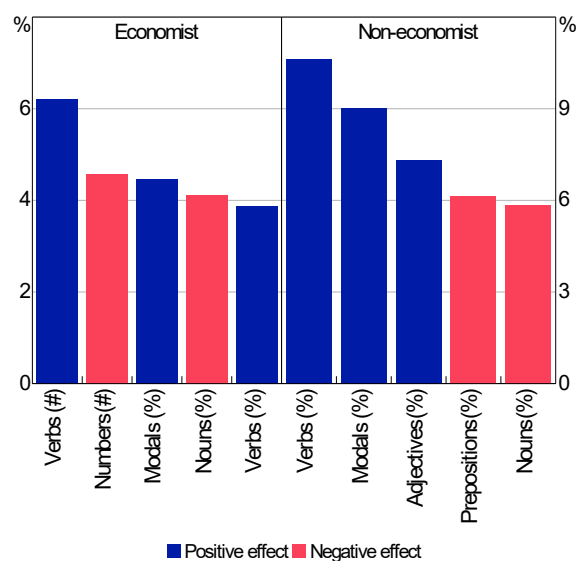
Figure 10 illustrates the top 5 features for the readability models, and Figure 11 shows them for the reasoning models. These top features are ranked based on how much information each variable contains to discriminate between the 2 categories.²²



Top Five Features for Reasoning ModelsFigure 11

21 The feature importance is extracted as a part of model outputs that is generated using the *caret* package in R. The importance value for each variable is calculated as the contribution of each variable based on the mean decrease in impurity (Gini) after removing this feature. Another way to calculate the feature importance is based on the mean decrease of accuracy.

22 The exact ranking for each variable may vary with different settings of parameters. However, the lists of top 5 variables for the models in this study are relatively stable based on our experiments.



Note: 'Prepositions' refers to preposition or subordinating conjunction
Source: Authors' calculations using survey results

It is not typically possible to determine whether the effect of these variables on the results are positive or negative. This is because RF models are capturing complex nonlinear relationships in the data. Notwithstanding this, we can get an idea of whether the average effect of a particular variable is positive or negative. To calculate this we run the models for a sample of 1,424 paragraphs and remove the top 5 variables one by one and regenerate the model prediction. Based on the difference between the 2 results, we classify the partial effect of a variable as positive or negative.²³ There are some similarities in the top features list between the 2 reasoning models but surprisingly little across the readability models.

Looking at the readability models first, we see that the FK grade level appears in the top 5 features for the non-economist model. However, the number of syllables, which contributes to the FK grade level negatively, appears with a positive sign. More generally, the model suggests that non-economists prefer paragraphs with more noun phrases, adjectives and determiners. Conversely, simple metrics don't show up in the economist model. The top feature is the proportion of coordinating conjunctions²⁴ and there seems to be a preference for paragraphs with fewer nouns and adverbs. One possible explanation for this difference is that economists hold more economic knowledge and, thus, may rely less on linguistic clues in the paragraphs (such as adjectives and determiners) to understand the importance of and

23 The partial effect of a variable on the target variable is positive if the prediction probability for 'high' is lower after removing this variable, and otherwise negative. We should not draw a conclusion on the effect of each feature on the final prediction results as the relationship between a feature and the output from the RF model is often nonlinear.

24 A coordinating conjunction is a word that joins two parts of a sentence. According to the 'Part-of-Speech Tagging Guidelines for the Penn Treebank Project' (Santorini 1990), the coordinating conjunction list includes *and*, *but*, *nor*, *or*, *yet*, as well as the mathematical operators *plus*, *minus*, *less*, *times* (in the sense of 'multiplied by') and *over* (in the sense of 'divided by'), when they are spelled out. The proportion of coordinating conjunctions is also an important feature for both readability and reasoning models of non-economists. As shown in Table C2, this features ranks seventh for the non-economist readability model and tenth for the reasoning model.

relationships between concepts. As noted by Gilliland (cited in Janan and Wray (2012, p 1)):

... in a scientific article, complex technical terms may be necessary to describe certain concepts. A knowledge of the subject will make it easier for a reader to cope with these terms and they, in turn, may help him to sort out his ideas, thus making the text more readable. This interaction between vocabulary and content will affect the extent to which some people can read the text with ease.

There is more similarity in the reasoning models. In particular, both economists and non-economists identify more verbs and fewer nouns with higher reasoning. This is natural because the verb phrase generally denotes eventualities, processes and states, and the roles that participants play in the events described (McRae, Ferretti and Amyote 1997). That is, the kind of terms you would use when expressing an argument or point of view rather than presenting facts. In addition, modal words, such as *might*, *could*, and *should*, are also important for both reasoning models. Modal words are normally associated with persuasive writing, and are often treated as an arguing feature in the study of linguistics (Farra, Somasundaran and Burstein 2015).

These findings are not too surprising but it is worth noting that in preliminary work we tried just using word lists to identify whether an argument was being made (e.g. counting uses of words like 'because') and this approach was relatively unsuccessful. That is, we have found that understanding the grammatical function of a word is more valuable in classifying text than the particular word that is used. Or, more poetically, and in the timeless words of Led Zeppelin, using word lists is more error-prone 'Cause you know sometimes words have two meanings'.

To help gain a greater sense of how the model works in practice, Table 3 presents 2 sample paragraphs, one rated high and one rated low, for each model.

Sample Paragraphs with Model Prediction Results and Actual Survey Scores
(continued next page) Table 3

| Model | Paragraph | Model results (a) | Survey scores ¹ (b) |
|---------------------|---|----------------------|-----------------------------------|
| Economist-Reasoning | The big question is whether we should expect these quirks to endure. Once a way to make above-market returns is identified, it ought to be harder to exploit. 'Large pools of opportunistic capital tend to move the market toward greater efficiency,' say Messrs White and Haghani. For all their flaws and behavioural quirks, people might be capable of learning from their costliest mistakes. The rapid growth of index funds, in which investors settle for an average return by holding all the market's leading stocks, suggests as much. | High (0.90) | 4.5 (1.01) |
| | Most of the sectors that declined as a share of non-mining output were capital intense. Agriculture, forestry and fishing, electricity, gas, water and waste services, information, media and telecommunications, and rental, hiring and real estate services declined by nearly 3.5 percentage points of non-mining output. Manufacturing declined by almost ten percentage points of non-mining output. | Low (0.27) | 2.0 (-0.91) |

| | | | |
|-----------------------------------|--|----------------|----------------|
| Economist– Readability | While household dwelling investment continued to decline over the first half of the year, there have been signs in recent months of a prospective improvement, partly in response to reductions in interest rates. Private residential building approvals, dwelling prices and auction clearance rates have all increased. The overall demand for housing finance has been broadly stable over the course of the year and many home owners are taking advantage of lower borrowing rates to pay off their loans more quickly. | High (0.90) | 4.5 (1.35) |
| | In any event, there is no strong economic rationale for a different tax rate for small companies. While compliance costs are higher for small companies (relative to their profits), it makes little sense to compensate them via a differentiated tax system. A lower tax rate compensates small companies with high profits much more than those with lower profits, for instance, even though the relative compliance costs are larger for companies with lower profits. The Government should ensure that the small and large company tax rate is equalised over the next few years. | Low (0.39) | 2.0 (–1.58) |
| Non- economist– Readability | In 2017, Australia's net foreign currency asset position amounted to 45 per cent of GDP (ABS 2017b). Around two-thirds of Australia's foreign liabilities were denominated in Australian dollars, compared with around 15 per cent of Australia's foreign assets. Since 2013, foreign currency assets and liabilities have both increased as a share of GDP. Since the dollar increase in assets has been greater than that in liabilities, there has been an increase in Australia's net foreign currency asset position of around 15 percentage points of GDP. | High (0.70) | 4.5 (1.05) |
| | Looking at more detailed data on cross-border bank lending from the Bank for International Settlements, it is evident that cross-border lending by European banks both increased most rapidly going into the crisis and subsequently contracted most sharply. Given that financial stress was concentrated in industrialised economies it is also noteworthy that lending to other industrialised economies peaked earlier than lending to emerging markets, which was curtailed only much later into the financial turbulence. This pattern is also evident in the sharp reversal of (net) flows between the United States and the United Kingdom as a result of reduced cross-border lending by European banks headquartered in London as institutions sought to unwind their exposures. | Low (0.28) | 2.0 (–0.83) |

Notes: (a) Numbers in parentheses are the probability results from RF models – essentially the strength of the model's prediction; the label is high if the probability is equal to or greater than 0.5 and low otherwise

(b) For paragraphs that are rated by multiple readers we report the average score; numbers presented in parentheses are standardised survey scores

Overall, given that many of the identified features appear to make sense linguistically, at least based on our knowledge and brief reading of the linguistic literature, we are fairly confident that our model has identified meaningful features rather than latched on to idiosyncratic features that have little true explanatory power. A key observation is that, because each model emphasises different features, making paragraphs readable for both economists and non-economists is not simple. For example, the correlation between predicted readability for economists and non-economists is 0.54 in our sample. While there is some correlation it is not straightforward and one size does not fit all. That said, simple metrics such as the FK grade level don't seem to be a good guide to readability. They have little correlation for the non-economist readability model and none at all for the economist model. This implies that targeting a particular FK grade level is unlikely to improve readability for either group.

Results

With the models trained we now apply them to a number of economic documents to demonstrate how they can be used to evaluate a large body of work that would otherwise be time consuming to classify manually. Most of the document we focus on, such as monetary policy statements and speeches, are from central banks, but we also include a sample of 20 articles from *The Economist* for comparison.

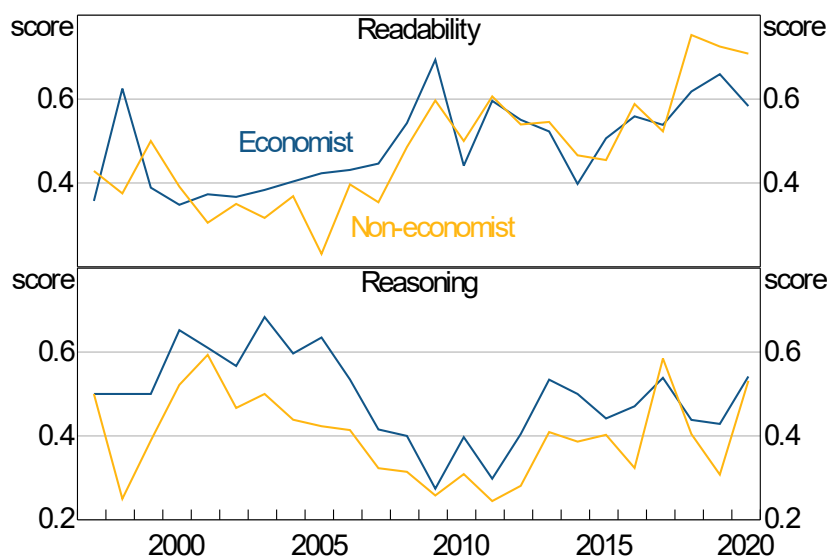
For modelling a document, we first break all documents into paragraphs using text mining tools and then convert each paragraph into a structured dataset that includes all variables shown in Table B1. Then, we predict the quality of each paragraph using our 4 models, and classify each paragraph as 'high' or 'low' for both readability and reasoning from economist and non-economist perspectives. Last, we measure the text quality of a document using the proportion of high-quality paragraphs in it:

$$\text{Document quality measure} = \frac{\text{count of paragraphs classified as high}}{\text{total number of paragraphs}}$$

Evaluating a document over time: SMP overviews

The most important documents that central banks use to communicate with external parties are typically the regularly released monetary policy reports. The RBA has published its *SMP* since 1997 and so the first texts we apply our models to are the *SMP* introduction section over the period from 1997 to 2020. This covers 87 issues of the *SMP* and 1,519 paragraphs. We choose the introduction/overview section because this section generally contains the explanation and justification for policy actions and, as such, is the most important section for understanding central bank policy. Other sections of the *SMP* tend to consist of more factual reporting of recent data. The results are shown in Figure 12.

Model Scores for Readability and Reasoning for SMP Overview Figure 12



Source: Authors' calculations using survey results

Our model results on readability, as shown in the top panel of Figure 12, suggest that the overview section of the *SMP* has become easier to read over time. Interestingly, our measure picks up more variation in readability over the years than the FK grade level (see Figure A1 for a comparison of the readability score for the *SMP* introduction and the FK grade level).

Conversely, the reasoning score has shown no particular trends over time. If anything, it has dropped in recent years. Indeed, there appears to be somewhat of a negative correlation between readability and reasoning with an obvious dip in reasoning around 2009 when readability scores jump higher. To the extent that transparency is affected by both readability and the degree of reasoning in documents, it does not necessarily follow that increases in the readability of the *SMP* have been associated with increases in transparency. While we can't make any statements about the absolute level of transparency in the *SMP*, these results suggest that evaluating the overall transparency of central bank documents requires a broader consideration than readability metrics alone can provide.

Comparing documents with each other

In addition to monetary policy statements, central banks also release other publications, such as speeches by senior staff, short articles and financial stability reports. In this section, we apply our models to some of these documents to see what they reveal about any variations in text quality across documents.

We choose a number of paragraphs from the Bank of England (BoE) *Inflation Report* introduction and boxes, RBA speeches and *SMP* introduction and boxes published in 2018 and 2019²⁵ and articles from *The Economist*²⁶. Figure 13 shows the results. The correlation between readability and reasoning is not significant in both panels, but the pattern is clearly different between economists and non-economists.

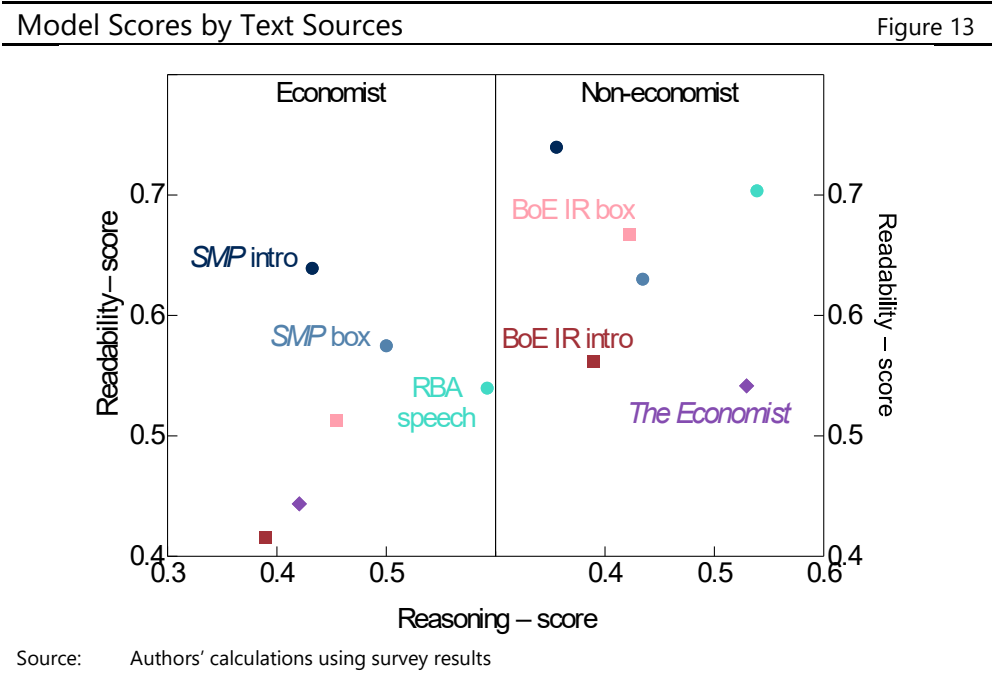
As assessed by economists, speeches have the highest reasoning rating but an average readability rating. Conversely, the introduction to the *SMP* in 2018–19 has a low reasoning rating but the highest readability rating. When assessed by non-economists, however, RBA speeches are found to have among the highest average readability and reasoning ratings. This may reflect the fact that spoken communication is different to written communication, but could also reflect the different objectives of these different documents. Speech givers seem to be communicating particular positions and arguments that are relatively clearer to non-economists, while the writers of boxes and the *SMP* seem to be more focused on communicating facts clearly.

Another interesting feature is the change in the relative ranking of the BoE samples between economists and non-economists. While RBA economists rated RBA documents more highly than BoE documents, non-economists rated BoE documents relatively higher and their ratings were less dispersed overall. This points towards a preference among RBA economists for the RBA 'house style'. We can't be certain, but

25 We restrict the *SMP* sample to the years 2018–19 to match the approximate time period covered by all the other sources considered. To the extent that the economic environment may affect the way content is communicated, this means that there is some comparability between the underlying documents, particularly those from the same institution.

26 We randomly selected 20 articles from the 'Finance & economics' section of *The Economist* that were published between 2019 and 2020.

given that topic and word choice do not affect our algorithms, this preference is unlikely to reflect greater familiarity among RBA economists with the topic matter of these publications – hence our suspicion that it reflects a ‘house style’ preference.

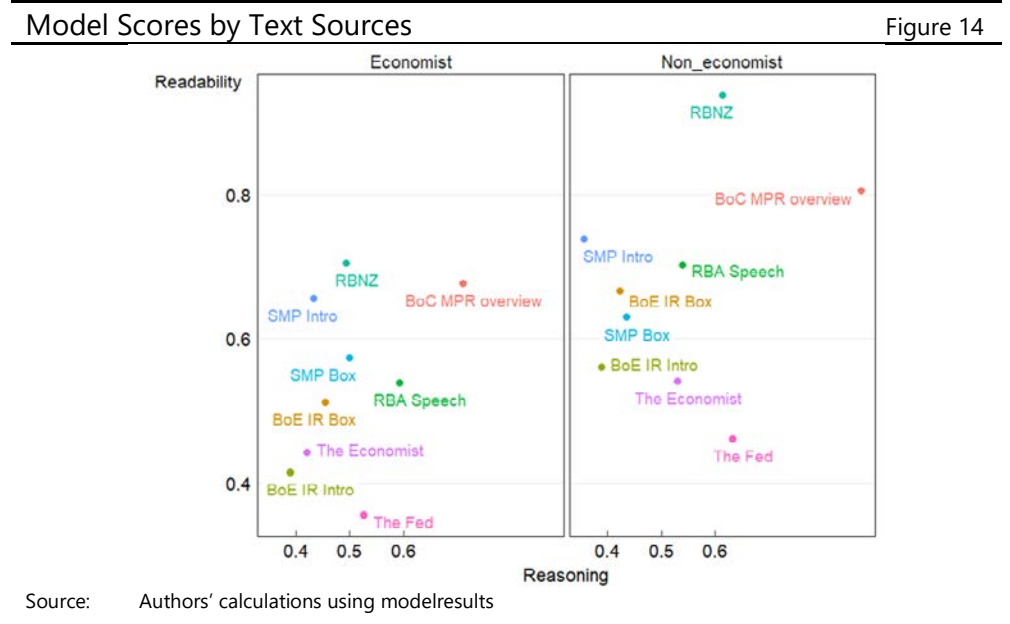


Finally, we see that *The Economist* is rated highly for reasoning by non-economists but not particularly highly for readability. While this reflects the fact that *The Economist* primarily presents analysis and opinions it does not seem to reflect its well-founded reputation for plain language. We see two possible explanations for this. One, our algorithm is reflecting a preference for a particularly Australian idiom or house style that *The Economist* does not conform to – which may also explain the low ratings from economists. Or two, by averaging the rating of paragraphs over a whole document we may be overemphasising the role of body paragraphs in a document and underemphasising the importance of introductory and concluding paragraphs. That is, the subjective assessment of a document’s overall quality may depend more heavily on the quality of the introduction and conclusion than our index does. We reflect on this point in the section below.

We have also scored some paragraphs from some central banks’ recent publications, including the Bank of Canada Monetary Policy Report overview (BoC MPR) published in 2020 and 2021, the Reserve Bank of New Zealand Monetary Policy Statement’s current economic assessment and key judgements (RBNZ MPS) and the Fed Monetary Policy Report summary (The Fed) published between 2019 and 2021²⁷. According to model results shown in Figure 14, economists and non-economists share similar views on the text quality of those documents. They both believe that the RBNZ MPS has the highest rating for readability but a moderate rating for reasoning. The Fed MPR overview, which has a similar reasoning rating to RBNZ MPS, has the lowest rating for readability. The BoC MPR overview is rated highly in both dimensions.

27 The Fed monetary policy reports are published biannually, while the other two are quarterly. The overview section in the BoC MPR is only available from report published after April 2020.

Notwithstanding these observations, the results are only preliminary and suggestive and are meant to be illustrative of the potential of these ML techniques rather than be definitive findings. Regardless, they re-emphasise our observation that: different documents are perceived differently by different audiences and this argues for clearly targeting one audience rather than attempting to reach multiple audiences with the one document.



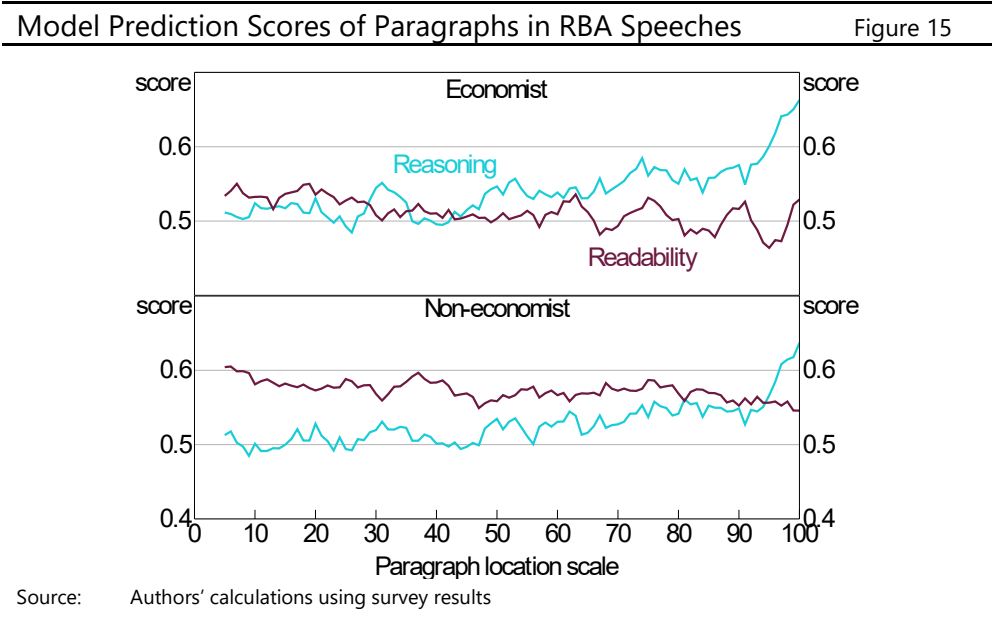
The variation of readability and reasoning within a document

So far, we have only assessed text quality differences at the aggregate level across documents, but we have not analysed text quality within a document. To investigate this aspect of communication we analysed 99 speeches that were given by RBA senior officers in 2018 and 2019. We first calculate the percentile position of each paragraph based on its location in a document. For example, if there are 20 paragraphs in a speech, the first paragraph's percentile position is 5 per cent, and the second is 10 per cent and so on.

Figure 15 shows the results from our 4 models. We can see that reasoning scores are much higher for paragraphs at the end of a speech, but readability scores are relatively higher for those at the beginning. This pattern seems to reflect a natural structure of a speech. The introduction is usually pleasantries and broad ideas, which are easy to understand as speakers want to grab the audience's attention and ensure they listen to the rest of it. The conclusion, conversely, is usually where the main arguments or opinions presented by the speaker are summarised.

This variation through the document, however, raises questions about the best way to assess overall document quality. Our method, by weighting paragraphs throughout a document equally, may penalise longer documents that contain more factual body paragraphs even though a human reader might judge them to be equally effective. We leave the question of which is the most effective way of rating the overall quality of a document for future research. Regardless, this suggests that targeting particular readability metrics may be useful for introductory paragraphs, but off target for conclusions. As with targeting different audiences with different documents, so

too different rhetorical objectives should be targeted with different styles – one size does not fit all.



Conclusion

In this study, we developed a novel approach of using survey data and machine learning models to assess the communication quality of central bank publications. To the extent that an important part of central bank transparency is to communicate ideas, positions and arguments we introduced a measure of reasoning in addition to the more commonly considered readability measure. Finally, recognising the multiplicity of readers for central bank documents, we considered how different audiences perceive the readability and reasoning of documents.

While our results are preliminary and subject to a number of limitations, they all point in a similar direction: communication needs to be adapted for different audiences and no single measure can do justice to the multiplicity of objectives communication has. There is little agreement between the economists and non-economists in our survey about the readability of paragraphs; there is little correlation between the readability of paragraphs we analysed and the reasoning contained in them; and readability alone is insufficient to capture the essence of transparent communication. Consequently, central banks aiming to improve transparency may need to present their core arguments in a range of formats with different expressions of those arguments in each. This may present a challenge for those central banks that have tended to emphasise verbatim consistency of their key messages as a way of reducing confusion. Regardless, we hope that better awareness of the various trade-offs involved in crafting communications, through the use of tools like those introduced in this paper, should lead to more effective communication in the future. We also hope that this research can serve as a foundation and catalyst for further investigation of the way different audiences perceive the multiplicity of elements that comprise effective and transparent central bank communication.

References

- Azar M (1999)**, 'Argumentative Text as Rhetorical Structure: An Application of Rhetorical Structure Theory', *Argumentation*, 13(1), pp 97–114.
- Bernanke BS (2010)**, 'Central Bank Independence, Transparency, and Accountability', Opening Remarks at the Bank of Japan–Institute for Monetary and Economic Studies Conference 'Future of Central Banking under Globalization', Tokyo, 26–27 May.
- Bholat D, J Brookes, C Cai, K Grundy and J Lund (2017)**, 'Sending Firm Messages: Text Mining Letters from PRA Supervisors to Banks and Building Societies They Regulate', Bank of England Staff Working Paper No 688.
- Bholat D, N Broughton, J Ter Meer and E Walczak (2019)**, 'Enhancing Central Bank Communications Using Simple and Relatable Information', *Journal of Monetary Economics*, 108, pp 1–15.
- Bini-Smaghi L and D Gros (2001)**, 'Is the ECB Sufficiently Accountable and Transparent?', European Network of Economic Policy Research Institutes, ENEPRI Working Paper No 7.
- Bjelobaba G, A Savic and H Stefanovic (2017)**, 'Analysis of Central Banks Platforms on Social Networks', Paper presented at the UBT 6th Annual International Conference, International Conference on Computer Science and Communication Engineering, Durrës, 27–29 October. Available at <<https://knowledgecenter.ubt-uni.net/conference/2017/all-events/81/>>.
- Blinder A, M Ehrmann, M Fratzscher, J de Haan and D-J Jansen (2008)**, 'Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence', *Journal of Economic Literature*, 46(4), pp 910–945.
- Blinder A, C Goodhart, P Hildebrand, D Lipton and C Wyplosz (2001)**, *How Do Central Banks Talk?*, Geneva Reports on the World Economy, 3, International Center for Monetary and Banking Studies, Geneva and Centre for Economic Policy Research, London.
- Born B, M Ehrmann and M Fratzscher (2011)**, 'Central Bank Communication on Financial Stability', European Central Bank Working Paper Series No 1332.
- Breiman L (2001)**, 'Random Forests', *Machine Learning*, 45(1), pp 5–32.
- Bulíř A, M Čihák and D-J Jansen (2012)**, 'Clarity of Central Bank Communication about Inflation', IMF Working Paper No WP/12/9.
- Cohen R (1984)**, 'A Computational Theory of the Function of Clue Words in Argument Understanding', in *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp 251–258. Available at <<http://www.aclweb.org/anthology/P84-1055>>.
- Davis JS and MA Wynne (2016)**, 'Central Bank Communication: A Case Study', Federal Reserve Bank of Dallas Globalization and Monetary Policy Institute Working Paper No 283.
- de Haan J, F Amtenbrink and S Waller (2004)**, 'The Transparency and Credibility of the European Central Bank', *Journal of Common Market Studies*, 42(4), pp 775–794.

- Dincer N and B Eichengreen (2009)**, 'Central Bank Transparency: Causes, Consequences and Updates', NBER Working Paper No 14791.
- Dincer N and B Eichengreen (2014)**, 'Central Bank Transparency and Independence: Updates and New Measures', *International Journal of Central Banking*, 10(1), pp 189–253.
- Eijffinger SCW and PM Geraats (2006)**, 'How Transparent are Central Banks?', *European Journal of Political Economy*, 22(1), pp 1–21.
- Farra N, S Somasundaran and J Burstein (2015)**, 'Scoring Persuasive Essays Using Opinions and their Targets', in J Tetreault, J Burstein and C Leacock (eds), *The Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, Proceedings of the Workshop, Association for Computational Linguistics, pp 64–74. Available at <<https://www.aclweb.org/anthology/W15-0608>>.
- Ferretti RP and S Graham (2019)**, 'Argumentative Writing: Theory, Assessment, and Instruction', *Reading and Writing*, 32(6), pp 1345–1357.
- Filardo A and D Guinigundo (2008)**, 'Transparency and Communication in Monetary Policy: A Survey of Asian Central Banks', Paper presented at the Bangko Sentral ng Pilipinas – Bank for International Settlements (BSP-BIS) High Level Conference on 'Transparency and Communication in Monetary Policy', Manila, 31 January, rev April 2008. Available at <<https://www.bis.org/repofficepubl/arpresearch200801.3.pdf>>.
- Fracasso A, H Genberg and C Wyplosz (2003)**, *How Do Central Banks Write? An Evaluation of Inflation Targeting Central Banks*, Geneva Reports on the World Economy Special Report 2, International Center for Monetary and Banking Studies, Geneva and Centre for Economic Policy Research, London.
- Fry M, D Julius, L Mahadeva, S Roger and G Sterne (2000)**, 'Key Issues in the Choice of Monetary Policy Framework', in L Mahadeva and G Sterne (eds), *Monetary Policy Frameworks in a Global Context*, Routledge, London, pp 1–216.
- Goldman SR and JA Rakestraw, Jr (2000)**, 'Structural Aspects of Constructing Meaning from Text', in ML Kamil, PB Mosenthal, PD Pearson and R Barr (eds), *Handbook of Reading Research: Volume III*, Routledge, New York, pp 311–335.
- Guyon I, J Weston, S Barnhill and V Vapnik (2002)**, 'Gene Selection for Cancer Classification Using Support Vector Machines', *Machine Learning*, 46(1–3), pp 389–422.
- Haldane A (2017)**, 'A Little More Conversation, a Little Less Action', Dinner Address given at the Federal Reserve Bank of San Francisco Macroeconomics and Monetary Policy Conference, San Francisco, 31 March.
- Haldane A and M McMahon (2018)**, 'Central Bank Communications and the General Public', *AEA Papers and Proceedings*, 108, pp 578–583.
- Hawkesby C (2019)**, 'Speaking, Listening and Understanding: The Art of Monetary Policy Communications', Address given at the 11th Annual Commonwealth Bank Global Markets Conference, Sydney, 28 October.
- Hornik K (2019)**, 'openNLP: Apache OpenNLP Tools Interface', R package version 0.2-7. Available at <<https://CRAN.R-project.org/package=openNLP>>.

Janan D and D Wray (2012), 'Readability: The Limitations of an Approach through Formulae', Paper presented at the British Educational Research Association Annual Conference 2012, Manchester, 4–6 September. Available at <<http://www.leeds.ac.uk/educol/documents/213296.pdf>>.

Jansen D-J (2011), 'Does the Clarity of Central Bank Communication Affect Volatility in Financial Markets? Evidence from Humphrey-Hawkins Testimonies', *Contemporary Economic Policy*, 29(4), pp 494–509.

Karimi D, H Dou, SK Warfield and A Gholipour (2020), 'Deep Learning with Noisy Labels: Exploring Techniques and Remedies in Medical Image Analysis', *Medical Image Analysis*, 65, Article 101759.

Kincaid JP, RP Fishburne, Jr, RL Rogers and BS Chissom (1975), 'Derivation of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) for Navy Enlisted Personnel', Naval Technical Training Command, Millington, Research Branch Report 8-75. Available at Institute for Simulation and Training, University of Central Florida <<https://stars.library.ucf.edu/istlibrary/56/>>.

Luangaram P and W Wongwachara (2017), 'More Than Words: A Textual Analysis of Monetary Policy Communication', Puey Ungphakorn Institute for Economic Research Discussion Paper No 54.

McRae K, TR Ferretti and L Amyote (1997), 'Thematic Roles as Verb-Specific Concepts', *Language and Cognitive Processes*, 12(2-3), pp 137–176.

Olsen LA and R Johnson (1989), 'Towards a Better Measure of Readability: Explanation of Empirical Performance Results', *Word*, 40(1-2), pp 223–234.

Preston B (2020), 'The Case for Reform of the Reserve Bank of Australia Policy and Communication Strategy', *The Australian Economic Review*, 53(1), pp 95–104.

Quinlan JR (1986), 'Induction of Decision Trees', *Machine Learning*, 1(1), pp 81–106.

Redish J (2000), 'Readability Formulas Have Even More Limitations Than Klare Discusses', *ACM Journal of Computer Documentation*, 24(3), pp 132–137.

Santorini B (1990), 'Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)', University of Pennsylvania, Department of Computer & Information Science Technical Report No MS-CIS-90-47. Available at <https://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports>.

Strunk, Jr W and EB White (1959), *The Elements of Style*, The Macmillan Publishing Company, New York, p 58.

Taylor A, M Marcus and B Santorini (2003), 'The Penn Treebank: An Overview', in A Abeillé (ed), *Treebanks: Building and Using Parsed Corpora*, Text, Speech and Language Technology, Vol 20, Kluwer Academic Publishers, Dordrecht, pp 5–22.

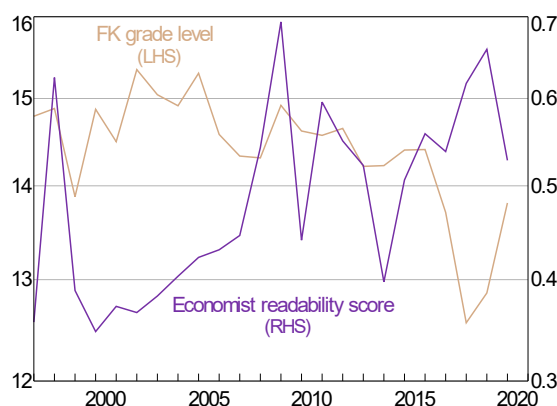
Woodford M (2005), 'Central Bank Communication and Policy Effectiveness', NBER Working Paper No 11898.

Yellen J (2012), 'Revolution and Evolution in Central Bank Communication', Speech given at the Haas School of Business, University of California, Berkeley, 13 November.

Appendix

Comparison of FK Grade Level and Economist Readability
SMP introduction, annual average

Figure A1



Source: Authors' calculations using survey results

Key Features Extracted from Sample Paragraphs

Table A1

| Category | Feature name | Description |
|------------------------|---|--|
| Textual ^(a) | Paragraph length | The count of words in a paragraph |
| | Sentence count | The count of sentences in a paragraph |
| | Number count | The count of numbers in a paragraph |
| | Comma count | The count of commas in a paragraph |
| | Other punctuation count | The count of any other punctuations except commas |
| Readability | First sentence with numbers | A Boolean value indicating the first sentence contains numbers |
| | First sentence with 'Table' or 'Figure/Graph' | A Boolean value indicating the first sentence refers to tables, figures or graphs |
| | Syllables count | The count of syllables |
| | Average word length | The average syllables of a word |
| | Count of complicated words | The count of words that have three and more syllables |
| Syntactic | FK grade level | Flesch–Kincaid grade level |
| | PoS count | The count of tokens marked with a certain part-of-speech tag in a paragraph |
| | PoS ratio | The percentage of tokens marked with a certain part-of-speech tag in a paragraph |
| | PoS count in the first sentence | The count of tokens marked with a certain part-of-speech tag in the first sentence of a paragraph |
| | PoS ratio in the first sentence | The percentage of tokens marked with a certain part-of-speech tag in the first sentence of a paragraph |
| | PoS count in the last sentence | The count of tokens marked with a certain part-of-speech tag in the last sentence of a paragraph |
| | PoS ratio in the last sentence | The percentage of tokens marked with a certain part-of-speech tag in the last sentence of a paragraph |
| | PoS for the first word in the first sentence | The type of PoS tag for the first word in the first sentence of a paragraph |
| | | |

| | | |
|--|--|--|
| Argument features | PoS for the first word in the second sentence | The type of PoS tag for the first word in the second sentence of a paragraph |
| | PoS for the first word in the third sentence | The type of PoS tag for the first word in the third sentence of a paragraph |
| | Parse tree types count for a paragraph | The count of parse tree types for each sentence in a paragraph |
| | Parse tree types count for the first sentence of a paragraph | The count of parse tree types for the first sentence of a paragraph |
| | Parse tree types count for the last sentence of a paragraph | The count of parse tree types for the last sentence of a paragraph |
| | Count of each type of clue words | Count of clue words by each type (summarise, informative, etc) |
| | Count of clue words in the first sentence | Count of clue words by each type in the first sentence of a paragraph |
| | Count of clue words in the last sentence of a paragraph | Count of clue words by each type in the last sentence of a paragraph |
| Note: (a) We deliberately exclude n-gram words in the feature list as our survey only includes economists working in the RBA, who have sufficient knowledge for all economic terms | | |

Alphabetical List of the Penn Treebank Part-of-Speech Tag Set Table A2

| Number | Tag | Description |
|--------|-------|--|
| 1 | CC | Coordinating conjunction |
| 2 | CD | Cardinal number |
| 3 | DT | Determiner |
| 4 | EX | Existential there |
| 5 | FW | Foreign word |
| 6 | IN | Preposition or subordinating conjunction |
| 7 | JJ | Adjective |
| 8 | JJR | Adjective, comparative |
| 9 | JJS | Adjective, superlative |
| 10 | LS | List item marker |
| 11 | MD | Modal |
| 12 | NN | Noun, singular or mass |
| 13 | NNS | Noun, plural |
| 14 | NNP | Proper noun, singular |
| 15 | NNPS | Proper noun, plural |
| 16 | PDT | Predeterminer |
| 17 | POS | Possessive ending |
| 18 | PRP | Personal pronoun |
| 19 | PRP\$ | Possessive pronoun |
| 20 | RB | Adverb |
| 21 | RBR | Adverb, comparative |
| 22 | RBS | Adverb, superlative |
| 23 | RP | Particle |
| 24 | SYM | Symbol |
| 25 | TO | to |
| 26 | UH | Interjection |
| 27 | VB | Verb, base form |
| 28 | VBD | Verb, past tense |
| 29 | VBG | Verb, gerund or present participle |
| 30 | VCN | Verb, past participle |
| 31 | VBP | Verb, non-3rd person singular present |
| 32 | VBZ | Verb, 3rd person singular present |

| | | |
|-------------------------------|------|-----------------------|
| 33 | WDT | Wh-determiner |
| 34 | WP | Wh-pronoun |
| 35 | WP\$ | Possessive wh-pronoun |
| 36 | WRB | Wh-adverb |
| Source: Santorini (1990, p 6) | | |

| Alphabetical List of the Penn Treebank Parse Tree Tag Set | | | Table A3 |
|---|--------|---|----------|
| Number | Tag | Description | |
| 1 | ADJP | Adjective phrase | |
| 2 | ADVP | Adverb phrase | |
| 3 | NP | Noun phrase | |
| 4 | PP | Prepositional phrase | |
| 5 | S | Simple declarative clause | |
| 6 | SBAR | Subordinate clause | |
| 7 | SBARQ | Direct question introduced by wh-element | |
| 8 | SINV | Declarative sentence with subject-aux inversion | |
| 9 | SQ | Yes/no questions and subconstituent of SBARQ excluding wh-element | |
| 10 | VP | Verb phrase | |
| 11 | WHADVP | Wh-adverb phrase | |
| 12 | WHNP | Wh-noun phrase | |
| 13 | WHPP | Wh-prepositional phrase | |
| Source: Table 1.2 in Taylor, Marcus and Santorini (2003, p 9) | | | |

| A Short List of Text Features for a Sample Sentence | | Table A4 |
|--|--|----------|
| 'The cat sat on the mat because it was warm.' | | |
| | Value | |
| Text features | | |
| Count of words | 10 | |
| Count of sentences | 1 | |
| Count of syllables | 11 | |
| Count of polysyllables (words with 3+ syllables) | 0 | |
| Syllables per word | 1.1 | |
| FK grade level | 1.29 | |
| Count of clue words ^(a) | 1 ('because') | |
| Syntactic features | | |
| PoS tags feature | DT = 2, NN = 2, VBD = 2, IN = 2, DT = 1, NN = 1, PRP = 1, JJ = 1 | |
| Syntactic parse features | S = 2, NP = 3, VP = 2, SBAR = 1, PP = 1, ADJP = 1 | |
| Note: (a) 'Clue words' is a list of words or phrases that link individual propositions to form one coherent presentation; please refer to Cohen (1984) for a full list | | |

Model Tuning Process

Feature selection process

In this study we adopt an automatic feature selection method, called recursive feature elimination (RFE) (Guyon *et al* 2002), to select the relevant features for each model. This helps ensure that each feature included in the final model has a minimum degree of predictive power. Otherwise, the models may mistake 'noise' for 'signal'. This algorithm is configured to explore all possible subsets of the features. The computing process is shown in Table A5

| Key Steps of a Recursive Feature Elimination Process | | Table A5 |
|---|---|----------|
| 1.1 | Train the model on training dataset using all features $\{X_1, X_2, \dots, X_n\}$ | |
| 1.2 | Calculate model performance | |
| 1.3 | Calculate variable performance | |
| 1.4 | For each subset size of $S_i, i = 1 \text{ to } n$ do | |
| | 1. Keep the S_i most important features | |
| | 2. Train the model on the training dataset using top S_i features | |
| | 3. Calculate the model performance | |
| 1.5 | End | |
| 1.6 | Calculate the performance profile over the S_i | |
| 1.7 | Determine the appropriate number of predictors | |
| 1.8 | Use the model corresponding to the optimal S_i | |
| Source: https://topepo.github.io/caret/recursive-feature-elimination.html | | |

Our model includes 292 features in total, so in the first step of the RFE process we include all features. Then, we run the model using 30 different subset feature sizes, that is (10, 20, ..., 290, 292). To minimising overfitting due to feature selection, we take the cross-validation resampling method to run the process listed in Table A5 on the testing dataset only and calculate the model performance using the validation dataset. We run this process 10 times and calculate the model performance (accuracy) for each subset of features using the average of the results from those 10 runs.

Tuning parameters process

To improve model performance, we tune 2 parameters:

- the number of trees that will be built for each model (n_{tree}), and
- the optimal number of variables that will be selected for each node in a tree (m_{try}).

The default value of n_{tree} is 500, and that of m_{try} is the root square of number of features. Different values of those 2 parameters may affect model performance. To find the optimal settings, we employ a grid search approach.

For the grid search, we choose 11 different n_{tree} values (10, 100, 200, 300,...,1,000) and, for m_{try} , as suggested by Breiman (2001), we choose 3 values: the default value ($m_{try} = 17$), half of the default ($m_{try} = 9$), and twice the default ($m_{try} = 34$). For each combination, we build 10 models using 10-fold cross-validation and repeat the process 3 times. The best combination of n_{tree} and m_{try} is selected based on the combination that returns the highest accuracy.

Top ten features for four models

| Top Ten Features for Four RF Models | | | | Table A6 |
|--|-------------------|---------------------------|----------------------|---------------------------|
| Rank | Reasoning model | | Readability model | |
| | Features | Importance ^(a) | Features | Importance ^(a) |
| Economist | | | | |
| 1 | Proportion of VB | 6.2 | Proportion of CC | 13.1 |
| 2 | Proportion of NNS | 4.6 | Proportion of RB | 9.7 |
| 3 | Proportion of MD | 4.5 | Proportion of VB | 7.6 |
| 4 | Count of digits | 4.1 | Proportion of VBP | 7.4 |
| 5 | Count of VB | 3.9 | Count of NN | 6.9 |
| 6 | Proportion of NN | 3.6 | Count of NP | 6.8 |
| 7 | Count of MD | 3.5 | Count of punctuation | 5.9 |
| 8 | Proportion of IN | 3.5 | Proportion of MD | 5.5 |
| 9 | Proportion of CD | 3.5 | Count of commas | 4.6 |
| 10 | Proportion of VBN | 2.8 | Count of SBAR | 4.6 |
| Non-economist | | | | |
| 1 | Proportion of VB | 10.6 | Proportion of DT | 5.2 |
| 2 | Proportion of MD | 9.0 | Proportion of JJ | 5.1 |
| 3 | Proportion of JJ | 7.3 | FK grade level | 4.7 |
| 4 | Proportion of IN | 6.1 | Count of NP | 4.7 |
| 5 | Proportion of NN | 5.9 | Count of syllables | 4.7 |
| 6 | Count of MD | 5.3 | Proportion of NN | 4.4 |
| 7 | Proportion of VBN | 5.3 | Proportion of CC | 4.4 |
| 8 | Count of VB | 5.2 | Proportion of VB | 4.3 |
| 9 | Proportion of TO | 5.1 | Proportion of IN | 4.3 |
| 10 | Proportion of CC | 5.1 | Proportion of NNS | 4.2 |
| Note: (a) The feature importance is extracted as a part of model outputs that is generated using the caret package in R; the importance value for each variable is calculated as the contribution of each variable based on mean decrease in impurity (Gini) after removing this feature | | | | |

Model Validation Results

Confusion matrix

We apply our fine-tuned RF models to the validation dataset and the prediction results are shown in the confusion matrices in Tables A7 and A8. Using the confusion matrix, we can calculate a number of performance metrics, such as:

- Accuracy: the proportion of the total number of predictions that were correct. That is the sum of true positive (TP) and true negative (TN) divided by the total observations ($TP + TN + FP + FN$). In Table A7 (reasoning panel), the accuracy is calculated as: $(33 + 23) / (33 + 13 + 4 + 23) = 76.71\%$.
- Sensitivity: the proportion of positives that are correctly predicted. In Table A7, the sensitivity is calculated as $(33) / (33 + 4) = 89.19\%$.
- Specificity: the proportion of negatives that were correctly predicted. In Table A7, the sensitivity is calculated as $(23) / (13 + 23) = 63.89\%$.

Kappa is another metric that can be calculated from the confusion matrix using the formula:

$$Kappa = \frac{accuracy - random\ accuracy}{1 - random\ accuracy}$$

where:

$$p_1 = \frac{TP + FN}{Total}$$

$$p_2 = \frac{TP + FP}{Total}$$

$$random\ accuracy = p_1 p_2 + (1 - p_1)(1 - p_2)$$

Accuracy is a fairly commonly used measure and it varies from 76.7 per cent for the economist content model to 65.2 per cent for the non-economist clarity model; 70 per cent is a threshold usually considered to indicate 'fair' performance.²⁸ A final validation measure included in these tables, but which can not be calculated directly from the confusion matrix because it focuses on the strength of the prediction, is LogLoss²⁹ – lower numbers are better for this metric. Overall, our results on this metric are relatively poor, reflecting the fact that our model does not make strong predictions about paragraph quality.

28 We also applied the other algorithms discussed in Section 6.1 to our validation dataset for the economist content model and their accuracy was worse than our final model (the fine-tuned RF model). For more details, please refer to the online supplementary information.

29 LogLoss is another metric that is widely used for assessing prediction performance of ML models. It

is calculated as: $Loss = -\frac{1}{N} \sum_{i=1}^N [y_i \times \log(p(y_i)) + (1 - y_i) \times \log(1 - p(y_i))]$, where y_i is the label and $p(y_i)$ is the predication probability. LogLoss penalises false classification, especially heavily on those that are confidently wrong. It ranges from zero to infinity.

Confusion Matrix for Economist RF Model

Cut-off threshold = 0.5

Table A7

| Reasoning | | | | Readability | | | |
|------------------|------|-----|-----------------------|------------------|------|-----|-----------------------|
| Confusion matrix | | | Performance measures | Confusion matrix | | | Performance measures |
| Reference | | | | Reference | | | |
| Predictio n | High | Low | Accuracy = 76.71 % | Predictio n | High | Low | Accuracy = 72.37 % |
| High | 33 | 13 | 95% CI: (65%, 86%) | High | 28 | 11 | 95% CI: (61%, 82%) |
| Low | 4 | 23 | Sensitivity = 89.19 % | Low | 10 | 27 | Sensitivity = 73.68 % |
| | | | Specificity = 63.89 % | | | | Specificity = 71.05 % |
| | | | Kappa = 0.53 | | | | Kappa = 0.45 |
| | | | LogLoss = 0.75 | | | | LogLoss = 0.80 |

Confusion Matrix for Non-economist RF Model

Cut-off threshold = 0.5

Table A8

| Reasoning | | | | Readability | | | |
|------------------|------|-----|-----------------------|------------------|------|-----|-----------------------|
| Confusion matrix | | | Performance measures | Confusion matrix | | | Performance measures |
| Reference | | | | Reference | | | |
| Predictio n | High | Low | Accuracy = 69.91 % | Predictio n | High | Low | Accuracy = 65.22 % |
| High | 41 | 18 | 95% CI: (61%, 78%) | High | 48 | 27 | 95% CI: (56%, 74%) |
| Low | 16 | 38 | Sensitivity = 71.93 % | Low | 13 | 27 | Sensitivity = 78.69 % |
| | | | Specificity = 67.86 % | | | | Specificity = 50% |
| | | | Kappa = 0.40 | | | | Kappa = 0.29 |
| | | | LogLoss = 0.82 | | | | LogLoss = 0.61 |

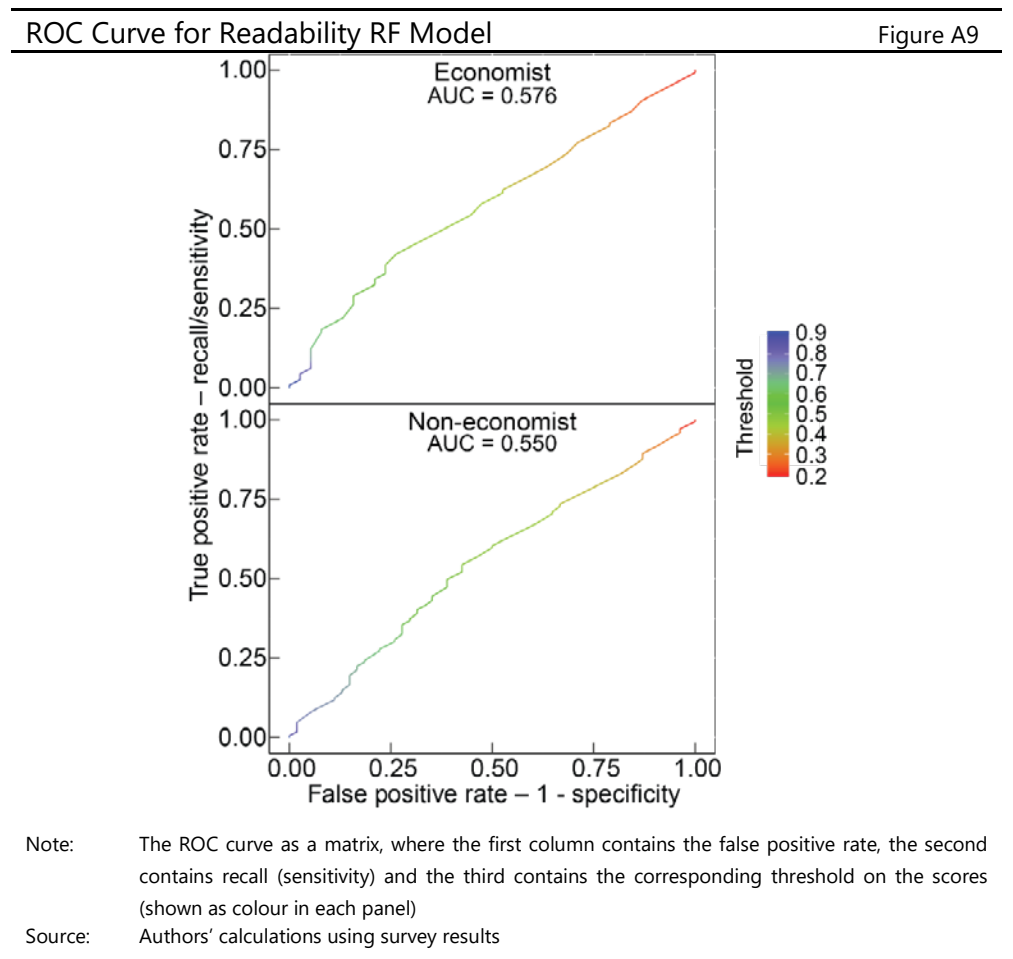
ROC-AUC

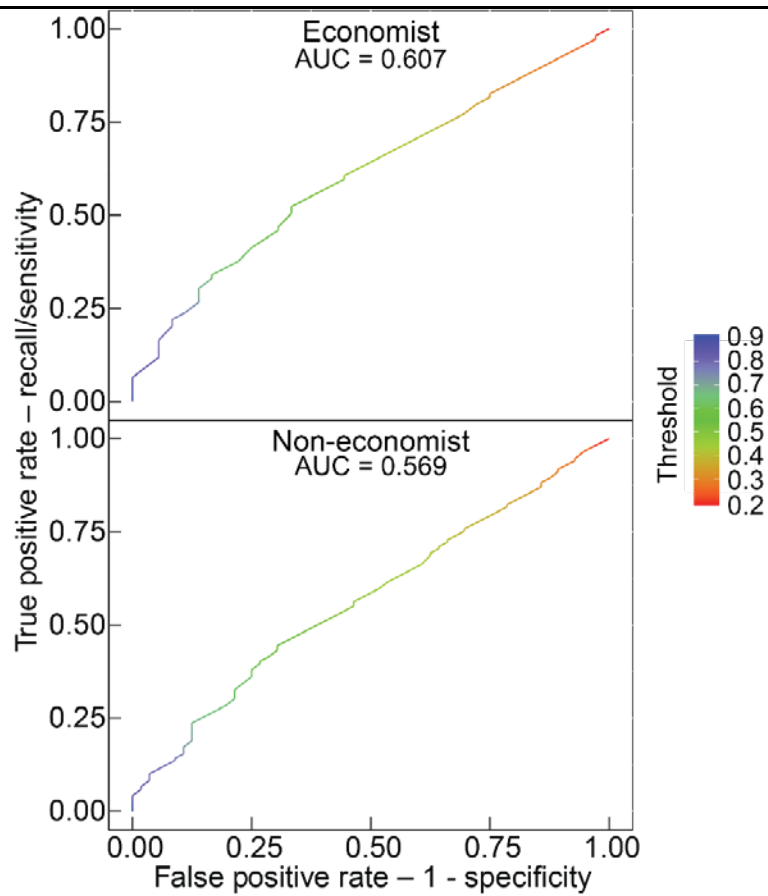
ROC is a probability curve that plots the false positive rate (FPR) on the x-axis and the true positive rate (TPR) on the y-axis for different probability cut-off thresholds. The area under the curve (AUC) is a measure of separability that is calculated as the area under the curve. The higher the AUC, the better the model is at definitively distinguishing between paragraphs with high quality and low quality. For a random classifier, such as a coin flip, there is a 50 per cent chance to get the classification right, so the FPR and TPR are the same no matter which threshold you choose. In this case the ROC curve is the 45-degree diagonal line and the area under the curve (AUC) equals 0.5. Thus, we would like to achieve an AUC of above 0.5. Our results, as shown in Figures A9 and A10, beat this benchmark but not substantially.

The fundamental problem we face in using the AUC metric is that the underlying quality of paragraphs is not cleanly separated into high and low, but has a large mass of inherently ambiguous paragraphs. What AUC requires is that a paragraph that is of '51% quality' is always perfectly classified as high while a paragraph of '49% quality'

is always classified as low, regardless of the cut-off threshold you use with your algorithm. For example, our algorithm may report that there is a 51 per cent chance that a given (truly 51% quality) paragraph is of high quality. We use a threshold of 50 per cent and this paragraph would be correctly classified as high with that cut-off. But, the AUC also asks what if you used a cut-off of 55 per cent, of 60 per cent and so on – it is calculated for all possible cut-offs between 0 per cent and 100 per cent. The AUC will find that if you use any threshold above 51 per cent it will misclassify the paragraph and this leads to a low AUC measure for this problem. Thus, while AUC is a standard metric, it is not a good metric for our particular problem given the underlying data is not a binary variable but closer to a continuous variable. For the same reason, the LogLoss values – an alternative metric – are a bit high, ranging from 0.6 to 0.8. Notwithstanding this, we anticipate that further refinements of the algorithm should be possible that improve its performance on these and other metrics.

It is also important to note that our models report a relatively high accuracy when we set the threshold at 0.5. That is, while our model does not do a good job at neatly separating high- and low-quality paragraphs at every threshold, it does a reasonable job of identifying paragraphs that are more likely than not to be high or low quality. In this respect it is quite 'human-like'. This suggests that the results for any given paragraph should not be given a large weight but, with a large enough sample, the results will still be useful.





Note: The ROC curve as a matrix, where the first column contains the false positive rate, the second contains recall (sensitivity) and the third contains the corresponding threshold on the scores (shown as colour in each panel)

Source: Authors' calculations using survey results



RESERVE BANK OF AUSTRALIA

Central Bank Communication: What Can a Machine Tell Us About the Art of Communication?

John Simon & Joan Huang

14-17 February 2022

IFC Workshop on “Data science in central banking”

Importance of central bank communication

“

Communication can be an *important* and *powerful* part of the central bank's toolkit since it has the ability to move financial markets, to enhance the predictability of monetary policy decisions, and potentially to help achieve central banks' macroeconomic objectives.

”

*By Alan S. Blinder, Michael Ehrmann, Marcel
Fratzscher, Jakob De Haan and David-Jan
Jansen, 2008*

What is effective communication?

- Three aspects
 - Readability
 - message is easily understood
 - Reasoning
 - Transparent central banks need to explain their reasoning (e.g. Preston 2020)
 - The substance of the message matters, not just its readability
 - Audience
 - What is clear to one audience may not be clear to another

Survey sample

Please read each paragraph and then rate them for their clarity and their content.

Clarity: When we say clarity we mean how easy the paragraph was to read. Use a scale from 1 to 5 to score it where 1 is very unclear and 5 is very clear.

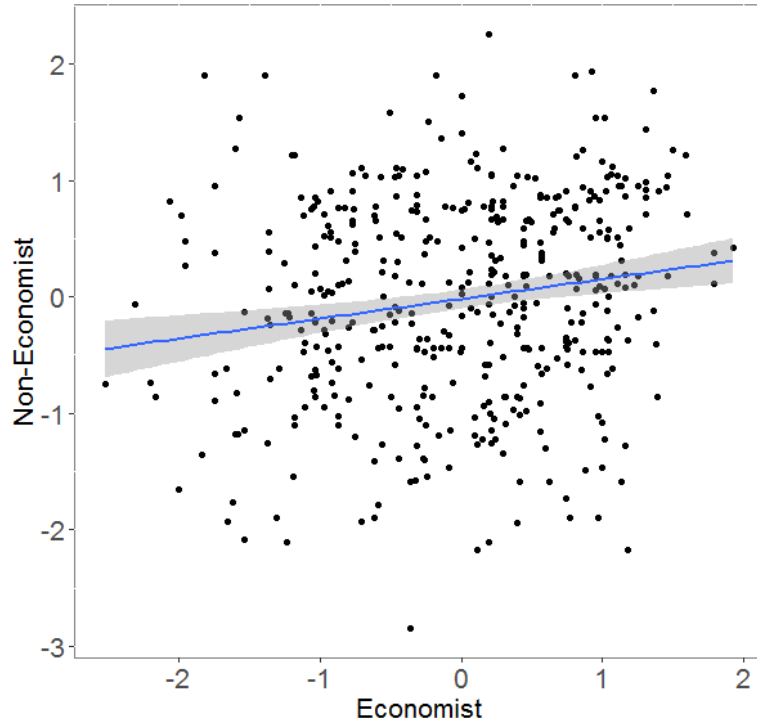
Content: When we say content we mean the extent to which the paragraph communicates ideas and arguments or why something is so. Use a scale from 1 to 5 to score it where 1 indicates a simple statement of facts and where 5 indicates that there is an idea, position, or explanation being given.

Many students are unaware that they can avoid paying for courses or subjects that they no longer want to take. Students usually have three or four weeks after teaching starts before they are charged or incur a Higher Education Loan Program (HELP) debt. To avoid paying, students must drop subjects prior to a 'census date'. A Grattan Institute survey found that fewer than 40 per cent of students surveyed understood the census date's significance; the others were unaware of it or confused it with some other university date. As a result, some students needlessly incur HELP debts.

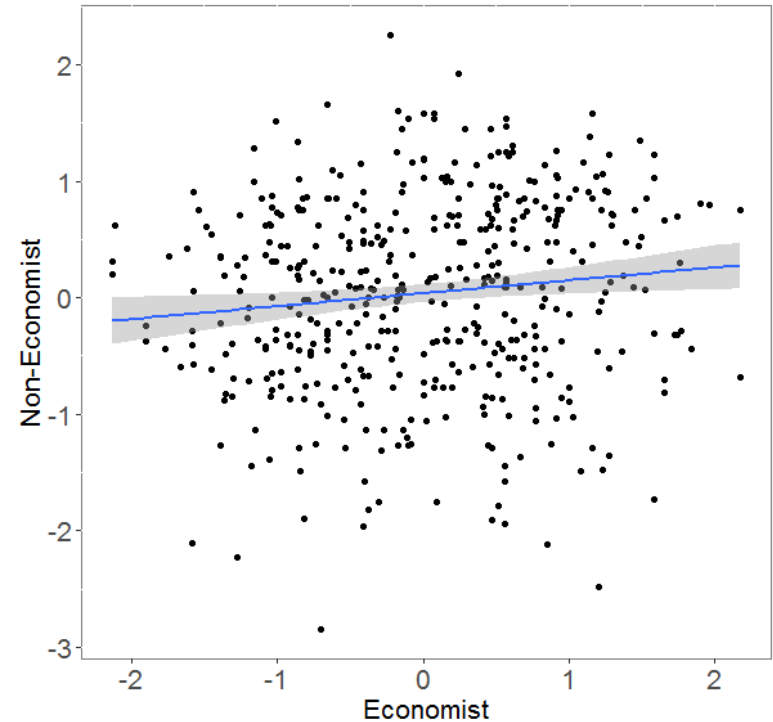
| | 1 | 2 | 3 | 4 | 5 |
|---|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Clarity (1=very unclear, 5=very clear) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Content (1=facts, 5=ideas) | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Economists vs. non-economists

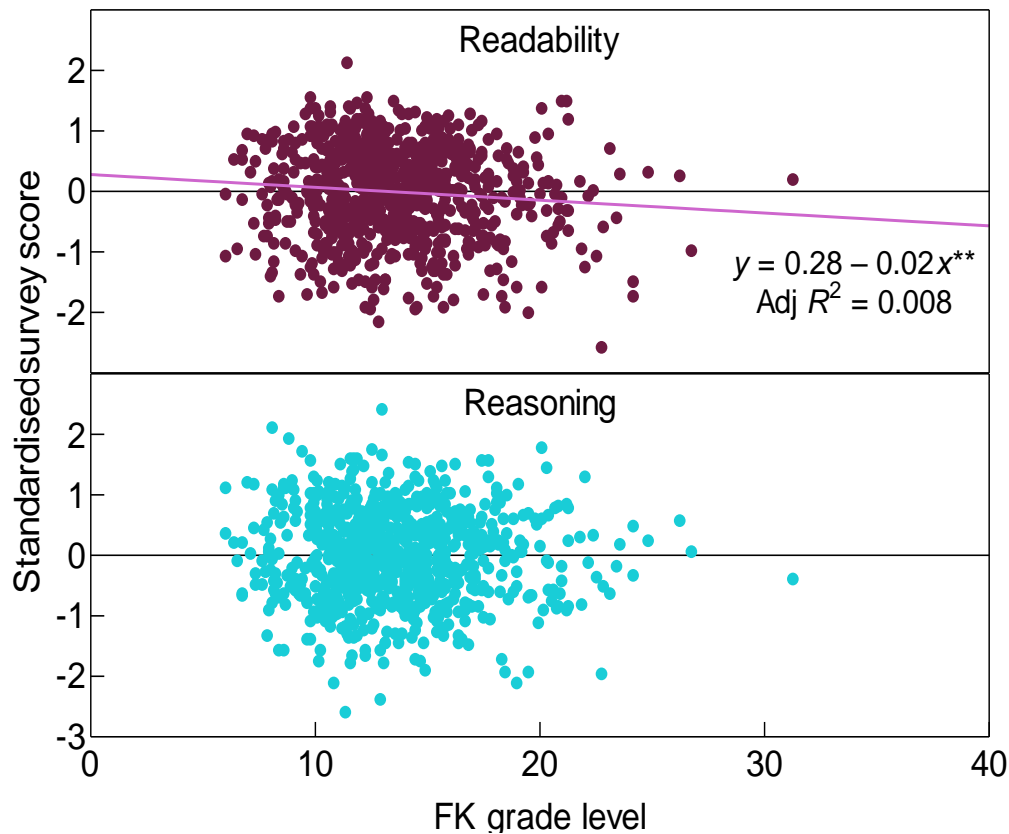
Readability



Reasoning



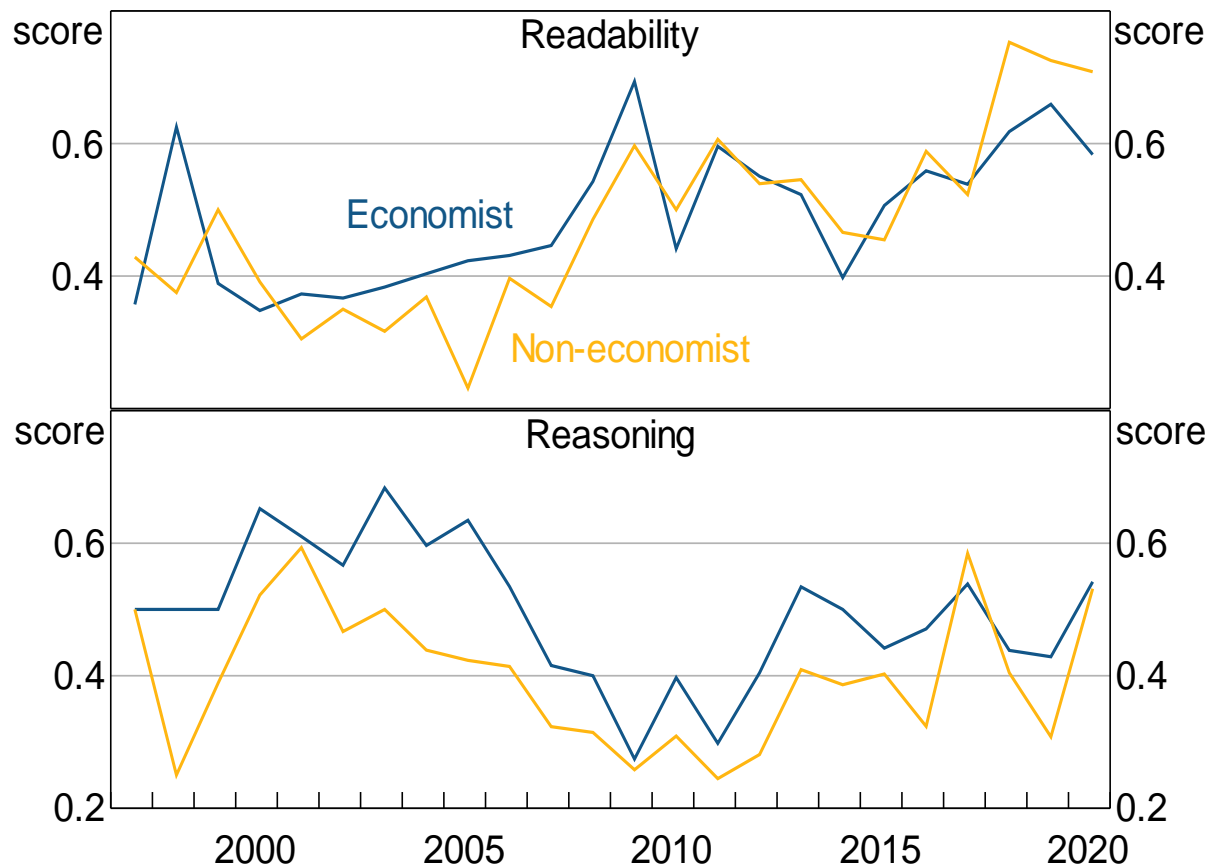
FK has little correlation to survey ratings



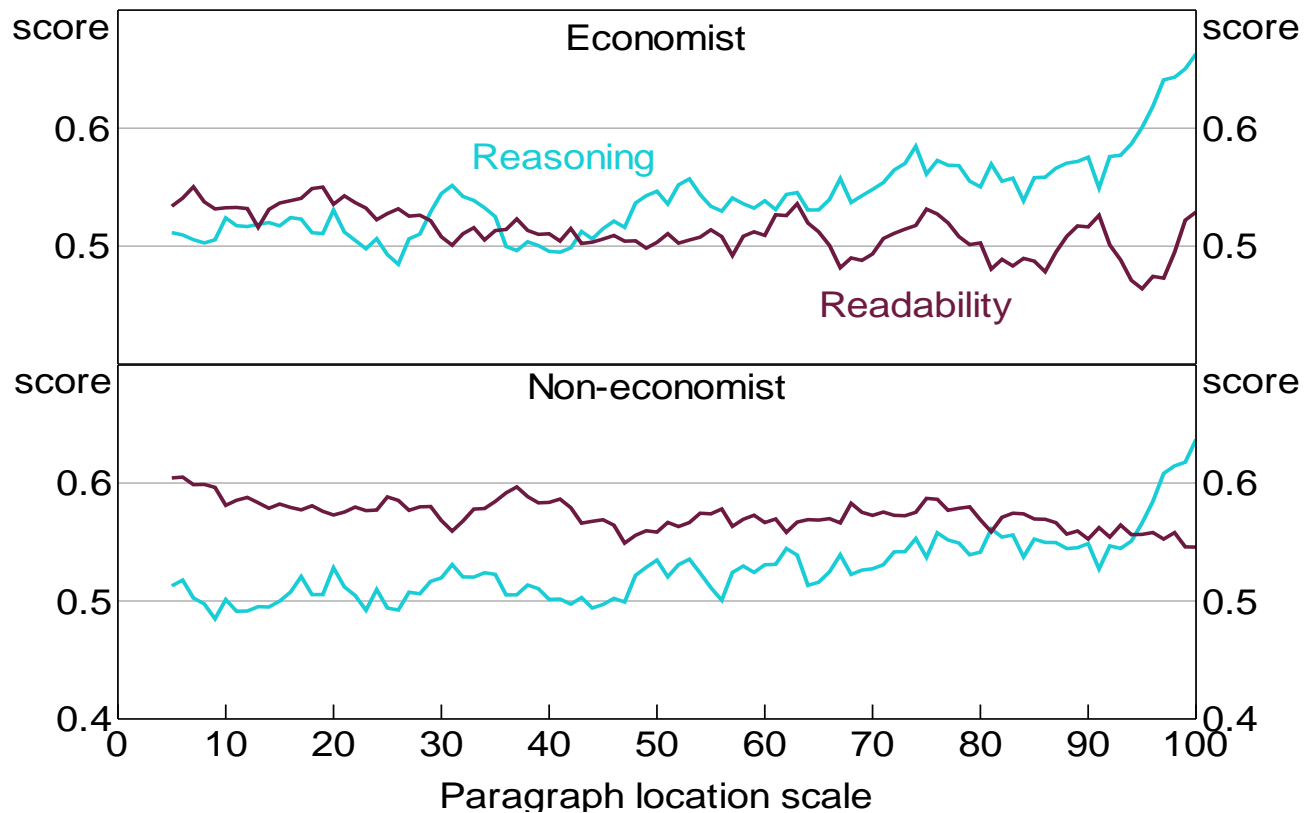
Three takeaways from survey results analysis

- Limitation of using readability formula
 - FK grade level is not sufficient for measuring either readability or reasoning
- Readability and reasoning
 - are ***independent*** measures of text quality
- Economists and non-economists
 - hold ***different*** views on text quality

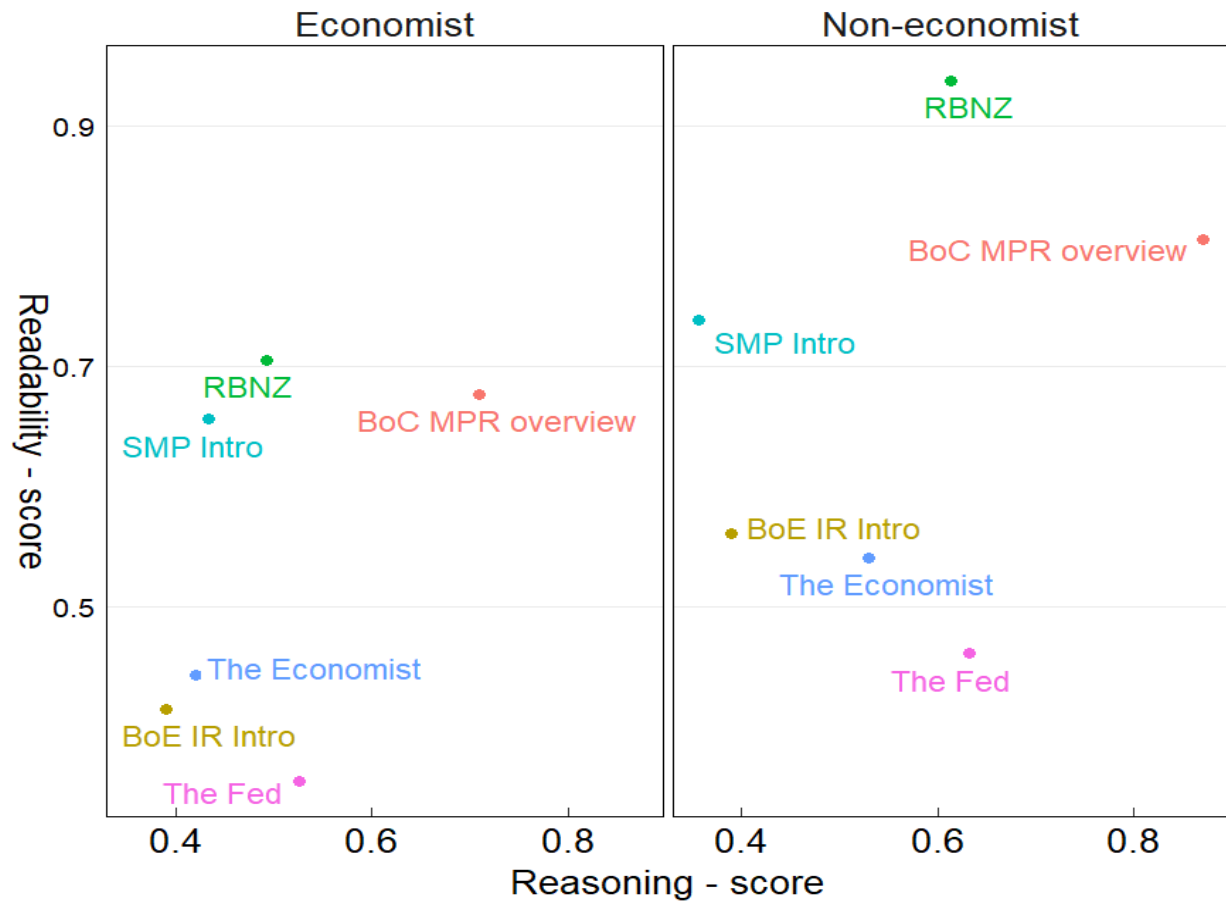
SMP overviews over time



within a document



Between organisations



Conclusion

- Existing readability measures limited
 - One-dimensional
 - Very weakly correlated with more comprehensive measures
- Readability and reasoning are two independent metrics
 - weakly correlated with each other
 - tradeoff between them
- Audiences matter: One size does not fit all
 - Central banks need to provide different documents for different audiences

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

News and banks' equities: do words have predictive power?¹

Valerio Astuti, Giuseppe Bruno, Sabina Marchetti and Juri Marcucci,
Bank of Italy

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

News and banks' equities: do words have predictive power?

Valerio Astuti*, Giuseppe Bruno*, Sabina Marchetti* and Juri Marcucci *

* Bank of Italy, DG for Economics, Statistics and Research, Via Nazionale 91, 00184 Rome, Italy.

Abstract

The employment of textual data from Italian newspapers can bring useful and timely insights into the economic conditions of banks and financial intermediaries. In this work we collect textual data from the most important national newspapers, we extract news sentiment and topics and investigate their role in predicting and explaining banking market variables such as stock volumes, yield and volatility. Different full and out-of-sample experiments show that many topics have predictive power for key banking market indicators. We show this by studying the performance of our model with respect to a simple autoregressive benchmark. Our model has smaller prediction errors over the period studied, and in addition it automatically selects topic and sentiment variables as useful for the predictions. Here our goal is twofold: on one hand we provide an empirical methodology to evaluate the polarity of newspaper articles written in Italian and secondly we establish a sound statistical framework to measure the causal links between sentiment and stock market series. We deem quite relevant a quantitative evaluation of the impact of the sentiment on financial markets in order to increase the timely awareness of the regulating institutions with respect to potentially critical microeconomic conditions.

JEL classification: C83, D84, E32.

Keywords: Latent Dirichlet Allocation (LDA), news aggregation, Topic Analysis, Sentiment Analysis.

December 2021

giuseppe.bruno@bancaditalia.it

The views expressed are the authors' only and do not imply those of the Bank of Italy.

1 Introduction and motivation

Veritas numquam perit.

Extracting useful signals from textual data taken from media outlets is an important topic in the field of Artificial Intelligence. The sheer amount of detailed online information streaming from social networks is increasingly attracting the attention of many kinds of researchers and practitioners. The linguistic analysis of social media, in different languages, has become a hot topic even for applied research [1, 2]. Detection of sentiments and opinions in social media is now a critical tool for monitoring such platforms. While the idea of news-driven economics forecasts is rather simple, evaluating its relevance could be quite challenging.

In this paper, we will focus on articles extracted from a comprehensive set of Italian newspapers starting at a different time period depending on the time an agreement between Dow Jones and the newspaper's editor was established. Among these newspapers we have included all of those mentioned in the Audiweb Internet ranking total November 2017.¹

Our analysis adds to the literature along two lines. The first one is the development of a sentiment analysis dictionary for the Banking-financial sector. The second consists of the definition and evaluation of a model for gauging the effects of news sentiments on stocks for the financial intermediation sector. Using only news sentiments, we achieved a mean directional accuracy of 80% in predicting the trends in short-term stock price movement.

The paper is arranged in the following way. In Section 2 we show how we assembled our *corpus* of Banking news. Section 3 describes the analysis for finding the number of topics and extracting them from our banking *corpus*. Section 4 explains the methodology and rules employed to carry on a sentiment analysis and extracting a banking sentiment index on the chosen *corpus*. Section 5 presents the results of the forecasting exercises for some balance sheet items, based on the sentiment index computed in the previous paragraph. Finally, Section 6 provides some concluding remarks and suggests some possible threads for future research.

¹<https://www.agcom.it/documents/10179/10214149/Studio-Ricerca+13-04-2018/4f2f5a5f-b76b-40f5-b07c-cb89359edecb?version=1.1>

2 Building a *corpus* of Banking news

To build our *corpus*, we have considered the features of Dow Jones Data, News and Analytics (DNA) information aggregator platform. For the purpose of our research we have designed a query (see the appendix for details) aimed at extracting most of the articles on banks and banking agglomerates in Italian language, and appeared on the major Italian newspapers in the period ranging from September 1996 to May 2018. The different newspapers from which the articles were extracted have different coverage over time: among the paper editions the oldest available newspaper is “La Stampa”, starting in 1996, whereas the most recent is “La Repubblica”, starting in 2005; the online editions started much later, the most recent being “Il Sole 24 Ore Online”, which started in 2013. The detail of the time coverage of the different newspapers can be found in Figure 2.

The result of the query submitted to the DNA platform consists of a total amount close to 220,000 articles, to which we applied some cleaning to remove duplicates and articles not suitable to textual analysis. More precisely, we removed duplicates and articles consisting mainly of tables and numerical data. For this purpose we employed a simple threshold based discrimination. For each article we computed the ratio R_d between the number of digits and alphabetic characters. A document is deemed suitable for our analysis when $R_d \leq .1$. In this way we excluded any document having more than 1 digit every 10 alphabetic characters. In addition we found and removed a small set of articles whose publication date was unknown. This preliminary filtering activity left us with a *corpus* of around 215,000 articles, one hundred million words and around 0.3 million of unique tokens.

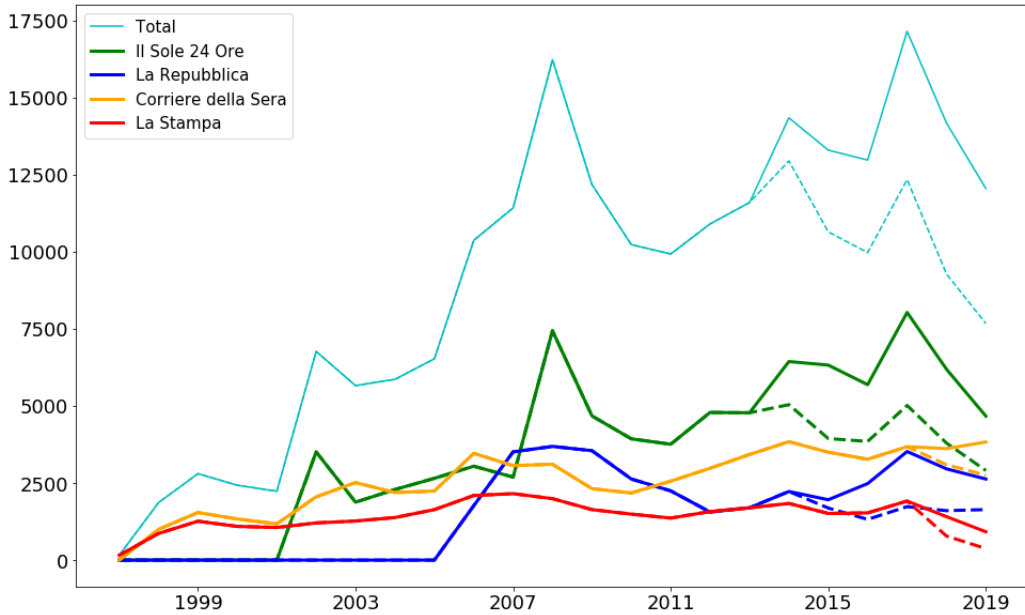


Figure 2.1: Number of articles per year (dashed lines are numbers excluding online editions)

As we will see in the following, the prediction algorithm we developed was applied only to a subset of articles, because the target variables we studied were available only from 2015 onward. This reduced the number of articles fed to the algorithm to 62,000.

The text mining preprocessing consisted of the following steps:

- **case normalization:** all capital letters were removed to discard difference between word given just by their position in the sentences;
- **punctuation removal:** non-alphanumeric characters were removed;
- **tokenization:** documents were transformed from whole strings to lists of words, treated as indepen-

dent objects. Extra spaces between words were removed;

- **stop-words removal:** a list of words carrying few informative content was selected and removed from all articles. Examples are articles and prepositions;
- **word stemming:** to reduce unjustified redundancies word-stemming was removed, identifying words like “prestito” and “prestiti” (“loan” and “loans”);
- **n-grams formation:** to retain some information about co-occurrences of words we considered not only single-word tokens, but also bi-grams: after the punctuation and stop-word removal we built the list of all couples of contiguous words;
- **bag-of-words analysis:** the daily content of the articles has been analyzed as an aggregate, so the final step of the text analysis was the formation of a bag-of-words (BoW) for any given day considered.

The previous steps are standard in any text mining analysis, and functional to any natural language processing application [3].

In order to perform the topic analysis detailed in the next section, articles were *vectorized*. Vectorization consists in the representation of every article as an element in a high dimensional vector space. We used one of the simplest mappings available, the so-called *term frequency* representation. This is a type of bag-of-words representation, that only accounts for the frequency of a given token in a document, discarding information about the position and the order of words in it. The first step is the construction of a vocabulary derived from our *corpus*. In principle we could include every word with at least an appearance in a document of the *corpus*; in this way however we would be forced to consider also typos or very rare and too common words. We will see that some kind of filter in the construction of the vocabulary can be employed to retain more useful information. Once a vocabulary is completed, every word contained in it will define a dimension of the vector space in which our articles are represented. In this representation every article is a vector having as many elements as the number of words in the vocabulary. Each component of this vector is the number of times the corresponding word appears in the statement or the whole document. As an example, consider the vocabulary:

$$V = \{\text{dog, sofa, cat, chair, table}\}$$

and the following sentences:

$$\begin{aligned}s_1 &= \text{The dog chased the cat over the sofa.} \\ s_2 &= \text{We should buy a sofa for us and a sofa for our dog.}\end{aligned}$$

The two sentences have the *term frequency* representation:

$$\begin{aligned}s_{1V} &= (1, 1, 1, 0, 0) \\ s_{2V} &= (1, 2, 0, 0, 0)\end{aligned}$$

In this representation every word in the vocabulary is an additional dimension in the documents vector space, so increasing the size of the vocabulary implies raising the complexity of the document representation. For this reason setting some filters in the construction of the vocabulary can improve the document representation. In particular words with very few appearances in the whole *corpus* - being them typos or very uncommon words - are not very important in the description of documents, the associated component being null in all but few documents. Conversely, very common words will be found in almost all the documents, so the associated components will not be useful in discriminating them.

We built our vocabulary discarding words appearing in more than 90% of the articles, and in less than 10% of the articles. Moreover we put a maximum limit on the size of the vocabulary, keeping only the 300'000 most frequent words. This latter selection discards less common words, so its effect goes in the same direction of putting a lower threshold on the number of documents in which a word has to appear so as to be considered.

3 Topic Distribution

Many different methodologies are available to quantify the content of news articles. Our application starts from a set of heterogeneous newspapers (some of them mostly focused on economic and financial news, other are generalist ones). We first pinned down the most relevant topics in the *corpus*.

Hierarchical mixture modeling is one among the most powerful methodology available to find patterns and structure in large collections of data. The main reason for which these models are useful is the large dimensionality reduction they achieve. Without any particular model every document would be described by the set of words by which it is composed. This would make each document a point in a space having dimension equal to the cardinality of the vocabulary. In our *corpus* we have an order of 10^5 words, so every article would be very sparse point in a huge dimensional space.

An interesting offspring of mixture modeling is *topic modeling*, where the data under study are large collections of documents. In this circumstance mixture modeling algorithms find the underlying patterns of words that are embedded in the collection. When using topics to identify documents, they become points in a space of dimension usually around the order of 10. In addition, with respect to other dimensionality reduction methods, the results obtained with topic modelling are more interpretable by humans. Pinning down these patterns - called in this setting *topics* - allows for effective clustering, searching, exploring, predicting and summarizing large corpora of documents.

To describe the topic content of the articles in our *corpus* we use Latent Dirichlet Allocation (LDA) [4], an unsupervised method which can describe at the same time the topic content of an article and the word content of a topic. LDA is a two level generative process, in which documents are linked to topic distributions and the *corpus* is modeled as a Dirichlet distribution on these latent topics. Given a vocabulary of V words this model assumes that each document in the *corpus* is generated by the following process:

1. The number of topics K , a K -dimensional vector α and a $K \times V$ matrix β are assumed as parameters of the model;
2. A K -dimensional random variable θ is selected from the $(K-1)$ -simplex with a Dirichlet probability density having parameters α :

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_K^{\alpha_K-1} \quad (1)$$

where $\Gamma(x)$ is the Gamma function;

3. Vector θ is used as a parameter for a multinomial distribution, used to pick the topic z from the K available;
4. Topic z is used to condition another multinomial distribution with parameter β , which is used to extract a word w from the vocabulary.

Along this process the K dimensional vector α has components $\alpha_i > 0$ and β is the probability matrix of selecting the word w_i once a topic z_j is chosen: $\beta_{ij} = p(w_i|z_j)$.

Given the parameters α and β , we obtain a joint probability distribution for the set of generated words w and the latent variables characterizing the topic z . Let the document be composed of N words, we obtain joint probability:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^N p(z_{k_n}|\theta) p(w_n|z_{k_n}, \beta) \quad (2)$$

where z_{k_n} denotes the topic selected for the word w_n . Integration over the latent variables yields the probability distribution for the set of N words w :

$$p(w|\alpha, \beta) = \int d\theta p(\theta|\alpha) \prod_{n=1}^N \sum_{k_n=1}^K p(z_{k_n}|\theta) p(w_n|z_{k_n}, \beta) . \quad (3)$$

LDA posits a fixed number of topics in a document collection and assumes that each document reflects a combination of those topics. The process can be reverted: in particular we are interested, given a set of documents, in deriving the topic distribution more suitable to synthesize their content. In principle we are interested in the distribution of the latent variables given the observed *corpus*:

$$p(\boldsymbol{\theta}, \mathbf{z} | \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} | \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w} | \boldsymbol{\alpha}, \boldsymbol{\beta})} \quad (4)$$

Even if the last expression is in general analytically intractable, a number of approximate inference algorithms can be used to find a solution. When a document collection is analyzed under these assumptions, these inference algorithms reveal an embedded thematic structure. With this structure, LDA provides a way to quickly summarize, explore, and search massive document collections.

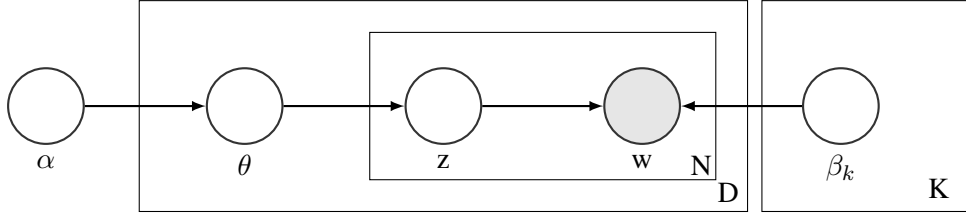


Figure 3.1: Graphical model representation of LDA.

The probabilistic graphical model in 3.1 reveals the nested structure of the LDA assumptions. LDA is composed of a hierarchy of mixture models. Each document is modeled with a finite mixture model, where the mixture proportions (i.e. the topic proportions) are drawn uniquely for each document but the mixture components (i.e. the topics) are shared across the collection. This is known as a grade of membership or mixed membership model in statistical theory [25]. LDA builds on seminal work in psychology [23] and machine learning [35]. It has close links to classical principal component analysis [18].

The most common way to evaluate a topic model is to compute the log-likelihood of a hold-out test set. This is usually done by splitting the dataset in two parts: one for training, the other for testing. For LDA, a test set is a collection of unseen documents $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$, and the model is described by the topic matrix $\boldsymbol{\beta}$ and the topic weights $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_d\}$ for every document. We need to evaluate the log-likelihood:

$$\mathcal{L}(\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}) = \log p(\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\} | \boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{j=1}^d \log p(\mathbf{w}_j | \boldsymbol{\beta}, \boldsymbol{\alpha}) \quad (5)$$

The measure traditionally used for gauging the goodness of fit of topic models is the *perplexity* of the held-out documents, defined as:

$$perplexity(\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}) = \exp\left\{-\frac{\mathcal{L}(\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\})}{\sum_{j=1}^d N_j}\right\} \quad (6)$$

where N_j is the number of words contained in the j -th document. The perplexity is a monotonic decreasing function of the log-likelihood $\mathcal{L}(\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\})$ of the unseen documents, such that minimizing the perplexity is tantamount to maximizing the likelihood function.

4 Italian Dictionary for Sentiment Analysis

One of the most basic information contained in a text is the sentiment expressed about a given subject. This is often also among the most interesting feature one would want to capture in any text analysis. For these reasons *sentiment analysis* is one of the most common applications of natural language processing [5, 6].

In its simplest formulation the goal of a sentiment analysis application is to transform any piece of text into a value on an ordered scale, representing the amount of positivity or negativity carried by the text. The usual

formats of the output are of two kinds: continuous, for example a number in the interval $[0, 1]$, or discrete, like a number of stars between 1 and 5.

The main approaches to sentiment analysis are of two kinds: rule-based algorithms and machine learning models. In the first category of applications the researcher identifies a set of keywords or features carrying the sentiment and maps each of them to a score value. One can then sum or average the scores for every identified feature to obtain a value representing the whole document. Some of the benefits of this approach - also called *vocabulary based* - are its transparency and ease of use: given that the vocabulary is built by the user he has complete control over it and can justify why each feature has a given score. In addition it can be very generic: keywords like “excellent” and “very good” have a positive connotation in most context, so a score based on such keywords will be mostly domain-independent. Another substantial advantage of a rule-based method over a machine learning model is that the latter requires a previously labeled set of documents to be trained, while the former can assign a score to documents without any previous knowledge (apart from the researcher’s knowledge).

Machine learning models leave the identification of features and the assignment of scores to an automated algorithm, trained to reproduce the scores of a given set of previously labeled documents. For example one could use as training set a sample of twitter feeds associated with one of the two hashtags *#good* or *#bad*, and train a supervised algorithm to replicate the given labeling. As already mentioned the use of this class of methods is bound to the availability of a large set of previously labeled documents, and the result can be domain specific. This can be of course both an advantage or an obstacle: the automated algorithm can pick nuances of the language which cannot be decoded in a vocabulary, but the interpretation of these nuances is usually dependent on the context in which they are used, and therefore on the document set used to train the model. A machine learning approach is useful only whenever the model can be trained on a class of documents similar to the ones the model has to label.

For our analysis we used a rule-based approach, employing a vocabulary built in [7], where the authors aimed at defining a vocabulary specialized on economic and financial language. This vocabulary is based on a self consistent algorithm which takes into account scores for synonyms and antonyms of any given word, in order to enforce a coherent score assignment. The authors evaluated the performance of their dictionary using as benchmark the Open Polarity Enhanced Named Entity Recognition (OpeNER) vocabulary [8], obtaining better results on every test performed.

5 Forecasting and Benchmarking Banking performances

Gauging the relevance of the information contained in news articles in explaining economic fluctuations of banking variables is of the utmost importance at both macro and micro-economic level. The sole yard stick we take into account is the ability of news to improve the predictive power toward balance sheet variables. We synthesize the information contained in the articles using the tools introduced in the previous sections: topic and sentiment analysis. With this approach we obtain a handful of variables representing the topic content and sentiment expressed in any given article, and these variables can be used as predictors for the behavior of indices related to four important Italian banks and the Italian stock index FTSE MIB.

To study the predictive power of the news variables we analyzed the period from the beginning of 2015 to May 2019².

Over this period we applied a 3-month wide moving window width to carry out an out-of-sample analysis: we performed an LDA analysis over the set of articles published in a first three month window, extracted the topics, and studied the weights of these topics in articles appearing in the following 3-month interval. Those weights, along with the sentiment extracted from the articles, are used to perform prediction of daily trading volumes and volatility.

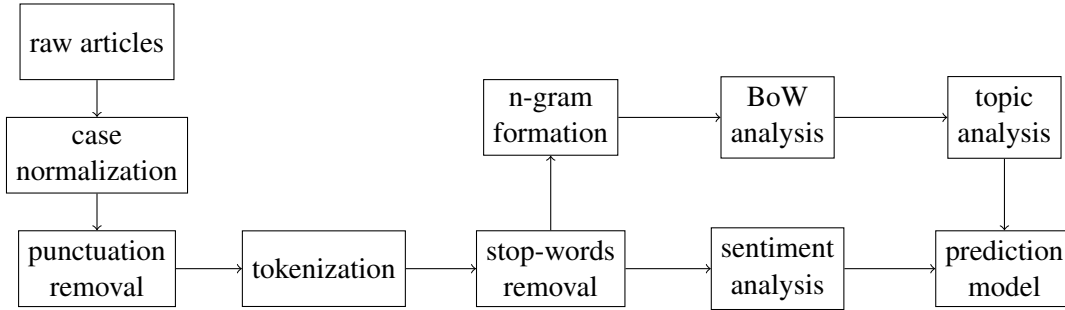
More in detail, the process consisted in dividing the whole period T going from Q1-2015 to Q2-2019 in

²This is the longest period for which all the data analyzed are available.

3-month windows m_j , $j \in \{1, 2, \dots, 18\}$ (we have 18 quarters in the period considered). Then, for every j going from 2 to 18, we applied the following steps:

- we fit the LDA model on the *corpus* of articles appearing in the window m_{j-1} , to obtain the most important topics in the period;
- we projected every article appearing in the window m_j on the topics obtained in the last step, to obtain a topic distribution for every article;
- we performed the vocabulary based sentiment analysis on every article in the window m_j ;
- we pooled all the articles relative to a single day, taking into consideration the average of the sentiment score and of the topic distribution for that day;
- we used the daily topic weights and sentiment score as predictor variables for the trading volumes and volatilities in the next day.

The following flow diagram shows a picture of the above described process, together with the preprocessing steps:



The topic analysis was performed with different choices of parameters, in order to minimize the perplexity score, enhance the interpretability of the topics, and increase the predictive power of the model. Removing the word stems in the preprocessing phase do not decrease sensibly the perplexity of the LDA, but in turn creates much less intelligible topics. For this reason we decided to skip the stemming step in the preprocessing phase.

We have a comparable number of articles in every time window, so we opted to select a constant number of topics over the whole period T . The minimum number of topics giving a satisfying perplexity was $n_{topics} = 3$, while the maximum was $n_{topics} = 5$, above this value the perplexity did not show any appreciable decrease. We took in consideration two types of models: one with single-word tokens, and a second in which bi-grams were considered. Both of them gave satisfactory results.

After the LDA and sentiment analysis the output variables were used as predictors in an enhanced autoregressive model, to test their forecasting accuracy using a benchmark autoregressive model of order one. The forecasting accuracy of the news was analyzed by running a battery of out of sample tests for the outcome variables volatility and trading volume for four of the most relevant Italian banks and the Italian stock index FTSE MIB. In a first phase both full-sample and out-of-sample predictive power were tested. The former approach however suffers from a possible look-ahead bias: the topics are obtained from an LDA performed on articles taken from the whole period, thus considering also articles appeared after the prediction date. In principle the topics obtained in this way could contain information about events in the future with respect of the day on which we want to make a forecast, thus introducing a possible look-ahead bias. Moreover the topics computed over the whole period T have another flaw: with a static description each topic has to describe some aspect of news appearing over the span of years. As such, these topics can only capture generic features, appearing in a large portion of the period T . Any “local” topic pattern would be lost with this approach, being them not persistent enough to be noticeable; this kind of information however is the most interesting for our goals, carrying the most predictive power for events in the short time range. For these reasons we focused on the results obtained with the out-of-sample method explained above in which

topics are defined using only a 3-month period preceding the prediction date.

In order to evaluate the forecasting power of the different selected topics we have compared the following regression models:

$$y_t = a + b y_{t-1} + c Sent_t + \sum_{j=1}^K d_j z_{j,t} \quad (7)$$

where the variables $z_{j,t}$ are the weights of each topic resulting from the LDA and $Sent_t$ is the average sentiment score for the day t . This equation has been estimated by the LASSO penalized regression [9, 10], in order to retain only variables with sufficient explanatory power. This method provides automatic variable selection, assigning a cost to any variable considered in the regression and thus forcing the selection of only the most useful ones. In the estimation step we fed the model with all the topic and sentiment variables, and then retained only the ones with predictive power as given by the LASSO regression. The forecasting accuracy of equation 7 has been compared with an AutoRegressive benchmark:

$$y_t = a + b y_{t-1} \quad (8)$$

If any of the variables among sentiment and topics are retained in the regression with some degree of statistical significance, we conclude they are good predictors of stock market movements for the day following the news coverage.

The increase of forecasting accuracy between the benchmark autoregression and our model including topics and sentiment has been evaluated using three classical indices. Let y_1, \dots, y_T denote *actual* values, and $\hat{y}_1, \dots, \hat{y}_T$ be the corresponding predictions:

1. Root Mean Square Error (RMSE)

$$\sqrt{\frac{1}{T} \sum_{t=1}^T (\hat{y}_t - y_t)^2};$$

2. Mean Absolute Percentage Error (MAPE):

$$\frac{1}{T} \sum_{t=1}^T \left| \frac{\hat{y}_t - y_t}{y_t} \right|;$$

3. Mean Directional Accuracy (MDA):

$$\frac{1}{T} \sum_{t=1}^T \mathbb{I}_{\text{sign}(y_t - y_{t-1}) = \text{sign}(\hat{y}_t - y_{t-1})}.$$

Additionally, we tested the hypothesis of equal forecasting accuracy between our models and the benchmark against the alternative of improved performance of the first with Diebold-Mariano (DM) test.

In table 1 and table 2 we show the previously mentioned statistical measures of the significance of the topics and sentiment coefficients in the regressions. As we can see the Diebold-Mariano test has always indicated a rejection of the null hypothesis of no difference between the forecasts produced by the AR model and those of the competing models which included the topic and the weighted sentiment:

Table 1: Comparison of forecasting accuracy for the Volume

| | Topics | n-gram | MAPE | Rel. RMSE | MDA | DM |
|-----------------|--------|--------|--------|-----------|------|----------|
| BMPS | 4 | 1 | 45.97% | 0.50 | 0.73 | 2.82 ** |
| | 5 | 1 | 46.18% | 0.51 | 0.73 | 2.77 ** |
| | 3 | 2 | 48.22% | 0.50 | 0.75 | 2.56 ** |
| | 4 | 2 | 46.40% | 0.50 | 0.73 | 2.80 ** |
| FTSE MIB | 4 | 1 | 19.18% | 0.76 | 0.87 | 3.93 *** |
| | 5 | 1 | 19.10% | 0.76 | 0.87 | 3.95 *** |
| | 3 | 2 | 19.29% | 0.76 | 0.86 | 4.06 *** |
| | 4 | 2 | 19.60% | 0.77 | 0.88 | 3.47 *** |
| ISP | 4 | 1 | 26.25% | 0.78 | 0.84 | 4.56 *** |
| | 5 | 1 | 26.22% | 0.78 | 0.84 | 4.55 *** |
| | 3 | 2 | 27.27% | 0.77 | 0.86 | 4.05 *** |
| | 4 | 2 | 27.49% | 0.81 | 0.87 | 3.41 *** |
| UBI | 4 | 1 | 30.18% | 0.72 | 0.85 | 3.65 *** |
| | 5 | 1 | 29.57% | 0.71 | 0.83 | 4.11 *** |
| | 3 | 2 | 31.16% | 0.71 | 0.85 | 3.55 *** |
| | 4 | 2 | 30.35% | 0.72 | 0.86 | 4.22 *** |
| UCG | 4 | 1 | 28.23% | 0.78 | 0.85 | 4.78 *** |
| | 5 | 1 | 28.26% | 0.79 | 0.86 | 4.01 *** |
| | 3 | 2 | 29.21% | 0.79 | 0.87 | 4.01 *** |
| | 4 | 2 | 28.55% | 0.79 | 0.86 | 4.22 *** |

Table 2: Comparison of forecasting accuracy for the Volatility

| | Topics | n-gram | MAPE | Rel. RMSE | MDA | DM |
|-----------------|--------|--------|--------|-----------|------|----------|
| BMPS | 4 | 1 | 41.35% | 0.55 | 0.40 | 3.04 ** |
| | 5 | 1 | 41.93% | 0.56 | 0.41 | 2.97 ** |
| | 3 | 2 | 44.68% | 0.56 | 0.44 | 2.90 ** |
| | 4 | 2 | 44.28% | 0.56 | 0.43 | 2.74 ** |
| FTSE MIB | 4 | 1 | 33.94% | 0.79 | 0.42 | 6.35 *** |
| | 5 | 1 | 33.95% | 0.80 | 0.40 | 6.19 *** |
| | 3 | 2 | 35.55% | 0.80 | 0.41 | 5.87 *** |
| | 4 | 2 | 36.20% | 0.83 | 0.41 | 5.03 *** |
| ISP | 4 | 1 | 35.19% | 0.82 | 0.39 | 6.27 *** |
| | 5 | 1 | 35.68% | 0.83 | 0.41 | 5.56 *** |
| | 3 | 2 | 37.59% | 0.84 | 0.42 | 5.11 *** |
| | 4 | 2 | 37.02% | 0.83 | 0.40 | 5.23 *** |
| UBI | 4 | 1 | 33.11% | 0.74 | 0.44 | 6.47 *** |
| | 5 | 1 | 33.98% | 0.75 | 0.41 | 5.86 *** |
| | 3 | 2 | 36.23% | 0.75 | 0.40 | 5.43 *** |
| | 4 | 2 | 35.02% | 0.75 | 0.42 | 5.62 *** |
| UCG | 4 | 1 | 34.90% | 0.82 | 0.40 | 5.27 *** |
| | 5 | 1 | 34.98% | 0.84 | 0.40 | 4.73 *** |
| | 3 | 2 | 35.72% | 0.82 | 0.41 | 5.09 *** |
| | 4 | 2 | 35.99% | 0.82 | 0.40 | 4.73 *** |

This means that in our model at least one among the topic and sentiment variables has a significant predictive power both for the volumes and volatilities over the period taken into consideration.

Another measure of the enhanced predictive power of our model over the AR benchmark are the *cumulative sum of squared error differences* (CSSSED). They show how over the period considered the errors of the topic-sentiment model are consistently lower than the ones of the benchmark model:

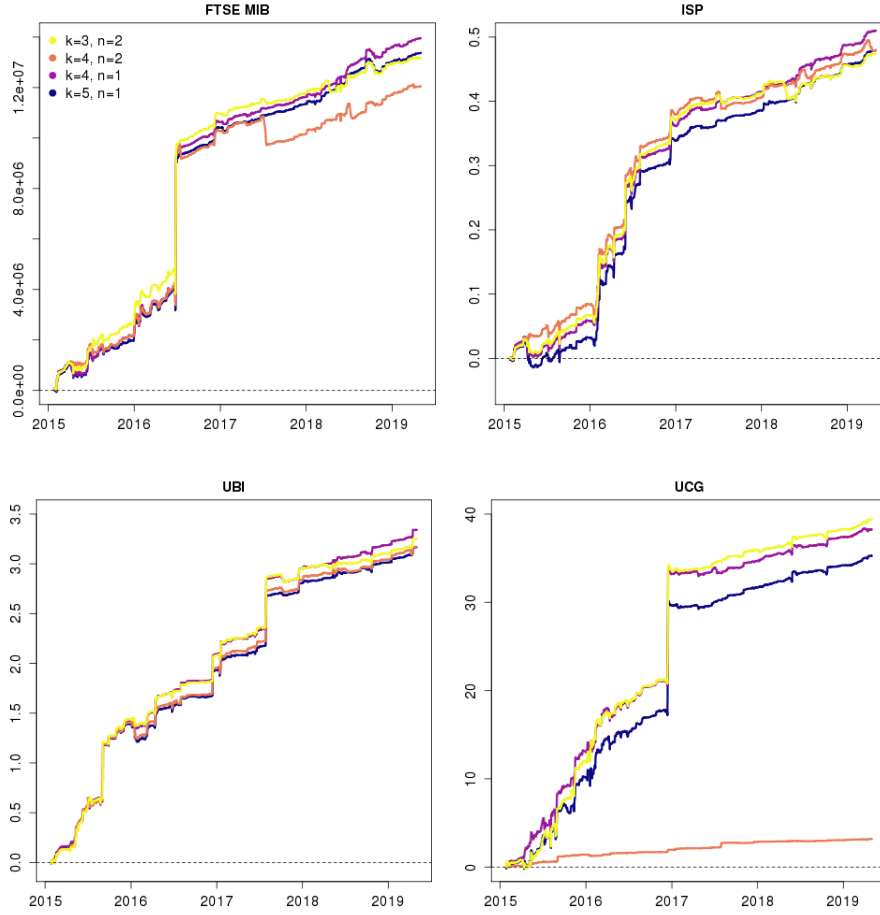


Figure 5.1: CSSED: Volatility

In the plots we show the CSSED for the volatility predictions for the class of models with $k = 3, 4$ and 5 as number of topics considered in the regression. The model with $n = 1$ is the one in which the tokens are single words, while in the model with $n = 2$ bigrams are considered in the topic definition. We show the results for the FTSE MIB index and three of the four banks studied³. As we can see the models comprising topic and sentiment variables always perform better than the benchmark on average, and performances are better in most of the sub-periods considered.

Finally, two of the most useful pieces of information we can extract from our model are the content of the topics with most predictive power, and the co-occurrences of variables in any given period. These information tell us what is really predicting the evolution of the target variables, allowing to make a connection between news content and market variables movements. As an example, in fig. 5.2, we show for the FTSE MIB index and the four banks considered a co-occurrence graph and a word-cloud with the content of the most predictive topic variable:

³The fourth was temporarily suspended from trading, hence the comparison is not available for the full period.

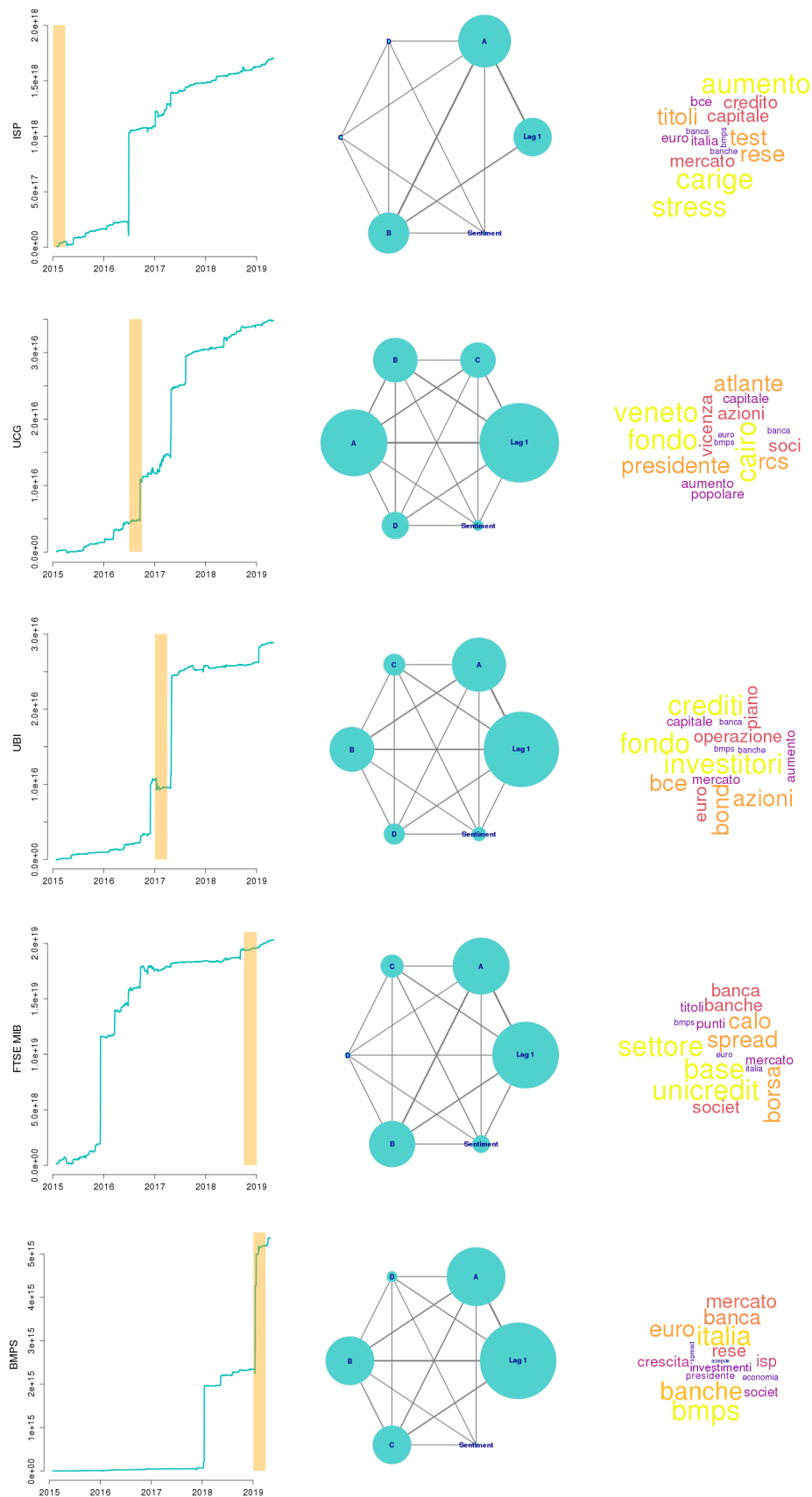


Figure 5.2: Variables co-occurrences and topics word-clouds

In the first column of the table the quarter in which the predictions are being made is highlighted, and the effect of the topic and sentiment variables on the predictive power is made clear by the CSSED. In the second column the variables with most predictive power and their connections are shown: the diameter of a circle is proportional to the number of days of the quarter in which the variable in question was retained by the LASSO, hence the number of days in which its presence was useful to make predictions. The thickness of the edges of the graph shows how often a couple of variables was retained in the model together: variables considered in the model often together are connected by wider links. We can see that though the autoregressive variable is kept in all the quarters showed, usually it comes together with other, news-related, variables.

Finally the third column of the table shows the words contained in the most predictive topic for that quarter, which allows us to check the relevant news content in the predictions. We can see that the model is able to pick and describe topics “trending” in the media during the periods considered. This confirms the fact that our model is able to exploit the most important news, as described by the sources considered, in order to perform better predictions.

6 Concluding Remarks

In this work we have shown how to extract relevant signals from textual data useful to forecast the main variables for the banking market in Italy.

We have compared the forecasting performance for the volatility, rate of return and exchanged volumes for four sistemically important banks and for the FTSE-MIB index for the Italian stock market.⁴ In all the examples, we have adopted an AR model whose order has been selected on the base of an Information criteria.

For the volatility and the rate of returns we have systematically achieved improvements for the MAPE and the Relative RMSE. The improvement is significant, though not dominant for the MDA.

When choosing 4 topics, we have seen that on about 90% of the times the Lasso regression selects at least one topic as significantly relevant. The sentiment results significant around the 30% of the shown regressions.

Robustness of our results has been checked by running the Diebold-Mariano test which has always indicated a rejection of the null hypothesis of no difference between the forecasts produced by the AR model and those of the competing models which included the topic and the weighted sentiment.

We aim to extend our work to explore also the role of the sentiment on the relevant variables composing the banks’ balance sheet. In addition the sentiment probed by our approach was extracted from articles appearing in specialized financial journals. Given the nature of the source the sentiment extracted is rarely strongly polarized. A more expressive sentiment variable could be obtained studying the public discourse regarding a given bank on social networks. This further direction is left for future studies.

⁴FTSE MIB is the benchmark stock market index for Italian national stock exchange. The index consists of the 40 most-traded stock classes on the exchange.

References

- [1] Paul C. Tetlock. “Giving content to investor sentiment: The role of media in the stock market”. In: *The Journal of finance* 62.3 (2007), pp. 1139–1168.
- [2] Tim Loughran and Bill McDonald. “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks”. In: *The Journal of finance* 66.1 (2011), pp. 35–65.
- [3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009.
- [4] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.
- [5] Minqing Hu and Bing Liu. “Mining and summarizing customer reviews”. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, pp. 168–177.
- [6] Bo Pang, Lillian Lee, et al. “Opinion mining and sentiment analysis”. In: *Foundations and Trends® in Information Retrieval* 2.1–2 (2008), pp. 1–135.
- [7] Giuseppe Bruno et al. “The Sentiment Hidden in Italian Texts Through the Lens of A New Dictionary”. In: *2nd International Conference on Advanced Research Methods and Analytics (CARMA 2018). Proceedings*. 2018.
- [8] Rodrigo Agerri et al. “OpeNER: Open polarity enhanced named entity recognition”. In: *Procesamiento del Lenguaje Natural* 51 (2013), pp. 215–218.
- [9] Fadil Santosa and William W Symes. “Linear inversion of band-limited reflection seismograms”. In: *SIAM Journal on Scientific and Statistical Computing* 7.4 (1986), pp. 1307–1330.
- [10] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

News and banks' equities: do words have predictive power?

Valerio Astuti, Giuseppe Bruno, Sabina Marchetti & Juri Marcucci¹

Bank of Italy

IFC-Bank of Italy workshop on “Data science in central banking” – Part 2: Data Science in Central Banking: Applications and tools

February 15, 2022



¹The views expressed in this presentation are the authors' only and do not necessarily reflect those of the Bank of Italy.

Motivations and Main steps

Motivations

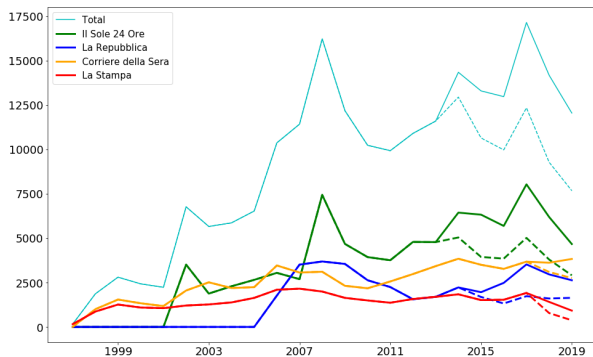
- Can we extract useful quantitative information from narrative content in newspaper?
- Are news a predictive factor in banks' equities trends?
- Is there any advantage in putting together text and classical balance sheet indicators?

Main Steps

- Starting point: building an archive of newspaper articles talking about the main Italian banks
- Pre-processing of articles, topic and sentiment analysis
- Application to predictive models for banks' trading volumes
- Memory-intensive tasks: Python and PySpark

Our Database

Number of articles per source



- From **Dow Jones Factiva** articles on the **100 most important Italian banks**
- *From September 1996 to May 2019* (article number not uniformly distributed over time)
- **Sources:** “Il Sole 24 Ore”, “La Stampa”, “La Repubblica”, “Corriere della Sera” plus online editions (long-dash lines!)
- **Corpus:** 217K articles, 100M words, 0.33M unique tokens (Zipf’s law)
- Case normalization, tokenization, stop-words removal, stemming
- Cut-off on minimum number of appearances of a term

Latent Dirichlet Allocation (LDA)

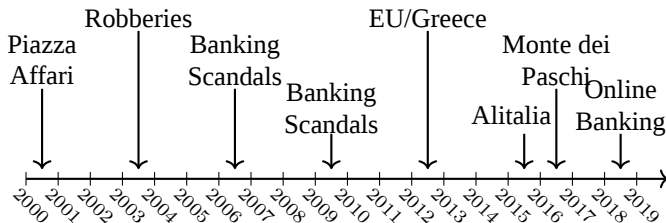
- Unsupervised, hierarchical probabilistic model to decompose a document in its most salient topics (probability distribution over words)
- Full sample (synchronic) and rolling subsample (diachronic): one defined over whole period, the other limited to three-year spans, rolling yearly. The number of topics is chosen to minimize perplexity
- Rolling sample was used to sidestep two possible problems: look-ahead bias and coarsening of topics over a 20-year span
- Full sample: minimum perplexity = 7.93, **number of topics = 15**
- Rolling sample: average perplexity = 7.83, **average number of topics = 8.75**

LDA Results: Main Topics → Full sample vs Rolling subsamples

Main Topics in Full Sample

| | |
|----------------------|---------------------|
| Economy-Politics | Italian Groups |
| Investigations | Public |
| Industry | Balance/Capital |
| Stock Exchange | Growth and Taxes |
| Local Activities | Investments |
| English articles | Stock Market Trends |
| News Reports | Boards |
| Financial Activities | |

Main Topics in Rolling Sub-samples

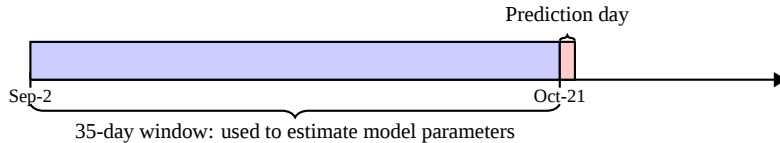


Model

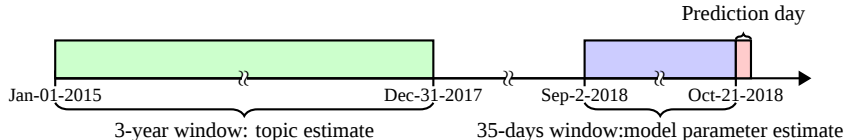
- Topics predictive power tested on stock market indices of 4 Italian banks and the Italian stock index FTSE MIB
- We analyzed returns, volatilities, and volumes. Volumes are the most reactive variables to news, and the ones topics forecast better
- Applied topic distributions with both static and rolling samples, with different results
- Our model is a LASSO with an adaptive number of topics k_t updated daily and the possibility to keep up to three lagged variables
- The benchmark model is an $AR(p_t)$ with p_t selected to minimize the BIC at each t

Topic distribution

- The number of variables used as predictors in a given day is selected by the LASSO methodology over the previous 35-day period;
- Possibility of look-ahead bias using topic estimated with future articles but on the other hand topics much coarser given the definition on a longer timespan;



- In the topic model with rolling sample the predictive variables are estimated over the 3 calendar years preceding the prediction day (weighted with the daily sentiment).
- Having defined the topics, the regression coefficients are estimated in the 35 days before the forecast.



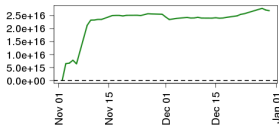
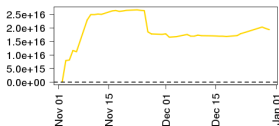
Out-of-sample performance: Cumulative Sum of Squared Error Differences (CSSED)

The performance of our models is evaluated through the difference between the CSSED of our models and the AR benchmark (if positive, our models are on average more accurate)

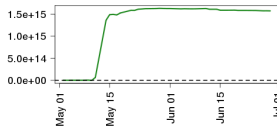
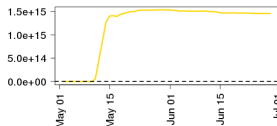
Upper panel → full sample

Lower panel → Rolling subsample

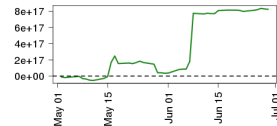
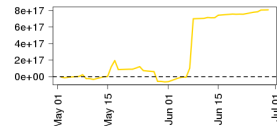
Bank 1, 2015



Bank 2, 2018



FTSE, 2018



Conclusions

- Topic analysis effectively captures relevant information content of newspaper articles
- Topics can be used as predictor variables for banks' equities
- Using the topics to make predictions, our models perform on average better than the AutoRegressive benchmark
- Tiny differences in terms of CSSED between using topics from the full sample and from rolling sub-samples

Thank You very much for Your Attention!

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Natural language processing for risk management¹

Bijan Sahamie,
Deutsche Bundesbank

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Natural Language Processing for Risk Management

Discussion of Use-Cases

Dr. Bijan Sahamie, Deutsche Bundesbank

Disclaimer

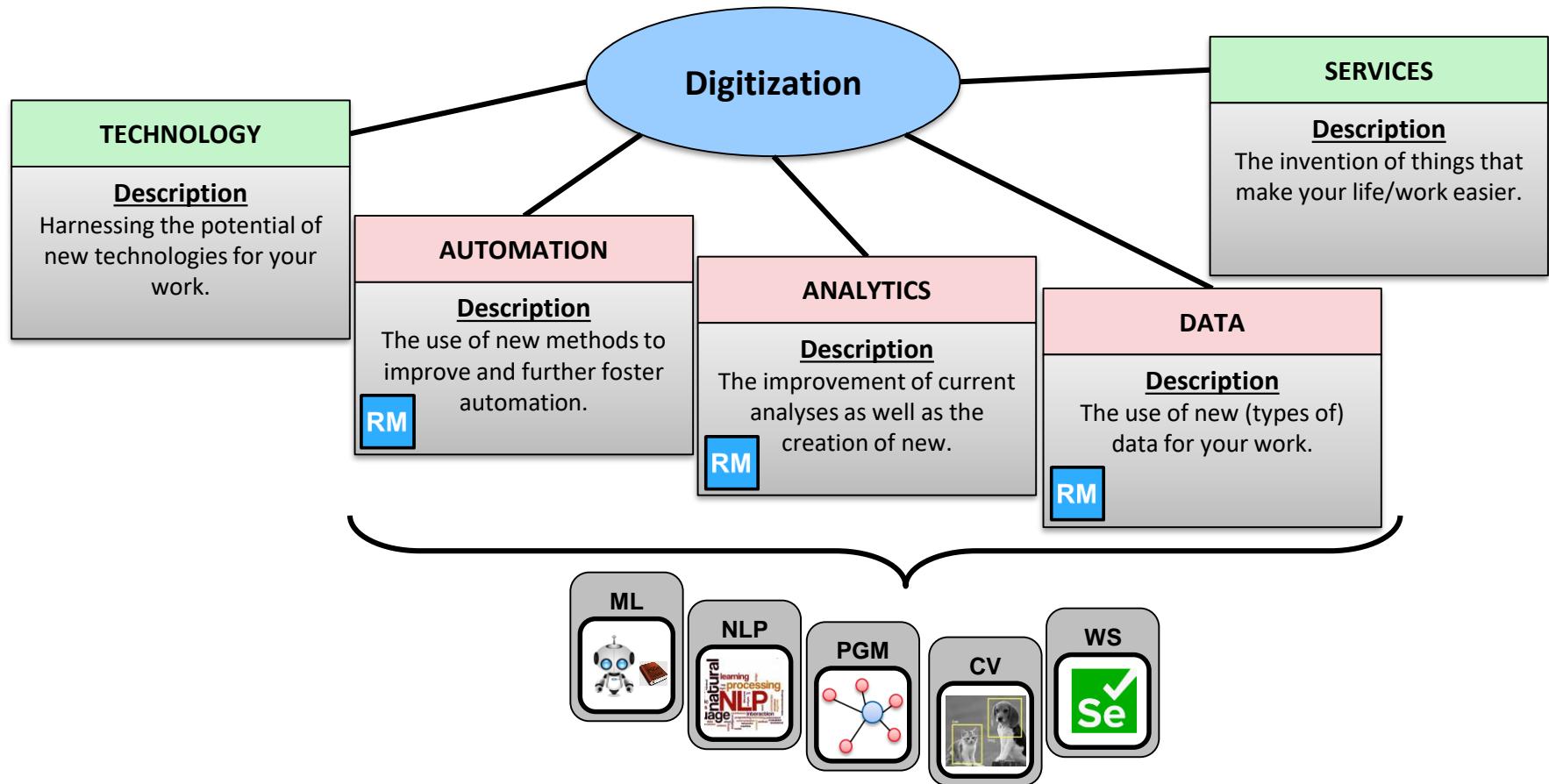
The opinions, views, facts, interpretations expressed within the confines of these slides are those of the speaker alone and have nothing to do with Deutsche Bundesbank. All opinions, views, facts, interpretations expressed by the speaker orally during the talk, before, or thereafter are explicitly those of the speaker himself and are not in any way connected with Deutsche Bundesbank.

All mistakes, bad impressions, all weirdness or non-professionalism you perceive either from these slides or the speaker do not in any way represent Deutsche Bundesbank and are the sole responsibility of the speaker.

Let's have a very high-level overview

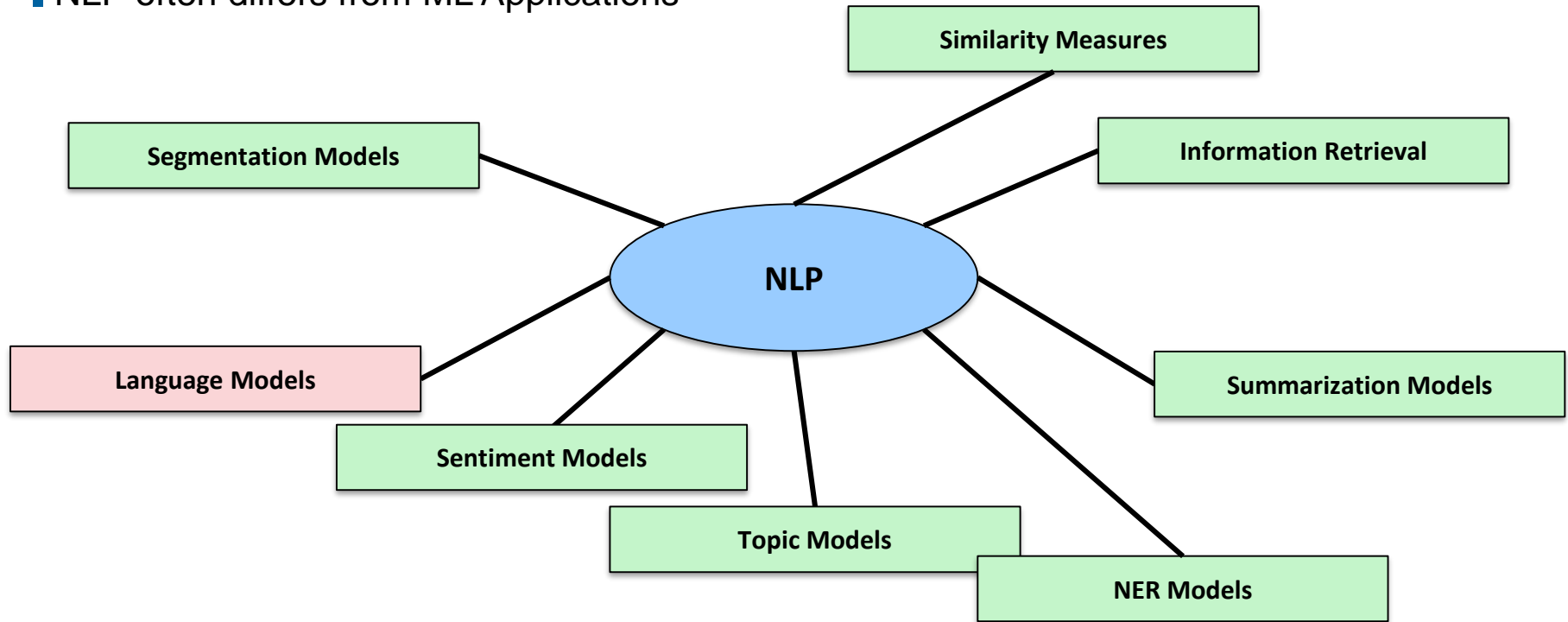
What are the main areas of interest and which technologies might be a good choice?

Let us have a look at the main areas of Digitization



Natural Language Processing

NLP often differs from ML Applications



Two Properties of NLP that stand out.

- Your use-case, although it might be specialized, is an (immediate) derivative of a general purpose task. Both are intimately related.
- Inside your IT-system there are several intertwined machine learning components. The machine learning system might even work as a hierarchy.

Counterparty Risk

Use-Cases in Risk Control

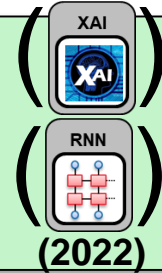
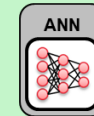
There are use-cases for both machine learning and natural language processing in Risk Control

Prediction of Financial Distress using Accounting Data

What's it about: „Classic“ forecasting using machine learning

Primary Utility: Generating a short-list for our analysts

Models: Ensemble of Neural Nets with calibrated thresholds (in the sense of a Outlier Detection)



ONLINE

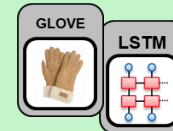
(2022)

(Fact-associated) Financial Sentiment of News Messages

What's it about: Automated assessment of news messages

Primary Utility : Workload reduction (i.e. improved efficiency)

Modelle: Mixture of Experts – Approach incl. LSTMs + Glove



ONLINE

(2022)

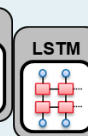
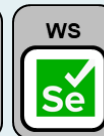
Filtering of News Messages sourced through webscraping

(Doing webscraping since nov 2019)

What's it about: Filter out uninteresting messages

Primary Utility: Better signal/noise-ratio, scalability of the approach

Models: (Self-Defined) Frequency Measures, Random Forests, LSTMs and/or GRUs + Glove.



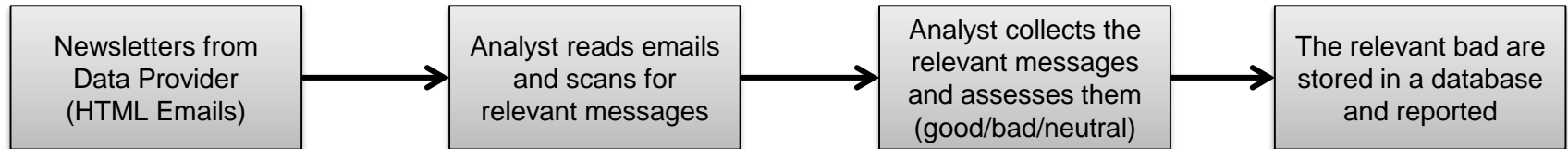
WORK IN
PROGRESS

(2022)

Financial Sentiment Analysis of News Data

Overview

Imagine the process. (very rough and simplified sketch, not entirely correct for the sake of brevity.)



Question. Can we make this more efficient?

Goal.

- Let a machine read the messages and just present to the analyst a pre-selected list (based on the system output)

Cost/Benefit compromise.

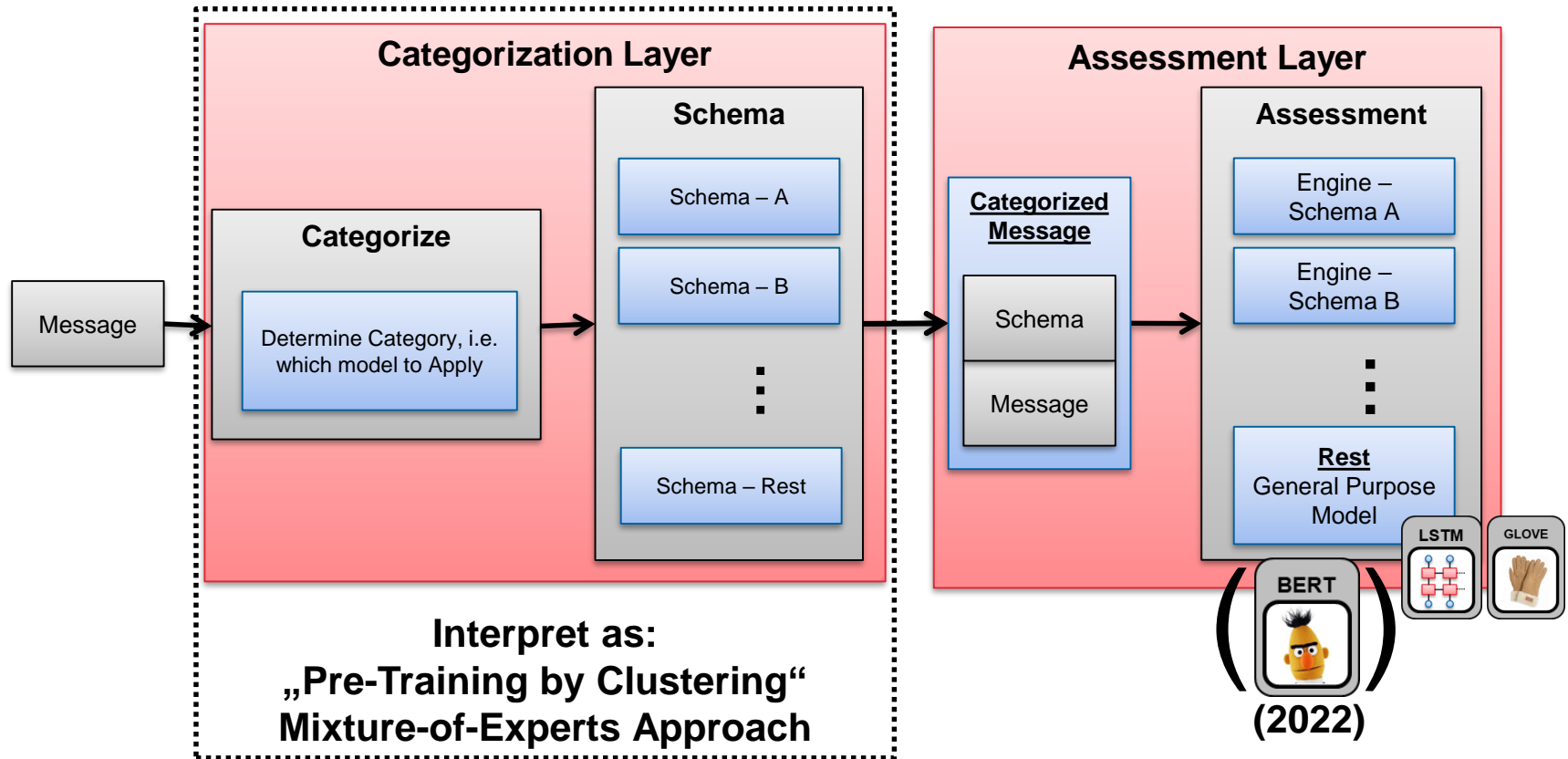
- Potential benefit: You lower the workload for the analyst
- Potential benefit: You raise the consistency of the process output
- Potential cost: You might lose some of the interesting messages.

ML-Dev Task.

- Find a sensible sweet-spot in the Cost/Benefit regime and optimize your system towards it

Financial Sentiment Analysis of News Data

The Model Concept




Financial Sentiment Analysis of News Data

A Screenshot

FWS News

FWS News Navigation



RICIS_News Write Lock

Status:

Locked by:

Lock

Unlock

Data Import and Log Checking

Analyst Assessment and Lookup

News Assessment and Lookup

Current User ██████████

Assessment and Lookup

| FileNo | No. | ImportTime Stamp | MessageDate | Section | Title | Message | Source | SystemAss | YourAss |
|--------|-------|---------------------|-------------|---------------------------|--|--|---------------|-----------|----------|
| 2739 | 47356 | 02.12.2021 10:06:59 | 02.12.2021 | BANKING | | Fitch Ratings removed Banca Monte dei Paschi di Siena's long-t... | Feature AI... | Negative | Not Set |
| 2739 | 47357 | 02.12.2021 10:06:59 | 02.12.2021 | BANKING | | KBC Bank Ireland PLC is engaging in detailed talks with Bank of ... | Feature AI... | Negative | Not Set |
| 2741 | 47372 | 03.12.2021 12:56:44 | 02.12.2021 | Capital Markets | Leveraged Commentary and Data: B2Holding postpones €300M ... | The deal is the second transaction in European high-yield to be p... | SNL MyS... | Negative | Not Set |
| 2741 | 47375 | 03.12.2021 12:56:44 | 02.12.2021 | Capital Markets | Leveraged Finance Trends: Sustainability-linked leveraged finan... | Fears regarding greenwashing have increased as the rate of Eur... | SNL MyS... | Negative | Not Set |
| 2741 | 47378 | 03.12.2021 12:56:44 | 02.12.2021 | Regulation, Policy & Law | Leveraged Commentary and Data: LMA and ELFA update best p... | Updates reflect important new guidance on sustainability provisio... | SNL MyS... | Negative | Not Set |
| 2742 | 47379 | 03.12.2021 12:56:44 | 03.12.2021 | TOP NEWS IN EUROPEAN ... | | The European Commission fined HSBC Holdings PLC, Credit Su... | Feature AI... | Negative | Negative |
| 2742 | 47382 | 03.12.2021 12:56:44 | 03.12.2021 | MARKET INTELLIGENCE IN... | | Insurers make progress on climate resilience at COP26, face fos... | Feature AI... | Negative | Not Set |
| 2742 | 47387 | 03.12.2021 12:56:44 | 03.12.2021 | BANKING | | HSBC CEO Noel Quinn warned of increased costs for banks und... | Feature AI... | Negative | Not Set |
| 2742 | 47391 | 03.12.2021 12:56:44 | 03.12.2021 | POLICY AND REGULATION | | The U.K. Financial Conduct Authority confirmed proposed chang... | Feature AI... | Negative | Not Set |
| 2742 | 47392 | 03.12.2021 12:56:44 | 03.12.2021 | POLICY AND REGULATION | | The latest stress test of Denmark's systemically important banks ... | Feature AI... | Negative | Not Set |
| 2742 | 47393 | 03.12.2021 12:56:44 | 03.12.2021 | POLICY AND REGULATION | | The Irish central bank fined Bank of Ireland Group PLC €24.5 mil... | Feature AI... | Negative | Not Set |

Purge Data from GUI

Load Messages

Date Selection

Selection

Date Type

from

till

Filter Properties

Sample

Checked

YourAss

Relevance

SystemAss

Counterparty Filter

CntrptNo.

CntrptName

Query

Reset

Chosen Counterparty

Message Details

| Positive | Neutral | Negative | SystemAss |
|----------|---------|----------|-----------|
| 16.6% | 18.2% | 65.2% | Negative |

Title

--

Message

The European Commission fined HSBC Holdings PLC, Credit Suisse Group AG, Barclays PLC and NatWest Group PLC a combined €261 million for participating in a foreign exchange spot trading cartel. Swiss bank UBS Group AG, which revealed the existence of the cartel, received full immunity and avoided a roughly €94 million fine, while U.K.-based peers HSBC, Barclays and NatWest Group received reduced fines for cooperating with the investigation. HSBC was hit with the largest fine at €174.3 million.

Edit Title

Edit Message

Mark for Deletion

Remove Mark

File Summary

Open File in Browser

Visualize

Type

System

System

System

System

System

RicisName

HSBC Holdings plc

Credit Suisse Group AG

Barclays plc

NatWest Group plc

UBS Group AG

Check out Entity

LinkNo

LinkType

Link

73942

TextBodyLink

<https://ec.europa.eu/commission/presscorner/detail/>

Open Link

Nothing to Save...

Down

Up

Relevance

Relevant

Not Relevant

Not Set

Sentiment

Positive

Neutral

Negative

Not Set

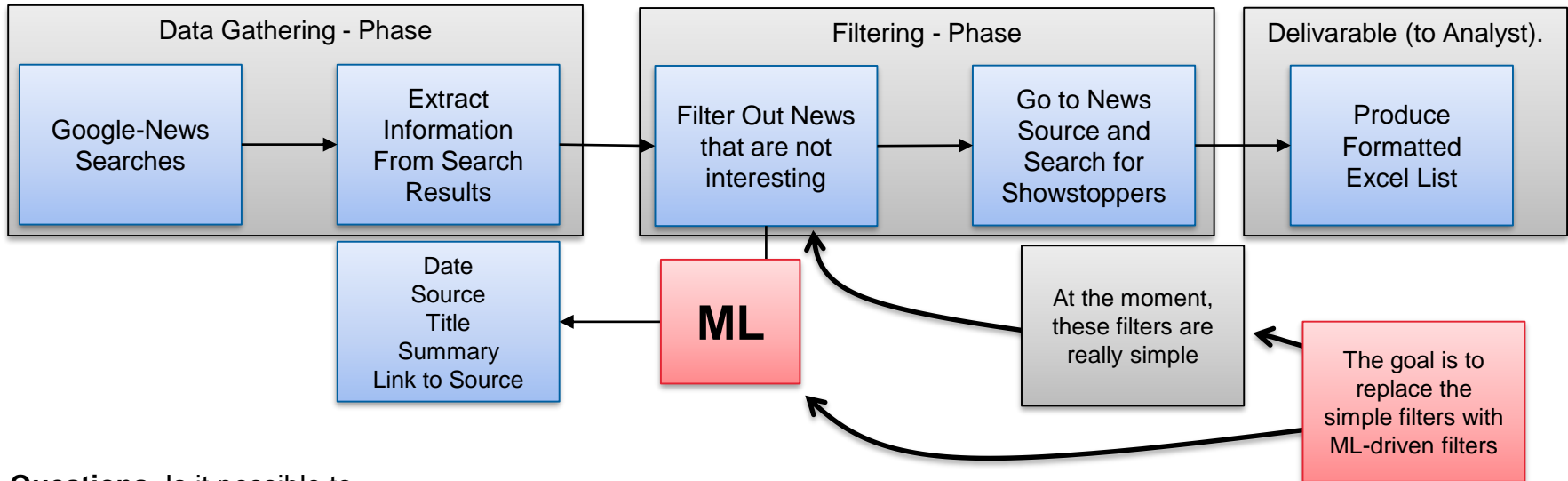
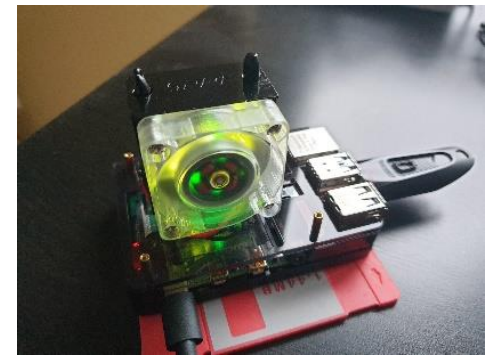
Dr. Bijar
01/02/2021
Page 6

Filtering Techniques for Webscraping of News

Filtering Techniques for Webscraping of News

Volume per Run (currently):

- ~15000 search results are scanned
- ~900 website jumps



Questions. Is it possible to...

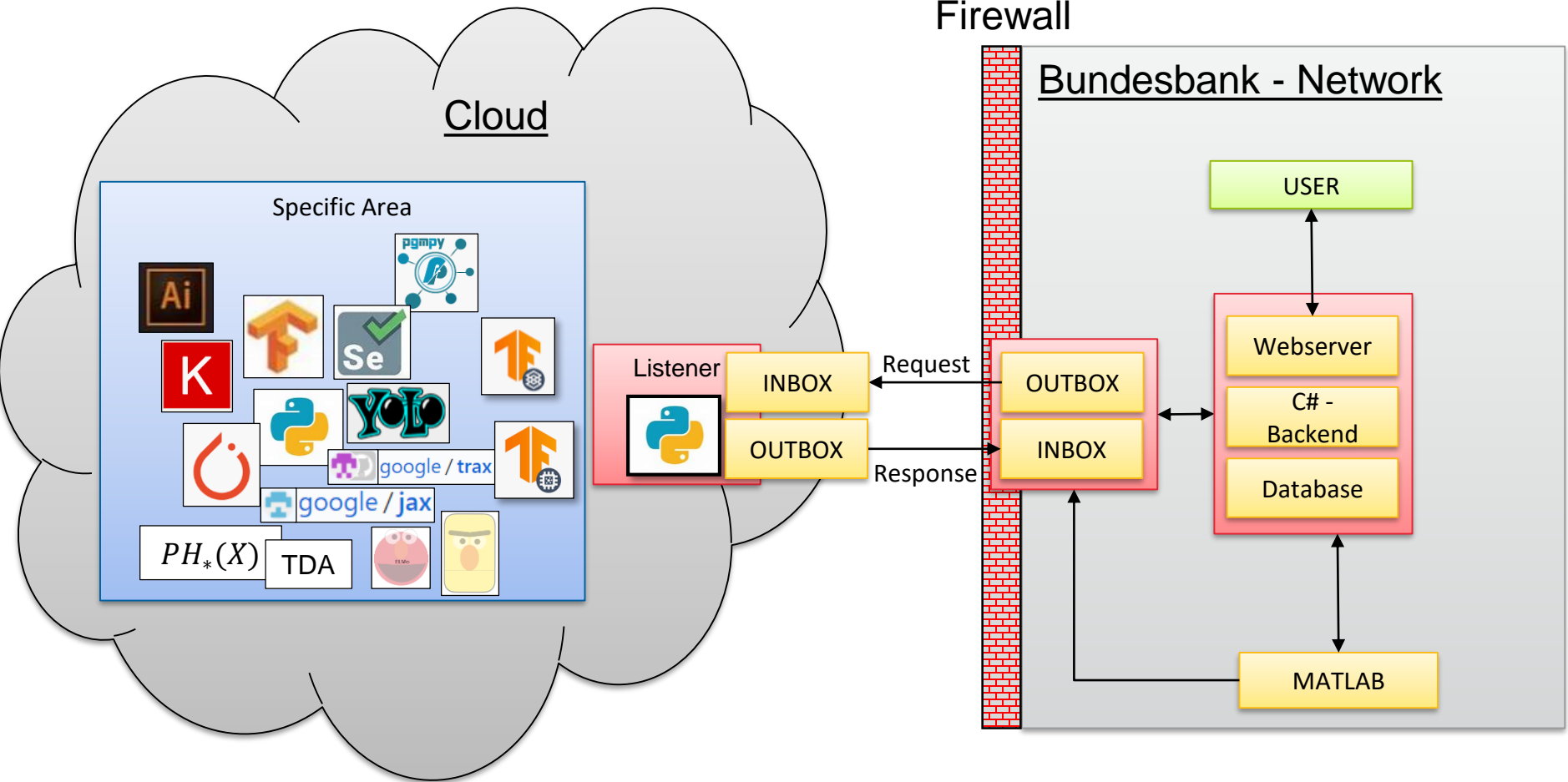
ML

- ... distinguish „useful“ from „non-useful“ sources by just looking at their name, and is this a learnable task?
- ... distinguish „useful“ from „non-useful“ sources by using frequency-measures?
- ... distinguish „useful“ from „non-useful“ messages by looking at the titles?
- ... distinguish „useful“ from „non-useful“ messages by looking at the summary?

Answer.
YES!

Filtering Techniques for Webscraping of News

(Target) IT – Architecture





The End

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Integrating natural language processing technologies to central bank operations at Bank of Thailand¹

Jiradett Kerdsri and Pucktada Treeratpituk,
Bank of Thailand

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Integrating Natural Language Processing To Central Bank Operations at Bank of Thailand

Pucktada Treeratpituk, Jiradett Kerdsri

Department of Data Management and Data Analytics
Bank of Thailand

15 Feb 2022



ธนาคารแห่งประเทศไทย
BANK OF THAILAND

NLP for BOT's Operations



Why NLP?

- Lots of textual data, ranging from unstructured economic data to financial documents to social media information that can be used to enhance efficiency & effectiveness of Central Bank operations.
- Additionally, due to the sheer volume, it is desirable to use ML to process these data.
- Data-driven Organization - data in many departments are mainly textual data.



Challenges

- Unstructured & Data Quality (eg. PDF scanned, social media, free-text data fields)
- Thai Language (characteristics of the language + lack of readily available tools).

ที่ผ่านมา ธปท. ได้ติดตามและเร่งบริษัทปรับระบบให้มีกระบวนการแสดงตนและพิสูจน์ตัวตนลูกค้าบุคคลธรรมดา (KYC) สอดคล้องตามหลักเกณฑ์ของกฎหมายป้องกันและปราบปรามการฟอกเงินและการสนับสนุนทางการเงินแก่การก่อการร้าย ซึ่งเป็นกระบวนการสำคัญที่จะช่วยดูแลให้การทำธุรกรรมทางการเงิน มีความปลอดภัย ไม่เป็นช่องทางให้เกิดการทุจริตหรือฟอกเงิน ที่เป็นภัยต่อประชาชนและระบบการเงินของประเทศ

example



Projects

1. Behavioral & Culture Supervision through Board of Directors' meeting minutes.
2. Automate Statistics Compilation of Government Expenditure.
3. Assessing and Managing BOT's Reputational Risk through social media analysis.



ธนาคารแห่งประเทศไทย
BANK OF THAILAND

1. Behavioral & Culture Supervision

Assessing organizational behavior and culture (B&C) of financial institutions to promote good corporate governance

B&C Toolbox



1. Self-Assessment



2. Board Observation



3. Interview & Dialogue



4. Desk Research



5. Expectation Framework



6. Enforcement



**Board effectiveness
& Risk Culture**

Because BOT cannot directly observe board meetings, BOT evaluates the Board's Effectiveness through **board survey** and analyzing board **meeting minutes**.

board of directors
meeting minutes



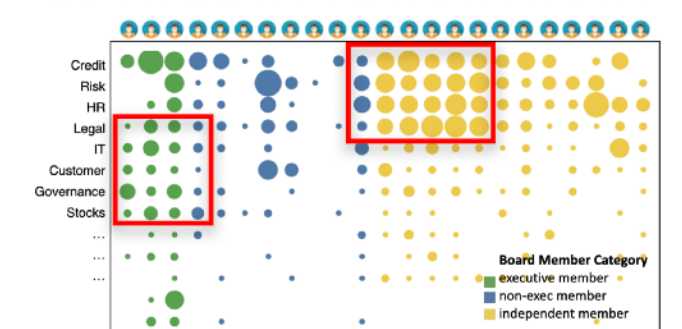
Thai
OCR



Document
Structure Parsing
(heuristics, **HMM**)



Thai Word
Segmentation,
Topic Modeling,
Name-Entity
Extraction



Topics of
Discussion



Participation of
Board Members

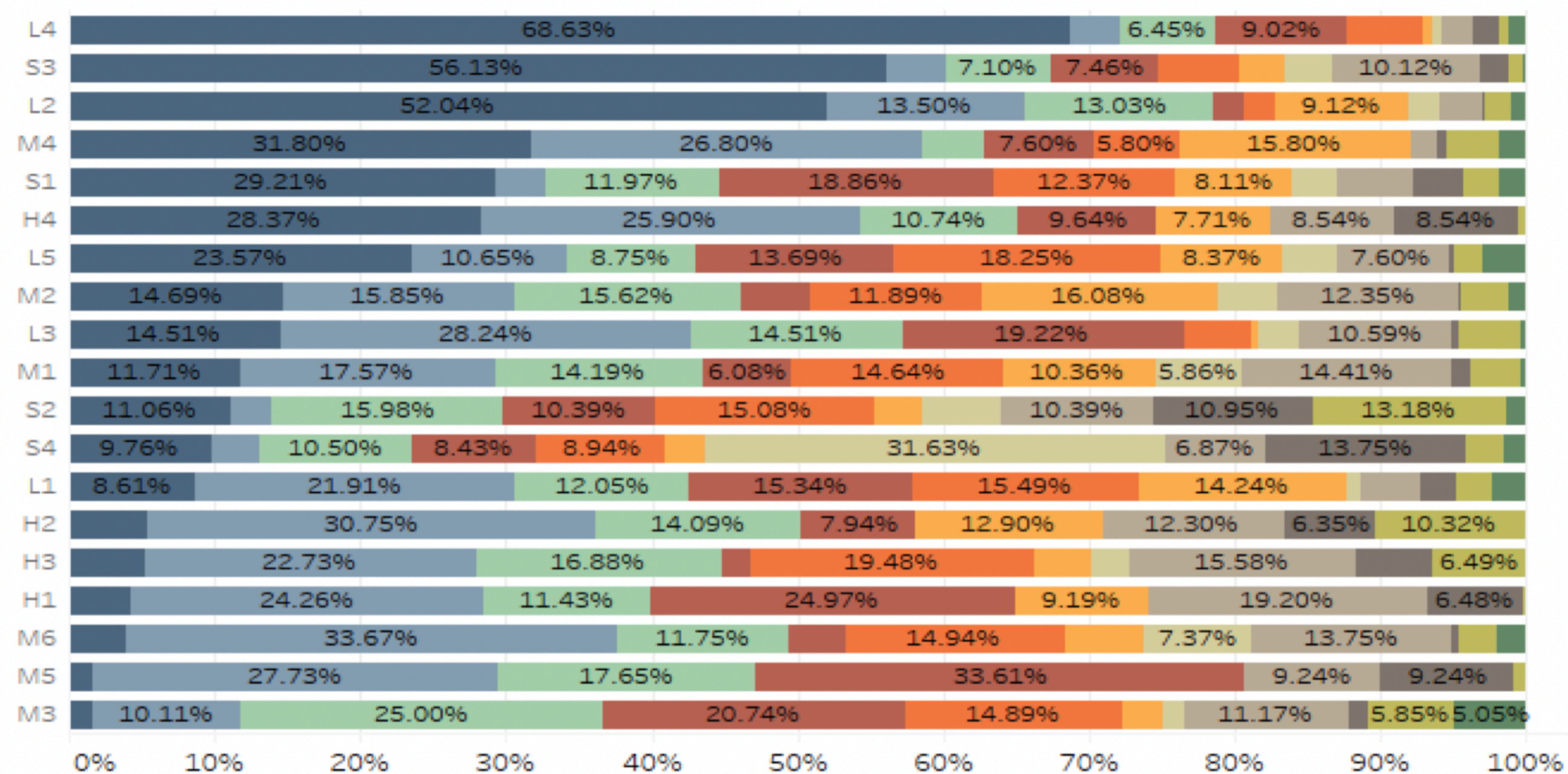


Suitability of
Time & Discussion

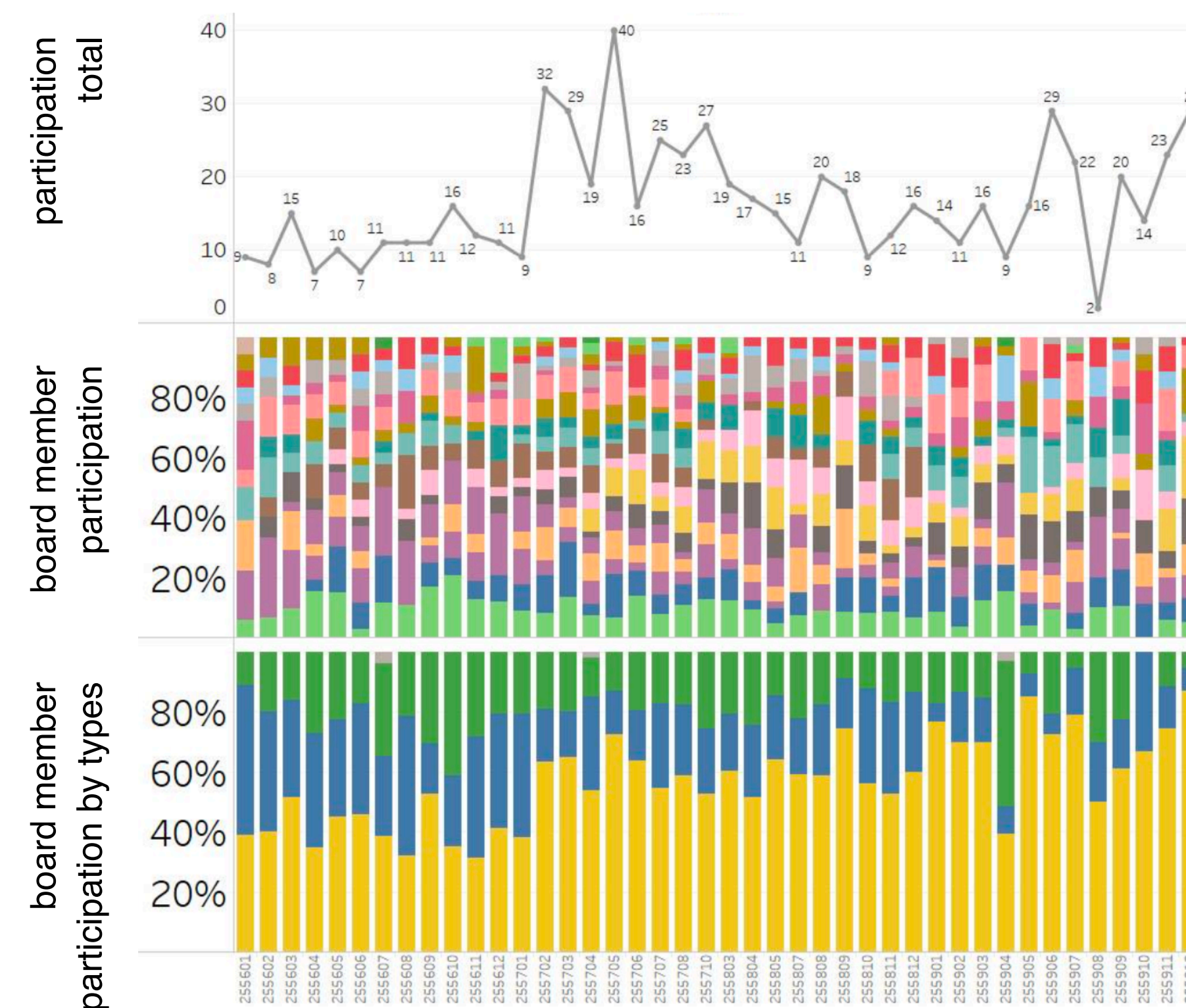


1. Behavioral & Culture Supervision

% of topics discussed by different FIs



board member participation according to meeting minutes

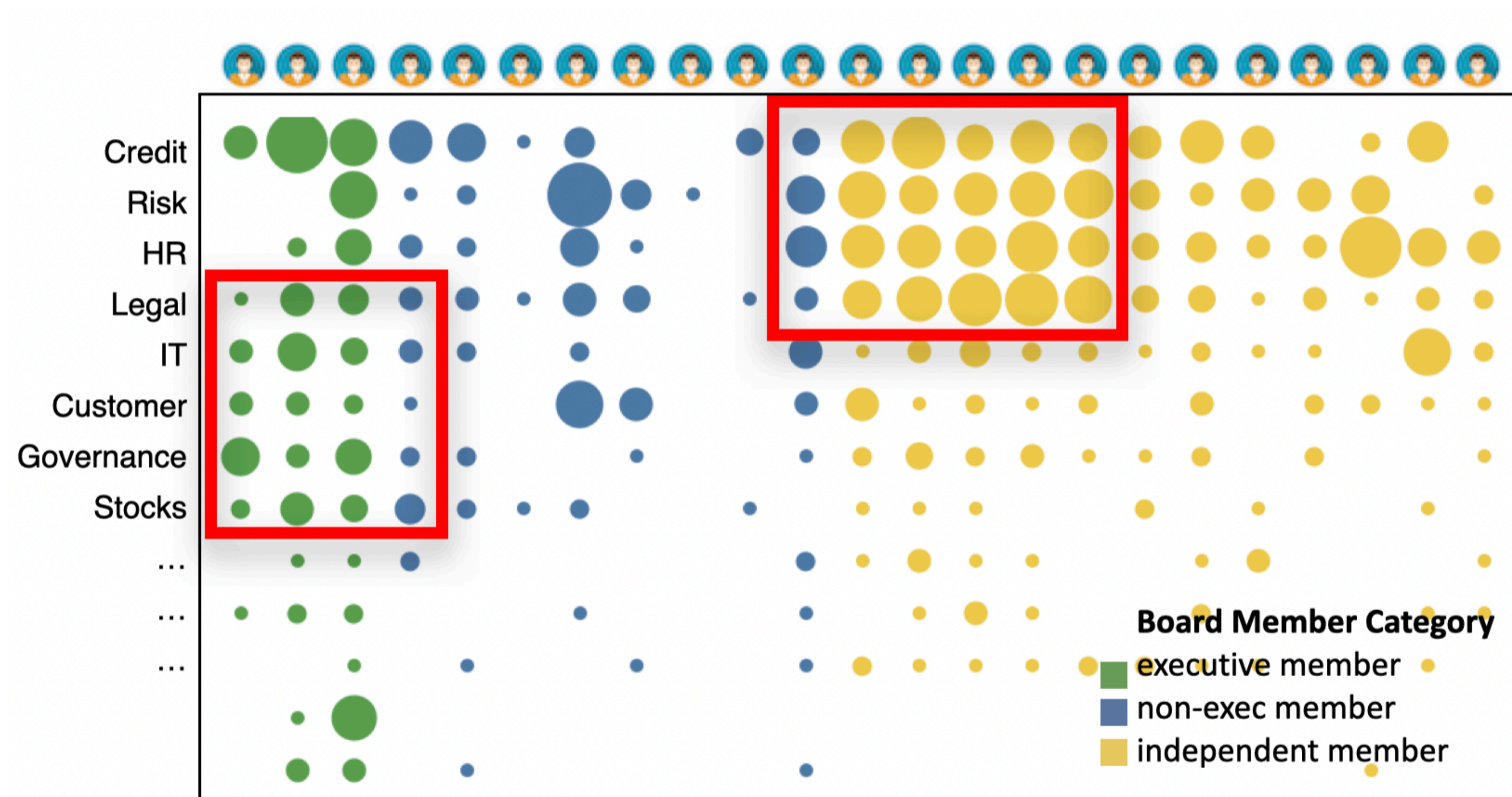


We track the percentage of topics discussed by different FIs in their board of director meeting minutes, together with the participation rate of each board members over time.

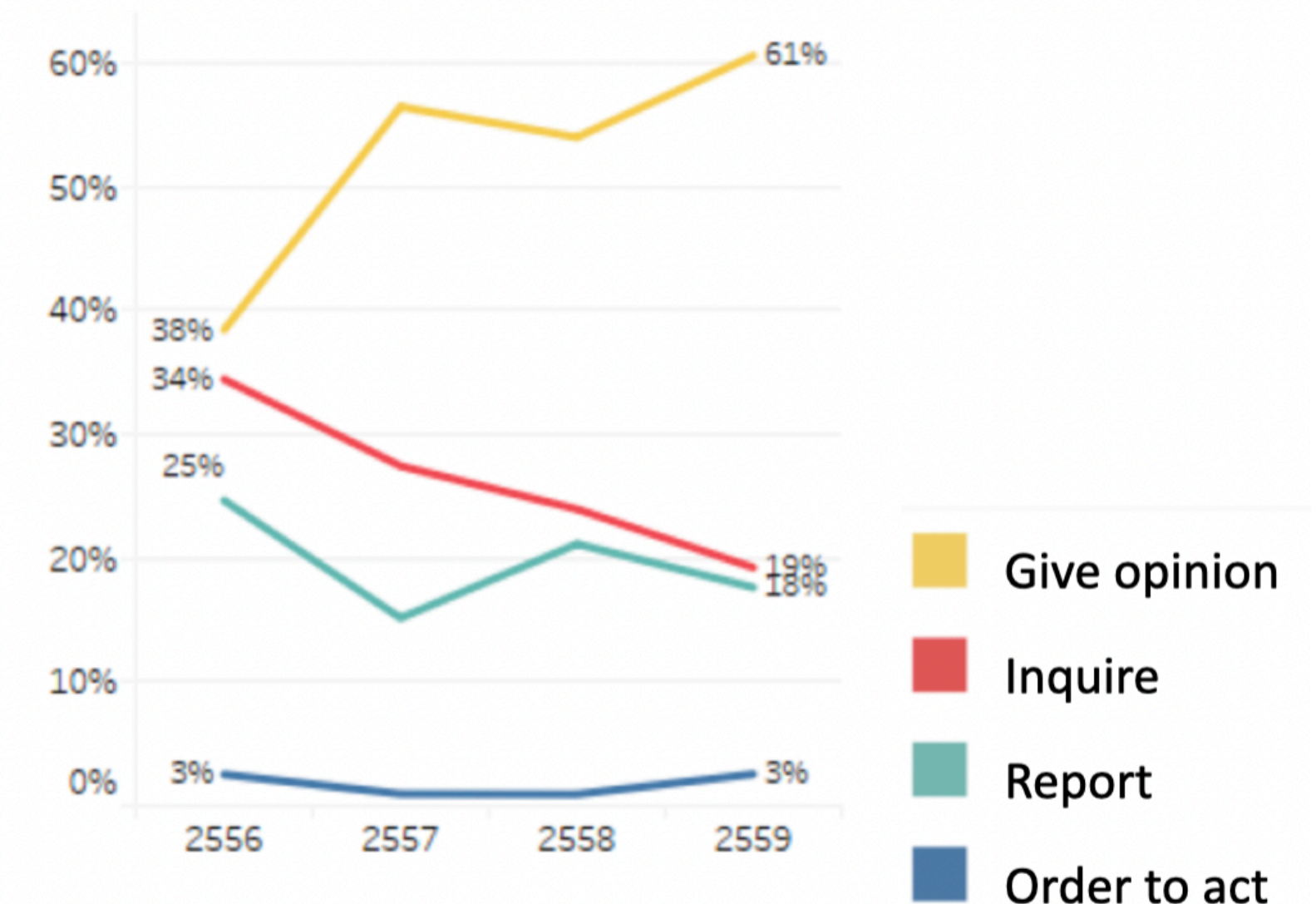


ธนาคารแห่งประเทศไทย
BANK OF THAILAND

1. Behavioral & Culture Supervision



Additionally, We track the participation of each board member of each FI, in different topic of discussion. This enables us to study and to monitor different culture of each FI board. For instance, in one FI, we see that for certain specialize topics such as IT and legal issues, only some specialize board members participate in the discussion (sparse circles). While in some topics such as credit and risk, there are more open discussion (dense circles in the box).

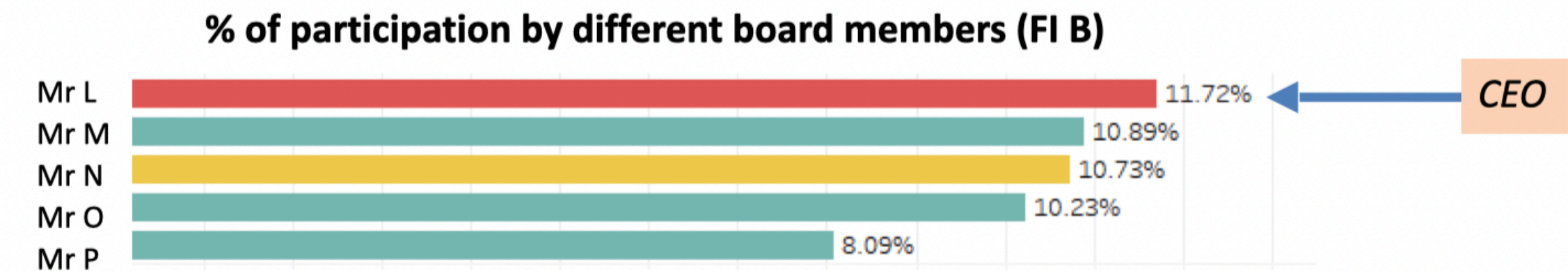
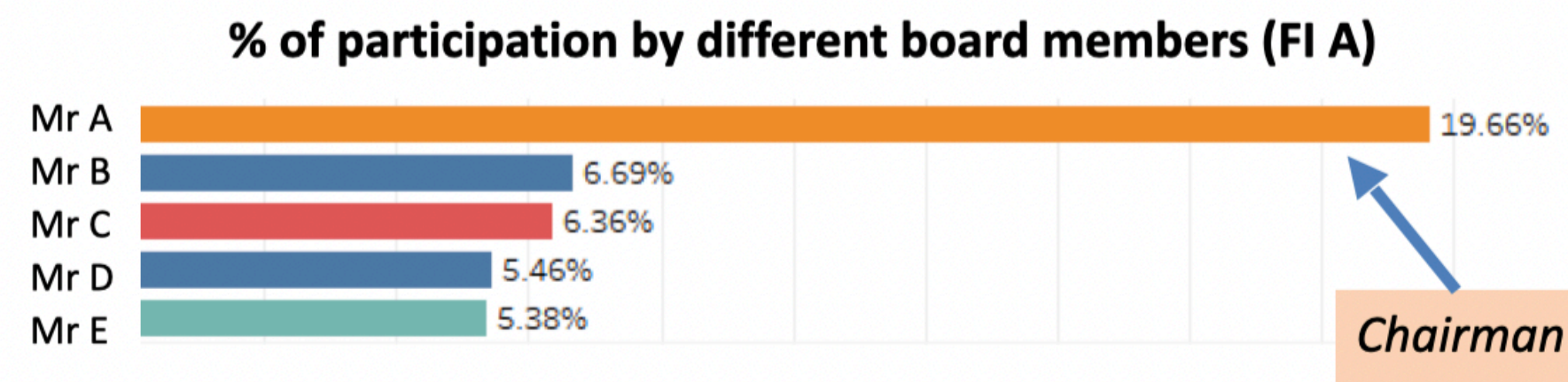


We also track “type” of participation by each board member whether they simply inquire, report, give opinion or giving order. This allow us to study the change in behavior of each board members. For instance, some board members become more opinionated the longer they stay on the board.



1. Behavioral & Culture Supervision

Examples of different pattern of board member participation between different FIs

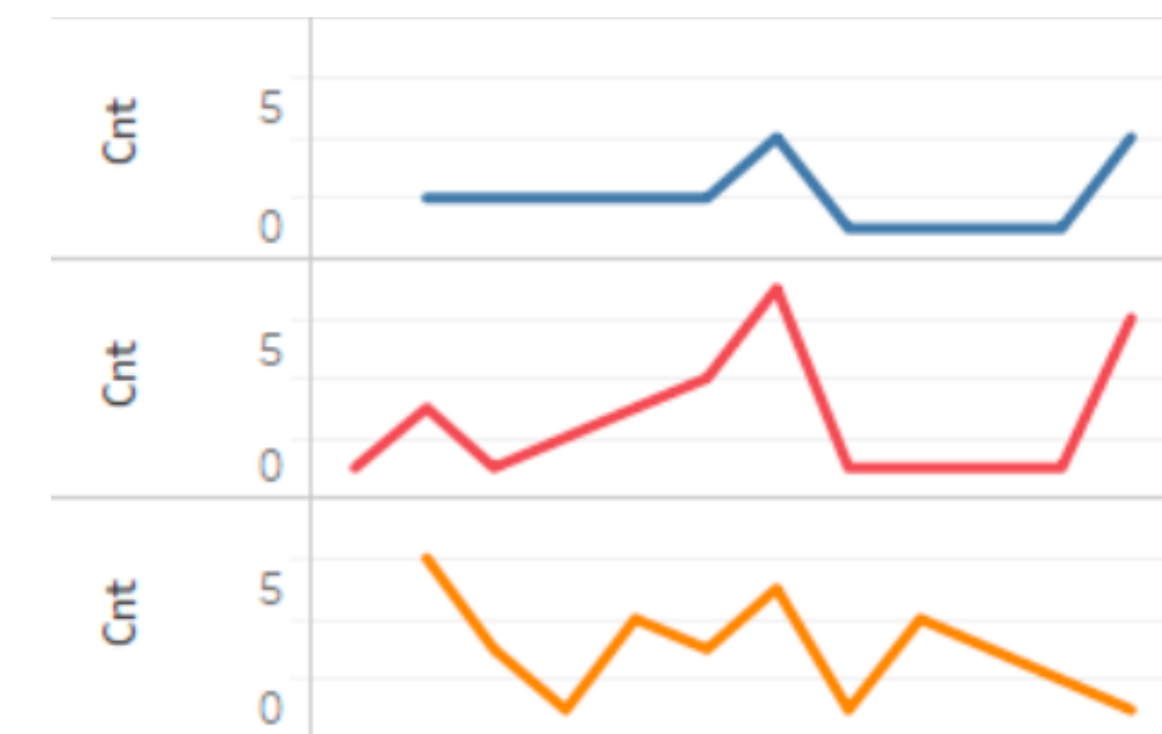


We found that each FI displays different behaviors of board member participations. Some FIs' board discussions are highly dominated by either the chairman or the CEO, while some FIs has high level of participation by all board members.

Results from the analysis of board meeting minutes were incorporated into B&C final report, which were sent to FIs.

Various components, eg. OCR, hmm document parser, word segmentation, were reused in other pilot efforts such as ESG dashboard, textual analysis on historical archives.

ESG dashboard (pilot)





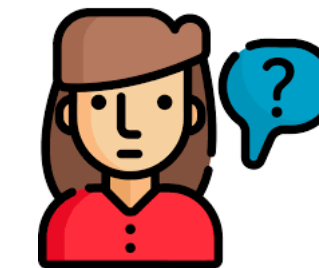
2. Government Expenditure Statistics Compilation

BOT routinely compile various statistics of data from other public or private agencies and share with other government and international agencies



Data from the Government Fiscal Management Information System (GFMIS), maintained by the Comptroller General's Department, Ministry of Finance, are among the most challenging ones.

| | Govt. Expenditure Item | Agency | Amount (Baht) |
|---|--|--|---------------|
| 1 | ค่าตอบแทนสมาชิกกองอาสารักษาดินแดน | กรมการปกครอง | 25,000,000 |
| 2 | ค่าใช้จ่ายการพัฒนาขีดความสามารถด้านข่าวกรองทางทะเล | กองทัพเรือ | 215,000,000 |
| 3 | ค่าปรับปรุงอาคารสำนักงาน กศน. ตำบลในเมือง อำเภอเมือง เพชรบูรณ์ จ. เพชรบูรณ์ | สำนักงานปลัดกระทรวงศึกษาธิการ | 42,000,000 |
| 4 | เตียงนอนพร้อมที่นอนผู้รับบริการขนาด 3.5 ฟุต บ้านพักเด็กและครอบครัวจังหวัดอำนาจเจริญ ตำบลโนนหนามแท่ง อำเภอเมืองอำนาจเจริญ จังหวัดอำนาจเจริญ | สำนักงานส่งเสริมสวัสดิภาพ เด็ก ผู้ด้อยโอกาสฯ | 500,000 |

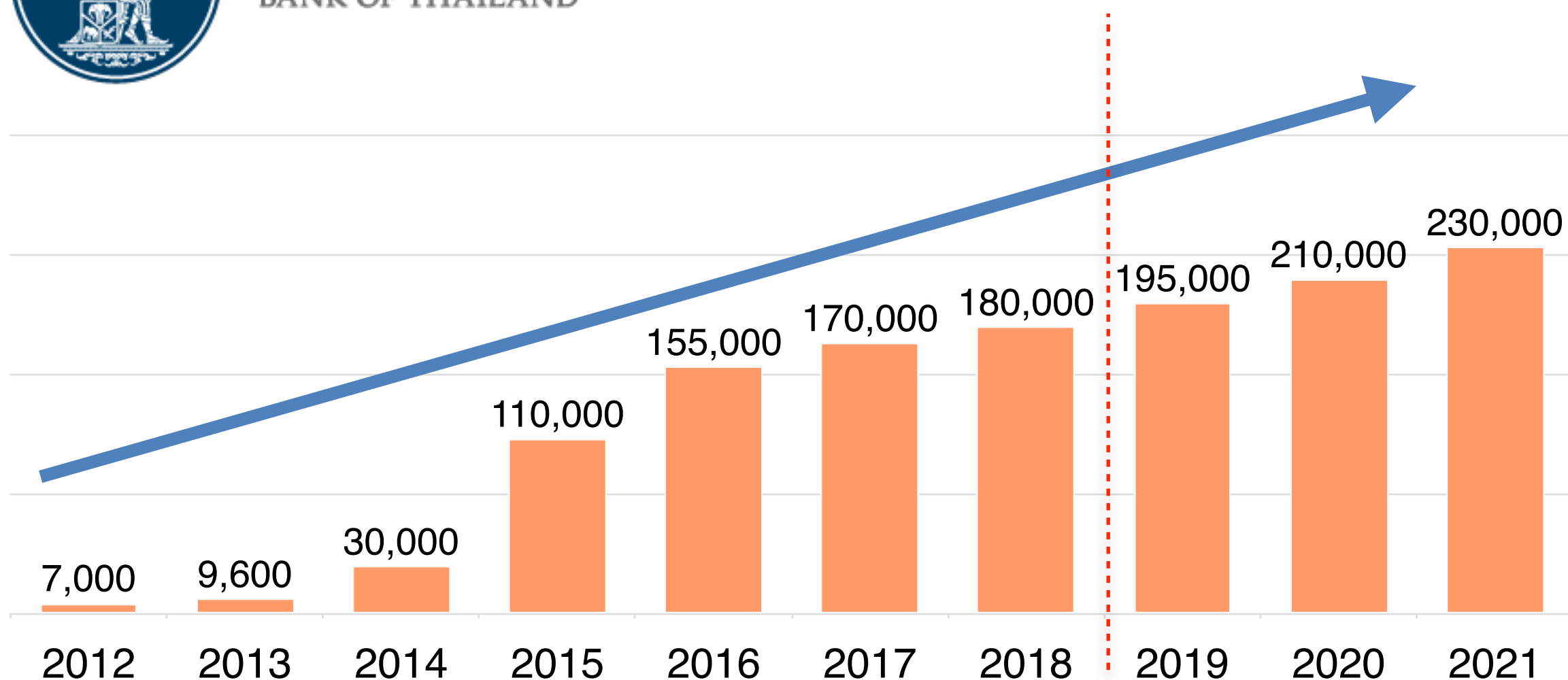


| Code GFS | Functional Code |
|----------|-------------------------|
| 01 | General public services |
| 02 | Defense |
| ≡ | |
| 09 | Education |
| 10 | Social Protection |

| Code GFS | Economic Code |
|----------|---------------------------------|
| 21 | Compensation of employees |
| 2111 | Wages and salaries in cash |
| 22 | Used of goods and services |
| ≡ | |
| 32 | Acquisition of financial assets |
| 33 | Principal Repayments |

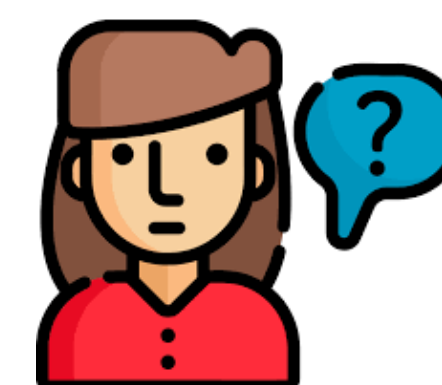


ธนาคารแห่งประเทศไทย
BANK OF THAILAND



2. Government Expenditure Statistics Compilation

deployed since 2019



Manual
Compilation

3 man-days

Automate
Compilation (ML)

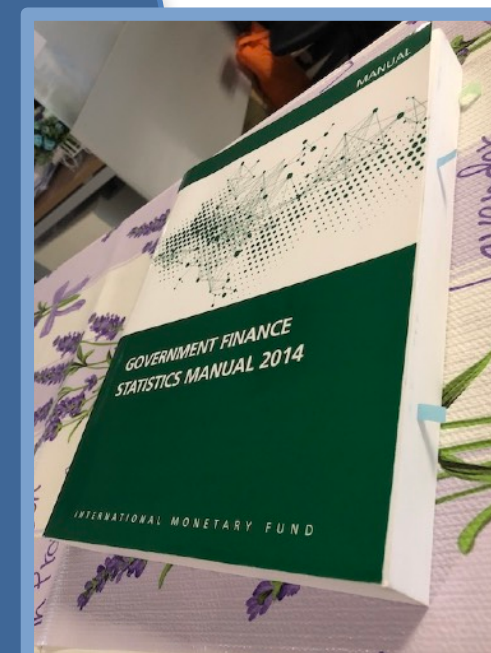
30 mins

Accuracy

Economic code 99.5%

Functional code 96.5%

- The number of government expenditure transactions in the system increase significantly over the last 7 years to ~ 230,000 items in the 2021 budgeting year.
- The assignment of Economic Code & Functional Code require domain experts to manually classify each items (also time-consuming).
- The data are not well formatted and do not have proper categorical information, and are in Thai.
- The Government expenditure statistics is also needed to be compiled monthly in the timely-manner to be shared with other govt. agencies.



Automation enable more time for further analysis of the government expenditure data (which was not possible before due to the time-constraint) such as analysis of government funding allocation in the local government level.

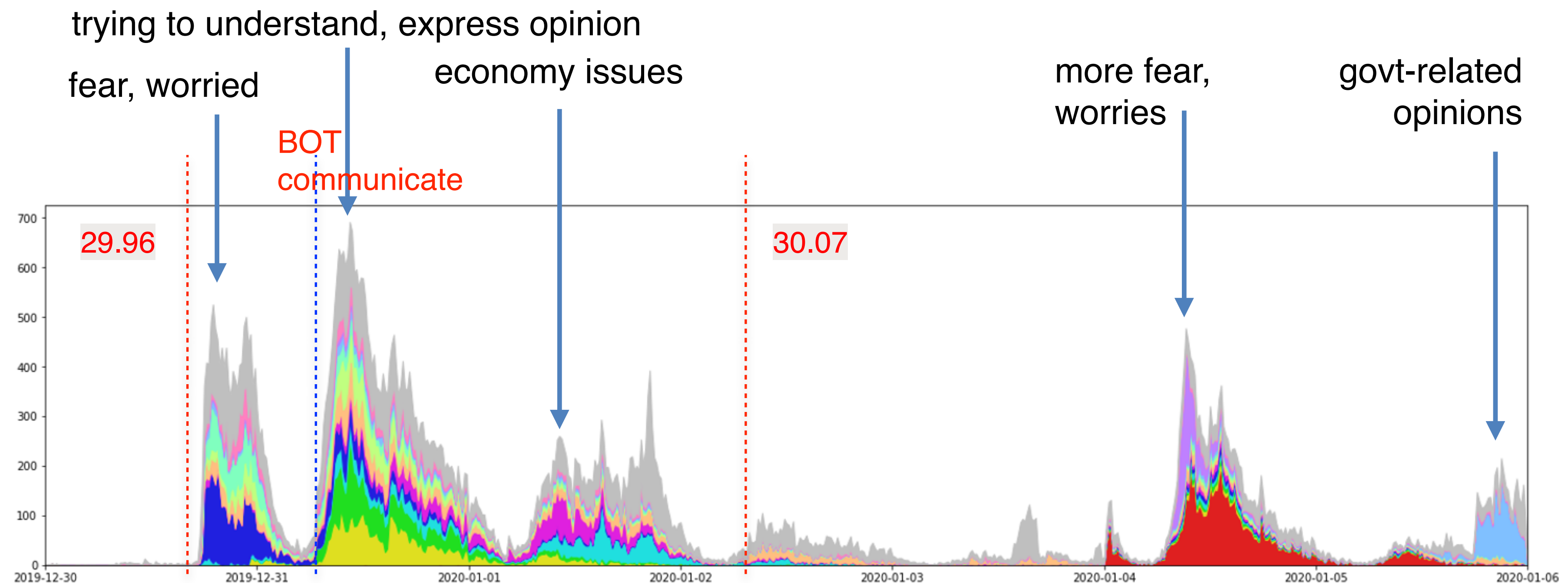
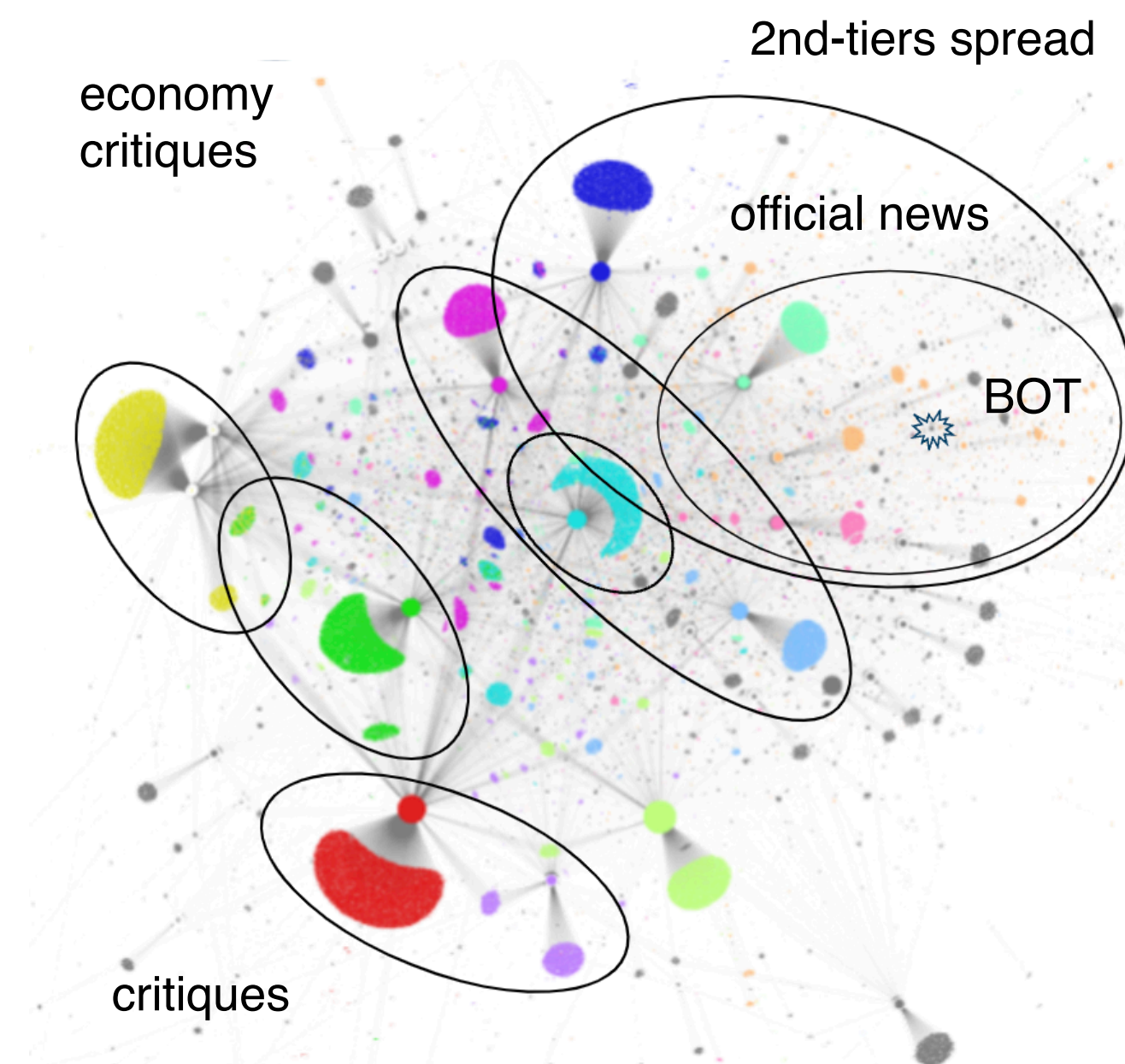


ธนาคารแห่งประเทศไทย
BANK OF THAILAND

3. Social Media Data for Reputational Risk

Monitoring engagement and public opinions on related issues and central bank policy in social media (FX, soft loan policy, counterfeit banknote, etc).

- clustering topics for public opinions monitoring
- monitoring virality of issues
- network analysis for community discovery (interested in neural voices and channel reach)





ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Thank You for you attention



Take Away

- Lots of Potential Use Cases for NLP in Central Bank Operations.
- Use Cases need to be explored collaboratively between Data Analytics Team and Operational Team.
- Long-Term Adoption can be a challenge.
- However, tools developed can still be re-repurposed as building-block for future projects

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Machine learning for measuring central bank policy credibility and communication from news¹

Muhammad Abdul Jabbar, Okiriza Wibisono,
Anggraini Widjanarti and Alvin Andhika Zulen,
Bank Indonesia

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Machine Learning for Measuring Central Bank Policy Credibility and Communication from News

Okiriza Wibisono¹, Muhammad Abdul Jabbar², Alvin Andhika Zulen³, Anggraini Widjanarti⁴

Abstract

Central bank's monetary policy will be effective only if the central bank is perceived to be credible. The relationship of monetary policy credibility and effectiveness has long been established by economists and central bankers, including credibility's significance in determining central bank's ability to manage the expectation of economic actors. Empirically, credibility is a qualitative concept, and thus is not straightforward to measure. In our previous research, we have constructed a machine learning-based index of central bank credibility from news data. The index is based on 4 component indexes: credibility of policy formulation, policy effectiveness, coordination with the government, and policy communication. The communication index measures how well public's expectation is aligned with the central bank's forward guidance, which in our case is traditionally related mainly to the policy rate. With the COVID-19 pandemic, quantitative easing (QE), along with the policy rate, plays a greater role in achieving central bank's monetary policy objectives. In this paper, we describe our machine learning-based credibility measure and propose an improvement to the communication index to accommodate the higher significance of QE. We test the indexes' effect on economists' expectation of short-term inflation.

Keywords: central bank credibility, quantitative easing, central bank communication, inflation expectation

JEL classification: E52, E31, D84

¹ Department of Statistics, Bank Indonesia. email: okiriza_w@bi.go.id

² Department of Statistics, Bank Indonesia, email: muhammad_abdul@bi.go.id

³ Department of Statistics, Bank Indonesia, email: alvin_az@bi.go.id

⁴ Department of Statistics, Bank Indonesia, email: anggraini_widjanarti@bi.go.id

The author would like to thank a colleague for insightful discussion on the econometrics methodology.

Contents

| | |
|--|----|
| 1. Background..... | 3 |
| 2. Literature Review | 4 |
| 2.1 Bank Indonesia’s Policy Credibility Survey | 4 |
| 2.2 Text mining of economic news..... | 5 |
| 3. Methodology | 6 |
| 3.1. Data | 6 |
| 3.1.1. News articles..... | 6 |
| 3.1.2. Inflation estimates..... | 6 |
| 3.2. Policy Credibility Index | 8 |
| 3.2.1. Annotation | 8 |
| 3.2.2. Data preprocessing..... | 9 |
| 3.2.3. Model training | 9 |
| 3.2.4. Text classification and index calculation | 10 |
| 3.3. Regression Setting..... | 11 |
| 3.3.1. Model and estimation | 11 |
| 3.3.2. Control variables..... | 12 |
| 4. Result and Discussion | 13 |
| 4.1 Index Results..... | 13 |
| 4.1.1 Policy credibility index and the 4 component indexes | 13 |
| 4.1.2 Comparison of communication indexes..... | 15 |
| 4.2 Regression Results..... | 16 |
| 5. Conclusion & Future Works | 18 |
| 5.1. Conclusion | 18 |
| 5.2. Future Works | 18 |
| References..... | 19 |
| Appendix A: Example annotated sentences..... | 20 |
| Appendix B: Regression results for heterogeneity of estimates..... | 21 |

1. Background

Central bank's monetary policy will be effective only if the central bank is perceived to be credible. The role of credibility in supporting policy effectiveness has long been established by economists and central bank practitioners. This is in-line with the rational expectation hypothesis which is believed to guide the behavior of economic agents. The credibility of monetary policy will determine the ability of central bank to manage the expectation of economic actors (Łyziak & Paloviita, 2016) and will be reflected in the expectation of long-term interest rates and other asset prices (Blinder, 2000).

Based on the theory, credibility is about consistency in providing accurate and valuable information or always being responsible (Sobel, 1985). In the terminology of policy makers, credibility is the expectation that policies that are announced will be implemented, not only reflecting the policy makers' wishes, but also capturing the situation and conditions behind the policy so that if there is a shock beyond expectation, it will affect the credibility (Drazen & Masson, 1994). Central bank credibility is formed by 7 (seven) important factors i.e. honesty, independence, consistency (history of fighting inflation), transparency, maintaining fiscal deficit, the use of policy rules, and incentives (personal loss) (Blinder, 2000).

Central bank credibility also plays a pivotal role in building public's trust on the institution. Trust is important for reasons of political accountability, ensuring operationally independent central banks are meeting the terms of their social contract with wider society. Another reason to try to build trust is that trust helps manage expectations. But, evidence suggests that public may never engage with central bank communication because it is written in a way that they cannot understand, which contributes to a lack of trust in the central bank as an independent institution (Haldane et al., 2020). As argued in Haldane and McMahon (2018), one of the reasons that a central bank may want to communicate more directly with the general public is to try to build public understanding as a means of establishing trust and credibility about central banks and their policies.

Empirically, credibility is a qualitative concept, which is not easy to measure, and there are several approaches have been used in measuring credibility. In Cukierman & Meltzer (1986) and Faust & Svensson (2001), central bank independence is used as a proxy for credibility. In Blinder (2000), the credibility of a central bank is measured using survey where economists choose the main characteristics that contribute to the central bank credibility, which is then quantified to produce an index. While in Łyziak (2016) measurement is made using an index, which is built from several indicators including the achievement of the central bank's target, the deviation of inflation expectation, transparent targets and indicators, independence, and accountability. The credibility measure based on deviation of inflation expectation strikes close to the objective of inflation targeting framework. However, it can be difficult to gauge especially as concerns households, who are often not even informed of the central bank's inflation target itself.

In our case, Bank Indonesia used to regularly conduct survey to external stakeholders to measure policy credibility, i.e. Bank Indonesia's Policy Credibility Survey. The Policy Credibility Survey is based on 6 aspects of credibility: formulation, independence, communication, accountability, coordination, and effectiveness.

In practice, the survey method (in general) has several weaknesses:

1. Survey fatigue: respondents experiencing burnout if surveyed repeatedly, so that surveys cannot be carried out too often (especially since the pool of economists or other stakeholders as respondents can be quite limited);
2. Desirability bias: respondents giving a response that is favorable for the surveyor, which is the central bank, hence the results are less objective;
3. Recency bias: respondents generally providing responses based on recent policies and/or events, hence the survey results are very dependent on the execution time; and
4. Survey cost & time.

Based on these considerations, we develop an alternative policy credibility measurement (indexes) by utilizing Big Data Analytics. The indexes are constructed from text mining of public perceptions toward central bank policy credibility that are reported in news media. In this paper, we explain the methodology, some of which have been covered in (Zulen, 2020), as well as describe an improvement to one of the policy credibility indexes and evaluate the indexes in econometric models.

The paper is organized as follows. In section 2, we provide literature reviews on Bank Indonesia's Policy Credibility Survey, text mining of economic and financial news, and inflation estimates, which we use for evaluating the indexes. In section 3, we discuss the data and methodology. In section 4, we provide a summary of the results and evaluation of the model. In section 5, we conclude the paper and offer some thoughts for future works.

2. Literature Review

2.1 Bank Indonesia's Policy Credibility Survey

From 2013 to 2018, Bank Indonesia conducted Bank Indonesia's Policy Credibility Survey, a semi-annual survey to measure policy credibility for all 3 sectors of policy: monetary, macroprudential, and payment system. The survey was aimed to provide a measure for policy credibility that is objective, accurate, reflecting broad view of stakeholders (including general public), and available timely. The survey was used to determine the effectiveness of policy communication as well as feedback for formulating future policy communication strategies.

The target respondent of the survey was approximately 1,000 respondents in 20 major cities in Indonesia, consisting of government personnel, bankers, industry players, academics, and general public. The survey measured 6 aspects of policy credibility, from which our indexes are derived:

1. Formulation: whether our policies are formulated carefully according to their objectives
2. Independence: whether we formulate our policies independently, without intervention from any party
3. Communication: whether our policies are well-communicated to the public
4. Accountability: whether our policies are well accounted for
5. Coordination: whether we always coordinate well with the government
6. Effectiveness: whether our policies are effective in achieving their objectives

2.2 Text mining of economic news

Text data have been widely used for research in economics and finance. Nowadays, text mining algorithms are growing rapidly along with the adoption of big data and machine learning. These algorithms can automatically "read" and "extract" relevant information from texts, such as person's name, topics, and sentiment. Compared to manual approach, text mining allows us to make use of much larger text data faster, including news, social media, and press releases.

As an example related to central banks, Sahminan (2008) identified keywords that reflect a tight, neutral, or loose monetary policy inclination in the press release statement of Bank Indonesia over the period from January 2004 to December 2007. Econometric analysis shows that monetary policy statements that contain loose or neutral policy inclination tend to lower interbank interest rates, while monetary policy statements with tight policy inclination tend to have no impact on interbank interest rates (asymmetric effect).

A closer research to ours is Tobback et al. (2017), who developed the Hawkish-Dovish (HD) index that measures media's perception of ECB communications. The HD index is computed by using two methods: semantic orientation (SO) and support vector machine (SVM). The HD index based on SO method is computed by counting the co-occurrences of strings with a pre-determined words/expressions that are normally associated with "hawkish" and "dovish" concepts to determine the tone of the document. For the SVM method, instead of using predefined set of keywords, the algorithm automatically looks for patterns in text documents to select the words with the highest discriminative power and determines the tone of a document based on them. Similar Hawkish-Dovish research has also been done earlier by Lucca & Trebbi (2009) for the FOMC statements, although it uses number of search hits, not directly mining (news) text data.

With regard to how we evaluate the policy credibility indexes, there are several research that test the impact of economic and central bank-related news on economic indicators. This is in line with the notion that sentiment affects economic conditions (self-fulfilling), in addition to the other way around that economic conditions affect sentiment (Algaba et al., 2020). For example, research by ter Ellen (2019) analyzes, among others, the impact of narrative monetary policy surprise in news media on macroeconomic indicators, including interest rate, CPI, and consumer confidence. The surprise is defined as the change in news topical coverage around the day of central bank policy announcements. A survey study by Hayo & Neuenkirch (2015) also shows that most financial market participants routinely monitor the media for central bank news, and thus the media may affect their decision-making.

Our evaluation methodology mostly refers the research by Lena (2011) which investigates the role of media on households' inflation perceptions and expectations, and research by Lamla & Maag (2012) which investigates the role of media on disagreement of inflation expectations across households and professional forecasters. Some of the main differences with our research are:

- We are interested in the effect of news on monetary policy credibility (more detail in the next section), while the previous research focus on news on inflation.
- We add two other explained variables: the accuracy of inflation expectations, and the anchoring of inflation expectations to inflation target of the government and Bank Indonesia. (We do not attempt to predict the inflation expectation itself).

- The coding of the whole set of news texts into credibility sentiment is done by machine learning models, after manually annotating several thousand example news sentences. In previous research, from what we understand the coding (of inflation news) was done manually or by a rules-based approach.

3. Methodology

3.1. Data

3.1.1. News articles

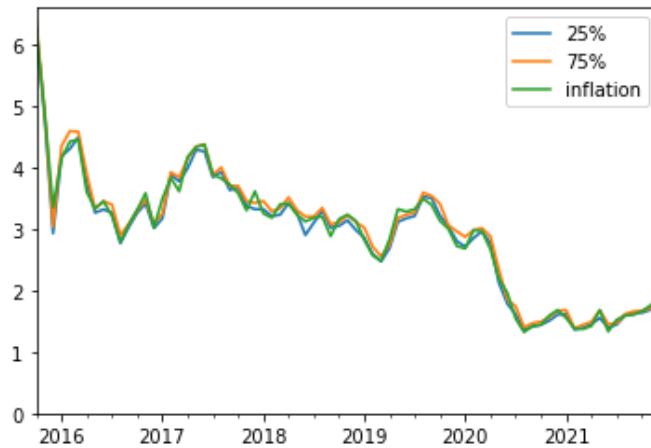
Source: News data serves as the main input for constructing the policy credibility indexes. We use news data from Bank Indonesia's Cyber Library, which is a curated internal repository of news articles related to economic and financial topics. There are more than 30 domestic news media, with an average of about 850 articles daily, although the number of news media and news articles can vary from month-to-month. The news data are available on a daily basis since 1999, but the news data that we use in this paper span from October 2015 to December 2021 (we will explain later about our choice of starting point). The news are in Bahasa Indonesia (Indonesian language).

Filtering: We filter out news that are not relevant for constructing the index. Specifically, we only keep news *sentences* that contain any of the keywords related to monetary policy: "inflation", "monetary", "exchange rate", "current account", "policy rate", "BI Rate", "BI 7-Day Reverse Repo Rate", and their variations in writing. Furthermore, the sentence or its previous/next sentence must mention "BI" or "Bank Indonesia".

Additional keywords for communication index: Starting from January 2020, we add more keywords in order to accommodate non-interest rate-related monetary policy, such as QE, which becomes more relevant since the onset of COVID-19 pandemic. Example keywords are "reserve requirement", "quantitative easing/QE", "accommodative", and "liquidity". We also include keywords related to Bank Indonesia's Joint Decrees with the Ministry of Finance on COVID-19 burden sharing mechanism: "burden sharing", "purchase of SBN (government securities)". These additional keywords are applied only to the communication index, since this index is the one most dependent on the specific monetary policy instruments being taken at the time.

3.1.2. Inflation estimates

As one means to evaluate the policy credibility indexes, we test the indexes in several econometric models of inflation estimates. The estimates are obtained from Bloomberg Economist Estimates, which, every month, asks a number of domestic and foreign economists about their prediction of Indonesia's CPI inflation for the current month. The CPI data itself is usually released by Indonesia National Statistics Office (BPS) on the first week of the next month.



As can be observed from Figure 1, in terms of level, the estimates track actual monthly inflation pretty well, with the monthly mean estimates' having mean absolute deviation of only 9 basis points. ($|\text{mean}(\text{estimate of inflation}_t) - \text{inflation}_t|$, averaged across all months in the sample). However, out of the 75 months sample period, there are 56 times or three-fourths where actual inflation is out of the 25-75th percentile of the estimates.

Since we are interested in whether news articles that make up our policy credibility indexes affect inflation estimates and the news are in Indonesian, we limit the sample to only include domestic economists.

Distribution of inflation estimates

Table 2

| | Sample statistic |
|--|------------------|
| Mean number of economists | 20.6 |
| Mean number of economists, domestic sample | 10.2 |
| Mean of monthly mean¹ | 2.951% |
| Mean of monthly standard deviation¹ | 0.097% |
| Mean of monthly absolute deviation from actual inflation¹ | 9 bp |
| Number of times actual inflation lower than percentile 25 estimate¹ | 33 (44%) |
| Number of times actual inflation higher than percentile 75 estimate¹ | 23 (30.3%) |
| Number of times mean estimate lower than lower bound of inflation target¹ | 21 (27.6%) |
| Number of times mean estimate higher than upper bound of inflation target¹ | 1 (1.3%) |

¹ All sample statistics are calculated using the estimates of domestic economists, not the whole sample.

Source: Bloomberg, author's calculation

3.2. Policy Credibility Index

3.2.1. Annotation

A random sample of the filtered news sentences are manually annotated to construct training data for “teaching” machine learning classification models. Each sentence is labelled with 4 (four) information representing public’s perception on the credibility aspects. The possible labels are positive, negative, or irrelevant, except for communication index, for which the possible labels are accommodative/dovish, neutral, tight/hawkish, or irrelevant.

There are some adjustments with respect to the credibility aspects in Policy Credibility Survey as explained in section 2.1:

1. Independence aspect is merged into coordination aspect, as they both capture Bank Indonesia’s coordination with the government.
2. Accountability aspect is excluded, as it is rarely discussed in news media.
3. Communication aspect is defined as public’s perception/expectation of monetary policy stance. In the survey, it measured whether our policies had been well-communicated. Arguably this original definition is more difficult to learn from news media.

Annotation is done by the authors and subject matter experts on monetary policy communication and central bank credibility within Bank Indonesia. Prior to annotation, we write out the guidelines on how to annotate the news sentences including specific examples, so that the result is more consistent across annotators. Each sentence is annotated by 2-3 annotators to minimize bias.

A total of 12,560 sentences from January 2010 to August 2019 are annotated (first four rows of Table 1). For the additional keywords for communication index about 7,367 sentences are annotated, for the period January 2020 to May 2021. Example annotated sentences are provided in Appendix A.

Distribution of annotated sentences

Table 1

| | Positive ¹ | Negative ² | Neutral | Irrelevant |
|-----------------------------------|-----------------------|-----------------------|----------------|----------------------------|
| Formulation | 1,595 (81.6%) | 360 (18.4%) | Not Applicable | 10,605 (84.5% of total) |
| Effectiveness | 2,072 (81.1%) | 482 (18.9%) | Not Applicable | 10,006 (79.8% of total) |
| Coordination | 303 (86.8%) | 46 (13.2%) | Not Applicable | 12,211 (97.2% of total) |
| Communication | 493 (27.6%) | 391 (33.0%) | 706 (39.4%) | 10,770 (14.2% of total) |
| Communication – additional | 3 (0.1%) | 3,358 (99.7%) | 6 (0.2%) | 4,000 (54.3% of total) |

¹ Tight/hawkish for communication index. ² Accommodative/dovish for communication index.

Sources: Authors’ calculation, from annotation

3.2.2. Data preprocessing

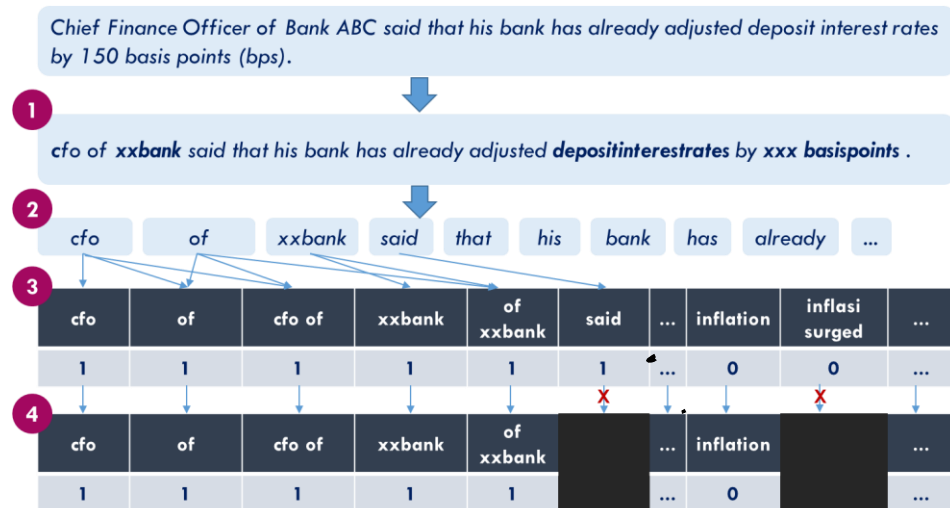
Each filtered sentence (not necessarily annotated) as described in section 3.1.1 is transformed from textual format into tabular-numeric so that it can be processed by machine learning algorithms, following the steps below:

1. Sentence cleansing: lowercasing, replacing synonyms, abbreviations, numbers, and common names in the sentence;
2. Tokenization: splitting sentence into words/tokens;
3. N-gram vectorization: creating n consecutive words (n-gram) as additional features; and
4. Sparse terms removal: removing rarely occurring terms from the feature list.

An example is shown in Figure 2 below.

Example data (sentence) preprocessing steps

Figure 2



3.2.3. Model training

From the preprocessed and annotated sentences, we train machine learning models to classify each sentence into one of possible labels (positive/negative/irrelevant, or accommodative/neutral/tight/irrelevant for communication index). We train one model for each aspect, so in total we have 4 sets of models representing the 4 aspects. We experimented with several learning algorithms: logistic regression, naïve bayes, decision tree, random forest, XGBoost, FastText, and deep learning – LSTM (long-short term memory network).

As can be seen from the annotation results (Table 1), the class distributions are quite imbalanced, with most sentences in the news media being irrelevant (even after filtering with keywords). For relevant sentences, the distributions are also imbalanced towards positive class. Considering these imbalances, we carry out the classification in 2 stages for each credibility aspect:

1. Classifying whether the filtered sentence contains perception about central bank credibility (credibility vs. non-credibility i.e. irrelevant); and

2. For relevant sentences, classifying the credibility sentiment in the sentence (positive vs. negative / hawkish vs. neutral vs. dovish).

More details on model training can be referred to in (Zulen, 2020).

Improvement on model for communication aspect: As an improvement to the previous procedure, we train a separate model for classifying sentences related to the additional keywords for communication index as described in section 3.1.1. Since most perceptions/expectations on monetary policy in news media since January 2020 is accommodative, which may reflect the monetary policy measures during COVID-19 pandemic, we simplify the classification model, i.e. the model only classifies whether each filtered sentence contains (accommodative) perception of monetary policy or whether it is irrelevant.

3.2.4. Text classification and index calculation

Having obtained the classification models for each credibility aspect, we apply the models to the whole (filtered) news data in Cyber Library. For each time period (monthly), we then tabulate the number of sentences classified as positive or negative for each aspect (formulation, effectiveness, and coordination). Besides on historical data, we also calculate the index calculation for ongoing periods, without the need for more manual annotation as the sentence classification models have been trained.

The index for formulation, effectiveness, and coordination aspects is calculated as the net balance of the number of positive and negative sentences in each time period.

$$index_{aspect,t} = \frac{\#positive_{aspect,t} - \#negative_{aspect,t}}{\#positive_{aspect,t} + \#negative_{aspect,t}}$$

For communication aspect, the calculation is different although still similar. First, we compute the stance perception index using net balance of the number of sentences classified into tight, accommodative, and neutral classes as below (note that the number of accommodative sentences include those related to the additional keywords as explained in the previous section). The time period is also different in that for the 3 other indexes, the monthly index are constructed from news at beginning of month (date 1) to end of month. For communication index, the beginning period is 1 day after monthly Board of Governors meeting to 1 day prior to the next monthly Board of Governors meeting.

$$index_{stance\ perception,t} = \frac{\#tight_t - \#accommodative_t}{\#tight_t + \#neutral_t + \#accommodative_t}$$

This index is then compared with the direction of forward guidance contained in Bank Indonesia's press release after each monthly Board of Governor meeting. In the formula below, forward guidance is 1 if the press release contains tight/hawkish stance, 0 if neutral, or -1 if accommodative/dovish. The codification of forward guidance is done manually.

$$index_{communication,t} = 1 - |forward\ guidance_t - index_{stance\ perception,t}|$$

The four indexes are averaged to obtain the (aggregate) policy credibility index, for each month. We refer to (Zulen, 2020) for more explanation on the indexes' characteristics.

3.3. Regression Setting

3.3.1. Model and estimation

We test the indexes in a regression setting. The hypothesis is that perceptions towards central bank credibility, as captured in news media, are read by economists and may affect their internal models or beliefs about the economy. We are interested in economists' estimates of inflation, operationalized specifically as 3 variables below:

1. Inaccuracy of estimate: how much the mean estimate deviate from actual inflation, in absolute terms.
2. Unanchoring of estimate: how much the mean estimate deviate from the midpoint of Bank Indonesia and the government's yearly inflation target range, in absolute terms.
3. Heterogeneity of estimate: how much the estimates vary between economists, as coefficient of variation (CV). We use CV as normalization since the level of variation itself may depend on the prevailing level of inflation.

The regression model is as below.

$$y_{k,t} = \beta_0 + \sum_i \beta_i \text{indexes}_{i,t} + \sum_j \beta_j \text{controls}_{j,t} + u_t$$

where $y_{k,t}$ are calculated as below:

$$y_{\text{inaccuracy},t} = |\hat{\pi}_{t,t-1} - \pi_t|$$

$$y_{\text{unanchoring},t} = |\hat{\pi}_{t,t-1} - \pi_t^*|$$

$$y_{\text{heterogeneity},t} = \frac{sd(\hat{\pi}_{t,t-1})}{|\hat{\pi}_{t,t-1}|}$$

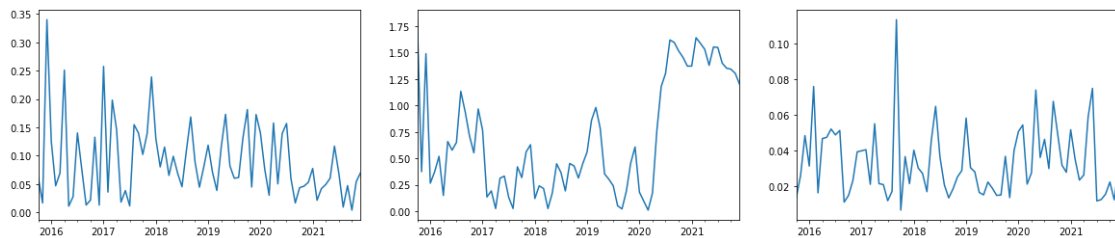
$\hat{\pi}_{t,t-1}$ = an economist's estimate of inflation for month t made at month t-1, π_t^* is the midpoint of inflation target range, and $sd(\dots)$ is the sample standard deviation function. We are interested in the significance of the β_i 's. Graphs of the explained variables are in Figure 3.

Dependent variables series

left: inaccuracy, middle: unanchoring, right: heterogeneity

Figure 3

The structural break in the middle graph coincides with one of the control variables (new COVID cases).



We try several variations of the policy credibility indexes as explanatory variables:

- Indexes = 4 component indexes
- Indexes = communication index
- Indexes = average policy credibility index

All policy credibility index variables are contemporaneous, except communication index which is lagged by 1 month due to difference in time period as explained in section 3.2.4.

In total, there are 9 regression models that we test. The models are estimated with OLS, with heteroskedasticity and autocorrelation consistent (HAC) standard errors for hypothesis testing⁵.

3.3.2. Control variables

We use several control variables in line with the literature, including lagged dependent variable as in (Lamla & Maag, 2012) and (Lena, 2011). We add USD/IDR (percent change, mtm) as exchange rate is one of the important macro indicators for EM economies. We also add new COVID cases variable, as developments in the number of COVID cases impact mobility and economic activity considerably.

The list of explained and explanatory variables, as well their transformation, lag, and unit of measurement are summarized in Table 3 below. The lag structure assures that the explanatory variables reflect the information set that is available to economists at the time they make their monthly inflation estimates.

Regression variables

Table 3

| Variable | Data Source | Transformations | Lag (months) | Unit | Mean | Std. Deviation |
|--------------------------------------|-------------------------------|---|--------------|---------|--------|----------------|
| y, inaccuracy¹ | Bloomberg Economist Estimates | absolute value of (monthly mean estimate – actual inflation) | - | % | 9 bp | 6.6 bp |
| y, unanchoring¹ | Bloomberg Economist Estimates | absolute value of (monthly mean estimate – midpoint of inflation target) | - | % | 70 bp | 54 bp |
| y, heterogeneity¹ | Bloomberg Economist Estimates | monthly standard deviation of estimates / absolute value of (monthly mean estimate) | - | % | 3.4 bp | 1.9 bp |
| Formulation index² | Cyber Library + our ML model | - | - | decimal | 82.8% | 12.9% |
| Effectiveness index | Cyber Library + our ML model | - | - | decimal | 63.8% | 10.0% |
| Coordination index | Cyber Library + our ML model | - | - | decimal | 93.4% | 11.6% |

⁵ R `lm` function for OLS, `coeftest` function in `lmtest` package for single variable hypothesis testing, `waldtest` function in `lmtest` package for joint hypothesis testing, and `vcovHAC` function in `sandwich` package for HAC standard errors.

| | | | | | | |
|--|----------------------------------|---|----------------|-----------------|-------|-------|
| Communication index | Cyber Library + our ML model | - | 1 | decimal | 75.8% | 14.7% |
| Avg credibility index² | Cyber Library + our ML model | - | - ³ | decimal | 79.0% | 7.8% |
| Lag of y | Bloomberg Economist Estimates | - | 1 | % | - | - |
| BI7DRR (policy rate) | Bank Indonesia | - | - | % | 4.78% | 0.86% |
| Inflation | Bank Indonesia | - | 1 | % | 3.01% | 1.08% |
| GDP | BPS (National Statistics Office) | Growth (yoy) + interpolation to monthly | 4 | decimal | 3.72% | 2.92% |
| USD/IDR | Yahoo! Finance | Growth (mtm), last day close | - | decimal | 0.00% | 2.75% |
| New COVID cases | WHO website | Natural log (set to 0 before COVID-19 pandemic) | - | - (natural log) | 3.30 | 5.22 |

Period: October 2015 to December 2021, monthly (75 observations).

¹ Only including domestic economists.

² Mean and std. deviation of formulation and avg. credibility indexes exclude outlying observation of formulation index in May 2017.

³ Communication index is lagged one month. The other 3 component indexes are not lagged.

Sources: Various, authors' calculation

4. Result and Discussion

4.1 Index Results

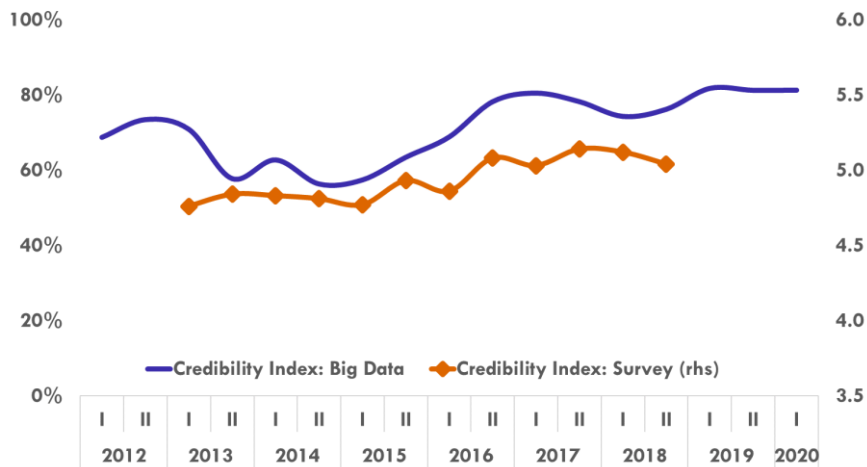
4.1.1 Policy credibility index and the 4 component indexes

For result evaluation, we calculate the correlation between the aggregate policy credibility index generated from news using Big Data Analytics and from our survey (which was decommissioned since 2018). The overall out-of-sample classification accuracy is 63.4% F1 score (averaged across the 4 indexes).

Graph of both indices from is presented in Figure 4. Although both indexes show upward trend, we do not expect them to have such (upward) trend since the indexes should change from time to time based on public's perception towards monetary policy credibility. The indexes have a correlation of 79.7%. The high correlation value indicates that the policy credibility indexes from news have potential to be used as a measure of public's perception on monetary policy credibility.

Aggregate policy credibility index comparison with survey

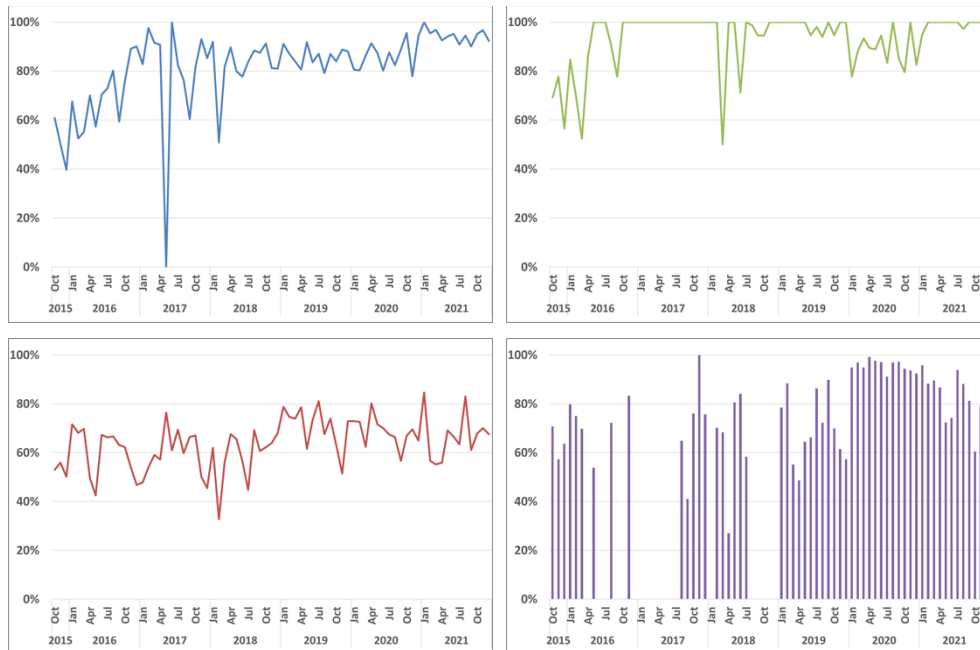
Figure 4



Policy credibility indexes

top left (blue): formulation, top right (green): coordination,
bottom left (red): effectiveness, bottom right (purple): communication

Figure 5



Referring to Figure 5 for the component indexes, we comment on some of the indexes' movements:

- All indexes are always positive. The coordination index is the highest, and it often touches the maximum possible value (100%). The effectiveness index is the lowest although still quite positive, with mean 64%.

- There is a stark outlier on formulation index in May 2017 (value: 0%). We double-checked the resulting sentence classifications and find that the machine learning model do correctly classify all sentences during this period.
- Only the formulation index shows upward trend for the whole study period (disregarding the outlying observation). The other indexes do not have clear upward/downward trend.
- In some periods communication index is blank. This is the case if there is no forward guidance in the press release of the previous period's Board of Governors monthly meeting.

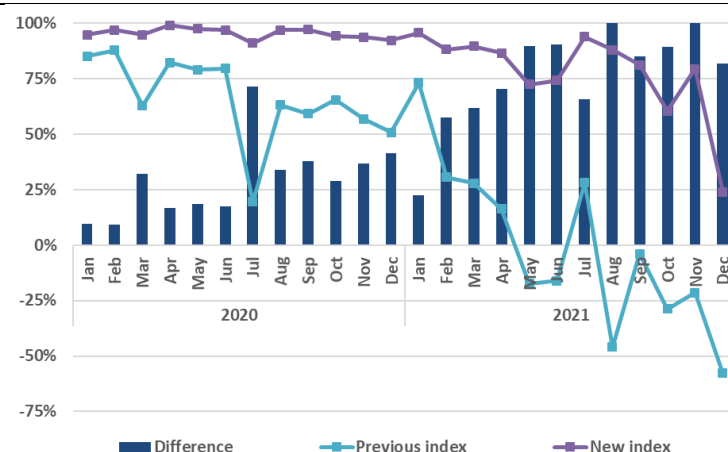
We refer to (Zulen et al., 2020) for more discussion on results of classification accuracy and event analyses of the indexes.

4.1.2 Comparison of communication indexes

In sections 3.1.1 and 3.2.4, we described an improvement to the communication index, namely by adding more keywords related to non-interest rate monetary policy and burden sharing arrangements since January 2020 (onset of COVID-19 pandemic).

Comparison of communication indexes

Figure 5



The new communication index is visibly higher than the previous index. The monthly difference is on average 54.3%, and the difference is as high as 134% (on August 2021). On the surface, this means that the new communication index is more aligned with forward guidance in the press releases of Board of Governors monthly meeting. The difference also has an increasing trend (mean difference for 2020 = 29.5%, for 2021 = 79.1%).

These observations show that the new indexes' potential ability to capture public's perceptions and expectations on the broader set of monetary policy, including not only policy rate but also reserve requirement and QE, as well as other coordinative policies with the government. As a note, Bank Indonesia's policy rate was last changed (decreased) in February 2021 to 3.50%. Public in the news media expect that the policy rate is going to be increased in 2022, thus the previous index has been negative since the latter half of 2021.

We have not analyzed these differences further e.g., in an econometric model.

4.2 Regression Results

Here we report the results of the regressions as described in section 3.3.

| Regression results | | | | Table 4 | | |
|----------------------------------|-------------------|---------------------|---------------------|----------------------|---------------------|---------------------|
| Variable | $y_{inaccuracy}$ | | | $y_{unanchoring}$ | | |
| | 4 indexes | Comm. index | Credib. index | 4 indexes | Comm. index | Credib. index |
| Formulation index (%) | 0.063 (0.053) | - | - | -0.446*** (0.144) | - | - |
| Effectiveness index (%) | -0.087 (0.083) | - | - | -0.864* (0.473) | - | - |
| Coordination index (%) | -0.089 (0.078) | - | - | -0.159 (0.337) | - | - |
| Communication index (% lag 1) | -0.044 (0.077) | -0.026 (0.076) | - | -0.190 (0.290) | -0.120 (0.282) | - |
| Avg credibility index (%) | - | - | -0.068 (0.211) | - | - | -1.457* (0.780) |
| Intercept (%) | 0.239 (0.199) | 0.1455** (0.068) | 0.188 (0.183) | 1.346 (0.888) | 0.220 (0.327) | 1.453 (0.791) |
| Lag of y (%) | -0.113 (0.125) | -0.113 (0.115) | -0.118 (0.114) | 0.445*** (0.133) | 0.482*** (0.128) | 0.458*** (0.121) |
| BI7DRR (policy rate) (%) | -0.001 (0.014) | -0.003 (0.0148) | -0.003 (0.015) | 0.094 (0.063) | 0.056 (0.057) | 0.056 (0.064) |
| Inflation (% lag 1) | 0.001 (0.018) | 0.004 (0.011) | 0.000 (0.017) | -0.111** (0.055) | -0.033 (0.049) | -0.097** (0.048) |
| GDP (yoy, decimal, lag 4) | -0.086 (0.266) | -0.176 (0.243) | -0.129 (0.26405) | -1.000 (0.843) | -1.941* (1.076) | -1.472* (0.807) |
| USD/IDR (mtm, decimal) | -0.127 (0.251) | -0.074 (0.275) | -0.092 (0.255) | -1.335 (1.145) | -0.553 (0.928) | -0.871 (0.968) |
| New COVID cases (ln) | -0.004 (0.004) | -0.004 (0.004) | -0.004 (0.003) | 0.048 (0.021) | 0.039* (0.022) | 0.043** (0.018) |
| R^2 | 0.135 | 0.095 | 0.096 | 0.785 | 0.752 | 0.775 |
| Adjusted R^2 | -0.000 | 0.000 | 0.002 | 0.751 | 0.726 | 0.752 |
| Wald test p-value | 0.088* | - | - | 0.013** | - | - |

Significant at: * 10% level, ** 5% level, *** 1% level

From the table above, several observations can be made:

- For $y = (\text{in})\text{accuracy}$ of estimates:
 - The models and the indexes do not well explain the accuracy of economists' inflation estimates. The highest R^2 is 0.135 when using all 4 indexes as explanatory variables.
 - In terms of significance, the 4 indexes are only jointly significant at 10% significance level.

- None of the control variables are also significant.
- It should be noted that the estimates are summarized as the sample mean across economists. It may be interesting to analyze these effects for individual economists, e.g. in a panel setting, to know whether the indexes affect estimate accuracy for some economists.
- For $y = (\text{un})\text{anchoring of estimates}$:
 - In general this variable is more readily explained by the indexes and the control variables, with R^2 's above 0.75.
 - For the model with 4 indexes, the formulation and effectiveness indexes are statistically significant, with the correct sign (negative). The model with the single aggregate policy credibility index is also significant, albeit only at 10% level. We may view this as indicating that the more positive perceptions on central bank credibility in news media, namely on monetary policy formulation and effectiveness, economists' inflation estimates in the very short run become more anchored to (midpoint of) inflation target.
 - In terms of magnitude, it could be interpreted that an increase of 20 percentage points in formulation index *ceteris paribus* would increase anchoring of inflation estimates by 9 bp on average, and similarly by 17 bp for effectiveness index, or by 29 bp if we take the single policy credibility index model.
 - For effectiveness index, the relevant sentences include news about inflation being within or outside inflation target, in expectation or realization (past inflation). In the case of the sentence discussing past inflation being within or outside inflation target, this information is similar to the lag of the explained variable, which is very statistically significant.
- For $y = \text{heterogeneity of inflation estimates}$, the results are reported in Appendix B. In summary, none of the index and control variables are significant for predicting heterogeneity of estimates, with R^2 as low as 0.061.

It should be noted that the results above are sensitive to the outlying observation of formulation index in May 2017. We also run the regressions with formulation index in May 2017 interpolated from April and June 2017. In this setting, we find that we cannot reject the null hypotheses for any of the 3 explained variables (results not included in this paper).

In light of these results and discussions, we side with the conclusion that the 4 indexes and the aggregate policy credibility index have, at best, weak effect on the accuracy of economists' inflation estimates. For the anchoring of inflation estimates, the formulation and effectiveness indexes are statistically significant. Further analysis is warranted since all the results are sensitive to the one outlying observation of formulation index.

5. Conclusion & Future Works

5.1. Conclusion

We develop a methodology for measuring Bank Indonesia's monetary policy credibility by utilizing news articles data and machine learning-based technique. From the out-of-sample evaluation results, we achieve an average F1-score of 63.4%. The aggregate policy credibility index also moves in-line with the index generated from our survey, with a correlation of 79.7%. The high correlation value indicates that the policy credibility indexes from news have potential to be used as a measure of public's perception on monetary policy credibility.

From the previous research (Zulen, 2020), we improved on the communication index by adding more keywords to capture non-interest rate policy such as QE and reserve requirement, as well as coordinative policies with the government related to COVID-19 pandemic. The new communication index is largely more aligned with forward guidance in the press releases of Board of Governors' monthly meeting.

We test the policy credibility indexes in econometric model of economists' inflation estimates, namely whether the indexes help explain (1) accuracy, (2) anchoring, or (3) heterogeneity of estimates. Initial results show that the indexes are at best weakly significant in explaining accuracy of estimates, but formulation and effectiveness indexes are significant in explaining anchoring of estimates. However, the results are sensitive to the outlying observation of formulation index, and more analyses should be conducted before drawing further conclusions. Heterogeneity of estimates are not explained by the indexes.

5.2. Future Works

Some possible research directions include:

- Analysis on other economic estimates

This paper focuses on estimates of inflation in the very short run (next month) by domestic economists. Other estimates that can be potentially explained by the policy credibility indexes are: estimates of inflation by the broader coverage of economists (including foreign economists), farther estimate horizon (e.g. end of year or 1 year), estimates of inflation by the general public (in line with the work by Munday & Brookes, 2021), or estimates of other macroeconomic indicators (e.g. GDP).

- More elaborate econometric models

We tested the policy credibility indexes using a simple linear regression model of the relationship between estimate variables and the indexes and control variables, with HAC standard errors. It could be interesting to test the indexes with more elaborate models, e.g. including more lags of the policy credibility index variables, panel data across economists.

- Data source/coverage addition

The current indexes cover news published by Indonesian media. A possible improvement is to add English news for constructing the indexes, which could then be tested against the estimates from foreign economists. In addition, currently the additional keywords are only applied to communication index. The formulation, effectiveness, and coordination indexes are also likely to change if we apply the additional keywords.

References

- Algaba, A., et al. (2020). Econometrics Meets Sentiment: An Overview of Methodology and Applications. *Journal of Economic Surveys*, Vol. 34, Issue 3, pp. 512-547.
- Blinder, A. S. (2000). Central bank credibility: Why do we care? How do we build it? *The American Economic Review*, 90(5), 1421-1431.
- Cukierman, A., & Meltzer, A. H. (1986). The Theory of Ambiguity, Credibility, and Inflation under Discretion and Asymmetric Information. *Econometrica*, 54(5), 1099-1128.
- Drazen, A., & Masson, P. R. (1994). Credibility of Policies Versus Credibility of Policymakers. *The Quarterly Journal of Economics*, 109(3), 735-754.
- Faust, J., & Svensson, L. E. (2001). Transparency and Credibility: Monetary Policy with Unobservable Goals. *International Economic Review*, 42(2), 369-397.
- Haldane, A., & McMahon, M. (2018). Central Bank Communications and the General Public. *AEA Papers and Proceedings*, 108, pp. 578-583.
- Haldane, A., Macaulay, A., & McMahon, M. (2020). The 3 E's of central bank communication with the public. *Bank of England Staff Working Paper* No. 847.
- Hayo, B., & Neuenkirch, M. (2015). Self-monitoring or reliance on media reporting: How do financial market participants process central bank news? *Journal of Banking & Finance*, vol. 59, pp. 27-37.
- Lucca, D. O., & Trebbi, F. (2009). Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements. *NBER Working Papers* 15367.
- Łyziak, T., & Paloviita, M. (2016). Anchoring of inflation expectations in the euro area: recent evidence based on survey data. *ECB Working Paper Series* No. 1945.
- Munda, T., & Brookes, J. (2021). Mark my words: the transmission of central bank communication to the general public via the print media. *Bank of England Staff Working Paper* No. 944.
- Sahminan. (2008). Effectiveness of Monetary Policy Communication in Indonesia and Thailand. *BIS Working Paper* No. 262.
- Sobel, J. (1985). A Theory of Credibility. *The Review of Economic Studies*, 52(4), 557-573.
- Tobback, E., Nardelli, S., & Martens, D. (2017). Between Hawks and Doves: Measuring Central Bank Communication. *ECB Working Paper Series* No. 2085.
- Zulen, A. A., & Wibisono, O. (2018). Measuring stakeholders' expectation on central bank's policy rate. *Proceedings of the Ninth IFC Conference*, pp. 1507-1534.
- Zulen, A. A., Wibisono, O., Widjanarti, A. (2020). Developing Machine Learning Technique for Measuring Central Bank Credibility. *2020 Data for Policy Conference*.

Appendix A: Example annotated sentences

Example annotated sentences from news

Table B.1

| No. | Sentence* | Credibility Aspect | Label |
|-----|---|---|--------------------------|
| 1. | <i>BI also decided to hold the policy rate as there are still external and domestic risks that need to be monitored.</i> | Formulation | Positive |
| 2. | <i>Due to BI's "wrong" exchange rate policy, rupiah (IDR) depreciated by more than 75% since September 2011 or about 15% per year.</i> | Formulation | Negative |
| 3. | <i>The increase in BI's policy rate managed to appreciate rupiah.</i> | Effectiveness | Positive |
| 4. | <i>Although BI Rate has been lowered, interest rate on bank loans in Indonesia is entirely not competitive.</i> | Effectiveness | Negative |
| 5. | <i>This policy is in line with the government and Bank Indonesia's aim to manage current account deficit toward more healthy figures.</i> | Coordination | Positive |
| 6. | <i>Some regard that there is political pressure on the monetary authority to lower BI Rate, related to the government's ambition to pursue 5.7% economic growth, even though this presumption has been denied by BI Governor in multiple occasions.</i> | Coordination | Negative |
| 7. | <i>Within 2013 there is still room for BI Rate hike, 50 bps at most.</i> | Expectation / Communication | Tight/ Hawkish |
| 8. | <i>BI Rate this year can be held steady at 5.75% as in last year, as long as there is no increase in inflation.</i> | Expectation / Communication | Neutral |
| 9. | <i>If inflation becomes more controlled, then BI Rate will likely be lowered again.</i> | Expectation / Communication | Accommodative/ Dovish |
| 10. | <i>Exchange rate stability policy and quantitative easing will be continued.</i> | Expectation / Communication – additional | Accommodative/ Dovish |
| 11. | <i>The government and the central bank has agreed to share the burden through the National Economic Recovery Program (PEN), as a response to COVID-19 pandemic.</i> | Expectation / Communication – additional | Accommodative/ Dovish |

*) Translated by the authors from Bahasa Indonesia to English

Appendix B: Regression results for heterogeneity of estimates

Regression results for y = heterogeneity of estimates

Table B.1

| Variable | $y_{\text{heterogeneity}}$ | | |
|-----------------------------------|----------------------------|-------------------|-------------------|
| | 4 indexes | Comm. index | Credib. index |
| Formulation index (%) | -0.017 (0.025) | - | - |
| Effectiveness index (%) | 0.009 (0.025) | - | - |
| Coordination index (%) | -0.020 (0.022) | - | - |
| Communication index (%) | 0.008 (0.018) | 0.009 (0.020) | - |
| Avg. credibility index (%) | - | - | -0.032 (0.035) |
| Intercept (%) | 0.086 (0.051) | 0.042 (0.034) | 0.079 (0.048) |
| Lag of y (%) | -0.063 (0.151) | -0.010 (0.126) | -0.039 (0.137) |
| BI7DRR (policy rate) (%) | -0.003 (0.005) | -0.001 (0.005) | -0.001 (0.005) |
| Inflation (%) | -0.002 (0.004) | 0.001 (0.003) | -0.001 (0.004) |
| GDP (growth, decimal) | -0.206 (0.097) | -0.215 (0.094) | -0.220 (0.091) |
| USD/IDR (growth, decimal) | 0.025 (0.069) | 0.017 (0.070) | 0.017 (0.065) |
| New COVID cases (ln) | -0.001 (0.001) | -0.001 (0.001) | -0.001 (0.001) |
| R² | 0.089 | 0.061 | 0.066 |
| Adjusted R² | -0.053 | -0.038 | -0.031 |
| Wald test p-value | 0.752 | - | - |

Machine Learning for Measuring Central Bank Credibility and Communication

Okiriza Wibisono, Muhammad Abdul Jabbar,
Anggraini Widjanarti, Alvin Andhika Zulen

IFC-Bank of Italy Workshop: Data science in central banking
14-17 February 2022

The views and results expressed here are those of the authors and do not necessarily represent Bank Indonesia.



Perception on **central bank (policy) credibility** plays a **pivotal role** in building public's trust on the institution and managing expectation.



Need an **objective** measure of central bank credibility. Existing approaches (direct survey, inflation expectation) have drawbacks.



Utilizing **Big Data Analytics** – text mining to gather public perception regarding central bank policy credibility.

This research:

- Constructing monetary policy credibility indexes from news
- Evaluating the indexes in a regression setting



News articles

Source: Cyber Library (internal repository of curated economic and financial news)

~30 domestic news (in Bahasa Indonesia)
~850 articles daily

Whole period: since Jan 1999
Training data: Jan 2010 – May 2021
For regression: Oct 2015 – Dec 2021

Keywords for filtering sentences:

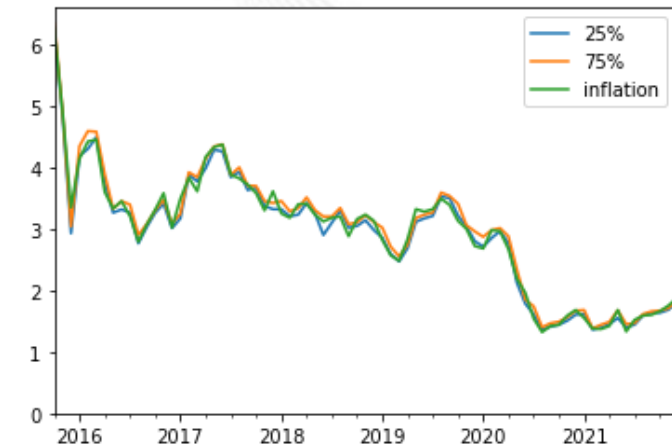
- inflation, monetary, exchange rate, current account, policy rate, BI Rate, BI 77-Day Reverse Repo Rate, (+ their variations)
- Addition for communication idx: QE, reserve requirement, accommodative, liquidity, burden sharing, government bond purchase
- AND the sentence must mention BI/Bank Indonesia (or its previous/next sentence)

Inflation estimates

Source: Bloomberg (Economist Estimates)

Estimate of current month's inflation
~20 economists per month, ~10 domestic

| | |
|--|---------------|
| Mean of monthly mean | 2.951% |
| Mean of monthly standard deviation | 0.097% |
| Mean of monthly absolute deviation from actual inflation | 9 bp |
| Number of times actual inflation lower than percentile 25 estimate | 33 (44%) |
| Number of times actual inflation higher than percentile 75 estimate | 23 (30.3%) |
| Number of times mean estimate lower than lower bound of inflation target | 21 (27.6%) |
| Number of times mean estimate higher than upper bound of inflation target | 1 (1.3%) |



Sample statistics from domestic economists only

1. Annotation

A sample of filtered sentences are annotated as training data for ML classification models.

- Annotated by authors and experts within BI.
- Guidelines incl. examples
- 2-3 annotators per sentence

| Distribution | Pos* | Neg* | Neutral |
|----------------------------|------------------|------------------|----------------|
| Formulation | 1,595 (81.6%) | 360 (18.4%) | N/A |
| Effectiveness | 2,072 (81.1%) | 482 (18.9%) | N/A |
| Coordination | 303 (86.8%) | 46 (13.2%) | N/A |
| Communication | 493 (27.6%) | 391 (33.0%) | 706 (39.4%) |
| Communication – additional | 3 (0.1%) | 3,358 (99.7%) | 6 (0.2%) |

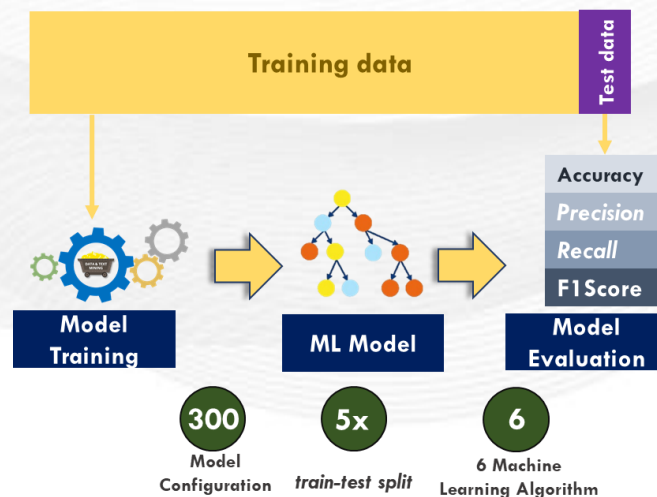
A total of ~20,000 sentences are annotated, majority are irrelevant (neither positive/ negative/ neutral)

*) For communication index, pos = tight, neg = accommodative.

3. Model training

ML model is trained for classifying sentences into pos/neg labels, for each aspect.

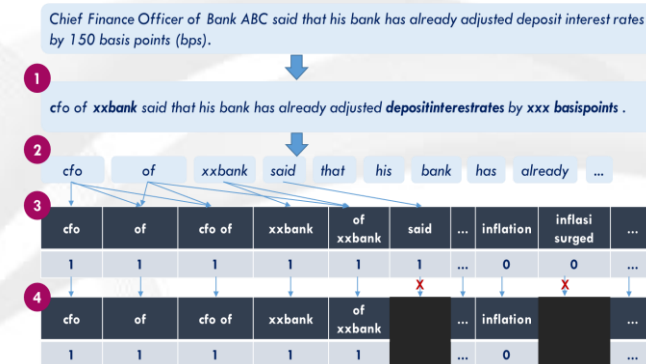
- 90-10 train test split, repeated 5 times
- >300 model configurations
- 6 ML algorithms
- Avg F1 of best model: 63.4%



2. Data preprocessing

Each sentence is transformed from text into tabular-numeric format for training ML models.

- Sentence cleansing
- Tokenization
- N-gram vectorization
- Remove sparse terms



4. Index calculation

The ML models are applied to all sentences, to construct monthly indexes.

$$Index_{aspect\ k,t} = \frac{\#positive_{k,t} - \#negative_{k,t}}{\#positive_{k,t} + \#negative_{k,t}}$$

$$Index_{stance\ perception,t} = \frac{\#tight_t - \#accommodative_t}{\#tight_t + \#neutral_t + \#accommodative_t}$$

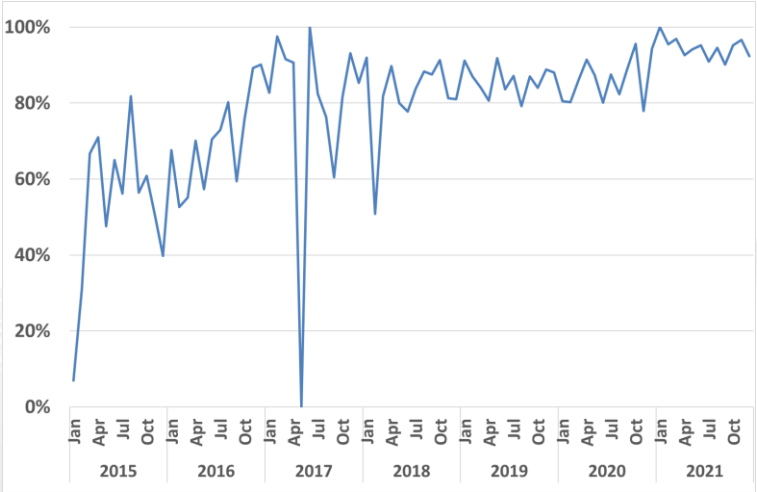
$$Index_{communication,t} = 1 - |Fwd\ Guidance_t - Indeks_{stance\ perception,t}|$$

| No | Sentence | Credibility Aspect | Label |
|----|---|--|------------------------|
| 1 | <i>BI also decided to hold the policy rate as there are still external and domestic risks that need to be monitored.</i> | Formulation | Positive |
| 2 | <i>Due to BI's "wrong" exchange rate policy, rupiah (IDR) depreciated by more than 75% since September 2011 or about 15% per year.</i> | Formulation | Negative |
| 3 | <i>The increase in BI's policy rate managed to appreciate rupiah.</i> | Effectiveness | Positive |
| 4 | <i>Although BI Rate has been lowered, interest rate on bank loans in Indonesia is entirely not competitive.</i> | Effectiveness | Negative |
| 5 | <i>This policy is in line with the government and Bank Indonesia's aim to manage current account deficit toward more healthy figures.</i> | Coordination | Positive |
| 6 | <i>Some regard that there is political pressure on the monetary authority to lower BI Rate, related to the government's ambition to pursue 5.7% economic growth, even though this presumption has been denied by BI Governor in multiple occasions.</i> | Coordination | Negative |
| 7 | <i>Within 2013 there is still room for BI Rate hike, 50 bps at most.</i> | Expectation / Communication | Tight/Hawkish |
| 8 | <i>BI Rate this year can be held steady at 5.75% as in last year, as long as there is no increase in inflation.</i> | Expectation / Communication | Neutral |
| 9 | <i>If inflation becomes more controlled, then BI Rate will likely be lowered again.</i> | Expectation / Communication | Accommodative / Dovish |
| 10 | <i>Exchange rate stability policy and quantitative easing will be continued.</i> | Expectation / Communication – additional | Accommodative / Dovish |
| 11 | <i>The government and the central bank has agreed to share the burden through the National Economic Recovery Program (PEN), as a response to COVID-19 pandemic.</i> | Expectation / Communication – additional | Accommodative / Dovish |

Formulation

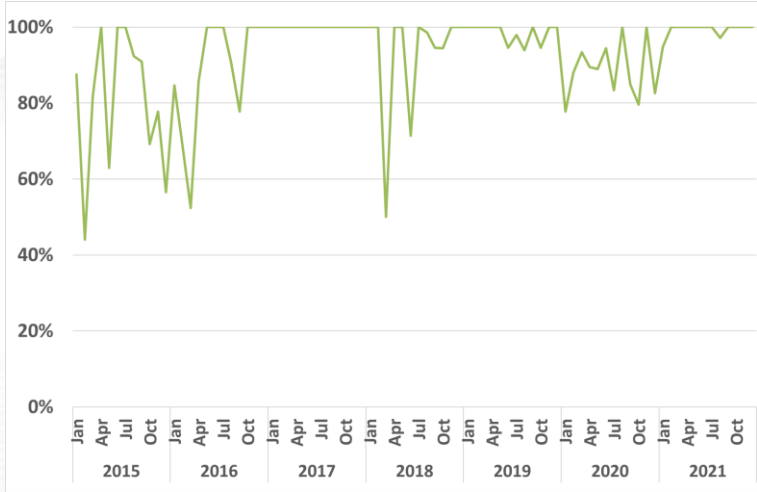
Correlation of avg. policy credibility index with survey results (2013-2018): 79.7%

Coordination



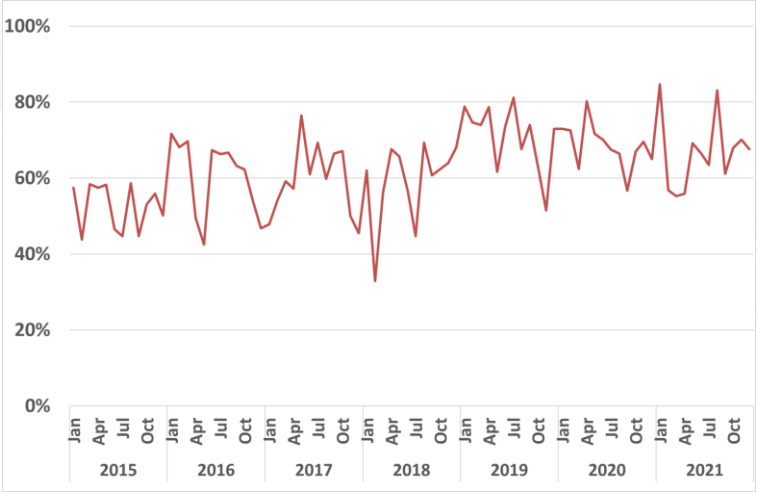
| | |
|-------|--------|
| Min | 0.0% |
| Mean* | 82.8% |
| Max | 100.0% |
| Std* | 12.9% |

*) Excluding outlying observation in May 2017



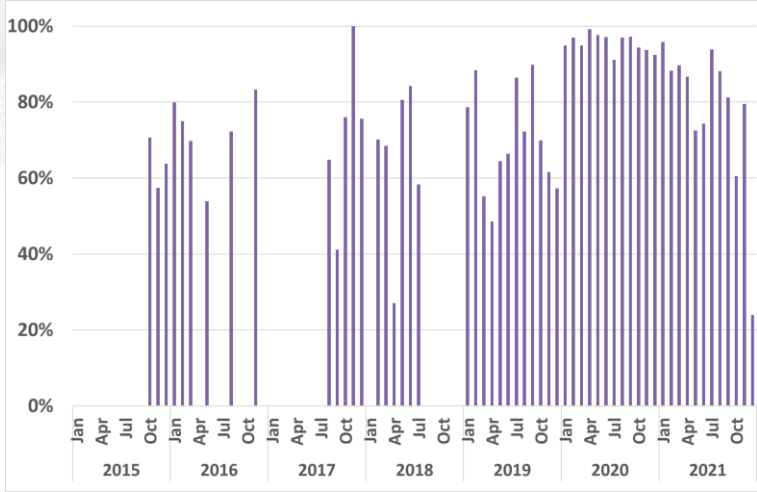
| | |
|------|--------|
| Min | 50.0% |
| Mean | 93.4% |
| Max | 100.0% |
| Std | 11.6% |

Effectiveness



| | |
|------|-------|
| Min | 32.8% |
| Mean | 63.8% |
| Max | 84.7% |
| Std | 10.0% |

Communication



| | |
|------|--------|
| Min | 27.0% |
| Mean | 75.8% |
| Max | 100.0% |
| Std | 14.7% |

We test whether news content, as summarized by the four indexes and the average policy credibility index, are incorporated into economists' estimates of inflation in that they make the estimates more accurate and/or anchored*.

$$y_{inaccuracy,t} = |\bar{\pi}_{t,t-1} - \pi_t|$$

$$y_{unanchoring,t} = |\bar{\pi}_{t,t-1} - \pi_t^*|$$

- Period: Oct 2015 to Dec 2021, monthly
- Variables of interest (3 sets of equations):
 - a. 4 policy credibility indexes
 - b. communication index (not shown*)
 - c. average policy credibility index
- Controls: lag of y, BI7DRR, lag inflation, GDP growth yoy (interpolated), USD/IDR % change mtm, ln of new COVID cases
- Communication index lagged by 1 month to allow its distinctive time period (from one BoG monthly meeting to the next, instead of from date 1 to end of month)
- OLS + HAC standard errors
- Result sensitive to outlying observation of formulation index in May 2017

| Variable/Statistic | Y inaccuracy | | Y unanchoring | |
|---------------------------------|-------------------|-------------------|-----------------------------|----------------------------|
| Formulation index | 0.063 (0.053) | - | -0.446*** (0.144) | - |
| Effectiveness index | -0.087 (0.083) | - | -0.864* (0.473) | - |
| Coordination index | -0.089 (0.078) | - | -0.159 (0.337) | - |
| lag(Communication index) | -0.044 (0.077) | - | -0.190 (0.290) | - |
| Avg. credibility index | - | -0.068 (0.211) | - | -1.457* (0.780) |
| lag(Y) | -0.113 (0.125) | -0.118 (0.114) | 0.445*** (0.133) | 0.458*** (0.121) |
| lag(Inflation) | 0.001 (0.018) | 0.000 (0.017) | -0.111** (0.055) | -0.097** (0.048) |
| R2, Adj. R2 | 0.135, -0.000 | 0.096, 0.002 | 0.785, 0.751 | 0.775, 0.752 |
| F-statistic (Wald test) | 0.088* | - | 0.013** | - |

*) Results from only including communication index are all not statistically significant, as well as results on y = heterogeneity of estimates (as coefficient of variation).

Conclusion

- 1 **Machine learning methodology for measuring perceptions of monetary policy credibility from news**
- 2 **Communication index includes recent policy developments, e.g. non-interest rate policy (QE, reserve requirement), coordinative policies with the government**
- 3 **Evaluation of the policy credibility indexes in econometric model of economists' inflation estimates. Results sensitive to outlier of formulation index**

Future Works

- 1 **Analysis on effect on other economic estimates, e.g. further horizon, GDP, general public**
- 2 **More elaborate econometric models for evaluation, e.g. panel of economists**
- 3 **Data source/coverage addition, e.g. English news, additional keywords. Macroprudential and payment system policies?**

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

A machine learning approach to narrative retrieval in economic news: the case of oil price uncertainty^{1 2}

Donald Jay Bertulfo,
Delft University of Technology & Asian Development Bank

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

² The author prepared part of this research while working at the Asian Development Bank (ADB). Work on index construction and evaluation benefitted from close guidance by Dr. Abdul Abiad and Dr. Irfan Qureshi, extensive support from Dr. Elisabetta Gentile, all of which are ADB economists. Meanwhile, subsequent work on natural language processing was done under the close supervision of Dr. Rachelle Sambayan and Dr. Fredegusto David of the University of the Philippines-Diliman.

A Machine Learning Approach to Narrative Retrieval in Economic News

The Case of Oil Price Uncertainty

Donald Jay Bertulfo

Delft University of Technology

February 17, 2022

Overview

- 1 Introduction and Motivation
- 2 Literature Review and Contributions
- 3 Methodology
 - Text mining and document feature extraction
 - News-based OPU index construction
 - Text preprocessing and cleaning
 - Generating document embeddings
 - Measuring pairwise article similarities
 - Detecting article communities
 - Recasting document labels via NPMI
 - Topic Modelling
- 4 Results
- 5 Conclusion and Recommendations

Text as Data

Emerging research interest in economics: The analysis of large volumes of texts to uncover patterns in economic dynamics. Recent NLP applications include:



Estimating the impact of political statements on market outcomes



Nowcasting the macroeconomy



Measuring the sentiment embedded in economic news



Quantifying the level of uncertainty surrounding policies



Analyzing and forecasting stock prices

Analysis of economic shocks

The status quo: Pursued from a model-based, deterministic view of the macroeconomy

Our argument: The nature and magnitude of economic shocks are context-specific and temporally heterogeneous

This research: aims to shed light on this complexity by pursuing the question from a data-driven, inductive and grounded approach



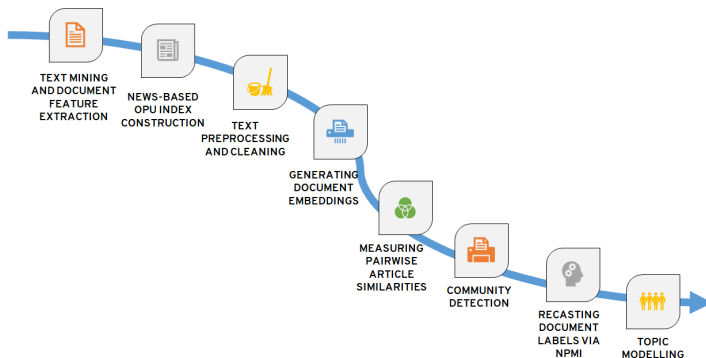
Extract salient themes that are common across episodes of economic shocks or crises



Retrieve salient narratives underlying shock episodes based on the content of talk in those moments

Analysis of economic shocks

Taking oil price uncertainty as a case study,



Literature Review and Contributions

- Linguistic information contained in texts reveals patterns that are useful in shedding light on the complexity of interactions among economic variables
- This research



contributes to literature that use news- or text-based methods to track economic movements



builds on applications of text-based dimension reduction techniques to extract salient themes in texts



extends literature on the use of neural networks to capture semantic regularities in text

Text Mining Procedure

News articles were mined from the Factiva search database, a Dow Jones news aggregator

The search algorithm takes articles



containing the trio of terms "oil or petrol or petroleum or gasoline," "gas", "price" and "uncertainty" (together with its synonyms)



which are no more than 2 words apart from each other



and with word count >99



+ restrictions on article type (not opinion, editorial, etc)

Downloaded full articles were stored in RTF format (100 articles per file)

Document feature extraction

- A rudimentary R code was used to extract important textual features such as title, date, publication, body of text and word count from each article.



Figure: Sample Labeled Article Snapshot

The news-based OPU index

- Following Baker, Bloom and Davis (2016), the following standardization and normalization procedures were applied:
 - ① For each month-year-newspaper, raw counts were scaled by the total number of articles.
 - ② The month-year-newspaper level series was standardized to unit standard deviation from January 1969 to January 2020. Then, averages were taken across the 50 newspapers by month-year.
 - ③ Finally, the 50-newspaper series was normalized to a mean of 100 from January 1969 to January 2020.
- A sanity check was performed by plotting the OPU index across time points and identifying episodes in the history of oil price shocks which may help explain spikes in the OPU index

The news-based OPU index

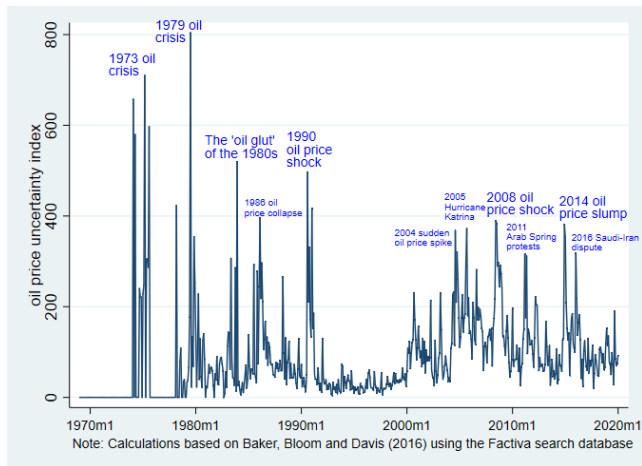


Figure: News-based global oil price uncertainty index, Jan 1969 - Jan 2020

Further validation

- The news-based OPU index was plotted vis-a-vis popular measures of oil price uncertainty to check for series co-movements. These measures include:
 - 60-day historical Brent volatility index (HVOLBREN)
 - 3-months implied volatility index (IVOLBREN)
 - 30-day volatility index (OVX)
- Correlation analysis suggests strong comovements between the news-based OPU index and other market-based measures of oil price volatility

Further validation

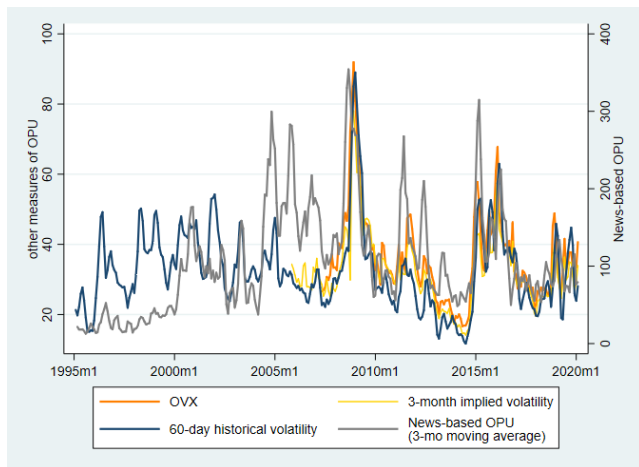


Figure: News-Based Global Oil Price Uncertainty (OPU) and Other Measures of OPU (January 1995 - January 2020)

Text preprocessing and cleaning

The following preprocessing procedures were implemented:



Tokenization

refers to the process of converting a document into a sequence of symbols called tokens. At this stage, each article was converted to unigrams (one-word tokens) using the tidytext package in R



Case folding and digit normalization

tokens were lowercased, punctuations and digits were removed from the corpus using regular expressions



Stopwords removal

high-frequency words with low information content (e.g. 'the', 'is', etc) were removed. These stopwords are documented in the R 'stopwords' package

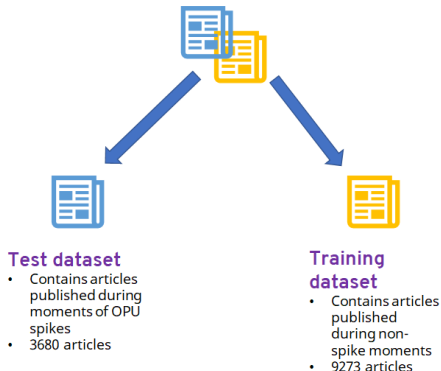


Text stemming

words were stemmed to their root using the Porter algorithm in R.

Text preprocessing and cleaning

- Further preprocessing involved reconstructing the cleaned and pruned articles (concatenation of word tokens).
- The entire text corpus was partitioned into two groups
- The cleaned and pruned training dataset was then fed into a doc2vec model (a type of paragraph embedding model)



Document embeddings

- **Distributional hypothesis:** words which occur in similar contexts tend to have similar meanings
- Ergo, documents that embed similar words tend to talk about the same things
- **Estimation Strategy:**
 - represent words as vectors;
 - input the word vectors, together with document labels into a neural network model;
 - model learns document semantics; pretrained model can be used to infer vectors for additional documents (i.e., articles in the test set)
- Here, a doc2vec model was implemented to learn vector space representation of documents

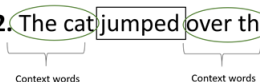
Document embeddings: the doc2vec models

- Doc2vec models are unsupervised neural network models that generate fixed-length vector representations (also called "paragraph vectors") from variable lengths of texts
- Two types:
 - **PV-DM/Distributed memory:** learns to predict the occurrence of a center word, given context words and a document label.
 - **PV-DBOW/Distributed bag-of-words:** learns to predict context words given document label

The cat jumped over the puddle

Center word: jumped

Assume context window=2. The cat jumped over the puddle.



Document embeddings: the doc2vec models

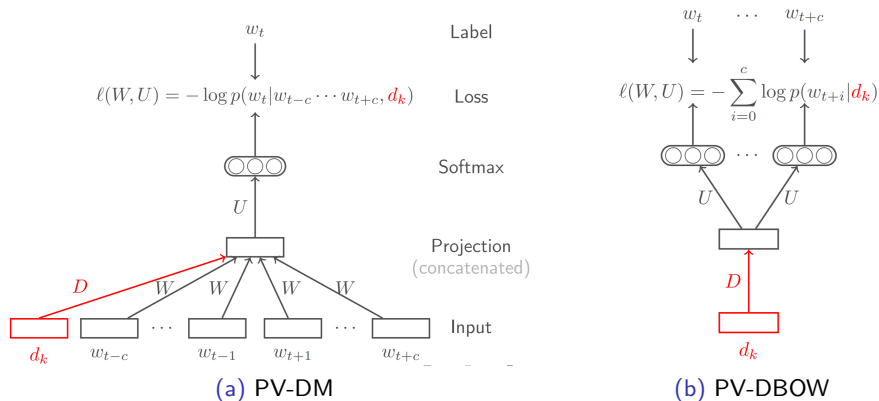


Figure: Doc2vec model architectures

Document embeddings: the doc2vec models

- The reconstructed cleaned and pruned articles in the training set were fed into the Doc2Vec DBOW model, using the following parameters: max epochs = 100, vector size = 300, alpha = 0.025, min alpha = 0.00025, number of negative samples = 5, window size = 5, sample = 0.001.
- After training, the pre-trained doc2vec model was used to *infer a paragraph vector for each document in the test corpus*
- This process yielded a matrix of size 3860 (number of articles in test set) \times 300 (number of features in hidden layer). For our purposes, we call this matrix the **paragraph matrix**

Pairwise document similarities

- Given inferred paragraph vectors for each article in the test set, we are now interested in knowing **how similar the articles in the test set are**.
- To compute for semantic closeness irrespective of document length, we invoke the concept of cosine similarity

Cosine similarity

Given two nonzero vectors \mathbf{x} and \mathbf{y} of length n , cosine similarity (i.e., $\text{cos_sim}(\mathbf{x}, \mathbf{y})$) is given by the formula

$$\text{cos_sim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

where $\|\cdot\|$ represents the Euclidean norm and $x_i, y_i, i = 1, 2, \dots, n$ are the components of \mathbf{x} and \mathbf{y} respectively.

Pairwise document similarity

- To estimate pairwise semantic closeness in the document space, cosine similarities were computed for each pair of paragraph vectors \mathbf{d}_i and \mathbf{d}_j .
- Values were stored in a matrix $\mathbf{P} = [\cos_sim(\mathbf{d}_i, \mathbf{d}_j)]$ to yield a symmetric matrix of pairwise cosine similarities (i.e., note that $\cos_sim(\mathbf{d}_i, \mathbf{d}_j) = \cos_sim(\mathbf{d}_j, \mathbf{d}_i)$ for any i, j).
- Establishing a link between matrix and graph theory, \mathbf{P} can also be regarded as the matrix representation of a weighted underlying similarity graph with documents as nodes

Community detection

- Taking \mathbf{P} as an underlying matrix representation of a similarity graph with articles as nodes, community detection was performed by inputting this adjacency matrix into the Louvain algorithm.
- Louvain is a popular unsupervised clustering algorithm that detects graph communities based on the principle of modularity maximization.
- **Modularity** is a measure of the density of links inside communities as compared to links between them.

Community detection

Modularity

For weighted networks, modularity is given by:

$$M = \frac{1}{2W} \sum_{i,j} \left[w_{i,j} - \frac{k_i k_j}{2W} \right] \delta(c_i, c_j)$$

where $w_{i,j}$ refers to the weight between edges i and j , k_i is the sum of the weights of links that connect to node i , W is the sum of all the links in the network, c_i refers to the community where node i belongs and $\delta(c_i, c_j)$ is a function that takes the value of 1 if $c_i = c_j$ and 0 otherwise.

Community Detection: The Louvain algorithm (Phase 1)

- 1 Assign each node i to a distinct community C_i .
- 2 For each node i , do the following:
 - 1 Identify the set of neighbors J of i
 - 2 For each neighbor $j \in J$, evaluate the gain in modularity that would take place by removing i from its community to C_j . The change is given by the following formula:

$$\Delta M = \left[\frac{\sum_{in} + 2k_{i,in}}{2W} - \left(\frac{\sum_{tot} + k_i}{2W} \right)^2 \right] + \left[\frac{\sum_{in}}{2W} - \left(\frac{\sum_{tot}}{2W} \right)^2 - \left(\frac{k_i}{2W} \right)^2 \right]$$

where \sum_{in} is the sum of the weights of the links inside community C_j , \sum_{tot} is the sum of the weights of links to all nodes in C_j , $k_{i,in}$ is the sum of the weights of links from node i to node C_j

- 3 Move node i to the community with the biggest modularity gain. If no positive gain is found, i stays in its original community
- 3 Repeat the process until no further improvement can be achieved

Community Detection: The Louvain algorithm (Phase 2)

- 1 Construct a new network whose nodes are the communities found in Phase 1. In this new network, the weights of the links between the new nodes are given by the sum of the weights of all links between the nodes in the corresponding communities. Weighted self-loops reflect links between nodes of the same community.

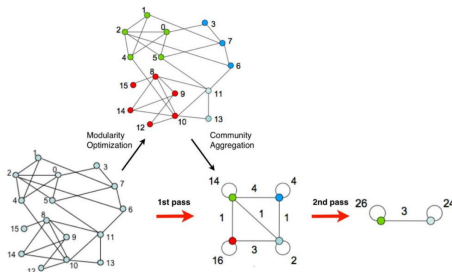


Figure: Louvain algorithm, a visualization

Community Detection: Implementation

- 1 **P** is then fed into the Louvain algorithm (network weights correspond to entries of **P**), which was ran in R via the igraph package.
- 2 The algorithm yielded a category label for each article $d_n \in \mathcal{D}_{test}$ wherein the category label reflects the community membership of d_n .
- 3 **Three document/article communities were detected by the Louvain algorithm.** To explore the semantic structure of these communities, the articles were grouped by community membership and visualized using wordclouds.

Sanity check: Wordclouds



Figure: Community 1 Wordcloud

Sanity check: Wordclouds

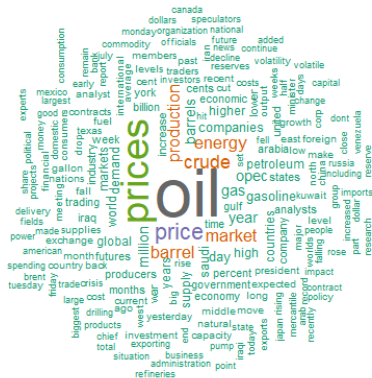


Figure: Community 2 Wordcloud

Sanity check: Wordclouds

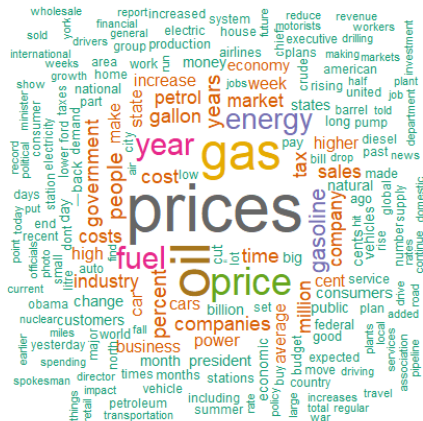


Figure: Community 3 Wordcloud

Recasting document labels via NPMI

- ❶ Constructing a word-by-community co-occurrence matrix
- ❷ Computing for normalized pointwise mutual information

$$NPMI(\mathcal{W}_i, C_k) = \frac{\log\left(\frac{p(\mathcal{W}_i, C_k)}{(p(\mathcal{W}_i)p(C_k))}\right)}{-\log p(\mathcal{W}_i, C_k)}$$

$p(\mathcal{W}_i)$ pertains to the relative salience of word \mathcal{W}_i in the entire corpus, $p(C_k)$ denotes the share of community C_k in the total number of articles in the entire corpus and $p(\mathcal{W}_i, C_k)$ represents the probability that word \mathcal{W}_i and community C_k co-occur.

- ❸ Computing for summary scores

$$S_{i,k} = NPMI(\mathcal{W}_i, C_k) - \sum_{n \neq k} NPMI(\mathcal{W}_i, C_n).$$

for each unique word \mathcal{W}_i in the corpus

- ❹ Aggregating summary scores
- ❺ Revising community labels of each article in \mathcal{D}_{test} according to the following decision criterion: $\hat{C}_k(d_n) = \max_k \mathbb{E}_n[S_{i,k}]_{\mathcal{W}_i \in V}$

Topic Modelling

- A **topic** is a "distribution over a fixed vocabulary"
- Topics consist of words, which are spread out in texts
- The goal of topic modelling is to discover, in an unsupervised manner, latent themes from documents
- Topic extraction using latent Dirichlet allocation was implemented using the Mallet package in Python
- Using regression analysis, we identify which among the topics extracted were most associated with which document community

Topic Classification

| Topic | Top keywords | C1 | C2 | C3 |
|-------|---|----|----|----|
| 1 | market, share, pc, price, group, year, time, ftse, bank, oil | ✓ | | |
| 2 | compani, share, deal, offer, million, sell, valu, bid, plan, board | ✓ | | |
| 3 | war, iraq, kuwait, crisi, natio, gulf, oil, presid, middle_east, russia | | ✓ | |
| 4 | gas, energi, natur, year, price, oil, state, texas, pipelin, drill | | | ✓ |
| 5 | year, job, busi, work, worker, cut, incom, small, pay, budget | | | ✓ |
| 6 | energi, electr, power, heat, fuel, cost, home, effici, gas, system | | | ✓ |
| 7 | airlin, fuel, year, cost, carrier, fare, air, hedg, travel, flight | | | ✓ |
| 8 | time, peopl, good, thing, long, make, back, mani, lot, reason | | | ✓ |

Topic Classification

| Topic | Top keywords | C1 | C2 | C3 |
|-------|---|----|----|----|
| 9 | car, vehicl, sale, fuel, year, auto, truck, ford, gas, model | | | ✓ |
| 10 | cent, dollar, yesterday, currenc, euro, gold, close, european, week, lowe | ✓ | | |
| 11 | oil, product, opec, price, produc, barrel, countri, saudi_arabia, world, market | | ✓ | |
| 12 | uk, britain, pound, british, govern, cut, energi, warn, industri, brown | | | ✓ |
| 13 | canada, canadian, govern, today, billion, cent, year, report, busi, price | ✓ | | |
| 14 | stock, index, market, fell, point, gain, rose, investor, close, share | ✓ | | |
| 15 | presid, state, obama, republican, bill, feder, bush, congress, senat, admi | | | ✓ |
| 16 | price, gasolin, gas, gallon, station, averag, pump, refinери, week, day | | | ✓ |

Topic Classification

| Topic | Top keywords | C1 | C2 | C3 |
|-------|--|----|----|----|
| 17 | price, cost, increas, compani, custom, consum, rise, pay, higher, bill | | | ✓ |
| 18 | rate, economi, growth, inflat, econom, bank, year, interest_r, rise, econo | ✓ | | |
| 19 | project, develop, industri, plant, build, power, plan, invest, product, busi | | | ✓ |
| 20 | price, year, sinc, drop, month, fall, week, expect, declin, lower | | | ✓ |
| 21 | time, show, page, univers, call, work, offic, art, polic, peopl | | | ✓ |
| 22 | market, oil, price, trade, futur, contract, trader, crude, specul, commod | | ✓ | |
| 23 | market, invest, investor, fund, stock, bond, year, manag, money, equiti | ✓ | | |
| 24 | percent, year, sale, month, report, retail, consum, spend, juli, increas | ✓ | | |

Topic Classification

| Topic | Top keywords | C1 | C2 | C3 |
|-------|---|----|----|----|
| 25 | oil, price, barrel, crude, demand, energi, high, rise, higher, suppli | | ✓ | |
| 26 | govern, elect, polit, world, meet, econom, leader, countri, issu, parti | | | ✓ |
| 27 | china, global, japan, world, economi, export, year, growth, econom, countri | ✓ | | |
| 28 | govern, bank, india, market, sector, year, polici, increas, current, credit | ✓ | | |
| 29 | oil, compani, industri, bp, explor, product, billion, shell, field, analyst | | ✓ | |
| 30 | tax, price, petrol, fuel, budget, litr, govern, cost, increas, diesel | | | ✓ |
| 31 | citi, peopl, day, servic, area, road, transport, drive, home, car | | | ✓ |
| 32 | million, year, billion, compani, profit, quarter, share, earn, revenu, expenditur | ✓ | | |

Topic Salience During Shock Moments

- **1974-1975:** OPEC, economic growth, politics and jobs
- **1979:** oil demand/supply, innovation
- **1983-1985:** OPEC, natural gas
- **1986:** oil supply/demand, OPEC, decline in gas prices, speculation in oil prices
- **1990:** stock prices, oil demand/supply, war, speculation in oil prices
- **2004:** retail gasoline prices, consumer prices, stock market, economic growth, oil demand/supply, profits, consumption
- **2006:** retail gasoline prices, tax
- **2008:** travel, decline in oil prices, automotives, jobs
- **2012:** retail gasoline prices, American politics, natural gas, innovation
- **2015-2016:** decline in gas prices, share price

Conclusion

- Machine learning has immense potential to uncover underlying narratives behind economic shock episodes
- The talk about economic shocks is embedded in multiple, interlocking topics
- Context matters in policy!

Recommendations

- Hyperparameter tuning and sensitivity analysis
- Linking computationally-derived narratives with macroeconomic time series
- Apply the same narrative extraction technique to unpack stories underlying other types of shocks (e.g. COVID-19 pandemic, geopolitical conflicts)

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Creation of a structured sustainability database from company reports – a web application prototype for information retrieval and storage¹

Eugenia Koblents and Alejandro Morales,
Bank of Spain

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Creation of a structured sustainability database from company reports

A web application prototype for information retrieval and storage

Alejandro Morales

Eugenia Koblents

Statistics Department, Banco de España

Abstract

A web application for semi-automated information extraction and storage has been developed at the Banco de España for retrieval of sustainability indicators from annual financial statements reported by Spanish non-financial corporations. The goal is to assist business users in the process of extracting sustainability indicators from large document databases and storing them in a structured format. The tool developed incorporates a set of pre-defined search terms for each indicator, which have been selected based on domain knowledge. For each company and indicator, the tool suggests the most relevant text snippets to the user, who identifies the correct indicator's value and stores it in the database via the user web interface.

This tool has been developed by two data scientists in three months, with the continuous support of a domain expert team for definition of requirements and refinements, input data collection and tool validation and testing.

The tool has been entirely implemented in Python and provides for interaction with the user by means of a user web interface. The web application has already been used by 20 people in the Banco de España's Statistics Department to create an initial version of the sustainability database, which currently contains more than 15,000 records, retrieved from a total of 800 reports submitted by 300 of the largest Spanish companies. To date, the new sustainability database contains information on 39 environmental, social and governance indicators, with plans for it to be extended in the near future.

This paper describes the technical approach adopted and the main modules of the prototype implemented, including text extraction, indexing and search, data storage and visualisation. It also presents an overview of the first version of the sustainability database created.

Keywords: sustainability, climate change, full-text search, OCR, databases, web application, Python.

JEL classification: Q5 (Environmental economics).

Index

- A web application prototype for information retrieval and storage 1
- 1. Introduction 3
 - Target information..... 3
- 2. Technical approach..... 6
 - Input data description..... 7
 - Digital and scanned text extraction 8
 - Full-text indexing and search 9
 - Data storage..... 10
 - User interface..... 11
- 3. First data ingestion 13
- 4. Conclusions..... 14
 - Next steps 14
- References..... 16

1. Introduction

The Banco de España, the Spanish central bank, embarked on several avenues of research on sustainable finance in its 2021-2022 analytical programme. In this context, in March 2021 Statistics Department staff (data scientists and accounting experts), together with members of other departments, launched a project to create a structured database containing sustainability information reported by non-financial companies in their annual financial reports.

Although some sustainability indicators have become mandatory in the non-financial reports annually submitted by Spanish non-financial corporations to the Mercantile Registers, they are often reported in an unstructured format, such as tables or images contained in the annexes of their annual non-financial statements.

To allow for the creation of the new database in a short period of time (March to July 2021), an experimental prototype has been developed out of the Banco de España's regular IT environment. A web application for semi-automated information extraction and storage has been developed, which implements (digital and scanned) text extraction, indexing, search and storage in a relational database. The tool suggests the most relevant text fragments to the user, who needs to validate the search results by selecting the correct value for each indicator and then stores it in the database.

The tool developed has recently been used by 20 business experts to create the first version of the new sustainability database, which is hosted at the Central Balance Sheet Data Office (CBSO), a division of the Banco de España's Statistics Department. After the first phase of this project, the new sustainability database contains 39 environmental, social and governance (ESG) indicators (selected from a list of over 100), 77% of which are included in the Global Reporting Initiative (GRI) standard.

This paper describes the technical approach designed to create the new sustainability database and the results obtained during the first ingestion phase. Section 1 presents the motivation and goals of this project and describes the target information and the main challenges faced. Section 2 describes the main modules and technical details of this experimental prototype. Section 3 sets out the results obtained during the first data ingestion process, creating the first version of the sustainability database. Lastly, Section 4 summarises the conclusions and future avenues of research envisaged.

Target information

In the first phase of this project, completed during 2021, the goal was to retrieve the value of the 39 continuous and categorical sustainability indicators listed in Table 1.

| Environmental indicators (17): |
|--|
| <ul style="list-style-type: none">▪ Energy consumption and reduction (MWh, GJ, etc.)▪ Total water consumption (m3, Hm3, mega litres, etc.)▪ Greenhouse gas emissions (Scope 1, 2 and 3) (tCO2e, etc.)▪ Greenhouse gas emission reduction (tCO2e, etc.)▪ Circular economy (yes, no)▪ Greenhouse gas emission intensity (ratio) |

| |
|--|
| <ul style="list-style-type: none"> ▪ Environmental policy (yes, no) ▪ Total waste generated (t) ▪ Waste not destined for disposal (t) ▪ Hazardous waste (t) ▪ Percentage of renewable energy (%) ▪ Company located in a stress area regarding water? (yes, no) ▪ ISO 14001 certification (yes, no) ▪ Non-hazardous waste (t) |
| Social indicators (18): |
| <ul style="list-style-type: none"> ▪ Diversity plan (yes, no) ▪ Number of employees ▪ Number of employees with disabilities ▪ Gender diversity (number of women employed) ▪ Gender diversity on the board (% of women) ▪ Permanent employees (%) ▪ Equality plan (yes, no) ▪ Health and safety policy (yes, no) ▪ Human rights policy (yes, no) ▪ Average age of the workforce (years) ▪ Number of dismissals ▪ Average pay gap (%) ▪ Employee absenteeism (days, hours) ▪ Employee turnover (number of employees who leave voluntarily) ▪ Employee training (hours) ▪ Work-life balance measures (yes, no) ▪ Payments to suppliers (days) ▪ Customer satisfaction level (%) |
| Governance indicators (4): |
| <ul style="list-style-type: none"> ▪ Number of corruption and bribery complaints ▪ Channel for complaints (yes, no) ▪ Average board remuneration (€) ▪ Crime prevention policy (yes, no) |

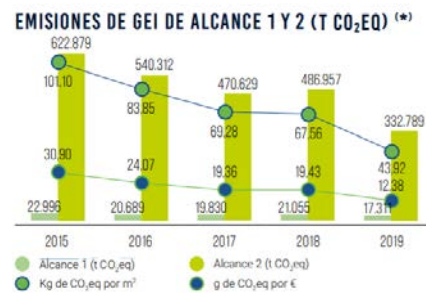
Table 1 ESG indicators collected during the first data ingestion phase.

The target information is presented in highly heterogeneous formats within the company reports, including plain text, tables, graphics and images. Figure 1 shows examples of the ways in which the information is presented. To date, no standard has been defined on how this information should be presented or how the metrics should be used. For this reason, a flexible tool capable of extracting the information in all possible formats had to be designed. European Union regulations are adapting and clear rules on how information should be reported are expected to be defined in the near future, making the information extraction process much easier.

Owing to the variety of formats in which information is reported, a semi-automated approach for information retrieval has been preferred, requiring user validation, in order to guarantee that only high quality data populate the new database. Documents usually contain both text in digital format and (often low-quality) scanned images, which would make manual information extraction extremely costly. A fully automated information retrieval approach was ruled out owing to the

impossibility of handling such a complex and heterogeneous information retrieval problem without requiring human validation.

- Reducimos nuestras emisiones de carbono un 49,6% respecto a 2015 (alcances 1+2) y 18,5% las de nuestra cadena de valor (alcance 3) respecto a 2016
- Reducimos las emisiones de nuestra cadena de suministro por euro comprado un 24,6% respecto a 2016
- Con nuestros servicios evitamos más de 3,2 millones de tCO₂, 3,3 veces nuestra huella de carbono
- Redujimos un 71,8% nuestro consumo de energía por unidad de tráfico



| | | |
|--|-----------|-----------|
| Emisiones Alcance 1 (tCO ₂ e) | 297.042 | 291.787 |
| Emisiones Alcance 2 (basado en método de mercado) (tCO ₂ e) | 1.615.146 | 1.153.046 |
| Emisiones Alcance 1 y 2 (tCO ₂ e) | 1.912.188 | 1.444.833 |



Figure 1 Examples of the ways information is presented in the reports.

Previous attempts have been made to extract sustainability information from unstructured company reports. However, to the best of our knowledge, none of these have produced a database of sustainability indicators, which is the main goal of the present work. In [5] the authors apply text mining techniques to analyse the Task Force on Climate-related Financial Disclosures (TCFD) recommendations on climate-related disclosures of the 12 Spanish significant financial institutions, using publicly available corporate reports from 2014 to 2019. In [6] the authors present an extension of this work to Pillar 3 reports.

In its 2018 and 2019 Status Reports, the TCFD also used supervised machine learning techniques to identify areas of the corporate reports potentially containing information related to each one of 11 recommended disclosures related to the four recommendations [5]. In [1] and [8] the authors train ClimateBert, a deep neural language model, on thousands of sentences related to climate-risk disclosures aligned with the TCFD recommendations. This model can be used for various climate-related downstream tasks like text classification, sentiment analysis and fact-checking.

Other attempts to analyse climate-related documents using machine learning include [2], [3], [4] and [7]. In [2] the authors use machine learning to automatically identify disclosures of five different types of climate-related risks, creating a dataset of over 120 manually-annotated annual reports by European firms. In [3] natural language processing (NLP) techniques are applied to pinpoint the companies that divulge their climate risks and those that do not, identify the types of vulnerabilities that are disclosed and follow the evolution of these risks over time. In [4] a custom NLP model named ClimateQA is proposed, which enables analysis of financial reports in order to identify climate-relevant sections based on a question answering approach. Finally, in [7] the authors explore the performance indicators disclosed in the GRI-based Sustainability Reports (SRs) produced by the companies of three different countries: Italy, Spain and Greece. They use regression trees to describe how the companies' variables explain a different use of the indicators. Their findings show that Spanish companies, on average, disclose the greatest number of indicators. Labour-related social indicators are those most frequently reported in the SRs of the three countries.

2. Technical approach

The technical goal of this project is to transform the non-structured information contained in large collections of documents into a structured relational database, in order to be able to combine it with additional information, to obtain meaningful insights into sustainability and climate change.

A semi-automated tool with full-text search and storage capabilities has been developed to fulfil this goal. The tool combines an offline and an online processing part and provides a user web interface to allow for easy interaction with the end user. Figure 2 summarises the main modules of the tool, which are described in the next subsections.

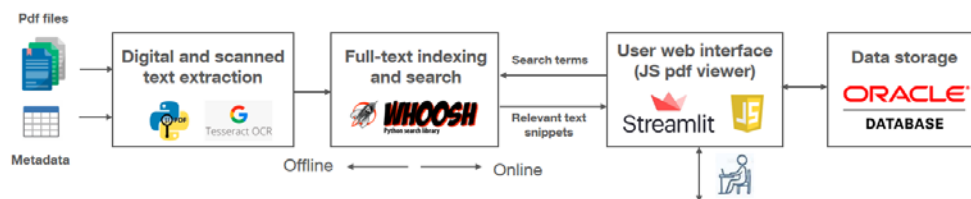


Figure 2 Main modules of the sustainability information retrieval and storage tool.

The input data to the system consists of a collection of pdf documents, the related metadata and the pre-defined taxonomy of search terms.

The offline processing part implements pre-processing, text extraction and indexing, and generates a search index for each company as a result. Then, search and data storage are performed online by the end user through the interactive user web interface.

The user interface is a web application that allows end users to interact with all the tool functionalities in order to extract relevant information from a large collection of documents. Figure 3 shows the main view of the user web interface, which incorporates multiple filters and selectors, advanced search and storage capabilities and a control panel to track data ingestion.

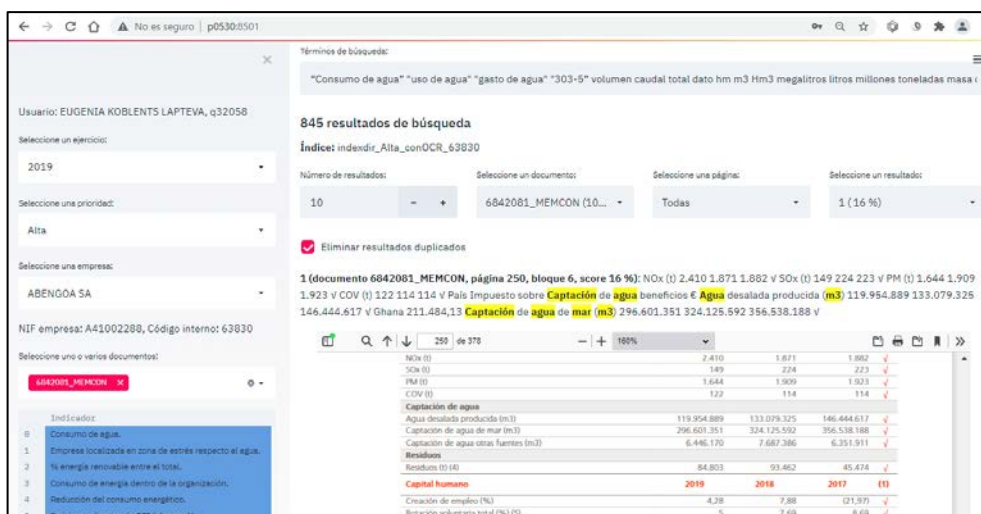


Figure 3 Main view of the user web interface.

Once all documents have been pre-processed and indexed in an offline processing phase, users can perform online searches and storage using the interactive web interface. Users log into the web application using their personal credentials. They then select a year, a company name and a sustainability indicator and the tool automatically loads all the information related to those selections. In particular, an index associated to the company selected and a pre-defined list of search terms are loaded. The tool then searches for all the pre-defined search terms in the corresponding index and retrieves a list of relevant text snippets sorted by a relevance score. Users can filter the results and select that which contains the indicator of interest. Lastly, they fill in a form with the information related to that indicator that needs to be stored. Additionally, the tool automatically stores complete context information on the selected search result (location in the document, surrounding text, etc.). All the modules involved in this process are described in more detail in the next subsections.

Input data description

The main source of input data for the tool developed is a collection of pdf documents reported by Spanish non-financial companies. Highly heterogeneous documents of variable length are available for each company. These documents contain text both in digital format, which is readily extractable, and text contained in (often low-quality) scanned images, which usually contains errors due to the Optical Character Recognition (OCR) process. Figure 4 shows an example of text fragments in digital format (left) and text contained in a scanned image (right). Over one thousand documents (6GB) have been processed to date. Only documents in the Spanish language have been considered so far; multilingual processing has not been addressed.

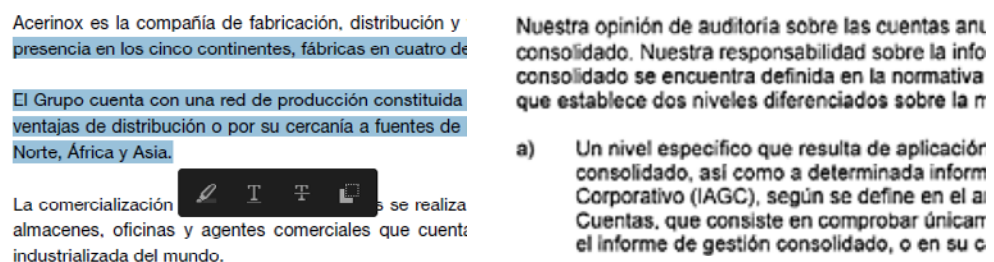


Figure 4 Examples of text in digital format (left) and text contained in scanned images (right).

The metadata related to the collection of documents and reporting companies are also required for the tool to operate. In particular, the metadata matrix contains a unique document identifier, type and date, as well as a company name and identifier, among other additional information. The metadata matrix is updated every time a new document is indexed and is currently stored on file together with the corresponding documents.

Using the input data sources described, the text extraction and indexing modules generate an index search for each company, which is also stored on file with a unique identifier. Lastly, a taxonomy of search terms needs to be defined for each sustainability indicator of interest; it is stored in the database and is fed into the search module.

Digital and scanned text extraction

This module extracts the textual content of the collection of documents, which will later be indexed by the indexing module. The module inputs are the collection of documents and the metadata file. Documents usually contain both text in digital format and text contained in scanned images, two text data sources that need to be processed in different ways.

Text in digital format can be readily extracted from documents and usually contains no errors. In this work, the Python library `pymupdf` was used to extract digital text from documents. This library enables extraction of text in multiple formats, including plain text, blocks, words, json, xml, and other formats. JSON format was preferred since it contains rich format information (text location, font and font size, etc.) additionally to the text content itself. Figure 5 shows an example of digital text extraction with `pymupdf`, where the output JSON structure contains rich information describing the extracted text, including the text bounding box coordinates, font and font size, colour, etc.

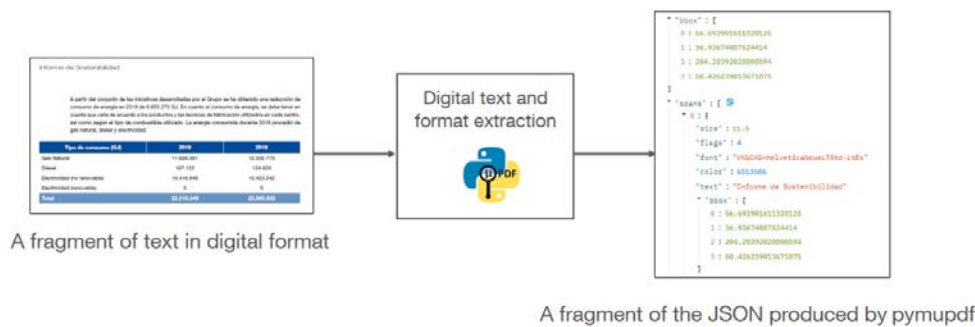


Figure 5 Digital text extraction with `pymupdf`.

Images, possibly containing scanned text, appear as base64 encoded strings in the JSON structure provided by `pymupdf`. These strings need to be decoded and converted into PIL images before being processed with an OCR system. In this work, the Tesseract OCR engine, implemented in the Python library `pytesseract`, was applied to all images in order to extract scanned text present in the documents. Figure 6 shows the scanned text extraction workflow process.

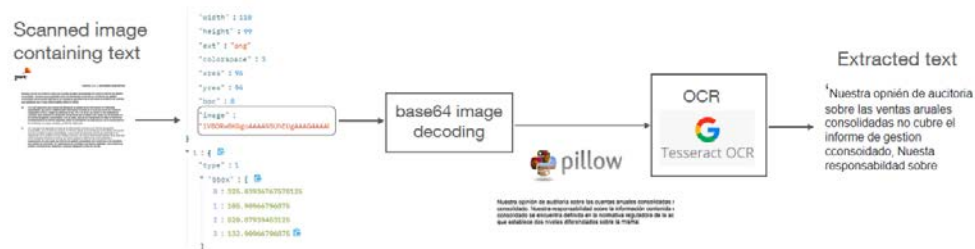


Figure 6 Scanned text extraction with `pymupdf`, base64 image decoding and Tesseract OCR.

The quality of scanned images is often low, yielding OCR errors that can sometimes be corrected using available error correction tools. In particular, the `pyspellchecker` and `hunspell` libraries are available in Python. The indexing and search engine implemented in the Python library `whoosh` implements fuzzy search, which may also be used as an alternative technique to account for errors in the extracted text. However, in this work, no error correction tools were used, since they can

potentially induce errors in correct but non-standard terms, such as company names and other words.

Both text extraction and indexing modules are executed offline every time a new document becomes available. They are not accessible from the user web interface.

Full-text indexing and search

The indexing and search module has been developed using the Python library whoosh, which provides a fast, full-text indexing and search engine implementation in pure Python, suitable for moderate-sized document databases. Full-text search engines allow for quick searches of high volumes of text for a custom textual query, as opposed to metadata-based or exact-match search engines. Full-text search systems rely on an inverted index that indicates in which fragments each term appears.

In this work, digital and scanned text was indexed in blocks of more than 50 words, sorted by y-coordinate. An index schema and a language analyser including a tokenizer, a stemmer, a stop-word filter, etc., had to be defined. For scalability and robustness reasons, an index was created for each company. This indexing process is performed offline by the tool developers.

Once the textual content has been extracted from the documents and indexed, end users can search for (sustainability) information in real-time using the user web interface and the search indices generated.

To standardise the search terminology, a taxonomy of search terms was pre-defined for each sustainability indicator of interest based on expert knowledge. This taxonomy is stored in an Oracle database and is accessible from the user web interface. When the user selects a sustainability indicator of interest in the user interface, the pre-defined list of search terms is automatically loaded and a search query is built based on those terms. By default, OR operators have been used for the query creation, but other logical operators can be used to build custom search queries.

Figure 7 shows the section of the user interface dedicated to the search of the sustainability indicators. The first selector allows the user to choose a sustainability indicator from a pre-defined list and a short description is loaded beside it in an expandable box. The list of pre-defined search terms is automatically loaded into the search bar. The user can also modify the search terms in real time by typing them into the search bar using the tool as a standard full-text search engine. The pre-defined search query can be restored by clicking on the corresponding button.

The tool automatically searches the index associated with the selected company for the terms listed in the search bar. The total number of search results and the index name are shown below the search bar. The tool allows the user to select the number of search results to be shown and filter them by document and page. The filtered list of the most relevant text snippets is shown below, sorted by the scoring metric computed by the whoosh library. In particular, the BM-25 scoring algorithm was used in this work. Each search result is listed together with the document name, page, text block and relevance score.

The tool automatically eliminates duplicated search results, when multiple copies of some fragments of text appear within the documents or when multiple OCR systems have been applied to the images contained in the documents.

Búsqueda de indicadores

Seleccione un indicador para búsqueda:

Consumo de agua

Restablecer términos de búsqueda

Términos de búsqueda:

"Consumo de agua" "uso de agua" "gasto de agua" "303-5" volumen caudal total dato hm m3 Hm3 megalitros litros millones toneladas masa (

Descripción del Indicador

Índice: [indexdir](#) Alta conOCR 63830

Seleccione un resultado:

Todos

1 (documento 6842081_MEMCON, página 250, bloque 6, score 16 %): NOx (t) 2.410.1871.882 v SOx (t) 149.224.223 v PM (t) 1.644.1909.1.923 v COV (t) 122.114.114 v País Impuesto sobre **Captación de agua** beneficios F Agua desalada producida (m3) 119.954.089 133.079.325 146.444.617 v Ghana 211.484,13 **Captación de agua de mar** (m3) 296.601.351 324.125.592 356.538.188 v

2 (documento 6842081_MEMCON, página 250, bloque 7, score 15 %): **Captación de agua** otras fuentes (m3) 6.446.170 7.687.386 6.351.91 v India (191.542,4) Residuos Israel - Residuos (t) 4) 84.801.93.462 45.474 v Luxemburgo 535,00 Capital humano 2019 2018 2017 (1) Marruecos 1.575.091,67 Creación de empleo (%) 4.28 7.88 (21,97) v México 1.344.343,56 Rotación voluntaria **total** (%) (5) 5 7,69 8,69 v



Data storage

Second, a relational database provides for simultaneous read-write access for a high number of users, which was an important requirement for implementation. In particular, an Oracle database was used in this project because it is the standard software supported by the Banco de España's IT team for this type of applications. Alternative solutions have not been explored. The storage solution implemented based on an Oracle relational database has yielded very good performance in terms of speed, simultaneous access, data protection and the possibility to recover past versions of the database. etc.

Creation of a structured sustainability database from company reports

Other functionalities satisfied by the application are a control panel to track data insertion and detect which companies are pending completion by each worker, and other filters and selectors to navigate within the range of documents and companies. Also, the web app allows users to delete records directly, instead of having to write SQL sentences.

A deeper look into the previously mentioned modules in the app can be seen in Figure 10 (authentication), Figure 11 (search), Figure 12 (pdf viewer) and Figure 13 (storage).

Figure 10 Users have to introduce their credentials to log into the app.

Figure 11 For each indicator, users can customise the search terms. They can then navigate through the search results and choose the correct one.

Figure 12 The pdf navigator is important to verify at source that the result found in the previous step is correct.

Figure 13 All the information is sent to the database in a fixed and structured format.

3. First data ingestion

Approximately 20 users worked part-time during two months in 2021 to make the first data ingestion. The 39 indicators were required to be filled in for 139 corporate groups. A total of 515 documents were available for these groups (making an average of approximately 2.9 documents per entity). Some 10,500 records were stored in this first data ingestion and 4,500 more in a second batch, making approximately 4,500 rows.

In 73% of cases the indicator was found for the given company, compared with 27% where it was not found. Listed groups provided more information than unlisted ones: 81% of the records belonging to listed companies were found. Social indicators were found for 80% of the data searched, whereas only 66% of the environmental indicators were found. 75% of government data was found. The documents were from 2019 and 2020, although some contain information from earlier years. This explains why historical series have been achieved for some indicators.

Some reports have been made and shown to the public. These reports are mainly dashboards made by PowerBI, as can be seen in Figure 14.

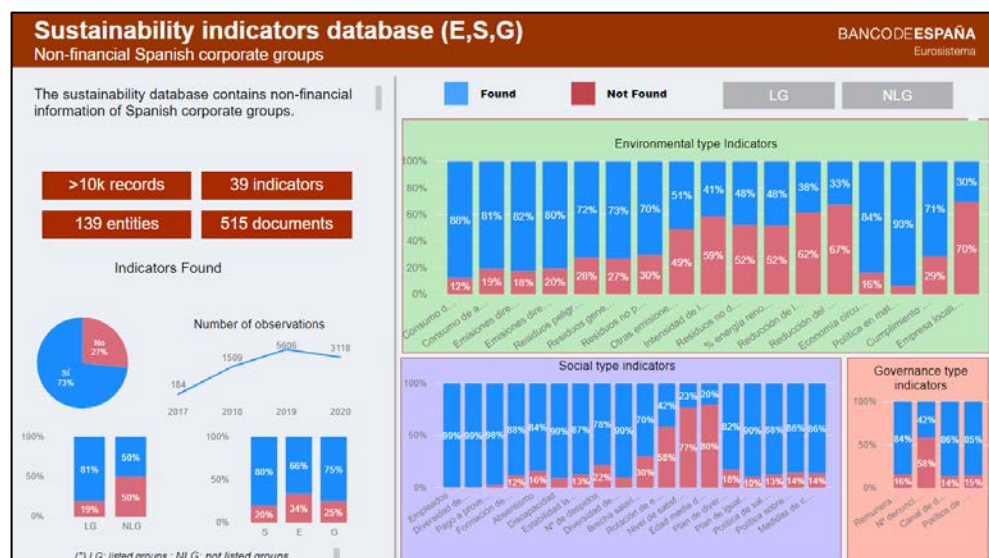


Figure 14 Several dashboards have been made to show the effectiveness of the product.

4. Conclusions

A web application has been developed and deployed for the creation of a new database of sustainability information from large collections of documents (Spanish corporate reports). This information has been obtained using a semi-automated approach that reduces the human effort required thanks to the search tools included.

The prototype was implemented by two data scientists in the Banco de España's Statistics Department, with the support of domain experts for requirements definition, testing, etc. The prototype is planned to be productionised with the help of the IT Department.

The tool incorporates authentication, data filters and selectors, full-text search in multiple documents, data storage, deletion and download and a control panel to track data insertion. The app allows the operators to customise the search terms (although default terms are given for each indicator).

Next steps

The project for storing companies' sustainability information is a long-term one as new regulations are arising and will arise in different legal frameworks. This means it is an ongoing project which will have to adapt to new documents, regulations, etc.

The work performed so far is a complete prototype prepared to be taken to production. Real data has already been introduced using the full life-cycle of the web app. Nevertheless, some lines of work are open for the short, medium and long term:

- **Database extension:** Currently, the database has been populated with information retrieved from documents of listed companies and consolidated groups. The plan is to expand this to other enterprises, and also to include financial entities.
- **Improvement of ontology using NLP:** Together with the above-mentioned 10,500 records, additional context information, such as the search terms, the location of the selected search result in the document, and its textual content has also been stored. NLP techniques could be used to automatically optimize the set of search terms for each indicator based on that context information.
- **Database publication within the Banco de España:** The database is accessible to the CBSO operators, but it has not been made available to Banco de España staff in general or to the public. The plan is for it to be released within the Banco de España environment and to external researchers within the Banco de España's data laboratory (BELab [9]) in the short term.
- **Migration to dash:** Dash is the most common Python tool for visualisation and is the official software for web development in Python at the Banco de España. For this reason, the current Streamlit prototype will be migrated to Dash in the near future.
- **Migration to production servers:** The app has been deployed on local servers in the Statistics Department but it will soon be migrated to official production servers.

- **Exploration of alternative OCR systems:** The IT Department is currently exploring the Ulpath system, which provides different OCR engines (Tesseract, Omnipage, etc.). Currently, the library pytesseract is the package used in this prototype to extract text from image formats, but Ulpath might be used in the future.

References

- [1] Bingler, Julia Anna, Mathias Kraus and Markus Leippold. "Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures." Available at SSRN (2021).
- [2] Friederich, David et al. "Automated Identification of Climate Risk Disclosures in Annual Corporate Reports." arXiv preprint arXiv:2108.01415 (2021).
- [3] Luccioni, Alexandra and Hector Palacios. "Using Natural Language Processing to Analyze Financial Climate Disclosures." *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California*. 2019.
- [4] Luccioni, Alexandra, Emily Baylor and Nicolas Duchene. "Analyzing Sustainability Reports Using Natural Language Processing." arXiv preprint arXiv:2011.08073 (2020).
- [5] Moreno, Ángel Iván and Teresa Caminero. "Application of Text Mining to the Analysis of Climate-Related Disclosures." (2020).
- [6] Moreno, Ángel Iván and Teresa Caminero. "Analysis of ESG Disclosures in Pillar 3 Reports. A Text Mining Approach". *Future Finance* (forthcoming 2022).
- [7] Tarquinio, Lara, Domenico Raucci and Roberto Benedetti. "An Investigation of Global Reporting Initiative Performance Indicators in Corporate Sustainability Reports: Greek, Italian and Spanish Evidence." *Sustainability*, 2018.
- [8] Webersinke, Nicolas et al. "ClimateBert: A Pretrained Language Model for Climate-Related Text." arXiv preprint arXiv:2110.12010 (2021).
- [9] <https://www.bde.es/bde/en/areas/analisis-economi/otros/que-es-belab/>.

A WEB APPLICATION PROTOTYPE TO RETRIEVE AND STORE SUSTAINABILITY INFORMATION FROM UNSTRUCTURED TEXT

Alejandro Morales (alejandro.morales@bde.es)
Eugenia Koblents (eugenia.koblents@bde.es)

IFC-BANK OF ITALY WORKSHOP ON DATA SCIENCE IN CENTRAL BANKING
PART 2: DATA SCIENCE IN CENTRAL BANKING: APPLICATIONS AND TOOLS

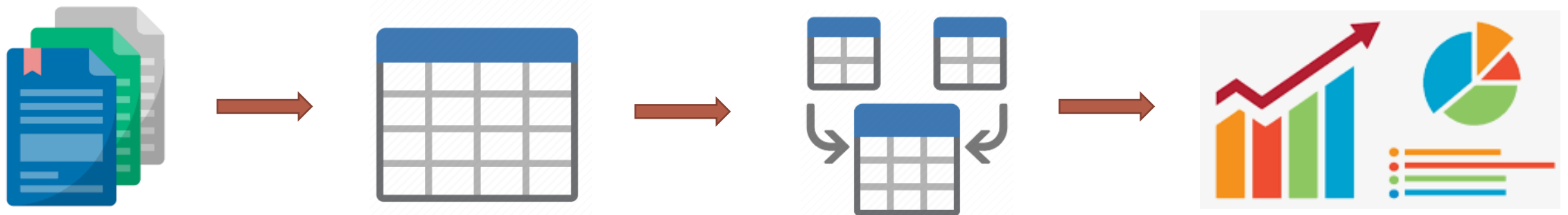
17/02/2022

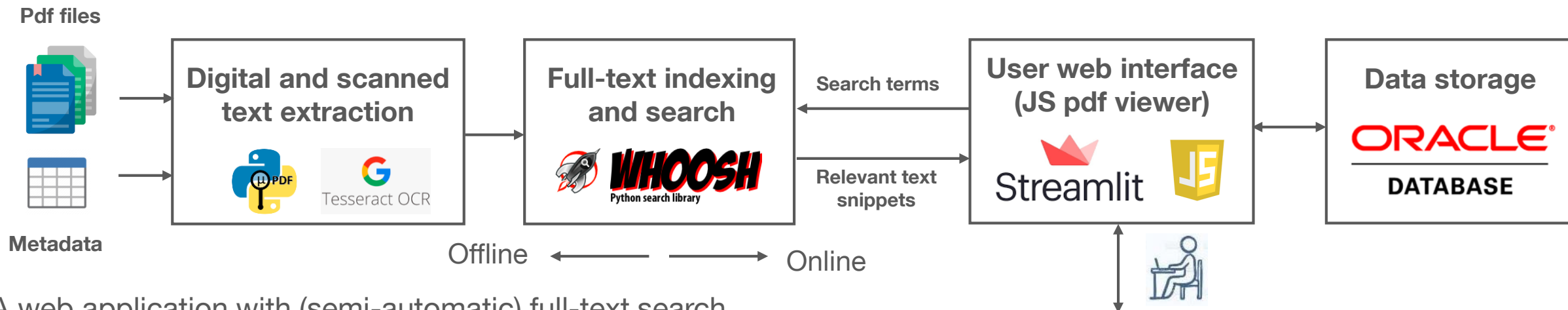


INDEX

1. Introduction
2. Main modules:
 - Text extraction
 - Indexing and search
 - Data storage
 - User interface demo
3. First data ingestion
4. Conclusions and next steps

- ❑ The goal of this project is the creation of a new database containing microdata information on sustainability extracted from **Spanish companies** reports in order to address **climate change**.
- ❑ The target information includes **39 environmental** (energy, water, emissions, residues, etc), **social** (employees, age, gender, etc) and **governance** (corruption, bribery, complaints, etc) **indicators**.
- ❑ To date, this information is only mandatory for **large corporate groups** but EU regulation is adapting.
- ❑ The reported information is presented in **highly heterogeneous formats**: plain text, tables, graphics and images. A broad variety of **metrics** are used for each indicator. **Manual extraction** is very costly.
- ❑ **Technical goal**: transform **non-structured information** (documents) into a **structured database** (table) to **merge** it with additional information and generate **statistics** on sustainability.
- ❑ **Project time span**: April-July 2021. Next phase planned for first semester of 2022.





A web application with (semi-automatic) full-text search and storage capabilities has been developed in **Python**:

1. The user selects a year, company and indicator.
2. The **tool searches** for pre-defined terms and retrieves a sorted list of relevant text snippets.
3. The **user validates** the search results and **stores** the indicator value into the database.
4. Complete **context information** is also stored (location, surrounding text, etc.).

Streamlit tool temporarily deployed on a local machine, not supported by IT. Migration to **dash** and final **deployment** on a corporate web server planned for 2022.

The screenshot shows the web application interface with search results and a data table.

Search Results:

- Términos de búsqueda:** "Consumo de agua" "uso de agua" "gasto de agua" "303-5" volumen caudal total dato hm m3 Hm3 megalitros litros millones toneladas masa
- 845 resultados de búsqueda**
- Índice:** indexdir_Alta_conOCR_63830
- Número de resultados:** 10
- Selección de documento:** 6842081_MEMCON (10...)
- Selección de página:** Todas
- Selección de resultado:** 1 (16 %)
- ☒ Eliminar resultados duplicados

Document Details:

1 (documento 6842081_MEMCON, página 250, bloque 6, score 16 %): NOx (t) 2.410 1.871 1.882 v SOx (t) 149 224 223 v PM (t) 1.644 1.909 1.923 v COV (t) 122 114 114 v País Impuesto sobre **Captación de agua** beneficios 6 **Agua** desalada producida (m3) 119.954.889 133.079.325 146.444.617 v Ghana 211.484,13 **Captación de agua de mar** (m3) 296.601.351 324.125.592 356.538.188 v

Data Table:

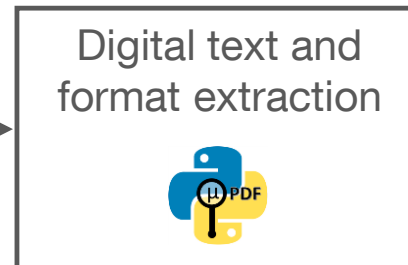
| | 2019 | 2018 | 2017 | (t) |
|--------------------------------------|-------------|-------------|-------------|-----|
| Captación de agua | | | | |
| Agua desalada producida (m3) | 119.954.889 | 133.079.325 | 146.444.617 | ✓ |
| Captación de agua de mar (m3) | 296.601.351 | 324.125.592 | 356.538.188 | ✓ |
| Captación de agua otras fuentes (m3) | 6.446.170 | 7.687.386 | 6.351.911 | ✓ |
| Residuos | | | | |
| Residuos (t) (4) | 84.803 | 93.462 | 45.474 | ✓ |
| Capital humano | | | | |
| Creación de empleo (%) | 4,28 | 7,88 | (21,97) | ✓ |
| Rotación voluntaria total (%) (5) | 5 | 7,69 | 8,69 | ✓ |

- ❑ **Input data:** Highly heterogeneous documents of variable length are available for each company, including text in **digital format** (readily extractable) and text contained in **scanned images** (requiring OCR).
- ❑ Text in **digital format** is readily extractable and (usually) contains **no errors**. The **pymupdf** library allows to extract digital text in **JSON** format, containing **rich format information** (text location, font, size, etc).

A partir del conjunto de las iniciativas desarrolladas por el Grupo se ha obtenido una reducción de consumo de energía en 2019 de 9.655.275 GJ. En cuanto al consumo de energía, se debe tener en cuenta que varía de acuerdo a los productos y las técnicas de fabricación utilizados en cada centro, así como según el tipo de combustible utilizado. La energía consumida durante 2019 procedió de gas natural, diésel y electricidad.

| Tipo de consumo (GJ) | 2019 | 2018 |
|-----------------------------|-------------------|-------------------|
| Gas Natural | 11.626.381 | 12.332.770 |
| Diésel | 167.122 | 124.620 |
| Electricidad (no renovable) | 10.416.846 | 10.423.542 |
| Electricidad (renovable) | 0 | 0 |
| Total | 22.210.349 | 22.880.932 |

A fragment of text in digital format



```

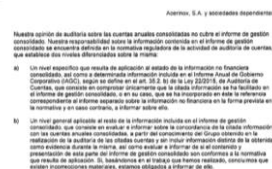
"spans": [
  0: {
    "size": 11.5
    "flags": 4
    "font": "VKGCXG+HelveticaNeueLTStd-LtEx"
    "color": 6513506
    "text": "Informe de Sostenibilidad"
    "bbox": [
      0: 56.692901611328125
      1: 36.93674087524414
    ]
  }
]

```

A fragment of the pymupdf JSON

- ❑ **(Scanned) images** appear as base64 encoded strings in the pymupdf JSON and need to be decoded and converted into PIL images to be fed into pytesseract OCR. The quality of scanned images is often low, yielding OCR errors.

Scanned text

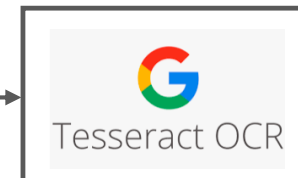


```

"xres": 96
"yres": 96
"bpc": 8
"image": "iVBORw0KGgoAAAANSUuEUGAAAG4AAAI"
}

```

base64 image decoding



Extracted text

‘Nuestra opinión de auditoría sobre las ventas aruales consolidadas no cubre el informe de gestión

- ❑ Digital and scanned text is indexed (offline) in **blocks of >50 words**. An **index** per company is created.
- ❑ A **language analyser** (tokenizer, stemmer, stop-word filter, etc) and an **index schema** need to be defined.
- ❑ A set of **search terms** has been defined for each indicator based on **expert knowledge**.
- ❑ The tool retrieves **relevant text snippets** in real time.
- ❑ The user can **filter results** by document and page.
- ❑ When a search result is selected, the corresponding text snippet is shown on a JavaScript **pdf viewer**.
- ❑ The tool can also be used as a standard **full-text search engine** by manually typing the search query.



Búsqueda de indicadores

Seleccione un indicador para búsqueda:

Consumo de agua

Restablecer términos de búsqueda

Términos de búsqueda:

"Consumo de agua" "uso de agua" "gasto de agua" "303-5" volumen caudal total dato hm m3 Hm3 megalitros litros millones toneladas masa (

845 resultados de búsqueda

Índice: indexdir_Alta_conOCR_63830

Número de resultados:

10

- +

Seleccione un documento:

6842081_MEMCON (10...

Seleccione una página:

Todas

Seleccione un resultado:

Todos

☒ Eliminar resultados duplicados

1 (documento 6842081_MEMCON, página 250, bloque 6, score 16 %): NOx (t) 2.410 1.871 1.882 v SOx (t) 149 224 223 v PM (t) 1.644 1.909 1.923 v COV (t) 122 114 114 v País Impuesto sobre **Captación de agua** beneficios € **Agua** desalada producida (m3) 119.954.889 133.079.325 146.444.617 v Ghana 211.484,13 **Captación de agua de mar** (m3) 296.601.351 324.125.592 356.538.188 v

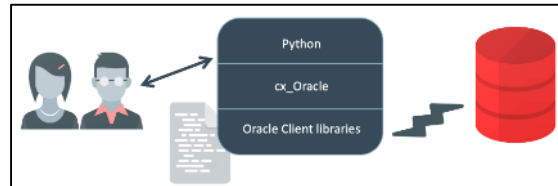
2 (documento 6842081_MEMCON, página 250, bloque 7, score 15 %): **Captación de agua** otras **fuentes** (m3) 6.446.170 7.687.386 6.351.911 v India (191.542,34) Residuos Israel - Residuos (t) (4) 84.803 93.462 45.474 v Luxemburgo 535,00 Capital humano 2019 2018 2017 (1) Marruecos 1.575.091,67 Creación de empleo (%) 4,28 7,88 (21,97) v México 1.344.343,56 Rotación voluntaria **total** (%) (5) 5 7,69 8,69 v

siendo el desglose por país.

| País | Impuesto sobre beneficios € | País | Impuesto sobre beneficios € |
|--------------|-----------------------------|--------------|-----------------------------|
| Arabia Saudí | - | Ghana | 211.484,13 |
| Argelia | 2.775.929,77 | India | (191.542,34) |
| Argentina | 854.558,40 | Israel | - |
| Belgica | - | Luxemburgo | 535,00 |
| Brasil | 92.363,54 | Marruecos | 1.575.091,67 |
| Chile | 181.612,42 | México | 1.344.343,56 |
| China | 164.774,57 | Países Bajos | 30.279,68 |
| Colombia | 121.352,68 | Perú | 1.708.222,37 |

| Indicador | 2019 | 2018 | 2017 | (1) |
|--------------------------------------|-------------|-------------|-------------|-----|
| NOx (t) | 2.410 | 1.871 | 1.882 | ✓ |
| SOx (t) | 149 | 224 | 223 | ✓ |
| PM (t) | 1.644 | 1.909 | 1.923 | ✓ |
| COV (t) | 122 | 114 | 114 | ✓ |
| Captación de agua | | | | |
| Agua desalada producida (m3) | 119.954.889 | 133.079.325 | 146.444.617 | ✓ |
| Captación de agua de mar (m3) | 296.601.351 | 324.125.592 | 356.538.188 | ✓ |
| Captación de agua otras fuentes (m3) | 6.446.170 | 7.687.386 | 6.351.911 | ✓ |
| Rotación | | | | |
| Rotación (t) (4) | 84.803 | 93.462 | 45.474 | ✓ |
| Capital humano | | | | |
| Creación de empleo (%) | 4,28 | 7,88 | 21,97 | ✓ |
| Rotación voluntaria total (%) (5) | 5 | 7,69 | 8,69 | ✓ |
| Mujeres en plantilla | | | | |
| En puestos directivos (%) | 11,34 | 11,52 | 10,04 | ✓ |

- ❑ The tool allows the user to **store** the value of the indicator in an **Oracle database**, together with additional information (metric, year, comments, etc). Connector: library **cx_Oracle** in python.
- ❑ Rich **context information** is also stored automatically (search terms, text snippets, user ID, date and time, etc).
- ❑ A **control panel** is automatically updated when new data is inserted into the database.
- ❑ The user interface allows to **download** the full as well as a filtered version of the dataset.
- ❑ More than 20 workers were required to work in this project and fill data into the database, good **simultaneous performance** has been a key requirement.
- ❑ The **authentication** in the database is used as registration in the interface.
- ❑ Backup and change **traceability** stored. Other intermediate tables are also kept in the database.



- The user web interface has been temporarily implemented in streamlit and deployed on a local machine.

- Streamlit** is an easy to use open-source app framework for data science visualization and web development.

- The tool incorporates integrated authentication, multiple filters and selectors, **full-text search**, **data storage** and a control panel to track the insertion of data.

- The user interface will soon be migrated to **dash** and deployed on a corporate **web server**.

Authentication

Introduzca su q de usuario: Ejemplo q32057

Introduzca su password a la bbdd

Search

Búsqueda de indicadores

Seleccione un indicador para búsqueda: **Consumo de agua** Descripción del indicador: +

Restablecer términos de búsqueda

Términos de búsqueda: "Consumo de agua" "uso de agua" "gasto de agua" "303-5" volumen caudal total dato hm m3 Hm3 megalitros litros millones toneladas masa

845 resultados de búsqueda
Índice: indexdir_Alta_conOCR_63830

Número de resultados: 10 Seleccione un documento: 6842081_MEMCON (10... Seleccione una página: Todas Seleccione un resultado: Todos

Eliminar resultados duplicados

1 (documento 6842081_MEMCON, página 250, bloque 6, score 16 %): NOx (t) 2.410 1.871 1.882 v SOx (t) 149 224 223 v PM (t) 1.644 1.909 1.923 v COV (t) 122 114 114 v País Impuesto sobre **Captación de agua** beneficios € **Agua** desalada producida (m3) 119.954.889 133.079.325 146.444.617 v Ghana 211.484,13 **Captación de agua de mar** (m3) 296.601.351 324.125.592 356.538.188 v

2 (documento 6842081_MEMCON, página 250, bloque 7, score 15 %): **Captación de agua** otras fuentes (m3) 6.446.170 7.687.386 6.351.911 v India (191.542,34) Residuos Israel - Residuos (t) (4) 84.803 93.462 45.474 v Luxemburgo 535,00 Capital humano 2019 2018 2017 (1) Marruecos 1.575.091,67 Creación de empleo (%) 4,28 7,88 (21,97) v México 1.344.343,56 Rotación voluntaria **total** (%) (5) **5** 7,69 8,69 v

Document check

250 de 378

señal el dígito por país

| País | Indicador | Valor | País | Indicador | Valor |
|--------------|--------------|----------------------|--------------|-----------|-------|
| Andorra | 2.175.000,00 | Chile | 211.484,13 | | |
| Argelia | 898.100,00 | India | 191.542,34 | | |
| Arabia Saudí | 1.644.190,9 | Israel | 84.803 | | |
| Austria | 1.871.188,2 | Italia | 1.344.343,56 | | |
| Bélgica | 1.909.192,3 | Países Bajos | 535,00 | | |
| Bulgaria | 1.882.149,2 | Perú | 1.575.091,67 | | |
| Canadá | 1.923.122,1 | Reino Unido | 1.344.343,56 | | |
| Chad | 1.882.149,2 | República Dominicana | 1.344.343,56 | | |
| China | 1.882.149,2 | República Dominicana | 1.344.343,56 | | |
| Colombia | 1.882.149,2 | República Dominicana | 1.344.343,56 | | |

Storage

Seleccione uno o varios documentos:

6842081_MEMCON **6842081_MEMCON** **6842081_MEMCON**

Almacenamiento de resultados

Seleccione un indicador para almacenamiento: **Consumo de agua**

Seleccione si ha encontrado el dato: **Encontrado** Tipo de Dato: **Consolidado (grupo)** Seleccione un año: **2019** Rellene el país: **GRUPO**

Rellene el valor del indicador: **Hm3** Seleccione la métrica: **Hm3**

Recuerde que los puntos denotan miles y las comas decimales

Añada un comentario opcional:

Actualizar Modificación

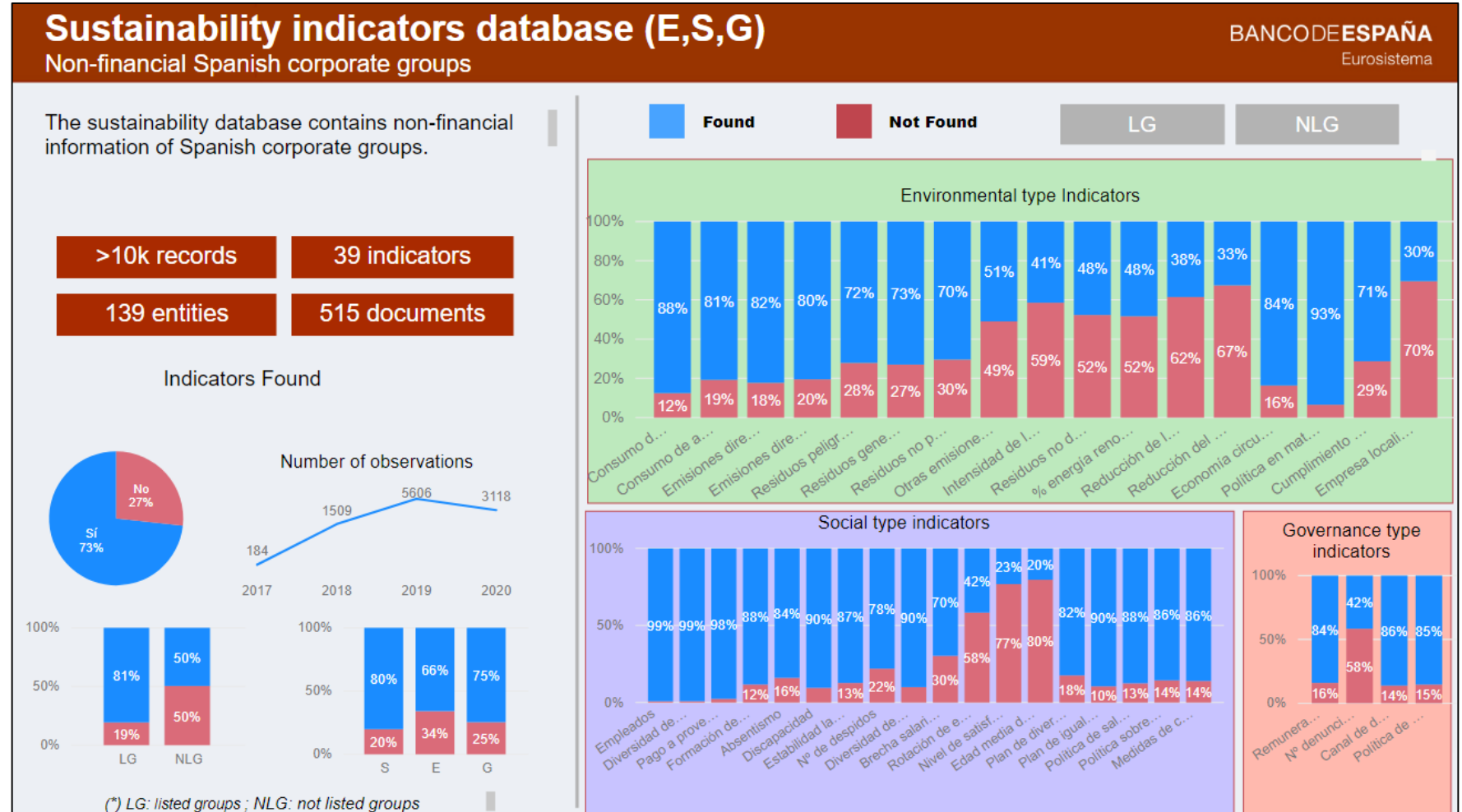


❑ 20 users were working part-time during 2 months to make the first data ingestion:

- > 10,000 records
- 39 indicators
- 139 companies
- 515 documents

❑ Listed companies provided more percentage of successfully found indicators.

❑ Historical series have been obtained for some indicators.



Conclusions:

- ❑ A **web application** has been developed and deployed for creating a new database of sustainability information from large collections of documents (Spanish corporate groups reports).
- ❑ The tool incorporates authentication, **full-text search**, **data storage** and a control panel to track the insertion of data. Prototype implemented in 3 months by two **data scientists** with the support of **domain experts** for requirements definition, testing, etc.
- ❑ The current prototype is **not supported by the IT Department** but has been implemented and deployed using temporary tools.
- ❑ The tool allows to **standardize the search terms** (each user makes the same searches by default).
- ❑ **Context information stored** (search terms, search results, etc.) will allow to train ML methods to optimize the search process.
- ❑ A **semi-automatic approach** reduces the human effort in populating the new database.

Next steps:

- ❑ **Database extension** to include other enterprises and financial entities: Currently, only listed companies and consolidated groups.
- ❑ **Improvement of ontology using NLP and machine learning.**
- ❑ **Database publication within the Bank of Spain:** The database is only accessible by the operators of the CBSO at the moment.
- ❑ **Migration to dash:** dash is the most common Python tool for visualization and is the official software for web development in Python in the Bank of Spain. The user interface will soon be migrated to **dash** enhancing its functionality.
- ❑ **Migration to corporate production servers:** So far the app has been deployed on local machines in the Statistics Department.
- ❑ Exploration of **alternative OCR systems.**

Thank you for your attention!



IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Measuring text-based sentiments from monetary policy statements – a Malaysian case study using natural language processing¹

Eilyn Chong and Sui-Jade Ho,
Central Bank of Malaysia

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Measuring Text-Based Sentiments from Monetary Policy Statements

A Malaysian Case Study using Natural Language Processing

Eilyn Chong and Sui-Jade Ho¹

Abstract

Central banks publish monetary policy statements (MPS) to provide insights into economic development and outlook, as well as to communicate judgment on the balance of risks and expectations on the future course of monetary policy. Using automated content analysis to extract the sentiment from each MPS published between August 2004 to September 2020 by the Central Bank of Malaysia, we analyse the relationship between the sentiment measures from these statements and financial market movements. These sentiment measures are derived from three dictionaries, including one specially developed in this paper for monetary policy analysis. We find the sentiment measures move in line with changes in the policy rate, the Overnight Policy Rate (OPR). Furthermore, using a high-frequency event-study methodology, we find evidence of an asymmetric impact of the sentiments on sovereign (Malaysian Government Securities, MGS) yields at higher maturity and interest rate swap rates. In particular, negative or dovish words are stronger predictors of a decline in yields and swap rates beyond the actual change in the OPR. Conversely, the relationship between positive or hawkish words and financial market movements is less evident. Our findings provide some evidence that the wording in the Central Bank of Malaysia's MPS is informative for the financial markets, especially during stress periods.

Keywords: Central bank communication, text analysis, dictionary, sentiment

JEL classification: D83, E52, E58, G14

¹ This paper was prepared for the IFC and Bank of Italy Workshop on "Data Science in Central Banking" which was held virtually on 14-17 February 2022. Both authors are from the Central Bank of Malaysia. Emails: eilyn@bnm.gov.my and jade@bnm.gov.my respectively. The views expressed in this paper are those of the authors and should not be interpreted as reflecting the views of the Central Bank of Malaysia, the Monetary Policy Committee (MPC) or anyone else associated with the Central Bank of Malaysia. The authors would like to thank Mohamad Hasni Sha'ari, Ong Li Ming, Dr Ong Hong Hoe, Nurashiqin Asri, Dian Hikmah Mohd Ibrahim and Nur Aimi Abdul Ghani for their invaluable comments on the paper, Sabrina Bashir Ahmad, Murshidah Sarbudeen and 'Aliya' Yasmin Hanafi for their contribution to the refinement of the Monetary Policy dictionary for this study, as well as Ivan Avannus Jacob Jimbangan for proofreading this paper.

1. Introduction

This paper documents our method of extracting the sentiments in the monetary policy statements (MPS) published by the Monetary Policy Committee (MPC) of the Central Bank of Malaysia and estimating the impact of these sentiments on financial markets. We employ an automated content analysis to extract sentiment measures from the published MPS texts. We then assess whether these measures influence the sovereign (Malaysian Government Securities, MGS) yields and interest rate swap (IRS) rates above and beyond the actual change in the policy rate, OPR. The relationship between policy statements and financial market reactions is important since it could have spillovers to real economic activity, thereby underscoring the need to pay attention to how the central bank conveys information.²

This study relates to the literature on the link between sentiments derived from text analytics and financial market responses. Nyman et al. (2021) show that sentiment derived from financial market text-based data is correlated with financial market activity during periods of stress. This is also corroborated by a study by García (2013) using financial news-derived sentiment. Related to this strand of the literature are studies that explore the information content specifically in central bank communications. These include Jegadeesh & Wu (2017), who quantify the tone in the Federal Reserve meeting minutes and find a significant relationship between the topic content and financial market volatility. The study by Bligh & Hess (2013) finds that speeches by the US Federal Reserve (Fed) Chairman (Alan Greenspan), testimonies and FOMC statements contribute to the prediction of financial market variables. Moniz & de Jong (2014) predict the impact of central bank communications on investors' interest rate expectations using Bank of England's Monetary Policy Committee Minutes as their corpus.

The paper is organised as follows: we first describe our methodology in Section 2. In this regard, we make several key contributions to the literature. First, we extract and analyse the sentiments from the MPS by the Central Bank of Malaysia, thus contributing to the small body of research for emerging market economies. Second, for the automated content analysis, we build upon existing dictionaries that are oriented towards finance or the financial stability context to develop a specially designed version for monetary policy context. We especially look at phrases that connote economic relationships that could convey a different sentiment than what the individual words suggest.

Using a high-frequency event-study methodology to demonstrate the ability of these sentiment measures in explaining financial market movements is our main contribution – especially as it is applicable in other contexts where MPS text is used as an input into modelling. Section 3 discusses the results. We find that MPS-derived sentiments exhibit relationships with financial market movements within the one-day window around the release of the MPS, and that sentiments contained within MPS can impact MGS yields at higher maturity and IRS rates. The sign of the marginal impact from the derived sentiment measures on MGS yields and IRS rates is generally consistent with the theoretical predictions. In addition, we found that negative word

² See Blinder et al. (2008), for example, for a survey of the literature on central bank communication with evidence suggesting the ability of central banks' communication to move financial markets, to enhance the predictability of monetary policy decisions, and potentially to help achieve central banks' macroeconomic objectives.

counts are stronger predictors of higher maturity MGS yields and IRS rates. Section 4 concludes the paper.

2. Methodology

A. Extracting sentiment measure from Monetary Policy Statements (MPS)

Our main text corpus used in this study are Monetary Policy Statements (MPS) from August 2004 to September 2020 that are published on the website of the Central Bank of Malaysia. These statements are retrieved via web-scraping — filtering out the subject headers and the paragraphs related to the schedule of the upcoming Monetary Policy Committee (MPC) meetings that are typically released every November. We further clean the text of each MPS by enforcing lowercase and removing punctuations, hyperlinks, HTML tags, and extra white spaces.

We then analyse the cleaned MPS text using automated content analysis, where a computer programme counts the frequency of certain words appearing in a text corpus based on a pre-specified wordlist or dictionary. Automated content analysis of texts offers some advantages over manual classification as it is less labour-intensive, less reliant on subjective judgment, and more likely to detect systematic patterns that would otherwise be missed. For this method, we use three dictionaries: Two are off-the-shelf dictionaries: one was developed by Loughran & McDonald (2011) (LM hereafter) tailored specifically to finance, and the other was developed by Correa et al. (2021) (Correa thereafter), which is a refinement of the LM dictionary catered to the financial stability context. Both dictionaries have a set of positive and negative words. Words that do not fall in either category are considered neutral. We then construct another dictionary that combines both LM and Correa dictionaries (Monetary Policy dictionary, MP thereafter) and refine it to better fit the monetary policy context.

Although the LM and Correa dictionaries are designed for financial-related context, many of their words could be used to describe the economy. These include words such as: *deterioration*, *recession*, *slowdown*, *improve* and *rebound*. For the MP dictionary, we reassign positive words from LM and Correa dictionaries as ‘hawkish’ words as a hawkish monetary policy stance is consistent with an overheating economy that can be described with positive words such as *strong* and *higher* in most cases. Similarly, we reassign negative words from LM and Correa dictionaries as ‘dovish’ words in the MP dictionary to describe a weak or weakening economy that could warrant a looser monetary policy.

Nevertheless, some cross-checking was undertaken and further refinements are included in the MP dictionary to incorporate nuances in the monetary policy context. The refinements consider:

1. Additional words such as *expansion* and *upside*.
2. Economic relationships that would render words or phrases to have different connotations than originally intended. Phrases such as *low unemployment* and *diminishing slack* would have been assigned a negative tone in the LM

and Correa dictionaries due to the presence of negative words *low* and *diminishing* but are reassigned as hawkish in our MP dictionary to account for the monetary policy context, along with other positive words such as *strong growth* and *high inflation*.

3. Dropping any words that describe OPR adjustments from the count of hawkish or dovish words, as we wish to assess the sentiment of the MPS independently of policy rate adjustments. For example, when *increase* is used to describe an OPR adjustment, we do not assign the word as a positive or hawkish tone.

Any positive or negative word used to describe changes in the stress, pressure or volatility in the financial market will be reassigned as neutral as it does not have a clear bearing on the monetary policy stance.³ Finally, the polarity of a word is swapped when negation words such as *not*, *not expected to*, *unlikely to*, *no reason to* and *despite* are used to negate a positive/hawkish or negative/dovish word. Each MPS can then be represented in terms of the frequency of words of specific tone based on these dictionaries. In doing so, we measure the intensity of the positive/hawkish and negative/dovish word usage.⁴ We then take a net count of words of opposing sentiments and normalise it by the total word count of each MPS to derive the sentiment measures.

$$\begin{aligned} sentiment_t &= \frac{\sum positive\ or\ hawkish\ words_t - \sum negative\ or\ dovish\ words_t}{\sum word\ count_t} \text{ for each MPS } t \\ &= sentiment_t^{(+)} - sentiment_t^{(-)} \text{ for each MPS } t \end{aligned}$$

where $sentiment_t^{(+)} = \frac{\sum positive\ or\ hawkish\ words_t}{\sum word\ count_t}$, which refers to positive/hawkish word count (normalised by total word count)

and $sentiment_t^{(-)} = \frac{\sum negative\ or\ dovish\ words_t}{\sum word\ count_t}$, which refers to negative/dovish word count (normalised by total word count)

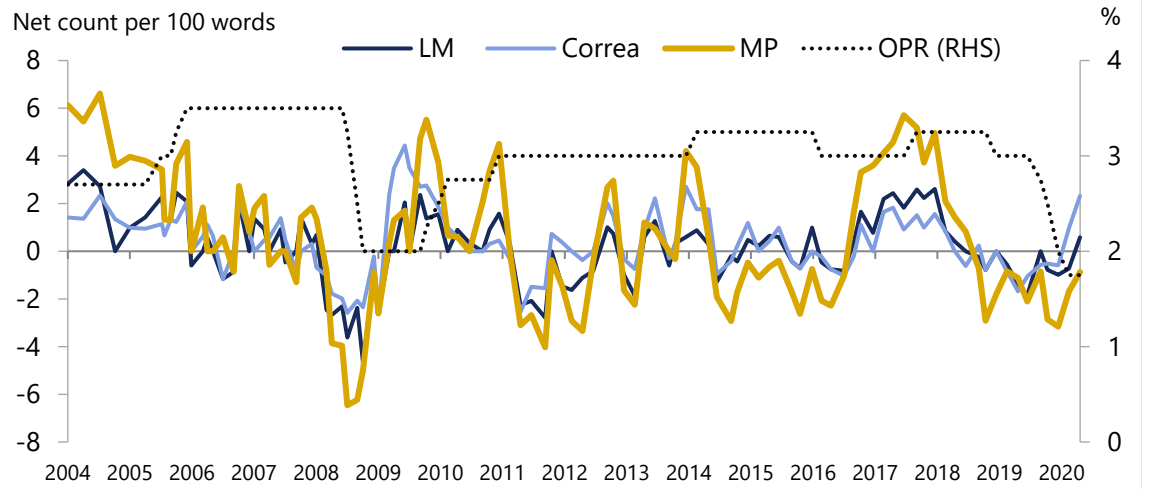
Figure 1 shows the sentiment measures produced using the three dictionaries. Of particular note is the sharp deterioration of sentiment leading up to the 2007-08 financial crisis and the recent COVID-19 pandemic, which reflect greater usage of dovish words. The decline in the derived sentiment measures moves in line with the imminent adjustments in the OPR in the sample period.

³ Instead of relying on word polarity, another approach is to train machine learning models on sentences from monetary policy statements mapped to sentiment ratings assigned by humans to better capture the specific contextual characteristics and nuances. However, this approach involves constructing a large, labelled training dataset which is resource intensive and requires familiarity with economic relationships and monetary policy jargon.

⁴ Note that our measure of intensity is derived purely from the frequency of hawkish or dovish words used in the MPS. We do not consider the intrinsic intensity of a given word, e.g. the different degree of dovishness between the words: *collapse* and *decline*.

Figure 1: Sentiment measures derived from MPS and OPR

(Positive net count indicates hawkish perceived sentiment, negative indicates dovish)



B. Relationship between sentiment measures and financial market

We employ a high-frequency event-study approach to analyse the relationship between the derived sentiment measures and the financial market. By estimating the impact of sentiments on financial market variables around a narrow window of one day around monetary policy announcement days, we measure the impact of the sentiment measures upon MPS release and limit other confounding factors that could influence asset prices.⁵

Specifically, we estimate the equation below:

$$\Delta y_t^m = \alpha + \beta_1 \text{sentiment}_t + \beta_2 \Delta \text{opr}_t + \varepsilon_t$$

where

Δy_t^m is the 24-hour window change in MGS yields of maturity with year, $m \in \{1, 2, \dots, 10\}$ or IRS rates of maturity with year, $m \in \{1, 3, 5\}$ around MPS announcement days;

sentiment_t is the sentiment measure derived from MPS using LM, Correa or MP dictionaries. All sentiment measures are normalised by the total number of words in the MPS;

Δopr_t is the change in OPR, with β_2 measuring the impact of OPR change on Δy_t^m

⁵ This approach is popular in the literature of estimating the impact of monetary policy shocks e.g. Kuttner (2001). Our approach in this paper is closely related to the literature on monetary policy shocks but with one key difference, given that our control variable is the actual daily change in the OPR and not the surprise component of the policy change. Nevertheless, as part of our robustness checks (not shown), we repeated this analysis using the daily change of the 3-month Kuala Lumpur Interbank Offered Rate (KLIBOR), which could move prior to an actual OPR change reflecting market expectations, as our control variable. The general pattern of our results remains in line.

Our variable of interest is β_1 . We assume that information available for market participants prior to the monetary policy meetings such as expectations on macroeconomic outlook have already been priced in asset valuations. According to the expectations hypothesis, the magnitude of the increase in MGS yields should be greater for MGS with longer maturity as changes in long-term rates are mainly determined by future monetary policy expectations, which would be inferred from the sentiments of the current period MPS.⁶ Figure 2 which shows the magnitude of β_1 suggests that the sentiment measures from MPS have an influence on MGS yields at longer maturity. The sign of the marginal impact is also consistent with the theoretical predictions, where a more positive (negative) net sentiment measure is associated with an increase (a decrease) in MGS yields controlling for OPR changes.

3. Results and Discussion

Relationship between sentiment measures and policy rate change

Table 1

| | Dependent variable: Δopr_t | | | | | |
|---------------------|------------------------------------|----------------------|---------------------|----------------------|---------------------|----------------------|
| | Dictionary: LM | | Correa | | MP | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| $sentiment_t$ | 4.819*** (1.240) | | 4.119*** (1.333) | | 3.112*** (0.725) | |
| $sentiment^{(+)}_t$ | | 3.368** (1.334) | | 0.553 (1.194) | | 2.204*** (0.776) |
| $sentiment^{(-)}_t$ | | -5.667*** (1.531) | | -7.999*** (2.376) | | -3.836*** (1.067) |
| Observations | 100 | 100 | 100 | 100 | 100 | 100 |
| R-squared | 0.236 | 0.245 | 0.155 | 0.243 | 0.331 | 0.340 |

¹ All regressions include a constant term. Robust standard errors in parentheses.

² Asterisks denote p-values; 1%: ***, 5%: **, 10%: *.

Sources: Authors' estimates

As a starting point, we assess the relationship between OPR adjustments and sentiment measures of the MPS derived from the three dictionaries. The positive coefficient for $sentiment_t$ confirms the observation from Figure 1, where the sentiment measures move in line with the changes in the OPR. Furthermore, when we regress the normalised count of positive words $sentiment^{(+)}_t$ and normalised count of negative words $sentiment^{(-)}_t$ separately (Columns 2, 4 and 6), $sentiment^{(-)}_t$ has a stronger association with OPR adjustments.

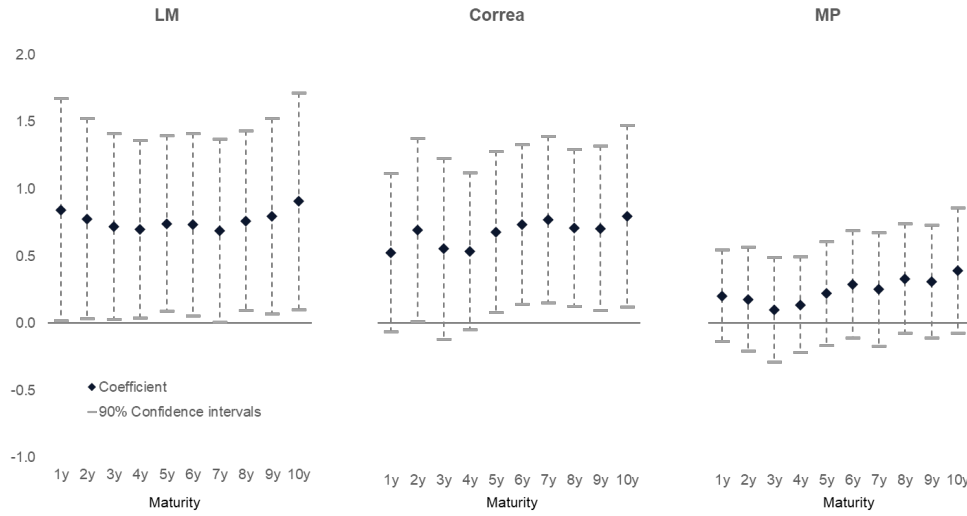
⁶ Nevertheless, there could be instances where certain market participants interpret a hawkish stance, especially a strong one, to be indicative of a potential slowdown in long-term growth. This may result in a more muted increase in longer-maturity bond yields than anticipated under the expectations hypothesis.

Figures 2 and 3 visualise the extent of the relationship between the derived sentiment measures and MGS yields and IRS rates, respectively, within a day of the MPS release. Panel B and C in both figures suggest evidence of an asymmetric impact of the sentiments on yields at higher maturity, especially for IRS rates. In particular, negative or dovish words are stronger predictors of a decline in yields and swap rates. Conversely, the relationship between positive or hawkish words and yield movements is less evident. These findings provide some evidence that the wording in the MPS is informative for the financial markets, especially during stress periods when negative or dovish words are more likely to be used in the MPS.

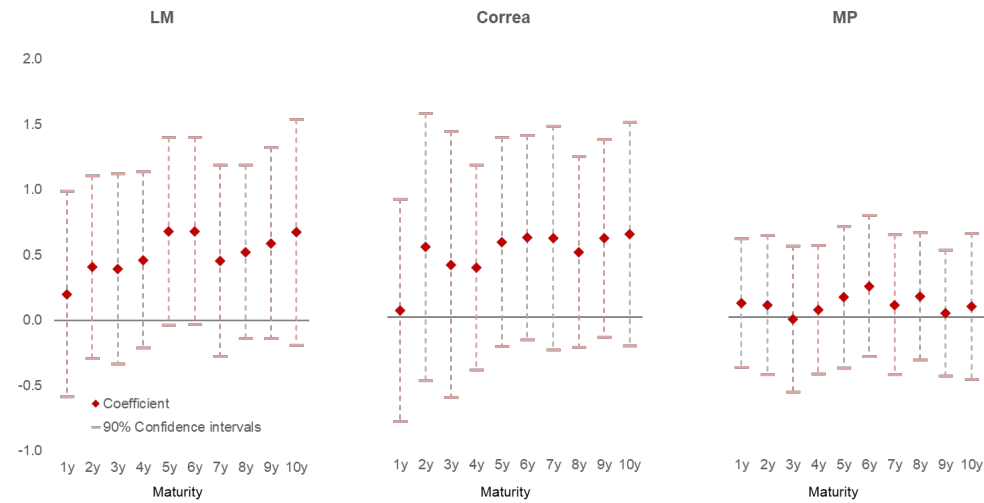
Notwithstanding these findings, one limitation of this study is that we cannot completely rule out the presence of other systematic confounding factors that could have influenced the financial markets especially during stress periods, given that our window of analysis is a 24-hour period, instead of minutes. In addition, our estimates rest on the precision of the sentiment indices themselves.

Figure 2: Estimated relationship between derived sentiment measures and bond yields, β_1 , across bond maturity

A. Using net sentiment measure, sentiment_t



B. Using positive/hawkish word count (normalised by total word count), $\text{sentiment}_t^{(+)}$



C. Using negative/dovish word count (normalised by total word count), $\text{sentiment}_t^{(-)}$

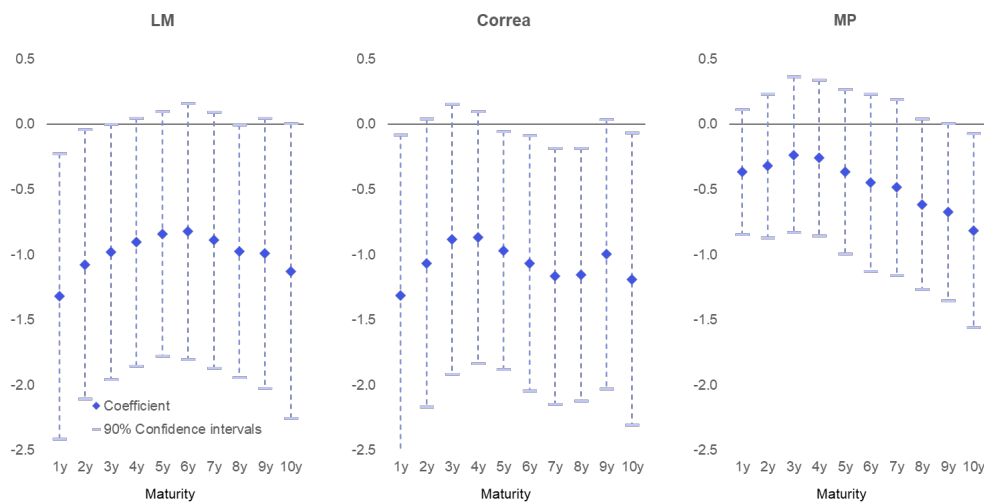
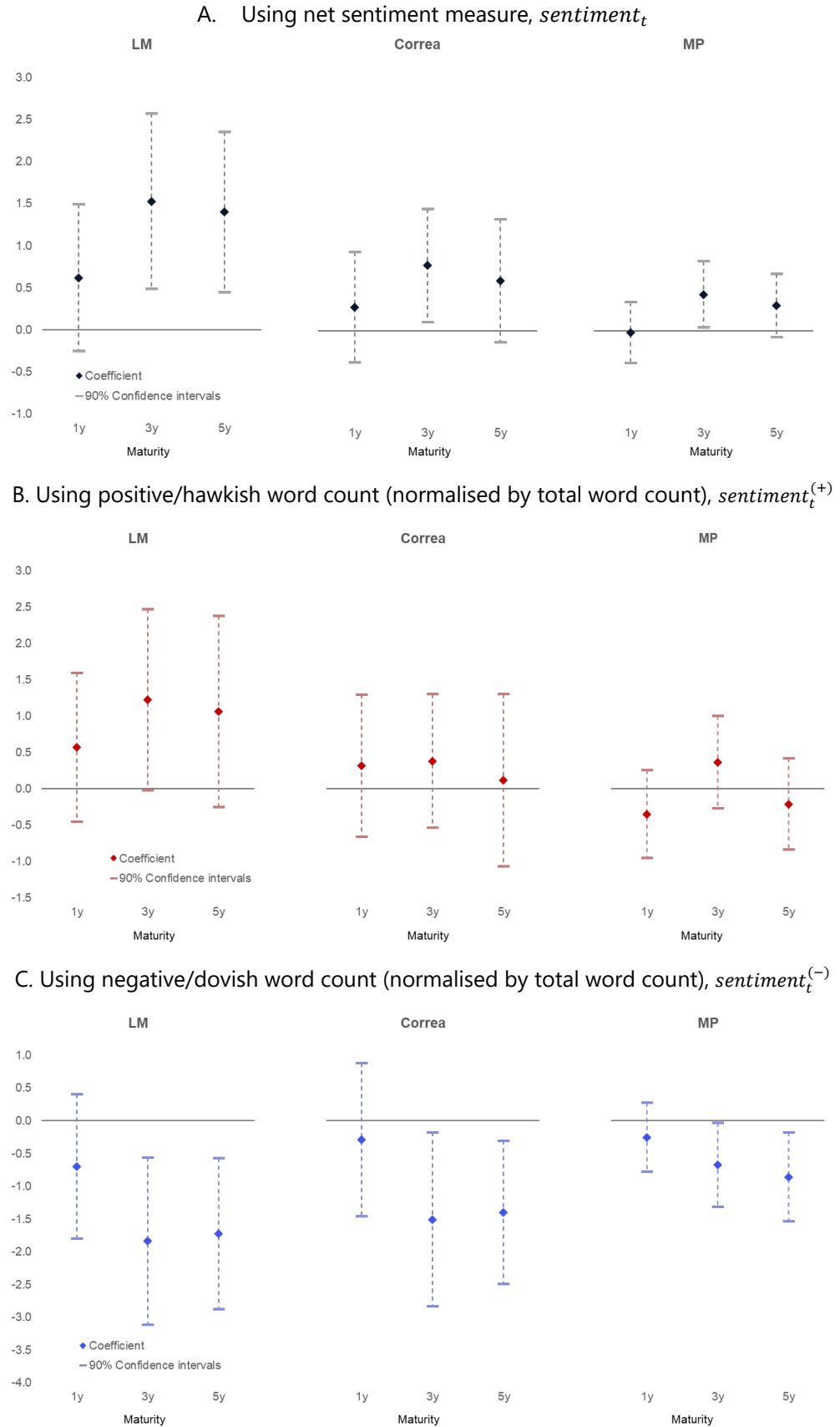


Figure 3: Estimated relationship between derived sentiment measures and IRS yields, β_1 , across swap maturity



4. Conclusion

In this paper, we build on the existing literature on text-sentiment analysis, which has mainly been conducted in the advanced economies, and make two contributions. First, we extract and analyse the sentiments from the MPS by the Central Bank of Malaysia. Second, for the automated content analysis, we build upon existing dictionaries that are oriented towards finance or the financial stability context to develop one that is specially designed for monetary policy context. We find the sentiment measures move in line with changes in the policy rate, the OPR. In addition, there is also some evidence of an asymmetric impact of the sentiments on sovereign (MGS) yields and interest rate swap rates. Notwithstanding these findings, there remain some limitations to this study, such as the potential presence of other systematic confounding factors and the precision of the sentiment indices themselves. These issues could be further explored in future research.

References

- Bligh, Michelle C., and Gregory Hess. 2013. "Deconstructing Alan: A Quantitative Assessment of the Qualitative Aspects of Chairman Greenspan's Communication." In *Central Bank Communication, Decision Making, and Governance: Issues, Challenges, and Case Studies*, edited by Pierre L. Siklos and Jan-Egbert Sturm, 123–47. Cambridge, MA: MIT Press.
- Blinder, Alan S, Michael Ehrmann, Marcel Fratzscher, Jakob de Haan, and David-Jan Jansen. 2008. "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence." *Journal of Economic Literature* 46 (4): 910–45. <https://doi.org/10.1257/jel.46.4.910>.
- Correa, Ricardo, Keshav Garud, Juan M Londono, and Nathan Mislang. 2021. "Sentiment in Central Banks' Financial Stability Reports*." *Review of Finance* 25 (1). <https://doi.org/10.1093/rof/rfaa014>.
- Garcia, Diego. 2013. "Sentiment during Recessions." *The Journal of Finance* 68 (3): 1267–1300. <https://doi.org/10.1111/jofi.12027>.
- Jegadeesh, Narasimhan, and Di (Andrew) Wu. 2017. "Deciphering Fedspeak: The Information Content of FOMC Meetings." *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2939937>.
- Kuttner, Kenneth N. 2001. "Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market." *Journal of Monetary Economics* 47 (3). [https://doi.org/10.1016/S0304-3932\(01\)00055-1](https://doi.org/10.1016/S0304-3932(01)00055-1).
- Loughran, Tim, and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66 (1). <https://doi.org/10.1111/j.1540-6261.2010.01625.x>.
- Moniz, Andy, and Franciska de Jong. 2014. "Predicting the Impact of Central Bank Communications on Financial Market Investors' Interest Rate Expectations." In . https://doi.org/10.1007/978-3-319-11955-7_12.
- Nyman, Rickard, Sujit Kapadia, and David Tuckett. 2021. "News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment." *Journal of Economic Dynamics and Control* 127 (June). <https://doi.org/10.1016/j.jedc.2021.104119>.

Measuring Text-based Sentiments from Monetary Policy Statements

A Malaysian Case Study using Natural Language Processing

Sui-Jade Ho and Eilyn Chong

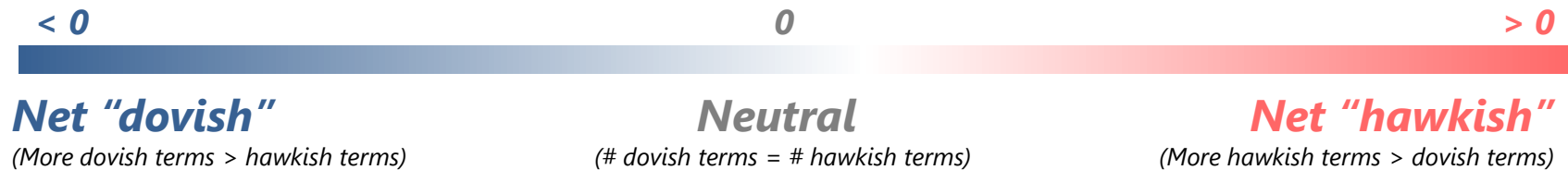
The views expressed are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Central Bank of Malaysia, the Monetary Policy Committee (MPC) or anyone else associated with the Central Bank of Malaysia.



Using automated content analysis, we derive sentiment measures to gauge how hawkish or dovish a monetary policy statement (MPS) is perceived to be

- Central banks publish MPS to provide insights into economic development and outlook, as well as to communicate judgment on the balance of risks and expectations on the future course of monetary policy.
- Using automated content analysis to extract the sentiment from each MPS published in 2004 – 2020 by the Central Bank of Malaysia, we derive a sentiment indicator to measure how often hawkish or dovish words are mentioned in each MPS.

$$sentiment_t = \frac{\sum \text{positive or hawkish words}_t - \sum \text{negative or dovish words}_t}{\sum \text{word count}_t} \text{ for each MPS } t$$



- Then, we analyse the relationship between the sentiment measures from these statements and financial market movements.

The count of positive/hawkish or negative/dovish words in the MPS is based on 3 dictionaries, including one specially refined to incorporate nuances in the monetary policy context ("MP")

Loughran-McDonald ("LM")

- Tailored specifically to finance
- Word lists by examining word usage in at least 5% of 10-Ks (i.e. annual reports) since 1994

Correa et al. ("Correa")

- Calibrated to the language of financial stability reports to generate a financial stability sentiment index

Monetary Policy ("MP")

- Combines the lexicon from LM and Correa, and reassigns words to *hawkish* or *dovish* tone
- Accounts for the monetary policy context

1 Additional words

For example: **Expansion/upside/robust**
(positive/hawkish)

2 Economic relationship

Low unemployment
(negative/dovish)

▶ **Low unemployment**
(positive/hawkish)

3 Tone to be independent of policy rate* adjustments

OPR increase
(positive/hawkish)

▶ **OPR increase**
(neutral)

4 Fin. mkt. movements with no clear bearing on MP stance

FM stress receding
(negative/dovish)

▶ **FM stress receding**
(neutral)

* The Overnight Policy Rate (OPR) is the indicator of the monetary policy stance.

Source: Correa, Ricardo, Keshav Garud, Juan-Miguel Londono-Yarce, and Nathan Mislav. 2017. "Constructing a Dictionary for Financial Stability." *IFDP Notes* 2017 (33).

Loughran, Tim, and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66 (1).

Impact of sentiment measures on financial markets through high-frequency event studies

Goal

- ▶ Investigate if the sentiment measures can explain movements in financial markets upon MPS release

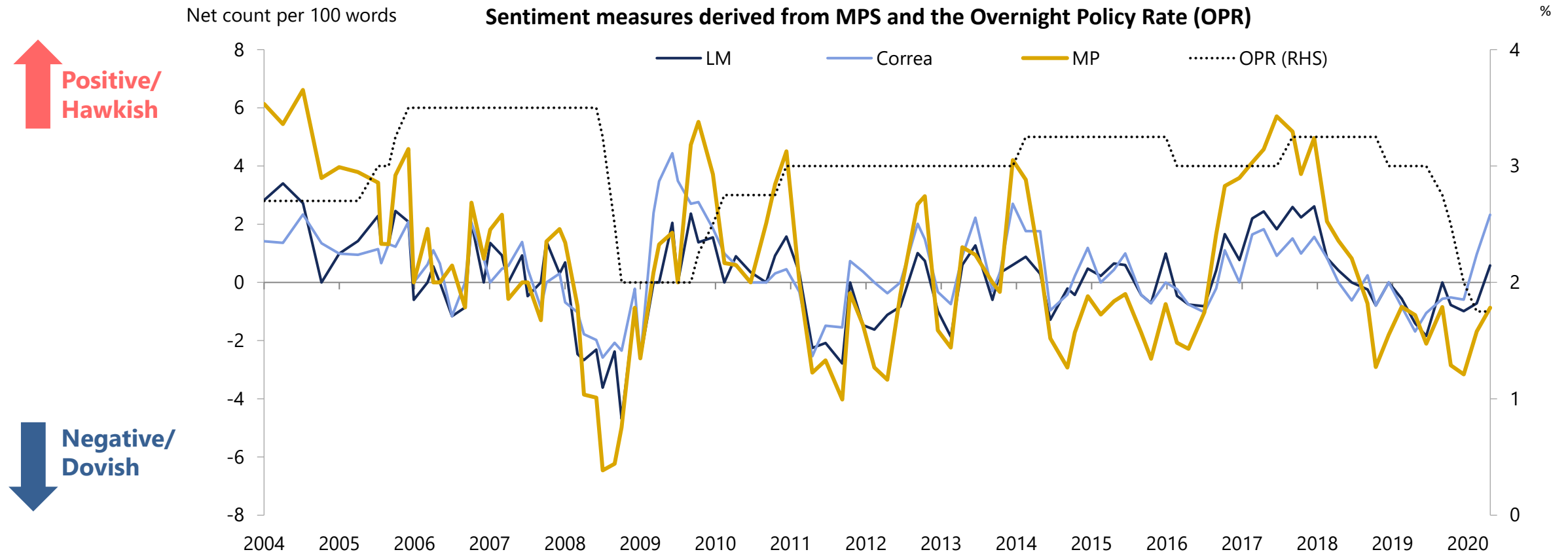
Hypothesis

- ▶ Controlling for OPR changes, a more positive (negative) net sentiment measure is associated with an increase (a decrease) in sovereign (Malaysian Government Securities, MGS) yields or interest rate swaps (IRS) rate

Empirical methodology

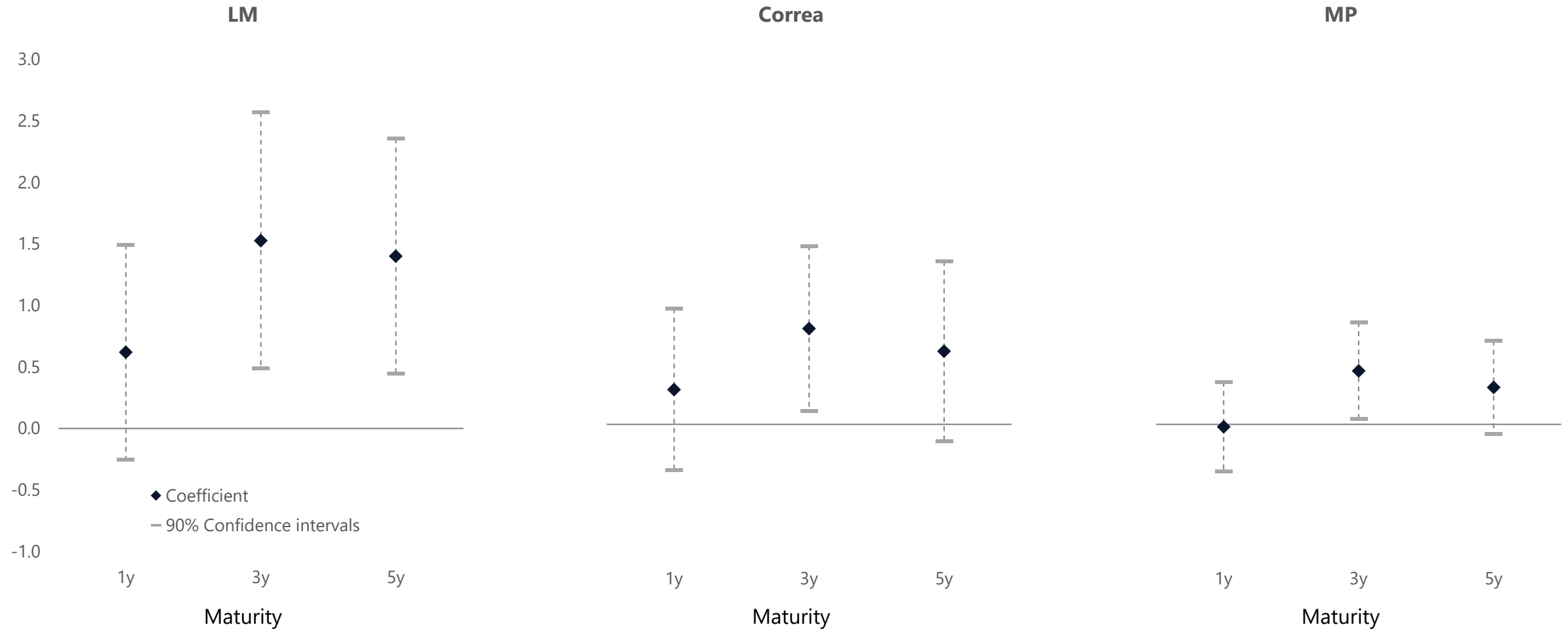
- ▶ Estimating equation : $\Delta y_t^m = \alpha + \beta_1 \text{sentiment}_t + \beta_2 \Delta \text{opr}_t + \varepsilon_t$
 - Δy_t^m is the 24-hour window change in MGS yields of maturity with year, $m \in \{1, 2, \dots, 10\}$ or IRS rates of maturity with year, $m \in \{1, 3, 5\}$ around MPS announcement days
 - $\text{sentiment} \in (LM, \text{Correa}, MP)$. All sentiment measures are normalised by total word count in the MPS
 - Δopr_t is the change in the Overnight Policy Rate (OPR). β_2 measures the impact of OPR change on Δy_t^m
 - Our coefficient of interest is β_1
 - Key identification assumption: Other factors affecting MGS yields and IRS rates within the short window are on average, orthogonal to MPS sentiment index

The sentiment measures generally move in line with changes in the policy rate



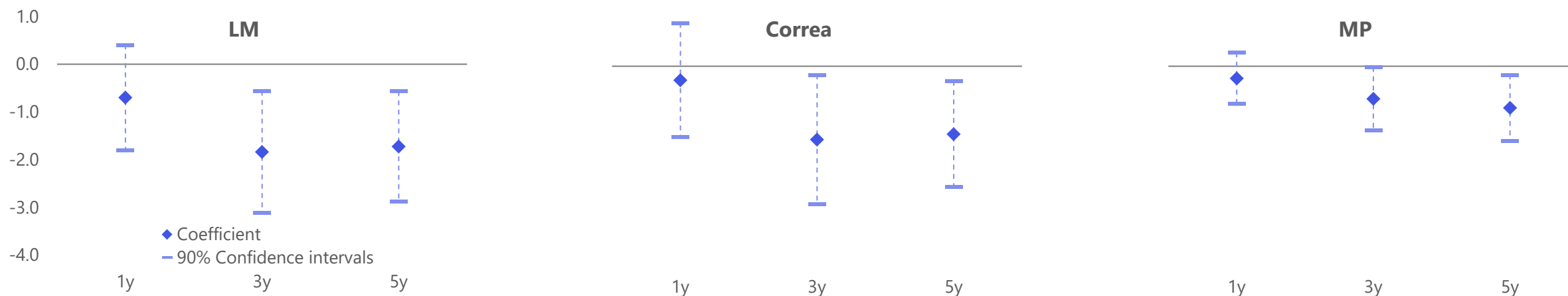
Some evidence that the sentiment measures can impact IRS rates

Estimated relationship between **net** sentiment measures and interest rate swap rates , β_1 across swap maturity

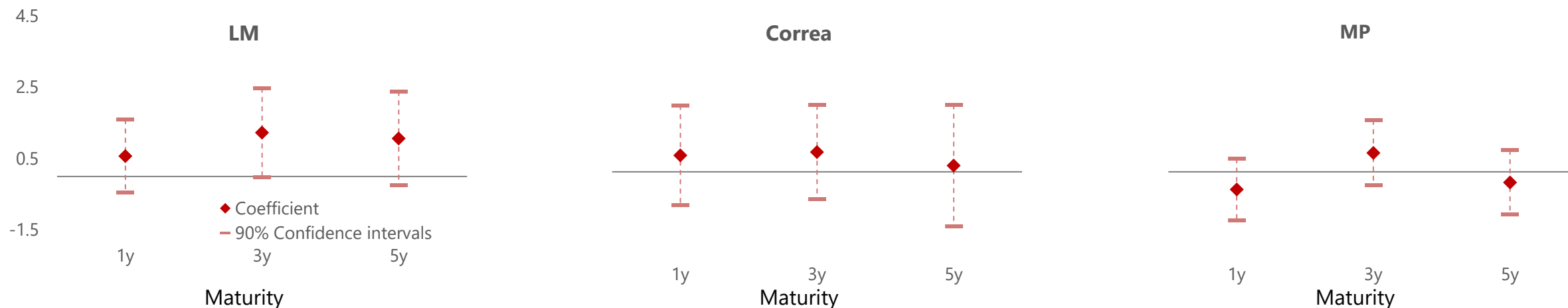


Dovish (negative) words appear to have stronger associations with a decline in swap rates

Estimated relationship between **negative** sentiment measures and interest rate swap rates , β_1 across swap maturity

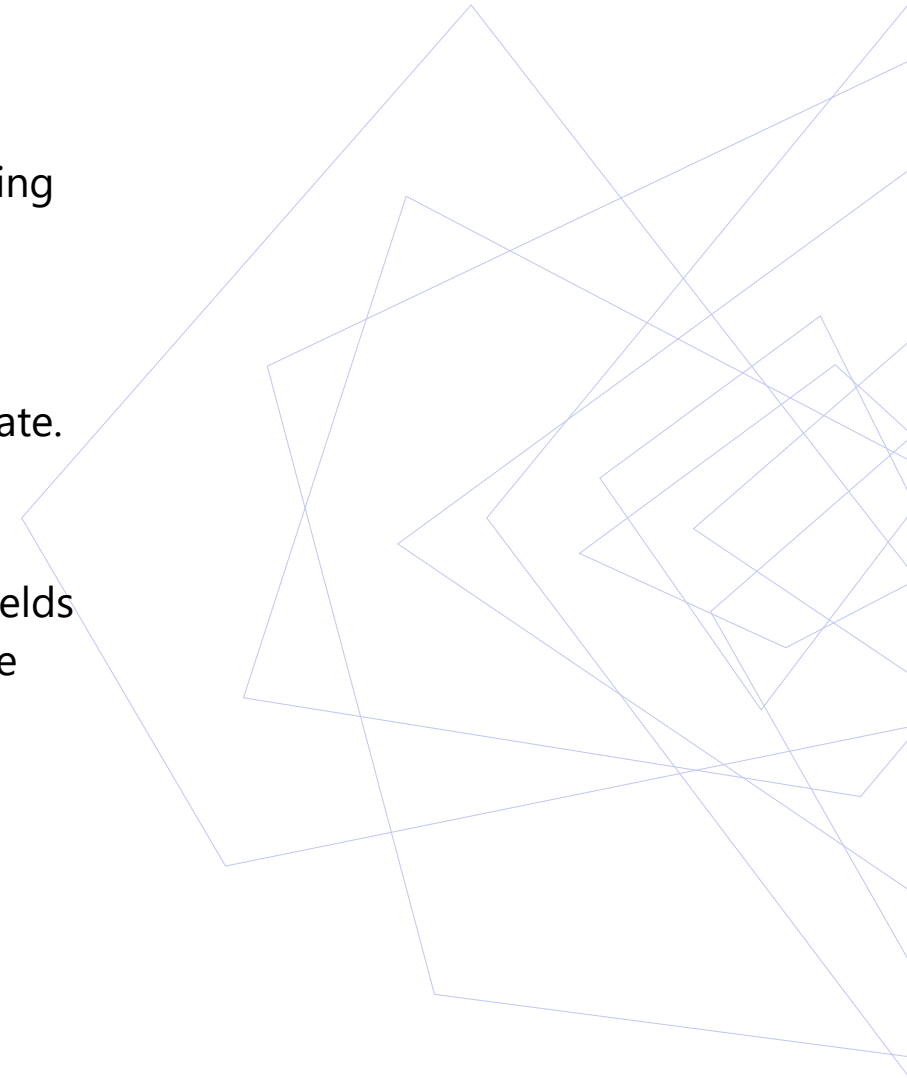


Estimated relationship between **positive** sentiment measures and interest rate swap rates , β_1 across swap maturity



Conclusion

1. We derive sentiment measures from the monetary policy statements published by the Central Bank of Malaysia using three dictionaries, including one specially developed for monetary policy context.
2. We find the sentiment measures move in line with changes in the policy rate.
3. There is an asymmetric impact of the sentiment measures on sovereign yields and interest rate swaps rates. Our findings provide some evidence that the wording in the monetary policy statement is informative for the financial markets, especially during economic stress periods.



IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Interactive visualization tool: outlier detection in large multidimensional datasets¹

Christoph Leitner, Thomas Kemetmueller and Philipp Reisinger,
National Bank of the Republic of Austria

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Interactive Visualization Tool

Outlier detection in large multidimensional datasets

Philipp Reisinger, Thomas Kemetmüller, Christoph Leitner¹

Abstract

The Oesterreichische Nationalbank (OeNB) collects granular credit data from several hundred domestic reporting agents (banks and financial institutions) in a Granular Credit Register (GCR) and in an integrated manner with the ESCB's AnaCredit. Per reporting date, the GCR contains approximately 1.2 million credit instruments with well beyond one hundred data dimensions. To assure data quality in such complex granular datasets, powerful approaches and tools are needed. At OeNB a variety of statistical methods (including statistical inference, regression models and outlier detection algorithms) is used to identify outliers and potentially incorrect values.

This paper presents visualisation techniques to support these statistical methods by additional results as well as simplifying these results by appropriate plots. The focus here are parallel coordinates plots which allow the visualization of multidimensional granular datasets but require a high level of interactivity. Hence, we developed a browser-based interactive visualization tool to appropriately utilize parallel coordinate plots and other techniques to explore large datasets. An example for the GCR demonstrates the usage of this tool to generate insights, identify anomalies, and thus helps to improve data quality.

Keywords: Data visualization, parallel coordinates, outlier detection

¹ Oesterreichische Nationalbank – Supervisory Statistics, Models and Credit Quality Assessment Division

1. Introduction

Under the ECB Regulation on AnaCredit (see EZB VO (EU) 2016) as well as the national banking legislation on granular credit data (see GKE-V 2018 and § 75 BWG 2022) the Oesterreichische Nationalbank (OeNB) collects monthly credit data from several hundred domestic reporting agents (banks and financial institutions) on a granular basis. The integrated data (implemented employing an entity relationship model) forms the henceforth called Granular Credit Register (GCR). The GCR contains approximately 1.2 million credit instruments with well beyond one hundred numerical and categorical data dimensions per reporting date.

Reported data must fulfil an extensive set of validation rules and plausibility checks upon transmission by reporting agents to avoid the acceptance of faulty data. The distinction between validation rules and plausibility checks is that the former identify unambiguous reporting errors with certainty, while the latter identify data that are likely to be incorrect. Still, accepted data may not necessarily be fully correct. Therefore, downstream quality assurance is required to improve overall quality of the data. In this step OeNB uses statistical methods like statistical inference, regression models and other effective outlier detection algorithms.

Following Anscombe (1973), who showed that summary statistics can obscure outliers, which could have been easily spotted by plotting the data, we experimented with visual methods to utilize human pattern recognition abilities. In the end, we developed a customized tool ("PARVIZ") to support outlier detection in the GCR. Chapter 2 describes the chart types incorporated in PARVIZ. Chapter 3 explains its design, set up and features. Chapter 4 demonstrates features of the tool in a use case identifying outlier credit instruments in the GCR.

2. Methods

In the context of analysing datasets (e.g., identifying outliers) the distribution of dimensions and the correlation between dimensions play an important role. There are a lot of possibilities to do this using visualization. The need to work with highly dimensional granular data limits this set. Parallel coordinates plots (PCPs) can show several data dimensions on parallel axes and are capable of handling large granular datasets but are intended to be used with numerical dimensions (see Figure 1).

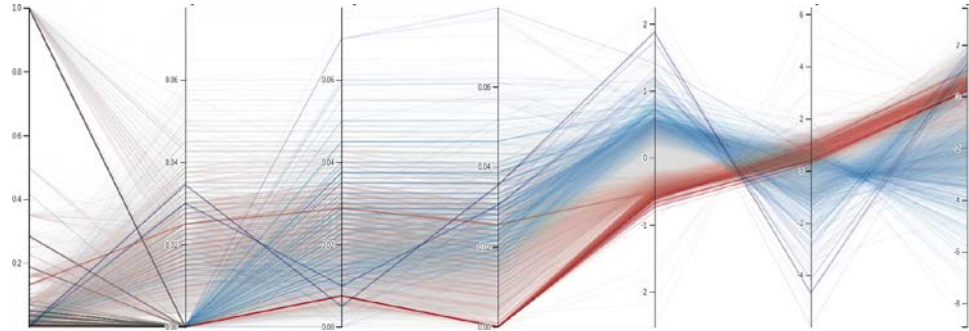


Figure 1: An example of parallel coordinates plot displaying data with seven numerical dimensions. Each observation is represented by a polyline, cutting the axes at the observation's value in the corresponding dimension. The colour of the polylines indicates z-scores of the 5th dimension (> 0.6745 : blue; < -0.6745 : red).

Other chart types, like parallel sets (see Kosara *et al*, 2006) or mosaic plots (see Hartigan and Kleiner, 1984), are better for categorical data, but were omitted due to the involved aggregation of values per category, lacking the required granularity. The PCPs shortcoming with categorical dimensions can be compensated by displaying aggregated values as bar charts for each respective axes, indicating the dimensions' distribution much like histograms. Sometimes box plots are used to show the distribution of numerical variables but since they only rely on simple summary statistics (and share the potential flaw presented by Anscombe, 1973), we chose violin plots to explicitly display the probability density. PCPs indicate the correlation of adjacent axes. However, Li *et al* (2010) showed that scatter plots are better for visual correlation analysis. As Figure 2 shows that patterns can only be identified between adjacent axes in PCPs, it would be useful to give users the possibility to easily modify axes order (along with their rotation and scaling). An interactive implementation of such a plot can solve this issue.

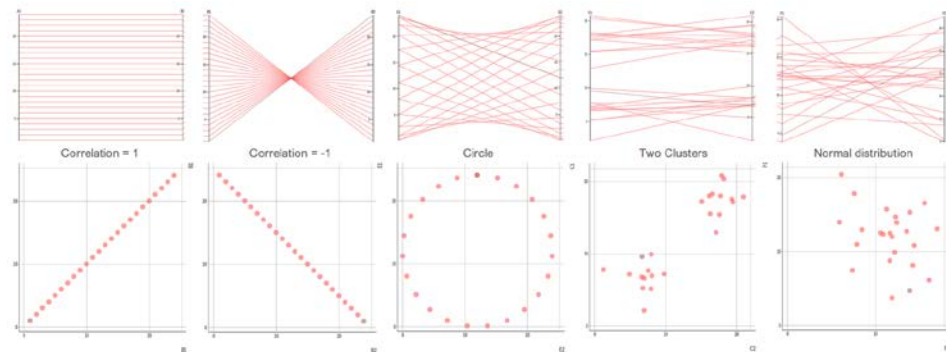


Figure 2: Parallel coordinates plots and scatter plots can be used to evaluate correlation in different data examples.

3. The interactive visualization tool PARVIZ

The interactive visualization tool (called “PARVIZ”) is designed as pure front-end web application, implemented with standard web technologies² and additional libraries: D3.js (Bostock, 2018), d3.parcoords.js (Chang, 2016 and Xing, 2019), d3.tip (Palmer, 2013), SlickGrid (Leibman, 2012), underscore.math (Chang, 2012), jQuery (jQueryTeam, 2016), jQueryUI (jQueryTeam, 2011), underscore.js (Ashkenas, 2012). The charts in PARVIZ are based on D3.js-examples by Bostock (2021), Davis (2021), Chang (2021), Petersson (2018), Galavotti (2017) and Holtz (2018).

The tools’ functionalities are contained in a single file (PARVIZ.html) that runs locally in a web browser (no backend). PARVIZ does not require an internet connection – all necessary libraries are included in the HTML-file (with only 1 MB). Data to be displayed is stored separately as JSON³-file in a secured folder on a network drive, preventing unauthorized access. The path to the secured data is contained in an additional JSON-file (2 KB), which allows to switch between different datasets without editing the main HTML-file. The specified dataset is then loaded when opening the HTML-file.

The initial view of the user interface contains a menu bar on the top, a PCP in the centre and a table at the bottom. All dimensions of the loaded data are included in the table, whereas only the selected (via menu) are displayed in the PCP. For each dimension with numerical values the distribution is shown using a violin plot on the axis. For each categorical dimensions a bar chart is shown on the axis. These axis plots can be hidden by changing their opacity in the menu bar. Missing/null values are displayed below the horizontal line at the bottom of the PCP.

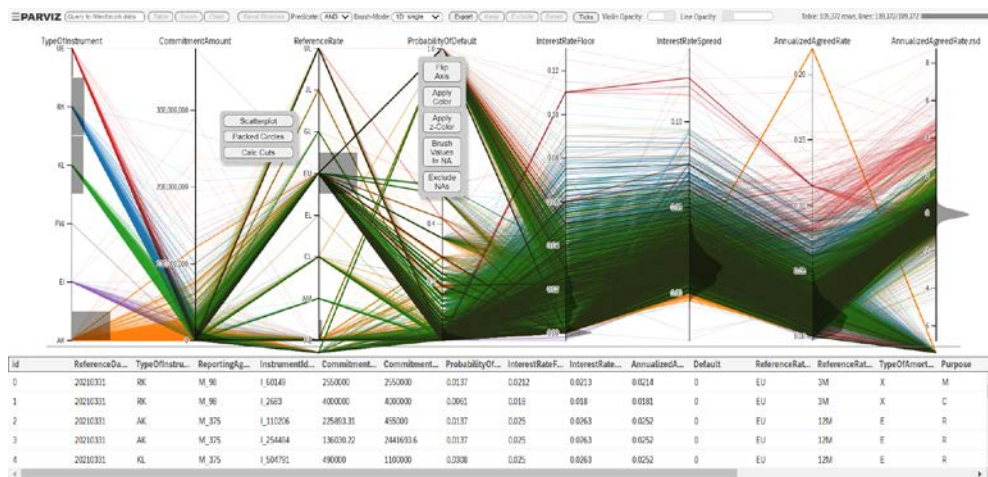


Figure 3: PARVIZ' user interface displays here two categorical and six numerical dimensions as parallel coordinates plot and a table with all dimensions of the corresponding dataset.

On hovering over table rows, the corresponding polyline is highlighted in the PCP (for persistent highlighting multiple rows can be “left-clicked”). Selecting and highlighting polylines is especially important when many polylines are plotted and individual polylines cannot be distinguished (“overplotting”). Brushes allow to select polyline bundles based on value ranges per axis (“single”: one range per axis, “multi”:

² JavaScript, CSS, HTML, SVG, Canvas

³ JavaScript Object Notation

multiple ranges per axis) or characteristics of lines between two adjacent axes ("strums": lines intersecting with a given strum, "angular": lines sloping within a given range specified by a circle segment). Figure 4 shows range-based brushes on the left (top: "single", bottom: "multi") and line-based brushes on the right (top: "strums", bottom: "angular"). Multiple brushes (of one type) can be joined using a logical "AND" or "OR" to allow for comprehensive data queries, filtering the tables rows accordingly.

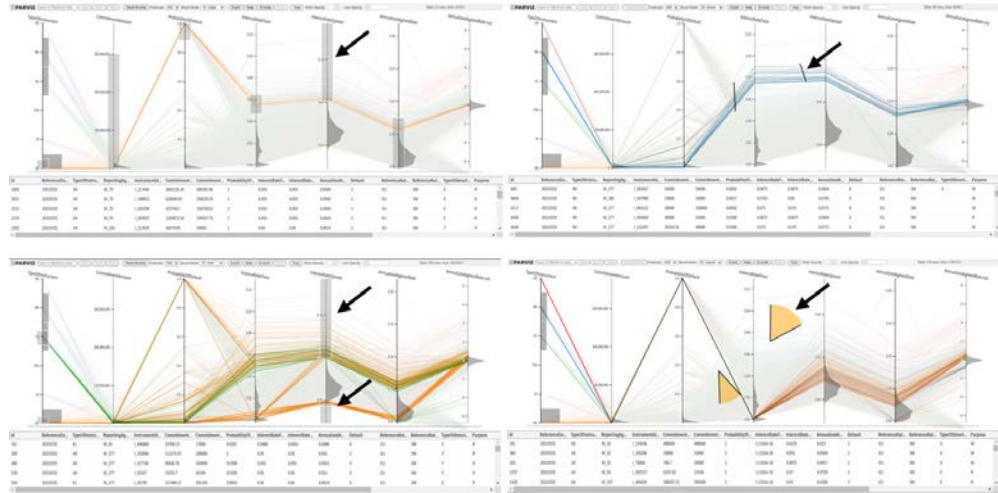


Figure 4: Different brushing methods (top left: "single", bottom left: "multi", top right: "strums", bottom right: "angular") allow targeted selections.

In addition to brushing the PCP via mouse interaction, an input box is available for a more precise selection and can be used to brush the plot or filter the table. The count of all currently selected rows and lines is displayed in the menu bar. For further reference, all rows of the selection are easily exportable in a CSV-format by a single click. Selections can be kept or excluded, triggering the rescaling of all axes, which is useful to zoom in on certain dimensions. Axes are reorderable (drag-and-drop) and invertible via an axis menu. Double-clicking on an axis header shows the axis' menu, which also enables users to quickly brush or exclude missing values and change the polyline colour to express the respective dimension – for categorical dimensions distinct colours are applied and with numerical dimension a colour gradient is used to differentiate high and low values or the z-scores⁴ of these values. When clicking on a column header of a numerical dimension in the table, a box plots is generated. By clicking on an element of the box plot the brushing in the PCP can be controlled (see Figure 5).

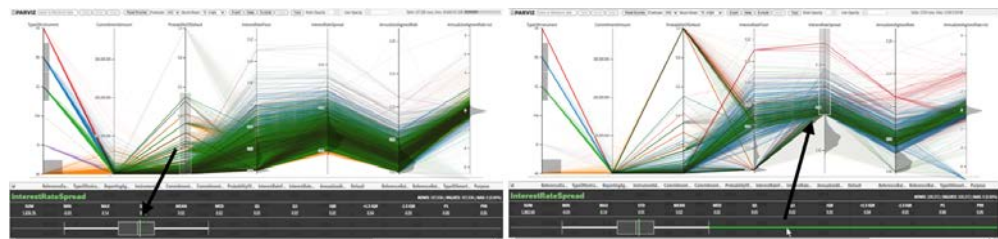


Figure 5: By clicking on a column header of numerical dimensions displays a box plot for all selected polylines (left). By clicking on an element of the box plot the brushing in the PCP can be controlled (right).

⁴ The z-score of a value is the standardized value, computed via $(\text{value} - \text{mean of all values of the same dimension}) / \text{standard deviation of all values of the same dimension}$.

Scatter plots are displayed by clicking between two adjacent axes. A tooltip shows configurable info for each dot (see Figure 6). Clicking on a dot filters the respective row in the data table. By hovering this row, the corresponding polyline in the PCP is highlighted. Packed circles show how a numerical dimension is distributed across multiple levels of categorical dimensions (see Figure 7). By clicking on the right of a numerical axis, its values are hierarchically aggregated for all categorical dimensions to the right of this numerical dimension, enabling users to quickly configure the plot by arranging the axes accordingly. Especially for the purpose of outlier detection, we also included a feature to identify polylines that diverge from the majority by showing the number of all other polylines that each polyline cuts (as total over all axes as well as for each adjacent axes-pair) on additional axes in the PCP. In this sense, the number of polyline cuts can be regarded as an outlier score where the polylines that “buck the trend” are potentially outliers and should be further investigated.

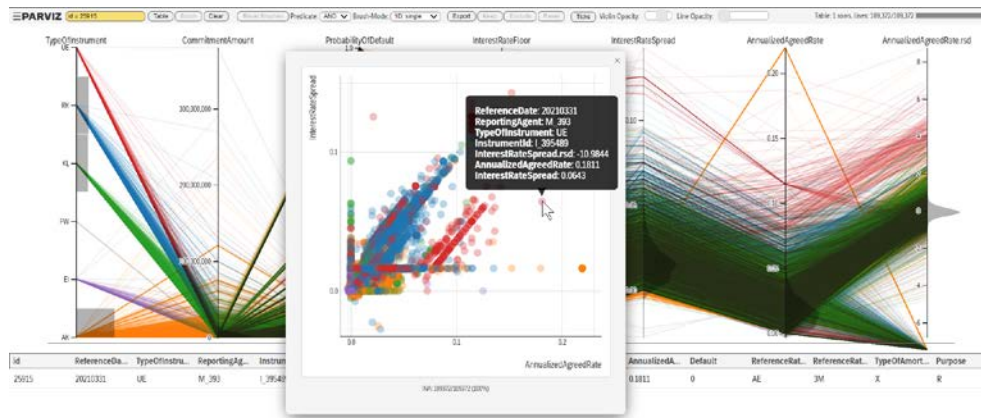


Figure 6: By clicking between two adjacent axes, a zoomable scatter plot (with tooltip) is displayed for those dimensions.

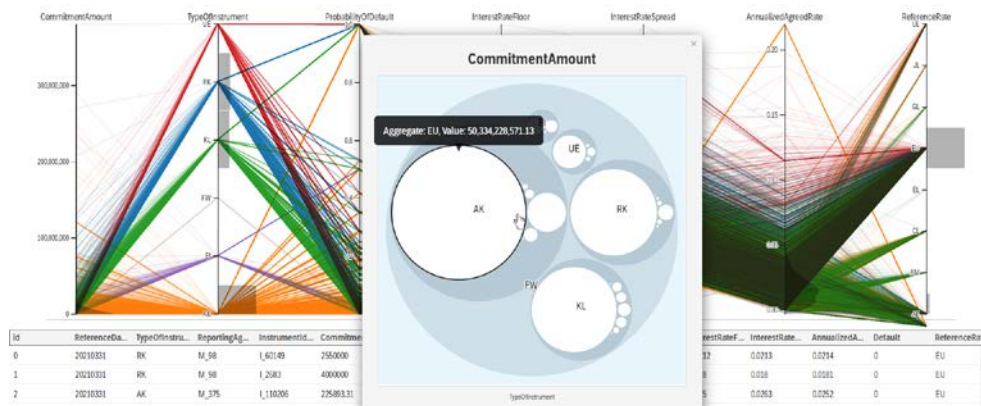


Figure 7: Zoomable packed circles plots of a numerical dimension aggregated over categorical dimensions (here “TypeOfInstrument” and “ReferenceRate”) help to explore categorical dimensions of datasets.

4. Using PARVIZ to identify outliers in the GCR

In this chapter we use a sample of the GCR to demonstrate features of the visualization tool with the goal to identify outlier credit instruments in the data. The data contains 109,372 rows (credit instruments) and 20 dimensions of one reporting date and is anonymized for this paper (see Table 1). The last three of those dimensions (indicated by the extension ".rsd") depict outlier scores. These scores are based on the residuals of a linear model (measured in multiples of the standard deviation of the residuals distribution) predicting the respective dimension using other dimensions as the explanatory variables. High absolute residuals reflect potential outliers (for more information on this method see Aggarwal, 2017).

Dimensions

Table 1

| Name | Type | Units | Example |
|-----------------------------|-------------|----------|------------------|
| ReferenceDate | categorical | - | "20210331" |
| TypeOfInstrument | categorical | - | "AK" |
| ReportingAgent | categorical | - | "M_12" |
| InstrumentId | categorical | - | "I_12345" |
| CommitmentAmount | numerical | € | 12,345.67 |
| CommitmentAmountAtInception | numerical | € | 12,345.67 |
| ProbabilityOfDefault | numerical | % | 0.2345 (=23,45%) |
| InterestRateFloor | numerical | % | 0.02 (=2%) |
| InterestRateSpread | numerical | % | 0.02 (=2%) |
| AnnualizedAgreedRate | numerical | % | 0.02 (=2%) |
| Default | numerical | 0/1 | 1 (=Default) |
| ReferenceRate | categorical | - | "EU" |
| ReferenceRateMaturity | categorical | - | "3M" |
| TypeOfAmortization | categorical | - | "X" |
| Purpose | categorical | - | "R" |
| CarryingAmount | numerical | € | 12,345.67 |
| ImpairmentAmount | numerical | € | 12,345.67 |
| AnnualizedAgreedRate.rsd | numerical | σ | 1.2345 |
| InterestRateSpread.rsd | numerical | σ | 1.2345 |
| InterestRateFloor.rsd | numerical | σ | 1.2345 |

Extreme values are natural candidates for potential outliers. In PARVIZ they are easily identified by examining polylines on each axis of the PCP and brushing relevant ranges. Brushing several axes quickly reveals local extreme values in subsets (cf. Figure 8).

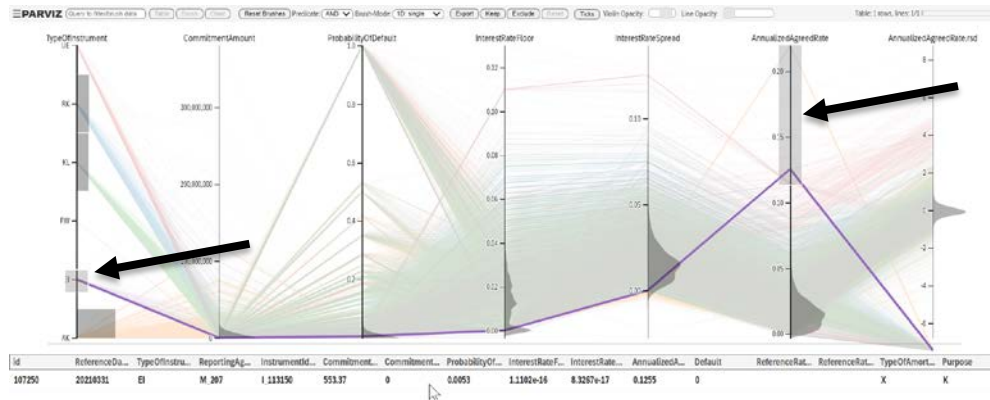


Figure 8: The maximum value of dimension “AnnualizedAgreedRate” for one credit instrument type (dimension “TypeOfInstrument” = “EI”) can be easily found by brushing two axes in the parallel coordinates plot.

The polyline of the highlighted instrument (id = 107250) in Figure 8 shows – despite having the highest annualized agreed rate of all instruments of the same type (“EI”) – that

- (1) the interest rate spread is very low, resulting in a line that goes against the flow of most other lines between the adjacent axes;
- (2) the instrument was not included in the outlier model for “AnnualizedAgreedRate” (indicated by the missing value on the last axis).

To further investigate the apparent correlation between the interest rate spread and the annualized agreed rate (indicated by the fact that most lines are parallel), a scatter plot for all credit instruments of this type is generated. The scatter plot confirms the relationship and detects a few outliers (including credit instrument with id = 107250; see Figure 9). Using the tools’ feature to count the intersections of each line with all other lines also identifies this credit instrument (see Figure 10).

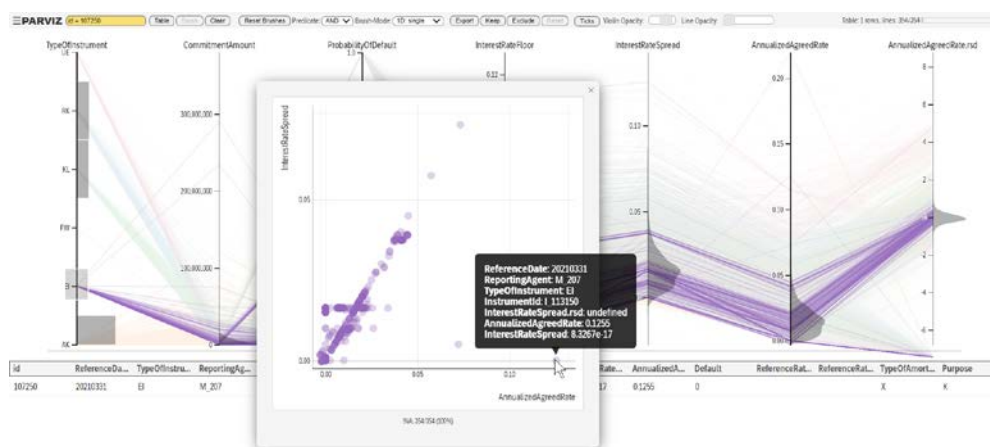


Figure 9: The scatter plot confirms the relationship and detects the instrument with the id = 107250 as outlier.

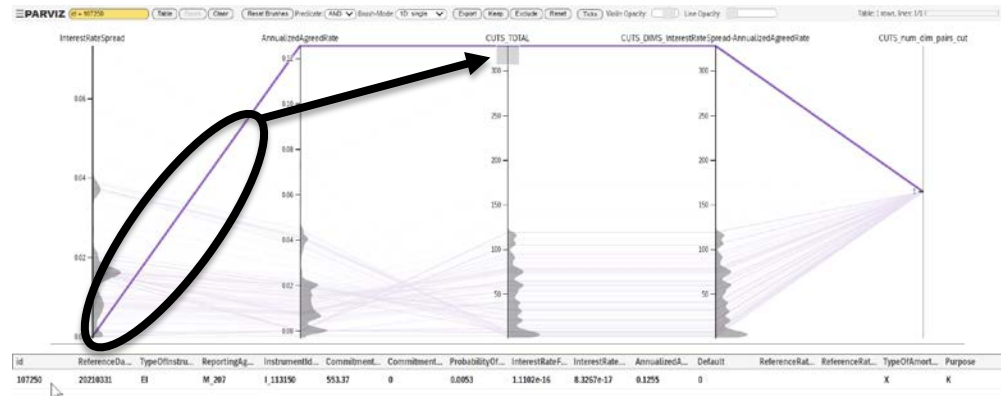


Figure 10: Counting the number of intersections of each line with all other lines shows the instrument with id 107250 as outlier (cutting far more lines than any other line/credit instrument).

The interactive box plot of the respective table column helps to verify the visually assessed anomaly of the inspected credit instrument. Selecting the box plots' upper outliers (> upper whisker) of the dimension "AnnualizedAgreedRate" and all lower outliers (< lower whisker) of the dimension "InterestRateSpread" results in two credit instruments: The known instrument with id = 107250 (see above) and another one (see Figure 11). The high residuals strengthen the finding that the combination of high annualized rates and low interest rate spreads are being suspicious. These residuals provide valuable guidelines not only for validating outliers but also to find them: By scrutinizing the lines slope in Figure 12, it seems that high residuals are connected to this unusual combination of interest rate attributes for all instrument types. This presumed relation can be exploited by using the angular brush to select all lines (including those with missing values on the residuals axis) with a similar slope to expand the models results to credit instruments not assessed by the model. Selecting all upward sloping lines of credit instruments with type "EI" leads to the same result as above. This time only by interpreting and analysing model results.

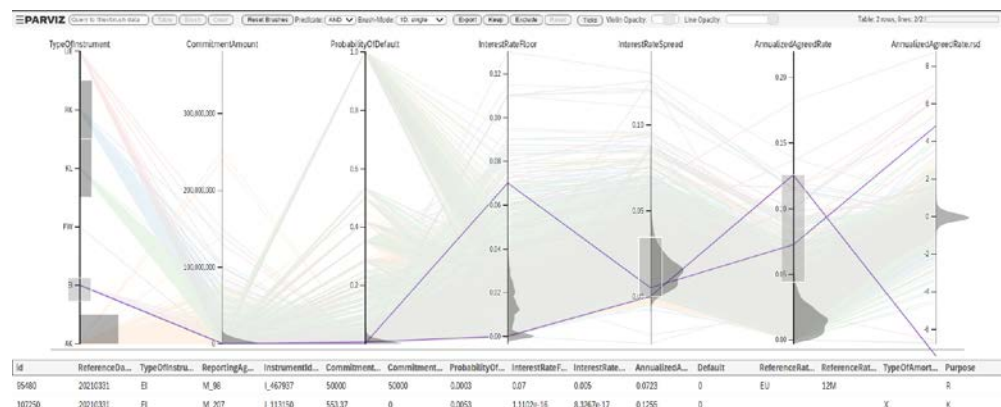


Figure 11: Two credit instruments identified as outliers by the combination of being an upper outlier (> upper whisker) of the dimension "AnnualizedAgreedRate" and a lower outlier (< lower whisker) of the dimension "InterestRateSpread".

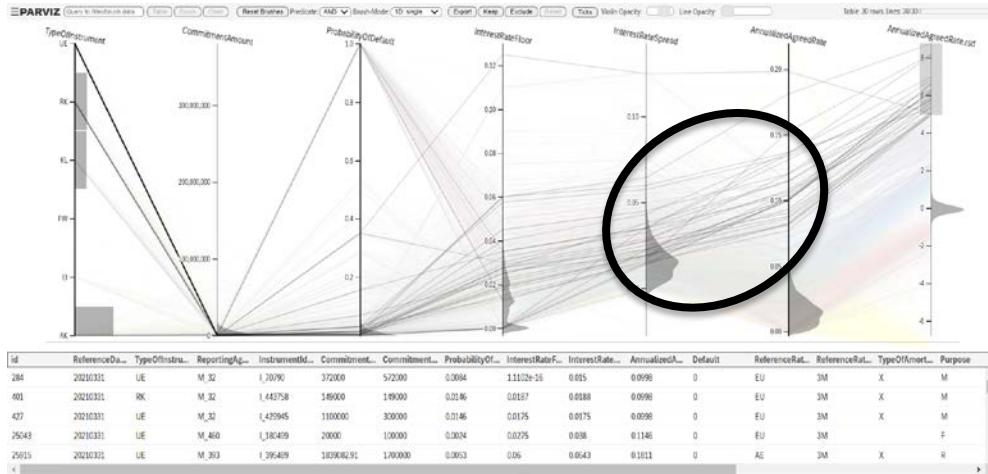


Figure 12: The high residuals on the last axis are apparently connected to the highlighted upward sloping lines.

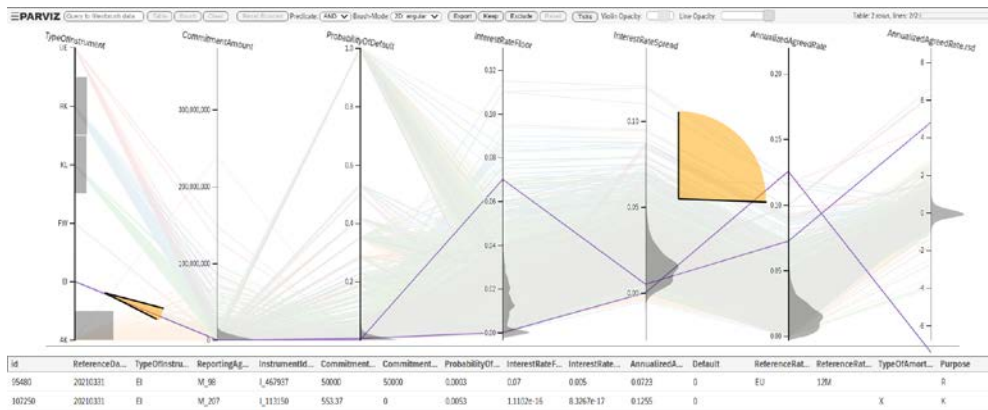


Figure 13: Selecting all upward sloping credit instruments (of type "EI") to identify the same two credit instruments as displayed in Figure 11.

Discussion and Conclusion

Visualizations can improve the understanding of large datasets by enabling users to quickly grasp the underlying data's structure. In addition, a selection of appropriate chart types combined with a fast interactive user interface helps identify potential anomalies in large datasets and subsequently increase data quality. Even though the results obtained from visually inspection and outlier detection do not guarantee that specific datasets are incorrect or erroneous data, they can justify further investigation. From our experience, derived indicators like outlier scores obtained from statistical modelling support the identification of outliers. PARVIZ combines all these aspects in one interactive visualization tool.

Interactive visualizations of big data can place considerable demands on computational resources. Since our tool runs in the browser on the user's machine, its performance is limited by local resources. As a consequence, loading large datasets and rendering all polylines of the PCP may take considerable time (several minutes). The data is only loaded once when PARVIZ is started, but polylines must be rendered frequently (e.g., on reordering or inversion of axes).

From our point of view, following the strategies to minimize the rendering duration of the PCP of the `parcoords`' library (see Chang, 2016 and Xing, 2019), like

- the use of Canvas to draw polylines (not of SVG),
- progressive rendering (to allow for interactions while lines are still being rendered),
- rendering prioritization (to render brushed polylines first),
- background rendering (to continue rendering after all brushed polylines are rendered);

is useful.

Additionally, the library allows to configure the render rate (number of polylines drawn at once). The higher the render rate, the faster the rendering process, but the greater the lag of user interactions. Lower render rates allow for more fluid user interactions but take longer to complete the plot. Concerning this trade off, we prefer to not inhibit users in their actions and opt for a low render rate (e.g., 100).

The design of our tool does not allow the use of larger decentralized server resources. Even the use of a powerful backend could possibly not alleviate the problem due to the high degree of interactivity that determines the plots appearance. For the time being, additional performance improvements only seem possible by increasing local resources and further optimizing the current code.

Performance requirements naturally increase with the size of data. So, we advise selecting useful subsets to reduce the computational burden and avoid overplotting. We recommend restricting plotting to only the data required for the current task. With advancements of the capabilities of modern web browsers and desktop computers performance will be further improved, allowing to handle increasing amounts of data.

Readily available and well documented web development techniques combined with open-source frameworks and a vast selection of visualization examples facilitate the development of tailor-made solutions to interact with data. We highly encourage to take advantage of these possibilities and create customized solutions.

References

- Anscombe, F. J.**, 1973, "Graphs in Statistical Analysis". *American Statistician*, 27 (1): 17–21.
- Aggarwal, C. C.**, 2017, "Outlier Analysis", Second Edition, Springer, Cham., pp. 70–74
- Ashkenas, J.**, 2012, "Underscore.js 1.3.1", Retrieved from <http://documentcloud.github.com/underscore>
- Bostock, M.**, 2018, "D3.js Version 5.4.0", Retrieved from <https://d3js.org/>
- Bostock, M.**, 2019, "Parallel Coordinates", *Mike Bostock's Block*, Retrieved from <https://bl.ocks.org/mbostock/1341021>
- BWG**, 2022, "Bankwesengesetz (Austrian Banking Act)", Version from 1.2.2022
- Davis, J.**, 2021, "Parallel Coordinates", *Jason Davies's Block*, Retrieved from <https://bl.ocks.org/jasondavies/1341281>
- Chang, K.**, 2012, "Underscore.math.js 0.1.2", Retrieved from <http://github.com/syntagmatic/underscore.math>
- Chang, K.**, 2016, "Parallel Coordinates (0.7.0)", Retrieved from <https://syntagmatic.github.io/parallel-coordinates/>
- Chang, K.**, 2021, "Nutrient Parallel Coordinates", *Kai's Block*, Retrieved from <http://bl.ocks.org/syntagmatic/3150059>
- EZB VO (EU)**, 2016, "Regulation (EU) 2016/867 of the European Central Bank of 18 May 2016 on the collection of granular credit and credit risk data (ECB/2016/13)", *Official Journal of the European Union*, 1.6.2016
- Galavotti, D.**, 2017, "Zoomable Circle Packing", *Davo Galavotti's Block*, Retrieved from <https://bl.ocks.org/davo/cd7261bd67581f284b6601fa4dd652b1>
- GKE-V 2018**, 2018, „Verordnung der Finanzmarktaufsichtsbehörde (FMA) über die Meldungen zur Erhebung granularer Kreditdaten (Granulare Kreditdatenerhebungs-Verordnung 2018 – GKE-V 2018)", *BGBL. II Nr. 170/2018*
- Hartigan, J.A. and Kleiner, B.**, 1984, „A Mosaic of Television Ratings", *The American Statistician*, 38, 32–35
- Holtz, Y.**, 2018, "Violin plot with jitter in d3.js", Retrieved from https://www.d3-graph-gallery.com/graph/violin_jitter.html
- jQueryTeam**, 2011, "jQuery v1.7", Retrieved from <http://jquery.com>
- jQueryTeam**, 2016, "jQuery UI - v1.12.1", Retrieved from <http://jqueryui.com>
- Kosara, R., F. Bendix, and H. Hauser**, 2006, "Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data.", *Transactions on Visualization and Computer Graphics (TVCG)*, vol. 12, no. 4, pp. 558–568
- Leibman, M.**, 2012, "SlickGrid v2.1", Retrieved from <http://github.com/mleibman/slickgrid>
- Li, J., J. Martens, and J. Wijk**, 2010, "Judging Correlation from Scatterplots and Parallel Coordinate Plots", *Information Visualization* 9(1):13–30

Palmer, J., 2013, "d3.tip v0.6.3", Retrieved from <https://github.com/caged/d3-tip>

Petersson, J., 2018, "D3 Zoomable Scatterplot", Jonas Petersson's *Block*, Retrieved from <http://bl.ocks.org/peterssonjonas/4a0e7cb8d23231243e0e>

Xing, Y., 2019, "parcoords-es v2.2.10", Retrieved from <https://github.com/BigFatDog/parcoords-es>



OESTERREICHISCHE NATIONALBANK
EUROSYSTEM

Interactive Visualization Tool

Outlier detection in large multidimensional datasets

Philipp Reisinger, Thomas Kemetmüller, Christoph Leitner

IFC Workshop (virtual event), 15 February 2022

Philipp.Reisinger@oenb.at

Thomas.Kemetmueller@oenb.at

Christoph.Leitner@oenb.at

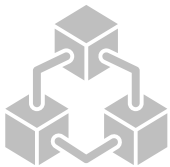
www.oenb.at



Background



Oesterreichische Nationalbank collects monthly data from several hundred domestic banks for a **granular credit register (GCR)**.



GCR contains over **1 million credit instruments** with over a **hundred** categorical and numerical **dimensions** (ER model) per reporting date.

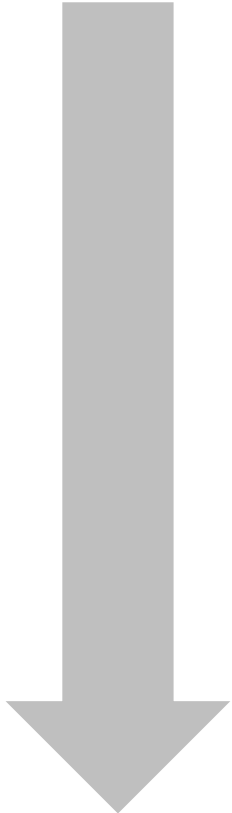


Requirements for **effective quality checks** increase with size and complexity of the collected data.



How can **incorrect data** be **identified**?

Quality Assurance



1. Validation rules (rejection of erroneous data on submission)

e.g. volume ≥ 0 , ...

2. Outlier detection using statistical models (downstream)

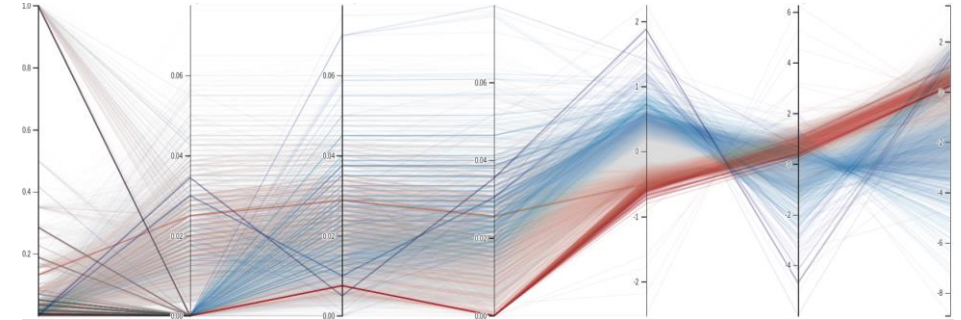
e.g. statistical inference, regression models, ...

3. Outlier detection using visual methods (downstream)

e.g. parallel coordinates plot, scatter plot, box plot, ...

Visualization

In **parallel coordinates plots**, each **observation** (credit instrument) is represented by a **polyline**, cutting the **axes** at its value in the corresponding **dimension** (colour can be used to highlight values of a certain dimension).

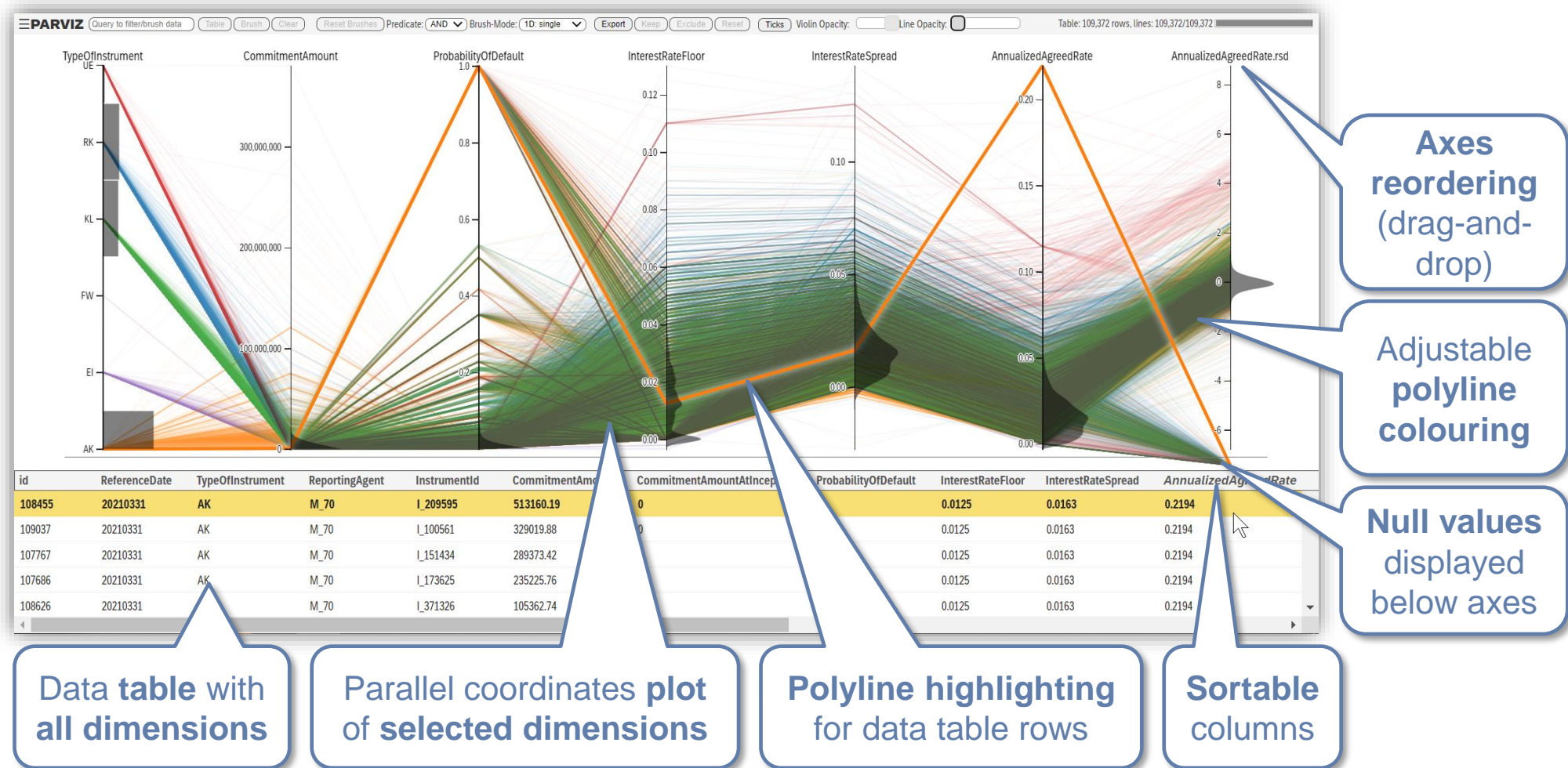


Parallel coordinates plots **require interactive implementations** to allow users to change the order and scaling of axes, polyline colour, ...

→ Customized tool for visual outlier detection:

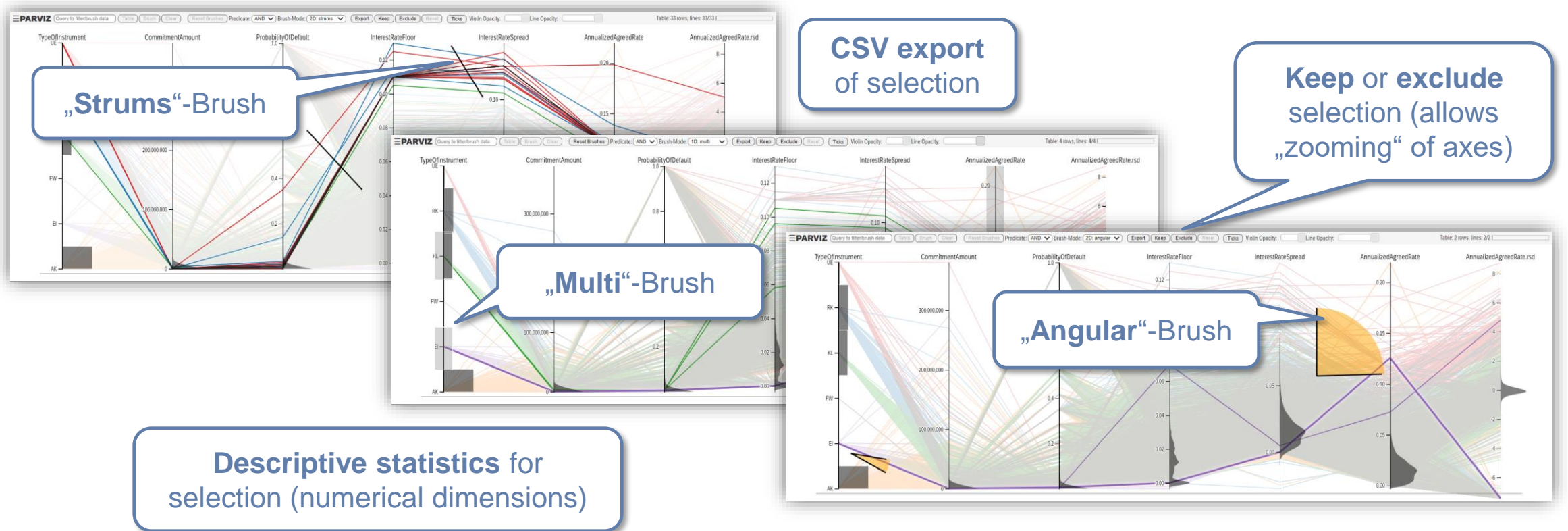
- Pure front-end **web application** using standard web technologies (HTML, JS, CSS, ...) plus additional libraries (D3.js, d3.parcoord.js, SlickGrid, ...).
- Build upon **chart examples** from Mike Bostock, Jason Davis, Kai Chang, Jonas Petersson, Davo Galavotti.
- Set up as a **stand alone HTML file** (all libraries included) loading data from a **secured network folder**.

Visualization Tool: User Interface

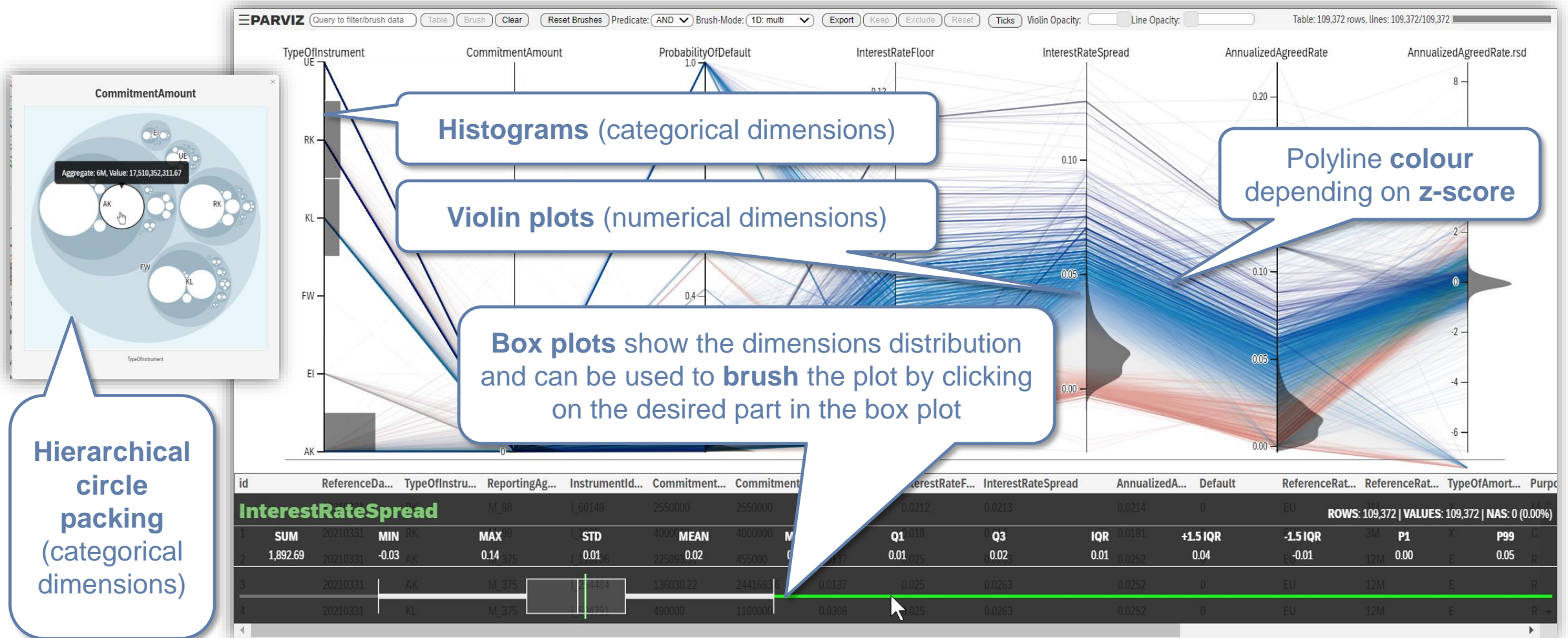


Visualization Tool: Data Queries

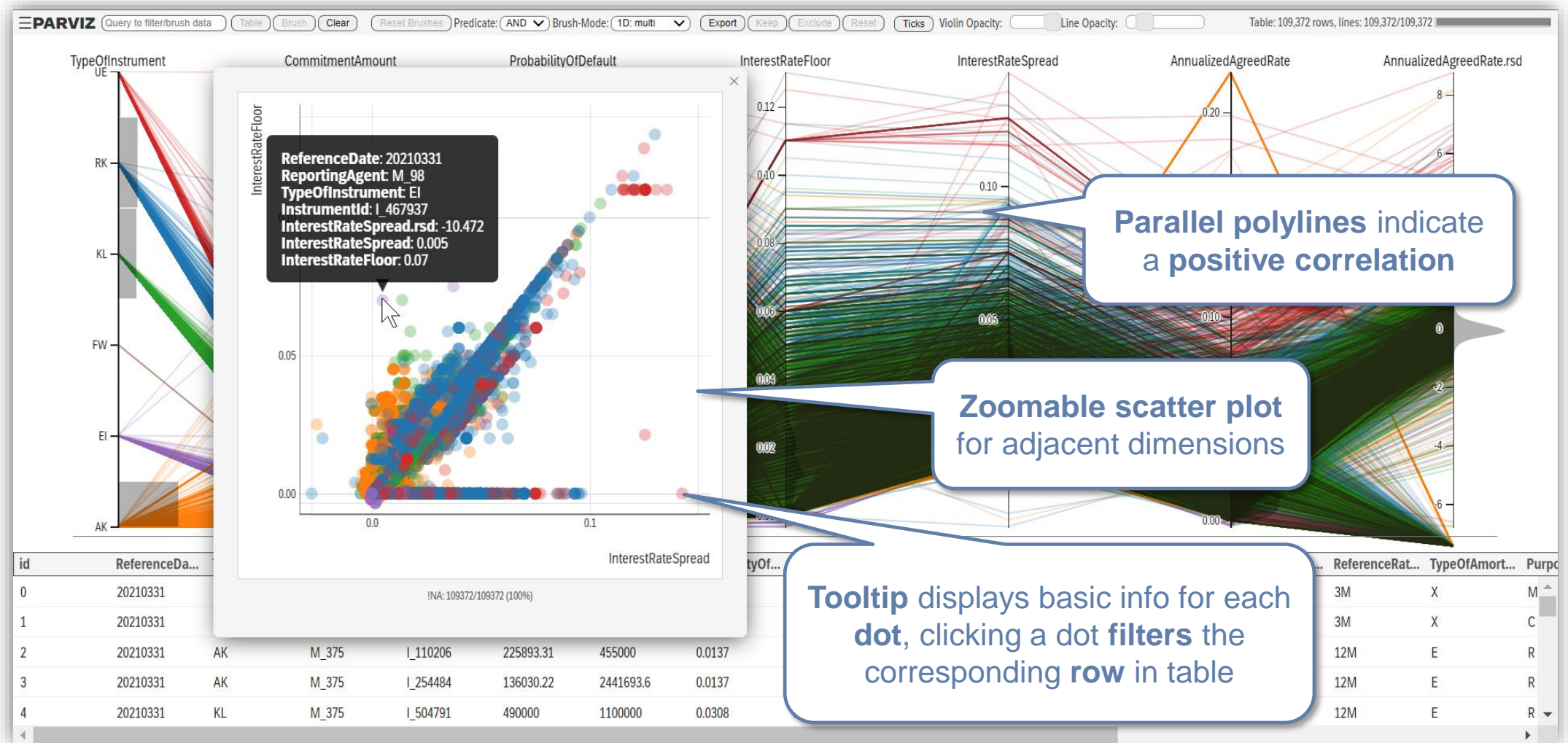
Polylines can be selected („brushed“) in the **plot** with different **brush-modes** or via an **input box** for **precise queries**.



Visualization Tool: Distributions

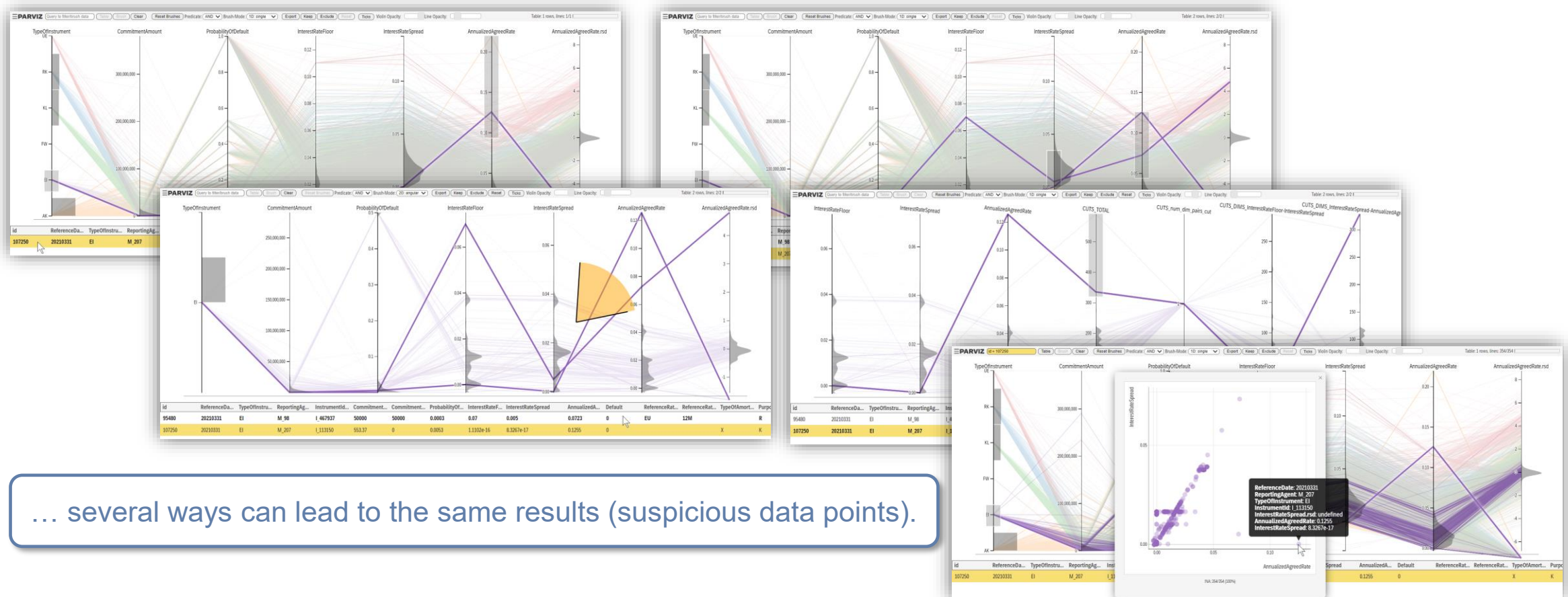


Visualization Tool: Correlations



Visualization Tool: Outlier Detection (example)

Features can be used independently or combined to find potential outliers...



... several ways can lead to the same results (suspicious data points).

Conclusion

- ✓ Interactive **parallel coordinate plots** are suitable to visualize **highly dimensional granular datasets** like the GCR.
- ✓ Various interactive visualizations for correlations and distributions help to quickly **grasp large datasets**, enabling users to **effectively identify potential data quality issues**.
- ✓ Visualizations improve the **understanding of complex statistical outlier detection methods**.
- ✓ Readily available and well documented **web development techniques** and chart examples **facilitate the development of tailor-made solutions**.

Danke für Ihre Aufmerksamkeit

Thank you for your attention

www.oenb.at

oenb.info@oenb.at

 [@oenb](https://twitter.com/oenb)

 [@nationalbank_oesterreich](https://www.instagram.com/nationalbank_oesterreich)

 [OeNB](https://www.youtube.com/OeNB)

 [Oesterreichische Nationalbank](https://www.linkedin.com/company/oesterreichische-nationalbank)



IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Spot the flaw – using power BI for quality control: an application to non-financial corporations’ data¹

José Alexandre Neves, Tiago Pinho Pereira and Ana Bárbara Pinto,
Banco de Portugal

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Spot the flaw – Using Power BI for quality control: an application to non-financial corporations' data

Ana Bárbara Pinto | Banco de Portugal

José Alexandre Neves | Banco de Portugal

Tiago Pinho Pereira | Banco de Portugal

Abstract

This paper shows the experience of the Central Balance Sheet Data Office (CBSO) of Banco de Portugal with Power BI dashboards, which represent a powerful tool to perform and monitor the quality control of data. CBSO databases contain quarterly data from the balance sheet and the income statement of firms and yearly data from all the financial statements and notes. Linking Power BI to CBSO databases allows to execute deep analyses and instantly through the dashboards.

As a powerful data visualisation tool, Power BI gives the opportunity to get information directly from the charts and drilling down data. This is of utmost importance in terms of quality control as it allows to quickly identify the situations that need to be tackled urgently to improve quality in non-financial corporations' statistics.

Keywords: firm-level databases; non-financial corporations; quality control; data visualisation; Power BI

JEL classification: C81; C88

Acknowledgements: We would like to thank Paula Casimiro and Nuno Azevedo for their valuable suggestions and comments, Duarte Santos for providing information and training on Power BI and Ana Teresa Oliveira and Joana Gonçalves for developing some of the dashboards presented in this article. The views expressed are those of the authors and do not reflect those of the Banco de Portugal or the Eurosystem.

Contents

| | |
|---|----|
| 1. Introduction..... | 3 |
| 2. Power BI for quality control purposes at the CBSO of Banco de Portugal | 3 |
| 3. Conclusion | 13 |
| References | 14 |

1. Introduction

We are in the golden era of data visualisation. Translating data into meaningful and appealing charts is essential to not only understand and communicate statistics better, but also to manage organisations.

Non-financial corporations' data available at Banco de Portugal come from two different main sources. The annual survey (IES, in the Portuguese acronym, standing for Informação Empresarial Simplificada), which is mandatory and covers all the non-financial corporations (NFCs) in Portugal, and the quarterly survey to NFCs (ITENF, in the Portuguese acronym, standing for Inquérito Trimestral às Empresas Não Financeiras), which is a sample-based survey. ITENF only contains a summarised balance sheet and income statement, while IES provides complete financial statements comprising the balance sheet, the income statement, the cash flow statement, the statement of changes in equity and detailed quantitative and qualitative data presented as notes to financial statements. Once at Banco de Portugal, data is subject to automatic and manual quality control procedures.

To perform and measure the quality control at the Central Balance Sheet Data Office (CBSO) and assess the quality control that remains to be done at any given moment, several Power BI dashboards were created. Setting up the data connections is easy, refreshing them is not demanding and it is possible to drill down the data by clicking in the visual and quickly check which firms are responsible for the result that is being shown, which is fundamental to boost and redefine the target of the manual validation, if it is necessary. Furthermore, because it is a visualisation tool, we can share the dashboards and deliver a presentation of the most relevant issues with no additional effort. In fact, Power BI dashboards allow performing, measuring and managing quality control and this is key.

The remainder of this work is as follows: Section 2 briefly describes the quality control of data received by the CBSO of Banco de Portugal and shows how Power BI is used for the quality control of CBSO data. Section 3 concludes.

2. Power BI for quality control purposes at the CBSO of Banco de Portugal

The Central Balance Sheet Database (CBSD) of Banco de Portugal is an economic and financial database on Portuguese NFCs. Its information is mostly based on annual and quarterly accounting data.

Annual data is obtained from the annual accounts of corporations reported under IES (Simplified Corporate Information – SCI in the English acronym), while quarterly data is drawn from the quarterly survey to NFCs (ITENF).

IES is a single electronic submission of accounting and fiscal information that meets the statistical needs of Banco de Portugal and the National Statistics Institute and is used by the Portuguese Ministries of Finance and Justice for fiscal purposes and legal deposit of accounts (Banco de Portugal, 2008). It comprises not only the detailed financial statements (balance sheet, income statement, statement of changes in equity and cash flow statement), but also the notes to them. It is mandatory and covers virtually all companies in Portugal (more than 450 000). ITENF, on the other hand, only includes a sample of NFCs (more than 4 500) and it covers a summary of the balance sheet and income statement.

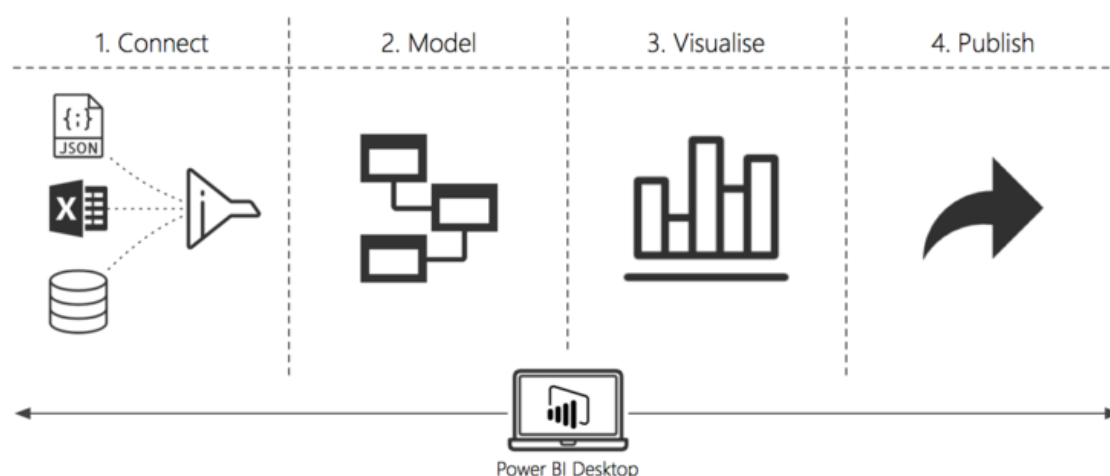
Both IES and ITENF are subject to quality control at the CBSO and Power BI is a key tool to perform and monitor quality control, given that it can rapidly show what is already accomplished, what has to be done and what are the priorities until the end of the process.

For each firm, quality control is performed to ensure that information is horizontally (across different years) and vertically (within financial statements) consistent¹. During quality control, data is usually compared with other internal and external data sources (Casimiro et al., 2017), and it is also possible to query NFCs for additional explanations. Automatic validation procedures are implemented and, for quality control purposes, only a sample of firms from IES and ITENF that violate the consistency checks implemented² are selected for manual validation.

Data from ITENF is received within 2/3 months after the end of the reference quarter, while IES is due 6.5 months after the end of the fiscal year, which corresponds to July 15 for the majority of companies resident in Portugal. Usually, around 700 NFCs are selected every quarter for manual quality control of ITENF and the CBSO staff validates all of them. Regarding IES, around 4 000 to 6 000 NFCs are selected for manual quality control. To validate them, the CBSO hires a group of undergraduate students³ from July to September, given that the period in which IES is available for quality control matches the students' summer break. It is a win-win situation as students develop their financial analysis skills and the CBSO assures the validation of a higher number of firms in a limited timeframe, that otherwise would not be possible.

Power BI allows connecting to data, transforming and cleaning them, as well as creating interactive and rich dashboards and reports. Premium versions of Power BI also enable publishing the dashboards (in websites, for example) and scheduling automatic updates (Figure 1). In the case of the CBSO of Banco de Portugal, Power BI dashboards are published in the integrated information system of the CBSO, which is an internal website that allows the analysis of individual firms' data and the insertion of any manual correction that is deemed necessary. Everyone working at Banco de Portugal without a Power BI software, but with an internet connection can access the dashboards, which is an advantage.

Figure 1: The life cycle of a dashboard created through Power BI



Power BI is a product from Microsoft and, thus, resembles Microsoft Excel very much. Reports may have several sheets and the way variables are selected to build the visuals is similar to the way pivot tables are built. You can add filters and switch from rows to columns to customise the visual.

Figure 2 displays a finished dashboard, designed for the quality control of annual data. Above, one can pick the year, size, institutional sector and economic activity of the firms and the charts will update accordingly, showing the validation and reporting status of the firms, with light green being the firms

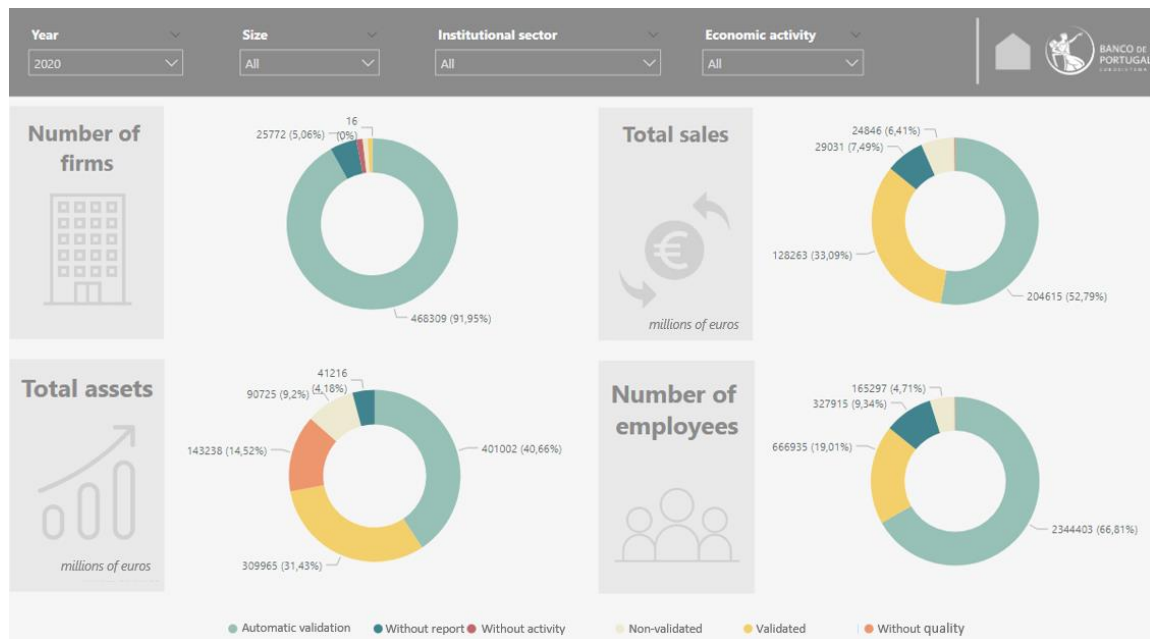
¹ For more details regarding the quality control that is performed at the CBSO of Banco de Portugal, please see Brites (2013)

² Some of the consistency checks implemented include absolute and relative annual/quarterly changes above a given threshold and differences between data submitted by NFCs and internal and/or external data sources above a given threshold.

³ In the area of economics, management, finance and accounting

that have been validated automatically, yellow the firms validated manually, beige the firms awaiting validation and dark green the firms that haven't reported yet. By changing the size and economic activity one can assess the firms that are manually validated or waiting validation for a subset of the universe, which constitutes relevant information for quality control. Power BI drilldown features also allow us to see the most relevant companies without report, which is useful so we can contact them in order to submit the missing information, or the most relevant companies that are awaiting validation in order to give them priority, if we are close to our deadlines and need to check more relevant firms first.

Figure 2: Status of validation and reporting of firms by number, total sales, total assets and number of employees



When one needs to focus in a certain timeframe, the range slicer, which is represented in Figure 3, referring to the number of manually validated firms is a very helpful visual, too. One can choose a specific year and, in addition, move the Days' range slicer to concentrate the analysis in the first days of the quality control, for example. This is particularly useful given that the analysts could be interested in a specific period instead of the whole period. At the beginning of the quality control process, the pace is slower, increasing over time as undergraduate students become more experienced and autonomous. It is always very important to bear this pace in mind to assess the level of manual validation that we can achieve at our deadlines and if we need to change the analysis strategy.

Figure 3: Number of validated firms

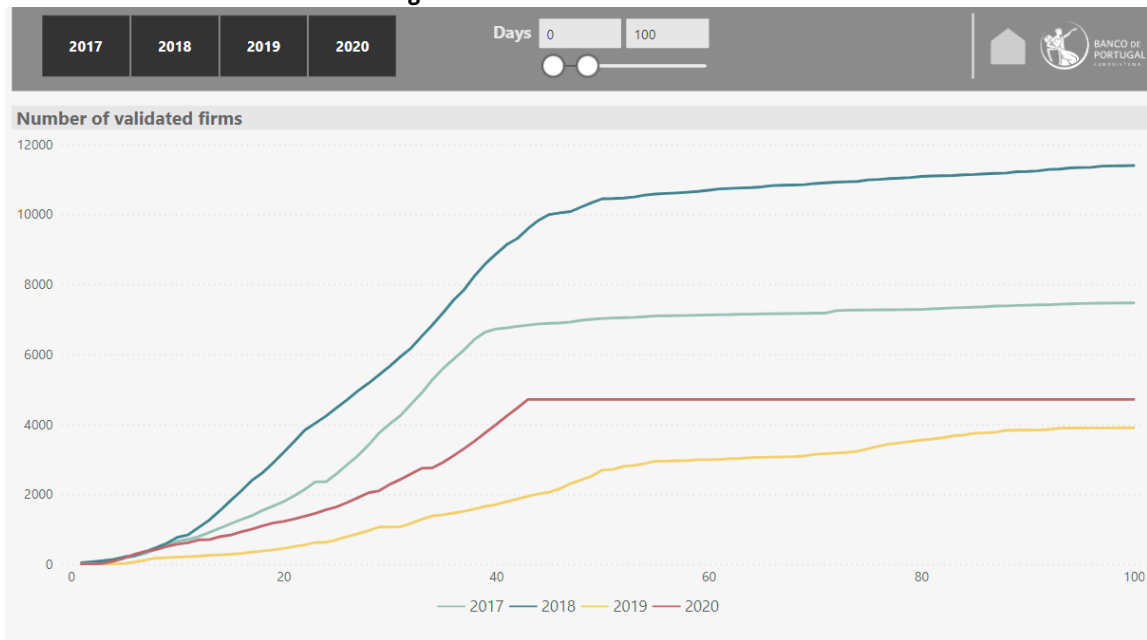


Figure 4 provides an overview of the validation of the quarterly survey of 2021Q1. This picture was taken before the ending of the quality control, which explains the decreases observed in the charts in the last period (2021Q1). The decrease is the first impression one has when looking to the dashboard and this is why visualisation tools like Power BI are very good to quickly identify the state of art at any given moment. The upper three charts show the difference between the total sample of ITENF (in blue) and the answers actually received from firms (in white), while the lower three charts focus on the manual validation of the firms that are selected through the consistency checks. The dashboard clearly shows that the validation level was still below in comparison to the previous quarters and that information is useful to give an idea of how far away of the goals one is and assign more resources to the task, if needed.

Figure 4: Overview of the quarterly survey validation

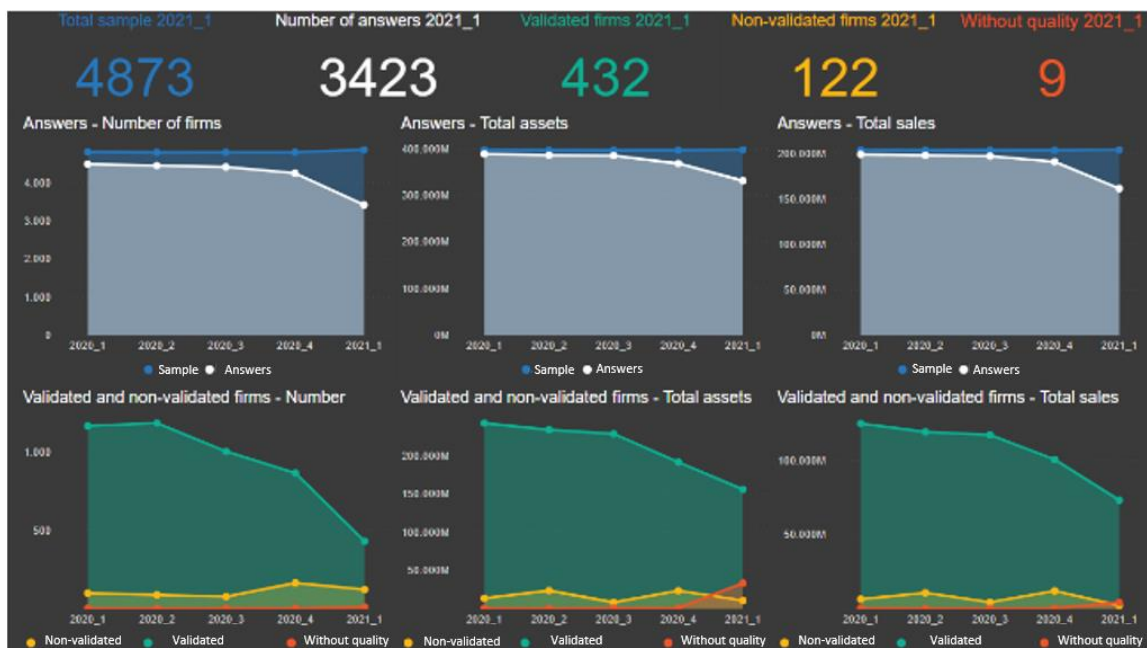


Figure 5 shows the percentage of validated firms by consistency check (in green) for the quarterly survey of 2021Q1. In yellow, it is displayed the portion of firms not validated yet. At the beginning of the quality control, the yellow portion of the bar prevails, of course, but the evolution of the green portion gives important information for monitoring the overall validation status. In order to meet the expected deadlines, this information is crucial to management. If the yellow portion is significant enough close to the deadlines, it is time to assign more resources to the task. Each bar represents a consistency check. As stressed before, the consistency checks refer to changes regarding the previous periods above a given threshold or to differences relative to other data sources above a given threshold, but also to coherency issues across the survey.

For example, if NFCs report meaningful sales or purchases to non-resident counterparts, it is expected that they will have trade credits respecting non-resident counterparts, unless we have a note indicating that the company actually does not have trade credits with non-resident entities. If it is not the case and firms do not report any value in trade credits neither there is a note explaining that, they violate the consistency check *External Trade Credits* and consequently are selected for manual validation. Another example of coherency issues across the survey is consistency check *Cost of Debt*. If a firm reports significant values in loans, it is very likely that it pays interest. If not, it is something to check.

Besides presenting the portion of validated and non-validated firms, Power BI is flexible enough to drill down data and show which NFCs are already validated (and those whose validation is missing). To see more details on the firms not validated yet (e.g., which ones and the username of the person in charge of the validation), it is only needed to click in the right button of the mouse above the yellow part of the bar and choose the option "Show data point as a table". This is valid for every dashboard in Power BI. It is always possible to click in the charts and access the firms that are contributing to that chart.

Figure 5: Validated firms by consistency check - quarterly survey



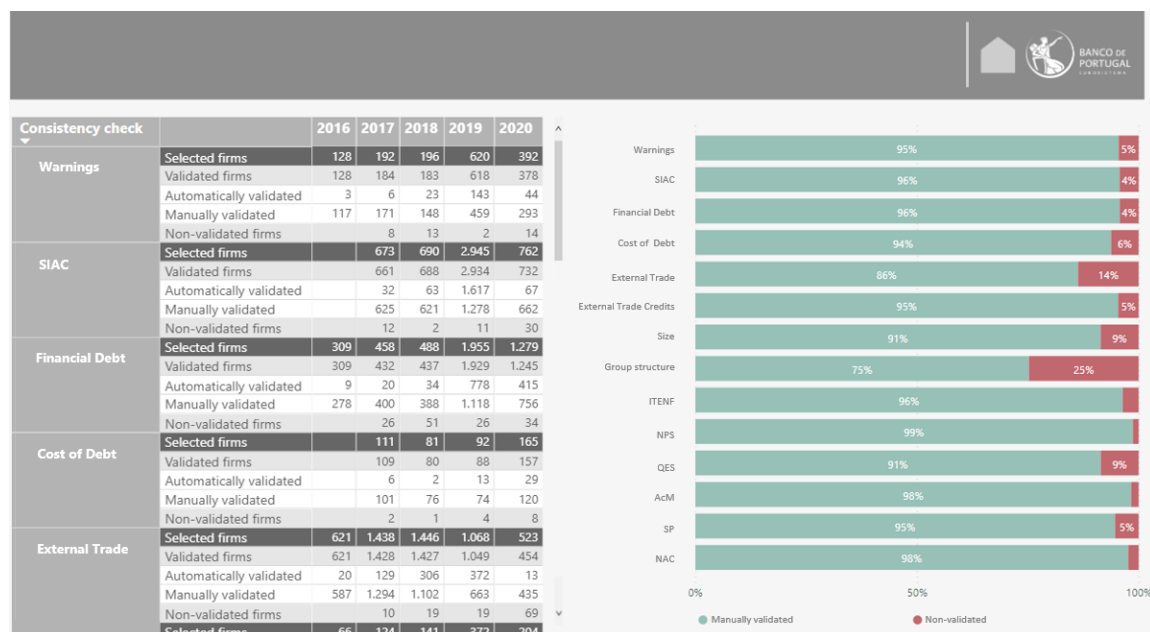
The quality control framework of the annual survey (IES) is similar to the framework of the quarterly survey (ITENF). However, given that the data collected by IES is more comprehensive, the timeframe to perform the quality control is larger and there are more human resources available, namely undergraduate students, as mentioned earlier, the number of validated companies is higher and the consistency checks to analyse are broader. In addition, because IES is submitted after the firms' annual accounts are approved and is available for the universe of Portuguese NFCs, it also contributes to

complete and correct data from ITENF and is essential to depict the performance of the NFCs sector as a whole.

Figure 6 shows the fraction of firms by consistency check already validated. The template is similar to the quarterly survey and, although the number of consistency checks is approximately the same as in Figure 5, their scope is broader. For example, in the quarterly survey, changes in the size class or in the group structure of the firms are not evaluated, while in the annual survey they are (consistency checks *Size* and *Group structure*, respectively).

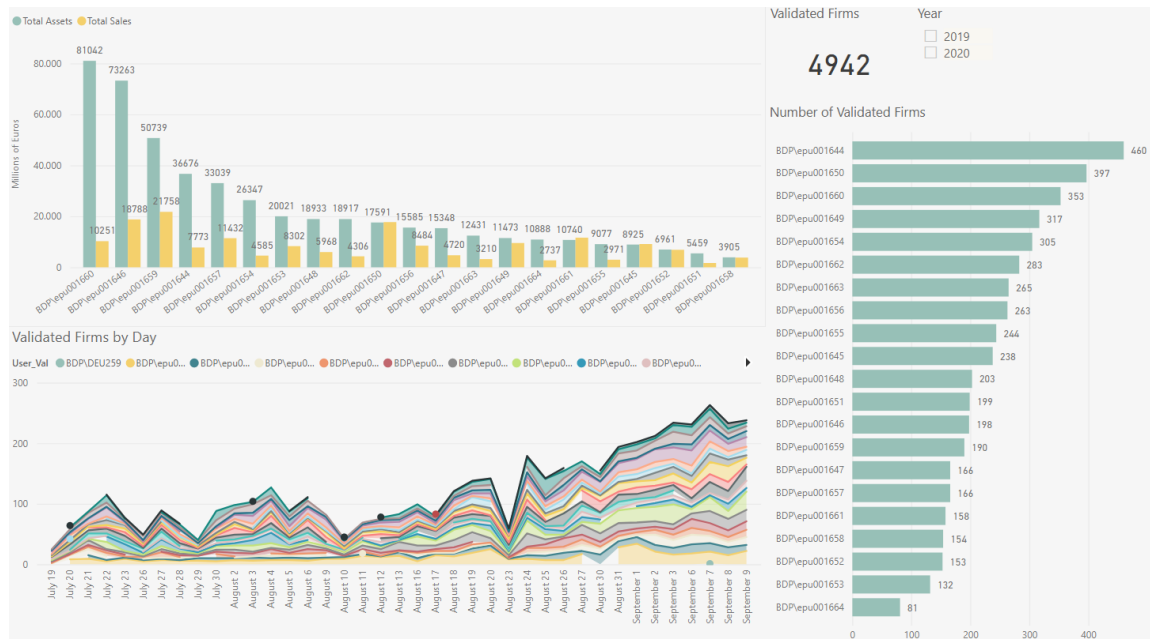
Again, it is possible to drill down to the firms that are behind each one of the consistency checks. This turns easier and faster to implement any change of priority that is needed during the quality control process.

Figure 6: Validated firms by consistency check – annual survey



The number of validated firms by user is shown in Figure 7 and gives very important information to management. This dashboard presents the total number of validated firms by day and user and it enables assessing whether performance is improving or not with time. It is expected that the pace of validation by the undergraduate students increases over time, as they gain more experience and knowledge and become more autonomous. One can see that the number of validated firms by day reaches a maximum in the first days of September, one and half months after the beginning of their work. Decreases in the number of validated firms by day are usually related to the assignment of more (or more complex) firms for validation. When a bar of a given user is selected, only information for that user is shown. With this dashboard we can early detect some issue with any student, and thus provide additional individual training in a timely manner in order to improve her/his contribution, as well as to follow closely the daily pace of validation as a whole.

Figure 7: Validated firms by user and details – annual survey

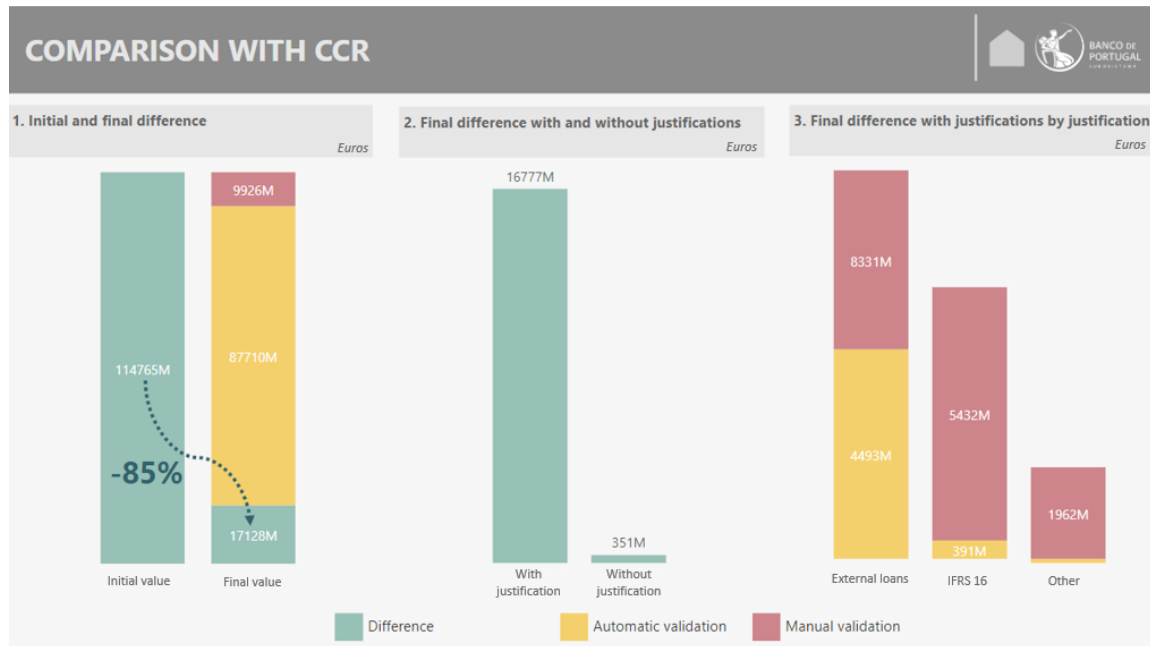


Comparison with other sources of information besides IES or ITENF is also an important part of quality control procedures. Of course, information can differ between data sources, but it should be consistent.

Figure 8 shows the difference between bank loans that are reported by firms through IES and the bank loans that are available from the Portuguese Central Credit Register (CCR) managed by Banco de Portugal. At the beginning of the quality control, the initial difference between sources is large, but as quality control evolves, the difference decreases and, at the end of the quality control, only 15% of the initial differences (green bar in Chart 1. *Initial and final difference*) remain. The majority of the differences are eliminated automatically by our algorithms (yellow bar in Chart 1. *Initial and final difference*) and a lower fraction is solved manually (red bar in Chart 1. *Initial and final difference*). Bank loans are overvalued in IES because many firms tend to misreport all their financial debt as bank loans. Once at Banco de Portugal, data are automatically or manually corrected. Most of the differences that remain have a justification (Chart 2. *Final difference with and without justifications*), with the most frequent being the existence of external bank loans, which are not covered by the CCR, and the adoption of the IFRS 16⁴ (Chart 3. *Final difference with justification by justification*), also not reflected in the CCR.

⁴ Affects data from 2019 onwards.

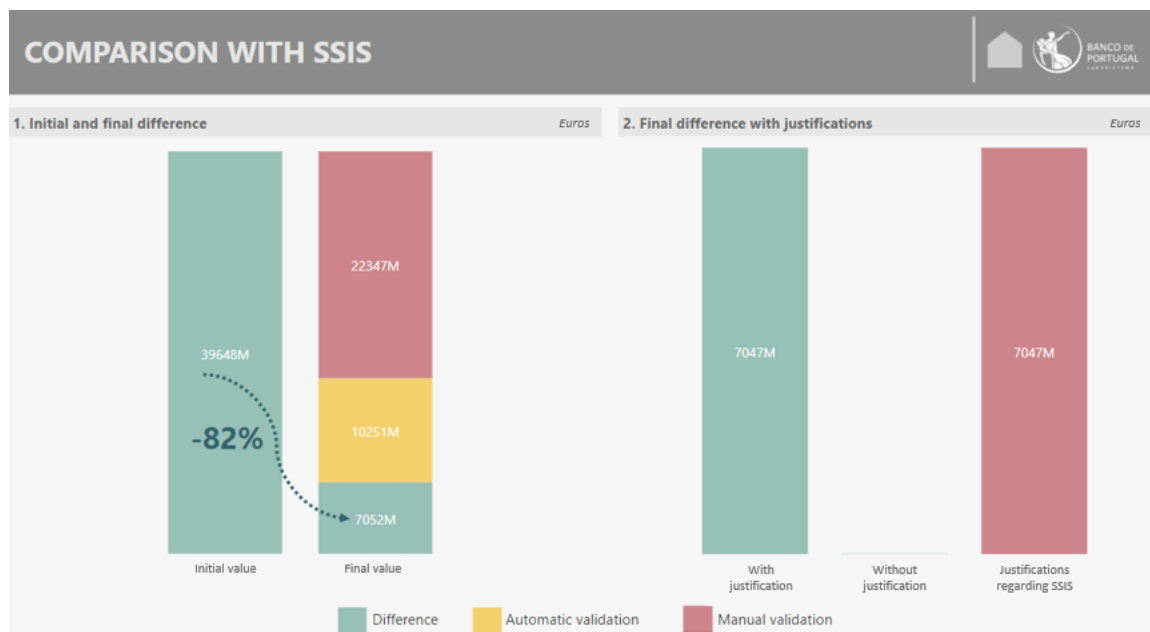
Figure 8: Comparison between CBSD (annual survey) and CCR



This dashboard is regularly updated at the time of the quality control process and allows monitoring the evolution of the differences regarding other sources. Given that one of the aims of the quality control process is to minimise differences regarding other sources, it is extremely helpful. Again, drilling down data is possible and one can check the firms with larger differences and quickly address this issue.

Differences between the amount of debt securities reported through IES and the amount of debt securities issued, which is available from the Securities Statistics Integrated System (SSIS) of Banco de Portugal, are also treated during the quality control process and the way they are monitored is similar to the comparison with the CCR (Figure 9).

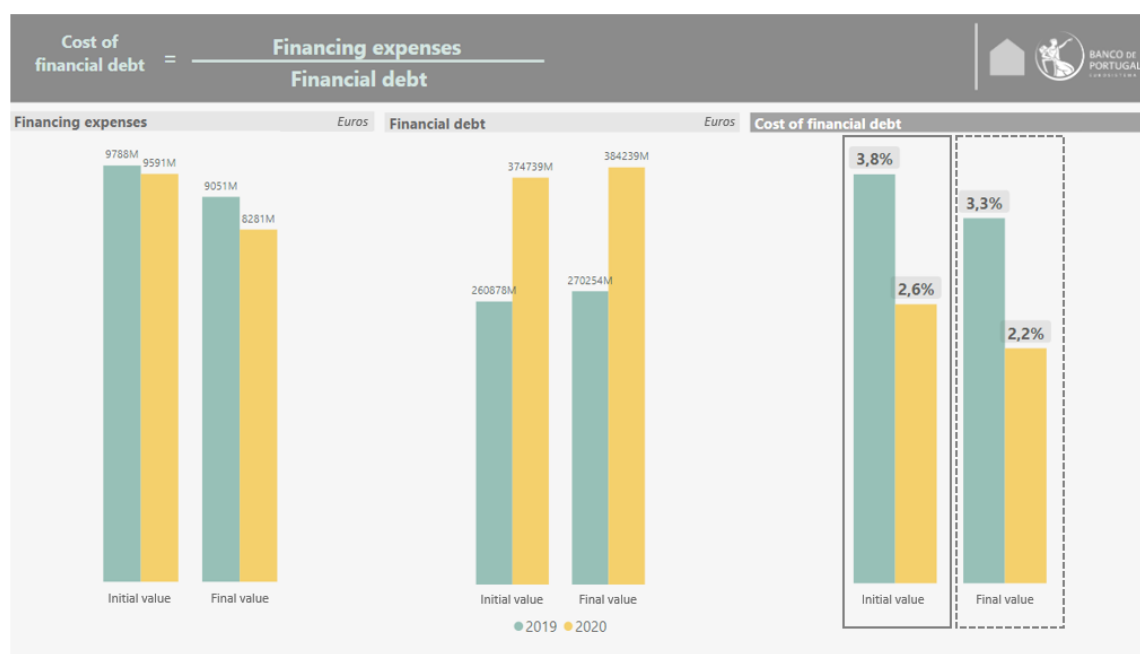
Figure 9: Comparison between CBSD (annual survey) and SSIS



Again, with this visual we can easily see that the initial difference between the two data sources decreases substantially by the end of the quality control process and that any remaining differences are explained and accounted for.

Another kind of analysis that we perform through Power BI is selecting some economic and financial indicators and monitor their evolution vis-à-vis the previous period in order to detect abnormal values. Figure 10 presents the cost of financial debt at the beginning of the quality control (bars on the left, inside the solid line rectangle, in the Chart *Cost of financial debt*) and at the end of the quality control (bars on the right, inside the dashed line rectangle, in the Chart *Cost of financial debt*). At the beginning of the quality control, the cost of financial debt is usually higher because firms incorrectly report expenses other than financing expenses as financing expenses.

Figure 10: Evolution of the cost of financial debt during the quality control process



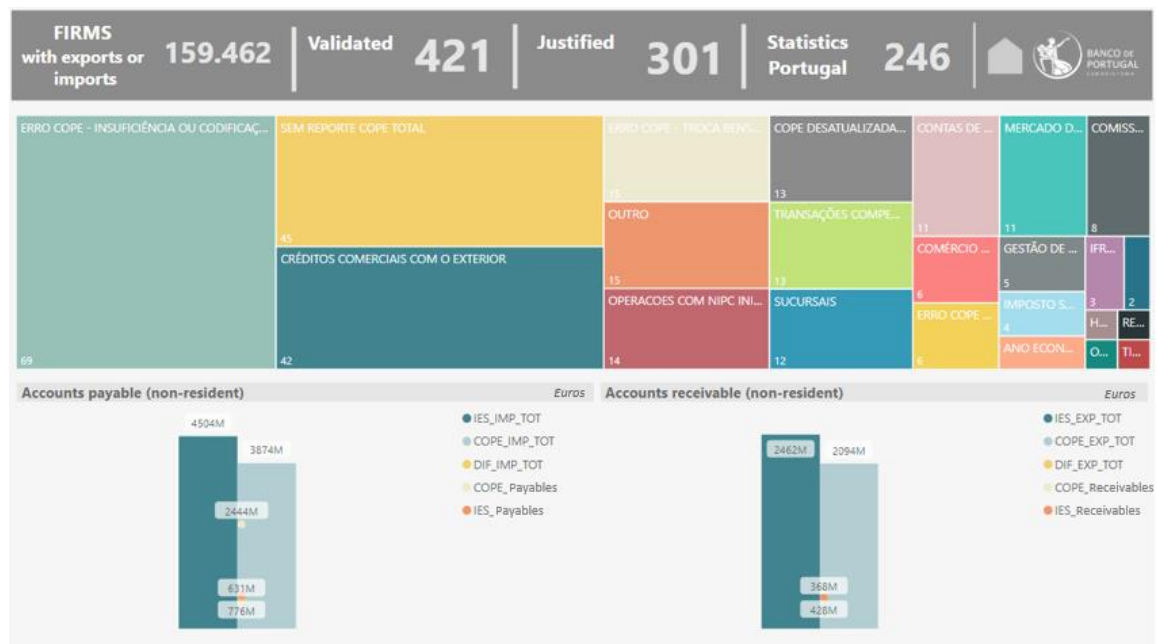
With this dashboard and the possibility of drilling down data, we can easily verify the firms that are influencing the cost of financial debt in a larger extent and correct them if data proves to be inaccurate. Conversely, we can also verify which firms had the larger corrections and double check these adjustments, which is especially important during the work of the undergraduate students.

Besides comparing data from IES with the CCR and the SSIS, exports and imports of goods and services communicated are also compared with Balance of Payments (BoP) data collected by Banco de Portugal. As in the case of CCR and the SSIS, differences arise and one should eliminate or justify them during the quality control process. Figure 11 presents a dashboard that is used to monitor the quality control regarding exports and imports of goods and services, namely the difference between the amounts reported in IES and those available from BoP that violate the consistency checks. The treemap shows the relevance of each justification for that difference. For example, from the 301 firms whose difference between data sources was justified, 69 are attributable to mistakes made by firms when reporting the BoP survey (the so-called COPE report⁵, as mentioned in Casimiro et al., 2017) and 45 to the absence of the COPE report for those firms. BoP figures not only contain exports and imports of

⁵ COPE is the Portuguese acronym for *Comunicação de Operações e Posições com o Exterior*. It is the main source of information for BoP statistics compiled by Banco de Portugal and it is a monthly report where entities in Portugal communicate their transactions and outstanding positions with the rest of the world.

goods and services, but also accounts payable and accounts receivable regarding non-resident counterparts and these data are also used for the quality control process at the CBSO.

Figure 11: Differences between IES and BoP in exports and imports of goods and services



In the course of the quality control of IES, information on the business group structure of firms is also analysed and corrected. In particular, data on the ultimate controlling institutional units (UCI) of NFCs and their location. To examine this issue, we use maps. Instead of presenting statistics for countries in bar charts, it is more appropriate showing a map. They are perfect to communicate when geography and space are important. Figure 12 presents the country of the UCI of firms in Portugal, split by categories, where dark orange represents the countries that control most of firms and light orange those that control the little. This visual rapidly shows that most of the firms in Portugal are controlled by domestic entities, followed by Spanish entities. Dark orange tones in countries that we don't expect to have such large influence or light orange tones in countries that we expect to have large influence represent a quick warning for quality control.

Figure 12: Location of ultimate controlling institutions, by number of firms controlled



3. Conclusion

This paper provides some examples to describe the experience of Banco de Portugal with Power BI for monitoring the quality control of quarterly and yearly accounting data at the CBSO.

Data from the CBSO of Banco de Portugal mostly relies on two sources: the quarterly survey (ITENF, in the Portuguese acronym) and the annual survey (IES, in the Portuguese acronym), which are subject to quality control. To assure the coherence and the consistency of information, data is matched with other data sources managed by the Statistics Department of Banco de Portugal, namely the CCR, the SSIS and the BoP. In fact, the extent of the quality control performed is only possible due to the availability and scope of these micro databases.

In order to assess the magnitude of quality control issues that arise after the implementation of the consistency checks, prioritise them and observe how the work of quality control is evolving, we use Power BI dashboards, which are a powerful visualisation tool that allows connecting quickly to the different data sources. Once the dashboard is created, one only needs to refresh data for evaluating the work in progress, which speeds up decision-making and ignites immediate action over the most urgent matters. Slight changes to visuals and customisation are also easier and faster to perform. Moreover, as a visualisation tool, Power BI not only meets the needs of performing and monitoring the quality control as can be used for delivering presentations. Power BI dashboards are also embedded in our quality control web application, and thus accessible to everyone who performs quality control. Putting all this together, the quality control of data becomes more effective as you can quickly look to the Power BI dashboards and clearly identify where the main problems to be tackled are and how long it takes to solve them. Indeed, Power BI dashboards contributed to fill some data gaps not solved before and to redefine priorities, as well as to decrease time spent doing presentations as they can be presented by themselves.

References

Banco de Portugal (2008), "Simplified reporting: inclusion of the Simplified Corporate Information in the Statistics on Non-Financial Corporations from the Central Balance-Sheet Database", Supplement to the Statistical Bulletin, 1/2008, May 2008.

Available at: <https://www.bportugal.pt/sites/default/files/anexos/pdf-boletim/sup-be-1-2008-en.pdf>

Brites, M. (2013), "Efficient ways of dealing with accounting data from enterprises", Workshop on Integrated Management of Micro-databases, Porto, June 2013

Casimiro, Paula, Pinto, Ana Bárbara and Pereira, Tiago Pinho (2017), Matching firm-level data sources at the Statistics Department of Banco de Portugal, *IFC Bulletin*, 45

Spot the flaw – Using Power BI for quality control: an application to non-financial corporations' data

**IFC Workshop on Data Science in Central Banking:
Applications and tools**

Ana Bárbara Pinto • José Alexandre Neves

• Tiago Pinho Pereira
Statistics Department

February 2022



BANCO DE
PORTUGAL
EUROSYSTEM

QUALITY CONTROL AT THE CBSO OF BANCO DE PORTUGAL

QUARTERLY SURVEY

- ☐ **SAMPLE BASED**
- ☐ **MORE THAN 4 500 FIRMS**
- ☐ **SUMMARISED BALANCE-SHEET AND INCOME STATEMENT**

ANNUAL SURVEY

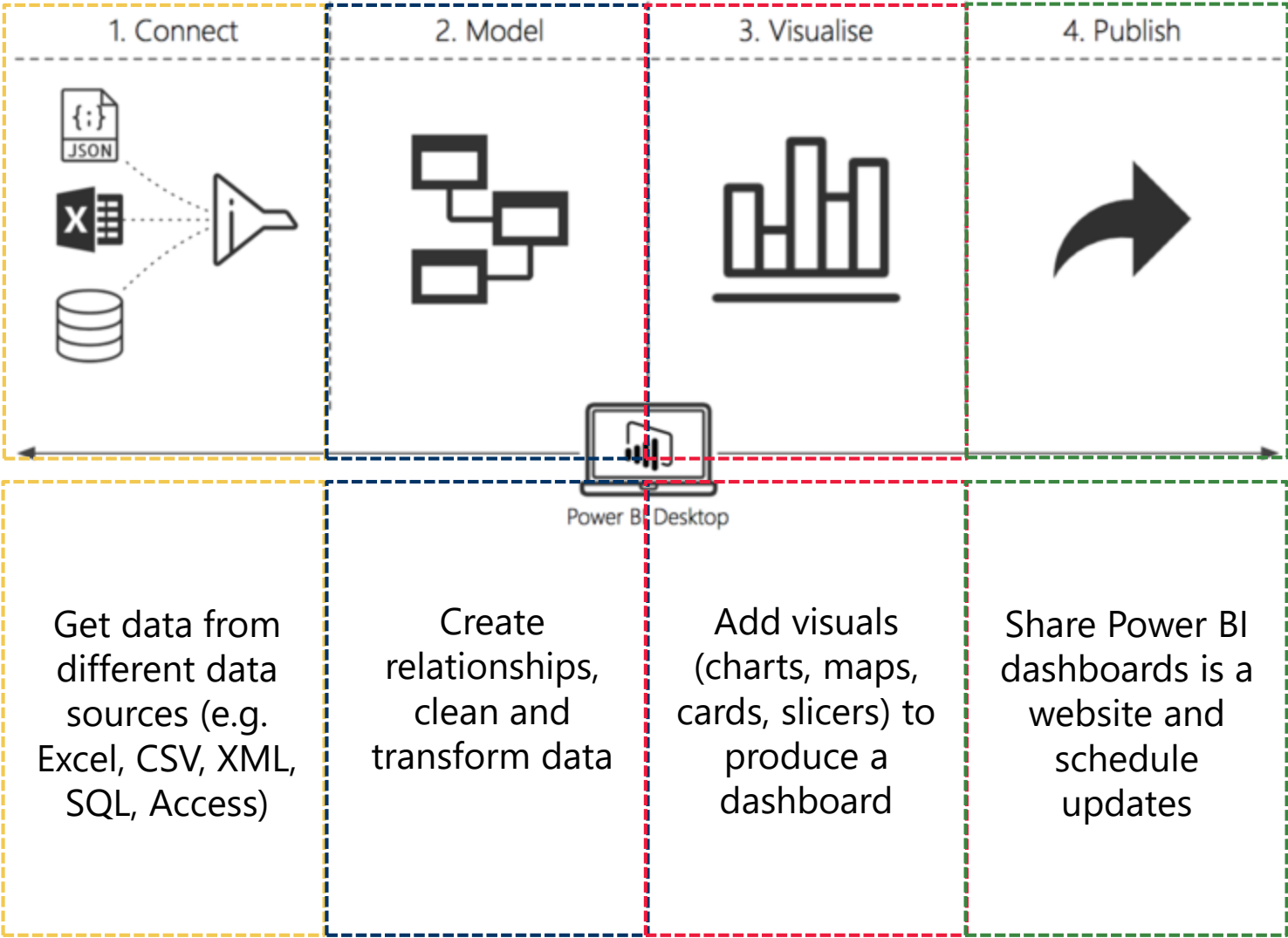
- ☐ **MANDATORY**
- ☐ **MORE THAN 450 000 FIRMS**
- ☐ **DETAILED BALANCE-SHEET AND INCOME STATEMENT**
- ☐ **STATEMENT OF CHANGES IN EQUITY**
- ☐ **CASH FLOW STATEMENT**
- ☐ **NOTES TO THE FINANCIAL STATEMENTS**

QUALITY CONTROL

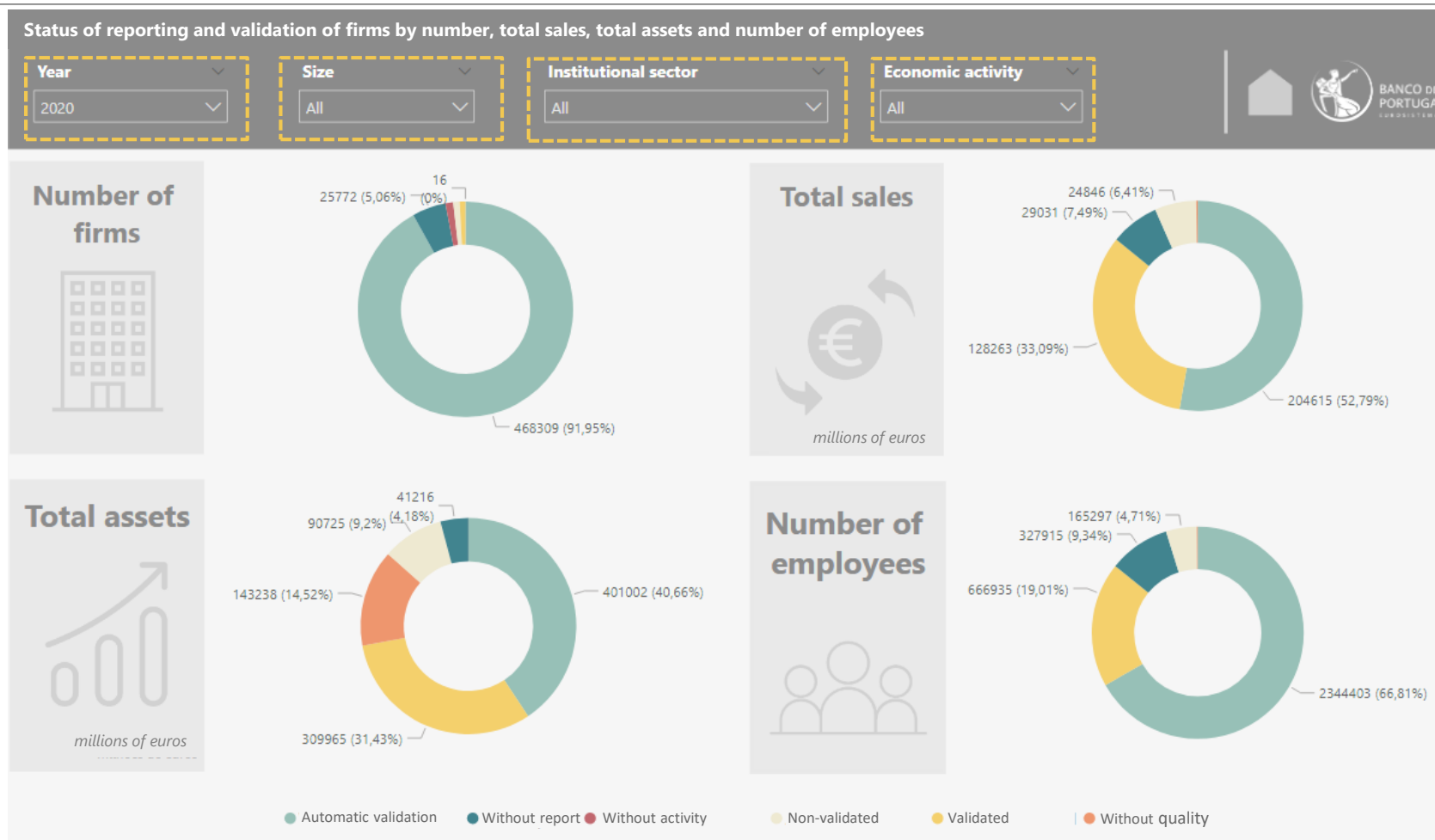
- ☐ **ENSURES THAT INFORMATION IS HORIZONTALLY AND VERTICALLY CONSISTENT**
- ☐ **COMPARISON WITH INTERNAL AND EXTERNAL DATA SOURCES**
- ☐ **AUTOMATIC CORRECTIONS APPLY**
- ☐ **FIRMS WITH LARGE ABSOLUTE OR RELATIVE DIFFERENCES ARE SELECTED FOR MANUAL VALIDATION**



POWER BI FOR QUALITY CONTROL PURPOSES AT THE CBSO OF BANCO DE PORTUGAL



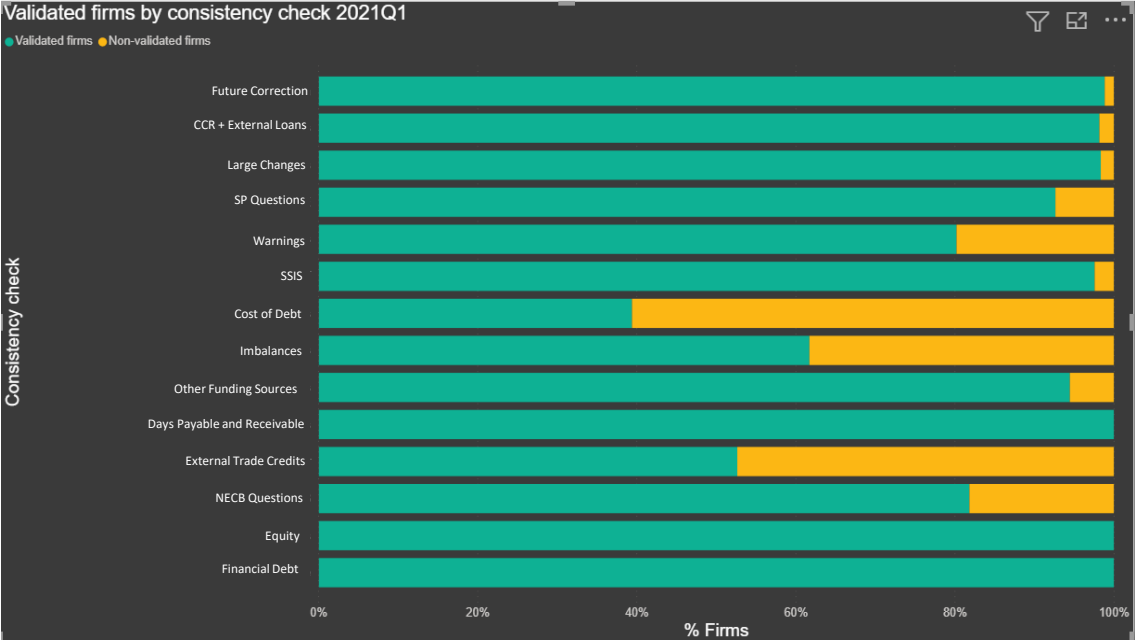
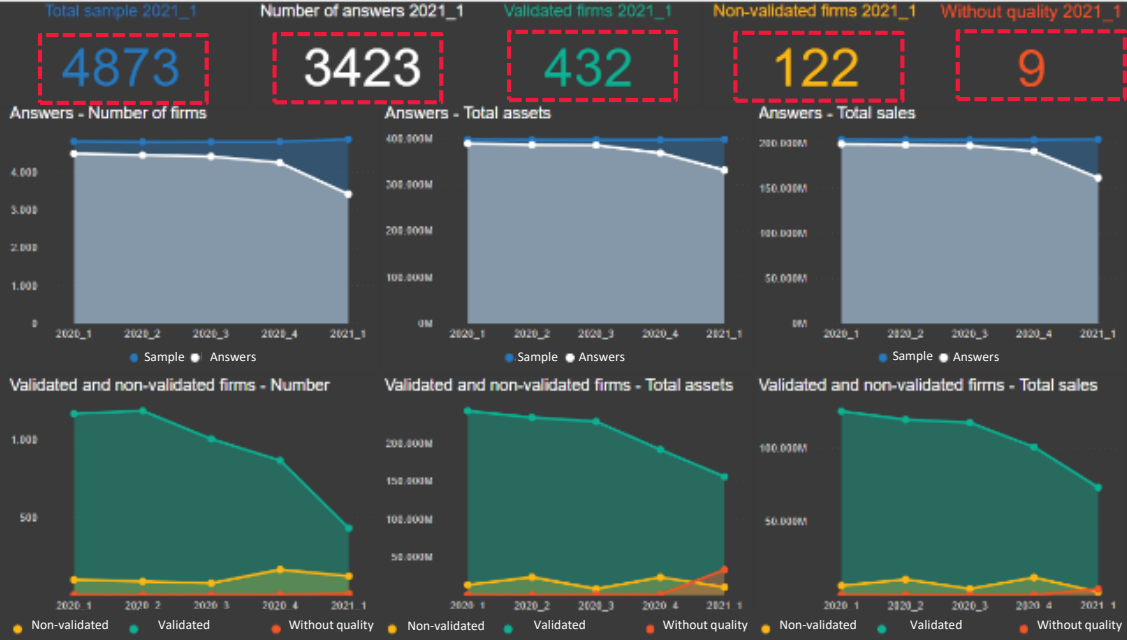
POWER BI FOR QUALITY CONTROL PURPOSES AT THE CBSO OF BANCO DE PORTUGAL



EVALUATE THE STATUS OF VALIDATION AND REPORTING OF FIRMS



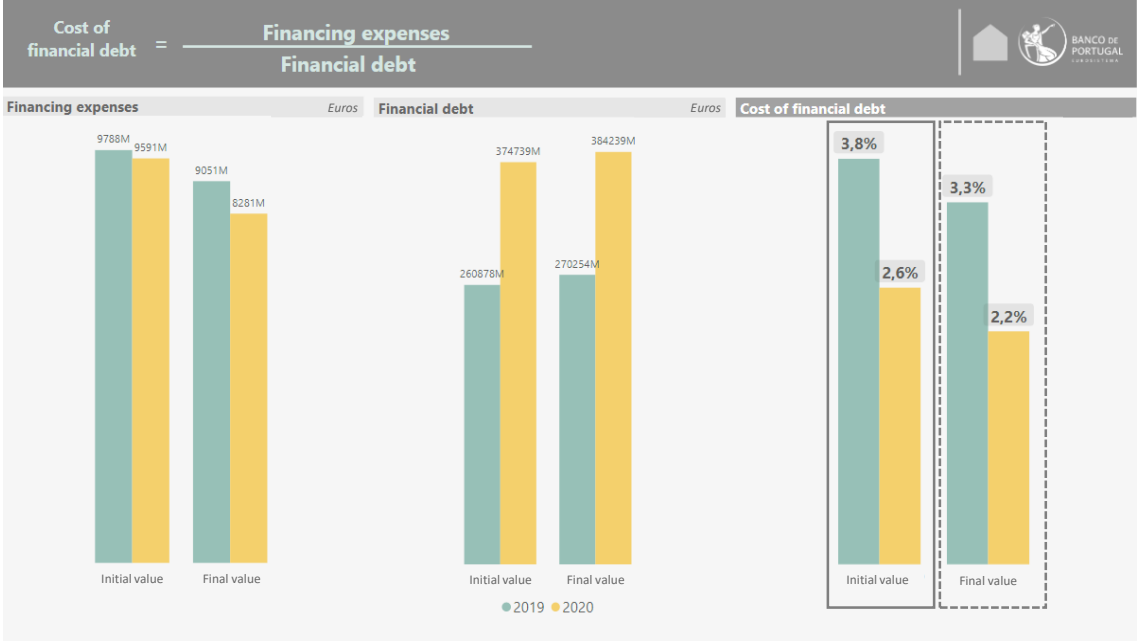
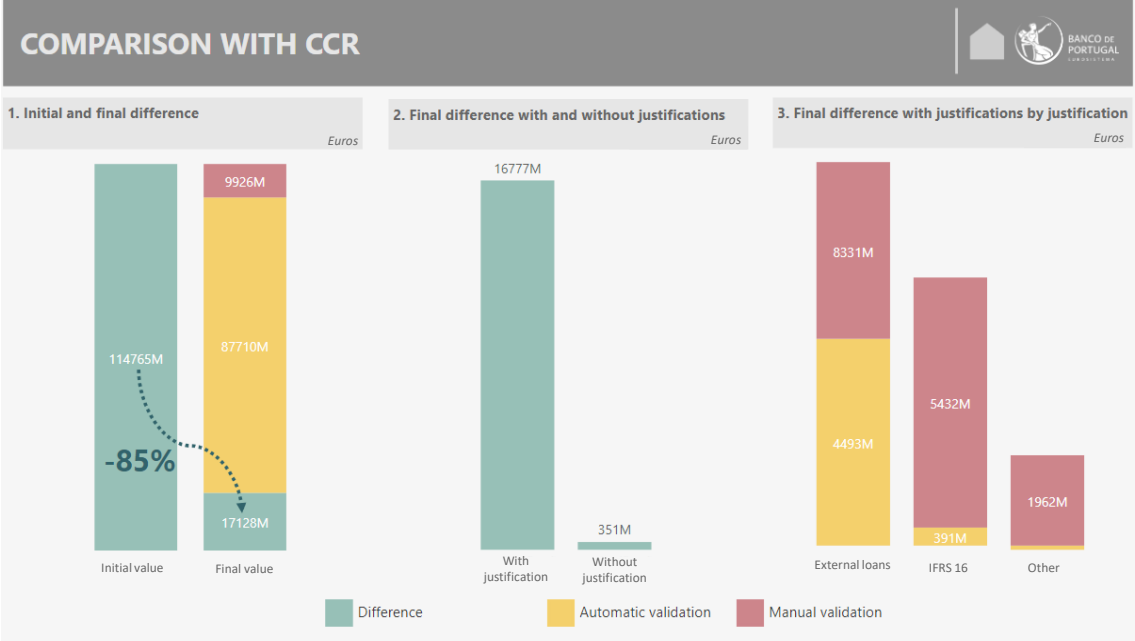
POWER BI FOR QUALITY CONTROL PURPOSES AT THE CBSO OF BANCO DE PORTUGAL



MEASURE THE EVOLUTION OF QUALITY CONTROL



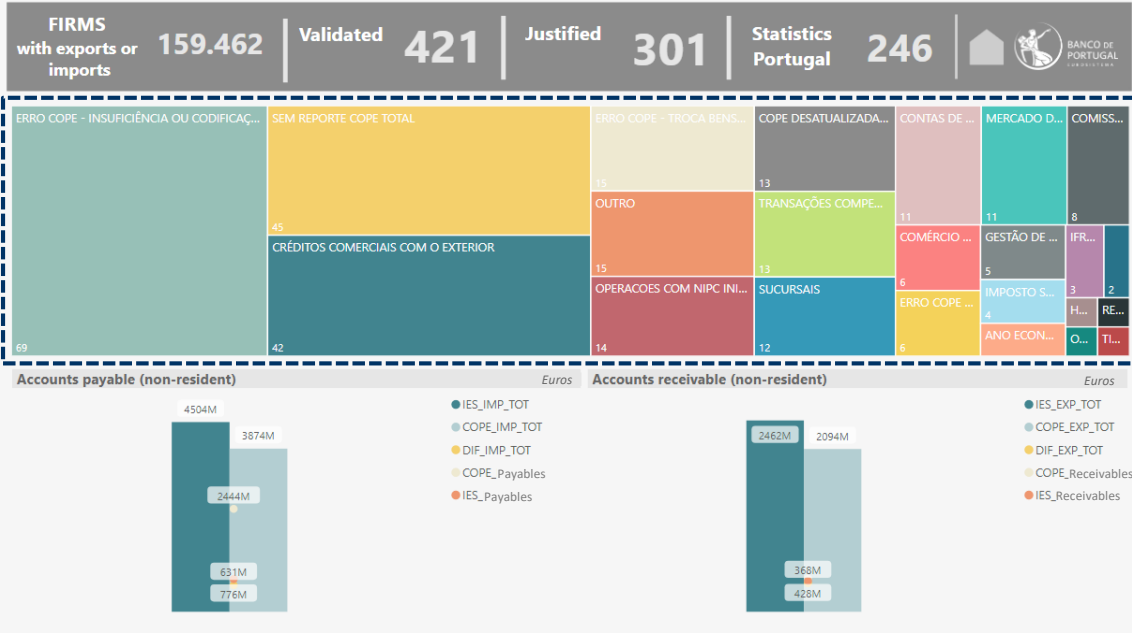
POWER BI FOR QUALITY CONTROL PURPOSES AT THE CBSO OF BANCO DE PORTUGAL



CHECK AND EXPLAIN THE IMPACT OF QUALITY CONTROL ON DATA



POWER BI FOR QUALITY CONTROL PURPOSES AT THE CBSO OF BANCO DE PORTUGAL



HIGHLIGHT RELEVANCE



MAIN CONCLUSIONS

TO SUM UP

- ❑ **ALLOWS CONNECTING QUICKLY TO THE DIFFERENT DATA SOURCES**
- ❑ **EMBEDDED IN OUR QUALITY CONTROL WEB APPLICATION**
- ❑ **DECREASED TIME SPENT DOING PRESENTATIONS AS POWER BI DASHBOARDS CAN BE PRESENTED BY THEMSELVES**
- ❑ **USED TO PERFORM AND MONITOR QUALITY CONTROL OF DATA AT THE CBSO OF BANCO DE PORTUGAL**
 - ✓ **QUICKLY SHOWS TRENDS AND DATA GAPS NOT SOLVED YET**
 - ✓ **HELPS REDEFINING PRIORITIES IN TERMS OF FIRMS AND CONSISTENCY CHECKS**



IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Leveraging the power of visualization for data exploration and insights communication - visual analytics with Tableau¹

Zunaira Rasheed,
BIS

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



Leveraging the power of visualization for data exploration and insights communication

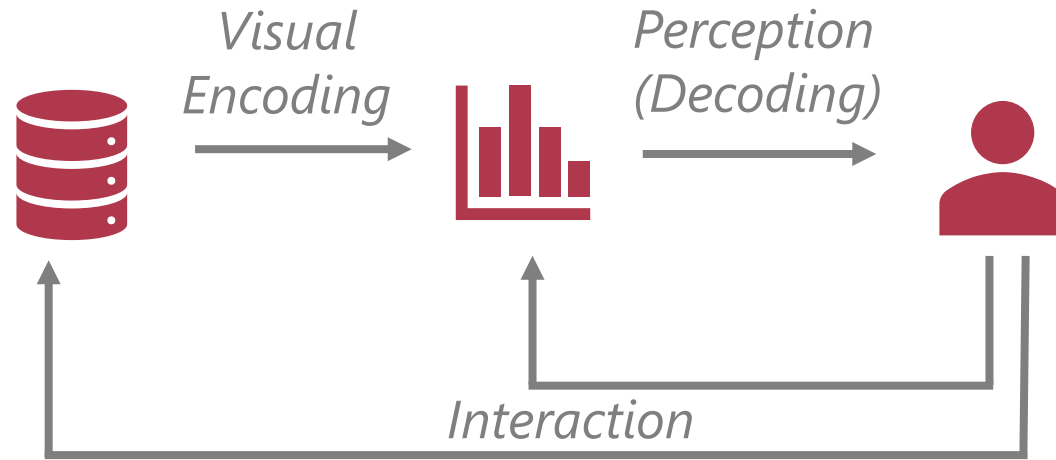
Visual analytics with Tableau

Zunaira Rasheed

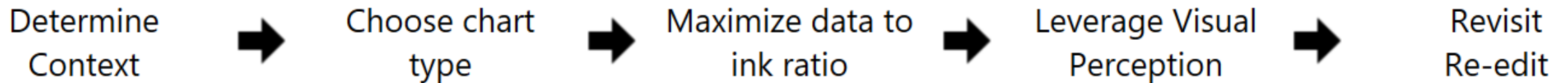
Tableau at BIS

- Tableau is being used for,
 - Business intelligence (Reports/Dashboards/Storyboards)
 - Self-service data discovery and exploration
 - Interactive graphs linked BIS research publications
- Tableau was introduced at the Bank in 2016 and since then adoption increased significantly
- In addition to the platforms, the centralized Data & Analytics also offers,
 - Center of excellence related to Tableau issues
 - Data visualization community platform to engage and discuss
 - Best practices and visual standards for BIS Tableau content
 - Tools and process for managing Tableau governance

Defining Visualization



Our Visualization Process



Determine Context

Who?

What?

How?

Choose chart type

Maximize data to ink ratio

Leverage Visual Perception

Revisit Re-edit

Deviation

Diverging bar
Diverging stacked bar
Spine
Surplus/deficit filled line

Correlation

Scatterplot
Column + line timeline
Connected scatterplot
Bubble
XY heatmap

Ranking

Ordered bar
Ordered column
Ordered proportional symbol
Dot strip plot
Slope
Lollipop
Bump

Flow

Sankey
Waterfall
Chord
Network

Distribution

Histogram
Dot plot
Dot strip plot
Barcode plot
Boxplot
Violin plot
Population pyramid
Cumulative curve
Frequency polygons

Change over Time

Line
Column
Column+ line timeline
Slope
Area chart
Candlestick
Fan chart
Connected scatterplot
Calendar heatmap
Seismogram
Streamgraph

Part-to-whole

Stacked column/bar
Marimekko
Pie
Treemap
Voronoi
Arc
Gridplot
Venn
Waterfall

Magnitude

Column
Bar
Paired column
Paired bar
Marimekko
Proportional symbol
Isotype
Lollipop
Radar
Parallel coordinates
Bullet

Spatial

Basic choropleth
Proportional symbol
Flow map
Contour map
Dot density
Heap map

Determine
Context



Choose chart
type



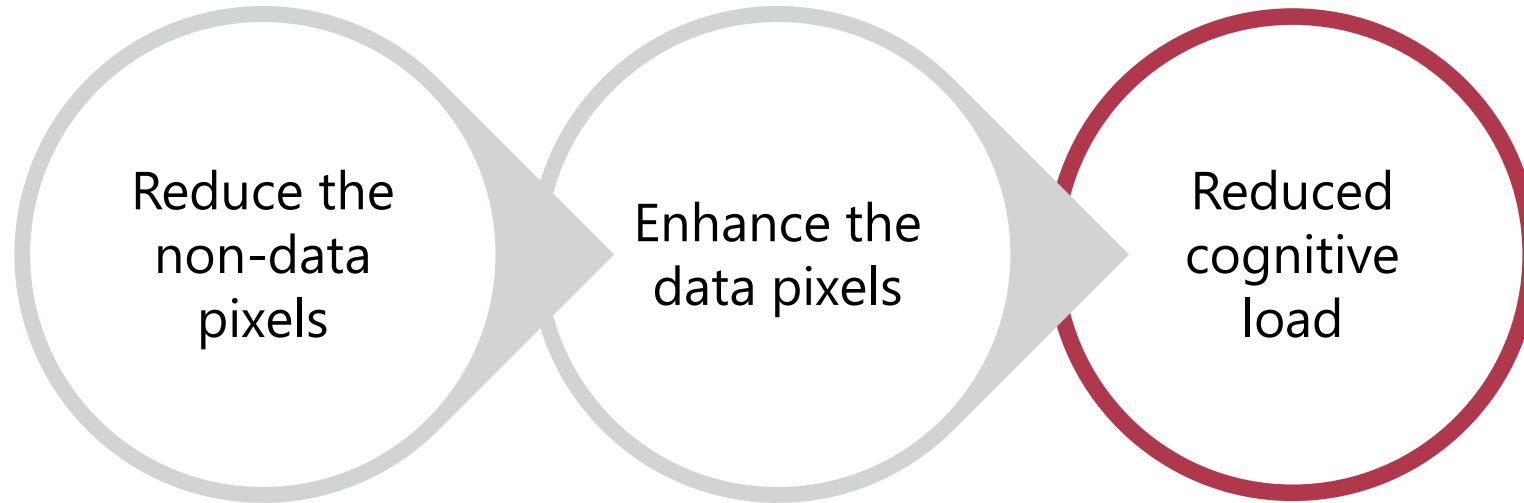
Maximize data to
ink ratio



Leverage Visual
Perception



Revisit
Re-edit



- Eliminate all unnecessary non-data pixels
- De-emphasize and regularize the non-data pixels that remain

- Eliminate all unnecessary data pixels
- Highlight the most important data pixels that remain

*Edward Tufte – The Visual Display of Quantitative Information
Stephen Few – Information Dashboard Design*

Determine
Context



Choose chart
type



Maximize data to
ink ratio



Leverage Visual
Perception



Revisit
Re-edit

Shape



Enclosure



Line Width



Saturation



Color



Size



Markings



Orientation



Position



3D



Length



Curvature



Density



Closure

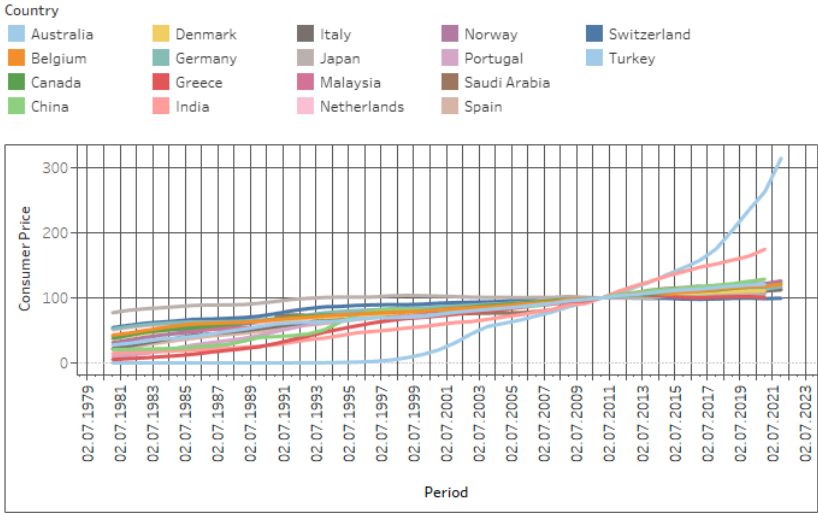
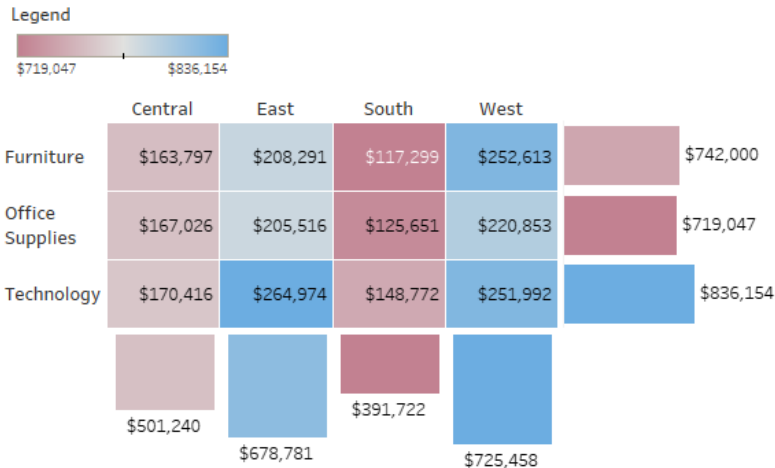


Sharpness

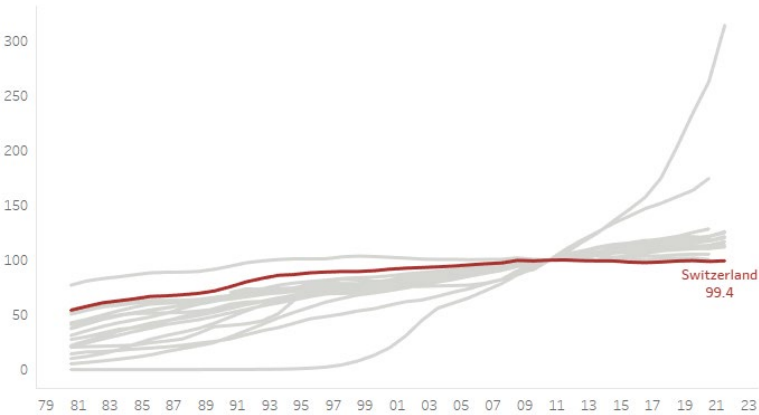


Bringing it all together - Examples

| Category | Region | | | | Grand Total |
|-----------------|-----------|-----------|-----------|-----------|-------------|
| | Central | East | South | West | |
| Furniture | \$163,797 | \$208,291 | \$117,299 | \$252,613 | \$742,000 |
| Office Supplies | \$167,026 | \$205,516 | \$125,651 | \$220,853 | \$719,047 |
| Technology | \$170,416 | \$264,974 | \$148,772 | \$251,992 | \$836,154 |
| Total | \$501,240 | \$678,781 | \$391,722 | \$725,458 | \$2,297,201 |



Consumer Price data for **Switzerland** from 1980 to 2020



Tools we use for facilitating the visualization process

Server Infrastructure

Deployment
Monitoring
Maintenance

Standardization - Style Guide

Corporate Identity
Template
Typography
Units & Abbreviations
Iconography
Colour

DataViz Community

Communication
Support
Meetups
Webinar
Competition
Trainings

Future Plans

Chart Glossary
Tableau Champions Program
Additional communication
& support channels



Thank you

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Defining business transformation rules in a standardised format – a practical case¹

Daniela Arru and Giulia Oddone,
European Central Bank

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Defining business transformation rules in a standardised format: a practical case¹

Arru Daniela, Oddone Giulia

Abstract

When talking about developing a data product at European level and collaboratively, the need and benefit of standardisation and harmonisation is often implicitly assumed, but this is not necessarily enough for a successful product. An important decision to be taken is the format that accompanies the standard, as the data product should combine business and technical aspects to gain a wider adoption and contribution. In this paper we explain why a standardised format with business characteristics is needed when it comes to regulatory reporting. The documentation of reporting instructions plays an important role in improving data reporting processes for financial institutions. Transformation rules are part of it and have been used as a way of formalising the regulatory instructions. Examples of documentation of transformation rules in Europe show that language and formats adopted often vary. In this paper we draw a set of principles for a standardised business-friendly format that bridges the needs of technical and business users. Such format should on one hand improve the ability of business experts to contribute and consume business transformation rules, and on the other hand support implementation of innovative solutions by technical experts.

Keywords: Banks' Integrated Reporting Dictionary (BIRD), transformation rules, Validation and Transformation Language (VTL), Extract Transform Load (ETL), business process, Business Process Model and Notation (BPMN), Integrated Reporting Framework (IReF)

JEL classification: G21, E50, C81, C82, C88, L17

Contents

| | |
|---|---|
| Defining business transformation rules in a standardised format: a practical case | 1 |
| 1. Introduction | 2 |
| 2. The role of transformation rules..... | 3 |
| 3. Formats – from technical towards business..... | 4 |
| 4. Conclusions..... | 8 |
| References | 8 |

¹The views expressed are solely those of the authors and do not necessarily reflect the opinion of the European Central Bank.

1. Introduction

Over the past decade and in response to the financial crisis, authorities have increased their demand of data for statistical and regulatory purposes. Data requirements are often based on different definitions within and across regulations. On the side of financial institutions, the lack of standardisation leads to a considerable effort to design and maintain the data reporting process.

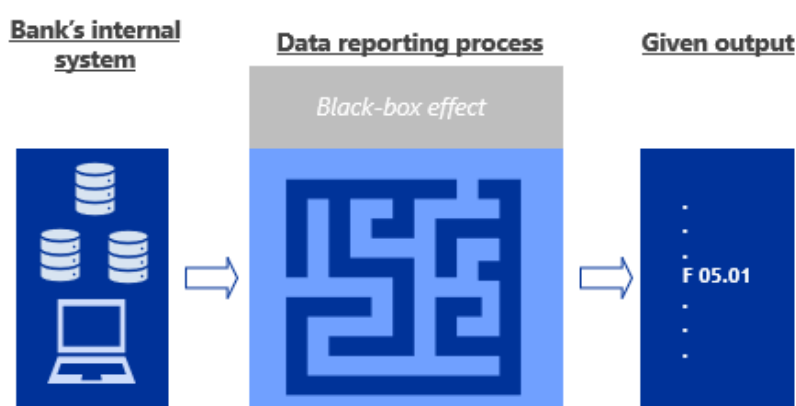
Every year each financial institution in Europe invests resources to interpret regulations and all associated manuals and instructions to fulfil their reporting obligations². Financial institutions are often supported by software vendors and are dependent on proprietary IT solutions. Software vendors that compete to sell user-friendly and advanced analytical reporting solutions to reporting agents not only engage in software development, but also analyse and interpret once again and in their own ways the same regulations, manuals, and other material. Currently most parties involved in the business of regulatory reporting develop transformation rules independently. In this scenario of multiplication of effort on non-competitive tasks, transformation rules developed collaboratively and made available to the public have the potential to unfold competition and innovation.

Their importance stems from the fact that well-defined transformation rules provide semantically consistent interpretations of regulations, therefore increasing compliance with the regulatory requirements, reducing reporting burden, permitting data and attribute lineage, and improving the consistency and quality of the data reported to authorities.

In this paper we focus on the development of transformation rules given our experience in collaborative initiatives, and we explain why we consider the format as an important factor when we aim at increasing standardisation and at fostering innovation. Formats that are too technical, e.g. programming languages, impose barriers to the collaboration by business experts and have long developing and testing cycles. The adoption of business language, logical constructs and automatable and test-driven approaches ensures the right support to business experts. It also forms a sound and standardised basis for any technical implementation. A suitable format for transformation rules provides a publicly available standard and overcome the perception that transformation rules are a costly black-box.

Figure 1. Black-box effect in data reporting process

Simplified representation



² European Banking Federation (2019)

2. The role of transformation rules

Currently in Europe reporting agents submit on a regular basis information of their non-trading loans and advances. One example of such reporting is the template F 05.01 for financial information reporting of the European Banking Authority. The task of producing F 05.01 implies potentially thousands of different implementations of the data reporting process. Each implementation has different features, e.g. technical language, data dictionary and data model, ETL process design, data lineage. These features are the results of multiple factors, such as independent interpretations of concepts in the regulatory reporting documentation, banks' legacy systems and expertise developed by proprietary software solutions. A multitude of different implementations also means that switching from one solution to another can be extremely costly for financial institutions due to customisation of the reporting solution and dependencies with their internal ETL processes and data systems.

In some European countries³ and at international level⁴ collaborative initiatives have been introduced where experts from financial institutions and authorities analyse regulations, manuals and reporting instructions, apply their expert judgment and make their interpretation available for other parties to make use of it. They don't only find synergies on subject-matter analyses of highly complex regulatory documentation, they also provide an agreed, acknowledged and understood set of information to national market participants which in turn use it as their standard.

As an example, the Banks' Integrated Reporting Dictionary (BIRD) is a project established as a collaborative initiative between authorities and the banking industry in the field of regulatory reporting. BIRD develops a database encompassing a harmonised data dictionary, and a harmonised data model, that specifies how data can be extracted from the banks' internal IT systems to generate the reports required by authorities. BIRD is not limited to the description of data that banks need to collect from their internal systems. Experts also document in the form of transformation rules the operations to be applied to input data to enrich them with derived concepts so to have the complete set of data requested by authorities.

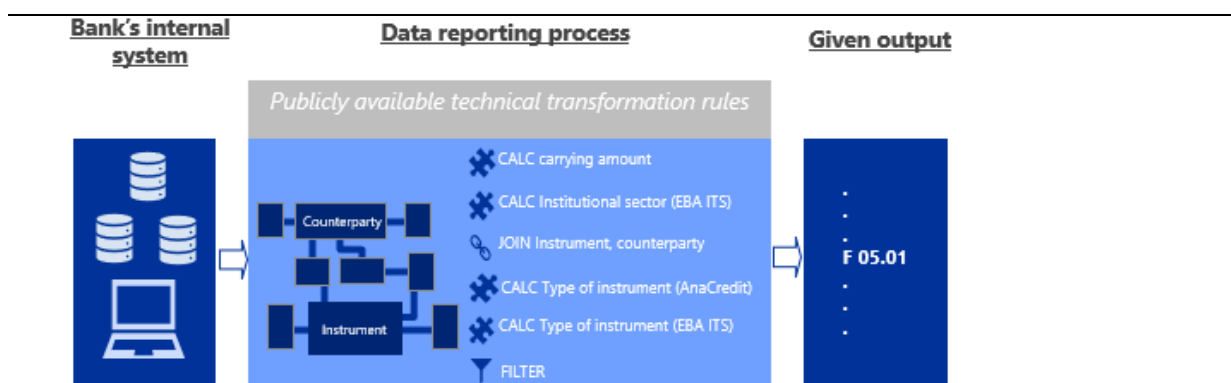
When thinking about template F 05.01, currently the process to create it can widely differ from bank to bank. BIRD provides one option of a harmonised data dictionary, including data model and transformation rules, for creating F 05.01 from a redundancy-free input. For example, in the BIRD database one can find calculations for the derivation of carrying amount, or for the institutional sector breakdown and the type of instrument according to European Banking Authority's reporting requirements, which may differ from the ones required for other reporting, e.g. AnaCredit. Output is generated by joins, filters, select statements, and by performing simple mathematical operations. Formal joins of counterparty and instrument's tables, selection of subset of variables and filtering out the assets held for trading or held for sale are also formally documented to generate template F 05.01. As shown in Figure 2, calculations are publicly available to market participants.

³ Bank of Italy Cooperazione PUMA and Central Bank of the Republic of Austria (OeNB) cooperation with Austrian banks. See also Kienecker et al (2018)

⁴ Banks' Integrated Reporting Dictionary (BIRD), https://www.ecb.europa.eu/stats/ecb_statistics/co-operation_and_standards/reporting/html/bird_content.en.html

Figure 2. Transformation rules are made publicly available in BIRD

Simplified representation



3. Formats – from technical towards business

Transformation rules that help the interpretation of the reporting requirements may be expressed in different formats: in semantic language, e.g. natural language, or in technical language, e.g. SQL, VTL, Python. In the case of the BIRD project, the Validation and Transformation Language (VTL)⁵ was chosen as the formal and technology neutral language to write end-to-end transformations rules. Figure 3 shows an example of transformation rule in VTL for the derivation of the carrying amount depending on the accounting classification⁶.

The choice of a technical language typically requires that transformation rules are successfully tested end-to-end with the benefit of ensuring high quality of the transformation rules, e.g. accuracy and consistency, of providing unambiguous interpretation and of encouraging technical implementations. The use of a technical language has therefore several advantages for standardisation.

However, business experts in collaborative initiatives are often required to learn the technical language before they could start writing transformation rules. The limited technical knowledge accompanied with the lack of freely or widely available tooling support may imply that business experts are not in the position to effectively and timely contribute nor to fully benefit from transformation rules. Moreover, end-to-end executable transformation rules are envisaged primarily for testing the quality and for potential implementation by software vendors. In practise however, for performance reasons of each IT environment, testing activities may require custom adaptations of transformation rules, resulting in a multiplication of effort, e.g. transformation rules completely rewritten in other languages, instead of being translated and directly used.

⁵ SDMX Technical Working Group, VTL Task Force (2015)

⁶ BIRD 5.0, https://www.ecb.europa.eu/stats/ecb_statistics/co-operation_and_standards/reporting/html/bird_content.en.html

Figure 3. Example of VTL transformation rule in the BIRD database

```

define operator D_CRRYNG_AMNT (ACCNTNG_CLSSFCTN component, FV component, GRSS_CRRYNG_AMNT_E_INTRST
component,
    ACCRD_INTRST component, FV_CHNG_HDG_ACCNTNG component, ACCMLTD_IMPRMNT component,
    CRRYNG_AMNT component, IS_CRRYNG_AMNT_DRVD component)
returns component is
    if IS_CRRYNG_AMNT_DRVD = "T"
    then
        if ACCNTNG_CLSSFCTN in {"14", "6", "8", "4", "2", "41"}
        then
            FV
        else
            GRSS_CRRYNG_AMNT_E_INTRST + ACCRD_INTRST - FV_CHNG_HDG_ACCNTNG - ACCMLTD_IMPRMNT
        else
            CRRYNG_AMNT
end operator;

```

Meanings of variables and codes used:

CRRYNG_AMNT: Carrying amount

ACCNTNG_CLSSFCTN: Accounting Classification (

2 = Financial assets held for trading; 6 = Financial assets at amortised cost; 8 = Financial assets at fair value through other comprehensive income; 4 = Financial assets designated at fair value through profit or loss; 14 = Cash balances at central banks and other demand deposits; 41 = Non-trading financial assets mandatorily at fair value through profit or loss)

FV: Fair value

GRSS_CRRYNG_AMNT_E_INTRST: Gross carrying amount excluding interest

ACCRD_INTRST: Accrued Interest

FV_CHNG_HDG_ACCNTNG: Fair value changes due to hedge accounting

IS_CRRYNG_AMNT_DRVD: Is carrying amount derived (parameter to trigger the generation of the Carrying amount)

ACCMLTD_IMPRMNT: Accumulated impairment

Experience had shown that a business component for transformation rules is needed on top of and prior to the technical one. The format of transformation rules needs to be suitable to business experts because they are often those in charge of their creation. Therefore, a design of transformation rules should meet their needs, as “producers” and as “users”. On the other hand, the format should also be useful and practical for technical implementation. IT solutions, e.g. in the form of prototypes, are essential to trigger quality control processes on the transformation rules, and to improve the consistency and accuracy of the content, e.g. via multiple decentralised testing processes.

Based on our experience, we conclude that the format of transformation rules should bridge business and technical users. We identified guiding principles for writing standardised and business-friendly transformation rules:

- functional classification and organisation depending on the type of result, e.g. derived attribute, output report
- attention to separation of concerns⁷, e.g. logical vs technical, to fit different purposes and users
- the selection of free, open-source and widely known formats to lower the barrier to contribute and to incentivise the development of tooling
- test-driven development approach where tests and testing activities are integral part of the development of transformation rules
- collaborative development and incremental contributions

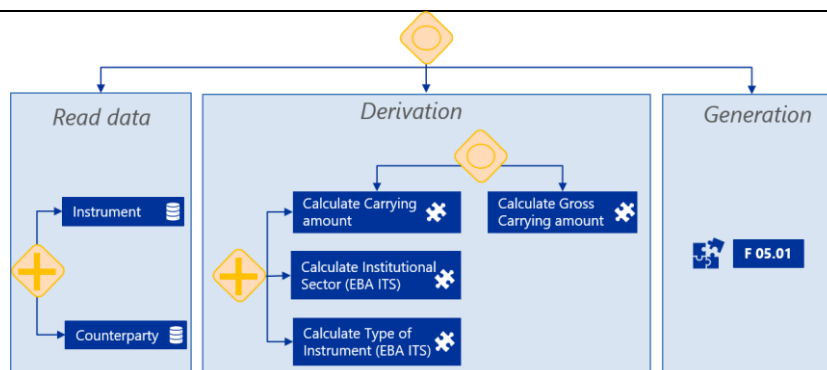
By following the guiding principles, we outline one possible way of writing transformation rules in a collaborative initiative that is favourable to become a standard, e.g. used by business and technical users.

⁷ Bos et al (2019)

Business experts provide first an illustration of the flow of transformation rules required from the input to the output using a known standard, e.g. Business Process Model and Notation (BPMN⁸). Such illustration, or diagram, depicts the sequential steps to be followed: starting from reading the input, enriching it with derived concepts, and finally generating the desired output. In our example of F 05.01, business experts sketch the diagram using the agreed convention; in this phase they also identify which concepts are already derived by an existing transformation rule and which need to be derived to feed the output report (Figure 4).

Figure 4. Example of workflow for F 05.01

Simplified representation

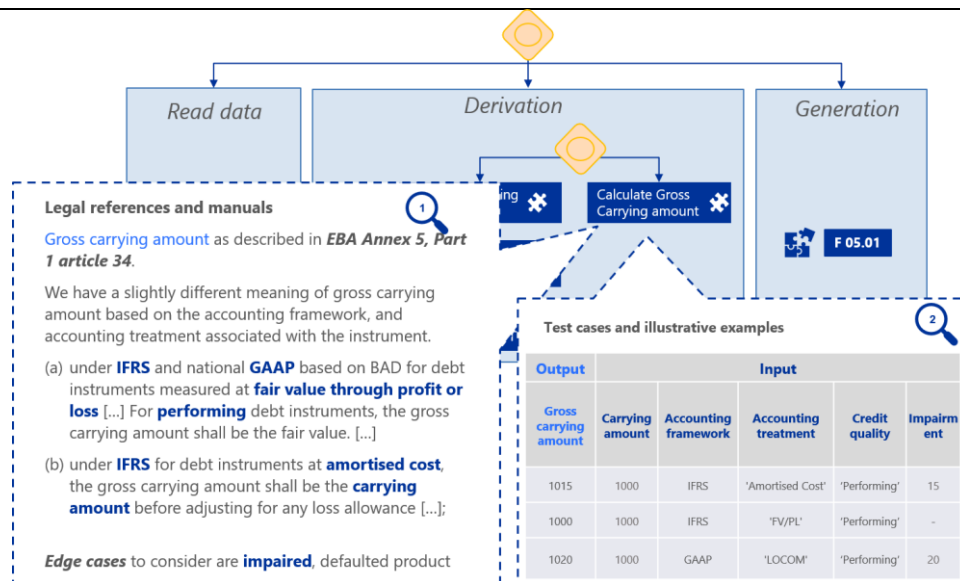


Under this approach, experts and users exploring transformation rules can rely on high-level lineage information and a workflow diagram that shows the order in which operations should be executed. This provides a business-oriented way of exploring a large set of transformation rules at a high level, together with the possibility of retrieving specific details when necessary.

Indeed, the high-level information of the diagram is then enriched with descriptions for each transformation rule. In particular, the business experts fill in, in a pre-defined and business-friendly format, relevant information for business and technical users. For example, a relevant set of information for deriving enriched concepts is: references to legal documents and reporting manuals, explanation of scenarios for illustrative cases, input and output variables, test cases and test data rooted in the same data dictionary, e.g. input and output data models in terms of tables, variables, and allowed values (Figure 5).

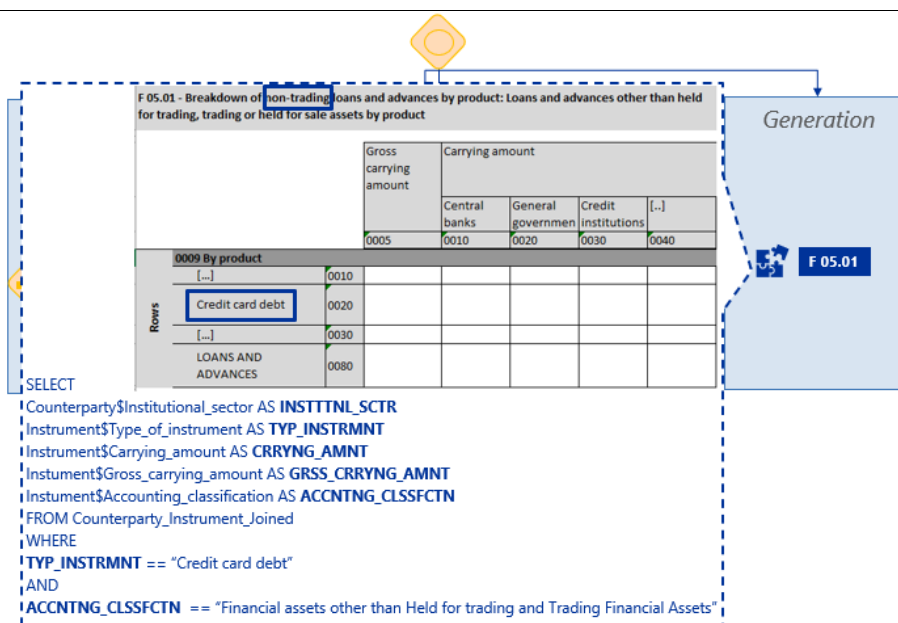
⁸ <https://www.bpmn.org/>

Figure 5. Focus on the derivation of gross carrying amount
Simplified representation



After all derived concepts are defined, business experts finalise the documentation of transformation rules for each desired output, e.g. they define which aggregation, grouping or filtering is needed to match the requirements expected for template F 05.01. The definition of steps for the generation of output requires simple mathematical operations which are therefore well documentable by non-technical experts also using a programming language, e.g. SQL, VTL, Python. The intention is to enable business experts to contribute time-effectively and with business value added, and not to write a fully testable and often not business-friendly piece of code. This can be achieved by accepting a simplification step, i.e. by assuming that correct join statements across all required input tables are applied and create a single flat table, thereby reducing the complexity of steps that business experts need to document (Figure 6).

Figure 6. Focus on the generation of F 05.01
Simplified representation



In this example, business experts do not create end-to-end executable transformation rules. They create all the business artefacts required for software vendors or reporting agents' IT developers for implementing an IT solution that uses the technology they find more suitable. Experts can carry out their contribution to the definition of transformation rules by analysing regulations and by using accessible tools and formats to write down their findings. The output is then published and available not only to other business experts for consultation and for potential validation and improvement, but also to technical experts, e.g. in the field of software development, for further enrichment and application, e.g. into an IT solution.

4. Conclusions

The field of regulatory reporting shows a high degree of heterogeneity of formats and semantical description, hence standardisation on how requirements are documented is constantly under authorities' consideration⁹. Possible gains for cost reduction and data quality can be achieved in the future with a European integration of requirements¹⁰ and a standardisation of formats. By means of standardised transformation rules, a considerable burden on financial institutions and software vendors is reduced and resources can be freed for more competitive tasks. Additionally, a standardised format with a focus on business language is important for a wide contribution, adoption and as a basis for innovation.

Our conclusions point to the need of designing a standardised and business-friendly format for business transformation rules prior to any documentation in technical language. Such solution overcomes barriers and long learning curves that may hinder effective contribution, and offers the market participants an environment that is favourable for experimenting and testing prototypes.

Future initiatives on standardisation in regulatory reporting can vastly benefit from a collaborative development of transformation rules. In this context, we agree that the two pillars of the ESCB long-term strategy for banks' data reporting, BIRD and IReF, are an opportunity of a playground at an early stage to continue developing synergies in aligning data models¹¹. Such synergies could be exploited in the future for what concerns standardised transformation rules, and therefore we identified principles to design a solution that bridges business and technical needs and creates a favourable requisite for wide adoption.

While a fully integrated reporting is a long-term objective, collaborative initiatives have the possibility to proceed at faster pace. They can absorb good national and international practices, set aside identified deficiencies of existing approaches, and step up their engagement on leading conceptual work and consolidation of practical experiences in the design of transformation rules.

References

Banks' Integrated Reporting Dictionary, https://www.ecb.europa.eu/stats/ecb_statistics/co-operation_and_standards/reporting/html/bird_dedicated.en.html

Bos, A. and van der Helm, R. (2019): "Heading for harmonization of data collection"

⁹ Crisanto et al (2020)

¹⁰ European Central Bank (2021)

¹¹ European Central Bank (2021), Colangelo et al (2021)

Business Process Model and Notation, <https://www.bpmn.org/>

Crisanto, J.C., Kienecker, K., Prenio, J. and E. Tan (2020): "From data reporting to data-sharing: how far can supotech and other innovations challenge the status quo of regulatory reporting?", FSI Insights on policy implementation, no 29, December.

Colangelo, A., Gross, F. and Schuster, F. (2021): "Effective measurement of the economy in the emerging digital age"

European Banking Federation (2019): "Boosting Europe: Building Trust and Supporting Growth in Europe. EBF recommendations for the EU 2019-2024 legislative cycle and beyond", May

European Central Bank (2021): The Eurosystem Integrated Reporting Framework: an overview, December

Kienecker, K, G Sedlacek and J Turner (2018): "Managing the processing chain from banks' source data to statistical and regulatory reports in Austria", Statistiken – Daten und Analysen, August.

PUMA, <https://www.cooperazionepuma.org/homepage/index.html>

SDMX Technical Working Group, VTL Task Force (2015): "Validation & Transformation Language, Part 1 – General description. Version 1.0"

DEFINING BUSINESS TRANSFORMATION RULES IN A STANDARDISED FORMAT*

A PRACTICAL CASE

IFC-BANK OF ITALY WORKSHOP ON "DATA
SCIENCE IN CENTRAL BANKING"

17/02/2022

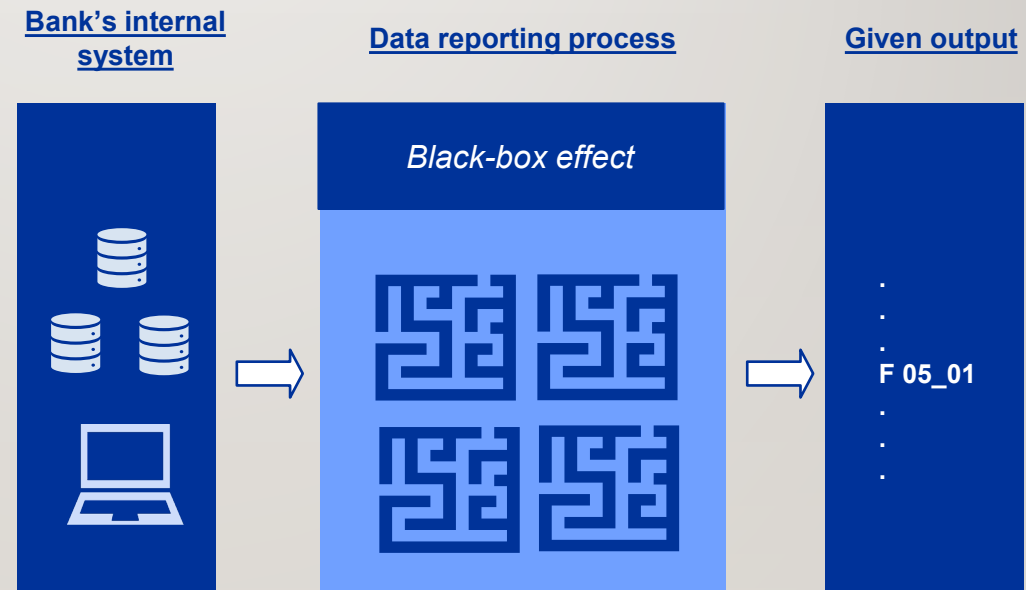
ARRU D., ODDONE G.

In the regulatory reporting fields scattered lack of standardisation has multiple consequences:

- Interpretation of regulations is a highly intense work for banks and software vendors
- Reporting processes are all different, customised to banks, and hard to change

Question: What can be done about it?

Example: to produce the same Finrep F 05_01 template, each bank utilises different input models and follows different data reporting processes

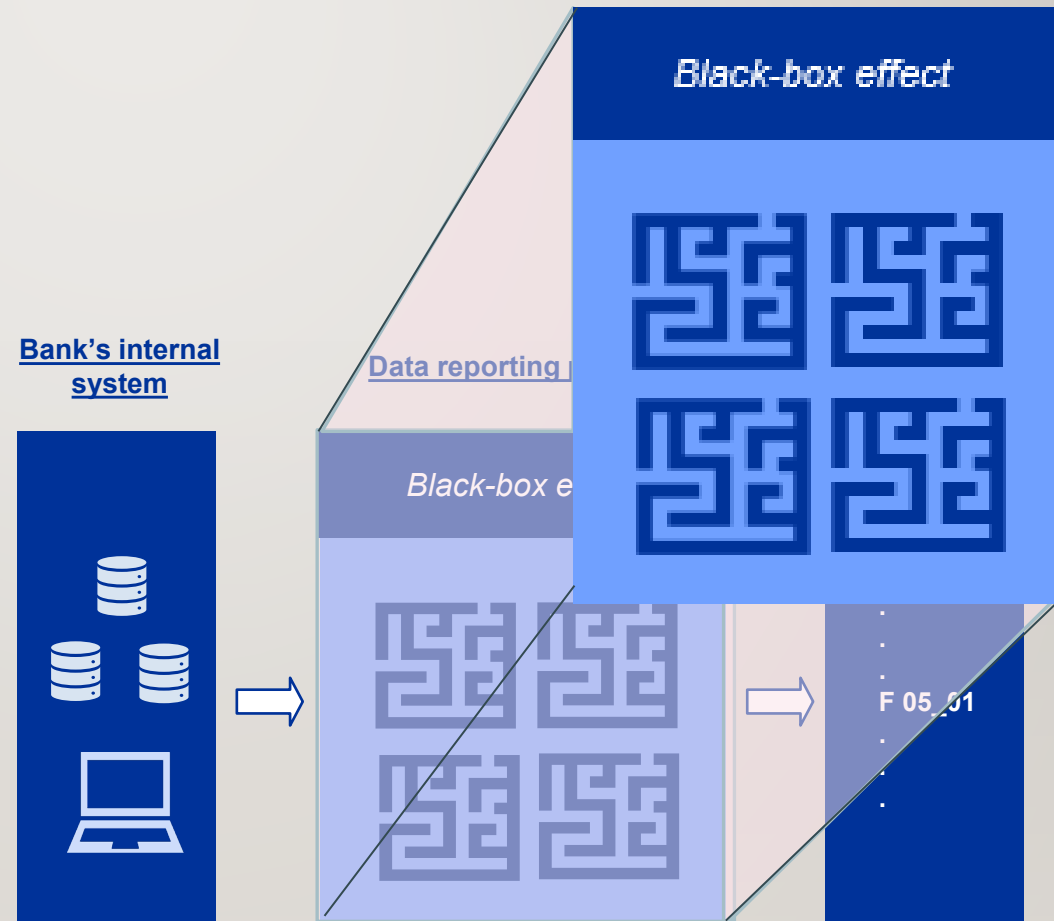


Transformation rules in data reporting processes play an important role in standardisation effort, since they describe operations to derive information and create reports.

The choice of the format of these descriptions is a key element to move away from the black-box effect.

Less technical and more business-friendly:

- ✓ More business contribution
- ✓ More usable and adoptable
- ✓ Less developing and testing



Let's see an everyday example...

Instructions to build wardrobes are highly standardised and we all benefit!

And when they are not following a standardised (and user-friendly) format, we as users experience difficulties to interpret them, e.g.:

Do we have all the pieces?

When do we assemble the drawer?

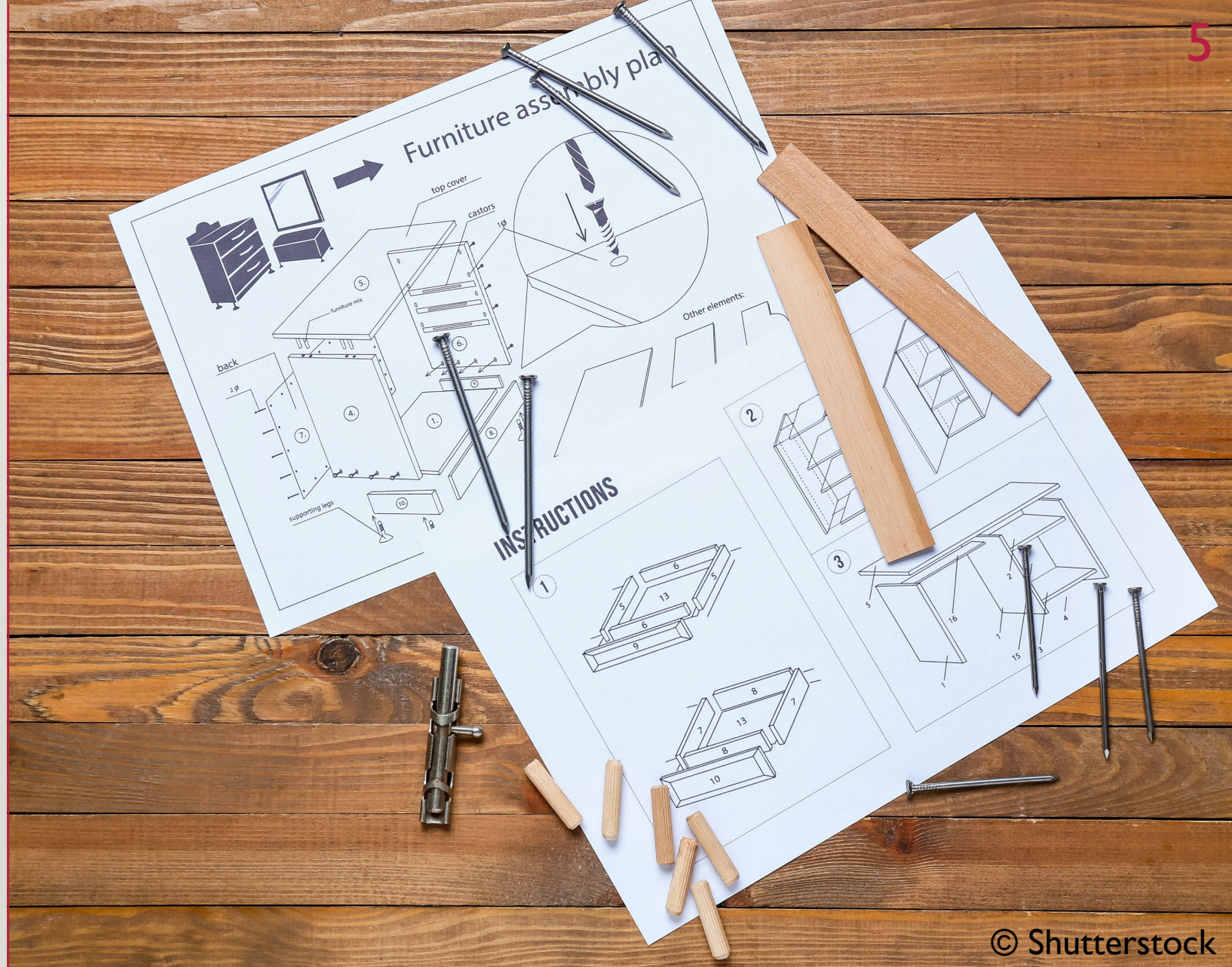
Before or after the mirror?



A standardised and user-friendly format for instructions leads to clarity and provide the basis for further improvements.

Likewise a standardised and business-friendly format for transformation rules offers many benefits:

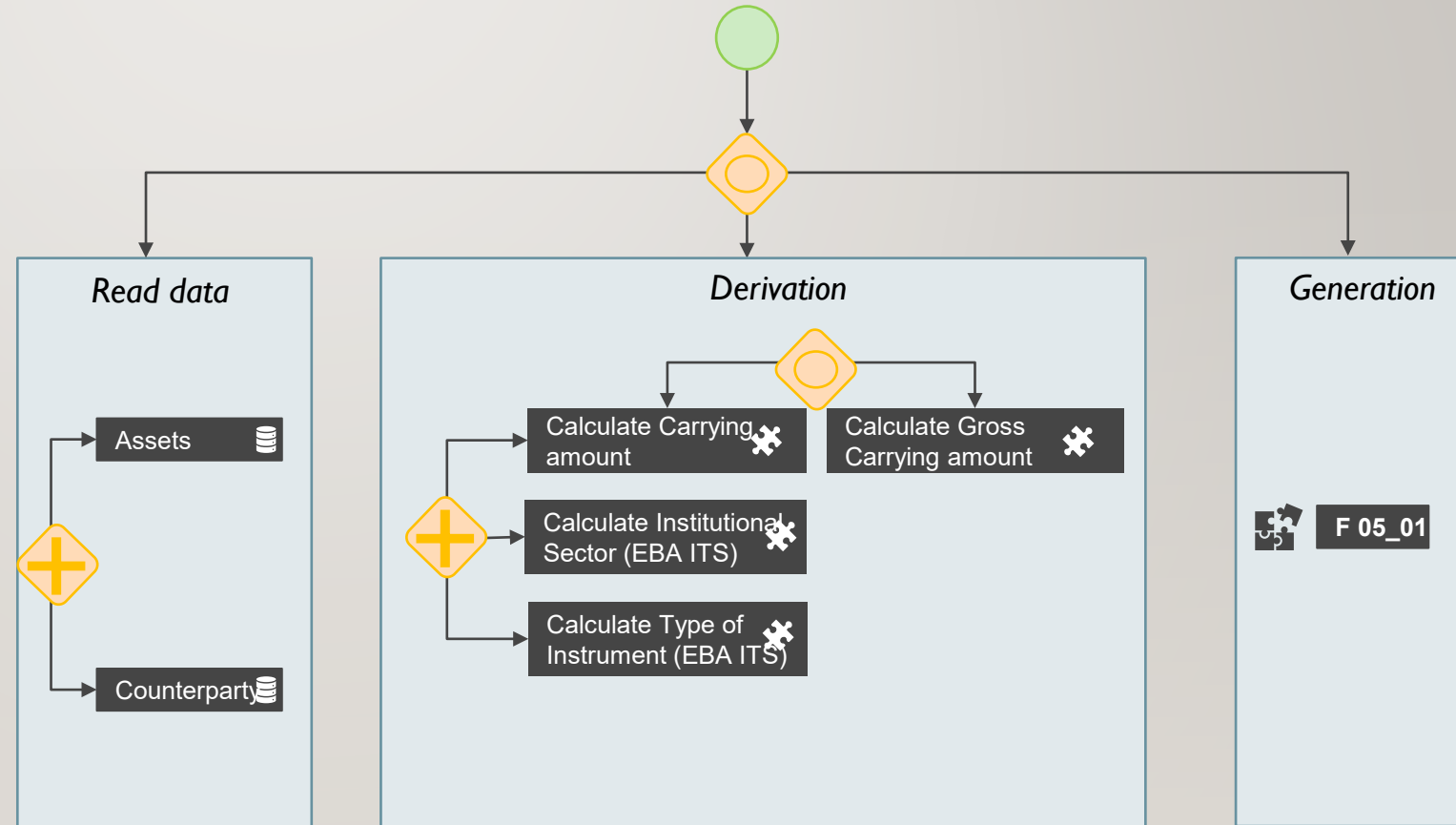
- ✓ adopted by market participants
- ✓ understood by experts
- ✓ further improved by technical experts
- ✓ IT solutions based on it



Principles for a standardised and business-friendly transformation rules to bring standardisation and incentivise innovation:

- ✓ Publicly documented in a collaboration effort between authorities and banks
- ✓ Respect the principle of separating concerns to better fit needs and focus contribution
- ✓ Founded on open-source and known formats for wider tools support
- ✓ Suitable basis for technical implementation

Example: to produce the same Finrep F 05_01 template, each bank has access to publicly available transformation rules organised and documented in business-friendly and standardised formats.



The choice of format for transformation rules should fit the needs of different interested parties and is an important element to feed in future initiatives.


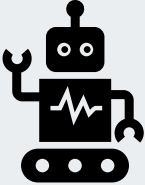
It facilitates contributions by:

- ✓ Internal business contributors
- ✓ External contributors

It provides useful descriptions and instructions of reporting requirements to:

- ✓ Business users
- ✓ Technical profiles

Example: standardised and business-friendly transformation rules for multiple benefits and different needs

| | |
|--|--|
|  | ✓ Legal references and manuals |
| | ✓ Diagram of the data transformation process |
| | ✓ Attribute lineage information |
| | ✓ Illustrative examples |
|  | ✓ Test cases |
| | ✓ Business artefacts required for implementation |
| | ✓ Technical format |

Some conclusions:

- Standardisation in the field of reporting and data processing is beneficial (and needed)
- Transformation rules can be documented in many formats and languages, e.g. natural, business, technical language
- The definition of a format can have relevant impact on future initiatives around reporting burden and integrated reporting requirements
- A standardised and business-friendly format for transformation rules improves the successfulness of the standardisation efforts. E.g. clarity for business experts to contribute and use, basis for technical experts to develop and improve

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Anomaly intersection: disentangling data quality and financial stability developments in a scalable way¹

Gemma Agostoni, ECB, Louis de Charsonville, McKinsey & Company,
Marco D’Errico, ECB, ESRB Secretariat, Cristina Leonte, BIS, and Grzegorz Skrzypczynski, ECB

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Anomaly *intersection*: disentangling data quality and financial (stability) developments in a scalable way

Gemma Agostoni (ECB), Louis de Charsonville (McKinsey & Company), Marco D'Errico (ECB, ESRB Secretariat), Cristina Stefana Leonte (BIS), Grzegorz Skrzypczynski (ECB)

The views expressed in this paper are of the authors and do not necessarily represent the views of the associated institutions.

Abstract

Enhancing the information available to policymakers is one of the pillars of the reforms following the global financial crisis. Several granular data collections have the potential to increase transparency and support effective policy responses. However, poor data quality is impairing this process, making it increasingly arduous to disentangle whether anomalies and risk build-ups are relevant from a policy perspective or the result of poor data quality. We term this problem *anomaly intersection* and propose a general framework to tackle it in a scalable and flexible way. The framework allows to build a set of automatable tools that analyse anomalies at all levels of aggregations, uncovering their nature. We show how this framework can be successfully applied to transaction-level data on derivatives collected under the European Market Infrastructure Regulation, the largest supervisory dataset to date.

Keywords: automatic, granular, derivatives, quality, EMIR, decision trees, financial stability

JEL classification: C18

Contents

| | | |
|-------|---|----|
| 1 | Introduction..... | 3 |
| 2 | EMIR data on derivatives and data quality..... | 5 |
| 2.1 | EMIR data – short overview..... | 5 |
| 2.2 | EU data quality framework for EMIR | 6 |
| 2.3 | Classification of broad types of quality issues..... | 7 |
| 2.3.1 | Data quality issues due to over-reporting | 8 |
| 2.3.2 | Data quality issues due to under-reporting | 8 |
| 2.3.3 | Data quality issues due to misreporting..... | 9 |
| 2.4 | Challenges in data quality assurance in large-scale financial datasets..... | 10 |
| 3 | Methods..... | 11 |
| 3.1 | Modelling framework | 11 |
| 3.2 | Algorithm..... | 13 |
| 3.2.1 | Entity-level analysis module | 13 |
| 3.2.2 | Dimensions’ analysis module..... | 14 |
| 3.3 | Application to double-sided reporting reconciliation..... | 18 |
| 4 | Application of the ADQ method to EMIR data..... | 20 |
| 4.1 | ADQ Process..... | 20 |
| 4.2 | What do we measure? | 21 |
| 4.3 | Extension of the work | 22 |
| 5 | Disentangling anomalies | 23 |
| 6 | Conclusions..... | 25 |
| | References..... | 27 |

1 Introduction

“Good data and good analysis are the **lifeblood of effective** surveillance and policy responses” (FSB, IMF 2009). Very few episodes exemplify this statement more than the global financial crisis and the more recent Covid-19 crisis. Indeed, one of the lessons we keep learning from crises is that prompt access to high quality data is key to develop an effective policy response. Indeed, “lack of timely, accurate information hinders the ability of policy makers and market participants to develop effective responses” (FSB, IMF 2009). Enhancing the set of information available to policymakers through the collection of new sources of data has been one of the pillars of the post crisis policy reforms, albeit a lesser studied one. Policy institutions’ work is now profoundly connected with the collection and analysis of data. For instance, the task of the European Systemic Risk Board¹ is to “**monitor and assess systemic risk** in normal times for the purpose of mitigating the exposure of the system to the risk of failure of systemic components” (ESRB Regulation).² Such monitoring activities “should be based **on a broad set of relevant macroeconomic and micro-financial data and indicators**”.

To improve their monitoring and analytical tasks, policymakers are increasingly collecting and analysing an unprecedented wealth of data. “Monitoring an interconnected financial system involves the availability of **detailed and granular transactions data**.” (Mario Draghi, 2018):³ indeed, these datasets are collected at a relatively high frequency, and with high level of granularity and details. One of the most well-known granular collections is represented by data reported under the European Market Infrastructure Regulation (EMIR).⁴ EMIR mandates entities in the EU to report details of their derivatives contracts to Trade Repositories (TRs), resulting in the collection and processing of about 100 million observations per day. Transaction-level derivatives data represent a particularly relevant instance, given the role the opacity of these instruments played in the amplification and transmission of the Global Financial Crisis. EMIR states that OTC derivatives “create a complex web of interdependence which can make it **difficult to identify the nature and level of risks involved**. The financial crisis has demonstrated that such characteristics increase uncertainty in times of market stress and, accordingly, pose risks to financial stability”.

Remarkably, since the inception of the reporting regime, these data have been characterised by substantial, pervasive, and persistent data quality issues. Data quality issues can be traced back to both reporting entities and trade repositories and apply to both large players (including Central Counterparties and large banking groups⁵)

¹ The European Systemic Risk Board (ESRB) was established in the aftermath of the global financial crisis, and is responsible for the macroprudential oversight of the EU financial system and the prevention and mitigation of systemic risk.

² Regulation (EU) No 1092/2010 of the European Parliament and of the Council of 24 November 2010 on European Union macro-prudential oversight of the financial system and establishing a European Systemic Risk Board (with further amendments)

³ Welcome remarks at the third annual conference of the ESRB, <https://www.ecb.europa.eu/press/key/date/2018/html/ecb.sp180927.en.html>

⁴ Regulation (EU) No 648/2012 of the European Parliament and of the Council of 4 July 2012 on OTC derivatives, central counterparties and trade repositories (with further amendments)

⁵ See also ECB Banking Supervision (2018)

and smaller players.⁶ They are very **heterogenous** in nature: from (i) missing reports (that is: a counterparty, including CCPs, not reporting all or part of their contracts), to (ii) limited use of agreed standards, such as the Legal Entity Identifier (LEI) or Unique Trade Identifier (UTI), to (iii) the reporting of values that are implausible, incorrect, not up to date or not reconciled between counterparties entering a contract. The first two root causes, while still rather concerning, can be and are addressed via institutional channels;⁷ the latter is left to the data scientist and policy analyst.

We argue that working in the presence of such data quality issues does not represent only a technical and statistical problem:

1. it represents a key hurdle to fully exploiting this wealth of data which, in turn, increases **opacity** for both policymakers and market participants;
2. it hampers the **scalability** of monitoring frameworks for policymakers and
3. it limits the **ability** of policymakers to analyse and study developments, as substantial resources have to be dedicated to solving these issues while, in numerous cases, leading to uncertain results.

Poor data quality introduces significant opacity and uncertainty, undermining the primary objective of data collection.

Systemic risk “builds up in the background before materialising”⁸: financial stability monitoring should prioritize frameworks that can proactively detect these risks. One of the cornerstones of effective risk detection is identifying specific anomalies in the data, such as pronounced risk concentrations or the emergence of large positions. This approach can be used in designing early warning systems, evaluating interconnectedness, assessing the implications of potential contagion risks, and evaluate possible tail events. For instance, daily monitoring of aggregate margin calls and their distribution during the March 2020 market turmoil has been key to identify risks⁹ and build policy recommendations.¹⁰

While it is possible, from a technological and analytical viewpoint, to develop scalable and continuous monitoring framework,¹¹ the challenge lies in understanding the implications of data quality. Specifically, it becomes essential to distinguish whether an observed development is pertinent from a policy systemic risk viewpoint or merely a consequence of suboptimal data quality.

Policymakers are therefore faced with a challenge: that of **disentangling** relevant development from data quality issues. Given the size and complexity of the data, they need to tackle this problem in a scalable and – to the extent possible – automatable way across various datasets, counterparties, and markets. Moreover, it requires

⁶ See ESRB (2020b)

⁷ The European Securities and Markets Authority (ESMA) and the National Competent Authorities (NCAs) have taken several actions to improve the quality of EMIR data. These include the Data Quality Action Plan and the Data Quality Assessment Framework. See ESMA (2021).

⁸ See Brunnermeier *et al.* (2012).

⁹ See ESRB (2020a).

¹⁰ See the Recommendation of European Systemic Risk Board of 25 May 2020 on liquidity risks arising from margin calls (ESRB/2020/6). Available at https://www.esrb.europa.eu/pub/pdf/recommendations/esrb.recommendation200608_on_liquidity_risks_arising_from_margin_calls~41c70f16b2.en.pdf

¹¹ See Abad (2016), Apicella et al (2022).

addressing data quality issues at any level of aggregation, thereby requiring approaching this problem in a way that allows to seamlessly shift across various levels of aggregation. For example, while the aggregate levels of a given quantity (e.g. notional amounts or initial margins) can be relatively stable in the aggregate, this may mask substantial concentration into one or few counterparties at the disaggregate level. Even upon identification, the presence of data quality issues does not allow to exclude the possibility that observed developments may arise from misreporting by an entity, such as inaccurate notional or margin amounts. This convergence of potential causes poses a challenge, which we refer to as "**anomaly intersection**".

In this paper, we introduce a framework to address this problem and apply it to granular, transaction-level data on derivatives collected under the European Market Infrastructure Regulation (EMIR) and available to the European Central Bank and the European Systemic Risk Board. This framework allows to make sense of observed anomalies in the data in two ways. First, it allows to set up rules at various levels of complexity to break down a potential signal across all its relevant dimensions. This facilitates the identification of which dimensions contribute to data quality issues or financial stability developments. Second, it provides a heuristic to reduce the likelihood of both false positives and false negatives by conditionally linking the detection of developments to the quality of the underlying data. More specifically, the probability that a development is relevant from a systemic risk viewpoint, increases in direct relation to the quality of the underlying data. Conversely, in those instances where a relevant financial development has been identified, its validation is contingent upon the framework indicating a superior data quality.

The framework allows to explore better the transaction-level datasets now available to policymaker: in the presence of dozens of dimensions, pinpointing the origin of an issue becomes a complex endeavor, requiring requires a robust system capable of efficiently navigating these dimensions, integrate expert judgment, and accounting for resource constraints. This paper proposes a first step to tackle this challenge.

2 EMIR data on derivatives and data quality

2.1 EMIR data – short overview

As a result of the turbulences caused by the global financial crisis, the 2009 G20 summit in Pittsburgh highlighted the need for greater transparency in the derivatives trading and put forward a set of measures intended to increase the stability of the international financial markets. Within the European Union, the objective was addressed through incentives to standardize derivatives contracts, introducing a mandate to centrally clear certain classes of derivatives via central counterparties (CCPs) and obligation to report them to trade repositories. The key legislation introduced to achieve those goals was the European Market Infrastructure Regulation (EMIR) (EU) No 648/2012.

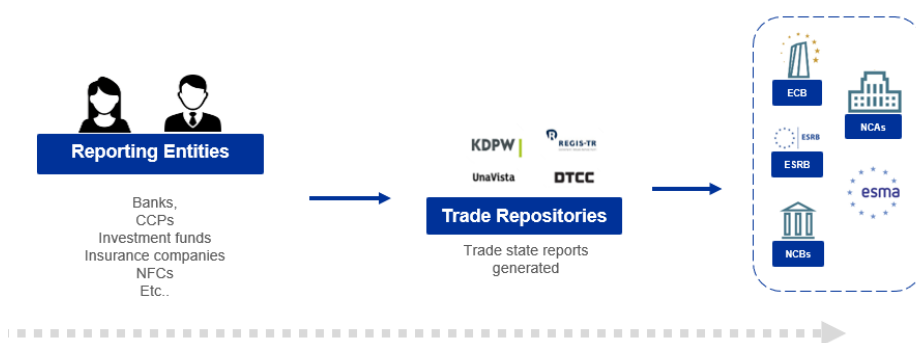
After the entry into force of the EMIR, EU competent authorities have been provided with an unprecedented amount of granular data. The data is reported daily

by all entities in the EU that are a counterparty to a derivative contract, both traded on exchanges as well as over the counter.

The decision to choose one or more of the TRs to which the trades will be reported is a free choice of the reporter and the deadline to report the transaction is the day after the transaction was executed, i.e. T+1. This implies that one reporter could submit the information of a particular trade either on the same day or the next date. In addition, the EMIR transaction reporting is a double-sided regime, meaning that every derivative trade has to be reported by both counterparties, as long as they are both resident in the EU. Thus, the two different reporting scenarios under EMIR from counterparties' perspective are (i) EU-EU and (ii) EU-non EU, while (iii) non EU-EU and (iv) non EU-non EU, are not reported under EMIR and may potentially fall under reporting rules of other jurisdictions. The typical transaction's reporting process is summarized in the Figure 1.

EMIR reporting process

Figure 1



Source: EMIR Regulation

Transaction reports received by the TRs consist of life-cycle events, that may represent, for instance, the conclusion, modification, valuation, and termination of a derivative. Owing to their volume, velocity, variety and veracity the EMIR data can be classified as "big data", which poses many challenges in using them.

2.2 EU data quality framework for EMIR

The entities reporting data under EMIR are responsible for delivering complete and correct information on concluded derivative contracts, in a timely manner. In practice, however, the data suffers from many inaccuracies, inconsistencies, or outright implausible values.

As a consequence, the authorities in the EU developed a comprehensive data quality framework to identify, exchange, prioritize and follow up on the issues found. The ESMA (European Securities and Markets Authority) assumes the leading, coordinating role in this process, directly supervising the TRs, and intermediating between authorities entitled to access EMIR data. The supervision of individual entities, however, lies within the remit of the relevant national supervisory agencies of EU member countries. This complex framework, together with predominantly cross-border nature of the derivatives trading in Europe, makes the task of dealing with reporting errors challenging.

An important tool in the supervisors' arsenal of methods to ensure data quality is the fact that the trades have to be reported by both counterparties to the trade, which allows the comparison of information transmitted in two separate reports, which, in turn, facilitates the identification of the incorrect reporting, and facilitates interpretation of potential anomalies discovered in the dataset. To support this endeavour, ESMA mandated the TRs to carry out a regular, weekly reconciliation exercise that identifies all the inconsistencies. The aggregated outcome is shared with the EU authorities.¹²

Nevertheless, despite the above efforts, the quality of EMIR data remains a significant problem. Given the size of the dataset, it's impracticable to expect that all data quality issues can be identified and addressed in a reactive manner. The importance of the regulatory reporting needs to be properly recognized by the reporting entities and should be further supported by clear and comprehensive reporting guidelines. Tackling the issues at the point of reporting data generation would be much more efficient than working on data quality on the receiving end and would enhance the quality of financial stability monitoring at the EU level.¹³

2.3 Classification of broad types of quality issues

Data quality is a multidimensional and complex concept. In the last decade, there has been a significant amount of work in the area of information and data quality management initiated by several research communities, ranging from techniques that assess information quality to build largescale data integration systems over heterogeneous data sources with different degrees of quality and trust. The development of established metrics to measure data quality is crucial to assess the significance of data-driven decisions.

EMIR requires market players to report an extensive set of characteristics for each derivative contract. Furthermore, financial companies (FCs), as well as non-financial companies (NFCs) above the so-called clearing threshold¹⁴, are obliged to report daily valuation data corresponding to their open trades and positions, as well as any relevant updates to the value of collateral exchanged. This information is reported to TRs, and this process is often intermediated by third-party entities, e.g. exchanges, trading platforms, or reporting software providers.¹⁵ The information received by the TRs is also used to create reports for the authorities. Along this reporting chain, data quality issues can emerge. Based on experience with data reported under EMIR, the following three main data quality categories can be identified and are discussed in this paper:

¹² As of 29 April 2024 (the go-live date of the so-called EMIR Refit) the TRs will also share the detailed reconciliation reports with reporting entities. See ESMA (2020a), section 6.2

¹³ See ESRB (2022)

¹⁴ As per EMIR Regulation (Article 4a and 10), NFCs and FCs are subject to the clearing obligation when exceeding predefined thresholds, see more details on ESMA's website: <https://www.esma.europa.eu/policy-activities/post-trading/clearing-thresholds>.

¹⁵ It is important to note that EMIR Regulation (Article 9(1f)) permits the delegation of reporting to other entities (including entities not directly involved in the trade).

2.3.1 Data quality issues due to over-reporting

Over-reporting data quality issues represent the non-required but reported records (derivative contracts or their valuation updates) by the market players which in the case of EMIR data are challenged by the double-reporting regime and/or reporting delegation framework; alternatively, they could be triggered by duplicated records produced while being processed on the side of the TRs.

For example, EMIR applies to entities resident in the EU, thus the trades concluded with counterparties outside the EU are expected to appear only once in the dataset. Therefore, a transaction reported by entities from non-EU jurisdictions could fall into this case, similarly to derivatives which have not been terminated appropriately or have already matured but are still reported.

Some of the data quality issues may be also generated during the portability process, a process for transferring data from one TR to another.¹⁶ The process was widely employed in the context of Brexit, but it is also frequently triggered by reporting entities on voluntary basis. Errors or inaccuracies in the transfer of information between the TRs can lead to duplicated transactions in the final dataset.

2.3.2 Data quality issues due to under-reporting

Under-reporting data quality issues represent the required but not reported records (derivative contracts or their valuation updates) by the market players. Alternatively, they could be existing records gone missing while processing the data on the side of the TRs.

A straightforward example could be a reporting entity that does not comply with EMIR reporting obligation and does not report its derivative contracts to the TR. Varying input formats required by the TRs could also be a potential data quality issue as the data needs to be transformed into the XML-based ISO20022 message, which the TRs are obliged to provide to authorities.¹⁷ In this conversion process, due to lack of standardisation of the information submitted or due to the incorrect mapping implemented by TRs, some records might be rejected from the final pool of transactions available for analysis, as they do not conform to the schema of the message to be transmitted to authorities. Similarly, transactions which fail the ESMA validation¹⁸ rules could be rejected by the TRs and may never reach the authorities.

Related to the abovementioned portability process, errors or inaccuracies in the transfer of information between the TRs can also lead to missing transactions in the final dataset.

Moreover, some submissions being reported late could be misinterpreted or not captured on a timely basis, and in turn they could bias the reconciliation of derivatives

¹⁶ See ESMA (2017)

¹⁷ This lack of input standardisation is expected to be significantly mitigated by the upcoming change of EMIR technical standards, so-called EMIR Refit, see: https://www.esma.europa.eu/sites/default/files/library/esma71-99-1490_press_release_emir_refit_final_report.pdf

¹⁸ See <https://www.esma.europa.eu/policy-rules/post-trading/trade-reporting>.

contracts. The reconciliation of derivatives contracts, described in section 2.2, is a relevant component for the data quality assessment, e.g. in the context of estimation of non-reported trades subject to double-sided reporting.

One exemption to keep in mind related to EMIR is that the regulation does not impose the reporting obligation on natural persons, hence for trades carried out by private individuals only the leg reported by the legal entity will be visible in the final dataset.

2.3.3 Data quality issues due to misreporting

Misreporting data quality issues arise in the required and reported records (derivative contracts or their valuation updates) containing erroneous information.

The erroneous information may come from the internal system of the reporting entities or be introduced in the process of the transformation of the details of the trades to TRs. It may consist in a variety of issues, including simple typos, incorrect categorisation or numerical values, as well as inaccurate interpretation of reporting guidelines¹⁹. The failure of CCPs, clearing members or more generally market players not coordinating on trade ID, counterparty ID or on position- versus transaction-level reporting²⁰ could also lead to the impossibility to reconcile trades subject to double-sided reporting, impacting the final data quality assessment and the overall analysis.

As in the above cases, the accuracy of the mapping from TRs' proprietary input formats to the XML message and the ESMA validation rules also plays a pivotal role in the accuracy and presence of the records in the final dataset.

Some of the data quality issues categorised above may be also caused by the complexities associated with the full life-cycle of a set of contracts. These may also include the post-trade processing, such as clearing, netting or compression of derivative contracts²¹. These complexities could lead to difficulties in correctly representing those events in the reporting template, and consequently to data quality issues listed above.

Distinguishing between data quality issues and genuine developments (e.g. market shifts) in EMIR is not straightforward. For example, a sudden increase in the volumes traded and reported submissions of a specific entity may have various reasons – it may represent a change in the trading behaviour, affected by the volatility of the market. It could, however, also stem from reporting errors. The anomaly-detection algorithms, controlling the changes of the outstanding amount over two specific dates, may not be in a position to tell apart those two scenarios, and may consequently generate misleading signals to competent authorities monitoring the

¹⁹ For regulatory technical standards and implementing technical standards, see <https://www.esma.europa.eu/policy-rules/post-trading/trade-reporting>.

²⁰ Position level reporting means net positions resulting from a set of contracts representing fungible products, rather than per individual trade. See TR question 17 in "ESMA Question and Answers on EMIR Implementation": https://www.esma.europa.eu/sites/default/files/library/esma70-1861941480-52_qa_on_emir_implementation.pdf

²¹ See ESMA (2020)

developments in the derivatives markets. This challenge will be further discussed in chapter 5 .

2.4 Challenges in data quality assurance in large-scale financial datasets

In the traditional data warehouse environment, comprehensive and manual data quality assessment and reporting was possible (if not ideal). Ensuring data quality could be thought of as a four-step approach: i. selecting data quality dimensions; ii. designing an assessment plan; iii. assessing the plan against the selected dimensions; iv. acting on results. The elements which are traditionally included in such an assessment plan are the following:²²

- **Validity:** the data is adherent with precision to a given real-world phenomenon
- **Reliability:** the data is defined, measured and collected in the same way all the time with a high degree of trust and reputation
- **Completeness:** the data contains all the information with pertinence at the set, record and element levels
- **Accuracy:** the data is detailed and correct and the units of measure are clear and valid
- **Timeliness:** the data is available on time
- **Integrity:** the data is internally consistent and not biased
- **Uniqueness:** the data does not contain the same information more than once.

However, in Big Data projects the scale of data makes the above process challenging. Thus, in many cases, the data quality measurements can at best be approximations, i.e. need to be described in probability and confidence intervals, and not in terms of absolute values. The challenges posed are mainly driven by the intrinsic features of large-scale financial datasets:²³

- **Volume:** the large-scale amount of data poses analytical challenges as it requires advanced handling techniques (e.g. parallelisation, partitioning and clustering) within a reasonable overhead on time and resources (storage, compute, human effort, etc.)
- **Velocity:** the high-speed of the reporting, collection, processing, visualisation and transformation of data into targeted insights poses timely challenges as by the time data quality assessment is completed, the output might be outdated, therefore it requires advanced processing techniques (e.g. sampling)
- **Variety:** for efficiency purposes, the data might include several data types (structured, semi-structured, and unstructured) coming in from different sources, therefore it requires advanced modelling techniques (e.g. structured metrics)
- **Veracity:** the amount of bias, noise and abnormality might hinder the accuracy and reliability of the dataset, making it difficult to add value created by

²² See Loschin (2010), Chapter "The Organizational Data Quality Program"

²³ See Du (2018), Chapter "Overview of Big Data and Hive"

identifying new patterns and fostering the decision-making process, therefore it requires advanced decision-making techniques (e.g. identification and classification)

- **Visualization:** visual loss due to noise of the excessive information available.

Differentiating the data quality dimensions is the key for matching potential issues against a business need and prioritizing which dimensions to assess and in which order becomes the problem to solve for large-scale datasets. In order to handle big datasets, another challenge is choosing which among the following data reduction strategies to apply:

- **Sampling:** every dataset can be viewed as a sample; the latter is featured by a probability value which can return the fraction of data with representative properties as a result
- **Filtering:** filtering for a specific dimension (e.g. timing) which meets specific conditions is another technique to query large datasets
- **Aggregation:** the dataset grouped by records falling within predefined bins into subsets.

The expression “Big Data” does not simply refer to its vast amount of information but it intrinsically recalls the technology, processes and techniques employed to store, manipulate and share the information on a large scale. On the basis of the difficulties posed by the size of the data and the intrinsic features of the underlying transactions, careful design is necessary in the systems used for data collection and analysis to ensure that the output actually produces some insightful content correctly interpreted.

3 Methods

3.1 Modelling framework

The purpose of the Automated Data Quality tool is to identify and classify the developments in numerical measures of granular, multi-dimensional datasets of financial information. Let’s assume that we have a collection of N_t observations, where each observation reflects details of an individual financial phenomenon, e.g. transaction, instrument, or lifecycle event. In the application to EMIR, each observation will represent the aggregate position of a counterparty on a specific type of derivatives. Furthermore, this information is available for two points in time, described as reference periods, denoted t and $t - 1$.

The dataset can be then described as the following matrix:²⁴

$$X_t = [c \quad d^1 \quad \dots \quad d^U \quad m]$$

²⁴ Please note that a wider dataset with additional columns or higher granularity can be easily converted to such representation by a series of SELECT, GROUP BY, and WHERE operations. The ADQ tool carries out such pre-processing of the dataset within the dimension and entity-level steps, described in section 4.1.

where:

c – a column (vector) identifying the entity involved in the observation, e.g. one of the counterparties; the values of c are alphanumeric strings, uniquely identifying the entity; the value cannot be empty. We denote the unique values in column c as ID^x , hence $c_i \in \{ID^1, ID^2, \dots, ID^x\}$;

d^u – a column (vector) representing dimension u of the U categorical dimensions describing the characteristics of the observation, $u \in \{1, \dots, U\}$; each element of the column can take a value from a predefined list $d_i^u \in \{v_1^u, v_2^u, \dots, v_{Z_u}^u, NULL\}$, where $NULL$ indicates that the given dimension is not available or not relevant for the observation in question;

m – a column (vector) representing a numerical measure²⁵ describing the observation, e.g. market value of the contract; it is assumed that a value of $NULL$ is equivalent to 0 for this column.

We calculate the total delta Δ as the difference of the sum of values of measure m , in two datasets pertaining to reference periods t and $t - 1$:

$$\Delta = \sum_{i=1}^{N_t} m_{i,t} - \sum_{j=1}^{N_{t-1}} m_{j,t-1}$$

However, given that the changes in different observations can have opposing direction, we also define total absolute delta Δ^{abs} , reflecting the sum of absolute differences between observations characterized by the same dimensions $[c, d^1, d^2, \dots, d^U]$. When determining Δ^{abs} , we need to ensure that we calculate the individual Δ 's between the measures referring to the same dimensions. For this purpose, each matrix is supplemented with missing sets of dimensions associated with measure equal to 0. As a consequence, both matrices will contain N' observations, where $N' \geq N_t$ and $N' \geq N_{t-1}$.

$$\Delta^{abs} = \sum_{i=1}^{N'} \Delta_i^{abs} = \sum_{i=1}^{N'} |m_{i,t} - m_{i,t-1}|$$

The goal of the tool is to identify a set of K row vectors $r_k, k \in \{1, \dots, K\}$, each consisting of a set of conditions²⁶ taking one of the following forms:

- $c = ID^x$
- $c \neq ID^x$
- $d^u = v_z^u$
- $d^u \neq v_z^u$

Each r_k identifies a set of conditions, which can be mapped to a set of observations $[ID^i \ v_i^1 \ \dots \ v_i^U]$ contained in the matrices X_t and X_{t-1} . In other words, vectors r_k partition the matrices X_t and X_{t-1} in K pairwise disjoint sets. The vectors r_k have to be mutually exclusive, i.e. none of the observations $[ID^i \ v_i^1 \ \dots \ v_i^U]$ should be mapped at the same time to two different vectors r' and r'' . Consequently, each vector r_k can be assigned a portion of Δ^{abs} ,

²⁵ The measure can take either positive or negative values.

²⁶ Not all dimensions have to be included in the r_k vector, and conversely one dimension may be subject to multiple conditions. See example below.

corresponding to the respective set of observations. If we denote the part of Δ^{abs} corresponding to the vector r_k as $f(r_k)$, we can write:

$$f(r_k) = \Delta_k^{abs}$$

Consequently, we are looking for an algorithm satisfying the following condition:

$$\Delta^{abs} = \sum_{k=1}^K f(r_k)$$

There exist multiple partitions $[r_1, r_2, \dots, r_k]$ of the matrices X_t and X_{t-1} . In the extreme, a collection of vectors representing each possible permutations of allowable values of c, d^1, d^2, \dots, d^U would explain the entirety of Δ . On the other hand, an empty vector would do the same.²⁷ Both solutions are not satisfactory from explanatory point of view. The purpose of the algorithm described in the following section is achieving certain explanatory power, i.e. find a partition that allows us to attribute anomalies in the dataset to a limited number of observation subsets, each of them described by a set of conditions imposed on the dataset dimensions.

3.2 Algorithm

In order to select the solution with optimal explanatory power, and able to tackle the enormous size of the granular regulatory datasets, the ADQ employs binary trees (see Nasiriany (2019), Chapter 7) to select the relevant dimensions d^u and their corresponding values (i.e. a partition $[r_1, r_2, \dots, r_k]$), which contribute to the developments in the dataset.

Nevertheless, this approach is not sufficiently performant for the entity identifiers, as their number can easily reach hundreds of thousands unique values in datasets available to the ECB and ESRB.²⁸ Therefore, the approach consists of two modules, which can be triggered independently, or consecutively, depending on need:

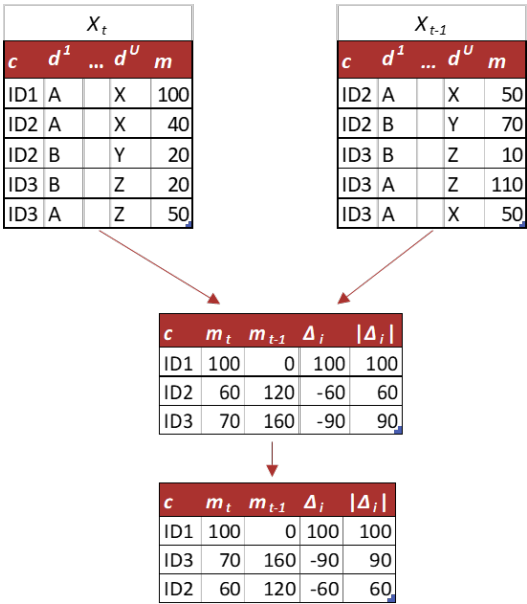
- entity-level analysis module,
- dimensions' analysis module.

3.2.1 Entity-level analysis module

Given that the number of entity columns is restricted to one, the selection of main entities explaining the developments in the dataset simplifies to a simple JOIN / GROUP BY operation, as illustrated in Figure 2.

²⁷ The empty vector should be interpreted as lack of filtering conditions imposed on the dataset, hence all allowable combinations would be covered by an empty vector r_k .

²⁸ Not taking into consideration the performance constraints, the problem could be simplified by treating the c column as one of the dimension column d^u .



Source: Own calculations

Notes: The chart presents a numerical example of the JOIN / GROUP BY operations performed on the dataset to identify entities contributing most to the change in the analysed measure.

The outcome of those operations clearly indicates the entities contributing most to the change in measure in question.

3.2.2 Dimensions’ analysis module

The analysis of dimensions is more complex. Following on the example in section 3.2.1 a set of JOIN / GROUP BY operations allows us to calculate Δ per combination of dimension values (we denote the outcome of this operation as $J(X_t, X_{t-1})$).²⁹

²⁹ Please note that the sum of absolute Δ varies between the entity-level and dimensions’ analysis modules. This is caused by the fact that for each module the absolute value is calculated for sub-aggregates on different level of aggregation.

| X_t | | | | | X_{t-1} | | | | |
|-------|-------|-----|-------|-----|-----------|-------|-----|-------|-----|
| c | d^1 | ... | d^U | m | c | d^1 | ... | d^U | m |
| ID1 | A | | X | 100 | ID2 | A | | X | 50 |
| ID2 | A | | X | 40 | ID2 | B | | Y | 70 |
| ID2 | B | | Y | 20 | ID3 | B | | Z | 10 |
| ID3 | B | | Z | 20 | ID3 | A | | Z | 110 |
| ID3 | A | | Z | 50 | ID3 | A | | X | 50 |

| $J(X_t, X_{t-1})$ | | | | | | |
|-------------------|-----|-------|-------|-----------|------------|--------------|
| d^1 | ... | d^U | m_t | m_{t-1} | Δ_i | $ \Delta_i $ |
| A | | X | 140 | 50 | -90 | 90 |
| B | | Y | 20 | 70 | 50 | 50 |
| B | | Z | 20 | 10 | -10 | 10 |
| A | | Z | 50 | 110 | 60 | 60 |
| A | | X | 0 | 50 | 50 | 50 |

Source: Own calculations

Notes: The figure presents an example of JOIN / GROUP BY operations, performed on an example dataset, in order to calculate the absolute changes characterising different dimension sets.

In this case, however, we have to count with multiple dimensions, and potentially hundreds of thousands of combinations of dimension values. Individual absolute deltas cannot give us any practical insight into which dimension contributes most to the change in the dataset.

To tackle this problem, we employ binary decision trees. We construct a tree where each decision reflects the dimension and corresponding value that best characterize the change in the investigated measure. At each node the tree is split into sub-trees representing subsets of $J(X_t, X_{t-1})$. In order to determine the optimal split, the tool measures the *Gini impurity index*.^{30,31}

$$\text{gini} = 1 - p^2 - (1 - p)^2 = 2p(1 - p)$$

where $p = \frac{|\Delta_i|}{\sum |\Delta_i|}$, i.e. the share of the absolute delta corresponding to particular combination of dimensions in the total absolute Δ .³² The split by dimension with minimum Gini impurity index value provides the highest contribution to the explanation of the development of the measure under examination.

To illustrate the procedure, let's assume that we consider only two dimensions d : "contract type" and "asset class".³³

³⁰ See Nasiriany (2019), p. 166

³¹ The tool can also apply the entropy measure in place of the Gini impurity index.

³² It is important to note that this method requires that the absolute value is calculated on the most granular level, and then summed up in the following steps. Otherwise, the results calculated on different nodes of the tree would not be additive and depending on the path taken by the algorithm we would arrive at different dataset splits.

³³ The following abbreviations apply: SWAP = swap, OPTN = option, INTR = interest rate, COMM = commodity, EQUI = equity.

Calculation of Gini impurity index – numerical example

Figure 4

| Contract type | Asset class | Δ Notional | Contract type | Δ Notional | p | |
|---------------|-------------|------------|---------------|------------|--------|------------------------|
| SWAP | INTR | 100 | SWAP | 270 | 39.71% | → <i>gini</i> = 0.4788 |
| OPTN | INTR | 200 | OPTN | 410 | 60.29% | |
| SWAP | COMM | 150 | | | | |
| OPTN | COMM | 90 | | | | |
| SWAP | EQUI | 20 | | | | |
| OPTN | EQUI | 120 | | | | |

| Asset class | Δ Notional | p | |
|-------------|------------|--------|------------------------|
| INTR | 300 | 44.12% | → <i>gini</i> = 0.4931 |
| COMM | 240 | 35.29% | |
| EQUI | 140 | 20.59% | |

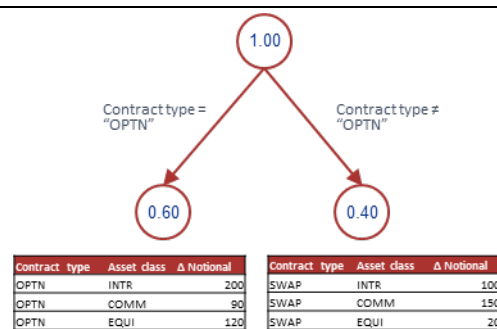
Source: Own calculations

Notes: The figure presents a calculation of the Gini impurity index for two possible groupings of the example dataset

As the Gini impurity index is lower for the "contract type" dimension, the tree is first split into two branches according to the criterion contract type = "OPTN" and contract type ≠ "OPTN". For each sub-tree the weight of the tree is calculated, reflecting the share of the total absolute Δ explained by the sub-tree. The procedure is applied recursively, according to some pre-determined stopping criteria.³⁴

Construction of the decision tree, top node – numerical example

Figure 5



Source: Own calculations.

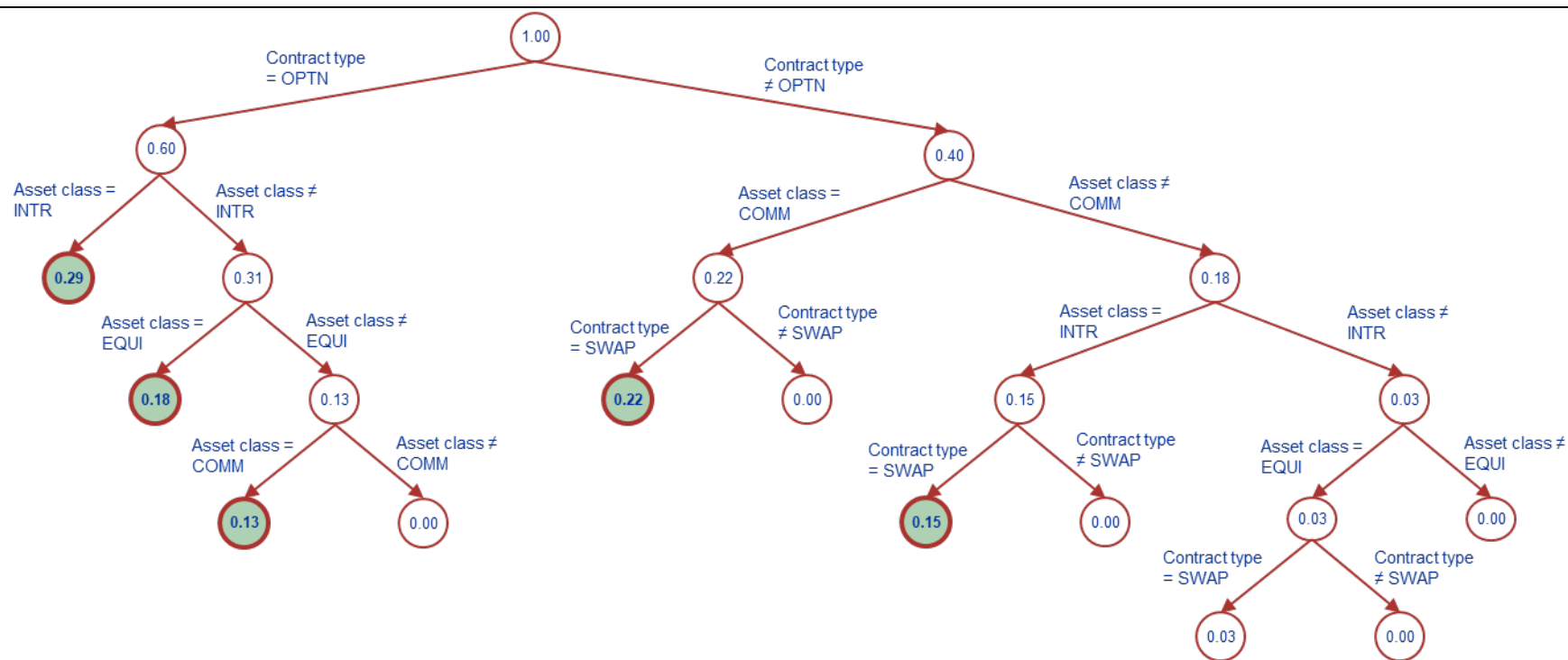
Notes: The figure presents a stylised example of a subtree, split by the contract type =/≠ OPTN. The numerical values inside the circles denote the share of the delta explained by the conditions on the path leading to the respective node. The tables below the child nodes represent the subsets of the dataset, according to the split criterion.

In the simple example above, the algorithm leads to the construction of the following tree, where the leaves with ultimate weight exceeding a predetermined threshold (in this case 0.1) are highlighted in green:

³⁴ The ADQ applies two customisable stopping criteria: maximum depth and minimum threshold change in impurity measure.

Construction of the full decision tree – numerical example

Figure 6



Source: Own calculations.

Notes: Notes: The figure presents a stylised example of a tree built on the basis of the dataset presented in the Figure 4. Each branch is split according to a criterion that minimises the Gini impurity index, until predetermined stopping criteria are fulfilled. The numerical values inside the circles denote the share of the delta explained by the conditions on the path leading to the respective node.

The weight of the leaf is also an indication of how much the conditions along the path leading to it contributed to the change in the total absolute Δ . From the diagram above we can see that the change in the measure over time was driven mainly by interest rate options (29%, INTR and OPTN) and commodity swaps (22%, COMM and SWAP).

Each path leading to a leaf in the constructed tree represents a vector r_k , as described in section 3.1, while the number in the circle is the share of the absolute delta, corresponding to this particular vector.³⁵

Obviously, in this simple example, the algorithm does not constitute a material advantage over visual inspection of the dataset. However, in a scenario with dozens of dimensions and millions of transactions, any manual or semi-manual approach clearly falls short.

3.3 Application to double-sided reporting reconciliation

As stated in section 2.1, some collections of granular financial data foresee the reporting of a particular financial phenomenon (e.g. derivative transaction) by both counterparties linked to the transactions. This type of collection is commonly described as “double-sided reporting”, and data reported under EMIR fall into this category.³⁶ Under the assumption of correct reporting, it can be expected that information referring to the characteristics of the trade is consistent between the sets of data reported by two involved counterparties. Similarly, the quantitative measures describing the contract should coincide if the measurement is made at the same point of time.

The double-sided reporting offers a unique possibility to benchmark the quality of the information reported by the counterparties. If the information reported differs significantly between the two reporting entities, it can be concluded that one of them (if not both) reports incorrect information, or does not report some data at all.

In the following analysis we will focus on the discrepancies in the reporting of the quantitative measures reported by the counterparties. For the purpose of assessing the quality of the reported information and understanding the underlying reasons, it is important to determine: (i) what are the pairs of entities that exhibit largest discrepancies, and (ii) if there are any specific characteristics of the observations underlying those differences that could give additional insight into the reasons for the discrepancies.

The procedure developed in section 3.1 can be easily adjusted to the case of double-sided reporting. Let’s assume that we have a dataset of financial information in the following form:

$$X = [c^R \quad c^O \quad d^1 \quad \dots \quad d^U \quad m]$$

The notation from section 3.1 applies accordingly with the following additions:

³⁵ To be precise, vector r_k , includes also the counterparty identifier c , identified in the entity-level analysis module (see section 3.2.1). See discussion in section 4.1 on how the two models interact in practice.

³⁶ This does not apply if one of the counterparties is resident outside of the EU, or is a private individual.

c^R – identification of the counterparty that reported the observation in question, also "reporting counterparty"

c^O – identification of the other counterparty that was involved in the observation, also "other counterparty"

The dataset X is double-sided, if and only if:

$$\forall i \exists j (c_i^R = c_j^O \wedge c_j^R = c_i^O)$$

In other words, for each observation there exists a corresponding one, with the same pair of counterparties, but in reverse. A tuple $\{i, j\}$, satisfying the condition $(c_i^R = c_j^O \wedge c_j^R = c_i^O)$ will be further described as a "paired position".

The dataset X can be split into two disjoint sets X' and X'' , such that elements of each paired position are separated, namely:³⁷

$$\forall i, j (c_i^R = c_j^O \wedge c_j^R = c_i^O) \rightarrow (X_i \in X' \wedge X_j \in X'') \vee (X_i \in X'' \wedge X_j \in X')$$

where c^R and c^O columns are replaced by a new identifier c^{pair} , which is a concatenation of the c^R and c^O identifiers, ordered alphabetically.³⁸ In this way the c^{pair} becomes a key, linking paired positions segregated into X' and X'' . The split criteria can be arbitrary, although, obviously, it is reasonable to assume the criteria following certain business logic. For instance, in case of a dataset that contains transactions between Central Clearing Counterparties (CCPs) and Clearing Members (CMs), it is reasonable to split X along the criterion $c^R \in \text{CCPs} / c^R \in \text{CMs}$. In other cases, an artificial splitting criterion may be needed, for example an alphabetical ordering of entities' IDs.

Consequently, we arrive at two datasets:

$$\begin{aligned} X' &= [c^{\text{pair}'} \quad d^{1'} \quad \dots \quad d^{U'} \quad m'] \\ X'' &= [c^{\text{pair}''} \quad d^{1''} \quad \dots \quad d^{U''} \quad m''] \end{aligned}$$

In this way the above problem reduces to the one described in section 3.1, with the datasets X' and X'' corresponding to X_t and X_{t-1} , respectively. Applying the algorithm described in section 3.2 results in identifying the largest differences in the information reported by counterparties, as well as explaining any patterns in characteristics of the trades, for which the differences occur.

The above reasoning can be also applied to reconciliation of other types of information, which is represented by two disjoint subsets of data referring to a particular reference period.³⁹ For brevity, we denote the algorithm described in section 3.2 as "time-series analysis", and the one characterised in section 3.3 as "intraday analysis".

³⁸ For example, if $c_i^R = \text{"ABC1"}$ and $c_i^O = \text{"XYZ2"}$, then $c_i^{\text{pair}} = \text{"ABC1;XYZ2"}$

³⁹ One example could be long and short positions of a CCP in specific products – the CCP by construction should have no net exposures, i.e. the absolute value of the short position should be equal to the long position.

4 Application of the ADQ method to EMIR data

4.1 ADQ Process

The ADQ method finds an application in a large-scale dataset such as EMIR. It allows to identify timely **data quality issues** and **developments of the derivatives market** overcoming the challenges posed by the size of the dataset. The application leverages on the EMIR IT system that is in place at the ECB since 2017. The ECB implementation relies on a Hadoop infrastructure and allows data consumers at the ECB to perform certain analytical activities that previously took hours or days, in a matter of minutes. The ADQ process applied to EMIR data is integrated in a set of automated daily activities, such as processing, enrichment,⁴⁰ and data quality management, carried out at the ECB to ensure the timely provision of EMIR data to users.

The method is applied to the trade state reports of EMIR⁴¹ that include the outstanding trades on a given date (i.e. reference period). The input to the process is provided by a set of parameters and by monitoring information resulting from the daily processing of the EMIR data. The measures, the dimensions, the concentration threshold, the input dataset, and the type of the analysis (time-series vs. intraday) are defined in the set of parameters that is provided to the process in the form of a JSON file. Varying the set of parameters allows running different jobs to analyse the dataset from different points of view. The monitoring information, in turn, allows to assess the completeness of the data reported by the trade repositories (TRs) on a given reference period. In case of uncomplete data submitted by a TR, all data from this TR are removed from the dataset that will be analysed.

Once these preliminary processes are carried out, we move to the analysis of the data according to the two modules: entity-level analysis and dimensions' analysis. These are executed sequentially and in two parallel workflows:

1. *Entity-level analysis followed by dimensions' analysis*: first the main entities contributing the most to the changes in the quantitative measure are selected and then the dimensions' analysis is applied to these entities to determine the driving factors of the changes observed.
2. *Dimensions' analysis followed by entity-level analysis*: first the combinations of dimensions with the largest explanatory power of the changes in the quantitative measure are selected and then the module of the entity-level analysis is applied to determine the main players responsible for the changes observed.

Once both workflows are concluded, the results are summarized in an HTML report circulated via e-mail to the stakeholders of the EMIR dataset. Such report

⁴⁰ The enrichment is a process, where the dataset is complemented with information from other reference datasets. At the ECB the data collected under EMIR is enriched with supplementary information on entities, benchmarks and underlying instruments from several internal and external sources.

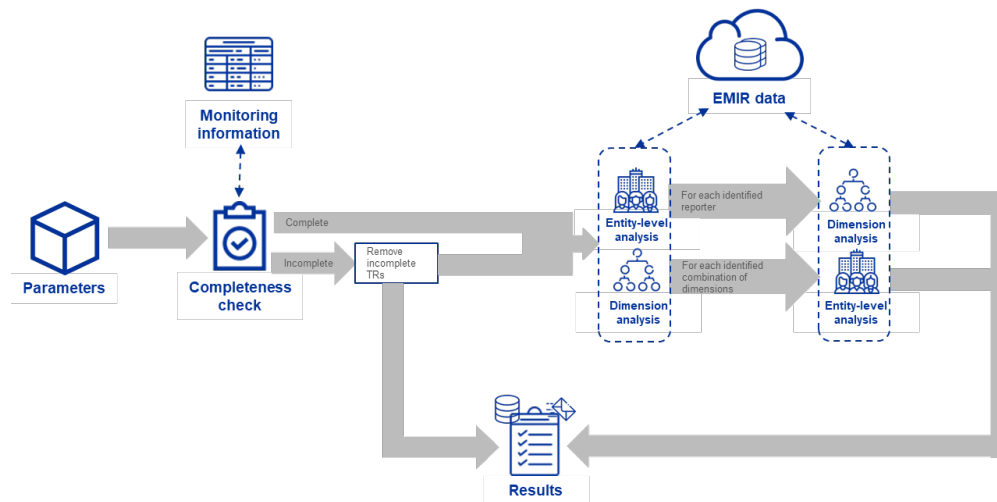
⁴¹ The TRs provide the authorities with two main types of reports:

- trade state: information on all derivative contracts outstanding on a given reference date;
- trade activity: information on new trades and lifecycle events affecting existing trades reported within a particular reference date.

provides information on the set of parameters applied, the main entities and dimensions that drive the changes observed in the quantitative measure resulting from the two workflows. In addition, the results of the ADQ algorithm are stored in a database that allows for feeding other processes and systems, for instance graphical tools to visualise the results in charts and dashboards.

ADQ architecture

Figure 7



Source: Own work

4.2 What do we measure?

The EMIR data include several quantitative measures that can be analysed through the ADQ method, such as the notional, the contract value and the initial margin received defined according to the EMIR Regulatory Technical Standards⁴² as follows.

- **Notional:** The reference amount from which contractual payments are determined. In case of partial terminations, amortisations and in case of contracts where the notional, due to the characteristics of the contract, varies over time, it shall reflect the remaining notional after the change took place.
- **Value of contract value:** Mark to market valuation of the contract, or mark to model valuation where applicable under Article 11(2) of Regulation (EU) No 648/2012. The CCP's valuation are to be used for a cleared trade.
- **Initial margin received:** Value of the initial margin received by the reporting counterparty from the other counterparty.

A set of parameters for each measure is created and passed to the ADQ process together with all the other relevant pieces of information, for instance the source

⁴² Commission Delegated Regulation (EU) No 148/2013 of 19 December 2012 supplementing Regulation (EU) No 648/2012 of the European Parliament and of the Council on OTC derivatives, central counterparties and trade repositories with regard to regulatory technical standards on the minimum details of the data to be reported to trade repositories: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02013R0148-20171101&from=EN>

dataset or the output database where the results will be stored. This results in different instances of the process that run independently for each measure.

Part of the initial set of parameters are the pieces of information to configure the two ADQ modules, namely the entity-level and the dimensions' analysis. The former requires as input information the identifier of the entity, for instance the **reporting counterparty** of a trade; the other module is specified by a set of dimensions. We identified 8 dimensions of the trade that are significant to explain the changes observed in EMIR data: the **asset class**, the **contract type**, the **currency** of the notional, the **clearing flag**, the **intragroup flag**, the **execution venue**, the **type of observation** reported in the trade activity report,⁴³ and the **trade repository** that submits the trade to authorities.

Another significant element characterising each instance of the ADQ process is the type of analysis. We differentiate between **time series** and **intraday** analysis:

- 1- *Time series analysis*: the trade state table at T is compared with the one at T-1. This type of analysis allows detecting data quality issues such as implausible values of notional and contract values for several reporting entities. In addition, the analysis over time of the data provides insights to market developments. This is the case of the increase of initial margins occurred with the outbreak of the pandemic (March 2020) or the movement of initial margins due to the developments in the European gas market in October 2021.
- 2- *Intraday analysis*: the information reported at T by the two legs of trades is compared. The intraday analysis is applied to two categories of reporting entities: CCPs and clearing members (CMs). The aim is to measure the discrepancy between the notional reported by pairs of CCPs and CMs. Therefore, the application of the method requires the aggregation of raw data computing the total notional by each pair CCP-CM and CM-CCP. This allows to detect issues of under-/over-reporting by one of the two sides for specific trades defined by the dimensions.

In the current set-up at the ECB, multiple workflows are triggered automatically every morning covering the above types of analysis with the run-time amounting to 5-10 minutes per workflow. The outcome of those workflows is shared with the group of EMIR operators, who then act on the findings. Conditional on the type of the issue, the matter may be further investigated, reported to an adequate authority (e.g. ESMA), raised to the attention of the respective trade repository and/or shared with internal users of the dataset.

4.3 Extension of the work

The flexibility and customisability of the process are ensured by its full parametrisation. This is not limited to the selection of the quantitative measures and dimensions to be explored but it also includes the possibility of running preliminary transformation to raw data that will serve as input data to be analysed. Therefore, the

⁴³ I.e. transaction- or position-level record, see also footnote 20.

method can be easily extended both in terms of instances within EMIR data perimeter and beyond to other datasets.

Considering further applications to EMIR, further workstreams could be implemented, such as

- additional module for the processing of data quality issues detected for transmission to ESMA as authority in charge of the EMIR data quality management.
- Application to trade state reports and trade activity reports to analyse the consistency between the two kinds of reports.
- Correcting for the potential temporal misalignment of reporting, e.g. when entities report the corresponding information on different days.
- Time series analysis applied with a larger lag, i.e. compare data at T with the data at T-30.
- Building complex customisable pipelines from ADQ modules, e.g. applying sequentially the entity-level analysis on reporting entities followed by the entity-level analysis on the other counterparties to the trades for the 3 main reporting entities and then concluding with the dimensions' analysis.

Regarding the extension to other datasets, the work started already and will be further complemented to apply the method to SFTR data both implementing the time series analysis and the intraday one to CCPs and CMs.

5 Disentangling anomalies

Detecting anomalies in the financial system through the analysis of a broad set of indicators represents the foundation of systemic risk monitoring. This set of indicators typically include the build-up of large exposures, concentration and interconnectedness, or the identification of specific exposures that are particularly sensitive under certain scenarios (e.g. to repricing and margin calls). Once detected, these anomalies could signal relevant financial stability developments and inform policymakers' actions.

However, in case of low quality of the data reported by market participants, these anomalies may simply reflect inaccurate information, rather than a development relevant from a financial stability viewpoint. In turn, this generates uncertainty in interpreting analytical results, which impairs monitoring capabilities, potentially leading to wrong conclusions. Additionally, uncertainty can lead to rely less on the data, thus spend less effort on its analysis, which can further worsen their quality as more issues are undetected. Avoiding this self-fulfilling spiral should therefore be a strategic objective of regulators.

From a research standpoint, low data quality can be often dealt with by narrowing the analysis by selectively restricting samples (e.g. to remove implausible observations), correcting outliers, or making specific assumptions. From a policy perspective, however, the downstream impact of both low data quality and the potential assumptions to deal with it can be significant if not carefully considered. In

fact, working in the presence of uncertainty creates both analytical and operative issues.

First, from an analytical perspective, the low data quality reduces the reliability of the results opens to potential false positives (e.g. when a substantially high value for an indicator measuring concentration is due to erroneous data) or false negatives (e.g. a low value of a relevant bilateral exposure due to missing or erroneous data). In this case, policymakers need to embark in a time-consuming case-by-case inspection to understand the potential root causes and gauge the impact of low-quality data. Moreover, they may have to judge whether the impact of low data quality is material or not, adding further assumptions to the analysis.

Second, from an operative perspective, low data quality makes financial stability monitoring substantially more challenging when it needs to be performed “at scale” and only partly automated: working case by case is not operationally feasible in the presence of very large datasets, reported with high frequency, e.g. daily. The rationale to scale up analytical systemic risk monitoring lies not only in the size of newly available data, but also on the evolving nature of risks in an increasingly complex, interconnected, and adaptive financial system. Analytical scaling also shows an intrinsic dimensionality problem: the number of potential indicators and their levels of aggregation can become easily extremely large.

In this section, we outline an approach to use the framework illustrated in this paper to mitigate this problem. The main intuition underpinning this application is straightforward. By *disentangling* between the two main sources of anomaly in the data, we can reduce the odds of encountering both false positives and false negatives. If the data is of high quality, the probability that an anomaly is a significant financial stability development is higher, whereas if the data is of low quality, this probability decreases. Leveraging insights gained from the ADQ framework, policymakers can discern genuine financial stability signals with less uncertainty.

The key feature of the ADQ framework is its capability to pinpoint the primary contributors to data quality issues by progressively breaking down along the relevant dimensions. This allows policymakers to make informed methodological decisions when interpreting a financial stability signal in the presence of suboptimal data quality.

Utilizing the ADQ framework can reduce uncertainties related to analytical outcomes. The framework offers at least two ways to achieve this: upstream and downstream.

- 1) **Upstream.** The first way is to start from the anomaly, as detected by the ADQ tool, and then analyse the quality of the underlying data. Once an anomaly is detected via the ADQ framework in the time-series mode, the ADQ can be applied in intraday mode on the two dates: if the ADQ in intraday mode does not lead to a data quality issue on both dates, then the anomaly detected in time-series mode is less likely due to a data quality issue. On the contrary, if the ADQ run in intraday mode returns a data quality issue, then the anomaly can be attributed to low data quality, according to which day it has appeared.
- 2) **Downstream.** The second way to reduce uncertainty via the ADQ is to start from a data quality issue and understand which analyses this may impact downstream: any anomaly detected which uses observations for which it is known that a data quality issue is present will have a higher likelihood to be due to a data quality issue.

It is important to remark that, while this use of the ADQ framework can be helpful to facilitate policymakers' work, it can never substitute the value of having high quality data. In the following two examples, we are going to illustrate how this approach can be used.

Example 1: margin calls. The first example uses the ADQ framework to disentangle the anomaly upstream. Let us imagine we observe a substantial increase in the margins reported by a Central Counterparty during a crisis period, detected by running the ADQ in time-series mode. While one may have anecdotal knowledge of potential margin calls, it is still unsure whether the margin call is in the order of magnitude signaled by the CCP and who are the clearing members, products, and clients affected by these margin calls. To this end, the first step of the disentangling procedure would be to run an ADQ procedure on the delta between two different dates to understand the relevant dimensions (e.g. the clearing members). Let us now imagine that the margin call is explained by a substantial fraction (e.g. more than 50% of the total margin increase) by one individual clearing member and the policymaker is unsure whether this is a relevant financial stability signal or it is due to a problem in the data reported by the CCP. If running a further ADQ process in intraday mode on the margin reported from the clearing members' perspective shows no data quality issue, then the anomaly is likely relevant from a financial stability viewpoint. If, on the contrary, a discrepancy between the CCP and the clearing member is detected in intraday mode, the anomaly is more likely to be explained by a data quality issue.

Example 2: concentration. The second example uses the ADQ framework to disentangle the anomaly downstream. Let us suppose we observe a discrepancy between the exposures (proxied by notional amounts) reported by two EU counterparties (A and B). Running the ADQ in intraday mode suggests that the issue is due to missing contracts from counterparty A. In this case, any further anomaly including data reported by counterparty A is more likely to be due to data quality issues, rather than being a financial stability signal.

6 Conclusions

Despite policymakers' efforts in collecting granular level data after the global financial crisis, persistent and pervasive data quality issues increase opacity, thereby hampering the ability to analyse data and produce effective policy responses. This paper describes a novel framework to identify factors underlying the developments in large, granular datasets of financial information and proposes several applications based on data on derivatives collected under the EMIR Regulation. We show how tools build on this framework have been successfully deployed on the ECB IT infrastructure, and are regularly used to decompose the changes in certain measures of derivative markets into data quality issues and genuine developments that may have potential impact on the financial stability.

One of the essential features of the tool is its customisability, allowing the relevant staff to apply the solution to various datasets, measures, and dimensions, as well as employ any initial filtering deemed necessary. Thus, while the original application was data collected under EMIR, we plan to apply the tool to other granular datasets available at the ECB and ESRB.

Given the structured output of the tool, this daily process can be further integrated into other daily monitoring operations on the granular datasets, significantly reducing efforts needed to ensure that the information ingested is correct, allowing also for considerable reduction of the time needed for identification and reporting of data quality issues.

References

- Abad J., Aldasoro I., Aymanns C., D'Errico M., Fache Rousová L., Hoffmann P., Langfield S., Neychev M., Roukny T. (2016). Shedding light on dark markets: First insights from the new EU-wide OTC derivatives dataset. ESRB Occasional Paper Series No 11, 2016.
- Apicella E., Ciullo A., Übelhör C., Marques P., D'Errico M. (2023). Monitoring *at scale*. Forthcoming.
- Brunnermeier, M.K., Gorton, G. and Krishnamurthy, A., 2012. Risk topography. NBER Macroeconomics Annual, 26(1), pp.149-176.
- Carraro T., Fache Rousová L., Furtuna O., Ghio M., Kallage K., O'Donnell C., Vacirca F, Zema S. M. (2021). Lessons learned from initial margin calls during the March 2020 market turmoil. *ECB Financial Stability Review*, November 2021
- Du D. (2018). Essential Techniques to Help You Process, and Get Unique Insights from, Big Data, 2nd Edition. Packt Publishing, Limited.
- Duffie D. (2011). A 10-by-10-by-10 Approach.
<https://www.darrellduffie.com/uploads/policy/Duffie10By10By10July2011.pdf>
- ECB Banking Supervision (2018). Report on the Thematic Review on effective risk data aggregation and risk reporting.
https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.BCBS_239_report_201805.pdf
- Erl T., Khattak W, Buhler P. (2016). Big Data Fundamentals: Concepts, Drivers & Techniques. Pearson
- ESMA (2021). EMIR and SFTR data quality report 2020,
https://www.esma.europa.eu/sites/default/files/library/esma80-193-1713_emir_and_sftr_data_quality_report.pdf
- ESMA (2020a). Final Report. Technical standards on reporting, data quality, data access and registration of Trade Repositories under EMIR REFIT
- ESMA (2020b). Report to the European Commission on post trade risk reduction services with regards to the clearing obligation under EMIR Article 85(3a).
https://www.esma.europa.eu/sites/default/files/library/esma70-156-3351_report_on_ptrr_services_with_regards_to_the_clearing_obligation_0.pdf
- ESMA (2017). Final Report – Guidelines on transfer of data between Trade Repositories. https://www.esma.europa.eu/sites/default/files/library/esma70-151-552_guidelines_on_transfer_of_data_between_trade_repositories.pdf
- European Systemic Risk Board (2022). ESRB's view regarding data quality issues and risks for financial stability.
https://www.esrb.europa.eu/pub/pdf/other/esrb.letter220713_on_data_quality_issues~18eccb6993.en.pdf
- European Systemic Risk Board (2020a). Liquidity risks arising from margin calls. Available at:
https://www.esrb.europa.eu/pub/pdf/reports/esrb.report200608_on_Liquidity_risks_arising_from_margin_calls_3~08542993cf.en.pdf

European Systemic Risk Board (2020b). Secretariat staff's response to ESMA's consultation paper on technical standards on reporting, data quality, data access and registration of trade repositories under EMIR Refit. https://www.esrb.europa.eu/pub/pdf/other/esrb.letter200812_response_to_ESMAs_consultation_paper~bef2263d90.en.pdf?0b34782ea7527a3e8a322cb3b124c097

FSB, IMF (2009). The Financial Crisis and Information Gaps, Report to the G-20 Finance Ministers and Central Bank Governors. https://www.fsb.org/wp-content/uploads/r_091029.pdf

Lai R., Potaczek B. (2019). Hands-On Big Data Analytics with Pyspark : Analyze Large Datasets and Discover Techniques for Testing, Immunizing, and Parallelizing Spark Jobs. Packt Publishing, Limited.

Loshin D. (2010). The Practitioner's Guide to Data Quality Improvement. Morgan Kaufman

Mahanti R. (2019). Data Quality: Dimensions, Measurement, Strategy, Management, and Governance. ASQ Quality Press

Nasiriany S. Thomas G., Wang W., Yang A. (2019). A Comprehensive Guide to Machine Learning. <https://www.eecs189.org/static/resources/comprehensive-guide.pdf>

Sambasivan N., Kapania S., Highfill H., Akrong D., Paritosh P., Aroyo L. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. <https://research.google/pubs/pub49953/>



EUROPEAN CENTRAL BANK

EUROSYSTEM

Anomaly intersection: disentangling data quality and financial stability developments in a scalable way

Gemma Agostoni (ECB)

Louis de Charsonville (McKinsey & Company)

Marco D'Errico (ECB, ESRB Secretariat)

Cristina Leone (BIS)

Grzegorz Skrzypczynski (ECB)

The views expressed here are of the authors and do not necessarily represent the views of the associated institutions

ECB-UNRESTRICTED
FINAL



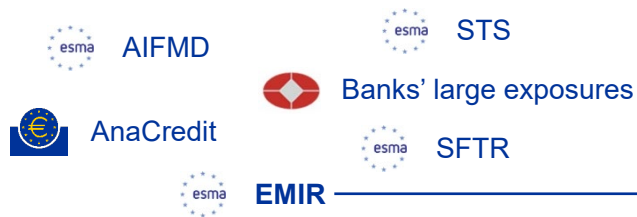
IFC-Bank of Italy workshop
Data Science in Central Banking: Applications and tools
15 February 2022

Background

- Following the financial crisis of 2008/2009, policymakers now have access to several large scale & granular-level datasets, implying the need to scale up monitoring and analytical work
- However, persistent and pervasive data quality issues (largely due to reporting agents and trade repositories) hamper this process, reducing transparency
- Policymakers are now facing a double challenge: how to disentangle developments that are relevant from a financial stability perspective from those resulting by bad data quality?

Quality in datasets of granular financial information

The financial crisis of 2008/2009 led to implementation of multiple high-frequency collections of **granular financial information**.

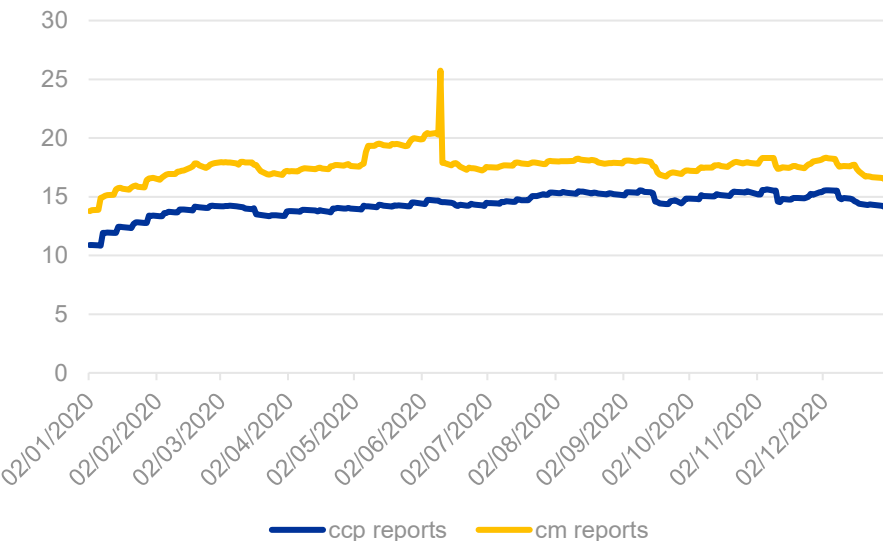


Those dataset pose a unique challenge to the regulators due to their enormous size and **insufficient data quality**

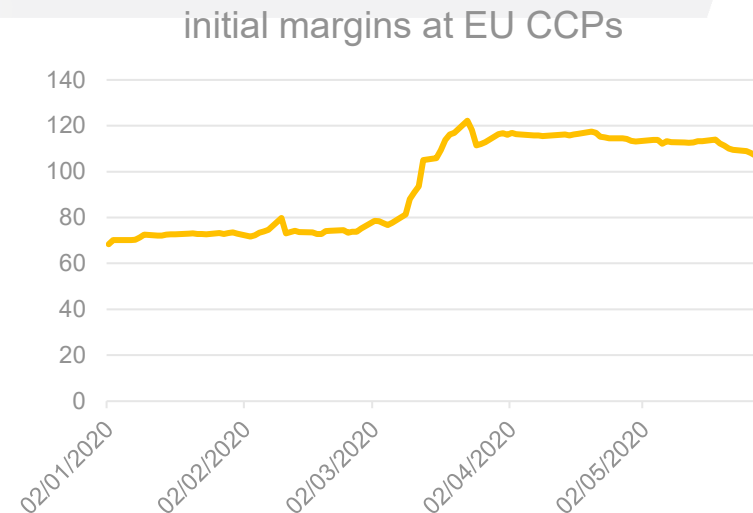
| IT and TR issues | Misreporting |
|---|---|
| 5 million duplicated trades sent daily over a month | Inconsistent information reported by CCPs and clearing members |
| Negative values incorrectly changed to absolute values for 1.5 years | Incorrect signs of contract values |
| Missing collateral reports (IM + VM) for large CCPs | Not following the reporting guidelines (e.g. collateral portfolio code, fx swaps) |
| Information reported by counterparties not passed onto the reports for authorities (e.g. "Asset class", "Contract type", collateral variables) | Implausible numerical values (reaching EUR trillions) – also by CCPs and other large entities |
| Disappearing negative rates | CCPs reporting no outstanding positions at end-day |



Disentangling data quality and financial stability developments

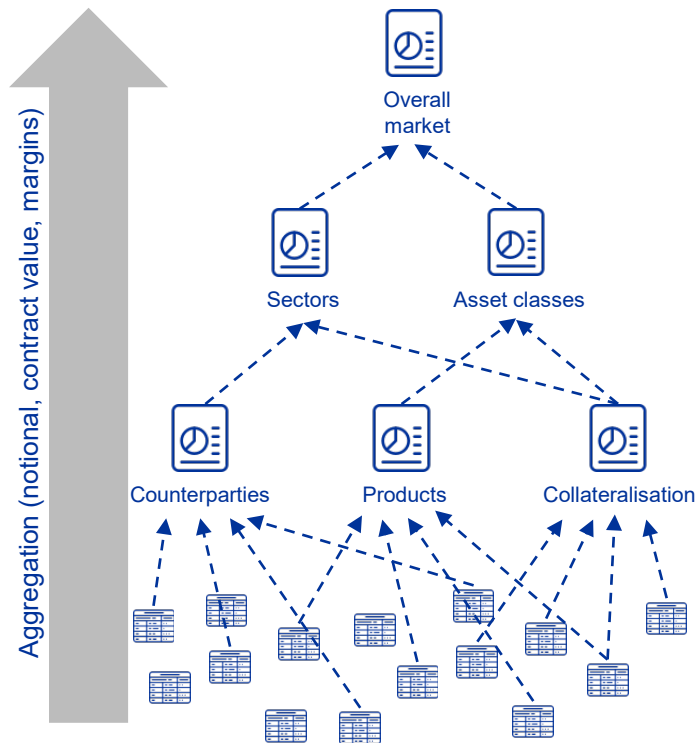


Notional amounts between EU CCPs vis-à-vis
1200+ EU clearing members differ substantially
-> data quality issue



Is the dramatic increase in initial margins at
EU CCPs during the March 2020 turmoil a
development or a data quality issue?

Bridge between micro and macro



Granular data like EMIR blurs the line between macro and micro – **the final users can seamlessly zoom-in and zoom-out across aggregation levels** from analysing individual trades to assessing the overall market.

But trying to apply **traditional manual or semi-automatic tools** to data quality management is like looking for a **needle in a haystack**.



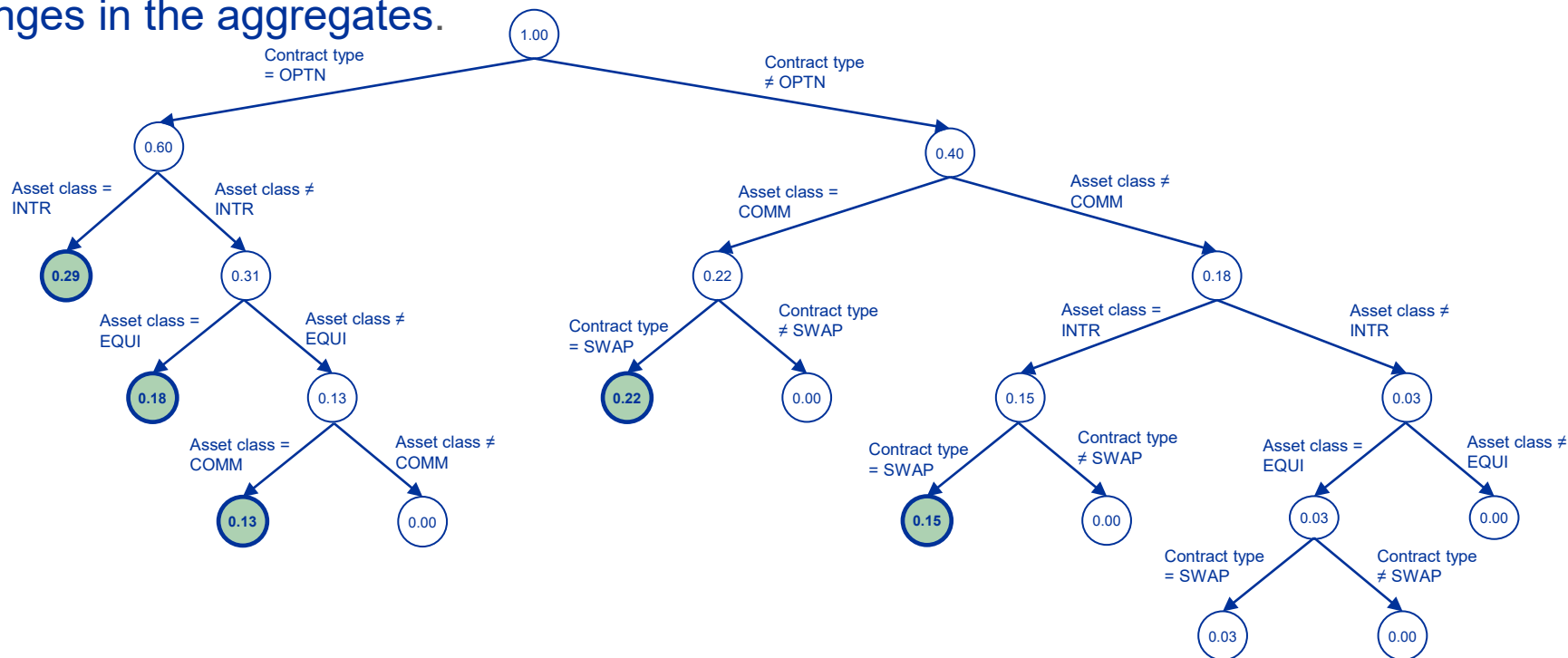
There's **no time** to laboriously look for the answers, when something unusual happens.

We want **the answers** to wait for us **in our mailboxes every morning** – before we even ask the question!



Dimension analysis

We use decision trees to determine dimensions that best characterize the changes in the aggregates.



Entity-level analysis

- **Aggregate the measure** reported by the relevant entities
- **Compare the values** reported in two reference periods analysed
- **Select the entities** with biggest impact on the change in the measure

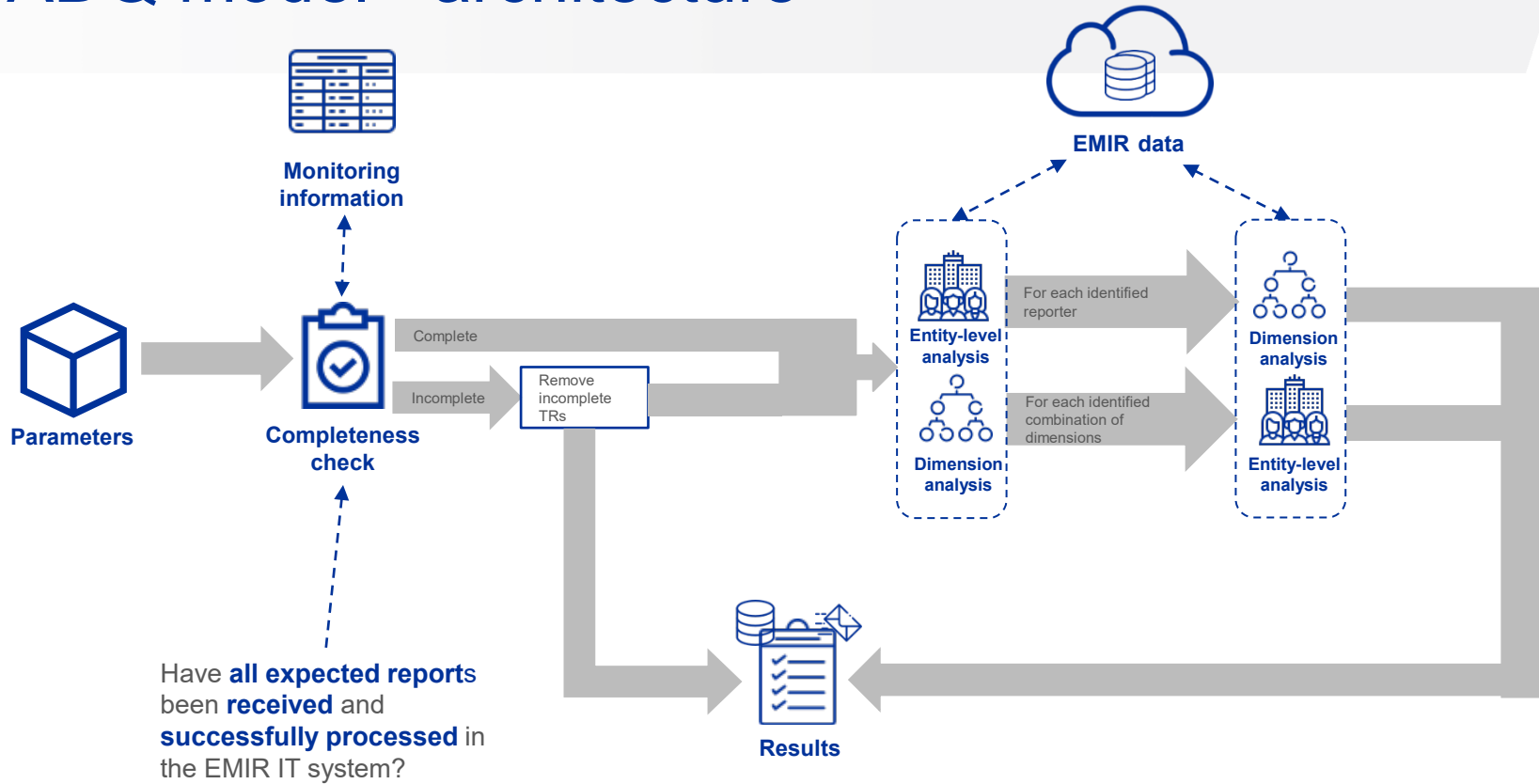
| t | | | | |
|--------|----------------|-----|----------------|----------|
| Entity | d ¹ | ... | d ^U | Notional |
| ID1 | A | | X | 100 |
| ID2 | A | | X | 40 |
| ID2 | B | | Y | 20 |
| ID3 | B | | Z | 20 |
| ID3 | A | | Z | 50 |

| t-1 | | | | |
|--------|----------------|-----|----------------|----------|
| Entity | d ¹ | ... | d ^U | Notional |
| ID2 | A | | X | 50 |
| ID2 | B | | Y | 70 |
| ID3 | B | | Z | 10 |
| ID3 | A | | Z | 110 |
| ID3 | A | | X | 40 |

| Entity | Notional _t | Notional _{t-1} | Δ | Δ |
|--------|-----------------------|-------------------------|-----|-----|
| ID1 | 100 | 0 | 100 | 100 |
| ID3 | 70 | 160 | -90 | 90 |
| ID2 | 60 | 120 | -60 | 60 |

The model allows the reporters to be treated as dimensions, however, given the number of unique reporters this would significantly reduce the performance

ADQ model - architecture



Conclusions & way forward

ADQ – main features



- Timeliness
- Flexibility
- Analytical support
- Dataset agnosticity

Way forward



- Supporting the analysis with information from **activity reports** (flow)
- Building **complex, configurable pipelines**
- **Semi-automated transmission** of issues to **ESMA**

THANK YOU FOR YOUR ATTENTION



IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Introducing explainable supervised machine learning into interactive feedback loops for statistical production systems¹

Thomas Gottron, Georgios Kanellos and Johannes Micheler, European Central Bank
José Martínez, Solenix,
Carlos Mougan, University of Southampton

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Introducing explainable supervised machine learning into interactive feedback loops for statistical production systems

Carlos Mougan¹ · George Kanellos² ·
Johannes Micheler² · Jose Martinez Heras³ ·
Thomas Gottron²

Abstract

Statistical production systems cover multiple steps from the collection, aggregation, and integration of data to tasks like data quality assurance and dissemination. While the context of data quality assurance is one of the most promising fields for applying machine learning, the lack of curated and labeled training data is often a limiting factor.

The statistical production system for the Centralised Securities Database features an interactive feedback loop between data collected by the European Central Bank and data quality assurance performed by data quality managers at National Central Banks. The quality assurance feedback loop is based on a set of rule-based checks for raising exceptions, upon which the user either confirms the data or corrects an actual error.

In this paper we use the information received from this feedback loop to optimize the exceptions presented to the National Central Banks, thereby improving the quality of exceptions generated and the time spent by the users for assessing those exceptions. For this approach we make use of explainable supervised machine learning to (a) identify the types of exceptions and (b) to prioritize which exceptions are more likely to require an intervention or correction by the NCBs. Furthermore, we provide an explainable AI taxonomy aiming to identify the different explainable AI needs that arose during the project.

¹ University of Southampton, United Kingdom
E-mail: C.Mougan-Navarro@soton.ac.uk

² Directorate of Statistics, European Central Bank
E-mail: George.Kanellos@ecb.europa.eu, Johannes.Micheler@ecb.europa.eu, Thomas.Gottron@ecb.europa.eu
This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

³ Solenix Deutschland GmbH
E-mail: Jose.Martinez@solenix.ch

1. Introduction

Providing statistical data and information of the highest quality is a core task of the European System of Central Banking (ESCB). Assuring the quality of data in statistical production systems is crucial as data is used in the policy decision-making processes. Statistical production systems rely on domain experts with an outstanding understanding of the data. These experts ensure that the information used in the compilation of statistical products is of the highest quality based on their expertise and domain knowledge.

The Centralised Securities Database (CSDB) [25] is a securities database with the aim of holding complete, accurate, consistent, and up-to-date information on all individual securities relevant for the statistical and, increasingly, non-statistical purposes of the ESCB. Ensuring quality in such a system is challenging given the amount of data that needs to be monitored. So far, quality assurance has been achieved through a combination of data aggregation and static quality rules. In addition, *business experts* with the role of Data Quality Manager use their experience for assessing data quality and manual intervention in case of issues.

The benefit of using expert knowledge for quality assurance of statistical products is two-fold: (i) all the knowledge and expertise of the domain experts is reflected in the final data and (ii) the overall confidence about the final product increases on the side of the producers and consumers of official statistical data. However, as the volume and granularity of the data increase, this process becomes progressively difficult to maintain. In the case of the CSDB, which contains information on approximately 7 million alive securities, it is almost impossible to manually assess the quality of granular data without substantially increasing the number of experts.

Over the years various approaches have been developed to support and automate quality assurance procedures. However, these approaches, based on the above mentioned aggregations and static rules, have certain limitations:

- **Data aggregation:** Assessing the quality of the data on an aggregated level condenses the information that needs to be checked to a manageable amount. The disadvantage of this approach is that aggregation might hide granular data quality issues on the level of individual observations. The risk of obfuscation depends on the type and level of aggregation used.
- **Static quality rules:** Another way to ensure data quality for big data, is using a fixed set of rules to identify potential quality issues. These rules are based on expert knowledge and formalize certain aspects of their domain knowledge. Such rules flag individual records that, for instance, exceed a predefined threshold. The flagged observations typically represent only a small fraction of the entire data set, allowing the field experts to focus their analysis on a limited amount of data. The shortcoming of this approach is that relying on a fixed set of static rules is not flexible enough to capture dynamics in data nor does it automatically adjust to new insights provided by users.

In this paper we make the following contributions:

- (i) First, we identify and describe the types of actions that experts have been performing historically in the context of their data quality assurance tasks. This interaction forms the basis for our feedback loop. The information obtained in this way permits the formulation of detecting data quality

issues as a supervised machine learning problem. The concrete objective is to predict the probability that a granular data instance is an outlier.

- (ii) Once we are able to identify the probability of a data quality issue, we process the results to produce a set of ranked exceptions. The ranking is based on the product of the amount outstanding or market capitalization for a certain financial instrument multiplied by the soft prediction of the machine learning model that the observation is an outlier. This ranking allows data quality managers to focus on those observations with a high risk of being wrong while keeping business priorities in mind.
- (iii) Furthermore, modern practices for machine learning models require (i) high predictive accuracy and (ii) model interpretability [18]. One of the barriers that artificial intelligence (AI) is facing regarding practical implementation in the financial and public policy sectors is the inability to explain or to fully understand the reasons why an algorithm takes certain decisions [1, 28]. With both reasons in mind, in Section 4, we provide desiderata for explainability in AI by categorising users and analysing potential needs.

The rest of the paper is organized as follows: First, we give an overview of the data and existing data quality assurance processes and tools. In Section 3 we describe our approach, including the data pre-processing and feature engineering steps, the methods used to solve the given task and an evaluation of performance. Finally, we provide a set of explainability needs for the model before concluding the paper with a summary and a review of current limitations towards future work.

2. Dataset Overview

The Centralised Securities Database (CSDB) [25] is a securities database with the aim of holding complete, accurate, consistent, and up-to-date information on all individual securities relevant for the statistical and, increasingly, non-statistical purposes of the European System of Central Banks (ESCB).

It is a single information technology infrastructure that contains master data on securities (e.g. outstanding amounts, issue and maturity dates, type of security, coupon and dividend information, statistical classifications, etc.), issuers (identifiers, name, country of residence, economic sector, etc.) and prices (market, estimated or defaulted) as well as information on ratings (covering securities, issuance programs, and all rated institutions independently of whether they are issuers of securities).

The CSDB covers securities issued by EU residents; securities likely to be held and transacted in by EU residents; and securities denominated in Euro, regardless of the residency of the issuer and holders. The CSDB currently contains information on over 7 million non-matured or “alive” debt securities, equities, and mutual fund shares/units plus approximately forty-nine million matured or “non-alive” (e.g. matured, early redeemed, or canceled) securities. Developed by the ECB, the CSDB is jointly operated by the members of the ESCB and it is only accessible by them, i.e. it is not available for public purposes. The CSDB uses data from commercial data providers, national central banks, and other sources. The most plausible value for each attribute and instrument is selected through a weight-based algorithm and gaps of missing information (in particular for prices and income) are filled with reliable

estimates. The system makes use of expertise within the ESCB to enhance data quality.

The CSDB provides consistent results and harmonization of concepts and calculations for all users together with efficiency in the data reporting, which reduces the burden from reporting agents, and improves the data compilation process.

2.1 iDQM Tool

Data quality management in the CSDB is based on the DQM framework⁴ that was established in 2012 and provided the guidelines for assessing data quality. Monitoring was performed once a month by using data snapshots. These snapshots focused on different dimensions such as changes in individual attributes, problems in issuer identification, and comparison with benchmark statistics, where certain target thresholds had to be fulfilled. Monitoring was performed via a Business Intelligence tool which allowed for data exploration but did not offer the possibility to amend or confirm potential issues. The only possibility to address them was via other tools and these changes were not reflected in the data snapshots.

To tackle these issues, a new tool was developed which allowed an interactive approach to data quality management: the iDQM. This tool visualises exceptions in the data that might require manual verification and potential intervention. Furthermore, the iDQM tool links the transactional part of the database (where all the data processing and calculations take place) with the data warehouse (where all the snapshots are taken from) and offers the possibility for data quality managers to amend errors and update the target metrics instantaneously. Moreover, the iDQM offers better prioritization in the exception generation, including an audit tool for tracing actions and confirmations.

2.2 Bulk Tool

The complementary bulk tool enables Data Quality Managers to change one or more instrument attributes directly. This tool is used for interventions on a larger scale and can be used to correct multiple values at the same time. These corrections do not necessarily correspond to detected exceptions.

2.3 Audit Records

Any data change performed either through the iDQM or the Bulk tool creates an audit record. Audit records contain information on how the data was before and after the change. By extracting a historical log of audit records we obtained a dataset of over 1 million changes for 25 different exception types. Please note, that the majority of the records were based on corrections made via the bulk tool.

3. Our Solution

Our objective is to identify data records in CSDB which might require a

⁴ <https://eur-lex.europa.eu/eli/guideline/2012/689/oj>

correction by a Data Quality Manager. Rather than explicitly modelling domain knowledge as static rules for identifying such records, we want to leverage the experts' knowledge implicitly contained in past data corrections. This means, we want use historic interventions of Data Quality Managers to generalise the patterns underlying the corrections of data records. These patterns can then be used to predict whether new data records might require a manual quality check, to prioritise the detected issues and to ideally propose which value requires an adjustment.

These objectives can be translated into three distinct tasks:

- (i) Identify if a record might need to be checked and potentially changed by a Data Quality Manager.
- (ii) Rank these suggestions to increase efficiency of the Data Quality Manager in reviewing data records by considering both: the economic relevance of an instrument and the probability that an observation is an outlier.
- (iii) Propose to the Data Quality Manager which is the field that is most likely to be incorrect.

3.1 Feature Engineering

The initial and very important step before modeling a machine learning prediction problem is to convert raw data into meaningful features. The features must capture information about the data that might be of relevance for the learning process. Only if the features actually capture the signals indicating data quality issues, a machine learning approach will be capable to learn from the data and achieve a good generalisation and predictive performance. This is why the steps of data preparation and feature engineering are crucial in the life cycle of any machine learning project [2,19]. In our particular case we created the following features:

- **Percentage change for numerical columns:** The relative change compared to the month before the prediction. Additional to this feature, we created the relative change to the median of the last three months, which is sometimes more reliable due to the statistical robustness of the median.
- **Lag features:** Boolean features denoting if the value of a column has been changed with respect to the previous month.
- **Time difference for date columns:** Date columns can lead to overfitting since the model might learn an exact date rather than a pattern. Instead of using dates or timestamps, they are replaced by the time difference between two events.
- **Categorical columns:** A high ratio of the problem features are categorical features with a high cardinality. To utilize these features in our model we used Target Encoders [19, 21, 15] with regularization to avoid overfitting or data leakage [17, 22].

3.2 Evaluation Setup

To evaluate and assess how well a model generalises and predict how it will perform in a production environment it is necessary to split the available data in subsets used for training, validating and testing the model. Given the setup of the CSDB system and the intended use of the classification model we decided to split the data along the temporal dimension. This means we used

the historically oldest data for training the model, and more recent data for validation and testing. A temporal split corresponds to the envisaged use case of applying the model to novel observations, arriving in the system after the model has been trained and deployed. This also addresses the specific challenge of distribution variability across different types of exceptions over the time axis. Our split is made following a ratio of 60:20:20 for train, validation and test data.

Furthermore, we created an additional gold-standard subset of our test data. This gold-standard subset was constrained to those exceptions that were corrected via the iDQM tool. The motivation for this additional evaluation dataset was to see how well the models trained on the overall corpus would perform in the context of the iDQM tool, where the data quality managers typically perform their work.

For the sub-tasks mentioned above, we used different evaluation metrics:

- (i) The learning task to detect which data records require a manual intervention corresponds to a binary classification problem. Hence, we use classical metrics based on a confusion matrix. By denoting true positive results as TP , false positive as FP and false negative results as FN we make use of the precision and recall metrics.

- **Precision** is a classification metric that measures the quality of the prediction. It is defined as:

$$precision = \frac{TP}{TP + FP} \quad (1)$$

In our case, precision tells us for which percentage of the cases the machine learning model was right when it predicted that a Data Quality Manager would change a record.

- **Recall** is a classification metric that measures the percentage of positive instances that were identified by the model, defined as:

$$recall = \frac{TP}{TP + FN} \quad (2)$$

In our case, recall tells us for which percentage of all the cases where a Data Quality Manager has actually changed a value, was the machine learning model able to predict this need of a change.

- (ii) The task of ordering the identified data records in such a way that a data quality manager encounters mainly items in the top positions which require an intervention is a ranking task.

As evaluation metric we employ the normalized discounted cumulative gain (NDCG). NDCG is, as its name suggests, a normalized version of the discounted cumulative gain (DCG). The traditional formula of DCG accumulated at a particular rank position p is defined as:

$$DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)} \quad (3)$$

$$NDCG_p = \frac{DCG_p}{IDCG_p} \quad (4)$$

where $IDCG_p$ is the ideal discounted cumulative gain, i.e. the cumulative gain value for an optimal ordering. Typically, NDCG is computed for

certain cut-off values K in the ranking, assuming a human user will not look at all items, but only at the top- K positions.

3.3 Predicting outliers

In the context of this project we use supervised classification methods to predict the probability that a data point will need to be amended by a Data Quality Manager. This corresponds to detecting outliers and can be modelled effectively as a binary classification task.

The methods we tried out for this problem were logistic regression, decision trees, support vector machines and gradient boosting (through CatBoost). For all methods we employed the standard Python implementations available in the scikit-learn package [23].

We found that not all of the 25 different exception types we observed in our data set were frequent enough to train and evaluate the machine learning methods. The other exceptions occurred too infrequently to provide sufficient data for inferring a generalised classification model. Furthermore, very early in the process it turned out that the CatBoost algorithm significantly outperformed the other approaches.

In Table 1 we present the performance of CatBoost on the full test dataset in terms of precision and recall for detecting different types of exceptions. What can be seen is that the performance varies across different types of exceptions. This is an indication that the patterns underlying those exceptions might be quite different and accordingly more or less challenging to detect.

| Exception Type | Precision | Recall |
|--------------------------|-----------|--------|
| <i>AmountOutstanding</i> | 0.409 | 0.762 |
| <i>CouponDate</i> | 0.063 | 0.305 |
| <i>SecurityStatus</i> | 0.456 | 0.801 |
| <i>MaturityDate</i> | 0.409 | 0.505 |
| <i>IssueDate</i> | 0.963 | 0.667 |
| <i>DividendAmount</i> | 0.369 | 0.551 |
| <i>ESAI2010</i> | 0.735 | 0.996 |

Table 1: Model performance of CatBoost across different types of anomaly detection on the overall test dataset.

To assess how suitable the models are for deploying them in the iDQM tool we additionally checked their performance on the gold-standard data set. The first observation we made was that only seven of the exception types appeared in the gold-standard data set. This might be explained by Data Quality Managers addressing only certain types of exceptions in the iDQM tool. For other types of exceptions, which might indicate some systematic errors, they use the bulk tool for correcting larger amounts of data records.

Furthermore, as shown in Table 2 the performance of the models differs quite a lot from the full data set. For some exceptions type the precision and recall for the binary classification task drop to 0, which means that no data record was marked as being an outlier. This deviation in the performance indicates that the two data sets behave quite differently and that the

observations from the iDQM tool follow a different distribution compared to the values from the more frequently used bulk tool. However, there still is some signal from the training data which can be leveraged to identify data records which might require a manual check and provide a relevance based ranking.

| Exception Type | Precision | Recall |
|--------------------------|-----------|--------|
| <i>AmountOutstanding</i> | 0.502 | 0.230 |
| <i>CouponDate</i> | 0.000 | 0.000 |
| <i>SecurityStatus</i> | 0.455 | 0.746 |
| <i>MaturityDate</i> | 0.000 | 0.000 |
| <i>IssueDate</i> | 0.000 | 0.000 |
| <i>DividendAmount</i> | 0.500 | 0.621 |
| <i>ESAI2010</i> | 0.249 | 0.365 |

Table 2: Model performance of CatBoost on the iDQM specific gold-standard dataset

3.4 Ranking Problem

Ultimately, the objective is to produce a ranking in order to prioritize the records that Data Quality Managers would need to investigate. Ideally, the order in this ranking would maximize the number of true positives in higher positions and reduce the amount of work for the Data Quality Managers.

The ranking for a certain record (R_p) is based on the product of the amount outstanding or market capitalization (AO_p) for a certain financial instrument multiplied by the soft prediction of the machine learning model that the observation is an outlier ($f(x_p)$) (cf. Equation 5). In this way we combine the insights from the machine learning model with the business priority to ensure correctness of data entries which are of higher economic relevance.

$$R_p = f(x_p) * AO_p \quad (5)$$

Table 3 shows the NDCG values for ranking the identified exceptions following equation (5). Ranking order is crucial for user experience in the context of the iDQM as it instills trust in the ranking algorithm and ensures that the limited resources dedicated to data quality will be focused on the most pertinent data issues. Applying the ranking brings several of the identified cases close to the top of the results, causing the data quality managers to be primarily exposed to those cases which actually require some manual intervention.

| Exception Type | K=10 | K=50 | K=100 | K=1000 |
|--------------------------|-------|-------|-------|--------|
| <i>AmountOutstanding</i> | 0.773 | 0.781 | 0.782 | 0.770 |
| <i>CouponDate</i> | 0.969 | 0.994 | 0.994 | 0.994 |
| <i>SecurityStatus</i> | 0.930 | 0.913 | 0.911 | 0.963 |
| <i>MaturityDate</i> | 0.500 | 0.500 | 0.543 | 0.843 |

| | | | | |
|-----------------------|-------|-------|-------|-------|
| <i>IssueDate</i> | 0.395 | 0.664 | 0.664 | 0.664 |
| <i>DividendAmount</i> | 0.000 | 0.000 | 0.000 | 0.000 |
| <i>ESAI2010</i> | 0.000 | 0.000 | 0.000 | 0.000 |

Table 3: NDCG ranking results on gold dataset for different exception types and ranking cut off values K .

4. Explanations Taxonomy

In this section we first overview the different types of users that interact with the CSDB and expand on their profiles. Afterwards, we analyse the different explainable AI desiderata that arise through ML implementation in a statistical production system. For more details about the desiderata of explainable AI in statistical production systems we suggest [18].

4.1 Users

During the project, we came across several generic user roles which help to classify the needs for solutions of explainable and responsible AI. A key question driving this classification is *Who needs an explanation of an AI method?* This helps to clearly define and distinguish different desiderata for explanations.

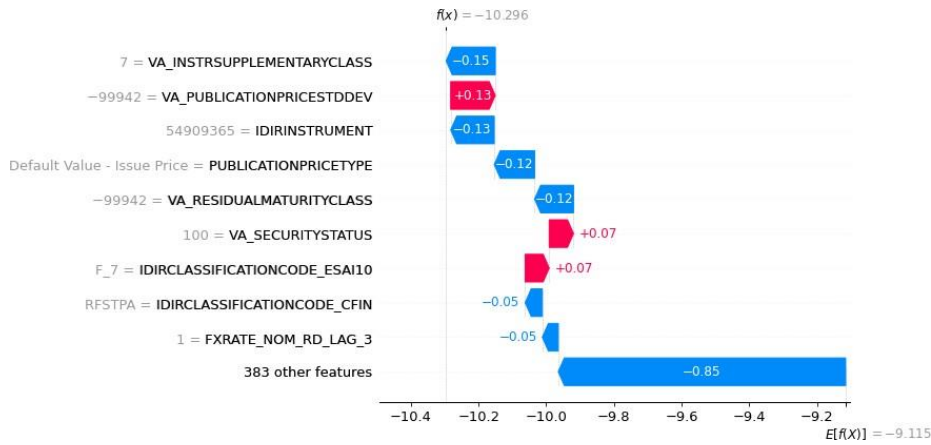
The following profiles define users that can potentially interact with a machine learning system all of whom have quite different needs for explainability:

- (i) **Data scientists and AI engineers:** This role corresponds to members of the team who build, model, and run a machine learning application. This type of user has technical expertise but does not necessarily have business expertise. They are in charge of the full life cycle of the machine learning application, from development to maintenance in production.
- (ii) **Business experts:** Users with this role provide the use case and domain expertise for a machine learning solution. They define the business activity or process which is supported by the AI solution. In our case, they are finance, economics, and statistics experts from the European System of Central Banks who act or intervene in business processes based on the recommendations of the models. This type of user might not have a technical background and is not a machine learning expert.
- (iii) **High stake decision makers:** This type of user determines whether to use and incorporate a machine learning model in the decision-making process. They typically have a management position, a high-level understanding of the business objectives, and a responsibility to deliver value. They need to understand and assess the potential risk and impact of incorporating the machine learning model into production.
- (iv) **End users:** Users which are affected by or make use of the final results belong to the group of end-users. The knowledge and potential expertise of this user group vary significantly. There might be cases where the group of end-users overlaps or is even identical to the group of *business experts*, e.g. when the machine learning solution is primarily serving internal business processes. Examples of end-users in our domain are the business areas or even the general public making use of data compiled by the Directorate General Statistics at the ECB.

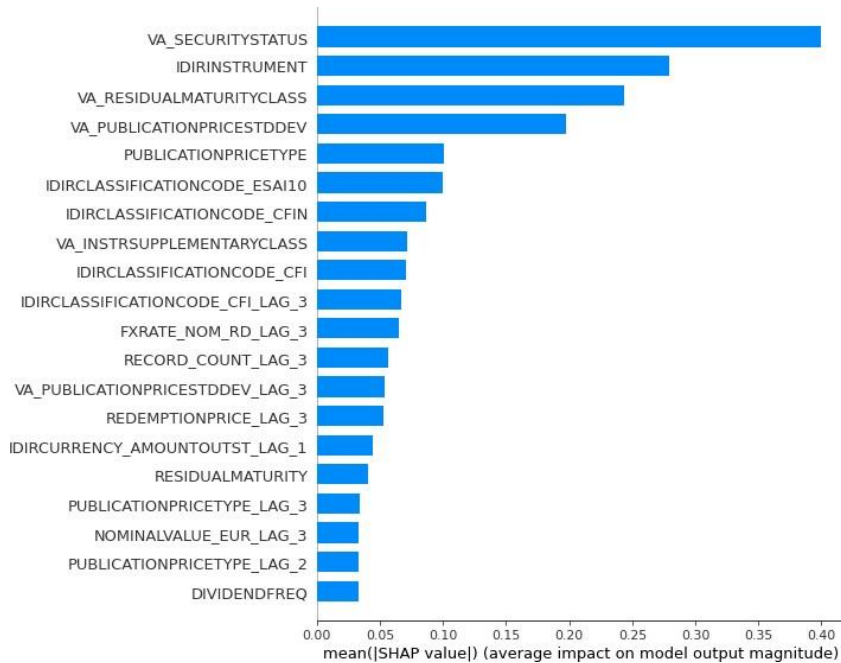
4.2 Building Trust

Trust is essential for the adoption of a machine learning application. Depending on the user, the meaning of trust and the way to obtain it differ [18]. Despite the careful testing and calibration of the machine learning process by the *data scientists*, *end users* of the data can identify potential data issues that have not been raised by the algorithm. These issues are communicated via the Data Quality Managers to the statistical production team (comprising of *business experts* and *data scientists*) responsible for the CSDB data quality.

The first step for the statistical production team is to verify that indeed the issues identified are not flagged by the algorithm. The next step is to investigate the reasoning behind this choice from the perspective of the algorithm by determining: (i) similar instances in the training dataset [3], (ii) which features contributed to the decision [10, 26] (cf. Figure 1) and (iii) logical decision rules to better understand the model logic [7, 27, 29].



(a) Local feature importance for given instance [12]



(b) Global Feature Relevance

Fig. 1: Shapley feature values by TreeExplainer [10, 11, 13]. Feature relevance explanations are useful to detect data leakage and prevent undesired model behaviour.

4.3 Actionable Insights

Understanding a machine learning algorithm is usually not an end in itself. The explanation offered through this understanding supports business processes and leads to actionable insights. Such insights enable the *business expert* to understand how to change a decision by manually intervening in the data. For instance, when data is identified to belong to a certain class, providing a set of actionable changes that would lead to a different decision can assist an expert in correcting or modifying the data.

Counterfactual generation aims to address this issue by proposing to the *business experts the minimal feasible change in the data in order to change the output of the algorithm*. Such a process enhances the understanding of the experts (and might further foster trust in the system). We formulate the problem following the counterfactual recourse formulation by Ustun et al. [33, 9] and the open source python package DICE [16] that quantifies the relative difficulty in changing a feature via feature weights.

We quantified the relative difficulty in changing a feature through the application of weights to the counterfactual explanations algorithm. For instance, in our case, recommendations should not ask the business expert (Data Quality Manager) to modify the country in which a financial instrument was issued or change the issue date to a time before the creation of the issuing company (cf. Table 4).

| Feature | Original Value | Modified Value | Initial Pred. | Modified Pred. |
|----------------------|----------------|----------------|---------------|----------------|
| SecurityStatus | 100 | 201 | 1 | 0 |
| ESAI2010 | F 31 | F 32 | 0 | 1 |
| SecurityStatus | 101 | 203 | 0 | 1 |
| PublicationPriceType | CLC | PAY | 1 | 1 |

Table 4: Set of counterfactual decisions generated [7, 16]. Counterfactual explanations can be understood as *What is the minimal feasible change in the data in order to change the output of the algorithm?*

4.4 Model Monitoring

Model monitoring aims to ensure that a machine learning application in a production environment displays consistent behavior over time. Monitoring is mainly performed by the *data scientist and AI engineer* and is crucial, as a drop in model performance will affect all the users. Two common challenges in model monitoring are (i) distribution shifts in the input data that can degrademodel performance (ii) changes in the machine learning algorithm due to a model retraining that can alter the individual explanations for decisions.

Detecting when the underlying distribution of the data changes is paramount for this use case, since failing to predict outliers or errors in the data will lead to a drop in the trust of the machine learning model. Also, the risk of having an incoherent explanation through time caused by the continual learning process is important as a discrepancy will lead to a decrease of trust by the *business experts*.

Diverse types of model monitoring scenarios require different supervision

techniques. We can distinguish two main groups: Supervised learning and unsupervised learning. Supervised learning is the appealing one from a monitoring perspective, where performance metrics can easily be tracked. Whilst attractive, these techniques are often unfeasible as they rely either on having ground truth labeled data available or maintaining a hold-out set, which leaves the challenge of how to monitor ML models to the realm of unsupervised learning [4, 14, 20, 6].

In this work since we are in the realm of unsupervised learning we have considered two possible solutions:

- Obtaining model uncertainty estimates via non-parametric bootstrap as an indicator of model performance when the deployed data is not available [20]. This monitoring technique allows to have an indicator of the model performance, detecting increases uncertainty lead to identifying indicators of when the model performance deteriorates in unsupervised data scenarios.
- Another approach suggested by Lundberg et al. [10] is to monitor the SHAP value contribution of input features over time together with decomposing the loss function across input features in order to identify possible bugs in the pipeline.

4.5 Fostering Explanations through Simple Models

Copying [31, 32] or distilling [8] machine learning models can greatly contribute to model explainability. Overly complex models tend to be difficult to explain [2] and can become unaccountable [28]. Model agnostic copies with a simple model might be able to achieve global explainability [30] which can be useful to build trust and gain knowledge by the *business expert*. Furthermore, in some deployment scenarios involving incompatible research and deployment versions [31], copying the ML model can ease the deployment task for the *data scientist*.

For our case, the original model $f_o(X)$ is a CatBoost classifier [24] which is a gradient boosting model [5] that often achieves state of the art results in many different types of problems and the copied models $f_c(X)$ are a scikit-learn [23] decision tree classifier and a Generalized Linear Model (GLM). The simpler models are of slightly inferior quality (cf. Table 5). However, having two simple models helps to improve the overall global explainability for the *data scientist* and to simplify deployment.

| | Catboost | Decision Tree | GLM |
|-----------|----------|---------------|-------|
| AUC | 0.816 | 0.771 | 0.741 |
| Precision | 0.805 | 0.741 | 0.685 |
| Recall | 0.833 | 0.824 | 0.733 |

Table 5: CatBoost is the original model $f_o(X)$ and the Decision Tree and the Generalized Linear Model the copied classifier $f_c(X)$. The creation of model surrogates or copies can distil the knowledge in the machine learning process thereby improving explainability.

5. Conclusions

Statistical production systems are one of the most promising fields for the adoption of machine learning processes within the context of central banking. Its maturity is still in the very early stages. In this paper, we have introduced our approach to improving the data quality exceptions presented to the Data Quality Managers of the Centralised Securities Database. This approach aims to first identify potential exceptions and then rank them based on the probability that they would need manual intervention by a Data Quality Manager.

Due to the importance of the quality assurance process and following the requirements of modern responsible machine learning, and explainable machine learning pipeline is needed to support the stakeholders involved in the data quality assurance task. With this reason in mind, a series of potential Machine Learning accountability desiderata have been presented tackling possible explainability needs that may arise.

The correct analysis, development, and introduction of machine learning in statistical production systems can lead to an optimization of the interventions needed to maintain data quality thereby making efficient usage of the limited resources available. This efficiency gain can translate into reduced time effort for Data Quality Managers or an increase in data quality. Eventually and over time such an ML pipeline can also provide new insights into the data as well as greater trust in the underlying processes.

5.1 Limitations

Using Machine Learning solutions for data quality assurance processes in granular statistical data collections is a very promising approach. An important insight from our project on the CSDB illustrates the potential of such approaches but also highlights some of the limitations.

Most importantly, the availability of sufficient and high-quality training data is crucial for successfully employing supervised machine learning methods. The volume, characteristics, and information contained in a data set with labeled outliers are essential for the quality of a trained model. We have seen in our case, that if the data used for training deviates from the data observed in a production setting this can negatively influence the performance of the trained model. This also underlines the importance of a sound evaluation methodology. The setup we chose allowed us to detect some limitations of the performance early on and before moving a full-fledged system to a production environment.

In conclusion, the main limiting factor currently is too short historicity of interventions of data quality managers for most of the exception types. While for some cases of exceptions the data permitted to train models of sufficient quality, many exceptions suffered from sparsity in the training models.

Furthermore, the errors corrected using the bulk tool deviate to a certain extent from the corrections made in iDQM. These similar but distinct feedback loops might serve as the basis for two different types of machine learning tasks, addressing different types of quality issues. Transferring the insights from one tool to the other seems to provide only limited insights.

Acknowledgements

This work was partially funded by the European Commission under contract numbers NoBIAS — H2020-MSCA-ITN-2019 project GA No. 860630.

References

1. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion* 58, 82–115 (2020)
2. Burkov, A.: *Machine Learning Engineering*. True Positive Incorporated (2020)
3. Caruana, R., Kangaroo, H., Dionisio, J.D., Sinha, U., Johnson, D.: Case-based explanation of non-case-based learning methods. In: *Proceedings of the AMIA Symposium*, p. 212. American Medical Informatics Association (1999)
4. Diethe, T., Borchert, T., Thereska, E., Balle, B., Lawrence, N.: Continual learning in practice. *stat* 1050, 18 (2019)
5. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29(5), 1189 – 1232 (2001)
6. Garg, S., Balakrishnan, S., Lipton, Z.C., Neyshabur, B., Sedghi, H.: Leveraging unlabeled data to predict out-of-distribution performance. In: *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications* (2021)
7. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local rule-based explanations of black box decision systems (2018)
8. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *CoRR* abs/1503.02531 (2015). URL <http://arxiv.org/abs/1503.02531>
9. Karimi, A.H., Schölkopf, B., Valera, I.: Algorithmic recourse: from counterfactual explanations to interventions. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362 (2021)
10. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610* (2019)
11. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence* 2(1), 56–67 (2020)
12. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777 (2017)
13. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model

predictions. In: Proceedings of the 31st international conference on neural information processing systems, pp. 4768–4777 (2017)

14. Malinin, A., Band, N., Gal, Y., Gales, M., Ganshin, A., Chesnokov, G., Noskov, A., Ploskonosov, A., Prokhorenkova, L., Provilkov, I., et al.: Shifts: A dataset of real distributional shift across multiple large-scale tasks. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
15. Micci-Barreca, D.: A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explor. Newsl.* 3(1), 27–32 (2001)
16. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, p. 607–617. Association for Computing Machinery, New York, NY, USA (2020)
17. Mougan, C., Alvarez, J.M., Patro, G.K., Ruggieri, S., Staab, S.: Fairness implications of encoding protected categorical attributes (2022)
18. Mougan, C., Kanellos, G., Gottron, T.: Desiderata for explainable AI in statistical production systems of the european central bank. *CoRR abs/2107.08045* (2021). URL <https://arxiv.org/abs/2107.08045>
19. Mougan, C., Masip, D., Nin, J., Pujol, O.: Quantile encoder: Tackling high cardinality categorical features in regression problems. In: V. Torra, Y. Narukawa (eds.) *Modeling Decisions for Artificial Intelligence*, pp. 168–180. Springer International Publishing, Cham (2021)
20. Mougan, C., Nielsen, D.S.: Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap (2022)
21. Pargent, F., Bischl, B., Thomas, J.: A benchmark experiment on how to encode categorical features in predictive modeling. Master's thesis, School of Statistics (2019)
22. Pargent, F., Pfisterer, F., Thomas, J., Bischl, B.: Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features (2021)
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830 (2011)
24. Prokhorenkova, L.O., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Cat-boost: unbiased boosting with categorical features. In: S. Bengio, H.M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (eds.) *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6639–6649 (2018). URL <https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html>
25. Pérez, A.C., Huerga, J.: The centralised securities database (csdb) - standardised microdata for financial stability purposes. In: B.f.l. Settlements (ed.) *Combining micro and macro data for financial stability analysis*, vol. 41. Bank for International Settlements (2016). URL

<https://EconPapers.repec.org/RePEc:bis:bisifc:41-15>

26. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144 (2016)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: S.A. McIlraith, K.Q. Weinberger (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pp. 1527–1535. AAAI Press (2018). URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982>
28. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5), 206–215 (2019)
29. State, L.: Logic programming for XAI: A technical perspective. In: J. Arias, F.A. D'Asaro, A. Dyoub, G. Gupta, M. Hecher, E. LeBlanc, R. Pen˜aloza, E. Salazar, Saptawijaya, F. Weikˆamper, J. Zangari (eds.) Proceedings of the International Conference on Logic Programming 2021 Workshops co-located with the 37th International Conference on Logic Programming (ICLP 2021), Porto, Portugal (virtual), September 20th-21st, 2021, CEUR Workshop Proceedings, vol. 2970. CEUR-WS.org (2021). URL <http://ceur-ws.org/Vol-2970/meepaper1.pdf>
30. Unceta, I., Nin, J., Pujol, O.: Towards global explanations for credit risk scoring. *CoRR* abs/1811.07698 (2018). URL <http://arxiv.org/abs/1811.07698>
31. Unceta, I., Nin, J., Pujol, O.: Copying machine learning classifiers. *IEEE Access* 8, 160268–160284 (2020)
32. Unceta, I., Palacios, D., Nin, J., Pujol, O.: Sampling unknown decision functions to build classifier copies. In: V. Torra, Y. Narukawa, J. Nin, N. Agell (eds.) *Modeling Decisions for Artificial Intelligence*, pp. 192–204. Springer International Publishing, Cham (2020)
33. Ustun, B., Spangher, A., Liu, Y.: Actionable recourse in linear classification. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19 (2019)



EUROPEAN CENTRAL BANK

EUROSYSTEM

Supervised machine learning in interactive feedback loops for statistical production systems



15/02/2022

**Carlos Mougán (University of Southampton),
Georgios Kanellos (ECB), Johannes Micheler (ECB),
Jose Martínez (Solenix), Thomas Gottron (ECB)**

Overview

- 1 Introduction to the CSDB
- 2 Problem statement and dataset preparation
- 3 Evaluating the results
- 4 Understanding explainability needs
- 5 Conclusions

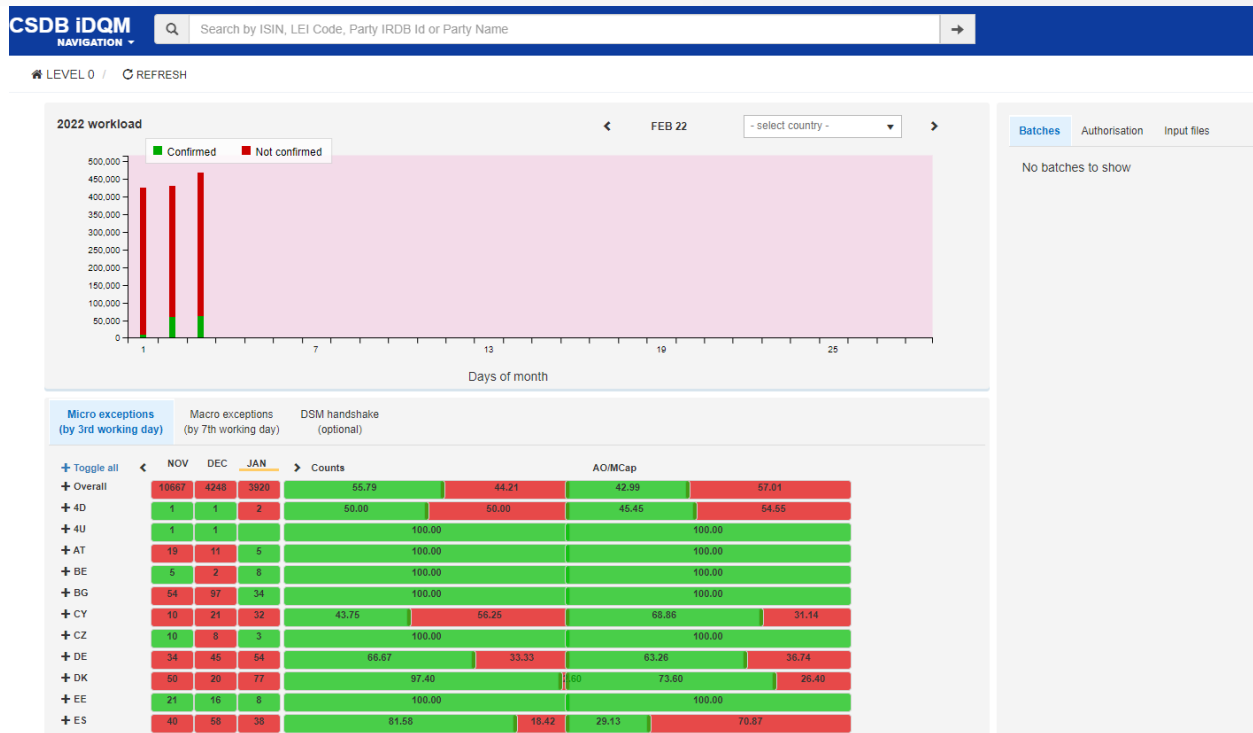
Centralised Securities Database

Complete, accurate, consistent and up to date information on all individual securities relevant for the statistical purposes of the ESCB

Includes information from multiple data sources:

- ☐ Commercial Data Providers + Rating agencies
- ☐ 26 National Central Banks (NCBs) and
- ☐ ECB internal sources

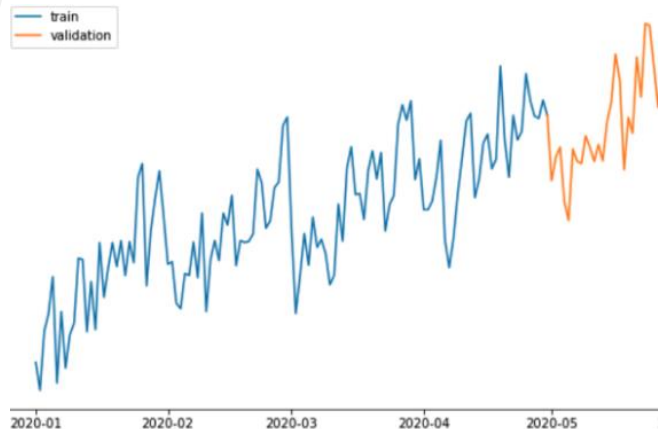
A unified tool for data exceptions



Problem statement

1. **Identify records** that might need to be checked and potentially changed
2. **Rank** these suggestions to minimize false-positives
3. **Suggest fields** that might need to be corrected to the DQ manager

Data preparation



Temporal model validation

- Percentage change for numerical columns
- Time difference for date columns
- Encode categorical columns

Feature engineering

Modeling

Our selected model $f(x)$ is a CatBoost classifier.

Amount outstanding (AO)

R – Final ranking of exceptions

$$R_p = f(x_p) \cdot AO_p$$

Ranking of exceptions

| Exception Type | K=10 | K=50 | K=100 | K=1000 |
|--------------------------|-------|-------|-------|--------|
| <i>AmountOutstanding</i> | 0.773 | 0.781 | 0.782 | 0.770 |
| <i>CouponDate</i> | 0.969 | 0.994 | 0.994 | 0.994 |
| <i>SecurityStatus</i> | 0.930 | 0.913 | 0.911 | 0.963 |
| <i>MaturityDate</i> | 0.500 | 0.500 | 0.543 | 0.843 |
| <i>IssueDate</i> | 0.395 | 0.664 | 0.664 | 0.664 |
| <i>DividendAmount</i> | 0.000 | 0.000 | 0.000 | 0.000 |
| <i>ESAI2010</i> | 0.000 | 0.000 | 0.000 | 0.000 |



EUROPEAN CENTRAL BANK

EUROSYSTEM

Explainability Needs

-
- Users
 - Classification





xAI USER PROFILING

Who requires an
explanation?

- ▶ Data scientists
 - + Technical expertise
 - Little business expertise
- ▶ Business experts
 - + Business expertise
 - Little technical expertise
- ▶ High Stake Decision Makers
 - Assess potential risk and impact
- ▶ Users
 - High variation in their profiles

Classification of explainability needs

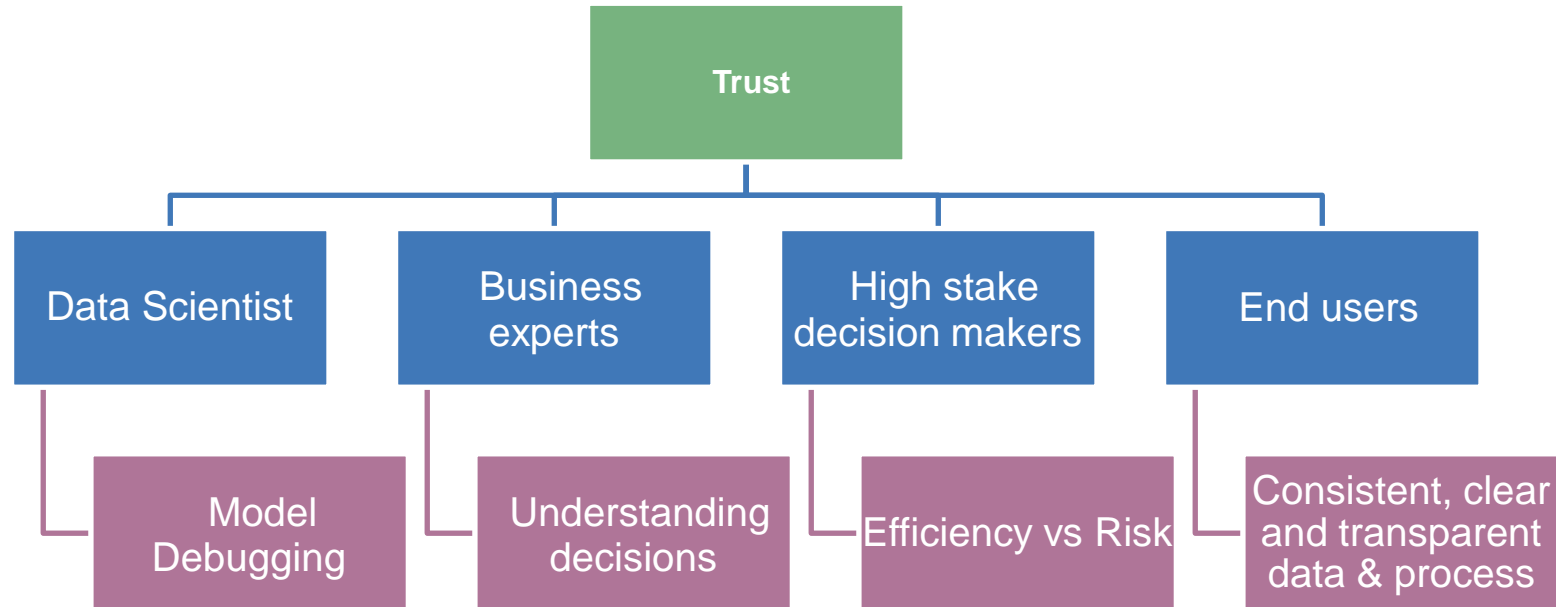
When do we
need explainability?



1. Building **trust**
2. Model **monitoring**
3. Actionable **insights**

1. Building trust

Different users, require different explanations with different tools

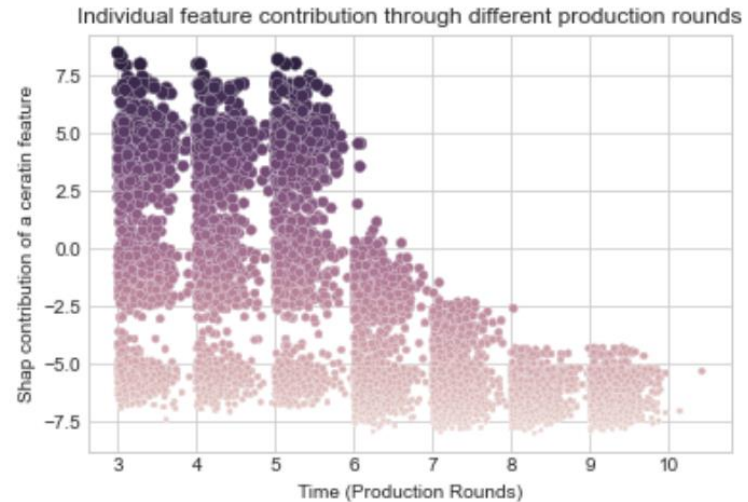


2. Model monitoring

*"Data is not static, it evolves
and so does the model"*

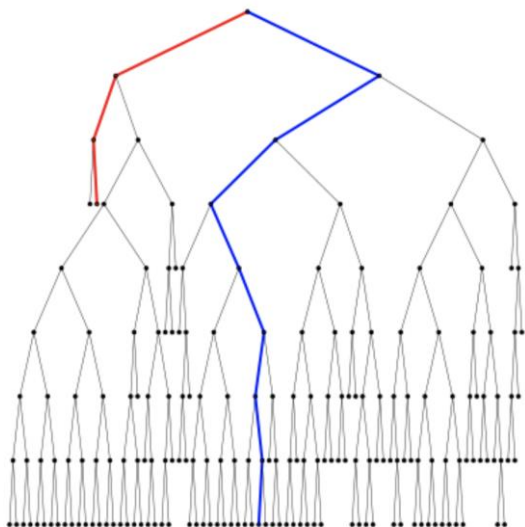
Distribution Shift: Train and Test data are statistically different.

Distributions change over time. Monitoring through explanations.



3. Actionable insights

What can we change so the data is no longer an outlier?



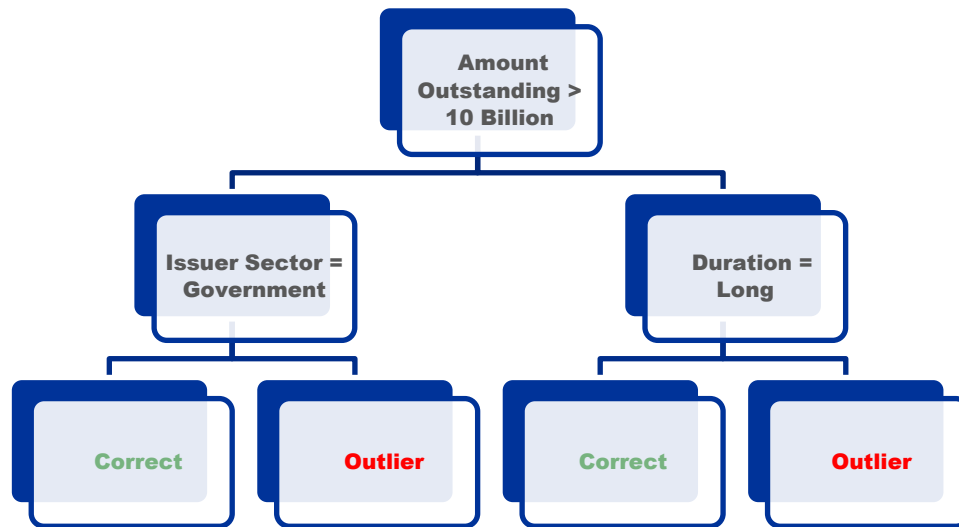
Explanation for Business experts

Counterfactuals that are:

- Feasible actions
- Plausible changes

Counterfactual Explanations

What does the user
need to change in order
to modify a decision?



Conclusions



EUROPEAN CENTRAL BANK

EUROSYSTEM

Towards machine learning in statistical productions systems

- ❖ Identify exceptions generated in the CSDB
- ❖ Rank those exceptions
- ❖ Classified xAI needs



Thanks for listening

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Stacking machine-learning models for anomaly detection: comparing AnaCredit to other banking datasets¹

Davide Nicola Continanza, Andrea del Monaco, Marco di Lucido, Daniele Figoli,
Pasquale Maddaloni, Filippo Quarta and Giuseppe Turturiello,
Bank of Italy

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Stacking machine-learning models for anomaly detection:
comparing AnaCredit to other banking datasets

by Pasquale Maddaloni, Davide Nicola Continanza, Andrea del Monaco,
Daniele Figoli, Marco di Lucido, Filippo Quarta and Giuseppe Turturiello

April 2022

Number

689



BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Stacking machine-learning models for anomaly detection:
comparing AnaCredit to other banking datasets

by Pasquale Maddaloni, Davide Nicola Continanza, Andrea del Monaco,
Daniele Figoli, Marco di Lucido, Filippo Quarta and Giuseppe Turturiello

Number 689 – April 2022

The series Occasional Papers presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The Occasional Papers appear alongside the Working Papers series which are specifically aimed at providing original contributions to economic research.

The Occasional Papers include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.

The series is available online at www.bancaditalia.it .

ISSN 1972-6627 (print)

ISSN 1972-6643 (online)

Printed by the Printing and Publishing Division of the Bank of Italy

STACKING MACHINE-LEARNING MODELS FOR ANOMALY DETECTION: COMPARING ANACREDIT TO OTHER BANKING DATASETS

by Pasquale Maddaloni*, Davide Nicola Continanza*, Andrea del Monaco*, Daniele Figoli*,
Marco di Lucido*, Filippo Quarta**, Giuseppe Turturiello**

Abstract

This paper addresses the issue of assessing the quality of granular datasets reported by banks via machine learning models. In particular, it investigates how supervised and unsupervised learning algorithms can exploit patterns that can be recognized in other data sources dealing with similar phenomena (although these phenomena are available at a different level of aggregation), in order to detect potential outliers to be submitted to banks for their own checks. The above machine learning algorithms are finally *stacked* in a semi-supervised fashion in order to enhance their individual outlier detection ability.

The described methodology is applied to compare the granular AnaCredit dataset, firstly with the Balance Sheet Items statistics (BSI), and secondly with the harmonised supervisory statistics of the Financial Reporting (FinRep), which are compiled for the Eurosystem and the Single Supervisory Mechanism, respectively. In both cases, we show that the performance of the stacking technique, in terms of F1-score, is higher than in each algorithm alone.

JEL Classification: C18, C81, G21.

Keywords: banking data, data quality management, outlier and anomaly detection, machine learning, auto-encoder, robust regression, pseudo labelling.

DOI: 10.32057/0.QEF.2022.0689

Contents

| | | |
|-----|---|----|
| 1 | Introduction | 5 |
| 2 | Data..... | 6 |
| 2.1 | AnaCredit vs. BSI..... | 7 |
| 2.2 | AnaCredit vs. FinRep | 8 |
| 3 | Anomaly detection strategies: definitions and estimation procedures | 9 |
| 3.1 | Robust Regressions | 10 |
| 3.2 | Autoencoders..... | 15 |
| 3.3 | Stacking predictions in a semi-supervised learning setting..... | 19 |
| 4 | Results and discussion..... | 23 |
| 4.1 | BSI vs. AnaCredit: empirical evaluation..... | 23 |
| 4.2 | FinRep vs. AnaCredit: empirical evaluation | 25 |
| 5 | Summary and conclusions..... | 28 |
| | References | 30 |
| | Appendix A - Tables and Charts | 34 |
| | Appendix B - Robust regression equation..... | 39 |

* Bank of Italy, Statistical Data Collection and Processing Directorate.

** Bank of Italy, IT Development Directorate.

1 Introduction¹

Big-data analytics is increasingly being adopted within the community of central banks, for several purposes (Cagala, 2017; Chakraborty *et al.*, 2017). An important area regards the application of machine learning techniques in order to improve the quality of data collected on the basis of regulatory reporting. Over the last few years, such surveys have become more granular and complex, in order to allow a better understanding of economic developments and, more in general, to improve the assessment of the actual and potential impact of policies on the economy². As regards banking data, a key role is played by credit disbursement to the economy that, in Italy, represents more than two thirds of banks' total assets.

The main sources of credit data currently used at the Bank of Italy are the Eurosystem's collection of Balance Sheet Items (BSI), the EU harmonized Financial Reporting (FinRep), the Italian Central Credit Register data (CCR) and, for a couple of years now, the Eurosystem's granular collection AnaCredit.

This paper investigates the possibility of building statistically founded cross-checking between the highly granular AnaCredit survey and the aggregated BSI and FinRep statistics by exploiting the similarities shared by the three surveys with respect to the phenomena that are covered. Originally, the three surveys were designed for different purposes and so the actual data collections follow different reporting rules and definitions with regard to the types of loans that are collected, the reporting population, the data model and the transformation rules. More importantly for our purposes, BSI and FinRep are very well established and mature data collections, whereas AnaCredit is quite a recent one, so it might not have achieved the same high quality standards of the other two yet. This is why in this paper, in defining a new set of quality checks, we try to exploit the information available in BSI and FinRep to improve the quality of AnaCredit data through outlier detection techniques.

To set up a new set of data quality checks, the expertise of the analysts needs to be complemented with the use of advanced statistical tools that allow us to handle the complexity of a highly granular survey such as AnaCredit. In this respect, the basic idea of the paper is to resort to machine learning techniques to carry out systematic cross-checking between series on the same phenomena although pertaining to different data collections in order to identify potential outliers to be submitted to reporting banks for their own checks³.

From a methodological point of view, we rely on machine learning methods (Bishop, 2011; Hastie *et al.*, 2001 and 2013) in order to overcome some of the limits recognized in the statistical literature on outlier detection as regards the identification of the boundary separating 'normal' observations from outliers. These limits are related both to the possibility that 'normal behaviour' might not be static but, rather, evolve over time and also to the lack of labelled data for training models (Chandola *et al.*, 2009). Within this research field (Cusano *et*

¹ The authors are grateful to Gianluca Cubadda and Alessio Farcomeni (University of Tor Vergata, Rome), Francesca Monacelli and Roberto Sabbatini (Bank of Italy) for their useful comments and fruitful discussions on a preliminary draft of the paper. The views expressed herein are those of the authors and do not necessarily reflect those of the Bank of Italy.

² For a recent discussion see Cœuré, 2017.

³ Namely, records that are considered anomalous because they are significantly different from the other points of the dataset (Aggarwal, 2017).

al., 2021; Zambuto *et al.*, 2020; Farnè *et al.*, 2018; Goldstein *et al.*, 2016), the novelty of this paper lies in the development of a general approach that makes a pairwise comparison between datasets containing information on similar phenomena.

We show that the proposed methodology, based on an ensemble learning technique, detects anomalies with a higher level of precision than the single methods used as baselines. Since anomalous observations are rare, the main metric considered to evaluate the performance of our developed models is the F1-score. With reference to this metric, the ensemble technique adopted also yields better results than the single baselines. In sum, we will show that the actual implementation of this methodology can contribute to improving the quality of AnaCredit data to the extent that the pairwise comparison with BSI and FinRep databases can lead to a more accurate list of potential outliers to be submitted to the cross-checking of reporting banks. It is worth remarking how the approach developed in this paper can be applied, more generally, to all those situations in which it is possible to exploit the information contained in aggregated datasets to detect potential outliers in a highly granular dataset.

The paper is organized as follows. Section 2 describes the three datasets under consideration and the deterministic pre-processing treatment carried out in order to make it possible to compare the aggregated series available for each of them. Section 3 explores the different strategies considered for detecting outliers and illustrates the developed ensemble machine learning techniques within a semi-supervised setting. Section 4 presents the results of the proposed approach. Section 5 summarizes the main conclusions, outlining the advantages of the proposed method and the possible directions for future research.

2 Data

Bank of Italy, in the context of the harmonized collections at the European level, collects aggregated credit information mainly within the scope of two ‘surveys’⁴: the monthly Balance Sheet Items (BSI), which is used for the common monetary policy analysis, and the quarterly Financial Reporting (FinRep) used for SSM supervisory purposes. Both surveys capture credit phenomena at aggregated level; different contractual forms of loans (i.e. overdrafts, mortgages, repurchase agreements) are added together by the amount paid out and then they are broken down by the relevant characteristics of the borrowers (i.e. sector and the residence) and by the main contractual features (currency, maturity, etc.). The global financial crisis of 2007-08 and the European debt crisis of 2009-10 showed that such aggregated data had been not sufficient to fulfill users’ need. This consideration led to the issue of Regulation (EU) 2016/867 on the collection of monthly granular credit and credit risk data (ECB/2016/13), the so-called AnaCredit Regulation, aimed at making available a new granular and multipurpose dataset containing loan-by-loan information on credit. Indeed, AnaCredit focuses

⁴ For the purpose of this paper, a ‘survey’ is a collection of homogenous data for a given purpose and disciplined by a reporting framework.

on the single credit instrument issued by credit institutions, within a contract stipulated vis-à-vis a given borrower (Di Noia et al., 2020). The main innovation brought by AnaCredit, as compared to BSI and FinRep data, rely on a larger number of details, at the level of single loan granted to counterparty provided that it is above the reporting threshold of 25,000 euros. This new unprecedented granular credit data collection allows the European System of Central Banks (ESCB) to carry out its tasks having a view of the entire distribution of this financial phenomenon. Furthermore, this data is more suitable to shed light on lending dynamics to legal entities and on the accumulation of risky debts in the banking sector.

In the following sub-sections, we describe the pre-processing steps carried out to build the two datasets used for our analysis. In particular, we use for our comparison only AnaCredit data starting from December 2018, although the first reporting date was September 2018. We decided to skip first reporting dates that, as it is often the case, present a very high degree of instability in terms of the quality of data, which is typically connected to the effective implementation and settlement of the compilation rules by reporting banks.

2.1 AnaCredit vs. BSI

The first comparison we carry out is between AnaCredit and BSI data collections from Italian banks. The latter refers to monthly aggregated stocks on assets and liabilities of Italian banks' balance sheets and it is used to compile the national contribution to Eurosystem's monetary statistics. BSI loans aggregates are based on data provided by reporting banks, which are then aggregated by amount according to some relevant loan information: the characteristics of the underlying contracts (type of instrument, duration and currency) and some classification variables of the contract counterparty (sector and residence). As anticipated, loans are particularly relevant being the core business of banks as well as the largest fraction of their assets.

For the purpose of this work, we take into account the main BSI time series of loans broken down by original maturity of credit instruments, the residence and the institutional sector of the borrower. In particular, we consider the following breakdowns: 1) Domestic Monetary Financial Institutions (MFIs), excluding Central banks; 2) Domestic Central Banks; 3) Other Euro area MFIs; 4) Domestic General Government; 5) Other Euro area General Government; 6) Euro area Other Financial Institutions and non-Money Market investment funds; 7) Euro area Insurance Corporations and Pension Funds; 8) Domestic Non-Financial Corporations (NFCs), original maturity up to 1 year; 9) Domestic NFCs, original maturity over 1 to 5 years; 10) Domestic NFCs, original maturity over 5 years; 11) Other Euro area NFCs, original maturity up to 1 year; 12) Other Euro area NFCs, original maturity over 1 to 5 years; 13) Other Euro area NFCs, original maturity over 5 years.

The ECB and the National Central Banks (NCBs) have already developed cross-checks between BSI and AnaCredit based on a deterministic approach: outliers are identified when the figures of interest exceed a pre-specified threshold that, for each BSI time series, is the same across all banks and reference dates. Typically, such thresholds are expressed in terms of percentage changes in the values detected in the two surveys. The current quality-control system would largely benefit of a statistical approach aimed at identifying acceptance thresholds that are bank-specific and can change over time.

The comparison between AnaCredit and BSI surveys requires the pre-processing of their differences in order to make data more comparable. To this end, we build a joint dataset and then we focus on the following differences. Firstly, we drop out from AnaCredit all those loans that are not recognized in the bank's individual balance sheet, since BSI includes only loans for which banks bear credit risk. Secondly, since AnaCredit contains accounting information only for end-of-quarter months, we impute the status of recognition of loans for the other two months of the quarter⁵. Thirdly, as new loans purchased on market are present in BSI at purchase price but in AnaCredit they are reported at nominal value, we discount the corresponding AnaCredit values by the difference between the nominal value and the price at the time of purchase. Fourthly, we derive in AnaCredit the classes of original maturity of loans present in BSI as the time between the settlement and the final legal maturity date of the contract expressed in years. Finally, reconciliation of data structures is performed in order to obtain comparable aggregates, by mapping the same subportfolio (e.g. interbank loans) of loans. Following the above preliminary adjustments, we can aggregate AnaCredit data by amount for the same bank and reference date and for the same characteristics of the BSI series, then obtaining the 'AnaCredit equivalent' specification of the 13 BSI series listed above. For both BSI and 'AnaCredit equivalent', the 13 considered series are further broken down by the sector of economic activity (NACE⁶) of the counterparty and the currency of the instrument⁷. The comparison between the two sets of indicators is carried out over the time span December 2018-March 2020 (monthly observations).

2.2 AnaCredit vs. FinRep

FinRep is the harmonized supervisory financial reporting that each credit institution must report on a quarterly basis according to the instructions of Regulation EU 680/2014 (Implementing Technical Standards - ITS) and International Financial Reporting Standards (IFRS). FinRep comprises accounting data on assets, liabilities, equity and statement of profit and loss. Within the assets of the balance sheet statements, reporting banks are also required to provide detailed information on loans, broken down by accounting portfolio, institutional sector and economic activity (NACE classification) of the counterparty, type of instrument, credit quality status and past due bands. It is relevant to underline that we exclude from the comparison all FinRep loans referred to households (institutional sectors S.14 and S.15 according to ESA 2010 classification) since not in all cases they are reported in AnaCredit. In order to derive the counterparty sector in FinRep, we resort to the detailed reporting rules defined by PUMA2 documentation⁸.

⁵ We assume that the last accounting evaluation is still valid for the non-end-of-quarter months and that all the new financial instruments are recognized. In general, it is quite rare that the recognition status of a financial instrument change from a month to another. Furthermore, the quota of new financial instrument is small and these new instruments are almost surely recognized.

⁶ See: Statistical Classification of Economic Activities in the European Community, Rev. 2 (2008) (NACE Rev. 2).

⁷ See Figures A1, A2 in Appendix A.

⁸ The main goal of PUMA2 process is to generate financial information for the production of several different statistical and supervisory reports. PUMA2 documentation provides detailed transformation rules to generate the final statistical reports from a granular input layers. For more details, see <https://www.cooperazionepuma.org/>.

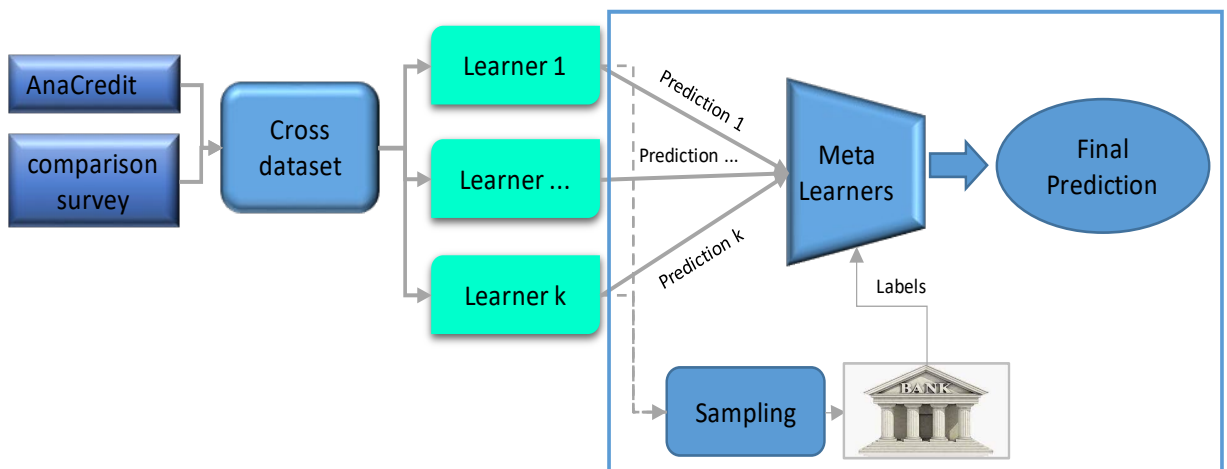
The three main measures of the accounting AnaCredit data, i.e. (1) net and gross carrying amount, (2) accumulated impairment amount and (3) accumulated changes in fair value due to credit risk, are compared to the equivalent measures of FinRep. We report some examples of disaggregated series elaborated for the comparison of the two dataset in Figure A3 of Appendix A.

As for the BSI comparison, a pre-processing of data is necessary to overcome a few differences between the two surveys. In particular, the main information that needs to be reconciled refers to: counterparty sector; type of instruments; the evaluation of past-due bands; the evaluation of gross carrying amount; the evaluation of accumulated negative changes on fair value due to credit risk on non-performing exposures. As in the case of BSI, the output is a joint dataset containing FinRep series and their equivalent (reconstructed) aggregation based on AnaCredit data. The comparison is carried out with reference to end-of-quarter dates, over the time span December 2018-March 2020.

3 Anomaly detection strategies: definitions and estimation procedures

Our cross-checking is based on two preliminary considerations. Firstly, BSI, FinRep and AnaCredit contain similar information on loans; therefore, we can assume that the patterns of the series referred to the same phenomena are similar. Secondly, the quality of BSI and FinRep datasets is very high, as improvements have been introduced over many years. So we assume that the potential outliers identified on the basis of ‘divergences’ between the compared series – BSI vs. AnaCredit and FinRep vs. AnaCredit, respectively – can be attributed to anomalies in AnaCredit data. The statistical approach that is followed combines supervised and unsupervised methods for the identification of regular patterns. In particular, for the supervised approach we develop a robust regression model, whereas for the unsupervised approach we resort to two autoencoder models. The three base models above (‘learners’) are then combined via a ‘stacking algorithm’ consisting of an additional classifier (‘meta-classifier’) trained on the base models’ outputs (Figure 1).

Figure 1: Workflow*



* In our work only 3 learners are used as base models.

The meta-classifier allows us to synthesize the complementary insights derived from the different base models and to outperform each one of them in making the final prediction⁹. In particular, the meta-classifier is trained in a semi-supervised setting by using a dataset enriched with the binary labels ‘anomalous’ or ‘not-anomalous’ that, with reference to sample cases¹⁰, are attached to each observation on the basis of cross-checking with the intermediaries and pre-assessments based on the domain knowledge.

It is worth anticipating that the strength of the above approach lies in its versatility in terms of use cases to which it can be applied and ‘learners’ that can be considered. Actually, this method can be easily adapted to any comparison between data collections sharing similar information and each specific base models (‘learners’) could be swapped with others yielding better predictions and their number can also be changed.

It is worth noting that the prediction of the model, i.e. the list of potential outliers, refers to aggregates which are themselves a decomposition of BSI or FinRep aggregates. This detail allows the anomaly to be contextualized by elements that better explain it (such as the information that helps the reporting banks to identify the erroneous records in their own archives and the reasons behind the data verification request), allowing the intermediaries for a more effective and faster evaluation of the case.

3.1 Robust Regressions

As mentioned in previous paragraphs, loans to legal entities reported in AnaCredit, BSI and FinRep refer to the same information content, although at a different level of aggregation. The conceptual relationship between the phenomena, confirmed empirically by the high correlation between AnaCredit aggregates, on one side, and BSI and FinRep series, on the other¹¹, can be statistically exploited within a linear regression framework. In our model BSI (or FinRep) series represents the independent variable, given its ascertained high level of quality, whereas the equivalent AnaCredit aggregate is regarded as the dependent variable.

Despite the pre-processing steps described in Section 2 to build comparable credit statistics, there are inevitable and permanent structural differences between the two datasets under comparison (e.g. the reporting threshold effect and the exclusion of natural persons in AnaCredit). Therefore our linear regression introduces a specific explanatory variable to capture such structural differences. This variable is able to consider such differences as intrinsic and normal instead of as reporting mistakes.

We end up with the following equation, using a log transformation of the original variables¹²:

$$\log(A_{i,j,t}) = \beta_0 + \beta_1 \log(F_{i,j,t}) + \beta_2 \log(F_{i,j,t-1}/A_{i,j,t-1}) + \epsilon_{i,j,t}, \quad (1)$$

where I denotes a bank, j a sub-portfolio of loans and t is a reference date. The AnaCredit aggregate for a particular sub-portfolio of loans at time t ($A_{i,j,t}$) is compared with the correspondent amount of BSI ($F_{i,j,t}$)

⁹ See Lessmann *et al.*, 2015.

¹⁰ In future developments of this paper, data revisions could also be considered to further enrich the prior knowledge.

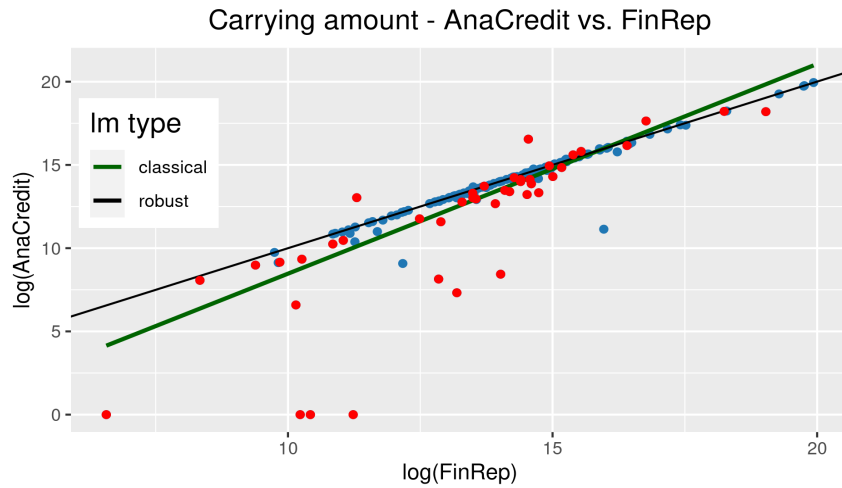
¹¹ See Figure 4 in Appendix A for the distributions of such index.

¹² For the derivation of the complete model see Appendix B.

(the same comparison holds for FinRep). The second explanatory variable is added to capture the definitional and structural differences between the datasets. The reporting mistakes remain isolated and they are contained only in the error component $\epsilon_{i,j,t}$. Unfortunately, we cannot identify the structural difference at time t , because of the presence of (potential) reporting errors in $A_{i,j,t}$. Instead, such differences are well identified at previous times, $t-1$, $t-2$, etc., on the basis of the findings of the data quality validation process in those periods. For the sake of simplicity, our model considers only the differences at time $t-1$. This way, the equation (1) could be read as an error correction model for the cross comparison of the two aggregates at time t (for more details, see Appendix B). We do not consider some seasonality form in equation (1), as we have short series available: only 5 dates for the FinRep/AnaCredit and 16 for BSI/AnaCredit comparison. Indeed in our stock data, we do not observe such element to a significant extent¹³ (see Figures A2 and A3 in Appendix A).

In an ‘ideal’ context, i.e. without anomalous data, the relationship in equation (1) would be correctly estimated. The presence of anomalous data in the dependent variable spoils practically this relationship. However, the literature on statistical robust estimation helps us to handle this issue (Hampel 1985; Hampel *et al.*, 1986; Farcomeni, 2015; Gschwandtner, 2012; Maechler, 2021), as shown in Figure 2, where the logarithm of AnaCredit carrying amounts (y-axis) and the logarithm of FinRep carrying amounts (x-axis) for the time series of ‘loans versus central governments, evaluated at amortized cost’ are plotted according to classical and robust linear predictions.

Figure 2: Classical vs. Robust regression



The robust regression (black line) is not affected by anomalous data like high leverage data points - such as red dots that are lying on the x-axis- as it happens in the classical linear regression (green line). The presence of high leverage data points has a significant impact on the parameters of the regression.

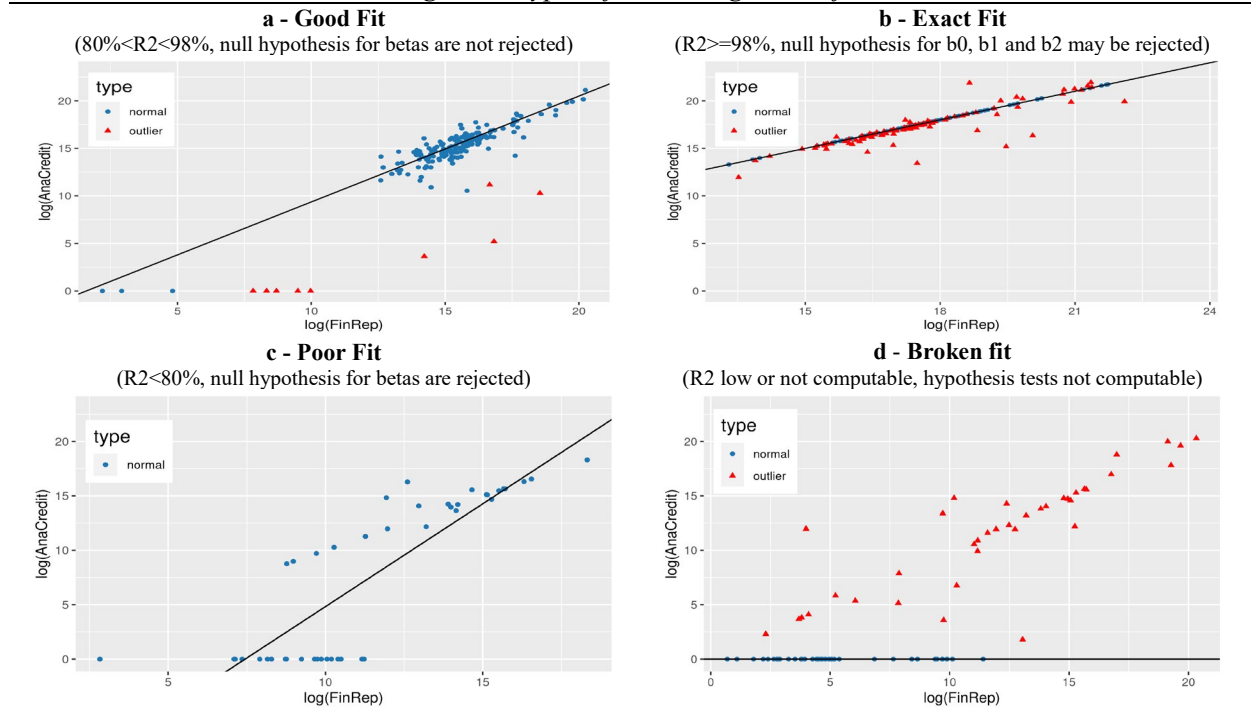
¹³ When more data is available, as future work we will be able to model also the seasonality of the series.

In particular, we consider the SMDM estimation proposed by Koller e Stahel (2011) that shows both high asymptotic efficiency and high breakdown point (BP; see Hampel *et al.* 1985). Furthermore, as derived in Appendix B we expect that the coefficients β_0 and β_1 should be equal, respectively, to zero and one, while the β_2 term should be less or equal to zero. Using the robust covariance matrix it is possible to test the null hypothesis that these conditions are met ($\beta_0 = 0, \beta_1 = 1$ and $\beta_2 \leq 0$). When these conditions are not met, we cannot consider the corresponding regression as reliable, and then we do not analyze the correspondent aggregates that are compared.

Four types of outcomes are obtained from our robust regressions (Figure 3):

- ‘Good fit’ (top left): the regression estimates are coherent with the prior knowledge on the betas and the R-squared is high;
- ‘Exact fit’ (top right): the SMDM algorithm tends to classify as ‘outlier’ those observations close to the regression line;
- ‘Poor fit’ (bottom left): the robust regression is affected by a high number of leverage points (BP is almost at 50%);
- ‘Broken fit’: there are more outliers than good observations (BP greater than 50%).

Figure 3: Types of robust regression fit



In our analysis, we face mainly the problem known as ‘exact fit’ (see, for example, Maronna *et al.*, 2006). Because of the high correlation between the AnaCredit aggregates and the corresponding BSI (and FinRep) series, our regressions often show a very high *R-squared* value (close to one), as the observed points are concentrated around a very small radius of the regression line. As a consequence, even in presence of a small

distance between an observed point and the regression line, that observation is marked as an outlier although it does not show an anomalous behavior. This is due to the fact that the distance is greater than the one recognized as ‘normal’ by the model in such situation. A second consequence has to do with the impossibility to perform reliable tests of hypotheses on the regression coefficients, as the robust estimates of their standard errors are close to zero, leading to the wrong rejection of the theoretically expected relationship (i.e. the model under the null hypothesis) when instead it should be regarded as valid.

To cope with the cases of exact fit, together with the cases of ‘poor fit’ and ‘broken fit’ when outliers are clustered (see Figures 4 and 5), we investigate three possible solutions:

- (1) resorting to the ‘Bonferroni correction’ for the *chi-squared* test of residuals (see Cerioli and Farcomeni, 2011);
- (2) adding Gaussian noise to the dependent variable (a procedure known as *jittering*);
- (3) de-noising through the removal of observations (a procedure known as *thinning*; see Cerioli and Perrotta, 2013).

Approach (1) relies on the assumption that the squared regression residuals follow approximately a *chi-squared* distribution with one degree of freedom. As discussed in Cerioli and Farcomeni (2011), in order to control for the probability of making one or more false rejections, a simple but effective method is to adopt the ‘Bonferroni correction’ for the confidence level over the sample size α/n . The observation is labelled as outlier if the statistics $T_{ijt} = \hat{e}_{ijt}^2 > \chi_{(1-\alpha/n, 1)}^2$.

Strategy (2) consists in adding a Gaussian random noise ε , $\varepsilon \sim N(0, \sigma_\varepsilon)$, to the dependent variable and, then, perform a robust regression on the new transformed variable. For the calibration of σ_ε , first we run a robust regression without adding noise and compute the Mean Absolute Deviation (MAD) of residuals, then we multiply it by a constant factor k , considering a floor positive value δ : $\sigma_\varepsilon = \min(\delta, k * MAD(\hat{e}_{ijt}))$. The empirical values of k and δ depend on the nature of the datasets under inspection: in this application, according to the abovementioned literature, we use $\delta \geq 0.1$ and $k \in [1.48; 5]$.

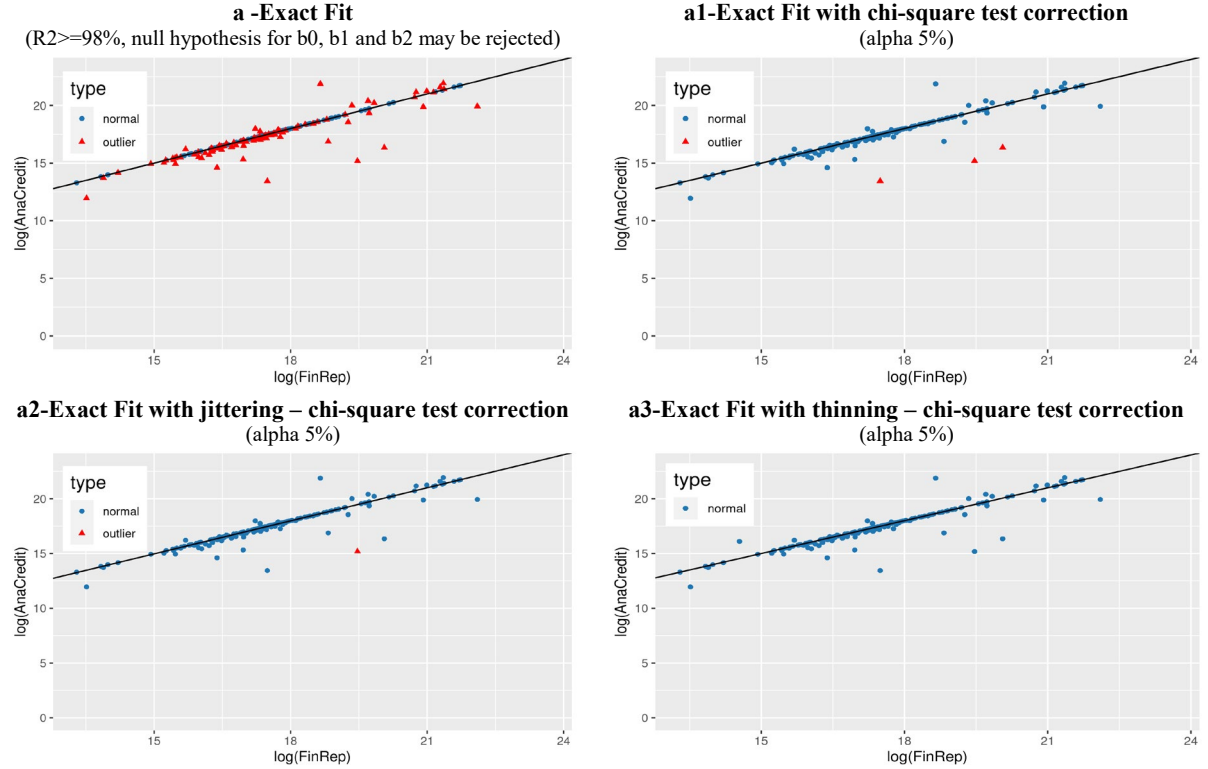
Finally, the procedure (3) consists in removing observations in order to down-weight the influence of high-density regions. To this end, it is necessary to define and evaluate a retain probability for each observation $\{y_i, x_i\}$, i.e. the logarithm of A_{ijt} and F_{ijt} . Following Cerioli and Perrotta (2013), we consider the retain probability $p(y_i, x_i) \propto 1 - \lambda_d(y_i, x_i)$, so that points will be deleted mainly in high-density regions; for the estimation of $\lambda_d(\cdot)$ we use the same isotropic Gaussian kernel¹⁴ introduced by the authors.

To detect outliers from procedure (2) and (3) we apply the *chi-square* test as defined in (1).

The ‘exact fit’ problem is well addressed by all the three solutions, as shown in Figure 4, where panel *a* illustrates the standard robust regression and panels *a1* to *a3* indicate the corresponding, robust regression when applying the three abovementioned strategies, respectively.

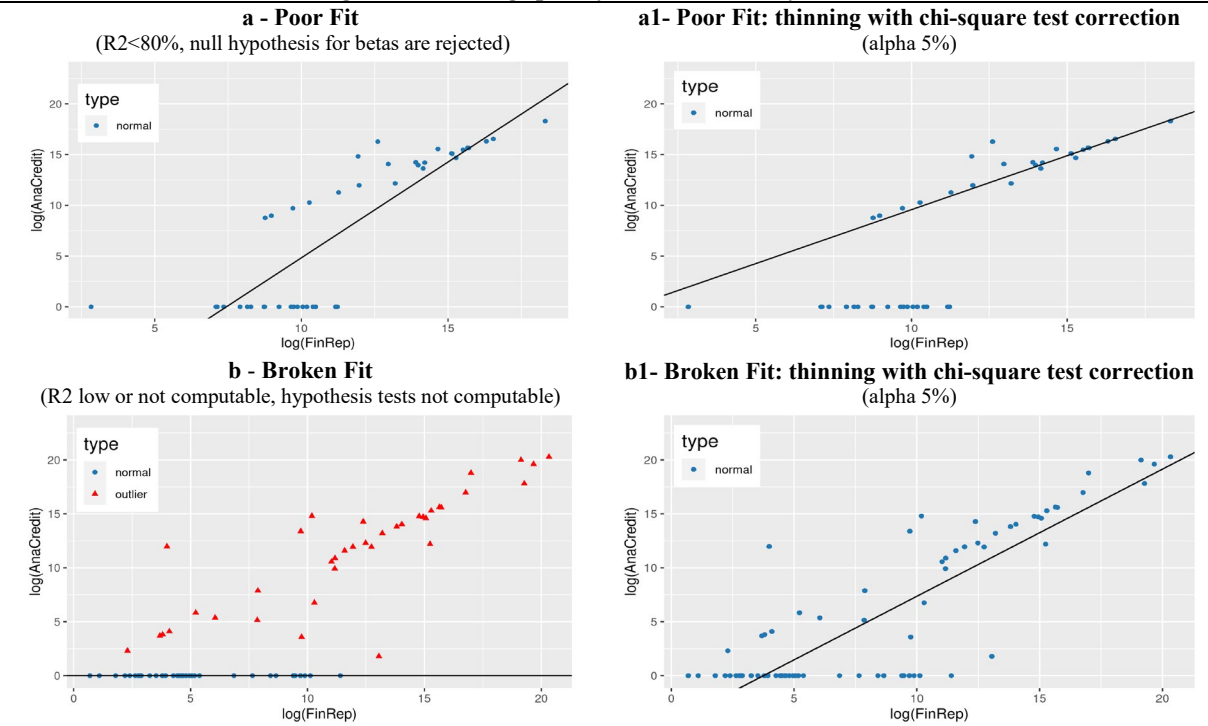
¹⁴ For its implementation, see function *density.ppp* of R package ‘spatstat’.

Figure 4: Solving 'exact fit' issue



The thinning procedure also allows to handle 'broken' and 'poor' fits (Figure 5): when extreme observations are clustered, the expected regression lines are correctly estimated both for 'broken fits' and for 'poor fits', allowing to correctly evaluate the hypothesis of presence of outliers.

Figure 5: Solving 'poor fit' and 'broken fit' cases

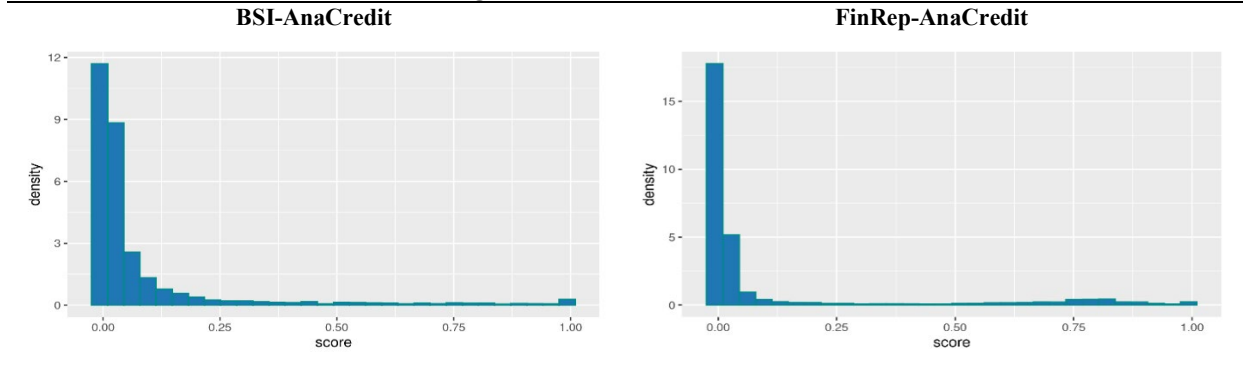


Since the series analyzed are different in terms of sample sizes, variance of residuals and level of contamination, and since each of the three methodologies has its own pros and cons, in order to take most out of them, we apply them according to the following hierarchical order:

- i) we run the jittering procedure, check that the null hypothesis for $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ is not rejected and the R-squared is above 80%. If such conditions are met, we classify the observation as an outlier or not on the basis of the chi-square test on residuals; if not, we move to (ii);
- ii) we run the standard robust regression, check that the null hypothesis for $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ is not rejected and the R-squared is above 80%. If such conditions are met, we classify the observation as an outlier or not on the basis of the chi-square test on residuals; if not, we move to (iii);
- iii) we run the thinning procedure, check that null hypothesis for $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ is not rejected and the R-squared is above 80%. If such conditions are met, then we classify the observation as outlier or not on the basis of the chi-square test on residuals; if not, we do not evaluate the observation.

Finally, in order to put the results of the robust regression in stacking with those of other models, we introduce a measure of anomaly for each data point inspected (score). Such scores are obtained by applying a min-max scaler to the absolute value of residuals of each ‘acceptable’ regression (i.e. that satisfies the conditions above). In Figure 6 the final scores of all the series for the BSI-AnaCredit and FinRep-AnaCredit comparison are reported.

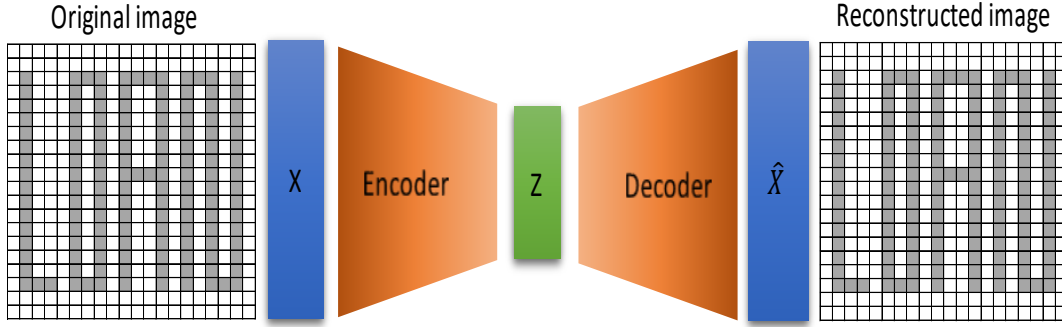
Figure 6: Scores distributions



3.2 Autoencoders

An autoencoder (AE) is a special type of multi-layer neural network performing hierarchical and nonlinear dimensionality reduction of data. The goal of an autoencoder is to replicate a given input as an output. Therefore, the output is the input itself, reconstructed. Typically, the model architecture is layered and symmetric, with the same number of nodes in the output and in the input layer, while nodes in middle layers are fewer in number. Therefore, the only way to reconstruct the input is by learning weights so that the intermediate outputs of the nodes in the middle layers consist in reduced representations. Figure 7 illustrates a fully connected autoencoder architecture.

Figure 7: Autoencoder architecture



Autoencoders are unsupervised models, as they do not need labels since the target variable is the input itself. Note that the outputs of the bottleneck layer represent the reduced representation, also known as ‘compressed representation’.

Because of its reduced representation of data, autoencoders represent a useful approach to detect outliers (Russo *et al.* 2019). The basic idea is that, in such dimensionality reduction, it is much harder to reproduce outliers than inliers (normal points), so the error of outliers’ reconstruction will be larger and, therefore, better identifiable.

Formally, autoencoders attempt to reconstruct an input image $x \in R^{k \times h \times w}$ through a bottleneck, effectively projecting the input (image) into a lower-dimensional space, called ‘latent space’. The projection (dimensionality reduction) occurs through an encoder function $E: R^{k \times h \times w} \rightarrow R^d$ and the reconstruction through an inverse decoder function $D: R^d \rightarrow R^{k \times h \times w}$, where d denotes the dimensionality of the latent space and k , h and w denote, respectively, the number of channels (equal to 3 in the case of Red, Green and Blue - RGB - images, to 1 for grayscale images), the height and the width of the input image. Choosing $d \ll k \times h \times w$ prevents the architecture from simply copying its input and forces the encoder to extract meaningful features from the input patches that facilitate accurate reconstruction by the decoder. The overall process can be summarized as

$$\hat{x} = D(E(x)) = D(z), \quad (2)$$

where z is the latent vector and \hat{x} the reconstruction of the input x . In our project, two models parameterize the functions E and D : the convolutional autoencoder (AE-CNN) and the dense autoencoder (AE-DNN). In the AE-CNN, ‘strided’ convolutions are used to downsample the input feature maps in the encoder and to upsample them in the decoder, while in the AE-DNN, dense layers are used for the same tasks.

We propose to measure the reconstruction accuracy with the Structural Similarity Index Metric (SSIM), as in Wang *et al.* (2004). This measure is designed to capture perceptual similarity; it captures inter-dependencies between local pixel regions that are disregarded by the current state-of-the-art unsupervised defect

segmentation methods based on autoencoders with per-pixel losses. The measure is not very sensitive to edge alignment and attaches importance to salient differences between input and its re-construction. The SSIM works in the spatial domain and, given two image patches $x = \{x_i | i = 1, \dots, P\}$ and $y = \{y_i | i = 1, \dots, P\}$, it is defined as

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2\mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (3)$$

where μ, σ are, respectively, the sample mean and sample standard deviation, σ_{xy} is the sample covariance between x and y and (c_1, c_2) are two positive stabilizing constants. The resultant SSIM index is a real value that could be normalized within the range $[0, 1]$, where 1 indicates perfect structural similarity that can be achieved only in case of two identical sets of data, while 0 denotes the absence of any degree of structural similarity. Following this approach, a loss function is derived as a mean of structural dissimilarity indexes (DSSIM), i.e. as the mean of the complement to 1 of the SSIM¹⁵ over all images

$$\mathcal{L}^{SSIM}(I) = \frac{1}{N} \sum_{i \in I} (1 - SSIM(i, \hat{i})), \quad (4)$$

where I is the set of all images. This loss function is used for AE-CCN models, while in the case of AE-DNNs the mean square error (MSE) is used.

The key idea of our model is to train an autoencoder on BSI (and FinRep) data in order to learn their ‘normal’ structure and to use it to identify abnormal structures (i.e. anomalous data) in AnaCredit. This method provides an *overall assessment* of all data reported by each bank for a given reference date. The identification of the reporting components that are anomalous occurs based on a score function, as described below.

The input data for the two neural networks are, respectively, the complete report of all series of a bank at a given reference date for the AE-DNN and their transformation in image for the AE-CNN. To create the image we use the collected scaled data to derive the auto cross-product: the result of this operation is a matrix whose elements are in the $[0, 1]$ range. This matrix can be regarded as an image in grey scale where each pixel is originated from each value of the matrix; each value of the matrix gives the grey level to color the pixel. The entries of the matrix have also a statistical meaning: each element gives the contribution of the interaction of the single component of the data collected with respect to the others. The final selected architecture is different for AE-DNN and AE-CNN. In the case of AE-DNN, the encoder and decoder networks consist of 2 fully-connected layers with respectively 150 and 28 hidden units, with LeakyReLU as activation function. The first layer also contains a dropout (Srivastava *et al.*, 2014) of 5%. The bottleneck layer is set as one fully connected layer with 20 hidden units, resulting in a 20-dimensional latent space.

In the case of AE-CNN, the encoder and decoder networks are comprised of convolutional layers with batch normalization and a max-pooling window of 2x2 after each convolution. The sigmoid as activation function is

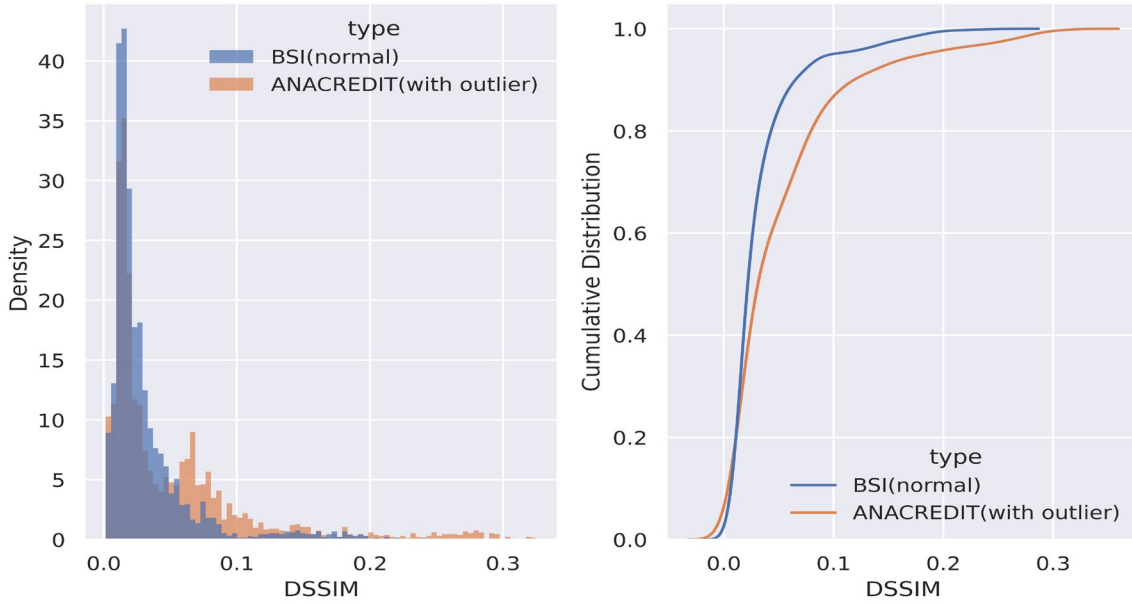
¹⁵ For more details, see Brunet *et al.* (2012).

used and the padding to reproduce the same dimensions. The encoder network is structured in a stack of three hidden layers with convolutional filters of respectively 16, 32, 16 units and kernel sizes of 3x3. The bottleneck consists of a convolutional layer with 4 convolutional filters of 23x23 size. Regarding the decoder network, its structure mirrors the encoding part having an up-sampling step in substitution of the max-pooling.

Training these two networks on BSI (and FinRep) dataset, which we assume are of a better quality, generates a distribution of the losses (MSE or DSSIM) associated with the entire dataset reported by each bank. This distribution is compared with the one derived from the application of the model to the AnaCredit dataset.

Figure 8 illustrates, as an example, the loss distributions (and relative cumulative distribution functions) obtained by training the AE-CNN model on BSI and then applying it to AnaCredit. Similar distributions occur for the comparison between AnaCredit and FinRep and in the cases of AE-DNN networks. For low values of the dissimilarity index the two distributions overlap, as expected, while for high values we observe the difference that need to be investigated.

Figure 8: Structural dissimilarity index



The distribution of DSSIM or normalized MSE¹⁶ helps to label as anomalous or not anomalous the whole set of AnaCredit data reported by each bank at a reference date, but it does not provide information on what components (i.e. which of the BSI and the FinRep series) have contributed the most to such result. To this end, we consider a score function mixing the loss function (DSSIM or normalized RMSE - l) value with the absolute relative difference (f) between pairs of BSI (and FinRep) and equivalent AnaCredit series:

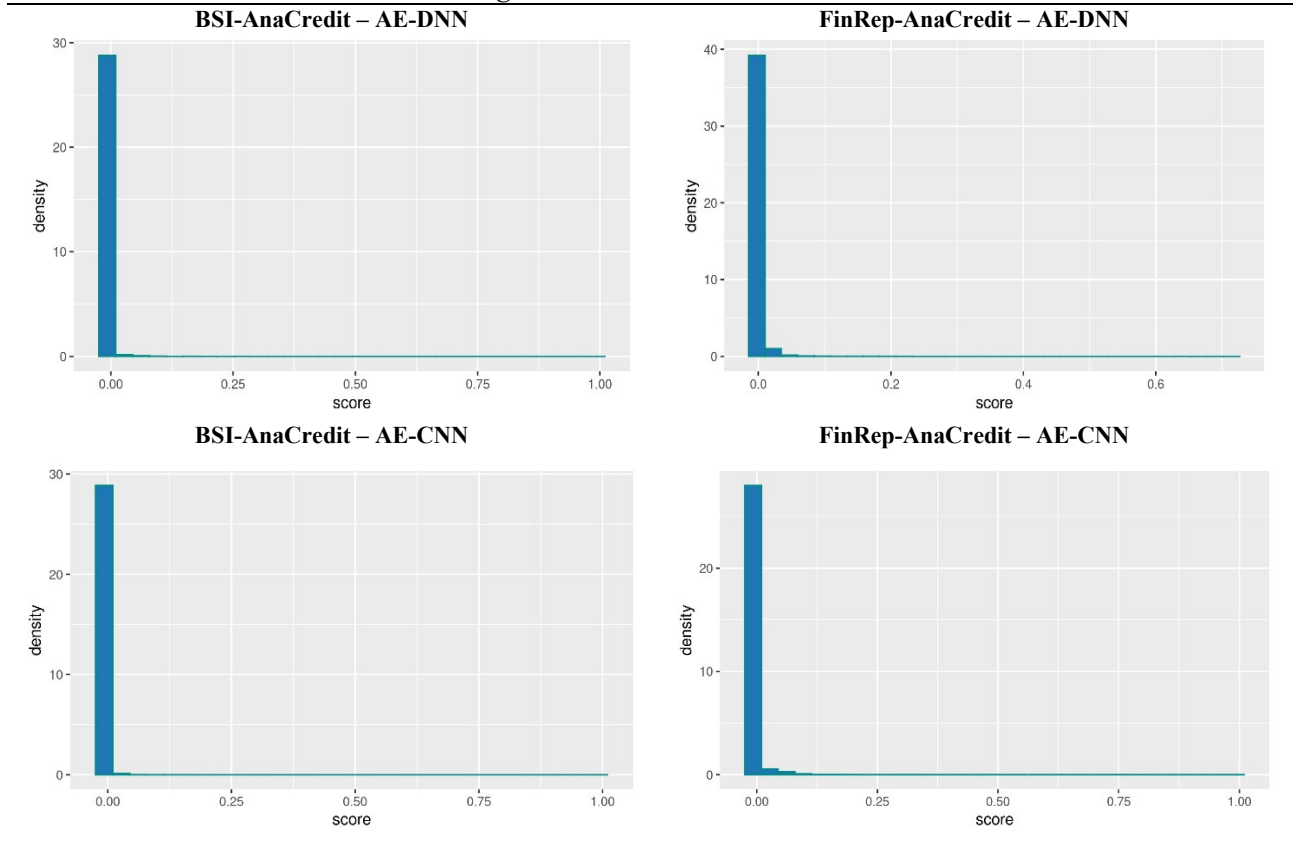
$$s(l, f) = l \cdot f^2 \quad (5)$$

¹⁶ We normalize the RMSE with the mean to obtain the coefficient of variation.

where l is the generic loss function for the model considered and f is the absolute relative difference between a generic BSI (and FinRep) aggregate and the equivalent AnaCredit, considered squared to emphasize big differences. Such a defined score function allows weighing the global evaluation (loss function) over all the pairwise comparisons. Besides, we rescale the score values in the interval $[0, 1]$ to have a normalized score, which can be compared to those produced by the other models developed.

Such normalized scores (see Figure 9) allow discriminating between good (low score values) and anomalous data (high score values).

Figure 9: Scores' distributions



In the following steps, the scores obtained above are used as input to the stacking models.

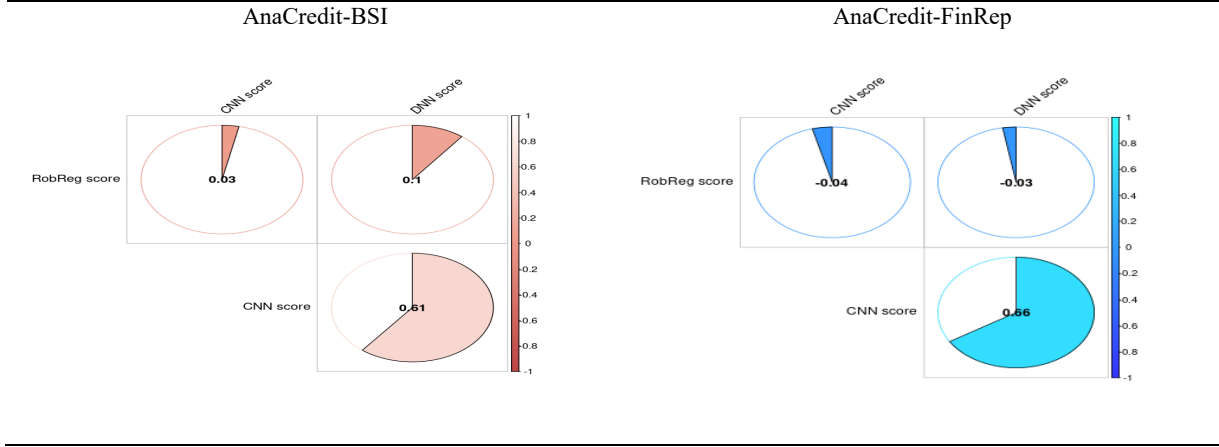
3.3 Stacking predictions in a semi-supervised learning setting

After training the basic models, for each observation (combination of reference date, reporting entity and compared aggregate) we have three anomaly scores, two produced by the autoencoders and one by the robust regression: these are the final predictions that are input to the meta-learners.

The robust regression scores exhibit low correlation with the autoencoders' scores, while the two autoencoders' scores do not have remarkable correlation among them. This is a common result in the two joint datasets (Figure 10). As the scores within the two comparisons map different information, they are combined together to produce a final forecast.

The combination takes place with the stacking technique (Wolpert, 1992; Dzeroski and Zenko, 2004), i.e. through a meta-learner using the abovementioned scores (predictions) as input for its forecast (Figure 1).

Figure 10: Correlation between learners' scores



Stacking models typically yields a better performance than each of the input learners. Stacking models require the existence of labels, i.e. a response variable that is attempted to replicate. To get a response variable, some cases have been sampled and verified with banks, while others have been pre-labelled on the basis of the domain knowledge of the analysts. Such information is bound to learner scores, obtaining a final partially labelled dataset on which we can train meta-learners in a semi-supervised learning paradigm (Chapelle *et al.*, 2006; González *et al.*, 2019).

With regard to the pre-labeling, based on domain knowledge, we mark as ‘correct’ (label 0) observations for which the difference between the amounts in BSI (FinRep) and the corresponding amounts in AnaCredit is deemed very small. In addition, we have marked as ‘anomalous’ (label 1) those for which the same distance is far negative, i.e. less than a chosen percentile of the empirical distribution.

The sampled cases are selected following a stratified sampling approach using the Neyman’s optimal criterion (Neyman, 1934). The stratification is obtained by considering the joint distribution of the scores appropriately discretized in binary values¹⁷, so that each observation is classified to a specific stratum. Neyman's optimal criterion is used with reference to the different variability of the average scores and to the sampling cost of each stratum, the latter being inversely proportional to the median distance of the difference between the amount of BSI (and FinRep) and the corresponding amount of AnaCredit. At this step, the sample size remains determined¹⁸ and the units are selected within each stratum with simple random sampling without replacement. Each sampled unit is analyzed with the bank that has reported it and this leads to the attribution of a response (R) that confirms or not the correctness of the AnaCredit data.

¹⁷ The scores are converted into binary coding, choosing the value of 0.2 as threshold.

¹⁸ For more details, see Table A1 in Appendix A.

Moving to the meta-learners, two different semi-supervised approaches are considered. Following the first semi-supervised approach, we develop and compare models in a context where the sample distribution of anomalous and not-anomalous cases within strata based on responses received have been previously reported to the universe of observations. In the second semi-supervised approach, the reporting of the sample responses to the universe and the estimation of the parameters of the models occurs in the training phase.

In the first setting, the completion of the labeling is performed following a Monte Carlo simulation approach. Pseudo labels (i.e. simulated responses) to the not-sampled and not-pre-labelled observations are assigned randomly by replicating the sample distribution of the responses received by banks within each stratum. By repeating this pseudo labeling over a sufficiently high number of times (we choose $N = 1000$), we obtain a dataset with this set of pseudo response variables, where each is composed by sampled/pre-evaluated labels (R) and simulated labels (Z).

On this enriched dataset, we train four different meta-learners: a logistic regression (*LOGIT*), a k -nearest neighbor (*KNN*), a random forest (*RF*) and a support vector machine (*SVM*). Each of these models has been optimized by cross-validation and the training is performed over the N simulated response variables. The final prediction is obtained as the central value of the N simulated predictions.

In this way, for a given reference date t and with respect to each bank i and sub-portfolio of loans j , we train four different models on the full vector of responses (R and Z) able to combine the predictions of the underlying scores exploiting different combination paths. Afterwards, we compare the model predictions and choose the best one with regard to some appropriate reference measures¹⁹. F1-score is our preferred reference measure to address the issue of the unbalanced classes (of 0 and 1) in the variables R and Y , but we provide the results also obtained for other measures, e.g. precision rate, recall, etc.

As regards the LOGIT, denoting by $p_{i,j,t} = \Pr(\hat{Y}_{i,j,t} = 1)$ the probability of an observation to be an outlier, where i, j, t represent, respectively, a bank, a sub-portfolio of loans and a given reference date, the model is described by the following equation:

$$l_{i,j,t} = \frac{p_{i,j,t}}{1-p_{i,j,t}} = \beta_0 + \beta_1 \text{ScoreRobReg}_{i,j,t} + \beta_2 \text{ScoreCNN}_{i,j,t} + \beta_3 \text{ScoreDNN}_{i,j,t} + \Theta\gamma + \varepsilon_{i,j,t}, \quad (6)$$

where Θ is a vector of control variables, i.e. dummy variables for the identification of re-aggregations of the analyzed aggregated series, and $\varepsilon_{i,j,t}$ is the noise term. In order to train the meta-learner, we use the weighted accuracy - with optimal weights²⁰ chosen for non-anomalous and anomalous data respectively - to cope with the imbalance in the responses. The optimal weights are obtained at the maximum value of the average F1-score calculated on the training set by employing the five-fold cross validation.

As regards the other three models, the generic function describing each of them is:

¹⁹ See Figure A5 in Appendix A.

²⁰ See more details in Figure A6 in Appendix A.

$$\hat{Y}_{i,j,t} = f(\text{ScoreRobReg}_{i,j,t}, \text{ScoreCNN}_{i,j,t}, \text{ScoreDNN}_{i,j,t}, \Theta) + \varepsilon_{i,j,t}, \quad (7)$$

where $f(\cdot)$ is a different function based on the model type and $\varepsilon_{i,j,t}$ is the noise term. Also for these models, the meta-parameters are calibrated on the training set by employing the five-fold cross validation, maximizing the average F1-score.

In the second setting, semi-supervised models for the equation 7 are trained by using only the sampled and pre-labelled variable (R) instead of \hat{Y} , as the labeling is carried out within the model estimation process. In this class of models, the following algorithms have been used: Self-training, SETRED, Tri-training, Co-Bagging and Democratic-Co.

The Self-training model (Yarowsky, 1995) is probably the earliest idea about using unlabelled data in classification. This wrapper-algorithm, starting from only the labelled data, iteratively uses a supervised learning method trained only on the part of the dataset labelled until the current iteration. At each step, it labels a part of the unlabelled points according to the current decision function until the whole dataset is labelled.

Similarly, the SETRED algorithm (Li and Zhou, 2005) first learns from labelled examples, and then iteratively chooses to label a few unlabelled cases on which the learner is most confident in prediction and adds them to its labelled set for further training at next step. However, at each iteration SETRED does not completely accept all the pre self-labelled examples, but it actively identifies the possibly mislabelled examples by testing a predefined null hypothesis with the local cut edge weight statistic associated with each self-labelled example. If the result of the test falls in a left rejection, the example is regarded as a good one; otherwise, it is a possible mislabelled example and it should not be included to the learner's training set.

The Tri-training algorithm (Zhou and Li, 2005) generates three classifiers from the original labelled set and labels an unlabelled observation if the other two classifiers agree on the labeling, under certain conditions. This procedure is repeated until convergence (generally with the complete labeling of the dataset).

The Co-Bagging method (Blum and Mitchell, 1998) assumes that the feature space can be split into two different conditionally independent views and that each view is able to predict the classes on its own. It trains one classifier in each specific view, and then the classifiers learn from each other the most confidently predicted examples from the unlabelled pool. The process continues until a predefined number of iterations is reached.

The Democratic-Co algorithm (Zhou and Goldman, 2004) uses multiple algorithms, instead of multiple views, to enable learners to label data from each other. This technique leverages off the fact that different learning algorithms have different inductive biases and that better predictions can be made by the majority vote.

In the next Section, we present the empirical results of the abovementioned models.

4 Results and discussion

The empirical results of the models introduced in the previous section are presented in the following two subsections, the first devoted to the comparison between AnaCredit and BSI, the second to that between AnaCredit and FinRep.

4.1 BSI vs. AnaCredit: empirical evaluation

The data dimension used for training and testing the models are reported in Table 1.

Table 1: Observations in training and test set

| | 1° semi-supervised setting | 2° semi-supervised setting |
|-----------------|----------------------------|----------------------------|
| <i>Training</i> | 5440 | 6199 |
| <i>Test</i> | 2331 | 1572 |
| <i>Total</i> | 7771 | 7771 |

The different size of the training and the test sets in the two approaches derives from the fact that in the first setting labels are available for all observations, being simulated via Monte Carlo. In the second setting, we need to consider only labelled observations in the test set; therefore, we assign all unlabelled observations to the training set and then split the pre-labelled observation between training and test set according to a 50% share.

Before evaluating the meta-learners, we measure the performance of the three learners (basic models) with standard performance metrics together with the balanced accuracy (in order to take into account the imbalance of the target variable Y in our dataset). Table 2 presents the performance metric results that can also be considered as benchmarks for the meta-learners results, presented in Section 3.

Table 2: Performance of the base models (learners)

| | Sample cases | | |
|--------------------|--------------|---------|---------|
| | RobReg | DNN | CNN |
| <i>Precision</i> | 0.09507 | 0.82353 | 1.00000 |
| <i>Recall</i> | 0.04500 | 0.02333 | 0.01333 |
| <i>Specificity</i> | 0.79965 | 0.81260 | 0.81122 |
| <i>Accuracy</i> | 0.73601 | 0.81266 | 0.81170 |
| <i>F1 score</i> | 0.06109 | 0.04538 | 0.02632 |
| Balanced accuracy | 0.42233 | 0.41797 | 0.41228 |

All the metrics considered in Table 2 are evaluated on a test set representing 30% of the available data. The Table clearly shows that the three basic models RobReg, DNN and CNN perform quite well in terms of accuracy, with values equal to 0.736, 0.813 and 0.811, respectively.. Unfortunately, since our input dataset is

biased towards non-anomalous cases (0.79 and 0.21) in Y variable, such models perform poorly when metrics taking into account the imbalance are considered: for instance, the balanced accuracy drops to 0.422, 0.418, and 0.412, respectively.

To strengthen the power of prediction of our models, a stacking step is further considered. In the first place, a baseline logistic model combining the predictions of the three basic models only on the sampled and pre-labelled data is developed. In particular, this baseline model is trained with a weighted accuracy in order to take into account the imbalance between non-anomalous and anomalous cases. Such model presents the following main results (Table 3): the F1 score is 0.997, the balanced accuracy 0.998, the precision 0.981 and the recall 0.988.

The performance over all the dataset is presented in Table 3 that shows the central values of the different meta-learners trained on the 1000 simulated pseudo labels.

Table 3: Performance of the models within the first semi-supervised setting

| | Baseline* | LOGIT1 | KNN | RF | SVM |
|-------------------|-------------|-------------|-------------|-------------|-------------|
| Precision | 0.981±0.019 | 0.901±0.016 | 0.848±0.014 | 0.864±0.018 | 0.830±0.021 |
| Recall | 0.988±0.011 | 0.433±0.161 | 0.459±0.016 | 0.463±0.014 | 0.514±0.017 |
| Specificity | 0.994±0.006 | 0.867±0.003 | 0.871±0.003 | 0.872±0.003 | 0.882±0.004 |
| Accuracy | 0.997±0.003 | 0.870±0.004 | 0.869±0.004 | 0.872±0.004 | 0.876±0.005 |
| F1 score | 0.997±0.002 | 0.585±0.017 | 0.596±0.015 | 0.603±0.015 | 0.635±0.017 |
| Balanced accuracy | 0.998±0.001 | 0.650±0.009 | 0.665±0.009 | 0.668±0.009 | 0.698±0.010 |

* Only on sampled cases.

All the metrics are evaluated on five test sets, each representing 30% of the data and obtained setting a different random seed: the central value of the different metrics is reported together with the deviation of this value from the minimum and maximum, in order to assess metrics' variability. The linear SVM yields somewhat better results in terms of F1 score and balanced accuracy when trying to replicate the baseline.

As regards the second setting of semi-supervised algorithms, the results highlighting their performances are shown in the following table:

Table 4: Performance of the models within the second semi-supervised setting

| | Self-training | SETRED | Tri-training | Co-Bagging | Democratic-Co |
|--------------------------|---------------|-------------|--------------|-------------|---------------|
| <i>Precision</i> | 0.995±0.005 | 0.993±0.007 | 0.993±0.007 | 0.989±0.004 | 0.993±0.007 |
| <i>Recall</i> | 0.995±0.005 | 0.988±0.012 | 0.987±0.011 | 0.995±0.005 | 0.989±0.008 |
| <i>Specificity</i> | 0.999±0.001 | 0.997±0.003 | 0.997±0.002 | 0.999±0.001 | 0.997±0.002 |
| <i>Accuracy</i> | 0.998±0.001 | 0.997±0.000 | 0.997±0.000 | 0.997±0.000 | 0.997±0.000 |
| <i>F1 score</i> | 0.995±0.003 | 0.991±0.003 | 0.991±0.002 | 0.992±0.002 | 0.992±0.002 |
| <i>Balanced accuracy</i> | 0.997±0.003 | 0.993±0.007 | 0.992±0.007 | 0.997±0.003 | 0.993±0.005 |

All the models use the support vector machine as learner (Democratic-Co use also a KNN and a C5.0) and all the metrics are evaluated on the test presented in Table 1. Additional four runs using different random seeds are used to generate new training and test sets, in order to assess metrics' variability. Based on the F1 score, the best performer is the self-training algorithm, showing a central value of 99.5%. If we compare it to the baseline model, we find a good replication in the F1-score and better results in terms of precision and recall. Pairwise comparison of the various models is carried out by using the McNemar's Test for binary classification. Table 5 contains the p-values for the null hypothesis that two models do not have significant differences in their label predictions. A small p-value denotes that there is a statistical significant difference in the power of prediction between the two models.

Table 5: Predictions comparison ()*

| | Self-training | SETRED | Tri-training | Co-Bagging | Democratic-Co |
|---------------|---------------|--------|--------------|------------|---------------|
| Self-training | | 0.009 | 0.138 | 0.003 | 0.013 |
| SETRED | | | 0.269 | 0.000 | 0.251 |
| Tri-training | | | | 0.002 | 0.025 |
| Co-Bagging | | | | | 0.000 |
| Democratic-Co | | | | | |

(*) P-value mean over the five test set. A p-value lower than 0.05 indicates a significant disagreement between the model predictions.

Table 5 shows that the self-training predictions are statistically equivalent to the tri-training ones. Since the self-training model gets the higher F1-score, and all the differences with SETRED, Co-Bagging and Democratic-Co are statistically different from zero, we can conclude that the self-training model is the best performer.

4.2 FinRep vs. AnaCredit: empirical evaluation

The data dimension used for the training and test set are reported in Table 6.

Table 6: Observations in training and test set

| | 1° semi-supervised setting | 2° semi-supervised setting |
|-----------------|----------------------------|----------------------------|
| <i>Training</i> | 26035 | 29680 |
| <i>Test</i> | 11201 | 7656 |
| <i>Total</i> | 37336 | 37336 |

As in the BSI comparison, the different size of the training and the test sets in the two approaches derive from the constraint of assigning, in the second semi-supervised setting, a share of 50% of pre-labelled observations to the test set and the rest of observations to the training set, while in the first setting it is not present.

In this case, we have a greater number compared to the BSI-AnaCredit case due to more disaggregated series considered in FinRep. As for BSI-AnaCredit, we first measure the performance of the three underlying models with standard performance metrics together with the balanced accuracy (imbalance of target variable Y). The results, reported in Table 7, are useful in terms of benchmark to the meta-learners presented in Section 3.

Table 7: Performance of the base models (learners)

| | RobReg | DNN | CNN |
|--------------------------|---------|---------|---------|
| <i>Precision</i> | 0.05330 | 0.83333 | 0.75000 |
| <i>Recall</i> | 0.05263 | 0.00283 | 0.00170 |
| <i>Specificity</i> | 0.82913 | 0.84726 | 0.84711 |
| <i>Accuracy</i> | 0.71184 | 0.84725 | 0.84708 |
| <i>F1 score</i> | 0.05296 | 0.00564 | 0.00339 |
| <i>Balanced accuracy</i> | 0.44088 | 0.42505 | 0.42441 |

All the metrics are evaluated on the test set representing 30% of the available data. The Table clearly shows that the three models RobReg, DNN and CNN perform quite well in terms of accuracy, with values equal to 0.712, 0.847 and 0.847, respectively. Therefore, their predictive capacity, considering the two classes of anomalous and non-anomalous data at the same time, is quite high. However, the dataset we are considering is unbalanced towards anomalous cases (only 16%). When we move to measures that take into account such imbalance, their performance decreases; for instance, the balanced accuracy falls to 0.441, 0.425, and 0.424, respectively.

Moving to the stacking step, the results of the performance measures for a weighted logistic baseline (only on sampled and pre-labelled data) and for LOGIT, RF, KNN and linear SVM models (over all the dataset) are reported in Table 8.

Table 8: Performance of the models within the first semi-supervised setting

| | Baseline* | LOGIT | KNN | RF | SVM |
|--------------------------|-------------|-------------|-------------|-------------|-------------|
| <i>Precision</i> | 0.998±0.002 | 0.833±0.008 | 0.501±0.005 | 0.747±0.010 | 0.503±0.006 |
| <i>Recall</i> | 0.999±0.001 | 0.367±0.009 | 0.204±0.004 | 0.434±0.008 | 0.380±0.013 |
| <i>Specificity</i> | 0.996±0.004 | 0.890±0.002 | 0.860±0.002 | 0.899±0.001 | 0.884±0.002 |
| <i>Accuracy</i> | 0.998±0.003 | 0.886±0.002 | 0.836±0.002 | 0.885±0.002 | 0.837±0.002 |
| <i>F1 score</i> | 0.997±0.003 | 0.508±0.009 | 0.290±0.005 | 0.548±0.009 | 0.432±0.010 |
| <i>Balanced accuracy</i> | 0.997±0.003 | 0.628±0.005 | 0.532±0.002 | 0.667±0.005 | 0.631±0.006 |

* Only on sampled cases.

All the metrics are evaluated on five test sets, each representing 30% of the data, obtained by setting a different random seed; the central value over the five test sets is reported for the different metrics and the deviation of

this value from the minimum and maximum value so as to assess metrics' variability. The various models attempt to replicate the baseline; although far from it, the model with the best results is the random forest, with an F1-score equal to 0.548 and a high precision of 0.747.

As regards the second approach of semi-supervised algorithms, the results highlighting their performances are shown in Table 9.

Table 9: Performance of the models within the second semi-supervised setting

| | Self-training | SETRED | Tri-training | Co-Bagging | Democratic-Co |
|--------------------------|---------------|-------------|--------------|-------------|---------------|
| <i>Precision</i> | 0.723±0.167 | 0.929±0.072 | 0.825±0.075 | 0.825±0.075 | 0.833±0.001 |
| <i>Recall</i> | 0.513±0.058 | 0.487±0.058 | 0.504±0.042 | 0.523±0.023 | 0.179±0.179 |
| <i>Specificity</i> | 0.605±0.079 | 0.656±0.106 | 0.653±0.097 | 0.653±0.097 | 0.500±0.107 |
| <i>Accuracy</i> | 0.697±0.054 | 0.733±0.054 | 0.715±0.036 | 0.715±0.036 | 0.518±0.125 |
| <i>F1 score</i> | 0.598±0.098 | 0.634±0.034 | 0.634±0.034 | 0.650±0.018 | 0.500±0.001 |
| <i>Balanced accuracy</i> | 0.563±0.065 | 0.582±0.072 | 0.595±0.053 | 0.595±0.053 | 0.335±0.139 |

All the models use C5.0 as learner (Democratic-Co use also a KNN and a SVM) and all the metrics are evaluated on the test presented in Table 6. Additional four runs using different random seeds are used to generate new training and test sets in order to assess metrics' variability. According to the F1-score, Co-Bagging is the best performers, followed by the SETRED and Tri-training algorithms.

Pairwise comparisons of various models is carried out by using the nonparametric McNemar's Test for binary classification. Table 10 contains the p-values for the null hypothesis that each pair of models does not show significant differences in their label predictions. A small p-value denotes a significant difference (improvement) in the prediction between the two models. Results reported in Table 10 show that the predictions of Co-Bagging is equivalent to the others. Therefore, there is not a clear winner from this competition and the Co-Bagging, SETRED and Tri-training algorithm seems to be equally efficient. Only the Democratic-co presents a rather poor performance.

Table 10: Predictions comparison ()*

| | Self-training | SETRED | Tri-training | Co-Bagging | Democratic-Co |
|---------------|---------------|--------|--------------|------------|---------------|
| Self-training | | 0.473 | 0.787 | 0.833 | 0.080 |
| SETRED | | | 0.488 | 0.573 | 0.184 |
| Tri-training | | | | 0.500 | 0.089 |
| Co-Bagging | | | | | 0.087 |
| Democratic-Co | | | | | |

(*) P-value mean over the five test set. A p-value lower than 0.05 indicates a significant disagreement between the model predictions.

5 Summary and conclusions

AnaCredit is a relatively recent ESCB dataset; it contains granular information (at contract and instrument level) on loans that banks grant to legal entities. Within the ESCB, two other ‘historical’, high quality datasets providing similar information on loans are the BSI and the FinRep, which are typically used in monetary policy and supervisory analyses respectively. Both of them can be exploited for a pairwise comparison with AnaCredit data to enhance the outlier detection process in the AnaCredit granular survey.

More specifically, we explore the use of machine learning techniques to carry out the above cross-checking with specific reference to loan portfolios, in a framework that takes the time dimension into account and is bank-specific. We resort to three models – a robust regression one and two autoencoder models – that grasp the existing relationships between each benchmark dataset – BSI and FinRep – and AnaCredit in order to identify potential outliers in the latter one.

Each model assigns an ‘anomaly score’ to each observation considered (uniquely identified by reporting date, entity, and loan aggregate). These anomaly scores are combined in order to yield a better forecast using a stacking approach under a semi-supervised learning context. Indeed, our approach lies in a semi-supervised environment having true anomalous or non-anomalous data labels only for a subset of the datasets. The true labels are based on both the domain knowledge of the analysts and the responses directly received from reporting entities on a number of observations, which are sampled by using a selection schema that is able to reproduce the distribution of scores assigned by the three models.

In this semi-supervised context, we consider two settings. In the first one, where the true labels have been reported to the universe of observations under a Monte Carlo simulation, we train logistic, random forest, KNN and SVM models and compare the results obtained from each of them. In the BSI-AnaCredit comparison, the SVM model has the highest F1-score, whereas in the FinRep-AnaCredit comparison the random forest is the better learner in terms of the same statistic. In the second setting, where the expansion of true labels takes place within the learning of the models themselves, we train Self-training, SETRED, Tri-training, Co-bagging and Democratic-co models and compare their results to the baseline. We find that for the BSI-AnaCredit comparison, the self-training gives the better F1-score, while for FinRep-AnaCredit comparison, the Co-Bagging model turns out to have the best performance. Considering all the models developed, we find that the algorithms of the second settings of semi-supervised models outperform those of the first settings.

Possible refinements of the paper, which are left to future developments, might consist in developing the current base learners: for the robust regressions we could move to a panel approach and for the autoencoders to the variational autoencoders. Further improvements are related to the optimization of the parameters underlying the second setting semi-supervised models and the use of other disaggregated BSI and FinRep series.

The framework developed in this paper is quite flexible and general and can be applied to carry out pairwise comparisons between datasets on similar phenomena but with different levels of granularity. As shown in our

empirical exercise, this approach exhibits important advantages not only in terms of a more accurate detection of potential outliers in a highly granular database, but also from the point of view of reporting banks that will have to cross-check such anomalies and decide whether to confirm or revise the data.

References

- Aggarwal C. (2017). “Outlier Analysis”, Springer.
- Bishop, C.M. (2011). “Pattern Recognition and Machine Learning”, Springer.
- Blum A. and Mitchell T. (1998). “Combining labeled and unlabeled data with co-training”, in Eleventh Annual Conference on Computational Learning Theory, COLT’ 98, pages 92–100, New York, NY, USA.
- Brunet D. and Vrscaj E. R. (2012). “On the Mathematical Properties of the Structural Similarity Index”, IEEE Transactions on Image Processing, vol. 21, no. 4.
- Cagala, T. (2017). “Improving Data Quality and Closing Data Gaps with Machine Learning”, IFC Bulletin, 46.
- Ceroli A. and Farcomeni A. (2011). “Error rates for multivariate outlier detection”, Computational Statistics and Data Analysis 55, pp. 544–553.
- Ceroli A. and Perrotta D. (2013). “Robust clustering around regression lines with high-density regions”, Springer, Advances in Data Analysis and Classification volume 8, pp. 5–26.
- Chakraborty C. and Joseph A. (2017). “Machine Learning at Central Banks”, Bank of England Staff Working Paper No. 674, <https://doi.org/10.2139/ssrn.3031796>
- Chapelle O., Scholkopf B. and Zien A. (2006). “Semi-supervised learning”, MIT Press
- Chandola V., Banerjee A. and Kumar V. (2009). “Anomaly detection: a survey”, ACM Computing Surveys, Vol. 41, No. 3, <http://doi.acm.org/10.1145/1541880.1541882>
- Cœuré B. (2017). “Setting standards for granular data”, Opening remarks by Benoît Cœuré, Member of the Executive Board of the ECB, at the Third OFR-ECB-Bank of England workshop on “Setting Global Standards for Granular Data: Sharing the Challenge”, Frankfurt am Main, 28 March 2017, <https://www.ecb.europa.eu/press/key/date/2017/html/sp170328.en.htm>
- Cusano F., Marinelli G. and Piermattei S. (2021). “Learning from revisions: a tool for detecting potential errors in banks’ balance sheet statistical reporting”, Bank of Italy, Working Papers, No. 611.
- Dzeroski, S. and Zenko, B. (2004). “Is combining classifiers with stacking better than selecting the best one?”, Machine Learning, 255–273.

- Di Noia M. and Moretti D. (2020). “Le informazioni statistiche della Banca d’Italia sul rischio di credito e la nuova rilevazione AnaCredit”, Banca d’Italia, Occasional Papers, No. 544.
- Farcomeni A. and Greco L. (2015). “Robust methods for data reduction”, CRC Press.
- Farnè M. and Vouldis A.T. (2018). “A methodology for automatised outlier detection in high-dimensional datasets: an application to euro area banks’ supervisory data”, ECB Working Paper N. 2171.
- Goldstein M. and Uchida S. (2016). “A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data”, Computer Science, Medicine, PLoS ONE.
- González M., Rosado O., Rodríguez J. D., Bergmeir C., Triguero I. and Benítez J. M. (2019). “ssc: An R Package for Semi-Supervised Classification”, R package version 2.1-0.
- Granger C.W.J. (1981). “Some Properties of Time Series Data and Their Use in Econometric Model Specification”, Journal of Econometrics, 28, 121-130.
- Gschwandtner M. and Filzmoser P. (2012). “Computing Robust Regression Estimators: Developments since Dutter 1977”, Austrian Journal of Statistics, Volume 41, Number 1, 45–58.
- Hampel, F. R. (1985). “The Breakdown Point of the Mean Combined With Some Rejection rules”, Technometrics, 27, 95-107.
- Hampel, F., Ronchetti E., Rousseeuw P. and Stahel W. (1986). “Robust Statistics: The Approach Based on Influence Functions”, N.Y.: Wiley
- Hastie T., Tibshirani R. and Friedman J. (2001). “The Elements of Statistical Learning”, Springer.
- Hastie T., James G., Tibshirani R. and Witten D. (2013). “An Introduction to Statistical Learning”, Springer.
- Koller, M. and Stahel W. A. (2011). “Sharpening wald-type inference in robust regression for small samples”, Computational Statistics & Data Analysis 55(8), 2504–2515
- Lessmann S., Baesens B., Seow H.V. and Thomas L.C. (2015). “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research”. *European Journal of Operational Research*, 247 (1), 124-136.
- Li M. and Zhou Z. (2005). “Setred: Self-training with editing. In Advances in Knowledge Discovery and Data Mining”, volume 3518 of Lecture Notes in Computer Science, pages 611–621. Springer Berlin Heidelberg.

- Maechler M., Rousseeuw P., Croux C., Todorov V., Ruckstuhl A., Salibian-Barrera M., Verbeke T, Koller M., Conceicao E.L. and Anna di Palma M. (2021). “robustbase: Basic Robust Statistics”, R package version 0.93-7, <http://robustbase.r-forge.r-project.org/>
- Maronna, R.A., Martin, D.R. and Yohai, V.J. (2006). “Robust Statistics: Theory and Methods”, Wiley, New York.
- Neyman, J. (1934). “On the two different aspects of the representative methods. The method stratified sampling and the method of purposive selection”, *Journal of Royal Statistical Society*, 97, 558-606.
- Russo S., Disch A., Blumensaat F. and Villez K. (2019). “Anomaly detection using deep autoencoders for in-situ wastewater systems monitoring data”. *Proceedings of the 10th IWA Symposium on Systems Analysis and Integrated Assessment (Watermatex2019)*, Copenhagen, Denmark, September 1-4.
- Srivastava N, Hinton G., Krizhevsky A., Sutskever I. and Salakhutdinov R. (2014). “Dropout: a simple way to prevent neural network from overfitting”, *Journal of Machine Learning Research*, 15.
- Tukey, J. W. (1977). “Exploratory Data Analysis”, Addison- Wesley, Reading, MA.
- Zambuto F., Buzzi M. R., Costanzo G., Di Lucido M., La Ganga B., Maddaloni P., Papale F. and Svezia E. (2020). “Quality checks on granular banking data: an experimental approach based on machine learning”, *Banca d’Italia, Occasional Papers*, No. 547.
- Zambuto F., Arcuti S., Sabatini R. and Zambuto D. (2020). “Application of classification algorithms for the assessment of confirmation to quality remarks”, *Banca d’Italia, Occasional Papers*, No. 631.
- Zhou and Goldman S. (2004). “Democratic co-learning”, in *IEEE 16th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 594–602.
- Zhou Z. and Li M. (2005). “Tri-training: exploiting unlabeled data using three classifiers”, *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541.
- Wang Z., Bovik A.C., Sheikh H.R. and Simoncelli E.P. (2004). “Image quality assessment: from error visibility to structural similarity”, *IEEE transactions on image processing*, 13(4):600–612.
- Wolpert, D. (1992). “Stacked generalization”, *Neural Networks*, 5, 241-260, [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).

Yarowsky D. (1995). “Unsupervised word sense disambiguation rivaling supervised methods”. In Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pages 189–196, Association for Computational Linguistics.

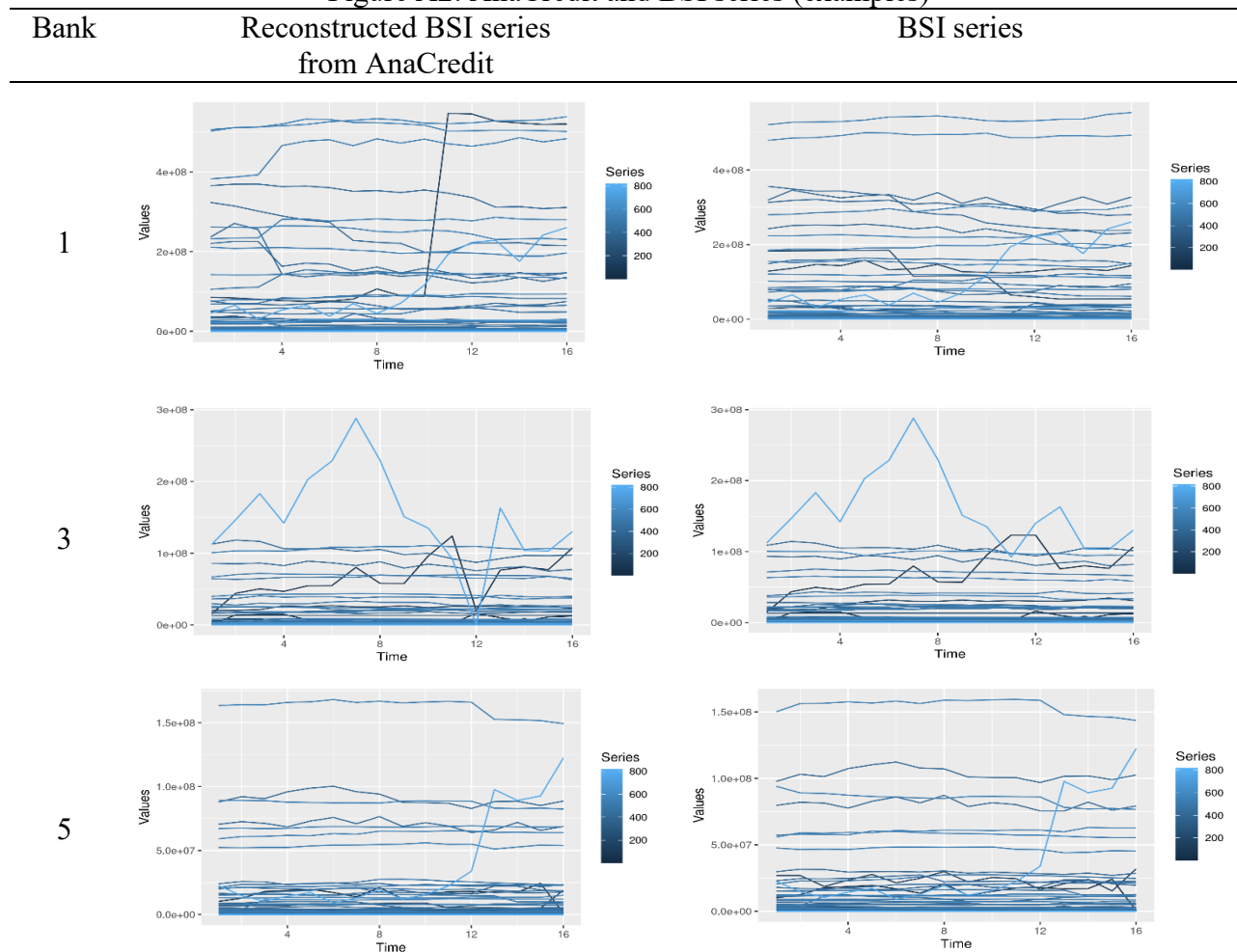
Appendix A - Tables and Charts

Figure A1: BSI aggregates



Amounts in billions of euros.

Figure A2: AnaCredit and BSI series (examples)



Amounts in billions of euros.

Figure A3: AnaCredit and FinRep series (examples)

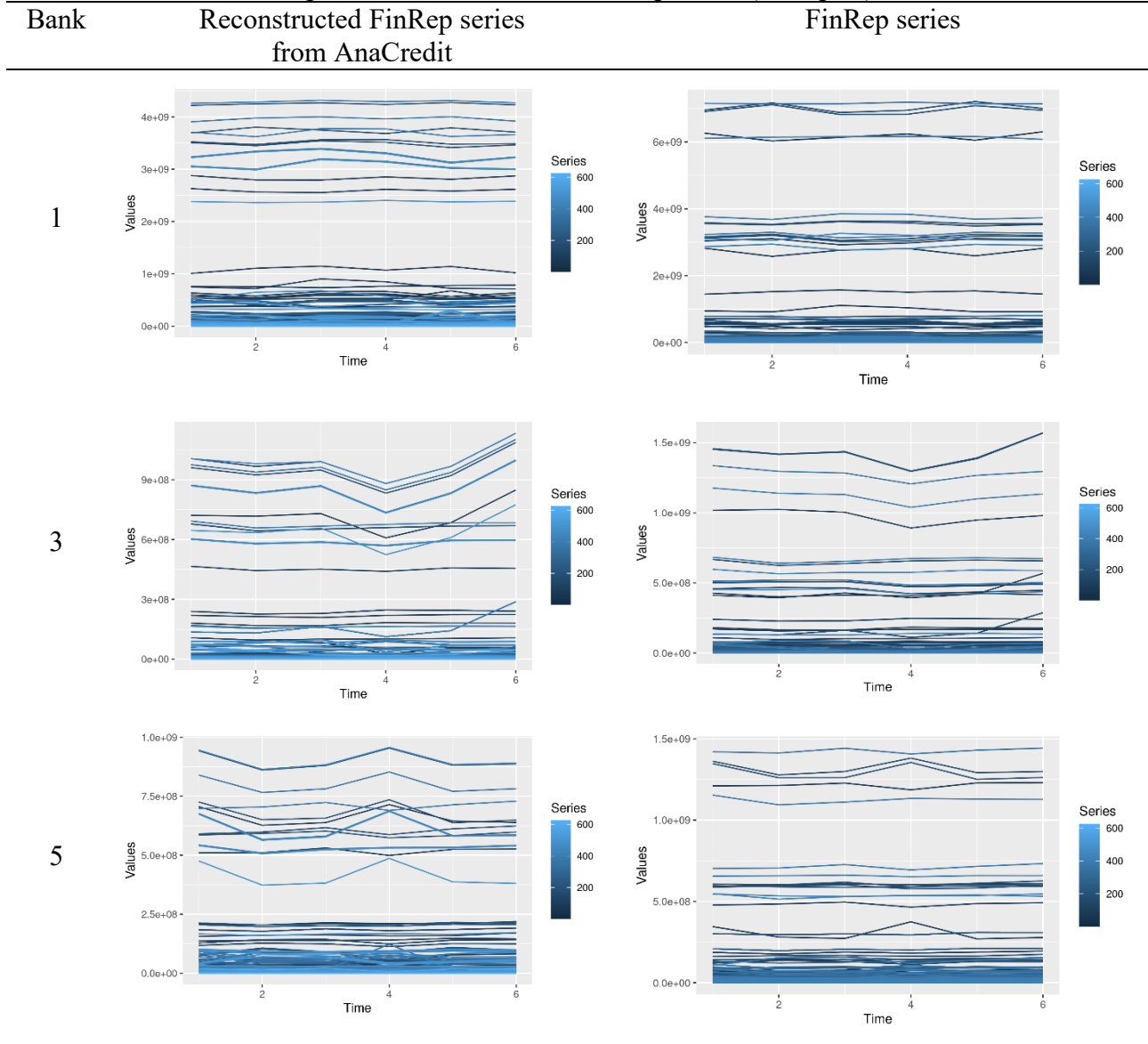


Figure A4: Distributions of the correlation between the compared aggregates

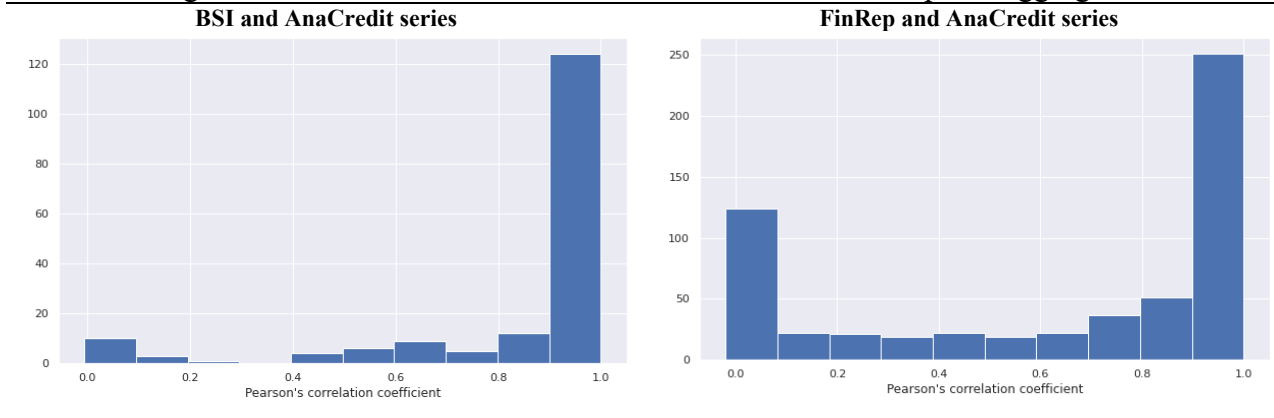


Table A1: stratified sampling dimension

| Strata | AnaCredit-BSI | | AnaCredit-FinRep | |
|--------------|---------------|-------------|------------------|--------------|
| | not sampled | sampled | not sampled | sampled |
| 0-0-0 | 4192 | 11 | 22210 | 9 |
| 0-0-1 | 2 | 2 | 35 | 2 |
| 0-1-0 | | | 13 | 2 |
| 0-1-1 | | | 4 | 2 |
| 1-0-0 | 425 | 11 | 3529 | 8 |
| 1-0-1 | 8 | 3 | 3 | 2 |
| 1-1-1 | 0 | 1 | | |
| 0-cases | | 2529 | | 9761 |
| 1-cases | | 587 | | 1756 |
| Total | 4627 | 3144 | 25794 | 11542 |

x-y-z stratum is identified respectively by Robust Regression (x), CNN (y), DNN (z) by binary prediction (0 not anomaly, 1 anomaly). 0-cases previously classified not anomalies and 1-cases previously classified anomalies

Figure A5: performance metrics

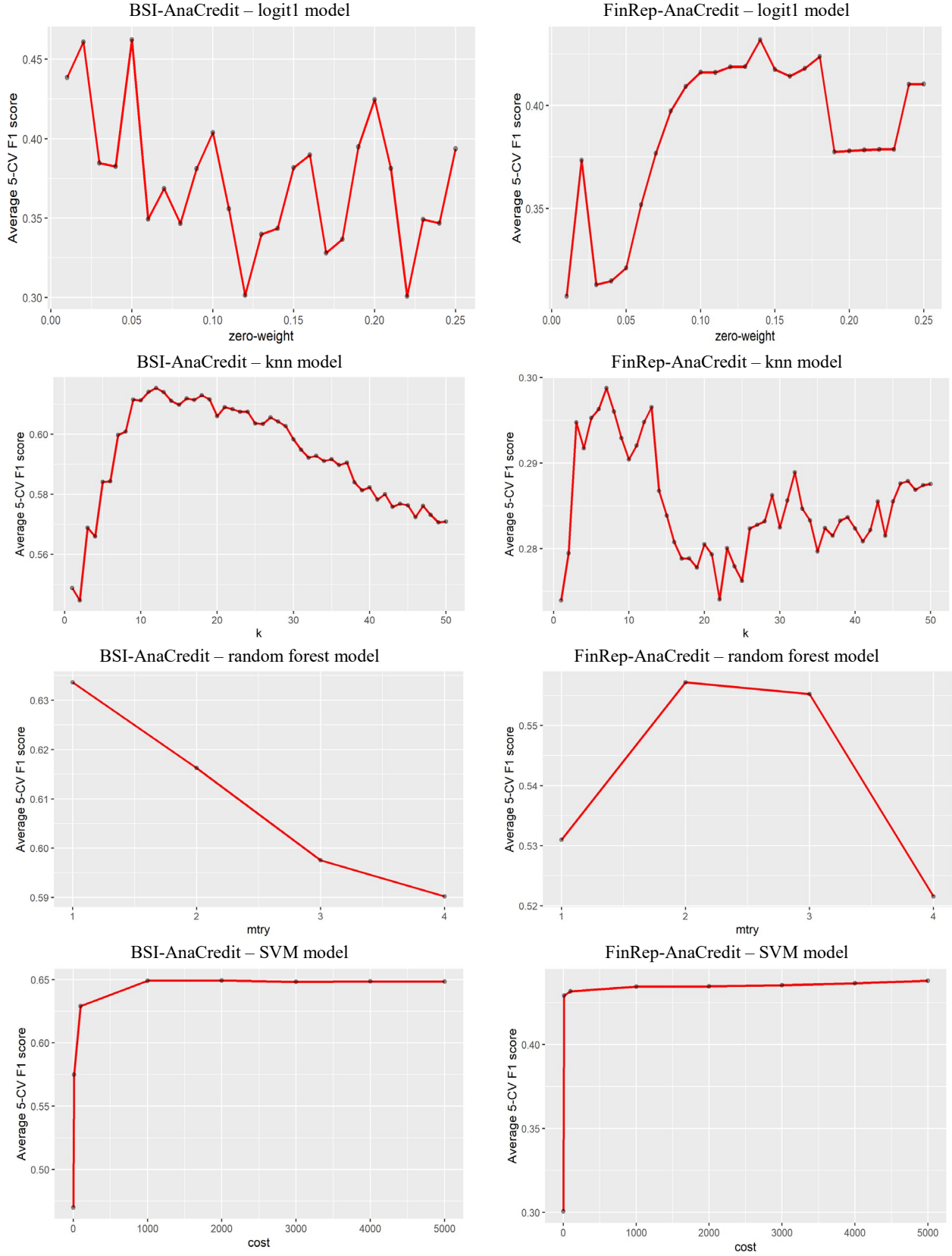
| | | Predicted | |
|--------|--------------|---------------------|---------------------|
| | | Positive (1) | Negative (0) |
| Actual | Positive (1) | True positive (tp) | False negative (fn) |
| | Negative (0) | False positive (fp) | True negative (tn) |

0=Not outlier, 1=Outlier

Metrics:

| | | |
|-------------------|---|------------------------|
| Precision | $\frac{tp}{(tp + fp)}$ | |
| Recall | $\frac{tp}{(tp + fn)} = \frac{p}{p}$ | <i>tpr sensitivity</i> |
| | $\frac{tn}{(tn + fn)} = \frac{tn}{n}$ | <i>tnr specificity</i> |
| Accuracy | $\frac{(tp + tn)}{(tp + fp + fn + tn)}$ | |
| F1 score | $2 * \frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$ | |
| Balanced accuracy | $\frac{(tpr + tnr)}{2}$ | |

Figure A6: optimizing meta-parameters



The parameters have been optimized with respect to the mean value of the simulated response variables.

Appendix B - Robust regression equation

Let $P_{i,j,t,s}$ be the amount of loan s of the subportfolio j granted from bank i at reference date t . Then the aggregated amount for bank i and the j -th subportfolio at date t is given by:

$$(B.1) \quad F_{i,j,t} = \sum_{s=1}^{J_{i,j,t}} P_{i,j,t,s},$$

where $J_{i,j,t}$ is the number of loans in the the j -th subportfolio.

In the AnaCredit framework, there are different criteria triggering a reporting obligation of loans (i.e. that regulatory threshold of 25,000 euros, the counterparty classified as legal persons, etc.). Therefore, the sum is only restricted to such eligible loans:

$$(B.2) \quad A_{i,j,t} = \sum_{s=1}^{J_{i,j,t}} P_{i,j,t,s} \cdot I(s: s \in \text{eligible loans}),$$

where $I(x)$ is an indicator function which is equal to 1 in case of eligible loan. $A_{i,j,t}$ is by definition less or equal to $F_{i,j,t}$. Since we have errors in AnaCredit data ($\xi_{i,j,t}$), we can express the amount observed: $A_{i,j,t}$ as the product of the ‘true’ value $A_{i,j,t}^*$ and an error $\xi_{i,j,t}$: $A_{i,j,t} = A_{i,j,t}^* \cdot \xi_{i,j,t}$. Therefore, the difference between the logarithm two aggregates of the two compared datasets can be expressed as follows:

$$(B.3) \quad \log(A_{i,j,t}) - \log(F_{i,j,t}) = \log(A_{i,j,t}^*) + \log(\xi_{i,j,t}) - \log(F_{i,j,t})$$

For an unbiased estimator $T_{i,j,t}$ of the ‘true’ difference, in the form $T_{i,j,t} = \log(A_{i,j,t}^*) - \log(F_{i,j,t}) - u_{i,j,t}$, so the equation (B.3) can be written,

$$(B.4) \quad \log(A_{i,j,t}) = \log(F_{i,j,t}) + T_{i,j,t} + \log(\xi_{i,j,t}) + u_{i,j,t}$$

The previous equation represents the theoretical model, in which the reporting error adds to a white noise $u_{i,j,t}$. An easy empirical specification to capture the relation between the two variables is by means of an Autoregressive Distributed Lag model (Granger, 1981) of order (1,1):

$$(B.5) \quad \log(A_{i,j,t}) = \alpha_0 + \alpha_1 \log(A_{i,j,t-1}) + \alpha_2 \log(F_{i,j,t}) + \alpha_3 \log(F_{i,j,t-1}) + \epsilon_{i,j,t}$$

As we are interested to capture the differences in the aggregated amounts as independent variable, in order to specify T as function of past differences, we impose the restriction $\alpha_3 = -\alpha_1$. This restriction allows to obtain the following restricted model:

$$(B.6) \quad \log(A_{i,j,t}) = \alpha_0 + \alpha_2 \log(F_{i,j,t}) + \alpha_3 \log(F_{i,j,t-1}/A_{i,j,t-1}) + \epsilon_{i,j,t}$$

When $E(T_{i,j,t}) = \alpha_3 \log(F_{i,j,t-1}/A_{i,j,t-1})$ and $\alpha_0 = 0$, $\alpha_2 = 1$, this last equation is equivalent to our theoretical model in B.4, where $\epsilon_{i,j,t} = \log(\xi_{i,j,t}) + u_{i,j,t}$. Since T is non-positive and the difference referred to $t-1$ is non-negative, then the coefficient α_3 must be non-positive. It’s worth to noting that the error reporting term ($\xi_{i,j,t}$) is only included in the error component term of equation B.6.

In our work, we have applied a log transformation²¹ of the original data values.

²¹ We adopt the $\log I p$ function of x , that is the natural logarithm of $x+I$.

Stacking machine-learning models for anomaly detection: comparing AnaCredit to other banking credit data

P. Maddaloni, D.N. Continanza, A. del Monaco, M. di Lucido,
D. Figoli, F. Quarta and G. Turturiello, Bank of Italy

February 2022

Overview

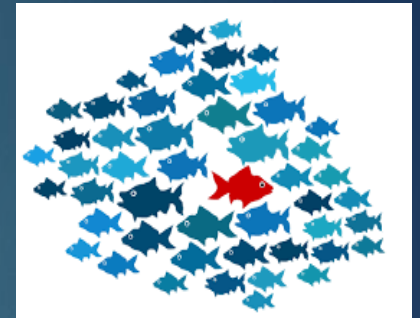
2

- ▶ Goal
- ▶ Data
- ▶ Proposed methodology
- ▶ Results
- ▶ Q&A

Goal

3

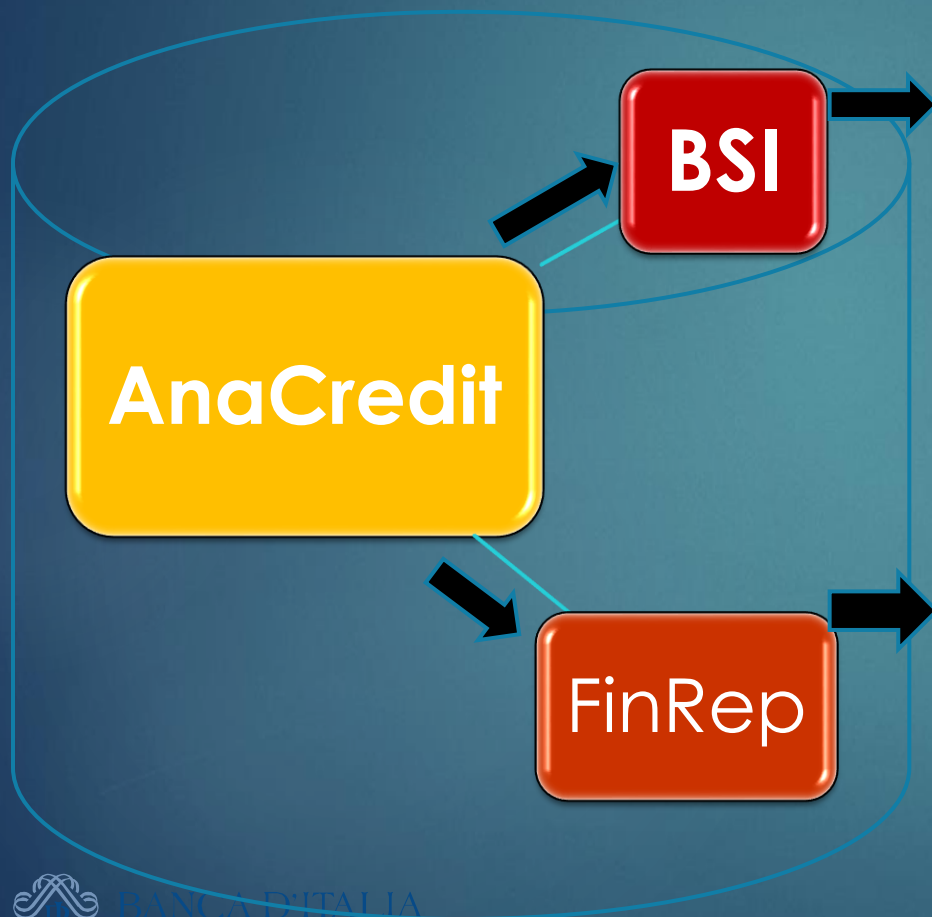
Strengthen the system of quality controls overseeing the AnaCredit, BSI and FinRep, surveys



Create cross- and personalized- checks for identifying potentially anomalous reports

Pair-wise comparison

Italian primary reporting: more granular series of BSI and FinRep



Monthly cross-dataset,
December 2018-March 2020

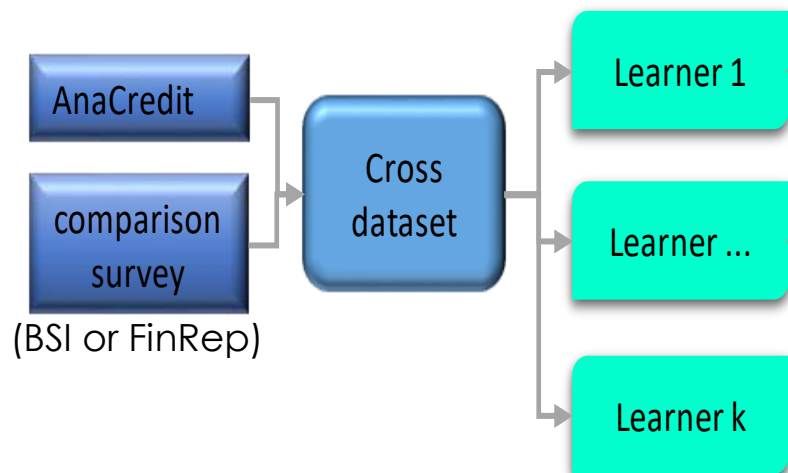
1. Outstanding nominal amount of loans

Quarterly cross-dataset,
December 2018-March 2020

1. the net and gross carrying amount,
2. the accumulated impairment amount,
3. the accumulated changes in fair value due to credit risk

Proposed methodology

5



We develop three ($k=3$) learners:

Robust regressions for linear relationship

Dense Autoencoder

Convolutional Autoencoder

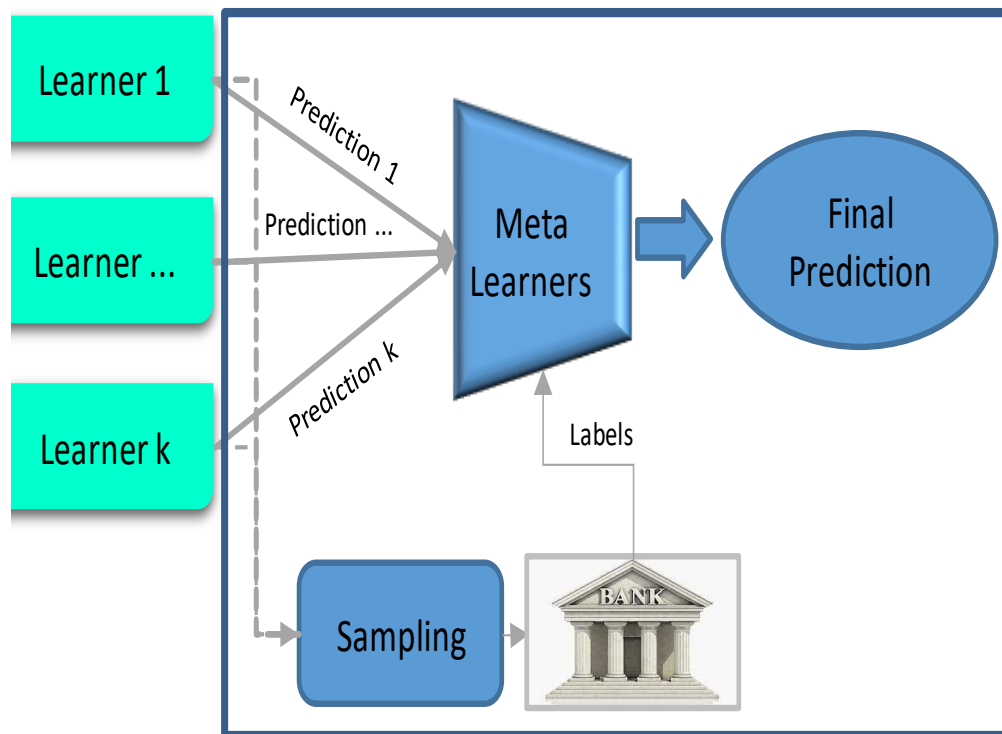
The predictions of each learner are **anomaly scores** normalized in the unitary interval

Proposed methodology

6

Stacking the prediction in semi-supervised settings

Pre-labelled cases: from a sample basis
(revisions for future extensions)



* In our work $k=3$

Proposed methodology

7

Stacking methods on partially labelled datasets:

First setting (static approach)

Pre-labelled cases are reported to the universe in a Montecarlo simulation

Second setting

Semi-supervised approach: iterative labelling

Results

8

First setting

| | AnaCredit-BSI | | AnaCredit-FinRep | |
|------------------|--------------------|--------------------|--------------------|--------------------|
| | Precision | F1 score | Precision | F1 score |
| Baseline* | 0.981±0.019 | 0.997±0.002 | 0.998±0.002 | 0.997±0.003 |
| LOGIT1 | 0.901±0.016 | 0.585±0.017 | 0.833±0.008 | 0.508±0.009 |
| KNN | 0.848±0.014 | 0.596±0.015 | 0.501±0.005 | 0.290±0.005 |
| RF | 0.864±0.018 | 0.603±0.015 | 0.747±0.010 | 0.548±0.009 |
| SVM | 0.830±0.021 | 0.635±0.017 | 0.503±0.006 | 0.432±0.010 |

Second setting

| | AnaCredit-BSI | | AnaCredit-FinRep | |
|----------------------|--------------------|--------------------|--------------------|--------------------|
| | Precision | F1 score | Precision | F1 score |
| Self-training | 0.995±0.005 | 0.995±0.003 | 0.723±0.167 | 0.598±0.098 |
| SETRED | 0.993±0.007 | 0.991±0.003 | 0.929±0.072 | 0.634±0.034 |
| Tri-training | 0.993±0.007 | 0.991±0.002 | 0.825±0.075 | 0.634±0.034 |
| Co-Bagging | 0.989±0.004 | 0.992±0.002 | 0.825±0.075 | 0.650±0.018 |
| Democratic-Co | 0.993±0.007 | 0.992±0.002 | 0.833±0.001 | 0.500±0.001 |

Questions?

*Thank you for
your
attention!*

Contact: pasquale.maddaloni@bancaditalia.it



BANCA D'ITALIA
EUROSISTEMA

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Now-casting business and financial cycles using low- and high-frequency data¹

Alberto Americo, Frederik Hering and Rukmani Vaithianathan,
BIS

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



Innovation  BIS2025

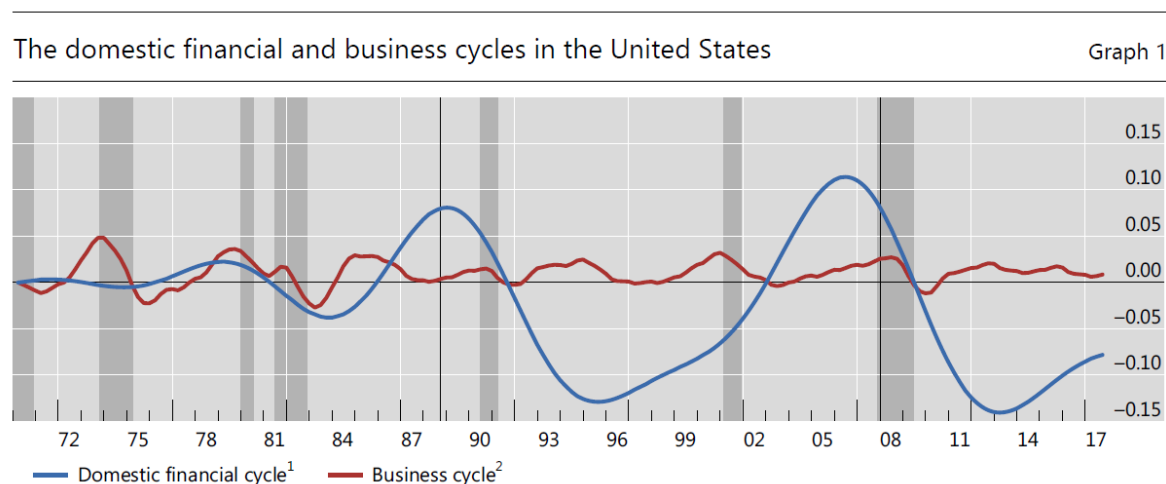
Shaping the Bank for tomorrow

Now-casting Business and Financial Cycles Using Low- and High-Frequency Data

Alberto Americo, Frederik Hering and Rukmani Vaithianathan (BIS)

Now-casting business and financial cycles: **objective of the project**

Objective: *Nowcasting and short-term forecasting* of economic and financial activity using a fully automated infrastructure. We seek to contribute to the understanding of these cycles using a wide range of macro and micro data, including qualitative information, from a wide range countries.



The shaded areas indicate recessions; the solid black lines indicate the start of a banking crisis as defined by Laeven and Valencia (2018).

¹ The financial cycle as measured by frequency-based (bandpass) filters capturing medium-term cycles in real credit, the credit-to-GDP ratio and real house prices. ² The business cycle as measured by a frequency-based (bandpass) filter capturing fluctuations in real GDP over a period from one to eight years.

Source:
Aldasoro, Avdjiev, Borio and Diyatat (2020)

Business cycles

- Recession dates (eg NBER)
- Business Climate (eg OECD)

Financial cycles

- Asset price cycles (equities and house prices)
- Financial stability cycle (banking crises)
- Leverage cycles (credit growth)
- Capital flows cycles
- Composite financial cycle

Now-casting business and financial cycles: **concept and data**

Methods:

- Econometrics models and machine learning:
 - Binary logistic regression models
 - Exhaustive feature selection (up to 5 variables)

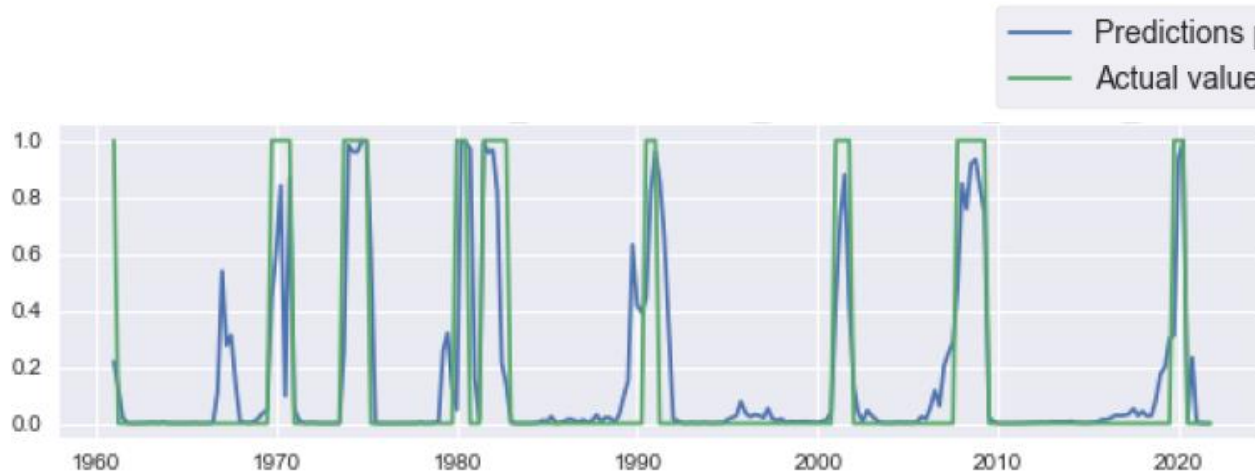
Data:

- *Countries*: OECD countries and EMEs (cluster of countries by data availability)
- *Explanatory variables*: more than 50 variables collected and updated daily from macroeconomic and financial market indicators to micro data.
- *Frequencies*: daily, monthly and quarterly.
- *Sample*: 1950 – Today

Now-casting business and financial cycles: **preview on results**

Business cycle for the US (1961-2021)

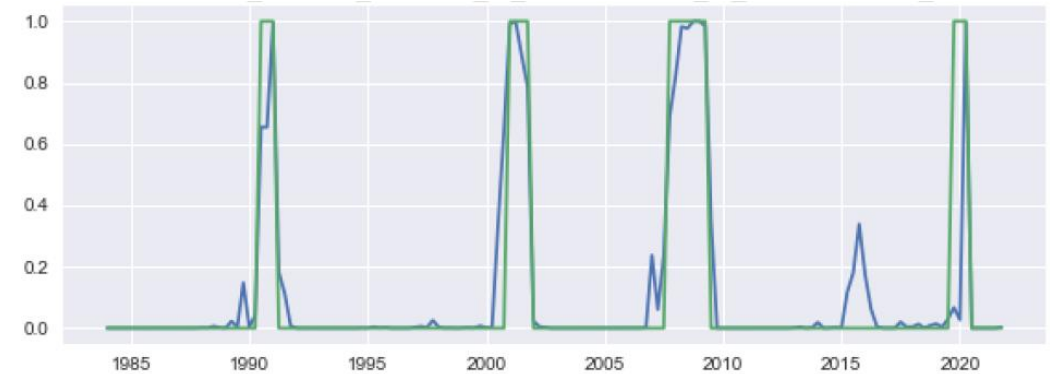
Logistic regression in-sample predictions



Explanatory variables:

- Policy rate
- Private consumption growth
- Unemployment rate ($t-1$)
- Policy rate ($t-3$)
- Current account balance ($t-2$)

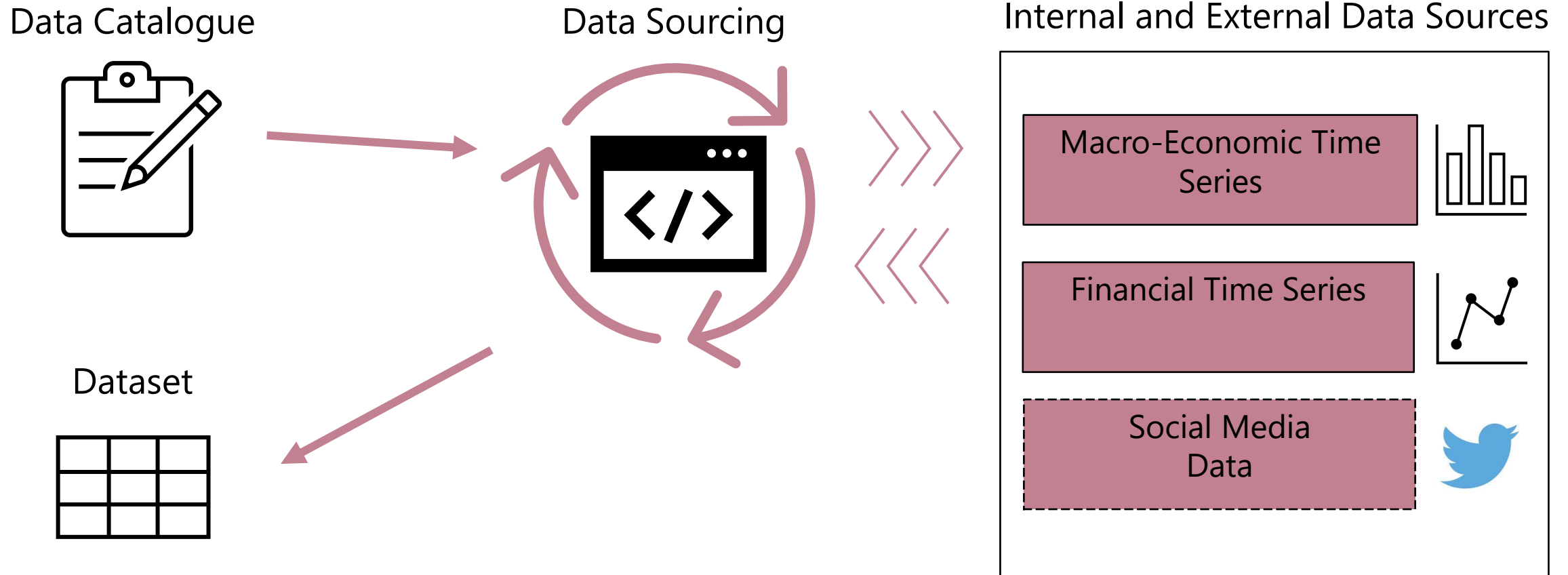
Business cycle for the US (1984-2021)



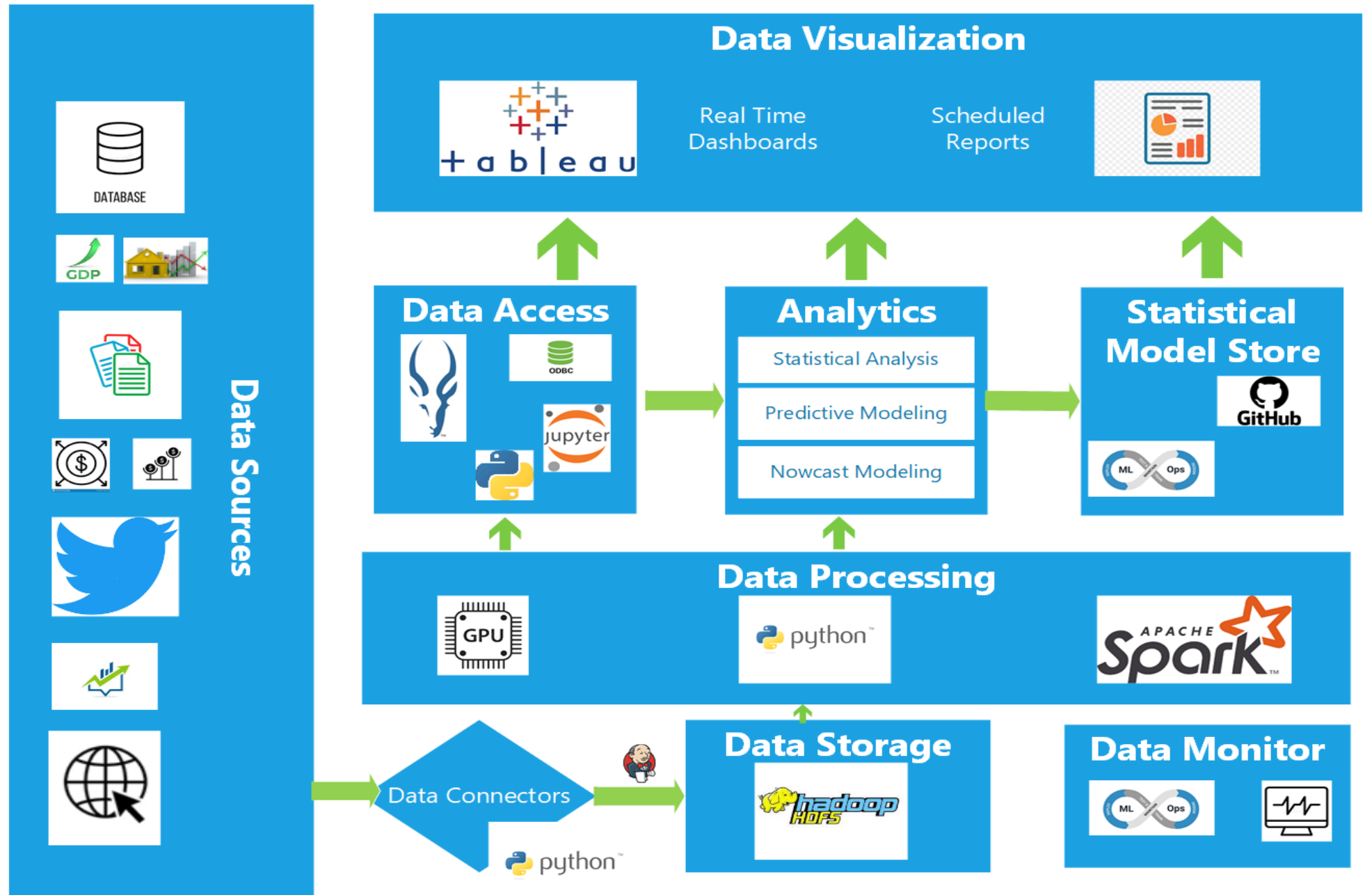
Explanatory variables:

- Credit to households growth
- Gross fixed capital formation growth
- GDP growth ($t-1$)
- Credit to government growth ($t-2$)
- Policy rate ($t-2$)

Now-casting dataset concept: Combining multiple data sources



NOWCAST ML PIPELINE ARCHITECTURE



Now-casting business and financial cycles: **next steps**

- Model calibration
- Compare different approaches (e.g. EFS vs step-wise)
- Out-of-sample simulations
- Cross-country comparison and robustness checks
- End-to-end pipeline
- Social media data ingestion

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Nowcasting economic activity with mobility data¹

Koji Takahashi, BIS,
Kohei Matsumura, Yusuke Oh and Tomohiro Sugo, Bank of Japan

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Nowcasting Economic Activity with Mobility Data^{*}

Kohei Matsumura[†]; Yusuke Oh[‡]; Tomohiro Sugo[§]; Koji Takahashi^{**}

Abstract

We develop high frequency indexes to measure sales in service industries and production activity in the manufacturing industry by using GPS mobility data from mobile applications. First, focusing on the possibility that the number of customers in service industries can be estimated using mobility data, we develop indicators to capture economic activity in amusement parks, shopping centres, and food services. We show that using GPS mobility data, it is possible to nowcast economic activity in the service industries, in real time, with a high level of precision---something which conventional statistics are largely unable to assist. In addition, by using a clustering method, we can construct an indicator with even better nowcasting performance. Second, in the manufacturing sector we identify the locations of large factories using factory-level data from the Economic Census and by utilizing hourly and daily mobility patterns such as a daytime ratio. We then construct indicators for nowcasting production based on the population in the specified areas. We find that we can nowcast production with a high level of precision for some labour-intensive industries including the transportation equipment and production machinery industries. These results suggest that mobility data are a useful tool for nowcasting macroeconomic activity.

Keywords: C49, E23, E27

JEL classification: mobility data, nowcasting, clustering

* The authors are grateful to Seisaku Kameda, Kenji Sakuta, Takuji Kawamoto, Kenichi Sakura, Kazushige Kamiyama, Jouchi Nakajima, Tomoyuki Iida and participants of the Big Data forum co-hosted by the University of Tokyo Center for Advanced Research in Finance and the Bank of Japan Research and Statistics Department, for helpful comments and discussions. Any remaining errors are ours. The views expressed in this paper are those of the authors and do not necessarily reflect the official views of the Bank of Japan or Bank for International Settlements.

[†] Bank of Japan, kouhei.matsumura@boj.or.jp

[‡] Bank of Japan, yuusuke.ou@boj.or.jp

[§] Bank of Japan, tomohiro.sugou@boj.or.jp

^{**} Bank for International Settlements, koji.takahashi@bis.org

1 Introduction

Since the 2010s, the rapid development of information and communication technology has made it possible to collect and use big data in business, which were previously unobtainable from conventional statistics and questionnaire surveys. In particular, in business marketing, big data based on the global positioning system (GPS) of cellular phones are widely used as a source to understand the consumption behavior of customers by taking into account their characteristics. Meanwhile, in the field of economics, especially in macroeconomics, point of sale (POS) data have been widely used in the analysis of supermarket sales and prices since the 1990s. The use of big data in macroeconomics has been rapidly spreading since the beginning of 2020, when the COVID-19 pandemic began to gather pace.

More specifically, the spread of COVID-19 has generated large economic fluctuations in short periods due to subsequent lockdowns and declarations of a state of emergency. The economic turmoil brought on by the pandemic has, therefore, massively increased the importance of understanding economic conditions in a timely manner for policy makers. However, the conventional statistics and questionnaire surveys on which central bankers have relied for economic analysis take at least several weeks for data to be released after the survey. This is due to the time which the process of collecting and compiling the data takes. To address this issue, the use of big data, especially high frequency data, has been rapidly spreading among central bankers as well as the utilization of information obtained through interviews with firm managers.¹

In this paper, focusing on GPS data, we show how to use big data to nowcast economic conditions in real time from macroeconomic perspectives. Specifically, by combining the GPS data with information on coordinates of commercial and public facilities such as shops and factories, we closely examine which sectors the data can be applied to for nowcasting with a high level of precision and frequency. We find that for some sectors, we can nowcast household consumption and firm production with high accuracy by extracting related information from the mobility data. This result implies that mobility data are useful not only for marketing but also for understanding macroeconomic conditions.

The remainder of this paper is organized as follows. Section 2 explains related literature and the dataset used in this paper. Section 3 shows that the GPS data are useful for nowcasting service industries. Section 4 demonstrates a methodology to nowcast the index of the industry production. Section 5 concludes by pointing out the caveats of using mobility data for these purposes.

2 Literature Review and Data

In this section, we discuss literature related to our paper and explain our dataset.

¹In this paper, big data include (1) high frequency data that are updated frequently, (2) high granular data such as transaction data between companies, and (3) text data such as those on social networking services.

2.1 Literature Review

Our paper is mainly related to two strands of literature. First, this paper relates to the literature on analyses of human mobility based on smartphone GPS data. Since the beginning of the COVID-19 pandemic, a growing number of papers study the effect of lockdowns using mobility data. For example, in the United States, Couture et al. (2021) develop a location exposure index and a device exposure index using smartphone mobility data. The former describes county-to-county movements of people and the latter quantifies people's exposure to others in commercial venues. Furthermore, using smartphone GPS data, Coven and Gupta (2020) show that since the pandemic, people with higher incomes tend to move away from urban areas whereas people with lower incomes are likely to continue going out and commuting to their work places as before.² On China, Fang et al. (2020) investigate the effects of lockdown on human mobility using smartphone GPS data. On Japan, Watanabe and Yabu (2020) develop a "stay-at-home" index and quantitatively investigate to what extent Japanese shelter-in-place behavior is explained by intervention effect and information effect. The former represents a direct effect resulting from an intervention such as the declaration of a state of emergency whereas the latter means an effect resulting from an announcement such as a news story about the number of infected people. While these studies focus on the mobility of people, they do not explicitly investigate the relationship between mobility data and economic activity.

The second strand of related literature is studies on the development of nowcast indexes. Among them, Cajner et al. (2019), together with other economists of the Federal Reserve Board (FRB), develop the ADP-FRB active employment index to understand the labor market conditions using data from a private company, Automatic Data Processing (ADP). The FRB uses the index in order to grasp current labor market conditions and wage. In addition, a growing number of studies use payment data for nowcasting. For example, Aprigliano et al. (2019) show that Business-to-Business and Business-to-Consumer payment data help to improve the accuracy of nowcasting GDP, business fixed investment, and consumption. Furthermore, Galbraith and Tkacz (2018) find that payment data for debit cards and checks are useful for nowcasting GDP and consumption and Aladangady et al. (2019) develop an index to nowcast consumption in real time by exploiting debit and credit card payment data. Since the occurrence of the pandemic, Chetty et al. (2020) developed economic indexes using big data compiled by private companies. They report that based on the indexes, a reduction in the service consumption of people with higher incomes leads to a decrease in income and the consumption of people with lower incomes working in those industries. These approaches are useful for practitioners as they allow them to grasp macroeconomic conditions before related public statistics are released. However, these studies use high frequency data that are directly linked to economic activities and do not use location data.

As for the studies on nowcasting economic activity using location data, Dong et al. (2017) develop indexes to nowcast sales of firms and consumption in some service industries in China. In addition, Arslanalp et al. (2019) and Cerdeiro et al. (2020) develop indexes to nowcast trade volume in the world using the traffic data of vessels.

²Using smartphone mobility data, Chen and Pope (2020) find that people with higher incomes travel longer distances and go to many places in the United States.

However, to date, few studies use location or mobility data for nowcasting economic activity.

Our paper extends these two bodies of literature and thereby contributes to the development of nowcasting indexes of economic conditions with high frequency and a high level of precision, which indicates that mobility data are a useful tool for nowcasting macroeconomic activity.

2.2 Data

We use mobility data from January 2017 to March 2020 compiled by Agoop. The data show the estimated number of people in every hour in each mesh element, which is defined as a 100m×100m square, dividing Japanese territory into about 20 million squares.³ The estimation is based on GPS data, which are collected through smart-phone applications with the approval of users.

However, we cannot measure any economic activities such as consumption and production simply using hourly population data because the data only include information on the coordinates of meshes in addition to hourly population. Therefore, in order to associate the number of people in a mesh with a specific economic activity, we combine the data with the 1) Economic Census for Business Activity compiled by the Ministry of Economy, Trade and Industry, 2) National Land Numerical Information compiled by the Ministry of Land, Infrastructure, Transport and Tourism and/or 3) point of interest data with the application programming interface (API). Thus, we develop indexes to capture the consumption or production activity based on the number of people in a mesh.⁴

More specifically, we identify a type of a facility and building located in each mesh and infer whether the hourly population in the mesh is associated with consumption such as retail, leisure, or food services, or with production in factories. By doing so, we capture the economic activity based on the number of people in each mesh.

2.3 Calculation of Indicators

We develop an economic indicator from GPS data (ELG) to capture economic activity such as sales in service sectors using population data by following the steps below.

First, we specify nowcasting sector J and select a set of meshes (I_t^J) related to the sector in time t . Then, we sum the number of people (Pop_{it}) in each selected mesh i using weight w_{it} , which is calculated with other data such as the Economic Census for Business Activity. Thus, we obtain the total population in the selected meshes ($TotalPop_t^J$) as follows,

$$TotalPop_t^J = \sum_{i \in I_t^J} w_{it} \times Pop_{it}. \quad (1)$$

³The data do not include meshes such as mountainous areas where the average population is almost zero.

⁴The Economic Census for Business Activity and API point of interest data include information on the business categories of firms in their specific meshes. The National Land Numerical Information provides information on the land use, which allows us to relate the number of people in some meshes to a specific economic activity.

Second, as set I_t^J does not necessarily include meshes that cover all the facilities and buildings related to the industry, we define EIG as a normalized index as follows,

$$EIG_t^J = \frac{TotalPop_t^J}{TotalPop_s^J} \times 100 \quad (2)$$

where $TotalPop_s^J$ indicates the total number of people at reference point of time s . It should be noted that the number of people in meshes are available on an hourly basis. Therefore, economic activity can be analyzed hourly or weekly, as well as on the monthly frequency used in typical conventional statistics.

3 Mobility Data and Service Industries

This section illustrates the methodology of constructing indexes to nowcast current economic conditions in service industries. In service industries, the number of people in a commercial facility is likely to represent the number of its customers. Therefore, if we count the number of people accurately, we can nowcast the number of customers or sales in that facility.

However, it is unusual that only one type of economic activity is conducted in a mesh. Suppose we are interested in food service industry. However, a mesh where a restaurant is located often includes a supermarket or a department store. It is therefore important to exclude some meshes that are affected by customers of the department store and the supermarket in order to nowcast economic activity in the food service industry accurately.

Against this background, in what follows we develop indexes for nowcasting economic activity in amusement parks, shopping centers, and food service industry by using a statistical method to get rid of noise if necessary and we investigate the plausibility of the indexes by comparing them with existing conventional statistics.⁵

3.1 Amusement Parks

The population in amusement parks is expected to be a proxy for the number of visitors. In addition, there are not many large amusement parks in Japan, which allows us to choose meshes that cover almost all amusement parks. We sum the number of people each day in meshes that cover the main amusement parks.

Figure 1 shows the hourly population by day of the week, indicating that the number of visitors is relatively high in the daytime on weekends as well as on Mondays and Fridays, which correspond to the days before and after weekends. This figure shows that the amusement park characteristic that visitors increase on weekends is captured well by the indicator. Figure 2 shows the year-over-year rate of change of the population in specific meshes and it appears to track the change in the number of visitors to amusement parks obtained from Current Survey of Selected Service Industries, which indicates that the GPS data are useful for nowcasting official statistics.

⁵As Chetty et al. (2020) pointed out, when we use alternative data for economic analysis, we need to modify biases arising from the system changes of data providers. Following the literature, we also modified discontinuity that arises from changes in the data collection process of the data provider.

3.2 Shopping Center

As a shopping center (SC) usually has a large site area, it is easy to extract meshes most related to consumption in SCs by excluding meshes where the hourly population is affected by other economic activities. First, we extract meshes that include the addresses of shopping centers listed by the Japan Council of Shopping Centers (3,203 meshes). Then, we exclude meshes including stations because the hourly population in stations is affected by many other economic activities rather than simply consumption in shopping centers. We also exclude meshes where the average population on weekends is smaller than that on weekdays, in order to exclude meshes with facilities such as offices. After this process, the number of selected meshes will be 2,361 meshes.

Figure 3 shows the average hourly population by day of the week in the selected meshes, indicating that the number of visitors increases in the afternoon on weekends and at 6 p.m. on Fridays. This pattern implies that the population captures the number of visitors to the shopping centers well. In fact, as shown in Figure 4, the year-over-year growth rate of the population from 10 a.m. to 8 p.m. shows similar patterns with those of sales at shopping centers, which are collected by Japan Council of Shopping Centers. The result shows that mobility data are useful for capturing consumption. However, we should note that the index based on the population does not capture fluctuations before and after the consumption tax rate hike in October 2019. This suggests that mobility data may not capture changes in the consumption value per person.

3.3 Food Service Industry

In the food service industry, a site area of each restaurant is expected to be relatively small in general, compared to the size of a mesh. In addition, commercial facilities such as pubs are located in areas where various different activities are conducted. This complicates the selection of meshes in the food service industry. Given these characteristics, we take the following three steps to identify meshes whose population represents activity in the food service industry. First, following Mizuno et al. (2020), in order to exclude residential areas, we extract meshes where the daytime population is greater than 80% of nighttime population (27,595 meshes). Second, we count the number of restaurants within a 300m radius from the center of each mesh and choose meshes where the number is greater than 300.⁶ In addition, we exclude meshes that include stations. The number of extracted meshes then becomes 758. Third, we categorize those meshes into five clusters by using a k-medoids method and then exclude clusters of meshes whose hourly population appears to be affected by other economic activities.⁷

Figure 5 illustrates the hourly population by cluster on weekdays, indicating that cluster 1 (435 meshes) has peaks at noon and 7 p.m. This pattern suggests that the population in cluster 1 successfully captures the visitors to restaurants. On the other hand, the population in other clusters has a peak at 3 p.m., which implies that the

⁶We use Gurus API to count the number of restaurants within a 300m radius from the center of each mesh.

⁷k-medoids is a partitioning method of clustering that splits the data points into k clusters by choosing a medoid of a cluster so that its average dissimilarity to all the objects in the cluster is minimal.

population in those meshes mainly includes workers in offices. As shown in Figure 6, the year-over-year changes in the population of cluster 1 and the consumption index in food services from JCB Consumption NOW show similar fluctuations over time. In addition, the correlation between them is higher than the correlation with the population before clustering as shown in Table 1. Note that even if we exclude the data in March 2020 when the consumption sharply declined because of the shortened hours of restaurants due to the pandemic, the correlation remains high. This indicates that the indicator successfully captures consumption in the food service industry even in normal times.

For the above analysis on the three sectors in the service industry, some may point out that after the pandemic, the number of visitors to commercial facilities in service industry has drastically declined, which increased the correlation between sales and the population. Therefore, some may claim that a high correlation in turbulent times does not guarantee high nowcasting performance in normal times. However, we only use data up to March 2020 when the effect of the pandemic on consumption was limited and show that even before the serious deterioration of the pandemic, the indicator demonstrates a high correlation with other statistics. This result implies that the indicators are a useful tool for nowcasting economic conditions in a timely manner not only in turbulent times when the economic situation could suddenly change, but also in normal times.

4 Mobility Data and Manufacturing Industry

In this section, we outline our methodology for using mobility data for nowcasting industrial production. Generally speaking, production in the manufacturing industry is determined by many different factors—including the utilization rate of capital and level of technology progress—and the impact of each factor differs across sectors. However, irrespective of other factors, labor input is thought of as an important factor for almost all sectors by economists.

If production can be approximated by a simple function of labor input and if the labor input can be captured by population in factories, an indicator based on the population would be a useful tool for nowcasting production and judging economic conditions.

As with the service industry discussed in the previous section, provided that the hourly population in a factory can be a proxy for labor input there, we develop indicators for industrial production. In this section, first, we show a relationship between the industrial production and labor input of conventional statistics. Then, we explain the methodology to construct indicators for nowcasting the industrial production.

4.1 Relationship between Labor Input and Production

Following the discussion above, we check the relationship between labor input and production before investigating the relationship between the population and production. Table 2 shows the correlations between year-over-year growth rates of labor input (hour \times number of workers) from the Monthly Labour Survey and the index of the industrial production, indicating that the correlations differ across sectors and they

are high especially in the transportation equipment and the production machinery industries. For those industries, due to the strong linearity of the relationship between production and labor input, the population data are expected to be useful for nowcasting the production if we can successfully capture the labor input by the population. On the other hand, for other sectors including the electronic parts, devices and electronic circuits industry and the food, beverages and tobacco industry the correlations are almost zero due to the progress of automation in factories in these industries. For those sectors, the labor input is not a major driving factor in determining the production level and therefore it would be difficult to nowcast production in those sectors using mobility data.

4.2 Identifying Meshes Representing Location of Factories

By using panel data from the Economic Census for Business Activity, we specify meshes where production takes place. More specifically, we associate factory information, including addresses and value-added production from the Economic Census for Business Activity with meshes and sum the population in those meshes. In addition, to reduce noise, we use only the top ten thousand factories in the manufacturing industry in terms of value-added production among factories with a site area larger than 10,000 m^2 .

The main issue for identifying meshes related to the production activity is that most of the factories have a site area larger than the minimum size of a mesh (100m by 100m square), but the census does not provide information on the shape of the factory. Nevertheless, in addition to the meshes corresponding to the address of factories in the panel data, we also need to choose additional meshes that sufficiently cover factories in order to nowcast production with a high level of precision. To address this issue, we employ a heuristic approach. After collecting meshes which sufficiently cover all areas that are expected to include factory sites, we exclude the less relevant meshes to production activity. Specifically, we choose meshes based on the three criterion: Sunday ratio, daytime ratio and average population. First, we use the Sunday ratio and daytime ratio as we think those ratios are useful measures to distinguish factories from residential areas. In fact, as shown in the upper panel of Figure 7, the Sunday ratio is low and the daytime ratio is high in a typical factory area. On the other hand, as shown in the lower panel of Figure 7, the Sunday ratio is high and the daytime ratio is low in a typical residential area. Thus, the Sunday ratio and daytime ratio should be good measures to separate factory areas from residential ones. In addition, by excluding meshes where the population is lower than specific thresholds, we mitigate the effects of noise arising from a small sample problem.

In this paper, considering the possibility that the meaningful thresholds of those ratios differ across sectors, we choose a combination of thresholds for the ratios so that the correlation between labor input and the population will be the highest. For example, in the transportation equipment industry, the correlation achieves the highest value of 0.86 when we set the Sunday ratio as 0.5 and the daytime ratio as 0.6 and the average population as 10 people, as shown in Table 3. Therefore, we use those values as the criterion for the transportation equipment industry. Even though the correlation between the population and the index of the industrial production differs across sectors, the population in the selected meshes can capture the labor input on

average, as shown in Figure 8.

4.3 Activity Indicator Based on the Hourly Population Data

In this subsection, we explain the methodology to construct a proxy for production activity using population in the selected meshes. When we aggregate the population in these meshes, we should note that it does not necessarily provide the best nowcast to sum population over all hours and days. This is because the hourly population can change even as a result of an increase in the number of workers irrelevant to production, such as facility maintenance workers. Taking this point into account, we investigate all combinations of choices of (1) whether we include holidays and weekends and (2) which hour we aggregate the population data with. Then, we figure out a combination that provides us with the highest correlation between the indicator based on the population and the industrial production index. Table 4 shows that the population in the evening is an appropriate measure to capture the industrial production on average although there is heterogeneity across sectors.

Figure 9 shows the sum of the population for the selected window of hours in the identified meshes, indicating it traces the development of year-on-year changes in the index of the industrial production well. In particular, in October 2019 when an enormous typhoon hit Japan, the indicator based on the population has massively declined in tandem with the industrial production index, which suggests that we can nowcast the production with a high level of precision using mobility data.

Figure 10 shows the indicators based on the population by industry, indicating that the indicators for the transportation equipment and the production machinery industries nowcast the corresponding indexes of the industrial production relatively well. The high nowcasting performance is due to the fact that labor can explain production because those industries both have a high level of labor intensiveness and the population can serve as a good proxy for labor input. On the other hand, the indicator based on the hourly population for the electronic parts and devices industry shows a poor nowcasting performance, especially in 2019 when production substantially declined. This is because the industry is a capital intensive sector and the population only poorly tracks the labor input.

4.4 Extension: High-frequency Analysis

One possible extension of the analysis in the previous section is to enhance the frequency of the indicator. Since the original data is hourly, it is possible to evaluate the current economic situation in a more timely manner by changing the frequency of the indicator to weekly or daily.

For example, Figure 11 shows the weekly production indicator, which is seasonally adjusted by a statistical method, based on the population data for the transportation equipment industry.⁸ The figure demonstrates that the indicator sharply dropped in the week of October 20th, 2019 in tandem with the industrial production index, which suggests that the indicator based on the population can capture the effect which the number 19 typhoon in 2019 had on the production activity in a timely manner. This result implies that mobility data are also helpful for capturing a sudden fluctuation of

⁸We use an open source software, called "Prophet," developed by Facebook for the seasonal adjustment.

economy even when the economy is hit by natural disasters such as heavy rains or typhoons.

5 Conclusion

In this paper, we develop indicators to capture sales in the service industries and production activity in manufacturing industries using the hourly population based on smartphone mobility data.

As the hourly population in a commercial facility correlates with the number of customers in the service industries, we develop activity indicators for amusement parks, shopping centers, and the food service industry. We find that the indicators based on population are useful for nowcasting activity in the service industry for which conventional statistics are unable to give an understanding of economic activity in real time. In addition, even for a sector where it is difficult to get rid of noise, we can construct an indicator with high nowcasting performance by using a statistical method such as clustering. Furthermore, in the manufacturing industry, we identify meshes related to production activity by using factory panel data from the Economic Census for Business Activity and by utilizing hourly and daily mobility patterns such as the daytime ratio. Then, we sum the population in specified meshes, thereby constructing indicators by industry for nowcasting. We find that we can nowcast with a high level of precision by using the indicators for labor intensive industries including the transportation equipment and production machinery industries.

These results suggest that mobility data are a useful tool for nowcasting macroeconomic activity.

We should note that the mobility data used in this paper include information on the hourly population in a 100m by 100m mesh but do not include information on characteristics such as age and gender of workers and consumers in the area. Therefore, we cannot analyze consumption behavior by attribute, such as that of elderly people, using mobility data alone. In addition, as the data used in this paper are collected through smartphones, they cannot reflect the behavior of people who do not use smartphones. Furthermore, the length of our data in terms of time series would be short to check the plausibility of our analysis. Moreover, as we combine the population data with the Economic Census in this paper, it is important to find other appropriate data for a targeted sector in order to construct an effective indicator. For example, if the relocation of a large factory from one of the selected meshes to another non-selected mesh occurred, the indicator would drastically drop. However, the reason for such a change cannot be found using population data alone. In such cases, we have to use other data in order to uncover the reason for the factory relocation and to control the effect on the indicator if needed. In addition, if we focus on production in a sector where the production process is highly automated, the utilization rate of capital and changes in productivity would be important determinants of production as well as labor input estimated by the population data. Disregarding these factors would degrade the accuracy of developed indicators. Investigating a methodology to address this issue will be a topic for our future research.

References

- Aladangady, A., S. Aron-Dine, W. Dunn, L. Feiveson, P. Lengermann, and C. Sahm (2019): "From transactions data to economic statistics: constructing real-time, high-frequency, geographic measures of consumer spending," NBER Working Paper 26253.
- Aprigliano, V., G. Ardizzi, and L. Monteforte (2019): "Using payment system data to forecast economic activity," *International Journal of Central Banking*, 15, 55–80.
- Arslanalp, S., M. Marini, and P. Tumbarello (2019): "Big data on vessel traffic: Nowcasting trade flows in real time," *International Monetary Fund Working Paper WP/19/275*.
- Cajner, T., L. D. Crane, R. A. Decker, A. Hamins-Puertolas, and C. Kurz (2019): "Improving the accuracy of economic measurement with multiple data sources: The case of payroll employment data," *Finance and Economics Discussion Series 2019-065*, Board of Governors of the Federal Reserve System.
- Cerdeiro, D. A., A. Komaromi, Y. Liu, and M. Saeed (2020): "World seaborne trade in real time: A proof of concept for building AIS-based nowcasts from scratch," *International Monetary Fund Working Paper WP/20/57*.
- Chen, M. K. and D. G. Pope (2020): "Geographic mobility in America: Evidence from cell phone data," NBER Working Paper 27072.
- Chetty, R., J. N. Friedman, N. Hendren, M. Stepner, and the Opportunity Insights Team (2020): "How did Covid-19 and stabilization policies affect spending and employment? a new real-time economic tracker based on private sector data," NBER Working Paper 27431.
- Couture, V., J. I. Dingel, A. Green, J. Handbury, and K. Williams (2021): "Measuring movement and social contact with smartphone data: a real-time application to COVID-19," *Journal of Urban Economics*, 103328.
- Coven, J. and A. Gupta (2020): "Disparities in mobility responses to Covid-19," NYU Stern Working Paper.
- Dong, L., S. Chen, Y. Cheng, Z. Wu, C. Li, and H. Wu (2017): "Measuring economic activity in China with mobile big data," *EPJ Data Science*, 6, 1–17.
- Fang, H., L. Wang, and Y. Yang (2020): "Human mobility restrictions and the spread of the novel coronavirus (2019-nCoV) in china," *Journal of Public Economics*, 191, 104272.
- Galbraith, J. W. and G. Tkacz (2018): "Nowcasting with payments system data," *International Journal of Forecasting*, 34, 366–376.
- Mizuno, T., T. Ohnishi, and T. Watanabe (2020): "Visualization of the rate of self-quarantine based on big population mobility data," *Artificial Intelligence (in Japanese)*.
- Watanabe, T. and T. Yabu (2020): "Japan's voluntary lockdown," CARF Working Paper.

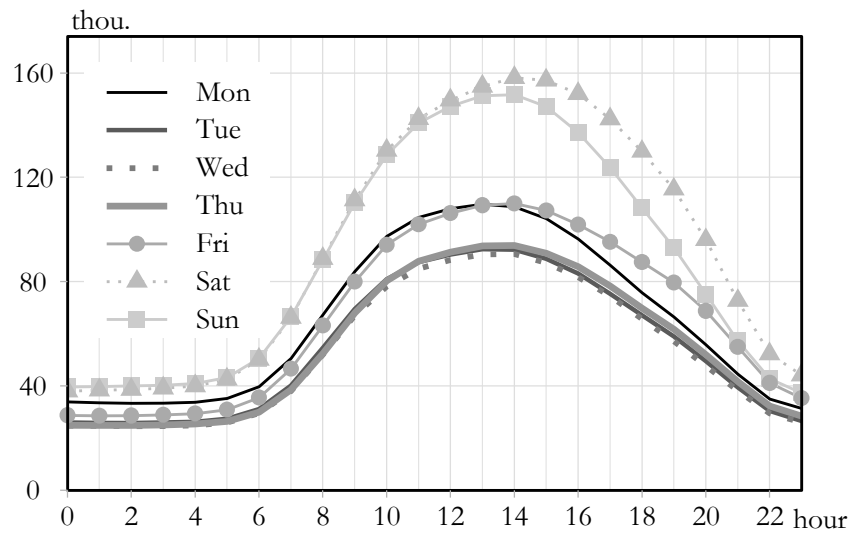


Figure 1: Hourly population by day of the week in amusement parks

Source: Agoop.

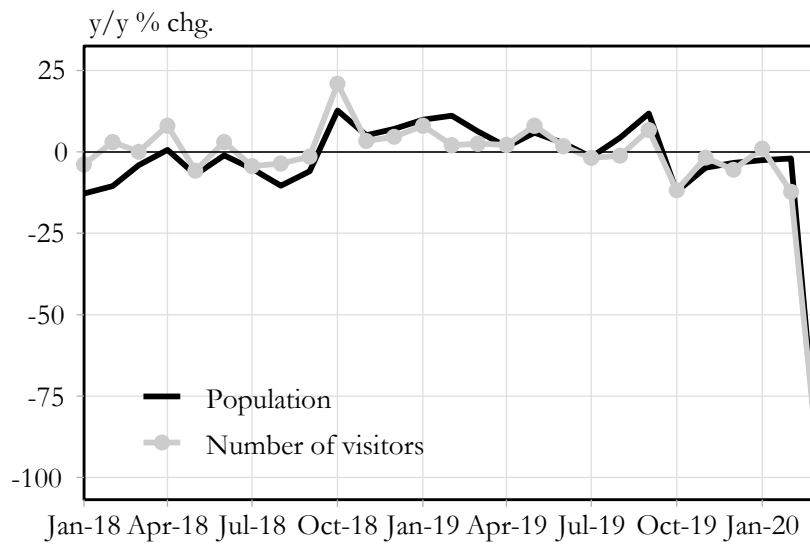


Figure 2: Indicator based on the hourly population in amusement parks

Note: "Number of visitors" indicates the number of visitors to amusement parks based on the Survey of Selected Service Industries.

Sources: Agoop; Ministry of Economy, Trade and Industry

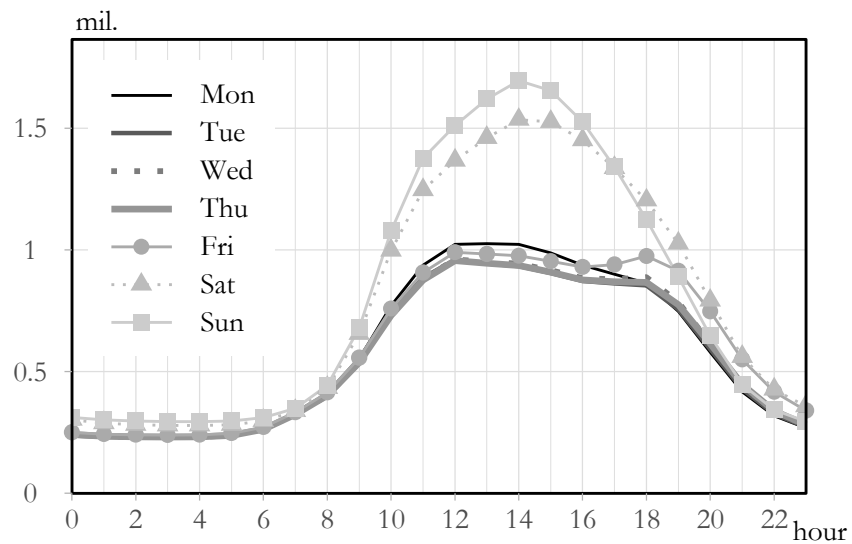


Figure 3: Hourly population by day of the week in shopping centers

Sources: Agoop; Japan Council of Shopping Centers; Ministry of Land, Infrastructure, Transport and Tourism.

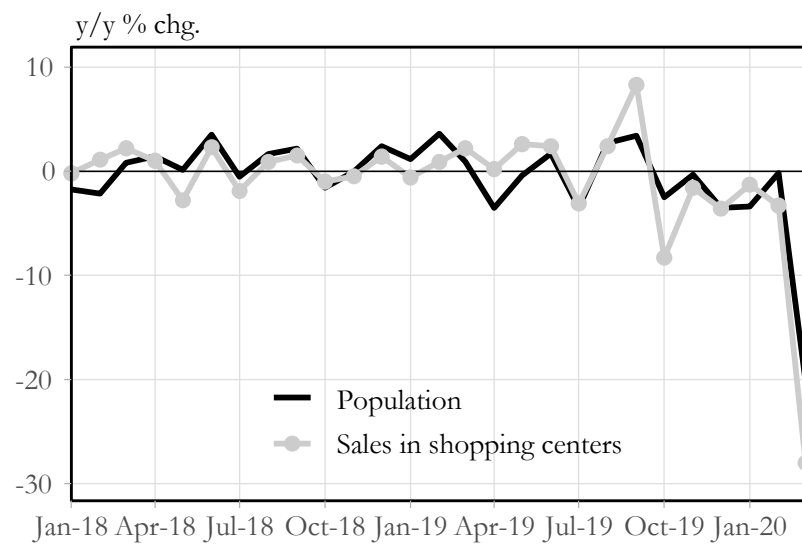


Figure 4: Indicator based on the hourly population in shopping centers

Note: "Sales in shopping centers" is based on the survey by Japan Council of Shopping Centers.

Sources: Agoop; Japan Council of Shopping Centers; Ministry of Land, Infrastructure, Transport and Tourism.

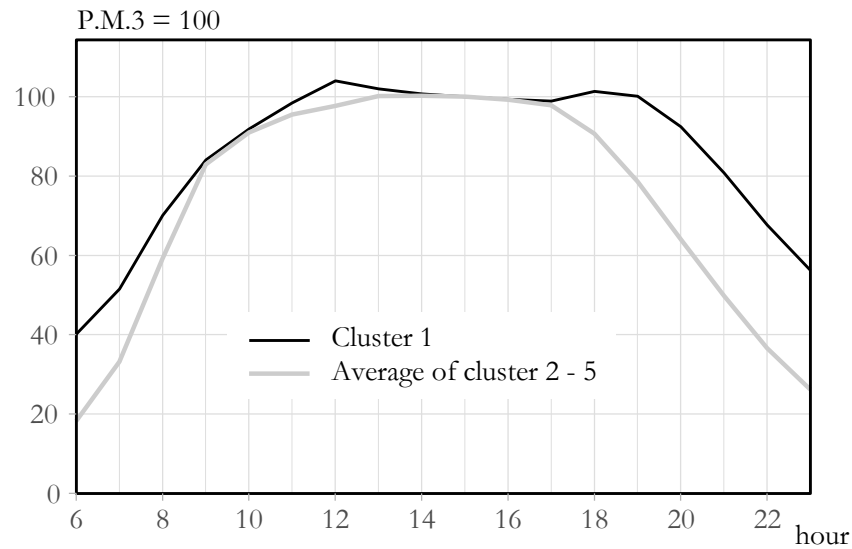


Figure 5: Hourly population on weekdays by cluster for food services

Sources: Agoop; Gurunavi; Ministry of Land, Infrastructure, Transport and Tourism.

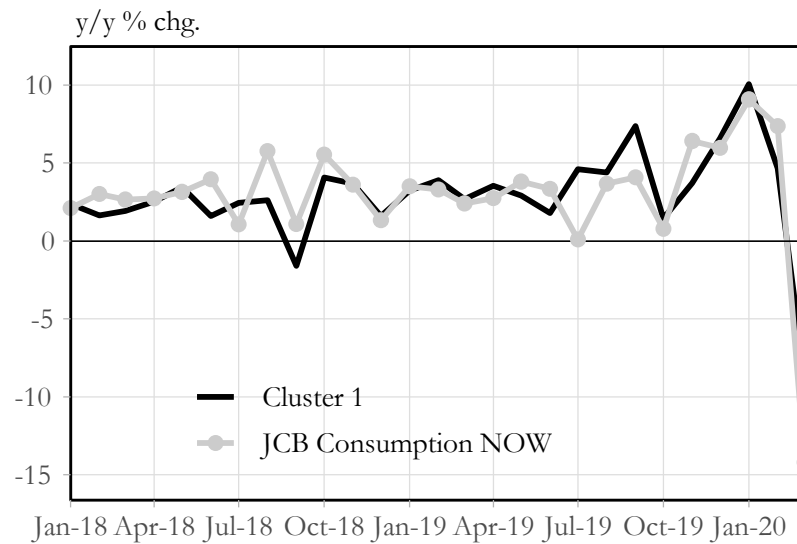


Figure 6: Indicator based on the hourly population for the food service industry

Note: "JCB Consumption NOW" indicates a consumption index for food services based on credit card transaction data.
 Sources: Agoop; Gurunavi; Ministry of Land, Infrastructure, Transport and Tourism; NOWCAST, Inc./ JCB, Co., Ltd., "JCB Consumption NOW."

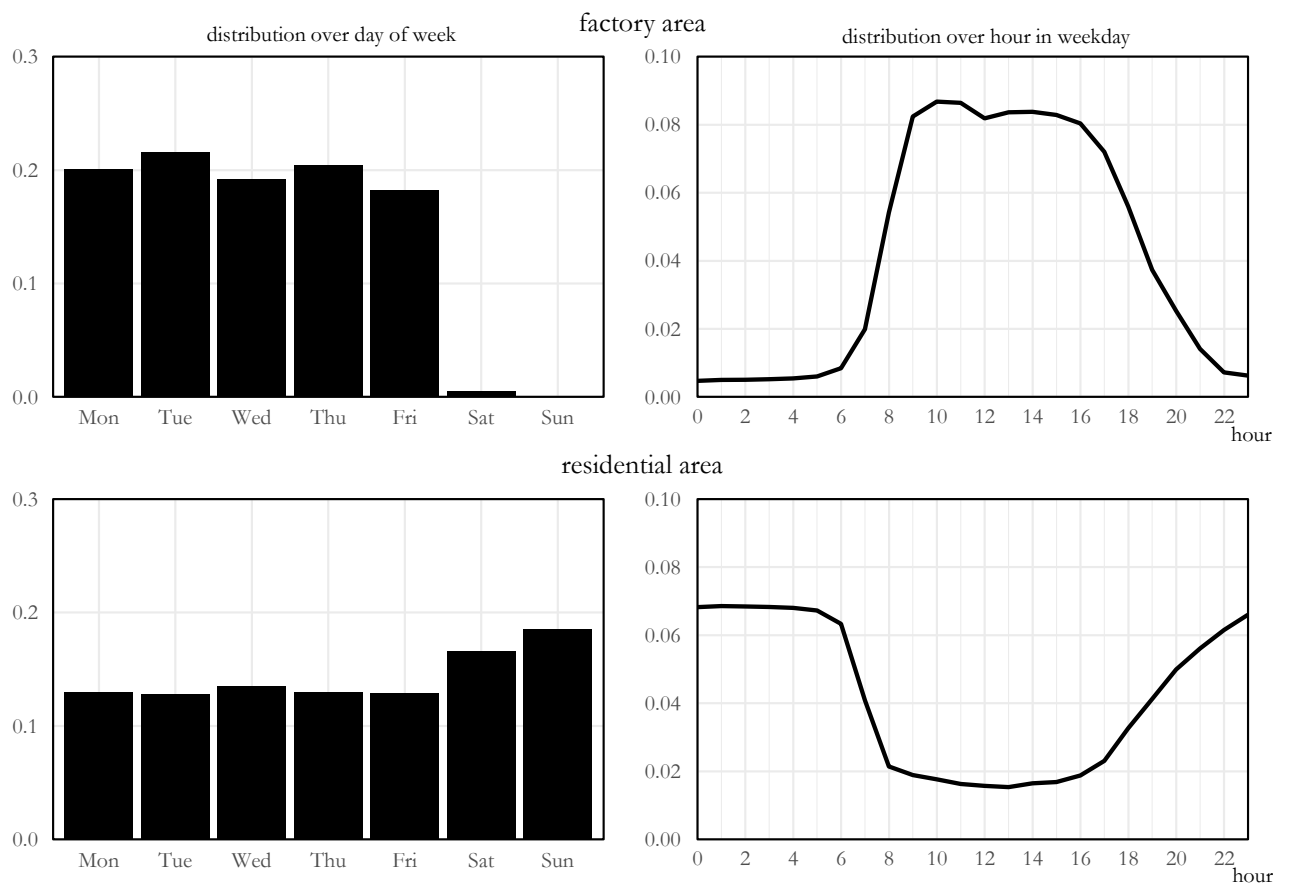


Figure 7: Hourly population in a typical factory area and residential area

Note: Figures show patterns of the hourly population in a typical factory area and residential area based on data on the weeks of which the previous and following weeks have five business days as well as those weeks themselves in order to eliminate the effect of public holidays. The left side panels show population by day of the week and the right side panels show average population by hour on weekdays.

Source: Agoop.

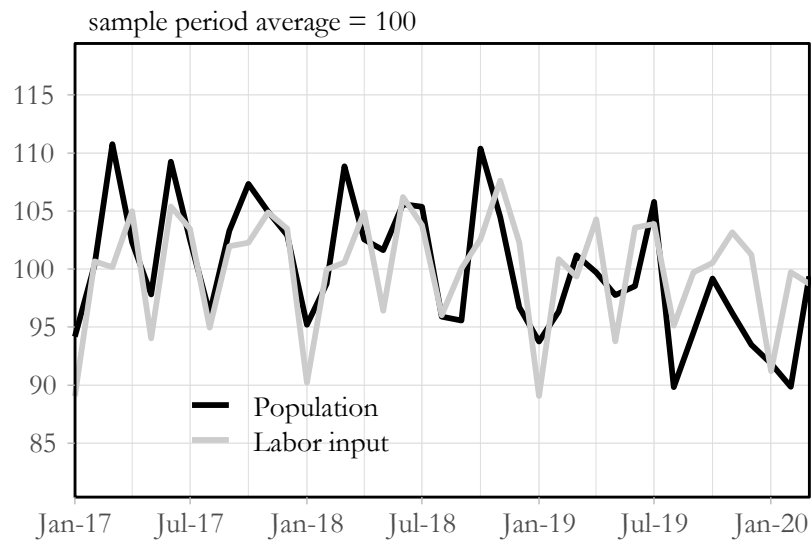


Figure 8: Population and labor input in the manufacturing industry

Sources: Agoop; Ministry of Health, Labour and Welfare.

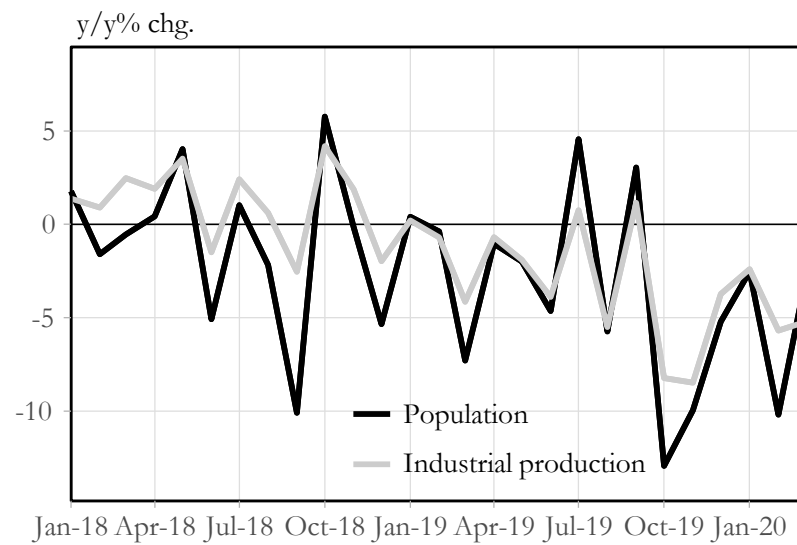


Figure 9: Indicator based on the hourly population for industrial production

Sources: Agoop; Ministry of Economy, Trade and Industry.

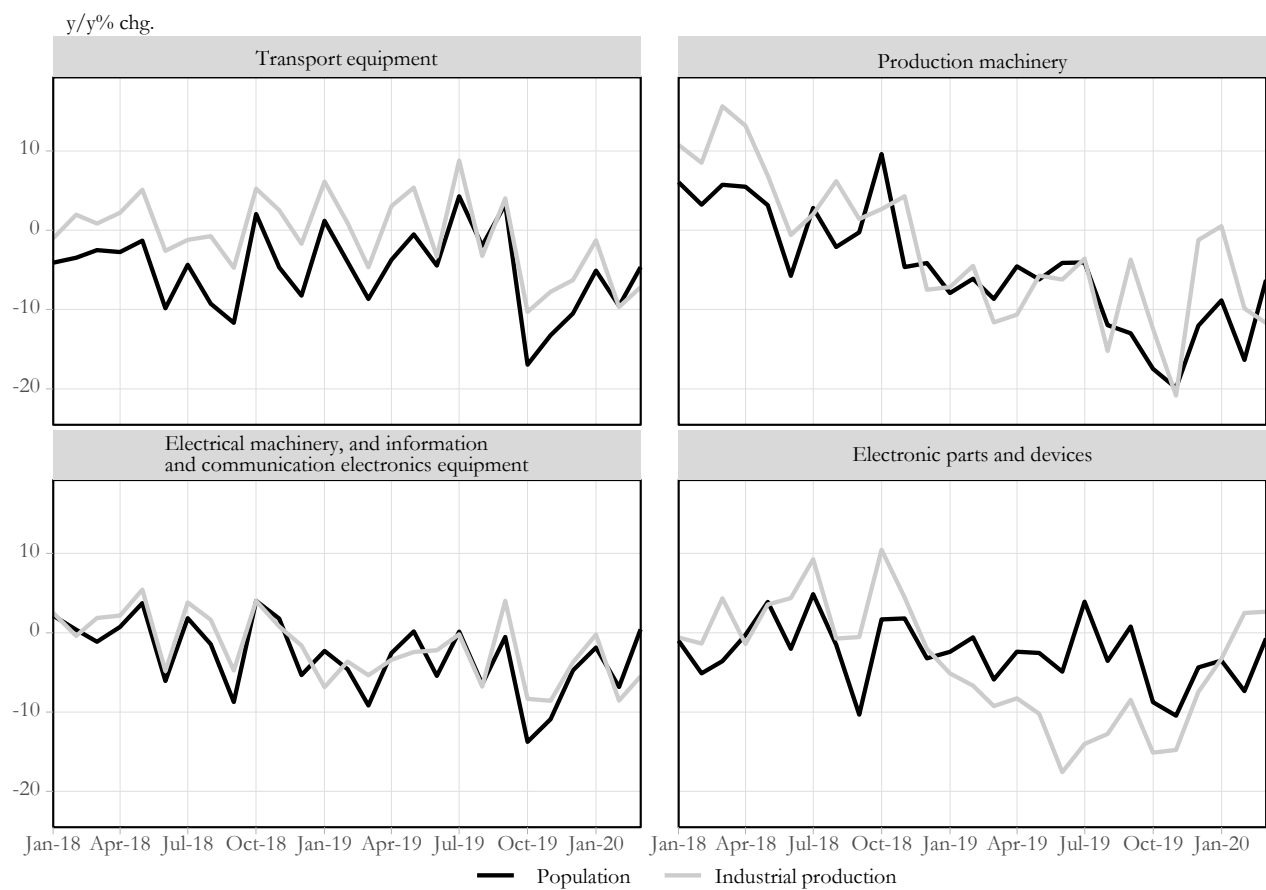


Figure 10: Production indicator based on the hourly population by industry

Sources: Agoop; Ministry of Economy, Trade and Industry.

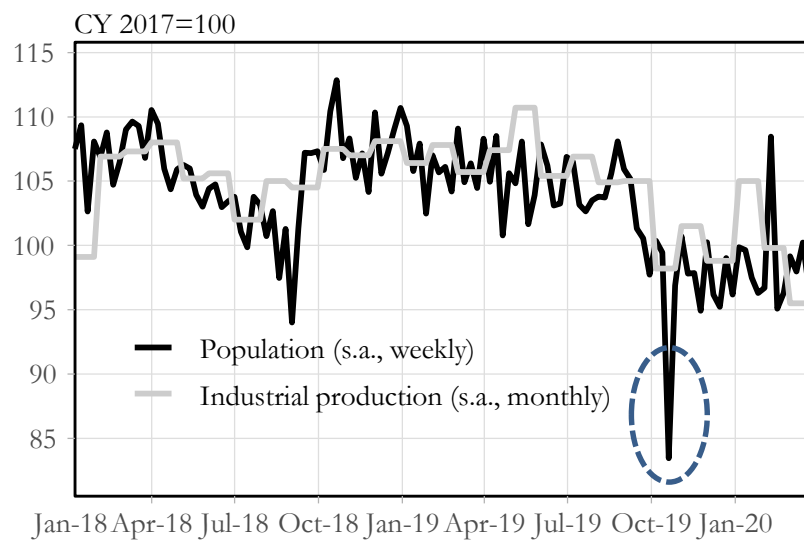


Figure 11: Weekly production indicator (transportation equipment)

Sources: Agoop; Ministry of Economy, Trade and Industry.

Table 1: Correlations between the EIG and consumption indexes for the food service industry

| | Before clustering | | Cluster 1 | |
|---------------|-------------------|-----------------|-----------------|-----------------|
| | up to Feb. 2020 | up to Mar. 2020 | up to Feb. 2020 | up to Mar. 2020 |
| Food services | 0.59 | 0.85 | 0.66 | 0.86 |
| Bars and pubs | 0.47 | 0.75 | 0.51 | 0.77 |

Notes: The table shows the correlations between the economic indicator from GPS data (EIG) for the food service industry and the corresponding indexes from JCB Consumption NOW based on credit card transaction data.

Sources: Agoop; Gurunavi; Ministry of Land, Infrastructure, Transport and Tourism; NOWCAST, Inc./ JCB, Co., Ltd., "JCB Consumption NOW" .

Table 2: Correlations between labor input and production

| Industry | Correlation |
|---|-------------|
| Production machinery | 0.75 |
| Transportation equipment | 0.72 |
| Fabricated metal products | 0.43 |
| Plastic products | 0.23 |
| Pulp, paper and paper products | 0.18 |
| Ceramic, stone and clay products | 0.15 |
| Electronic parts, devices and electronic circuits | 0.12 |
| Food, beverages, and tobacco | 0.05 |

Note: The table shows the correlation based on year-over-year changes in the period from January 2014 to December 2019. Labor input = Total hours worked × Regular employees (from the Monthly Labour Survey).

Sources: Ministry of Economy, Trade and Industry; Ministry of Health, Labour and Welfare.

Table 3: Criteria for Sunday and daytime ratios, and average population

| Industry | Sunday ratio upper threshold | Day time ratio lower threshold | Average population lower threshold | Correlation |
|---|---------------------------------|-----------------------------------|---------------------------------------|-------------|
| Transportation equipment | 0.5 | 0.6 | 10 | 0.86 |
| Production machinery | 0.1 | 2 | 40 | 0.69 |
| Information and communication electronics equipment | 0.9 | 1 | 80 | 0.68 |
| General-purpose machinery | 0.5 | 1.8 | 40 | 0.63 |
| Food, beverages and tobacco | 0.1 | 1 | 60 | 0.60 |
| Business oriented machinery | 0.5 | 2 | 10 | 0.54 |
| Non-ferrous metals and products | 0.1 | 1.4 | 20 | 0.49 |
| Electrical machinery, equipment and supplies | 0.1 | 0.8 | 10 | 0.42 |
| Electronic parts, devices and electronic circuits | 0.9 | 1.4 | 10 | 0.31 |
| Chemical and allied products | 0.9 | 0.6 | 10 | 0.20 |

Note: Labor input = Total hours worked \times Regular employees (from Monthly Labour Survey). Each ratio is calculated using mobility data on the weeks of which the previous and following weeks have five business days as well as those weeks themselves in order to eliminate the effect of public holidays. Sunday ratio = average population on Sunday / average population on weekdays. Daytime ratio = average population from 9:00 a.m. to 4:59 p.m. / average population from midnight to 4:59 a.m.

Sources: Agoop; Ministry of Health, Labour and Welfare.

Table 4: Criterion for holiday inclusion and hours

| Industry | Conditions | | | Correlation |
|---|-------------------|------------|----------|-------------|
| | Holidays included | Start hour | End hour | |
| Transportation equipment | yes | 17 | 18 | 0.85 |
| Production machinery | yes | 18 | 21 | 0.82 |
| Iron, steel and Non-ferrous metals | yes | 8 | 12 | 0.78 |
| Electrical machinery, and information and communication electronics equipment | no | 16 | 17 | 0.72 |
| Fabricated metals | no | 16 | 19 | 0.71 |
| Plastic products | yes | 17 | 18 | 0.69 |
| General-purpose and business oriented machinery | yes | 10 | 11 | 0.65 |
| Pulp, paper and paper products | no | 9 | 10 | 0.61 |
| Ceramics, stone and clay products | yes | 14 | 15 | 0.55 |
| Electronic parts and devices | yes | 8 | 9 | 0.36 |
| Chemicals | no | 16 | 17 | 0.35 |
| Foods and tobacco | no | 22 | 23 | 0.30 |

Note: "Start hour" and "end hour" indicate the start and end time of the selected sample window, respectively. For example, the row with a "start hour" of 18 and "end hour" of 21 means the data taken from 18:00 to 21:59.

Sources: Agoop; Ministry of Health, Labour and Welfare.

Nowcasting Economic Activity with Mobility Data*

Kohei Matsumura (Bank of Japan), Yusuke Oh (BoJ), Tomohiro Sugo (BoJ),
and Koji Takahashi (BIS)

IFC-Bank of Italy workshop

February, 2022

* The views expressed here are those of authors and do not necessarily
reflect those of the Bank of Japan or BIS.

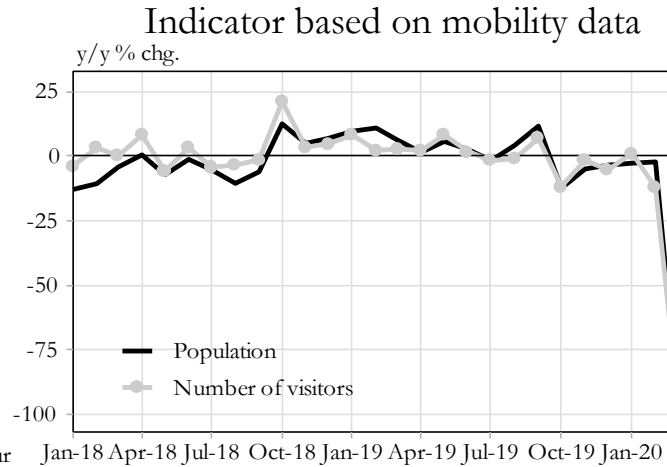
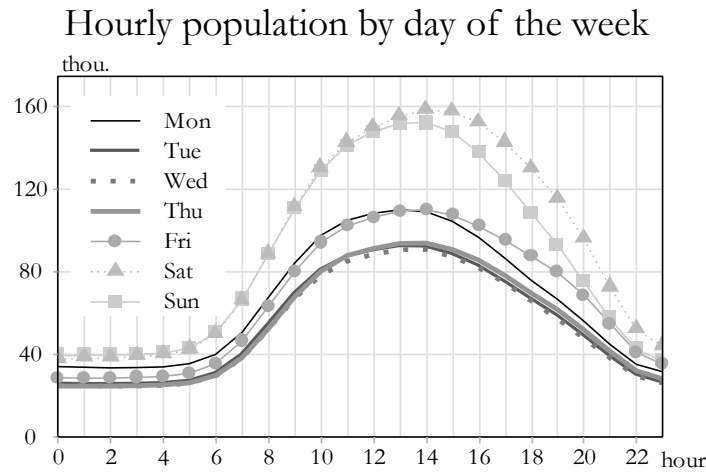
Data

- Collected by smart phone apps in Japan.
- The number of people who stayed in a mesh
- A mesh is defined as **100m × 100m** square.
 - ✓ 20 mil. Meshes in total.
- Hourly data from Jan. 2017 to Mar. 2020.
- Combine with other POI data
 - Economic Census
 - ✓ Factory addresses, number of employees, sales, etc.
 - National Land Numerical Information
 - ✓ Coordinates of stations and airports, and use of lands.
 - Information provided by private companies
 - ✓ Coordinates of facilities of interest (e.g. restaurants) based on their names and addresses.

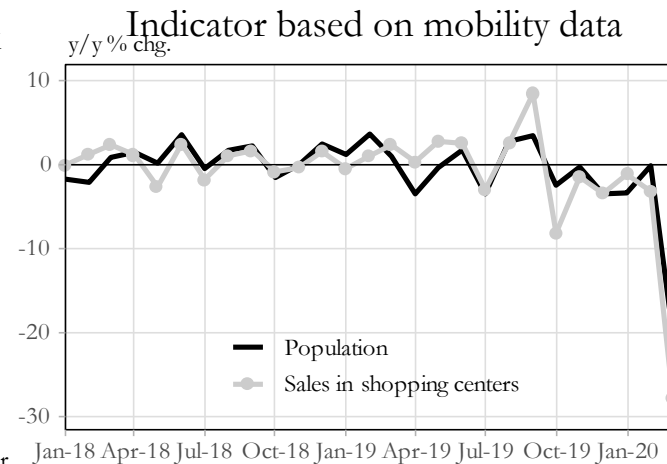
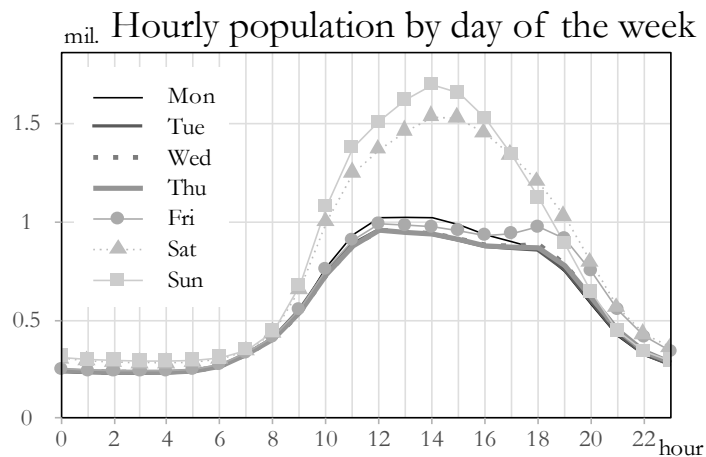
| Mesh ID | Year | month | day | hour | population |
|--------------------------|------|-------|-----|------|------------|
| XX | 2017 | 1 | 1 | 0 | 1 |
| XX | 2017 | 1 | 1 | 1 | 2 |
| XX | 2017 | 1 | 1 | 2 | 30 |
| 34 bil. records in total | | | | | |
| ZZ | 2020 | 3 | 31 | 20 | 1 |
| ZZ | 2020 | 3 | 31 | 23 | 10 |

Amusement parks and shopping malls

Amusement parks



Shopping malls



Restaurants

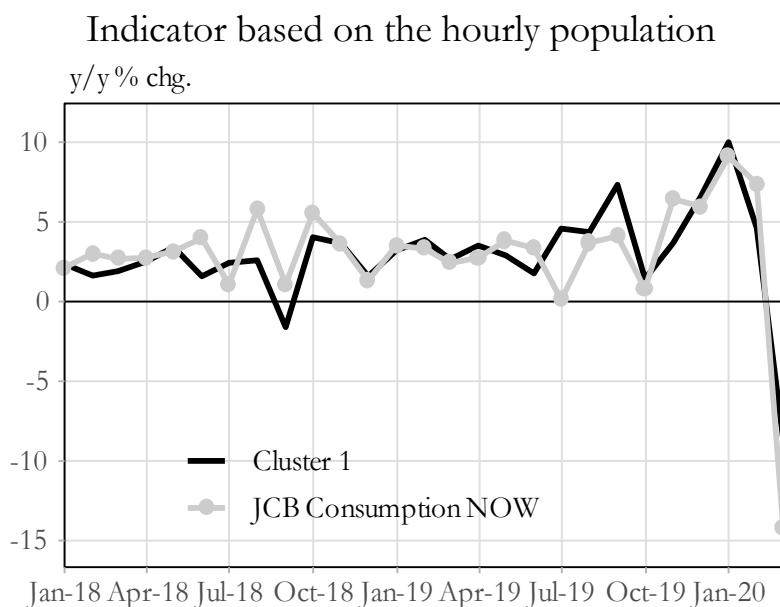
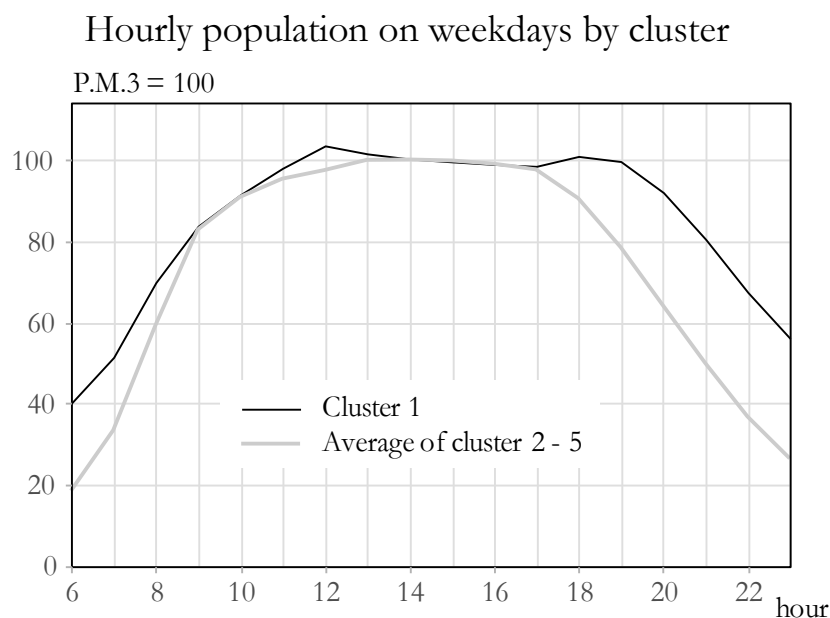
➤ Small size for each one but the number of restaurants is enormous.

➤ Located in areas where many various different activities are conducted.

1. Focus on non-residential areas based on the ratio of population in daytime

2. Use Grunavi (restaurant guide service) API and extract meshes that include more than 300 restaurants

3. Cluster the selected meshes into 5 groups, using k-medoids

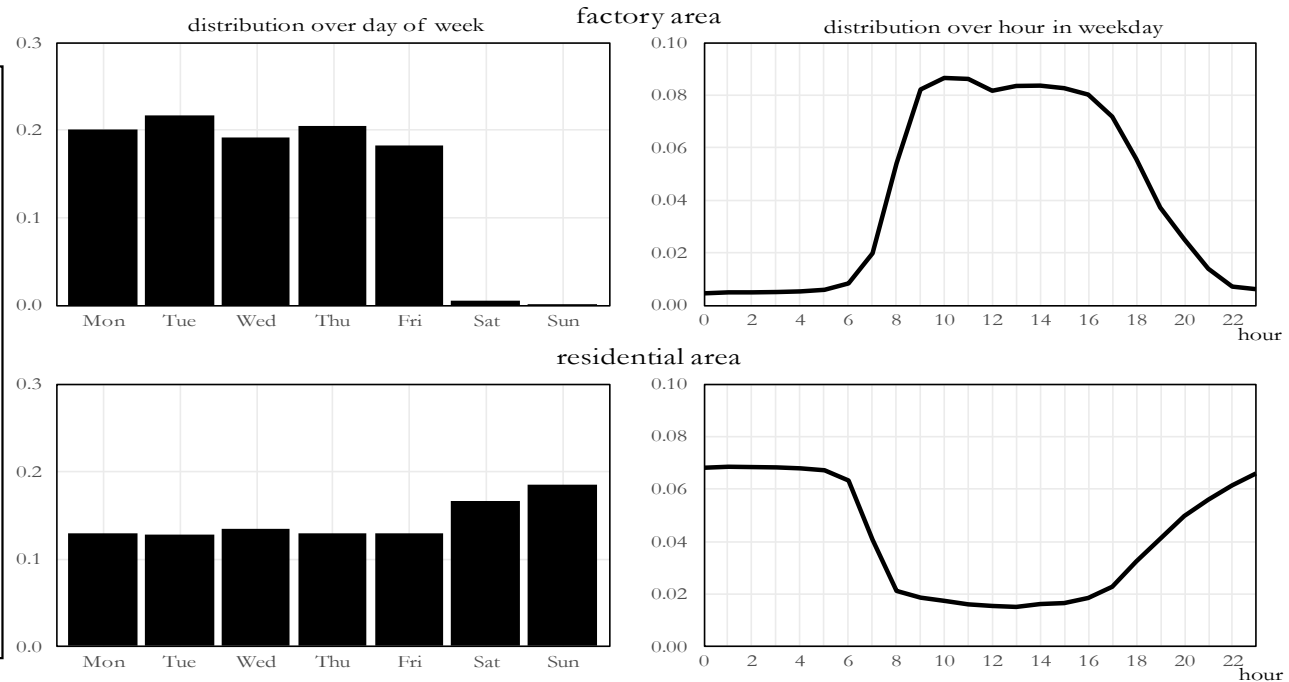


Sources: Agoop; Grunavi; Ministry of Land, Infrastructure, Transport and Tourism; NOWCAST, Inc./ JCB, Co., Ltd., "JCB Consumption NOW"

Production: Identifying meshes representing factories

1. Among factories listed in Economic Census for Business Activity whose area are above $10,000 \text{ m}^2$, select top 10,000 factories in terms of value-added.
2. We collect candidate meshes around the registered address by using information on its area.
3. Focus on several features (Sunday ratio, etc.) to remove unrelated ones.

For each industry, grid search three thresholds (day time ratio, Sunday ratio, and average population) such that the correlation between population and labor input (from official statistics) is highest.

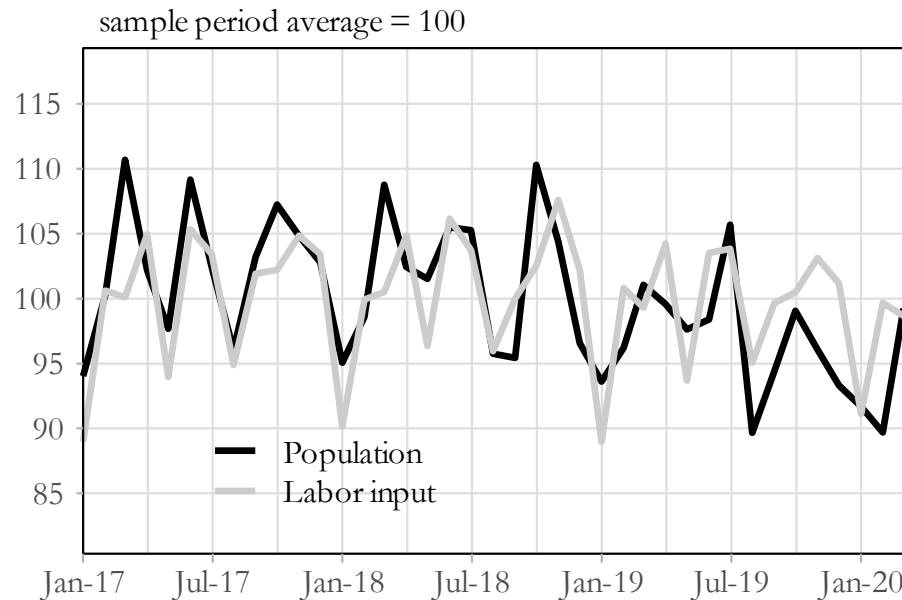


Mobility data and labor input

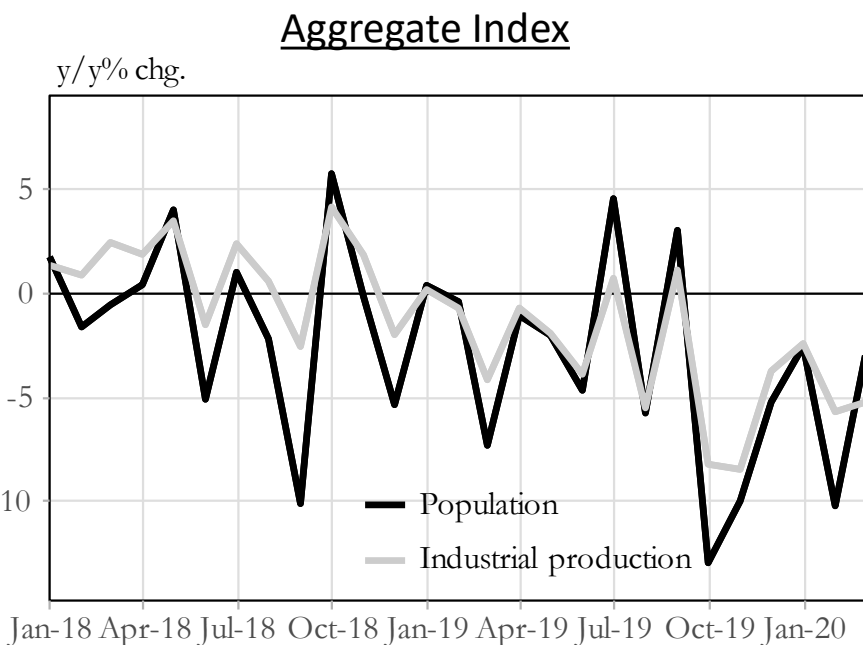
- Aggregating population over all hours and days does not necessarily approximate labor inputs well.

e.g. fluctuations can be driven by facility maintenance workers.

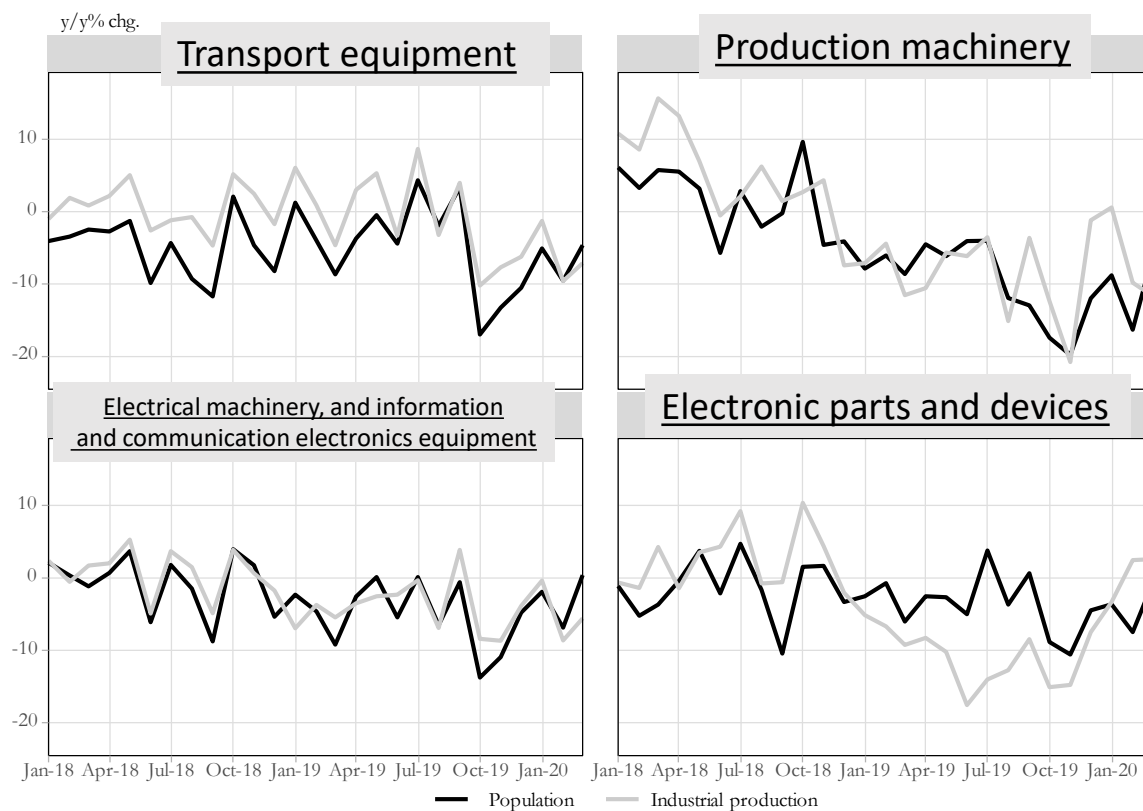
- Grid search three parameters (start and end time, and whether including weekends) to determine the best time window.



Mobility data and industrial production

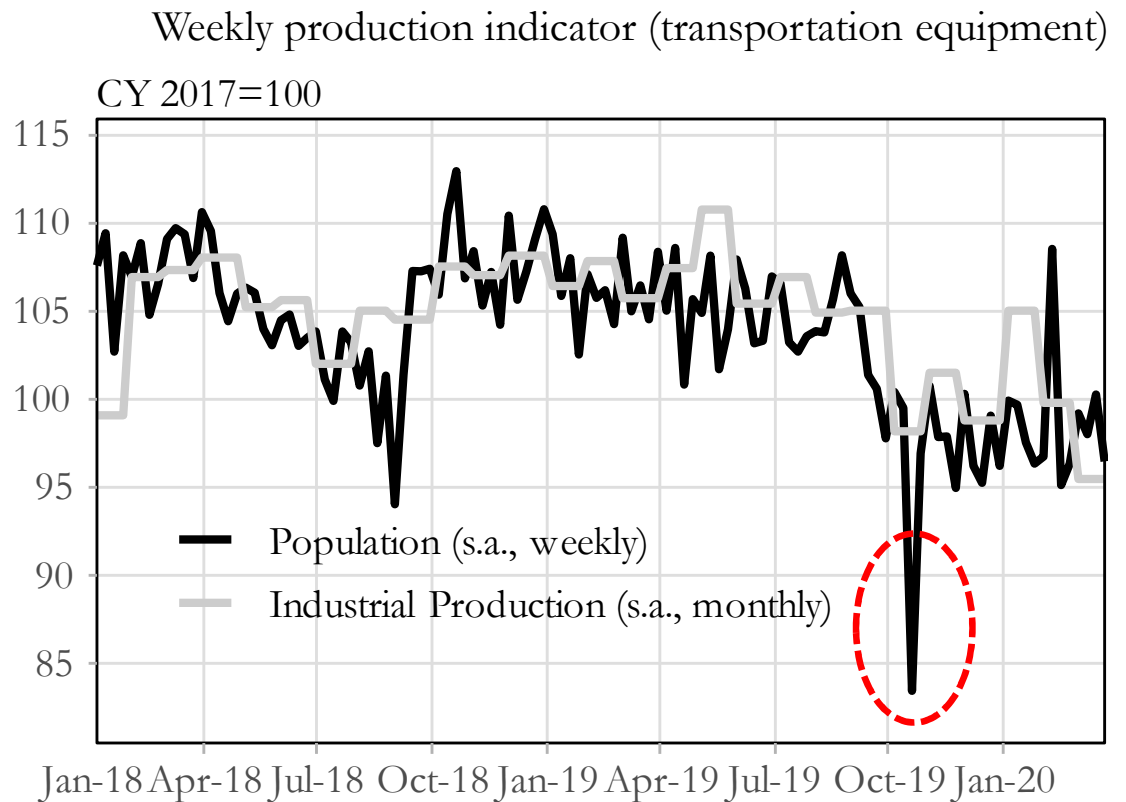


Sources: Agoop; Ministry of Economy, Trade and Industry.



Extension: High-frequency Analysis

- We construct weekly index by aggregating our indices at the weekly level and removing seasonal effects.
- ✓ The index capture the impact of Typhoon Hagibis on production in late Oct. 2019.



Sources: Agoop; Ministry of Economy, Trade and Industry.

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Extracting economic sentiment from news articles: the case of Korea¹

Younghwan Lee and Beomseok Seo,
Bank of Korea

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Extracting Economic Sentiment from News Articles: The Case of Korea

Younghwan Lee, Beomseok Seo

Abstract

In this study, I propose a News Sentiment Index as an approach to meet the growing need for timely economic statistics. Using a set of machine learning techniques, economic sentiments were extracted from news articles from 2005 to the present to construct this new index. The proposed index complements existing economic statistics in two ways. First, unlike many existing macroeconomic data, it can be available on a daily basis. Owing to recent advances in information technology, this index can be calculated quickly, whereas the calculation of traditional macroeconomic indices is time-consuming and costly. Second, the News Sentiment Index is a good predictor of important macro-variables; not only is it highly correlated with GDP, the Economic Sentiment Index, the Consumer Sentiment Index, and the Business Sentiment Index, but it also leads those indices by one to two months. Empirical evidence supports the hypothesis that news articles convey valuable information regarding economic prospects. The accompanying brief analysis of the 2020 COVID-19 crisis in South Korea proves the usefulness of the index.

Keywords: Big-data analysis, Economic sentiment, Sentiment index, Natural language processing, COVID-19

JEL classification: C45, E37

Table of Contents

| | |
|---|---|
| Extracting Economic Sentiment from News Articles: The Case of Korea | 1 |
| 1. Introduction | 2 |
| 2. Methodology | 3 |
| 2.1 Preprocessing | 4 |
| 2.2 Training | 4 |
| 3. Empirical Result | 5 |
| 4. Concluding Remarks | 7 |
| References | 8 |

1. Introduction

In this paper, a method for constructing an economic sentiment index using news articles is proposed, and the empirical validity of the new index is presented. In contrast to existing survey-based sentiment indexes such as the University of Michigan Consumer Sentiment Index (MCSI) of the US, and the Composite Consumer Sentiment Index (CCSI) of the Bank of Korea, the News Sentiment Index (NSI) uses web-scraped news articles as input data and applies a set of machine-learning techniques to extract economic sentiment from news articles. The two fundamental benefits of substituting survey data with news articles are timeliness and cost efficiency. To use survey data to measure economic sentiment, a large-scale survey must be conducted on a regular basis, which is a time-consuming and expensive task. Furthermore, the survey respondents have to be carefully selected to ensure the representativeness of the analysis and maintained for consistency of the index. On the other hand, a large volume of news articles focused on economic issues is created and stored on a daily basis and is easily accessible via the Internet. Streamlining of the entire process in this way allows us to access economic sentiment quickly with minimal effort.

Despite the apparent advantage of using text data to measure economic sentiment, it is difficult to apply in practice. The difficulties stem from two factors: the unobservability of sentiment and the unstructured nature of text data.

First, the unobservability of economic sentiment raises a question regarding what to measure. The idea of the NSI is to quantify the difference between the number of positive-sentiment sentences and the number of negative-sentiment sentences for a given time period. It is based on the assumption that economic sentiment and the sentiment revealed in economic news articles are highly correlated. Using Michigan Survey data Barsky and Sims (2012) show that consumer confidence reflects the prediction of the change in economic fundamentals. Thus, the assumption is justifiable as long as the economic news articles are based on facts. Specifically, for a given time period t , the proportion between the number of positive sentences and the number of negative sentences is quantified as follows:

$$X_t = \frac{N \text{ of Positive Sentences} - N \text{ of Negative Sentences}}{N \text{ of Positive Sentences} + N \text{ of Negative Sentences}} \quad (1)$$

The NSI at time t is defined as a scaled and translated version of X_t such that its long-term mean and standard deviation are equal to 100 and 10, respectively.

Second, because the input data are unstructured text data, a question arises as to how to measure the sentiment of each article sentence. A growing body of economic literature is concerned with incorporating text as an alternative data source. Shapiro et al. (2020) provide the closest example to this study. They proposed a method to extract the sentiment of news articles based on sentiment-labeled lexicons and grammatical rules, such as negation, to construct a new sentiment index using news articles. They show that the new index predicts existing sentiment indices such as the MCSI and Conference Board Consumer Confidence. The Economics Policy Uncertainty Index suggested by Baker et al. (2016) is another prominent example. They defined three groups of keywords related to economics, policy, and uncertainty. Their index is based on the proportion of articles that contain a predefined set of keywords.

In contrast to the aforementioned studies, the NSI is based on a machine-learning approach. Rather than using predefined rules, the sentiment classification rule is trained using data and algorithms. The advantage of this approach is its flexibility. The meaning of words is subject to change. For example, in economic news articles, the meaning of word "corona" has changed from a brand of beer to a name of virus that triggered severe economic slowdown. The machine learning approach allows for such changes to be updated via additional training.

A brief analysis provides empirical evidence that news articles convey valuable information regarding economic prospects. The NSI is a good predictor of important macroeconomic variables; not only is it highly correlated with the GDP, CCSI, Business Survey Index (BSI) and Economic Sentiment Index (ESI), but it also leads those indices by one to two months.

The rest of this paper is organized as follows. In Section 2, the methodology for obtaining the sentiment classifier is presented. Section 3 presents a brief analysis of the empirical validity of the NSI. Finally, Section 4 concludes the paper.

2. Methodology

The sentiment classifier is an essential component of the entire process undertaken in this study. Mathematically, it is a function f that takes a sentence as an input and returns the corresponding sentiment label. That is,

$$f : S \rightarrow \{Positive, Negative, Neutral\} \quad (2)$$

Here, S is the set of article sentences. The question is how to properly approximate the function f . To make it tractable, the classifier function f is deconstructed into two parts: preprocessing and training. The first part, preprocessing, is a step that converts article sentences into sequences of tokens so that they can be used as input for the training step. That is, the preprocessing is a function h such that

$$h : S \rightarrow V^M,$$

where V is the set of tokens, and M is the maximum length of the sequences. All sequences shorter than M are padded with null tokens until their length is equal to M .

The second part, the training, is to find a function that takes a sequence representation of a sentence as input and returns its corresponding sentiment label. Let $g(\cdot|\theta) : V^M \rightarrow \{Positive, Negative, Neutral\}$ be a classifier function parameterized by θ . Assume that it is a true classifier if $\theta = \theta_0$. Equation (2) can then be rewritten as follows:

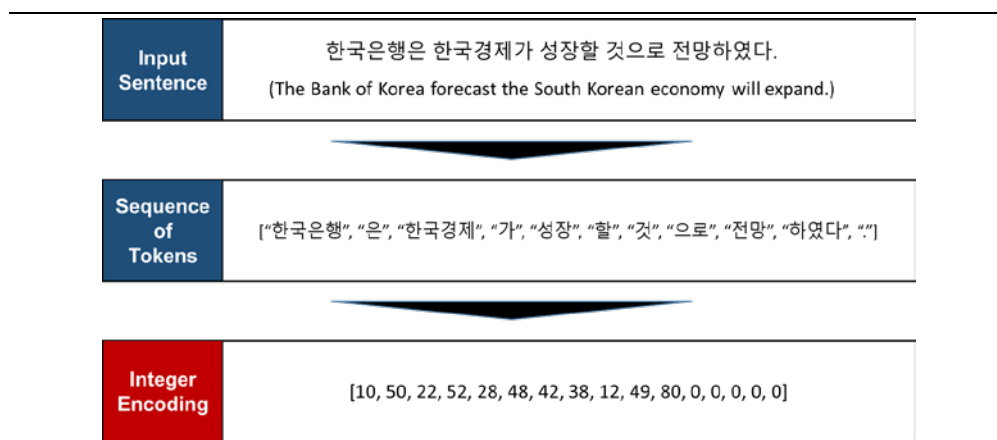
$$f(s_i|\theta_0) = g(h(s_i)|\theta_0),$$

where $s_i \in S$. Thus, given the preprocessor h , the remaining task is to find a proper proxy $\hat{\theta}$ of θ_0 to determine an approximation of f . It is performed by supervised learning with the transformer model proposed by Vaswani et al. (2017) and 450,000 lines of a human operator-labeled training set.

The obtained classifier is applied to categorize the sentiments of the input sentences. Based on the categorization result, the counts of positive and negative sentences are aggregated to calculate and update the NSI. The details are discussed in Section 3.

Example of Preprocessing¹

Figure 1.



¹ In this example, integer encoding is assumed in which the maximum length of a sequence is equal to 16 and the null token is represented by 0.

Source: Bank of Korea

2.1 Preprocessing

The objective of preprocessing is to convert the article sentences into token sequences. It begins with the tokenization of the target text. Tokenization is the process of separating text into smaller units called tokens, which is an atomic unit of analysis. A token can be a word, morpheme, or character. For example, the sentence "It is a pen" can be tokenized into a list of elements such as ["it", "is", "a", "pen"] or ["i", "t", "i", "s", "a", "p", "e", "n"]. The algorithm and the granularity of tokenization must be determined depending on the goal of the analysis¹. In this study, tokenization is based on morphemes to account for the nature of the Korean language.

Next, each token is converted into a numeric type, such as an integer value or a one-hot vector. That is, V is not a set of labels of tokens themselves but, rather, a numerical version of it. This step is necessary because the transformer algorithm does not operate directly on label data. Figure 1. summarizes the process.

2.2 Training

The transformer model (Vawani et al., 2017) was used to approximate the second component of the sentiment classifier $g(\cdot|\theta)$. This model has distinctive advantages compared to existing alternatives. First, it improves the accuracy of the support vector machine (SVM) in the sense that it takes positional and contextual information into account. The meaning of a sentence, a sequence of tokens, is more than the sum of the meanings of individual tokens. Rather, the meaning of a sentence is dependent on the order of words as well as on the other words included in the sentence. The transformer model considers positional information and the dependency between tokens to embed input sentences into a semantic space.

¹ For a detailed discussion, see Webster and Kit (1992).

Performance of the Classifier¹

Table 1.

| Predicted label | Human operator label | |
|-----------------|----------------------|----------|
| | Positive | Negative |
| | Positive | 0.95 |
| | Negative | 0.03 |
| | | 0.97 |

¹ To evaluate the accuracy of the classifier when it is used to calculate the NSI, the case where either l_i or \hat{l}_i is neutral is excluded from the test.

Source: Bank of Korea

Second, the transformer model improves the speed of the recurrent neural network (RNN) model because it allows for more parallelization in the computation process. The RNN model sequentially takes each token in a sequence and embeds it into a semantic space. The sequential nature of the model allows it to preserve the positional information of a sentence at the cost of computation time. However, the transformer model does not sequentially process the input sequence to incorporate the positional information of tests. Rather, it relies on an attention mechanism that can be easily parallelized. This characteristic enables the transformer model to exploit parallel computing power to speed up training.

Approximately 450,000 labeled instances were used to train the transformer model. Each labeled instance is the sentence-sentiment label pair $(s_i, l_i) \in S \times \{Positive, Negative, Neutral\}$ where each sentiment label l_i is manually labeled by human operators. The training sentences were randomly selected from economic news articles from 2005 to 2021. Based on the training set and the transformer model, the proxy $\hat{\theta}$ of θ_0 is calibrated. The composition of $g(\cdot)$ and $h(\cdot | \hat{\theta})$ results in the classifier $\hat{f}(\cdot)$.

To test the validity of the classifier, an additional 5,000 sentences were randomly sampled and labeled by human operators to construct the validation set. Table 1 summarizes the performance of the classifiers. Each entry is the corresponding conditional probability $\Pr[\hat{l}_i | l_i]$ where $\hat{l}_i = \hat{f}(s_i)$. Overall, it showed an acceptable level of reliability.

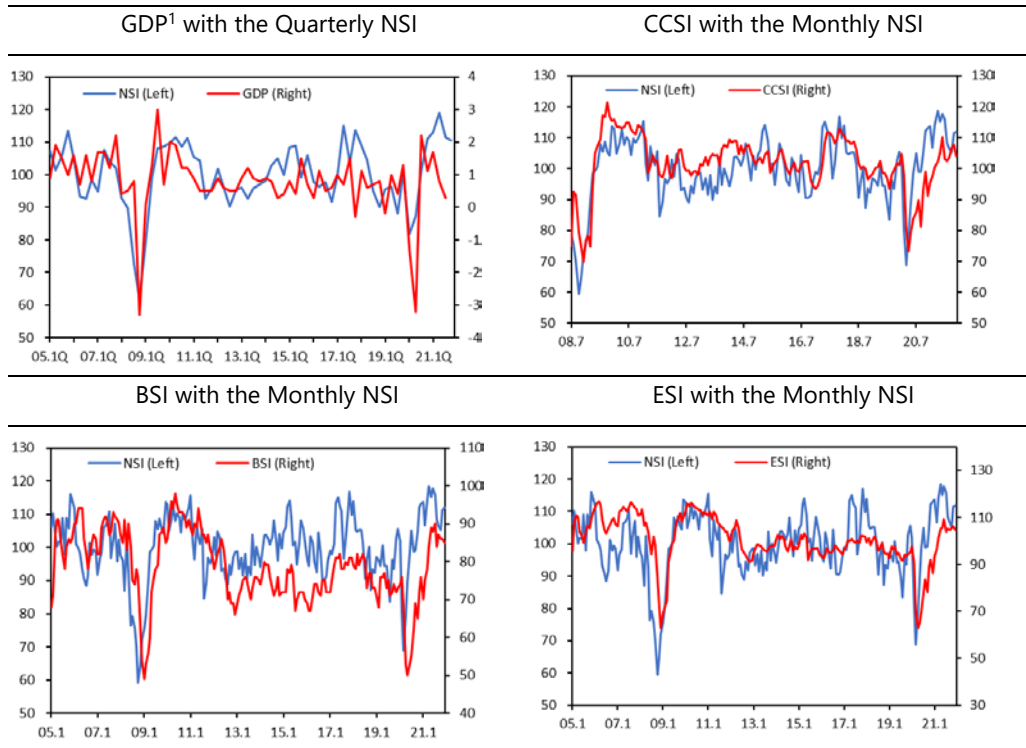
3. Empirical Result

To construct the target dataset for the NSI, all news articles accessible on the economic section of a web portal were crawled on a daily basis. It covers 3,500 articles from 50 newspapers a day on average as of 2021. After separating articles into sentences, 10,000 were randomly selected as the target sample. The time coverage of the dataset is from 2005 to the present, which includes several major events such as the financial crisis and COVID-19.

The classifier was applied to categorize the target sentences. For each day, the number of positive and negative sentences was counted. Using this sentiment count, the NSI can be calculated for an arbitrary period. For example, by applying equation (1) to the monthly and quarterly sum, the monthly and quarterly NSI is calculated after standardization. On the other hand, the daily sentiment NSI is based on the sum of the past seven days to remove the day of the week effect.

Macro-Variables with the NSI

Figure 2.



¹ GDP here is the seasonally adjusted real GDP growth rate.

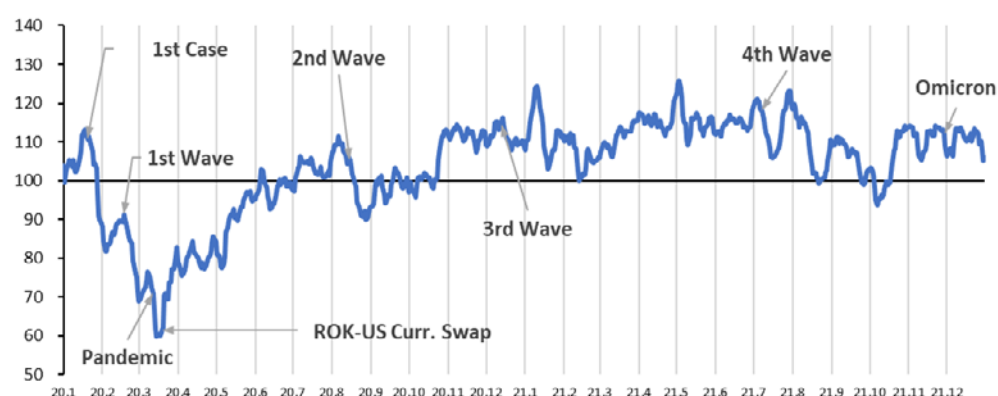
Source: Bank of Korea

To examine the empirical validity of the NSI, the correlations between the NSI and four macro-variables were analyzed: Gross Domestic Product (GDP), Composite Consumer Sentiment Index (CCSI), Future Business Condition Business Survey Index (BSI), and Economic Sentiment Index (ESI). Figure 2 shows a plot of the NSI and the macro-variables. The NSI closely follows all of these variables. The findings are twofold. First, GDP and the quarterly NSI are highly correlated. The correlation coefficient of the two variables was 0.55. This implies that news articles convey substantial information regarding economic fundamentals.

Second, the lag correlation analysis between the monthly NSI and sentiment-related macro-variables² suggests that the NSI can help econometricians predict changes in economic sentiment. The results of the analysis show that the NSI leads CCSI by one month, BSI by two months, and ESI one month, where the corresponding maximum correlation coefficients are 0.75, 0.61, and 0.61, respectively.

In Figure 3, the daily NSI and the dates of several important COVID-19-related events are indicated. This shows that the timeliness of the index helps identify the impact of specific events on economic sentiment. For example, after the first confirmed case was reported in the Republic of Korea, the NSI dramatically decreased until the Bank of Korea (BOK) and the US Federal Reserve signed a currency swap contract. After the contract, the NSI improved for five months until the second wave came.

² CCSI, BSI, and ESI.



¹ The daily NSI here is constructed based on the news articles of the previous seven days.

Source: Bank of Korea

4. Concluding Remarks

In this paper, a methodology to construct a sentiment index based on internet news data was proposed, and it was found that the new index (NSI) predicts the existing macro-variables, including sentiment indices. The new index allows policymakers to access current economic sentiment in a timely manner with little marginal effort because it is not based on a survey, which is costly to conduct.

Understanding the topics that drive economic sentiment remains a topic for future study. Refining the keywords exposed in the positively or negatively classified sentences will allow us to keep up with current events that drive current economic sentiment.

References

- Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics* 131 (4), 1593-1636.
- Barsky, R. B. and E. R. Sims (2012). Information, animal spirits, and the meaning of innovations in consumer confidence. *American Economic Review* 102 (4), 1343-77
- Shapiro, A. H., M. Sudhof, and D. J. Wilson (2020). Measuring news sentiment. *Journal of Econometrics*.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998-6008
- Webster, J. J. and C. Kit (1992). Tokenization as the initial phase in NLP. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*.

Short Bio

Younghwan Lee is an economist with the Bank of Korea, which joined in September 2019. He previously worked as an assistant research professor at Seoul National University. He earned his PhD in economics from Seoul National University and an undergraduate degree in Management Science from the Korea Advanced Institute of Science and Technology. His research interests lie in the fields of computational economics and finance.

Beomseok Seo is an economist in the Bank of Korea, where he joined in January 2011. He earned his PhD in statistics from Pennsylvania State University researching interpretable machine learning and bachelor's degree in economics and statistics from Korea University. His research interests lie primarily in the fields of statistical modeling and machine learning for the better human interpretation.

Extracting Economic Sentiment from News Articles: The Case of Korea

Younghwan Lee, Beomseok Seo

Economic Statistics Department,
Bank of Korea

February 10, 2022

Motivation

- Currently, economic conditions change very quickly.
- A timely assessment of economic conditions is especially valuable when the market is highly volatile and uncertain.
- Does official statistics fast enough? Think about survey based statistics such as Michigan Consumer Sentiment Index (MCSI).
- Collecting relevant data is time consuming and costly.
- On the other hand, tons of data is produced and stored at every second. Internet news articles data is one of the examples.
- This study proposes the ***News Sentiment Index (NSI)*** which is a timely available and cost-efficient measurement of economic sentiment.

The Idea

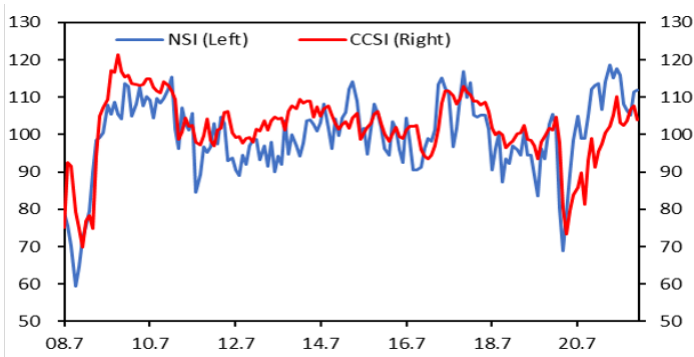
- Randomly sample 10,000 article sentences on a daily basis and classify their sentiments into three categories: positive, negative, and neutral.
- Counting period can be arbitrarily chosen.
- Quantify the number of positive and negative sentences of news articles as follow:

$$X_t = \frac{\# \text{ of pos. sentences} - \# \text{ of neg. sentences}}{\# \text{ of pos. sentences} + \# \text{ of neg. sentences}}$$

- Translate and scale X_t to make it looks like an index.
(Mean = 100, Std. = 10)

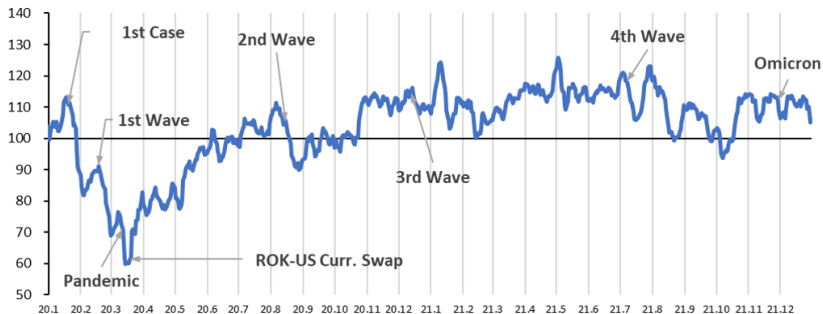
CCSI and NSI

- Composite Consumer Sentiment Index (CCSI) is a survey based consumer sentiment index in Korea.
- The monthly NSI lead it by 1 month and correlation is 0.75.



COVID-19 and NSI

- The daily NSI is able to react quickly to changes in economic conditions such as COVID-19 events.



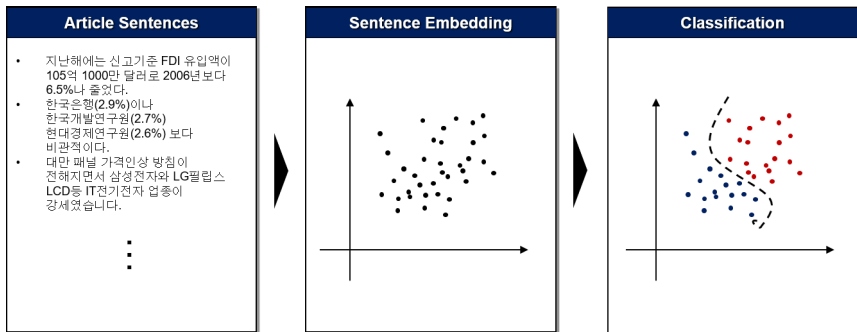
Methodology: Overview

- Supervised learning approach:
 - Input: 450,000 human operator labelled sentence-sentiment pairs.
 - Model: Transformer model
- Different from lexical approach from Shapiro et al.(2020), this approach is:
 - i) does not require pre-defined rules.
 - ii) change in meaning can be updated by additional training.
(Ex. Corona vs Corona)
- Work flow:



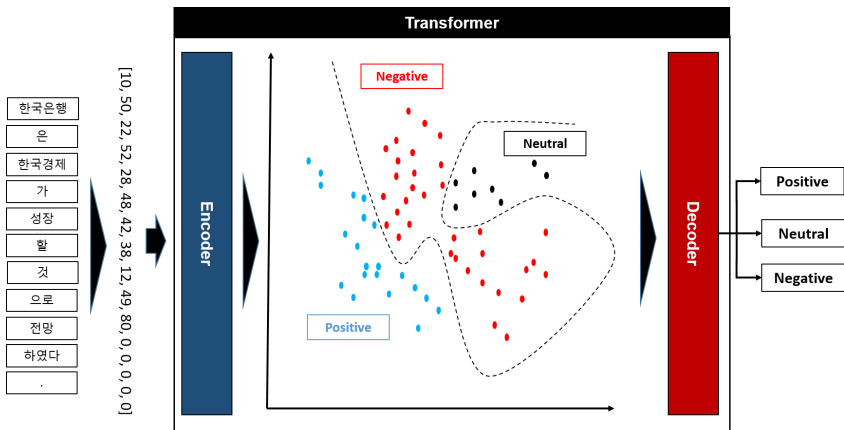
Methodology: Intuition

- We cannot directly apply classification models to the text data.
- The idea is to embed sentences into the ***semantic space*** that preserves senses of sentences.



Methodology: Transformer Model

- Vaswani et al. (2017) have proposed the transformer model.
- It takes positional information into account while faster than RNN.



Concluding Remarks

Implication:

- By using big-data, we can complement existing official statistics in terms of cost and time.
- Unstructured data such as text can help us to make better policy decision

Future work:

- What is the driving force of the change in economic sentiment?

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Big data analytics on real time gross settlement data for tracking corporate activity¹

Mohammad Khoyrul Hidayat, Amin Endah Sulistiawati and Alvin Andhika Zulen,
Bank Indonesia

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Big Data Analytics on RTGS Data for Tracking Corporate Activity

Amin Endah Sulistiawati¹, Mohammad Khoyrul Hidayat², Alvin Andhika Zulen³

Abstract

Corporate sector performance is one of the main indicators of financial system stability. Nevertheless, the available information is limited to corporate financial reports which are published a quarterly period with data lag of up to 2 months. The rapid evolution of technology is bringing powerful tools to bear with big data on payment system. This paper explores a new approach for measuring corporate sector activity using Bank Indonesia – Real Time Gross Settlement (BI-RTGS) transaction data by utilizing Big Data Analytics. We identify payment transaction activities (inflow and outflow) via BI-RTGS of more than 100 corporations (with the largest sales) from 9 sectors. We construct a new indicator for tracking corporate sector activity that can be produced on a monthly basis within few days after the reference period. This indicator also has good correlation with sectoral GDP data (for several sectors), which indicates the usefulness of the indicator for tracking economic activity.

Keywords: corporation, payment transaction, RTGS, big data analytics

Introduction

Financial stability is not easy to measure or define as it has complex interconnection of different elements of financial system. The analysis should examine the macroprudential aspects more closely, with more emphasis on the link between the real economy and the financial system. Corporate and household sectors, as well as macroeconomic developments, should be more closely integrated into credit risk assessments of the banking sector (Costeiu Neagu, 2013). By the time, such interactions among financial systems dimensions becomes more complex. The presence of excessive volatility, stress, or crises make financial system characteristics unstable.

Non financial corporations constitute a majority of banking sector's loan borrowers. Corporates' credit quality, loan standards, and lending conditions have significant influence in maintaining healthy economic activity. Fragilities that may

¹ Data Scientist at Digital Data & Big Data Analytics Development Division - Statistics Department, Bank Indonesia

² Data Scientist at Digital Data & Big Data Analytics Development Division - Statistics Department, Bank Indonesia

³ Analyst at Digital Data & Big Data Analytics Development Division - Statistics Department, Bank Indonesia

arise from corporate sector may leap into the banking sector by causing vulnerabilities and systemic risks (Güngör et al., 2016).

Corporate sector performance is one of the main indicators of financial system stability. Based on Bank Indonesia's Financial Stability Review report on 2021, stronger corporate sector performance had a knock-on effect of increasing household income and consumption. However, the available information related to corporate sector performance is limited to financial reports for listed companies. The reports are published in quarterly period with data lag of up to 2 months. Whereas, Bank Indonesia has weekly and monthly Board of Governor meetings. Hence, the need for a more timely indicator for tracking corporate sector activity is becoming more relevant.

On the other hand, the revolution of Big Data in payment systems offers richer data provision and techniques to understand and analyse financial system stability. This paper explores a new approach for measuring corporate sector activity using Bank Indonesia - Real Time Gross Settlement (BI-RTGS) transaction data by utilizing Big Data Analytics methodology. We identify payment transaction activities (inflow & outflow) via BI RTGS of more than 100 corporations from 9 sectors. The selected corporations are chosen to represent the largest share in sales for each sector.

Literature Reviews

For a long time, to have access on the state of the economy, policymakers and the private sector have relied on data released by official statistical institutions with usually a lag of several months or years. However, the emergence of technology changes the world and the way it is measured. Based on a recent survey conducted among the members of the Irving Fischer Committee (IFC), the majority of central banks discuss the topic of big data that is used with machine learning applications in a variety of areas, including research, monetary policy and financial stability as well as for supervision and regulation (suptech and regtech applications). Whereas, the quality of data, sampling and representativeness are major challenges for central banks, and so is legal uncertainty around data privacy and confidentiality (Doerr, 2021).

The new era of advanced data analytics is starting to be adopted by the financial sector. Deep learning models nowadays are considered into financial technology stacks. (Gensler and Bailey, 2020) used deep learning to explore financial sector and predict financial stability. By exploring 9 characteristics of deep learning (features of hyper-dimensionality, non-linearity, non-determinism, dynamism, and complexity; challenges of explainability, bias, and robustness; and an insatiable hunger for data), it marks a fundamental discontinuity that creates significant opportunities to enhance efficiency, financial inclusion, and risk mitigation. However, extensive adoption of deep learning may also increase uniformity, interconnectedness, and regulatory gaps, which could remain the financial system more fragile.

Another research using big data to analyse the complexity on financial stability is conducted. (Mertzanis, 2018) said that analysis on financial stability and policy should focus on accurate price discovery for complex instruments, realistic financial information generation processes, and system-wide risk materializing within complex financial networks. Thus, complexity analysis can make a useful contribution. In the

past year, the development on standardization on the global legal entity identifier system is made to uniquely identify parties to financial transactions across the globe. This is the first step towards a strong, flexible and adaptable global data infrastructure conducive to financial stability policy.

BI-RTGS Data

Bank Indonesia - Real Time Gross Settlement (BI-RTGS) is an electronic funds transfer system, amongst participants, in rupiah with real-time settlement on an individual transaction basis. BI-RTGS has played an important role in processing payment transactions, including High Value Payment System (HVPS) transactions exceeding Rp100 millions as well as urgent transactions. Data source used in this study is BI-RTGS Transactional Data. We use transaction type code (TTC) 100 and 101 also MT 103 which represents bank's transaction of ordering/beneficiary customer. These type of transaction has approximately 800.000 transactions/month (out of 1,2 millions transaction in total). Since BI-RTGS is in payment infrastructure of Bank Indonesia, we can capture the data on the following day of the settlement time.

Table 1 describes the data structure, which contains Settlement Time, Transaction Type Code, Sender/Receiver Bank Code, Nominal Transaction And Transaction Message (Block 4). While the first 5 fields are structured information, the last one, is in the form of unstructured text. However, Transaction Message field is the fundamental part of this study as it has substance essence information needed.

BI-RTGS Structure Data

Table 1

| Settlement Time | Transaction Type Code (TTC) | Sender Bank | Receiver Bank | Nominal | Block 4 |
|--|---|---|--|---|--|
| Transaction settlement time (MM/DD/YYYY HH:mm) | RTGS transaction code 100 : Transaction o/b of customer 101 : Transaction o/b of customer (w/o account) 111 : Forex Buy/Sell 112 : Interbank Money Market Etc. | Sender's Bank SWIFT Code (BIC) e.g.: CEINAJDJA | Receiver's Bank SWIFT Code (BIC) e.g.: BMRIIDJA | Transaction amount in Indonesian Rupiah | Transaction message, sent by Sender Bank into BI-RTGS system |

Transaction Message (Block 4) on data sample, shown in Table 2, is a free text of transaction message filled by sender bank into the system. We highlight on the line text preceded by code 50K for ordering customer information, 59 for beneficiary information and 70K for remittance information. Customer or beneficiary information may consist of account number, corporation name and/or address. Each transaction can have various ways of writing Blok 4 as we do not specify the standard. Hence, to extract this information is another challenge that we will discuss on the next chapter.

BI-RTGS Data Sample

Table 2

| Settlement Time | TTC | Sender Bank | Receiver Bank | Nominal | Block 4 |
|-----------------|-----|-------------|---------------|------------------|---|
| 5/2/2018 13:35 | 100 | DXXXIDJA | MXXXIDJA | 4,000,000,000.00 | :20:02RE201804304660#:23B:CRED#:23E:SDVA#:32A:180502IDR4000000000,00# :50K:PAM LYONNAISE JAYA,PT#JL. JEND GATOT SUBROTO KAV 51-52#10270 JAKARTA PUSAT#INDONESIA#:52D:BANK DXXX JAKARTA INDONESIA# :53A:/D/521067000990#DXXXIDJA#:57A:/C/520008000990#MXXXIDJA# :59:/1210005182096#PERSONEL ALIH DAYA, PT# :70:INV. 1804-0167#R/LOCAL# :71A:OUR#:72:/CODTYPTR/100#/CLRC/0670304# :77B:/FEAB/R /PTR/LOCAL |
| 5/2/2018 13:36 | 100 | DXXXIDJA | NXXXIDJA | 900,000,000.00 | :20:02RE201805020573#:23B:CRED#:23E:SDVA#:32A:180502IDR9000000000,00# :50K:HANJAYA MANDALA SAMPOERNA TBK.PT#RUNKUT INDUSTRI RAYA# NO.18#INDONESIA#:52D:BANK DXXX JAKARTA INDONESIA# :53A:/D/521067000990#DXXXIDJA#:57A:/C/520009000990#NXXXIDJA# :59:/10533549#ASTRA GRAPHIA TBK PT# :70:701625825S,BLACK AND WHITE PRINTED#PAGE#R/LOCAL#: :71A:OUR#:72:/CODTYPTR/100#/CLRC/0670304# :77B:/FEAB/R /PTR/LOCAL |

50K : Ordering Customer

59 : Beneficiary Information

70 : Remittance Information

Big Data Analytics Methodology

Big data analytics methodology is used in this study, as we will deal with large amount of payment transaction from BI-RTGS. Not only the volume that is large, but also the velocity is semi-real time. Hereafter, the variety information will be extracted from structure (text and number) and unstructured (free text). Transactional data from BI-RTGS is streamed from payment system infrastructure into Hadoop, an open source framework used for storing and processing big data.

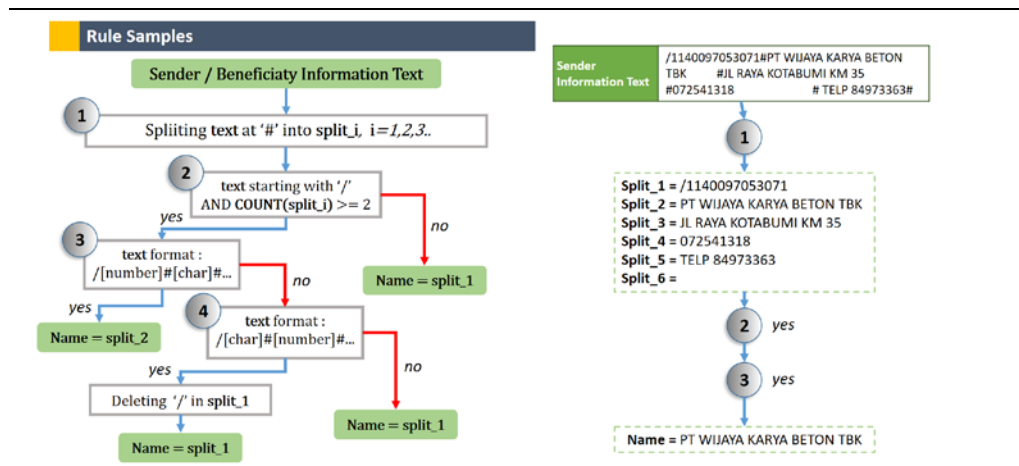
1. Parsing

There are two steps to extract ordering and beneficiary information from Transaction Message (block4) field. The first step is identifying patterns from

message. The second one is defining 61 rules based on identified patten to obtain to corporate name entity. Figure 1 visualize the example of rule based parsing method.

Rule Samples

Figure 1



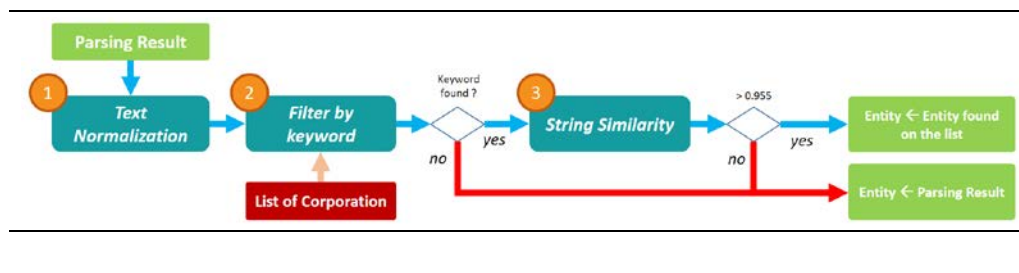
Random sampling of 1.500 transactions data (from 83.300 total transactions on first week period of August 2016) is used for evaluation. Validation on the ruled based parsing result has 96% accuracy. The corporate name entity results is not unique yet because there can be different writing refer to one name. Thus, we need to perform further method to obtain satisfactory result.

2. Entity Resolution

Real-world data, in this case transaction message, is far from perfect due to a plethora of corporate name data entered multiple times in unique sources by various people in their own ways. Entity resolution is a technique used to identify corporate name record in BI-RTGS transactional data that refer to the same corporate and to link the transaction records together. This method, visualized on Figure 1, matches corporate name from transaction message that are nearly identical but maybe not exactly the same.

Identifying similar entities, with different writing

Figure 2



The first step of entity resolution technique is text normalization. We remove non-alphabet characters e.g. dot, comma, space and tab. In order to simplify the writing of corporation name, we replace "and" to symbol "&" and removing stop words such as "PT", "Tbk", "Ltd", "CV", "Pte". The second step is filtering corporations. In this study, we have more than 100 corporation list names. The entity is selected when containing at least 1 (one) word.

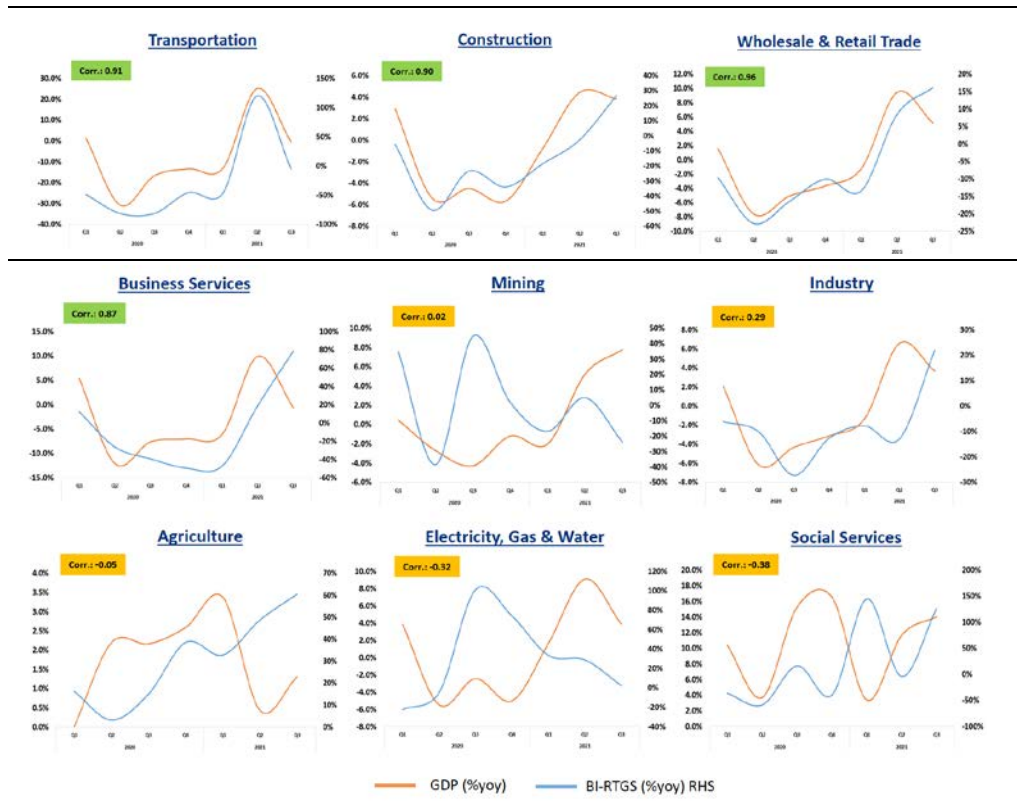
The last one is applying string similarity algorithm. Jaro wringler algorithm is used to calculate entities similarity with threshold more than equal to 0.955. The algorithm is evaluated using random sampling of 1.500 transaction records and results 98.9% accuracy.

Result and Analysis

We construct a new indicator for tracking corporate sector activity that can be produced on a monthly basis within few days after the reference period. We compare corporate sector indicator with sectoral gross domestic product (GDP) data, from Q1 2020 to Q3 2021, published quarterly by The National Statistics Office. The indicators for construction, wholesale and retail trade, transportation and business services sector, have good correlation (≥ 0.87) (Figure 3). Although, mining, industry, agriculture, electricity, gas and water, and social services have lower correlation with GDP.

Corporate sector indicator vs. Indonesian GDP Sector

Figure 3



Conclusion

Big data analytics play an important role to improve the quality of economic analysis and research, as increasingly recognised by policymakers in Bank Indonesia. Applying big data analytics on BI-RTGS transitional data, the new corporate sector indicator is created and (for several sectors) has good correlation with GDP sector. Thus, the indicator can be used for tracking economy activity that will be used as main indicator for financial stability analysis in Bank Indonesia.

References

- S Doerr, L Gambacorta, and J M Serena (2021): "Big data and machine learning in central banking", BIS Working Papers No 930.
- G Gensler and L Bailey (2020): "Deep Learning and Financial Stability", MIT Artificial Intelligence Global Policy Forum.
- C Mertzanis (2018): "Complexity, big data and financial stability", *Quantitative Finance and Economics (QFE)*, 2(3): 637–660.
- G Güngör, M D Özbekler and T P Sümer (2016): "Corporate sector financials from financial stability perspective", Irving Fisher Committee on Central Bank Statistics IFC-ECCBSO-CBRT.
- A Costeiu and F Neagu (2013): "Bridging the Banking Sector With The Real Economy A Financial Stability Perspective", European Central Bank Working Paper Series No 1592.



BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA



Big Data Analytics on Real Time Gross Settlement Data for Tracking Corporate Activity

Amin Endah Sulistiawati, Mohammad Khoyrul Hidayat, Alvin Andhika Zulen
Statistics Department – Bank Indonesia
Email: amin_endah@bi.go.id, moh_khoyrul@bi.go.id, alvin_az@bi.go.id

February 2022

*The views expressed here are those of the authors and
do not necessarily reflect the views of Bank Indonesia*



OUTLINE

1 Background

2 Proposed Idea

3 Data

4 Methodology

5 Result & Analysis





Background

Non-financial corporations constitute a majority of banking sector's loan borrowers. Corporates' credit quality, loan standards, and lending conditions have significant influence in maintaining healthy economic activity. **Fragilities that may arise from corporate sector may leap into the banking sector by causing vulnerabilities and systemic risks** (Güngör et al., 2016).

Financial stability analyses should examine the macroprudential aspects more closely, with more emphasis on the link between the real economy and the financial system. Corporate and household sectors, as well as macroeconomic developments, should be more closely integrated into credit risk assessments of the banking sector (Costeiu & Neagu, 2013).

Corporate sector performance is one of the main indicators of financial system stability. Based on Bank Indonesia's Financial Stability Review report on 2021, stronger corporate sector performance had a knock-on effect of increasing household income and consumption.

The available information related to corporate sector performance is limited to financial reports for listed companies. The reports are published in quarterly period with data lag of up to 2 months. Whereas, Bank Indonesia has weekly and monthly Board of Governor meeting. Hence, **the need for new & more timely indicator for tracking corporate sector activity is becoming more relevant.**

Proposed Idea

This paper explores a new approach for measuring corporate sector activity using Bank Indonesia – Real Time Gross Settlement (BI-RTGS) transaction data by utilizing Big Data Analytics methodology.

We identify payment transaction activities (inflow & outflow) via BI-RTGS of more than 100 corporations from 9 sectors. The selected corporations are chosen to represent the largest share in sales for each sector.



Agriculture



Mining



Industry



Electricity, gas, & water



Construction



Trade



Transportation



Business services



Social services



Data Source : BI-RTGS Transactional Data
Transaction Type : Banks' transactions o/b of customer (TTC 100 & 101) ~ MT103
Total Transactions : \cong 800.000 transactions/months (out of 1,2 mio transactions)

BI-RTGS



DATA STRUCTURE

| Settlement Time | Transaction Type Code (TTC) | Sender Bank | Receiver Bank | Nominal | Block4 |
|--|--|--|---|------------------------------|--|
| Transaction settlement time (MM/DD/YYYY HH:mm) | RTGS transaction code 100 : Transaction o/b of customer 101 : Transaction o/b of customer (w/o account) 111 : Forex Buy/Sell 112 : Interbank Money Msrket etc. | Sender's Bank SWIFT Code (BIC) e.g.: CENAI DJA | Receiver's Bank SWIFT Code (BIC) e.g.: BMRI DJA | Transaction amount in Rupiah | Transaction message, sent by Sender Bank into BI-RTGS system |

| Settlement Time | TTC | Sender Bank | Receiver Bank | Nominal | Block4 |
|-----------------|-----|-------------|---------------|------------------|---|
| 5/2/2018 13:35 | 100 | DXXXIDJA | MXXXIDJA | 4,000,000,000.00 | :20:02RE201804304660#:23B:CRED#:23E:SDVA#:32A:180502IDR4000000000,00# :50K:PAM LYONNAISE JAYA,PT#JL. JEND GATOT SUBROTO KAV 51-52#10270 JAKARTA PUSAT#INDONESIA#:52D:BANK DXXX JAKARTA INDONESIA# :53A:/D/521067000990#DXXXIDJA#:57A:/C/520008000990#MXXXIDJA# :59:/1210005182096#PERSONEL ALIH DAYA, PT# :70:INV. 1804-0167#R/LOCAL# :71A:OUR#:72:/CODTYPTR/100#/CLRC/0670304#:77B:/FEAB/R /PTR/LOCAL |

50K: Ordering Customer

59: Beneficiary Information

70: Remittance Information



Methodology

BI-RTGS



Hadoop



1. Parsing

Extracting ordering and beneficiary entity name from transaction message (block4)



2. Entity Resolution

Identifying similar entities, with different writing



3. Sectoral Mapping

Grouping corporations by sector



Data Extraction



4. Validation & Analysis

- We aggregate the data for incoming & outgoing transactions by sector.
- We calculate the correlation between the new indicator with sectoral GDP data.

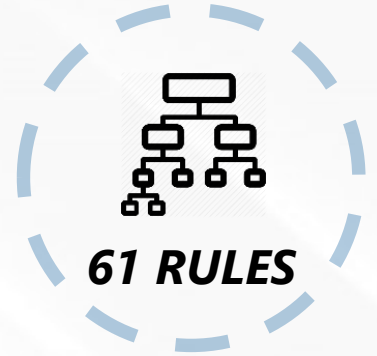


Methodology: Parsing

Extracting sender & beneficiary information from “Transaction Message (block4)” field

Method

- ✓ Identifying writing patterns for sender & beneficiary information .
- ✓ Define some rules in order to extract entities name in sender/beneficiary column based on writing patterns and sender bank code.

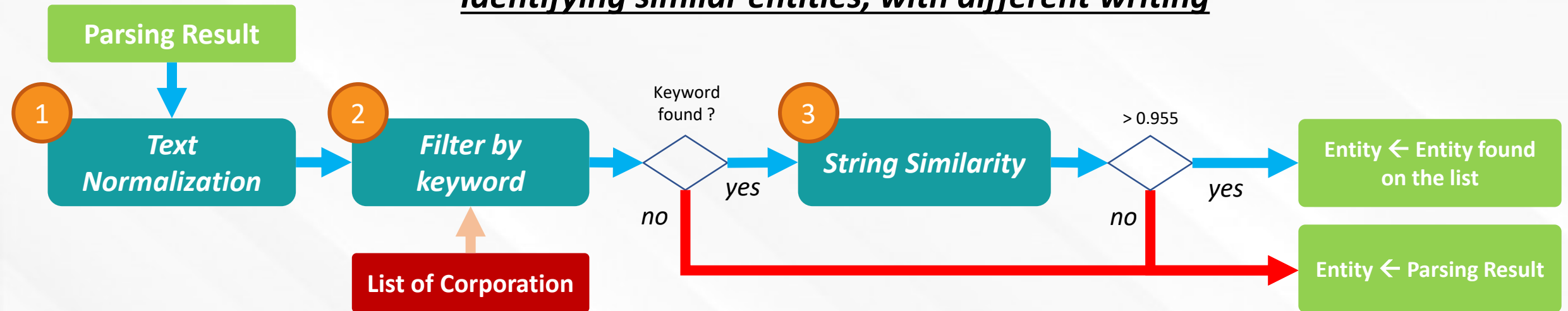


Pattern Samples

| RTGS Data (Sender / Beneficiary Information) | Pattern |
|--|--|
| /NO-ACC#PT SARANA INDAH PERKASA# | /NO-ACC#name# |
| /NO-ACC#PT LONTAR PAPYRUS PULP DAN PAPER IN#DUSTRY# | /NO-ACC#name#name# |
| /NO-ACC#PT PELABUHAN INDONESIA IV (PERSERO)# MAKASSAR# | /NO-ACC#name#address# |
| /NO-ACC#317104000100402170/ 071609261406260#001# | /NO-ACC#acc. number#acc. number# |
| /703096771400#SATYA MITRA SEJAHTERA#JALAN DIPONEGORO NO | /acc.number#name#address |
| /823000009900#ANEKA SAWIT LESTARI#THE PLAZA OFFICE TOWER 35TH FLOOR#JL MH THAMRIN KAV 28-30#10350# | /acc.numbe#name#address#address#address# |
| SWASTISIDDHI AMAGRA PT# | name# |
| /PT GEMA GRAHA SARANA TBK#09102878785# | /name#acc. number# |

Methodology: Entity Resolution

Identifying similar entities, with different writing



1. Text Normalization

- Removing *non-alphabet* character, e.g. “.” (dot) , “,” (comma) , “ ” (space), “\t” (tab)
- Replacing “and” & “dan” with “&”
- Removing stop words, e.g.: “PT”, “Tbk”, “Ltd”, “CV”, “Pte”

2. Filter by keyword

Filtering corporations from the list, containing at least 1 (one) word from text normalization result.

3. String Similarity

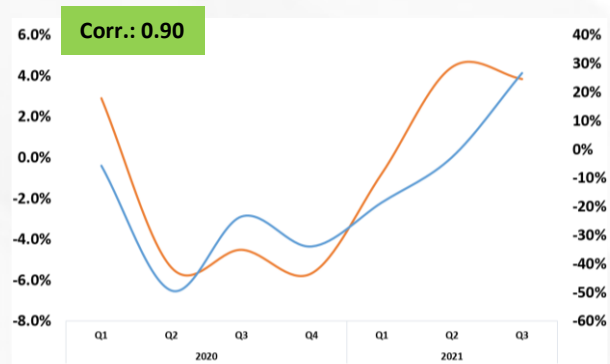
Using Jaro-Winkler algorithm to calculate similarity between entities.

Criteria : String similarity ratio ≥ 0.955

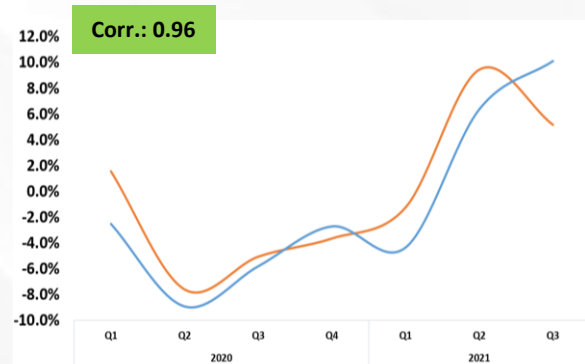
Result & Analysis

We construct a **new indicator for tracking corporate sector activity** that can be produced **on a monthly basis within few days after the reference period**. This indicator also has good correlation with sectoral GDP data (for several sectors), which indicates the usefulness of the indicator for tracking economic activity.

Construction



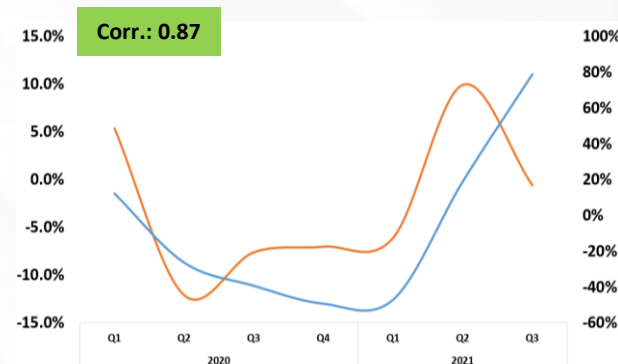
Wholesale & Retail Trade



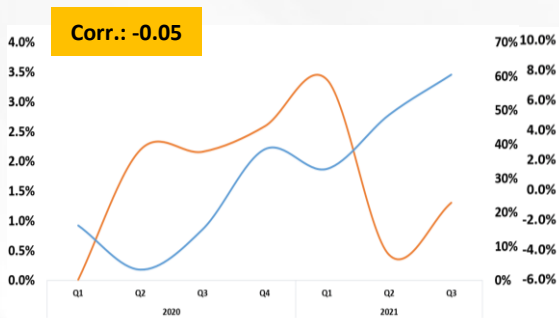
Transportation



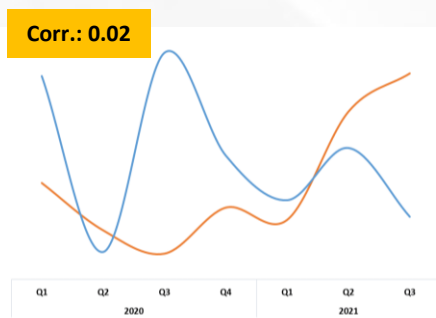
Business Services



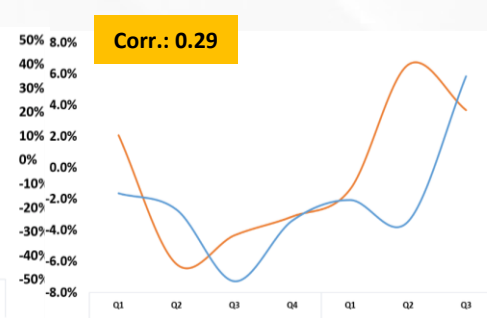
Agriculture



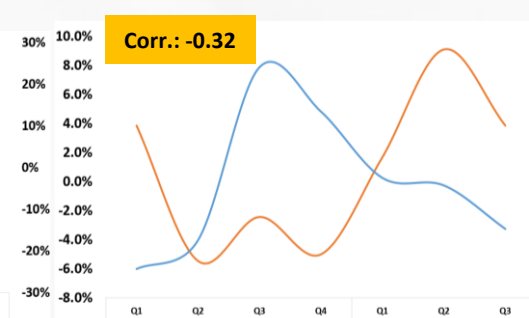
Mining



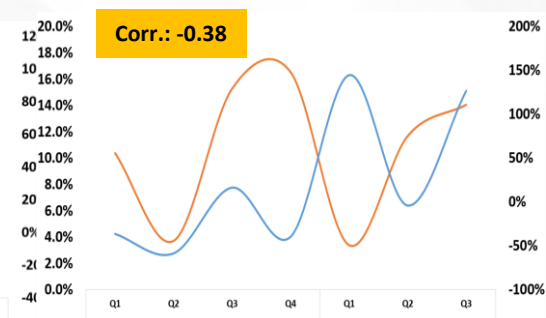
Industry



Electricity, Gas & Water



Social Services



— Sectoral Real GDP (%yoy) — Total Inflow + Outflow BI-RTGS (%yoy) RHS



BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

THANK YOU TERIMA KASIH

Amin Endah Sulistiawati, Mohammad Khoirul Hidayat, Alvin Andhika Zulen
Statistics Department – Bank Indonesia
Email: amin_endah@bi.go.id, moh_khoirul@bi.go.id, alvin_az@bi.go.id

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Deep vector autoregression for macroeconomic data¹

Marc Agustí and Ignacio Vidal-Quadras Costa, European Central Bank;
Patrick Altmeyer, Delft University of Technology

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Deep Vector Autoregression for Macroeconomic Data

Patrick Altmeyer¹, Marc Agusti², Ignacio Vidal-Quadras Costa³

Vector Autoregression is a popular choice for forecasting time series data. Due to its simplicity and success at modelling monetary economic indicators VAR has become a standard tool for central bankers to construct economic forecasts. A crucial assumption underlying the conventional VAR is that interactions between variables through time can be modelled linearly. We propose Deep VAR: a novel approach towards VAR that leverages the power of deep learning in order to model non-linear relationships. By modelling each equation of the VAR system as a deep neural network, our proposed extension outperforms its conventional benchmark in terms of in-sample fit, out-of-sample fit and point forecasting accuracy. In particular, we find that the Deep VAR is able to better capture the structural economic changes during periods of uncertainty and recession. By staying methodologically as close as possible to the original benchmark, we hope that our approach is more likely to find acceptance in the economics domain.

1 Introduction

As stated by the European Central Bank, the monetary transmission mechanism is the process through which monetary policy decisions affect the economy in general and the price level in particular. Uncertainty with respect to this transmission is generally huge, given that it is characterized by long, variable and uncertain dependencies through time and variables. Hence, it is typically challenging to predict how changes in monetary policy actions affect real economic outcomes. It is therefore of foremost importance for policy-makers to use adequate tools to model the underlying mechanisms.

With this in mind, a lot of research on the forecasting of time series has been developed to assess the effect of current policy decisions on future economic variables. Thanks to this, over the last decades policy makers have had more information when taking decisions. This information usually comes in the form of point estimates and interval forecasts. To come up with these estimates, several methodologies have been developed and applied in the time series forecasting literature.

At the time of writing, one the most common methodologies to produce these estimates is the so-called Vector Autoregression (VAR). This framework, which belongs to the traditional toolkit of econometric forecasting techniques, has been shown to provide policy-makers with fairly good and consistent point and interval

¹ Delft University of Technology, p.altmeyer@tudelft.nl

² European Central Bank, marc.agusti_i_torres@ecb.europa.eu

³ European Central Bank, ignacio.vidalquadrascosta@barcelonagse.eu

estimates. It has therefore been used extensively in the monetary policy divisions of central banks.

Simultaneously, with the recent advancements in computational power, and more importantly, the development of advanced machine learning algorithms and deep learning, interesting novel tools have become available that may be useful for forecasting time series. Whereas the good performance of techniques such as VAR is well established, it is still uncertain whether deep learning algorithms can be applied successfully to macroeconomic data.

To this end, this paper contributes a new and ground-breaking methodology that combines the VAR equation-by-equation structure with deep learning. We provide evidence that this improves the model's capacity to capture potentially highly non-linear relationships in the underlying data generating process. The primary objective of this paper is to develop a methodology that produces improved modelling outcomes while deviating as little as possible from the established VAR framework, thereby keeping things straight-forward and familiar to economists. We show that the existing VAR methodology can be easily extended to the broader class of Deep VAR models and provide solid empirical evidence that Deep VAR models consistently outperform the conventional approach.

To the best of our knowledge, this is the first paper to propose a Deep VAR framework of this structure, namely, to fit a deep neural network for each equation of the VAR process. Although previous work has explored the use of deep learning to forecast macroeconomic time series, previous proposed methodologies deviate more from the conventional VAR framework. For example, Verstyuk (2020) chooses to model the whole system through one unified deep neural network. We find that the equation-by-equation approach not only helps to maintain interpretability and simplicity, but also appears to produce better modelling outcomes. To enable researchers and practitioners to easily implement our proposed methodology, we have developed a unified framework for estimating Deep VAR models in R and plan to continue its development going forward.

We find that the Deep VAR methodology outperforms the traditional VAR framework in terms of in-sample and out-of-sample fit as well as with respect to forecasting accuracy. In particular, the Deep VAR appears to be better at capturing non-linear dynamics underlying the time series process. It therefore leads to consistently lower modelling errors than the VAR, especially during periods of economic downturn and uncertainty.

Arguably policy makers are not only interested in the forecasting accuracy of the model but are typically also concerned with inference. For example, central banks are often interested in knowing to what extent interest rates granger cause other variables within the monetary transmission mechanism. Another aspect policy makers and researchers care about is how the variables of the system evolve through time in response to innovations. This information is typically recovered using Impulse Response Functions (IRFs). The linear additive modelling assumption underlying the conventional VAR makes inference straight-forward. In the case of Deep VAR models inference is arguably more complicated, though promising avenues have recently been explored (Verstyuk 2020). We believe that the methodology proposed in this paper can be augmented to the inference realm in future work.

The remainder of the paper is structured as follows: in section 2 we present a literature review of prior research on the methodologies used to provide forecasts and on the monetary transmission mechanism in general. Section 3 provides a

detailed description of the data we use for our empirical exercises. In section 4 we present the traditional VAR methodology and develop our proposed Deep VAR model. Sections 5 and 6 present our empirical findings and possible extensions and caveats, respectively. Finally, section 7 concludes.

2 Literature review

There is broad agreement among economists on the fact that monetary policy affects economic activity in the short and medium term. Friedman and Schwartz (2008) found that monetary policy actions are followed by movements in real output that may last for two years or more (Romer and Romer (1989); Bernanke (1990)). The underlying forces that trigger these outcomes is of great interest to most economists. Central bankers in particular aim to understand the monetary transmission mechanism. If monetary policy affects the real economy, then what exactly is the transmission mechanism through which these effects occur? This is one of the questions which is among the most important and controversial topics in modern-day macroeconomics.

In the aftermath of the oil price shock in the 1970's, interest emerged in understanding business cycles. To this end economists initially made use of large-scale macroeconomic models, which was criticized by Lucas (1976), stating that the assumption of invariant behavioural equations was inconsistent with dynamic maximizing behaviour. Hence, **New Classical** economists started to make use of so-called market clearing models of economic fluctuations. With the goal of really taking into account productivity shocks, **Real Business Cycle** models were developed (Kydland and Prescott (1982)).

Following the failure of large-scale macroeconomic models when trying to predict business cycles, the economic profession resorted to structural vector autoregressive (VAR) models to analyse business cycles, which proved to be useful for capturing the impact of policy actions. Sims et al. (1986) suggested that VAR models were an efficient tool to evaluate macroeconomic models. One of the advantages of VAR models is their simplicity, which makes it easy to estimate and interpret them.

Yet, this simplicity comes at a cost: conventional VAR models are typically not able to capture non-linear relationships in the data, which might be a significant limitation. In the very short run many time series can be expected to behave more or less according to their past and a linear model may be efficient to capture dynamics, but for longer term dependencies this is typically not the case. With respect to the economic time series that form part of the monetary transmission mechanism, specifically output, inflation, interest rates and labour market variables, non-linear dependencies are likely to form part of the data generating process as shown by Brock et al. (1991). This is true in particular during times of abrupt and significant economic fluctuations.

During past years, economists have therefore started to add non-linear techniques to their forecasting tool kit. Machine Learning has contributed a lot to this field of research. Some of the most popular machine learning techniques which do not assume a linear relationship between inputs and outputs include K-Nearest Neighbours (first introduced by Fix and Hodges (1951)), Support Vector Machines (mostly developed by Cortes and Vapnik (1995)), Random Forests (first introduced in

1995 by Ho (1995)) and Deep Artificial Neural Networks (first proposed in 1943 by McCulloch and Pitts (1990)). The latter have been explored previously in the realm of time series forecasting (Hamzaçebi (2008), G. P. Zhang (2003), Kihoro, Otieno, and Wafula (2004)). Neural networks are non-parametric models that have been shown to be particularly successful at capturing non-linearities (G. Zhang, Patuwo, and Hu (1998), G. P. Zhang (2003)).

A particular subclass of neural networks used primarily for sequential data are recurrent neural networks (RNN). RNNs propagate previous outputs recursively allowing the model to learn persistent dependencies and thereby making them very efficient for time series data (Dorffner 1996). A recent staff working paper published by the Bank of England provides some empirical support for the argument that deep learning can be successfully applied to macroeconomic data (Joseph et al. 2021). The authors run a horse race for forecasting inflation across different time horizons comparing the performance of linear and non-linear models. They find that neural networks in particular and other common machine learning algorithms are useful for forecasting particularly at a longer horizon.

3 Data

To evaluate our proposed methodology empirically we use a sample of monthly US data on leading economic indicators, which spans the period of January 1959 through March 2021. We use the relatively novel FRED-MD data base which is updated monthly and publicly available (McCracken and Ng 2016). The sample spans from March, 1959 to March, 2021 providing us with a relatively rich data set of macroeconomic time series with $T = 745$ observations.

In order to investigate the monetary transmission mechanism, the literature typically focuses on variables related to economic output, inflation, short and long term interest rates as well as labour market indicators. Some go beyond to also include stock price indices, money and credit aggregates, balance of payments figures, confidence indicators and some cases foreign domestic indicators. In this paper we limit our attention to the four main indicators mentioned above. In particular we use changes in the industrial production index (IP) to measure output growth, changes in the growth of the (all items) consumer price index (CPI) to measure inflation, the Federal Funds Rate (FFR) as our interest rate and the unemployment rate (UR) as our labour market indicator. Note that we use IP rather than the gross domestic product as a proxy for output, because the latter is only available at quarterly frequency.

Another strength of the FRED-MD is the fact that the data is already pre-processed. Specifically, the industrial production index comes in log differences, the CPI in second-order log differences and both the Fed Funds Rate and the unemployment rate in first-order differences. This has allowed us to let the data enter our estimations without any further adjustments, which should facilitate the reproducibility of our results.

4 Methodology

In conventional Vector Autoregression (VAR) dependencies of any system variable on past realizations of itself and its covariates are modelled through linear equations. This corresponds to a particular case of the broader class of Deep Vector Autoregressions investigated here and will serve as the baseline for our analysis.

4.1 Vector Autoregression

Let \mathbf{y}_t denote the $(K \times 1)$ vector of variables at time t . Then the VAR(p) with p lags and a constant deterministic term is simply a linear system of stochastic equations of the following form:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{A}_2 \mathbf{y}_{t-2} + \dots + \mathbf{A}_p \mathbf{y}_{t-p} + \mathbf{u}_t, \quad \mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \Sigma_u) \quad (4.1)$$

The matrices $\mathbf{A}_m \in \mathbb{R}^{K \times K}$, where $m \in \{1, \dots, p\}$, contain the reduced form coefficients and $\mathbf{u}_t \in \mathbb{R}^{K \times 1}$ is a vector of errors for which $\mathbb{E} \mathbf{u}_t$, $\mathbb{E} \mathbf{u}_t \mathbf{u}_t^T = \Sigma$ and $\mathbb{E} \mathbf{u}_t \mathbf{u}_s^T = \mathbf{0}$ for all $t \neq s$. We refer to (4.1) as the **reduced form** representation of the VAR(p) because all right-hand side variables are predetermined (Kilian and Lütkepohl 2017).

We can restate (4.1) more compactly as

$$\mathbf{y}_t = \mathbf{A} \mathbf{z}_{t-1} + \mathbf{u}_t \quad (4.2)$$

where $\mathbf{A} = (\mathbf{c}, \mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_p) \in \mathbb{R}^{K \times (Kp+1)}$ and $\mathbf{z}_{t-1} = (1, \mathbf{y}_{t-1}^T, \dots, \mathbf{y}_{t-p}^T)^T \in \mathbb{R}^{(Kp+1) \times 1}$. The expression in (4.2) demonstrates that the VAR(p) can be considered as a **seemingly unrelated regression** (SUR) model composed of individual regressions with common regressors (Greene 2012). In fact, it is useful to note for our purposes that the VAR(p) can be estimated efficiently through equation-by-equation OLS regression. In particular, it follows from (4.2) that

$$y_{it} = c_i + \sum_{m=1}^p \sum_{j=1}^K a_{jm} y_{jt-m} + u_{it}, \quad \forall i = 1, \dots, K \quad (4.3)$$

which corresponds to the key modelling assumption that at any point in time t any time series $i \in 1, \dots, K$ is just a weighted sum of past realizations of itself and all other variables in the system. This assumption makes the estimation of VAR(p) processes remarkably simple. Perhaps more importantly, the assumption of linearity also greatly facilitates inference about VAR models.

For implementation purposes it is generally more useful to estimate the VAR(p) through one single OLS regression. To this end let $\tilde{\mathbf{A}} = \mathbf{A}^T$ and note that (4.2) can be restated even more compactly as

$$\mathbf{y} = \mathbf{Z} \tilde{\mathbf{A}} + \mathbf{u}_t \quad (4.4)$$

with $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_T)^T \in \mathbb{R}^{T \times K}$ and $\mathbf{Z} \in \mathbb{R}^{T \times (Kp+1)}$. Then the closed form solution for OLS is simply $\tilde{\mathbf{A}} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y}$ and hence

$$\mathbf{A} = \mathbf{y}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \quad (4.5)$$

4.2 Deep Vector Autoregression

We propose the term Deep Vector Autoregression to refer to the broad class of Vector Autoregressive models that use deep learning to model the dependences

between system variables through time. In particular, as before, we let \mathbf{y}_t denote the $(K \times 1)$ vector that describes the state of system at time t . Consistent with the conventional VAR structure we assume that each individual time series y_{it} can be modelled as a function of lagged realizations of all variables y_{jt-m} , $j = 1, \dots, K$, $m = 1, \dots, p$.

More specifically, we have

$$y_{it} = f_i(\mathbf{y}_{t-1:t-p}; \theta) + v_{it} \quad , \quad \forall i = 1, \dots, K \quad (4.6)$$

where $\mathbf{y}_{t-1:t-p} = \{y_{jt-m}\}_{j=1, \dots, K}^{m=1, \dots, p}$ is the vector of lagged realizations, f_i is a variable specific mapping from past lags to the present and θ is a vector of parameters. While in the conventional VAR above we assumed that the multivariate process can be modelled as a system of linear stochastic equations, our proposed Deep VAR(p) can similarly be understood as a system of potentially highly non-linear equations. As we argued earlier, Deep Learning has been shown to be remarkably successful at learning mappings of arbitrary functional forms (Goodfellow, Bengio, and Courville 2016).

Note that the input and output dimensions in (4.6) are exactly the same as in the conventional VAR(p) model (equation (4.3)): f_i maps from $\mathbf{y}_{t-1:t-p} \in \mathbb{R}^{Kp \times 1}$ to a scalar. Our proposed plain-vanilla approach to Deep VAR models diverges as little as possible from the conventional approach: it boils down to simply modelling each of the univariate outcomes in (4.6) as a deep neural network. We can restate this approach more compactly as

$$\mathbf{y}_t = \mathbf{f}(\mathbf{y}_{t-1:t-p}; \theta) + \mathbf{v}_t \quad (4.7)$$

where $\mathbf{f}(\cdot) = (f_1(\cdot), f_2(\cdot), \dots, f_K(\cdot))^T \in \mathbb{R}^{K \times 1}$ is just the stacked vector of mappings to univariate outcomes described in (4.6).

The notation in (4.7) gives rise to a more unified and general approach to Deep VAR models that would treat the whole process as one single dynamical system to be modelled through one deep neural network \mathbf{g} :

$$\mathbf{y}_t = \mathbf{g}(\mathbf{y}_{t-1:t-p}; \theta) + \mathbf{v}_t \quad (4.8)$$

This approach is in fact proposed and investigated by Verstyuk (2020) in his upcoming publication. We decided to go with the approach in (4.7) for two reasons: firstly, the link to conventional VAR models is made abundantly clear through this implementation and, secondly, we found that the equation-by-equation approach produces good modelling outcomes and is relatively easy to implement using state-of-the-art software.

Finally, note that if f_i in (4.3) is assumed to be linear and additive for all $i = 1, \dots, K$ then we are back to the conventional VAR(p). This illustrates the point we made earlier that the linear VAR(p) is just a particular case of a Deep VAR(p). Since the model described in equations (4.6) and (4.7) is less restrictive but otherwise consistent with the conventional VAR framework, we expect that it outperforms the traditional approach towards modelling multivariate time series processes.

4.3 Deep Neural Networks - a whistle-stop tour

So far we have been speaking about deep learning in rather general terms. For example, above we have referred to our model of choice for learning the mapping $f_i: \mathbf{y}_{t-1:t-p} \mapsto y_{it}$ as a **deep neural network**. The class of deep neural networks can

further be roughly divided into **feedforward neural networks** and **recurrent neural networks**. As the term suggests, the latter is generally used for sequential data and therefore our preferred model of choice. Nonetheless, below we will begin by briefly exploring feedforward neural networks first. This should serve as a good introduction to neural networks more generally and (even though we have not tested this empirically) there is good reason to believe that even Deep VAR models using feedforward neural networks perform well.

4.3.1 Deep Feedforward Neural Networks

The term **deep feedforward neural network** or **multilayer perceptron** (MLP) is used to describe a broad class of models that are composed of possibly many functions that together make up the directed acyclical graph. The functions $f_i(\cdot)$ - sometimes referred as layers \mathbf{h}_i - are chained together hierarchically with the first layer feeding forward its outputs to the second layer and so on (Goodfellow, Bengio, and Courville 2016). Applied to our case, an MLP with H hidden layers can be loosely defined as follows:

$$f_i(\mathbf{y}_{t-1:t-p}; \theta) = f_i^{(H)} \left(f_i^{(H-1)} \left(\dots f_i^{(1)}(\mathbf{y}_{t-1:t-p}) \right) \right) \quad (4.9)$$

The depth of the MLP is defined by the number of hidden layers H , where, generally speaking, deeper networks are more complex.

The desired outputs of any $f_i^{(h)}$ that will serve as inputs for $f_i^{(h+1)}$ cannot be inferred from the training data $\mathbf{y}_{t-1:t-p}$ ex-ante, which is where the term **hidden** layer stems from. Each $f_i^{(h)}$ is typically valued on a vector of hidden units, each of them receiving a vector of inputs from $f_i^{(h-1)}$ and returning a scalar that is referred to as activation value. This approach is inspired by neuroscience, hence the term **neural network** (Goodfellow, Bengio, and Courville 2016).

4.3.2 Deep Recurrent Neural Networks

Recurrent neural networks (RNN) are based on the idea of persistent learning: a continuous process that evolves gradually and at each step uses information about its prior states instead of continuously reinventing itself and starting from scratch. To this end, RNNs develop the basic concepts underlying feedforward neural networks by incorporating feedback loops. Formally the loop is typically made explicit as follows

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t; \theta) \quad (4.10)$$

where $\mathbf{h}_t \in \mathbb{R}^{N \times 1}$ corresponds to the hidden state of the dynamical system at time t that the RNN learns (Goodfellow, Bengio, and Courville 2016), and N corresponds to the number of hidden units in each hidden layer, known as the width of the layer. In the given context we have that $\mathbf{x}_t = \mathbf{y}_{t-1:t-p}$ as specified in (4.7). Given some random initial hidden state vector \mathbf{h}_0 the RNN updates parameters sequentially at each time step t as follows

$$\begin{aligned} \mathbf{a}_t &= \mathbf{b} + \mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\mathbf{h}_{t-1} \\ \mathbf{h}_t &= \tanh(\mathbf{a}_t) \\ \hat{y}_{it} &= c + \mathbf{v}^T \mathbf{h}_t \end{aligned} \quad (4.11)$$

where $\mathbf{b} \in \mathbb{R}^{N \times 1}$ is a vector of constants (biases), $c \in \mathbb{R}$ is a scalar that captures the deterministic term of the VAR, \tanh is the hyperbolic tangent activation function, $\mathbf{W}, \mathbf{U} \in \mathbb{R}^{N \times N}$ are coefficient matrices and $\mathbf{v} \in \mathbb{R}^{N \times 1}$ is a vector of coefficients. Note

that to simplify the notation we have omitted the layer index in (4.11): to be specific, \mathbf{h}_t really represents $\mathbf{h}_t^{(H)}$ (the ultimate hidden layer), \mathbf{h}_{-1} stands for $\mathbf{h}_t^{(H-1)}$ (the penultimate layer). Finally, at each step t the first layer $\mathbf{h}_t^{(0)}$ of the forward propagation corresponds to $\mathbf{y}_{t-1:t-p}$.

A shortfall of generic recurrent neural networks is that they fail to capture long-term dependencies. More specifically, if parameters are propagated over too many stages in a simple RNN, it typically suffers from the problem of **vanishing gradients** (Goodfellow, Bengio, and Courville 2016). Fortunately, there exist effective extensions of the RNN, most notably the long short-term memory (LSTM), which was introduced by Hochreiter and Schmidhuber (1997) and is our model of choice for Deep VAR models. The key idea underlying LSTMs is to regulate exactly how much information is propagated from one cell state vector \mathbf{C}_{t-1} to the next \mathbf{C}_t through the introduction of so called sigmoid gates:

“The LSTM [has] the ability to remove or add information to the cell state, carefully regulated by structures called gates. Gates are a way to optionally let information through.” — Olah (2015)

These regulating gate layers include a **forget gate** \mathbf{f}_t , an **input gate** \mathbf{i}_t and a **output gate** \mathbf{o}_t . Each of them are vector-valued sigmoid functions whose elements $\mathbf{f}_{it}, \mathbf{i}_{it}, \mathbf{o}_{it}$ are bound between 0 and 1. Their individual purposes are implied by their names: faced with \mathbf{h}_{t-1} and $\mathbf{y}_{t-1:t-p}$, the forget gate regulates how much of each individual unit in \mathbf{C}_{t-1} is retained. Then the input gate regulates which units of \mathbf{C}_{t-1} should be updated and to what candidate values $\tilde{\mathbf{C}}_{t-1}$. Using the previous two steps the actual update is performed according to the following rule

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{C}}_{t-1} \quad (4.12)$$

where \odot indicates the element-wise product. Finally, the output gate acts like a filter on \mathbf{C}_t : the new hidden state is computed as $\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{C}_t)$ where as before we use the hyperbolic tangent as our activation function.⁴ Formally, we can summarize the LSTM neural network underlying our Deep VAR framework as follows:

$$\begin{aligned} \mathbf{f}_t &= \sigma(\mathbf{b}_f + \mathbf{W}_f \mathbf{h}_{t-1} + \mathbf{U}_f \mathbf{h}_{-1}) \\ \mathbf{i}_t &= \sigma(\mathbf{b}_i + \mathbf{W}_i \mathbf{h}_{t-1} + \mathbf{U}_i \mathbf{h}_{-1}) \\ \mathbf{o}_t &= \sigma(\mathbf{b}_o + \mathbf{W}_o \mathbf{h}_{t-1} + \mathbf{U}_o \mathbf{h}_{-1}) \\ \mathbf{C}_t &= \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{b}_c + \mathbf{W}_c \mathbf{h}_{t-1} + \mathbf{U}_c \mathbf{h}_{-1}) \\ \mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{C}_t) \\ \hat{\mathbf{y}}_{it} &= c + \mathbf{v}^T \mathbf{h}_t \end{aligned} \quad (4.13)$$

which is best understood when read from top to bottom. Once again we have simplified the notation by omitting the layer index in (4.11). The same notation as before applies.

4.4 Model selection

There are at least two important modelling choices to be made in the context of conventional VAR models. The first choice concerns properties of the time series data itself, in particular the order of integration and cointegration. The second choice is about the lag order p . In order to arrive at appropriate decisions regarding these

⁴ For a clear and detailed exposition see Olah (2015).

choices the VAR literature provides a set of guiding principles. We propose to apply these same principles to the Deep VAR, firstly because they are intuitive and simple and secondly because treating both models equally to begin with allows for a better comparison of the two models at the subsequent modelling stages.

4.4.1 Stationarity

When working with time series we are generally concerned about stationarity. Broadly speaking stationarity ensures that the future is like the past and hence any predictions we make based on past data adequately describe future outcomes. In order to state stationarity conditions in the VAR context it is convenient to restate the K -dimensional VAR(p) process in companion form as

$$\mathbf{Y}_t = \begin{pmatrix} \mathbf{c} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \mathbf{A}\mathbf{Y}_{t-1} + \begin{pmatrix} \mathbf{u}_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (4.14)$$

where $\mathbf{Y}_t = (\mathbf{y}_t^T, \dots, \mathbf{y}_{t-p+1}^T)^T \in \mathbb{R}^{Kp \times 1}$ and $\mathbf{A} \in \mathbb{R}^{Kp \times Kp}$ is referred to as the companion matrix (Kilian and Lütkepohl 2017). Stationarity of the VAR(p) follows from stability: a VAR(p) is stable if the effects of shocks to the system eventually die out. Stability can be assessed through the system's autoregressive roots or equivalently by looking at the eigenvalues of the companion matrix \mathbf{A} (Kilian and Lütkepohl 2017). In particular, for the VAR(p) in (4.14) to be stable we condition that the Kp eigenvalues λ that satisfy

$$\det(\mathbf{A} - \lambda \mathbf{I}_{Kp}) = 0$$

are all of absolute value less than one. Stability implies that the first and second moments of the VAR(p) process are time-invariant, hence ensuring weak stationarity (Kilian and Lütkepohl 2017).

A straight-forward way to deal with stationarity of VAR models is to simply ensure that the individual time series entering the system are stationary. This usually involves differencing the time series until they are stationary: for any time series y_i that is integrated of order $I(\delta)$, there exists a δ -order difference that is stationary. An immediate drawback of this approach is the loss of information contained in the levels of the time series. Modelling approaches that take into account cointegration of individual time series can ensure system stationarity and still let individually non-stationary time series enter the system in levels (Hamilton 2020).

4.4.2 Lag order

The VAR's lag order p can to some extent be thought of as the persistency of the process: past innovations that still affect outcomes in time t happened at most p periods ago. From a pure model selection perspective we can also think of additional lags in terms of additional regressors that add to the model's complexity. From that perspective, choosing a lower lag order corresponds to a form of regularization as it pertains to a more parsimonious model.

Various strategies have been proposed to estimate the true or optimal lag order p empirically (Kilian and Lütkepohl 2017). Among the most common ones are sequential testing procedures and selection based on information criteria. The former involves sequentially adding or removing lags - **bottom-up** and **top-down** testing, respectively - and then testing model outcomes in each iteration. A common point of criticism of sequential procedures is that the order tests matters (Lütkepohl (2005)).

Here we will focus on selection based on information criteria, which to some extent makes the trade-off between bias and variance explicit (Kilian and Lütkepohl 2017). In particular, it generally involves minimizing information criteria of the following form

$$C(m) = \log \left(\det \left(\hat{\Sigma}(m) \right) \right) + \ell(m) \quad (4.15)$$

where $\hat{\Sigma}$ is just the sample estimate of the covariance matrix of errors and ℓ is a loss function that penalizes high lag orders. In particular, we have that our best estimate of the optimal lag order p is simply

$$\hat{p} = \underset{m \in \mathcal{P}}{\operatorname{argmin}} C(m) \quad (4.16)$$

where $\mathcal{P} = [m_{\min}, m_{\max}]$. We will consider all of the most common functional choices for (4.15).

4.4.3 Neural Network Architecture

By now it should be clear that deep neural networks come in many shapes and sizes. When thinking about the architecture of a neural network many different design choices can be made and networks can thus be tailored to specific use cases. Here, we intend to keep things simple and vary only the depth and width of the LSTMs underlying the Deep VAR. The number of hidden units per hidden layer is held constant across layers.

Figure 4.1 illustrates a simulated network architecture for the case of two lags ($p = 2$) and four variables ($K = 4$). We can see that the first layer corresponds to the inputs, that is, the input layer $\in \mathbb{R}^{Kp \times 1}$. This architecture consists of $H = 2$ hidden layers each counting twenty hidden units. Since we are modelling equation-by-equation, there is only one output unit, namely variable y_{it} .

With respect to network compilation, the popular Adam optimization algorithm is used (Kingma and Ba 2014). This algorithm can be used instead of the more traditional stochastic gradient descent to update network weights. There are several reasons to use this algorithm that are particularly appealing, among them its straightforward implementation and its computational efficiency. Adam distinguishes itself from classic stochastic gradient descent in that it uses adaptive learning rates.

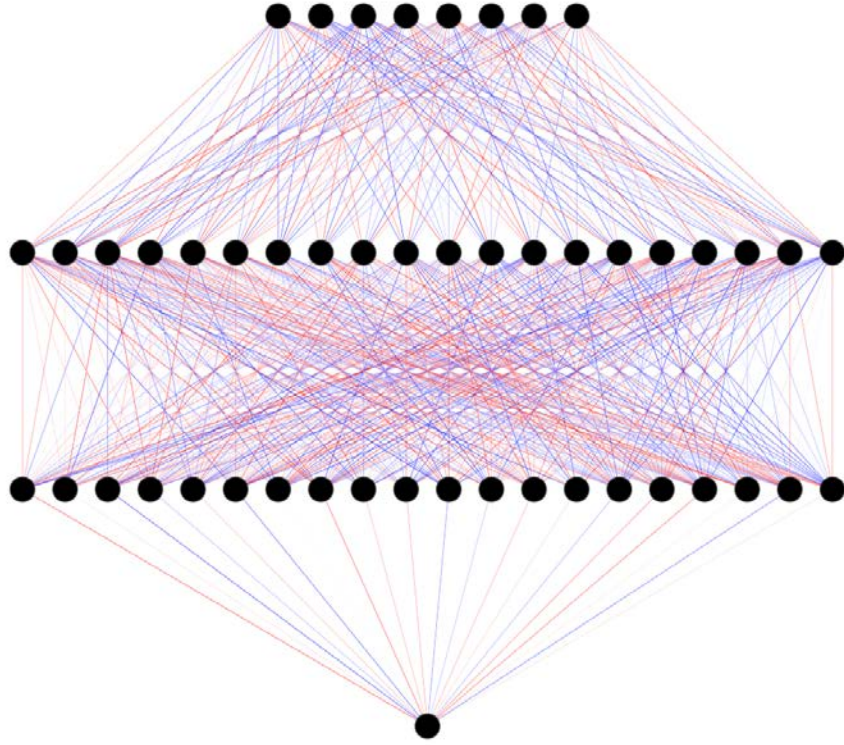


Figure 4.1: Neural Network Architecture.

As mentioned above, the estimation of deep neural networks involves a very large number of parameters and hence regularization is an important concern. One way to mitigate the risk of overfitting is to choose a neural network architecture that is neither excessively wide nor deep. Another way to regularize the neural network is to use the fact that optimization at the training phase is stochastic. One greedy way to reduce overfitting risk is therefore to simply retrain the network multiple times and then average over the obtained parameter estimates and predictions (Srivastava et al. 2014). While theoretically appealing, this approach is computationally prohibitive. Instead, another layer of stochasticity can be introduced at the training stage through **dropout**: at each training iteration and each stage of the forward propagation a share of the hidden units is simply dropped at random. This approach mimics the idea of repeated training. Dropout adds noise into the model and thereby avoids that hidden layers try to adapt to a mistake made by previous hidden layers.

5 Empirical results

We now proceed to benchmark the proposed Deep VAR model against the conventional VAR and other existing approaches using our macroeconomic time series data. To begin with, we compare the models in terms of their in-sample fit. For this part of the analysis the models will be strictly run under the same framing conditions. Due to the RNN's capacity to essentially model any possible function $f_i(\cdot)$ the Deep VAR dominates competing approaches in this realm. We investigate during what time periods the out-performance of the Deep VAR is particularly striking to gain a better understanding of when and why it pays off to relax the linearity constraint.

These findings with respect to in-sample performance provide some initial evidence in favour of the Deep VAR. But since a reduction in modelling bias is typically associated with an increase in variance, we are particularly interested in benchmarking the models with respect to their out-of-sample performance. To this end we split our sample into train and test subsamples. We then firstly benchmark the models in terms of their pseudo out-of-sample fit. Finally we also look at model performance with respect to n -step ahead pseudo out-of-sample forecasts.

The final part of this section relaxes the constraint on the framing conditions. In particular, we investigate how hyperparameter tuning with respect to the neural network architecture and lag length p can improve the performance of the Deep VAR.

5.1 In-sample fit

For this first empirical exercise all models are trained on the full sample. We have decided to include the post-Covid sample period despite the associated structural break, since it serves as interesting point of comparison. The optimal lag order as determined by the Akaike Information Criterion is $p = 6$, where we used a maximum possible lag of $p_{\max} = 12$ corresponding to one year. A look at the eigenvalues of the companion matrix showed that the VAR(6) is stable. The LSTM models underlying the Deep VAR are composed of $H = 2$ that count $N = 32$ hidden units each. The dropout rate is set to $p = 0.25$.

To assess the fit of our models we use squared residuals. Figure 5.1 shows the cumulative loss of the Deep VAR model and its conventional benchmarks for each of the time series over the whole sample period. Aside from the linear VAR, we have added another popular approach towards VAR models that addresses non-linearity (Threshold VAR). We have also added a Random Forest Regressor (RF) for comparison, which was trained on the entire FRED-MD database, so far more variables than the four output variables. Previous studies have shown that RF tends to well at high-dimensional time series modelling (Masini, Medeiros, and Mendes 2021).

The first thing we can observe is that the RMSE of the Deep VAR is consistently flatter than the RMSE of its benchmarks. With respect to model fit, the Deep VAR dominates throughout the almost the entire sample period and for all of the considered variables. This empirical observation seems to confirm our expectation that the vector autoregressive process is characterized by important non-linear dependencies across time and variables that the conventional VAR and even the TVAR and RF fail to capture.

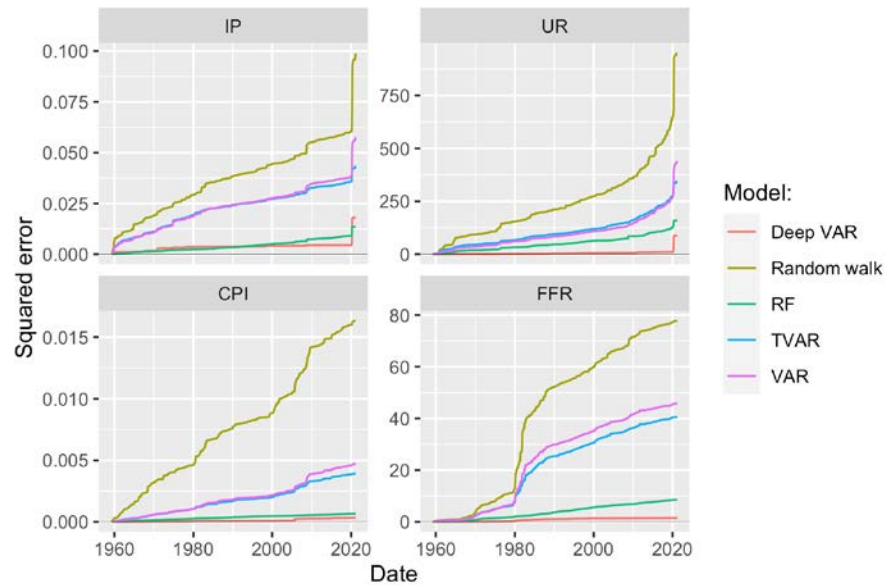


Figure 5.1: Comparison of cumulative loss over the entire sample period for Deep VAR and benchmarks.

Figure 5.1 is especially useful to assess in which specific periods the Deep VAR model fits the data better than alternative approaches. From the very beginning and across variables, we observe that the increase in cumulative loss for other models is greater than for the Deep VAR model.

The US economy during 1960s was influenced by John F. Kennedy's introduction of **New Economics**, which was informed by Keynesian ideas and characterized by increasing levels of inflation, a reduction in unemployment and output growth. The change in government certainly corresponded to a regime switch with respect to the economy (Perry and Tobin 2010) and in that sense it is interesting to observe that the Deep VAR appears to be doing a better job at capturing the underlying changes. The 1970s can be broadly thought of as a continuation of New Economics and loosely defined as a period of stagflation. The Deep VAR continues to outperform during that period.

The first truly interesting development we can observe in Figure 5.1 coincides with the onset of the Volcker disinflation period. Following years of sustained CPI growth, Paul Volcker set the Federal Reserve on course for a series of interest rate hikes as soon as he became chairperson of the central bank in August 1979. The shift in monetary policy triggered fundamental changes to the US economy and in particular the key economic indicators we are analysing here throughout the 1980s (Goodfriend and King 2005). Despite this structural break, the increase in the cumulative RMSE of the Deep VAR remains almost constant during this decade for most variables. The performance of the VAR on the other hand is unsurprisingly poor over the same period, in particularly so for the CPI and the Fed Funds Rate, which arguably were the two variables most directly affected by the change in policy. The Deep VAR also clearly dominates the VAR with respect to the output related variables (IP) and to a lesser extent unemployment.

These findings indicate that changes to the monetary transmission mechanism in response to sudden policy shifts are not well captured by a linear-additive vector autoregressive model. Instead they appear to unfold in a high-dimensional latent state space, which the Deep VAR by its very construction is designed to learn.

Following the Volcker disinflation period, Figure 5.1 does not reveal any clear outperformance of either of the models during the 1990s. Interestingly the dot-com bubble has little effect on either of the models, aside from a small pick-up in cumulative loss with respect to the CPI for both models. With all that noted, the Deep VAR still continuously outperforms the VAR since evidently its cumulative loss increases at a slower pace altogether.

As the Global Financial Crisis unfolds around 2007 the pattern we observed for the Volcker disinflation reemerges, albeit to a lesser extent: there is a marked jump in the difference between the cumulative loss of the VAR and the Deep VAR, in particular so for the CPI, the Fed Funds rate and industrial production. The gap for all these variables continues to widen during the aftermath of the crisis. The Deep VAR once again does a better job at modelling the changes that the dynamical system undergoes: post-crisis US monetary policy was characterized by very low interest rates, low levels of inflation as well as the introduction of a range of non-conventional monetary policy tools including quantitative easing and forward guidance.

Finally, it is also interesting to observe how the different models perform in response to the unprecedented exogenous shock that Covid-19 constitutes. All models exhibit an abrupt and substantial increase in loss with respect to both IP and UR - the two series that were arguably most strongly affected by Covid. Evidently, the magnitude of that sudden increase is somewhat larger in absolute terms for VAR than for Deep VAR. Still, it is also worth pointing out that both the Threshold VAR and the Random Forest Regressor are less adversely affected by the Covid shock than Deep VAR.

As a sanity check we also visually inspected the distributional properties of the model residuals for the full-sample fit. The outcomes are broadly consistent across models: while for some variables residuals are clearly not Gaussian, we see no evidence of serial autocorrelation of residuals (see Figures 9.2 and 9.3 in the appendix).

5.2 Out-of-sample fit

In order to assess if the Deep VAR's outperformance is a consequence of overfitting, we now repeat the previous exercise, but this time we train the models on a subsample of our data. The training sample spans from March, 1959 to October, 2008, whereas the test data (including validation period) goes from November, 2008 to March, 2021. This corresponds to training the model on 80 percent of the data and retaining the remaining 20 percent for testing purposes. The optimal lag order for the training subsample is $p = 7$ where we use the same criterion and maximum lag order as before. Once again we find this VAR specification to be stable.

Tables 5.1 shows the Root Mean Squared Error (RMSE) for the in-sample and the out-of-sample predictions of both the VAR model and the Deep VAR model. We can see that the RMSE for the Deep VAR is lower than for the conventional VAR for both the training data and the test data and for all time series. The fifth column of the table shows us the ratio between the RMSEs of the Deep VAR and the VAR: the lower the ratio, the better the Deep VAR compared to the VAR. With respect to the training sample, the RMSE of the Deep VAR model is consistently less than 75% of that of the conventional VAR reflecting to some extent the results of the previous sections. Turning to the test data, there is no evidence that the Deep VAR is more prone to

overfitting than the VAR. For both industrial production and unemployment, the Deep VAR yields an RMSE that is around half the size of that produced by the VAR. For inflation and interest rate predictions the out-performance on the test data is less striking, but still large.

Table 5.1: Root mean squared error (RMSE) for the two models across subsamples and variables.

| Sample | Variable | DVAR | VAR | Ratio (DVAR / VAR) |
|--------|----------|---------|---------|--------------------|
| test | IP | 0.00485 | 0.01484 | 0.32703 |
| test | UR | 0.90300 | 1.65170 | 0.54671 |
| test | CPI | 0.00225 | 0.00342 | 0.65892 |
| test | FFR | 0.15743 | 0.23974 | 0.65665 |
| train | IP | 0.00267 | 0.00727 | 0.36737 |
| train | UR | 0.03701 | 0.43322 | 0.08543 |
| train | CPI | 0.00035 | 0.00232 | 0.14925 |
| train | FFR | 0.03658 | 0.25780 | 0.14191 |

5.3 Forecasts

Up until now we have been assessing the model fit, which has provided some initial evidence in favour of Deep VAR. Typically though in the time series context we are more interested in out-of-sample forecasts. which we shall turn to next.

We begin with a single forecasting exercise, where forecasts are produced recursively both for the VAR and the Deep VAR. Specifically, we use the models we trained on the training data to recursively predict one time period ahead, concatenate the predictions to the training data and repeat the process. Note that for the Deep VAR an alternative approach is to work with a different output dimension for the underlying neural networks.⁵

We produce one-year ahead forecasts beginning from the first date in the test sample (November, 2008). Table 5.2 shows the resulting root mean squared forecast errors (RMSFE) along with correlation between forecasts and realizations. As we can see in the table, the RMSFE of the Deep VAR is consistently lower than the one for the VAR. Regarding correlations the VAR produces forecasts that are negatively correlated with actual outcomes for all time series: in other words, when the time series evolves in one direction, the VAR forecast tends to evolve in the opposite direction. For industrial production, the Deep VAR forecast also has a highly negative correlation with the actual values. For the rest of time series the Deep VAR forecasts correlate positively with actual outcome, albeit weakly. Another general observation we made with respect to these forecasts is that the forecasts from the conventional VAR are fairly volatile, while the Deep VAR forecasts swiftly reverts to steady levels (see Figures 9.8 and 9.9 in the appendix).

Table 5.2: Comparison of n-step ahead pseudo out-of-sample forecasts.

| Variable | VAR RMSFE | Deep VAR RMSFE | VAR correlations | Deep-VAR correlations |
|----------|--------------|-------------------|---------------------|--------------------------|
| IP | 0.01870 | 0.01673 | -0.30409 | -0.09175 |
| UR | 0.85984 | 0.73402 | -0.10093 | 0.31968 |

⁵ In future work we plan to assess this approach further.

| Variable | VAR RMSFE | Deep VAR RMSFE | VAR correlations | Deep-VAR correlations |
|----------|--------------|-------------------|---------------------|--------------------------|
| CPI | 0.00946 | 0.00710 | -0.33567 | 0.03954 |
| FFR | 0.52321 | 0.39851 | -0.55935 | -0.01335 |

Finally, we repeat the forecasting exercise above using a rolling window approach: we train our models on a window of 240 months, compute and store 12-month ahead forecasts out of the training sample, roll the window one period forward and repeat the previous steps. This allows us to benchmark the different models in terms of their forecasting performance over the entire sample period. Once again forecasts are for now computed recursively: in other words, neural networks underlying the Deep VAR are not explicitly trained to forecast 12-steps ahead.

In Figure 5.2 we have plotted the cumulative loss incurred by each model: the different output variables are faceted across columns; each row corresponds to a different forecast horizon. For example, the panel in row 2 of column 3 shows the cumulative mean squared error incurred by each model for forecasts up to the 3-month horizon.

While the results are less striking than what we observed above for the in-sample fit, the Deep VAR nonetheless dominates its conventional benchmark overall. For both inflation (CPI) and interest rates (FFR), the Deep VAR forecasts incur substantially lower loss over the entire sample period and in particular at short horizons. We also see somewhat better forecasts overall for industrial production, while for the unemployment rate the Deep VAR is at par with its conventional benchmark. It is not altogether surprising that losses converge at longer horizons since we would expect that forecasts from both autoregressive models converge to their unconditional expectations.

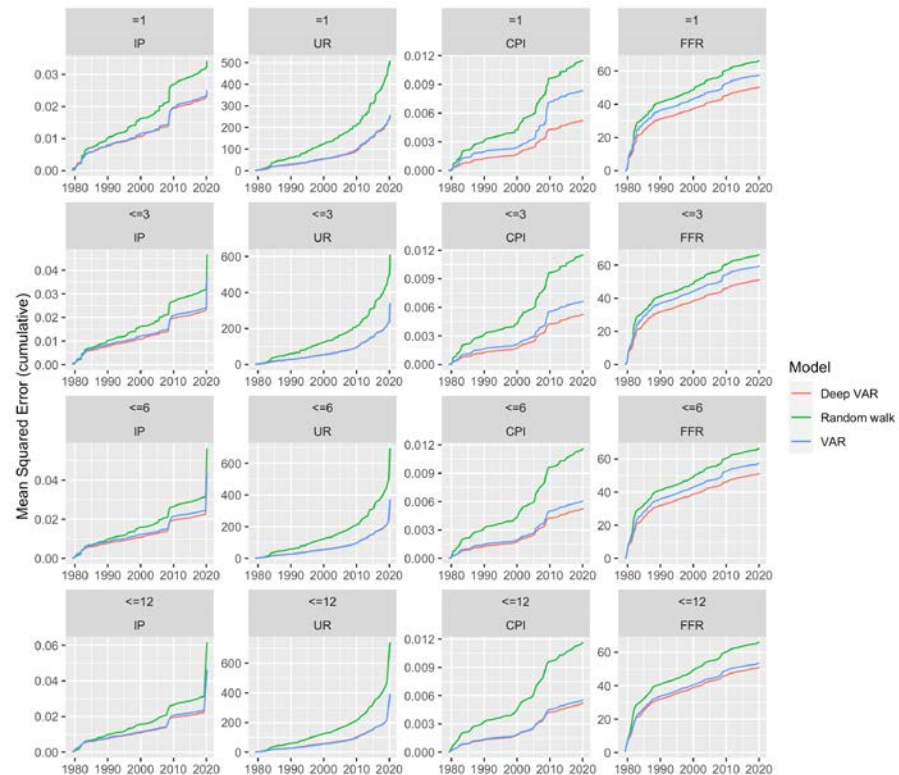


Figure 5.2: Comparison of cumulative rolling-window forecasting error over the entire sample period for Deep VAR and benchmarks. Forecasts are computed recursively.

5.4 Varying hyperparameters

While up until now with respect to model selection we have intentionally remained strictly within the conventional VAR framework, we will now relax that constraint and vary the lag length as well as hyperparameters of the Deep VAR. In particular, we perform a grid search where we vary the number of hidden layers (1,2,5), number of hidden units per layer (50,100,150), the dropout rate (0.3,0.5,0.7) and the lag order (10, 50, 100). For each combination of parameter choices we train the two models and compute the various performance measures introduced above.⁶ Our expectation is that the conventional VAR is prone to overfitting and will produce poor out-of-sample outcomes for higher lag orders. For the Deep VAR we expect to interesting variation in the outcomes for different lag order and hyperparameter choices. It is not clear ex-ante that the Deep VAR should suffer from the same issue of overfitting for higher lag orders. The bulk of the corresponding visualizations can be found in the appendix.

5.4.1 Tuning the Deep VAR

To begin with, we shall forget about benchmarking for a moment and focus on the outcomes for the Deep VAR as we vary parameters. Recall that a higher number of hidden layers (depth), a higher number of hidden units (width) and a smaller choice for the dropout rate all correspond to an increase in neural network complexity. Consistent with this intuition we find that the in-sample loss for the Deep VAR improve as complexity increases (Figure 9.10): higher complexity leads to a reduction in bias and as we noted earlier the underlying recurrent neural networks should in principle be able to model arbitrary functions (Goodfellow, Bengio, and Courville 2016). Conversely, we observe exactly the opposite pattern for out-of-sample loss: as evident from Figure 9.11 a higher choice for the dropout rate and lower choices for the depth and width of the neural networks generally yields a smaller out-of-sample RMSE across variables.

Interestingly, both in- and out-of-sample loss tend to decrease significantly as the number of lags increases. In other words, the Deep VAR seems to be relatively insensitive to overfitting with respect to the lag order. With that in mind, we find that using standard lag order selection tools such as the AIC above may in fact not be appropriate for Deep VARs.

Finally, Figure 9.12 provides an overview of how pseudo out-of-sample forecasting errors behave as we vary the hyperparameters. As before we produce one-year ahead forecasts starting from the end of the 80% training sample. In this context, the pattern is less clear and varies across variables. As the lag order increases, for example, the forecast performance for the unemployment rate deteriorates. For inflation, forecasts are poor for the medium lag choice of $p = 50$ and much better for the low and high lag orders. The exact opposite relationship appears to hold for the Fed Funds Rate. With respect to the choices for the Deep VAR hyperparameters it is difficult to establish any clear pattern at all. The magnitude of differences in RMSFE is

⁶ Of course, with respect to the conventional VAR only the lag order affects outcomes.

generally very small, so overall we conclude that to some extent the variation we do observe may be random.

In light of this evidence, we propose that for the purpose of hyperparameter tuning Deep VAR researchers should focus on the RMSE associated with the 1-step ahead fitted values. For the underlying data, a reasonable set of hyperparameter choices could be: 1 hidden layer, 50 hidden units and a dropout rate of 0.5.

5.4.2 Benchmark

Using the hyperparameter choices proposed above we now turn back to comparing the performance of the Deep VAR to the conventional VAR. Figure 5.3 shows the pseudo out-of-sample RMSE and RMSFE for both models across the different lag choices. For the sake of completeness we also include the performance measures we obtained when we initially ran both models in section 5.2 using the optimal lag order as determined by the AIC.

The first observation is that the Deep VAR outperforms the VAR across the board, reflecting our earlier findings. As expected, the VAR is subject to overfitting for when high lag order are chosen. This trend is observed both for the RMSE as well as the RMSFE. The fact that n -step ahead forecasts of the VAR are also subject to overfitting with respect to the lag order, while the Deep VAR appears unaffected, to some extent may reflect what we observed earlier: for the given data, Deep VAR forecasts swiftly converge to steady levels, while VAR forecasts are volatile, which may explain the relative outperformance of the Deep VAR. It appears that this effect is amplified for higher lag orders.

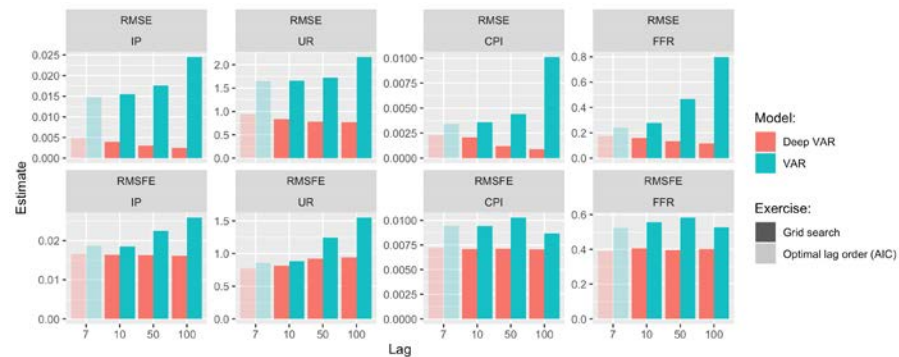


Figure 5.3: Pseudo out-of-sample RMSE and RMSFE for both models across the different lag choices. For the sake of completeness, we also include the performance measures we obtained when we initially ran both models using the optimal lag order as determined by the AIC.

To conclude this empirical section, we summarize our main findings:

1. We provide evidence that the conventional, linear VAR fails to capture important non-linear dependencies across time and variables that are typically used to model the monetary transmission mechanism.
2. Tapping into the broader class of Deep VAR leads to consistently better model performance.
3. Deep VAR appears to be relatively insensitive to very high lag orders at which conventional VAR models are prone to overfitting.

6 Caveats and extensions

In this work we have provided empirical evidence that the introduction of deep learning can lead to improved modelling and forecasting performance in the context of macroeconomic time series data. While we believe that our proposed methodology extends the conventional VAR framework quite naturally, it still comes with a lot of added complexity. Unfortunately, in the case of deep learning this added complexity also entails reduced interpretability: even though we have intentionally worked with a relatively small and simple neural network architecture, the number of parameters and interactions between neurons that they govern cannot possibly be interpreted by a human. This is why deep artificial neural networks are commonly referred to as black boxes.

Perhaps more importantly in the context of time series forecasting, it is also much harder to quantify predictive uncertainty of deep neural networks: while confidence intervals around point forecasts from a linear VAR can be computed using closed-form analytical expressions (Kilian and Lütkepohl 2017), no such expressions exist in the context of Deep VAR. Future work on this issue will most likely rely on probabilistic deep learning, which has gained popularity in recent years. Among the most widely used approaches to uncertainty quantification for deep learning are deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2016) and Monte Carlo dropout (Gal and Ghahramani 2016). The former boils down to training not just one but multiple networks and effectively averaging over predictions: since weights are initialized randomly, predictions are stochastic. The latter similarly introduces stochasticity by activating dropout not only during training but also at the testing stage. A common drawback of these and other approaches that rely on Monte Carlo is the increased computational burden. As an alternative to Monte Carlo Daxberger et al. (2021) have recently shown that Laplace approximation can be used for effortless Bayesian deep learning.

Support for the estimation of impulse response functions is another missing cornerstone in the current version of our proposed framework. IRFs are used to understand how system variables change in response to unit shocks to any of the system variables. When estimating the model with the traditional VAR, IRFs can be readily derived from the reduced form model coefficients. Generalized (or structural) IRFs require the system to be fully identified, which is typically achieved through restrictions on contemporaneous (and likely correlated) reduced-form errors. In the context of Deep VAR further research is required concerning both computation of IRFs and the identification problem. Verstyuk (2020) computes impulse response functions for their proposed MLSTM numerically and relies on a Cholesky decomposition of the reduced-form covariance matrix, just like in the conventional setting. A more desirable approach may once again involve probabilistic deep learning: Ish-Horowicz et al. (2019) proposes a straight-forward approach towards producing global feature importance measures for input features of Bayesian neural networks. It might be possible to leverage these importance measures as proxies for the conventional VAR's linear coefficients and produce approximate impulse response functions for Deep VAR models in the same way as for conventional VAR models. Of course, these are merely rough ideas for future research.

7 Conclusions

Our initial motivation for this study was to see if by incorporating some of the latest developments from the machine learning and deep learning domains in the conventional VAR framework, we could attain improvements in the modelling and forecasting performance. In an effort not to deviate too much from the established framework, we only relax one single assumption to move from the conventional linear VAR to a broader class of models that we refer to as Deep VAR models.

To assess the modelling performance of Deep VAR models compared to linear VAR models we investigate a sample of monthly US economic data in the period 1959-2021. In particular, we look at variables typically analysed in the context of the monetary transmission mechanism including output, inflation, interest rates and unemployment. Our empirical findings show a consistent and significant improvement in modelling performance associated with Deep VAR models. In particular, our proposed Deep VAR produces much lower cumulative loss measures than the VAR over the entire period and for all of the analysed time series. The improvements in modelling performance are particularly striking during subsample periods of economic downturn and uncertainty. This appears to confirm or initial hypothesis that by modelling time series through Deep VAR models it is possible to capture complex, non-linear dependencies that seem to characterize periods of structural economic change.

When it comes to the out-of-sample performance, a priori it may seem that the Deep VAR is prone to overfitting, since it is much less parsimonious than the conventional VAR. On the contrary, we find that by using default hyperparameters the Deep VAR clearly dominates the conventional VAR in terms of out-of-sample prediction and forecast errors. An exercise in hyperparameter tuning shows that its out-of-sample performance can be further improved by appropriate regularization through adequate dropout rates and appropriate choices for the width and depth of the neural. Interestingly, we also find that the Deep VAR actually benefits from very high lag order choices at which the conventional VAR is prone to overfitting. In summary, we provide solid evidence that the introduction of deep learning into the VAR framework can be expected to lead to a significant boost in overall modelling performance. With respect to the main question posed at the beginning of this work, we therefore conclude that deep learning may be leveraged effectively in the context of macroeconomic time series modelling and vector autoregression.

We also point out several shortcomings of our proposed Deep VAR framework, which we believe can be alleviated through future research. In particular, policy-makers are typically concerned with uncertainty quantification, inference and overall model interpretability. Future research on Deep VAR models should therefore address the estimation of confidence intervals, impulse response functions as well as variance decompositions typically analysed in the context of VAR models. We point to a few possible avenues that involve probabilistic deep learning. We very much recognize the need for model interpretability especially in the context of policy-making and believe that the Deep VAR framework proposed here can be augmented to meet these demands.

References

- Bernanke, Ben S. 1990. "The Federal Funds Rate and the Channels of Monetary Transmission." National Bureau of Economic Research Cambridge, Mass., USA.
- Brock, William Allen, William A Brock, David Arthur Hsieh, Blake Dean LeBaron, and William E Brock. 1991. *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence*. MIT press.
- Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3): 273–97.
- Daxberger, Erik, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. 2021. "Laplace Redux-Effortless Bayesian Deep Learning." *Advances in Neural Information Processing Systems* 34.
- Dorffner, Georg. 1996. "Neural Networks for Time Series Processing." In *Neural Network World*. Citeseer.
- Fix, E, and J Hodges. 1951. "An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation." *International Statistical Review* 3 (57): 233–38.
- Friedman, Milton, and Anna Jacobson Schwartz. 2008. *A Monetary History of the United States, 1867-1960*. Vol. 14. Princeton University Press.
- Gal, Yarin, and Zoubin Ghahramani. 2016. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In *International Conference on Machine Learning*, 1050–59. PMLR.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
- Goodfriend, Marvin, and Robert G King. 2005. "The Incredible Volcker Disinflation." *Journal of Monetary Economics* 52 (5): 981–1015.
- Greene, William H. 2012. "Econometric Analysis, 71e." *Stern School of Business, New York University*.
- Hamilton, James Douglas. 2020. *Time Series Analysis*. Princeton university press.
- Hamzaçebi, Coşkun. 2008. "Improving Artificial Neural Networks' Performance in Seasonal Time Series Forecasting." *Information Sciences* 178 (23): 4550–59.
- Ho, Tin Kam. 1995. "Random Decision Forests." In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–82. IEEE.
- Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80.
- Ish-Horowicz, Jonathan, Dana Udwin, Seth Flaxman, Sarah Filippi, and Lorin Crawford. 2019. "Interpreting Deep Neural Networks Through Variable Importance." *arXiv Preprint arXiv:1901.09839*.
- Joseph, Andreas, Eleni Kalamara, George Kapetanios, and Galina Potjagailo. 2021. "Forecasting Uk Inflation Bottom Up."
- Kihoro, J, RO Otieno, and C Wafula. 2004. "Seasonal Time Series Forecasting: A Comparative Study of ARIMA and ANN Models."
- Kilian, Lutz, and Helmut Lutkepohl. 2017. *Structural Vector Autoregressive Analysis*. Cambridge University Press.
- Kingma, Diederik P, and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." *arXiv Preprint arXiv:1412.6980*.
- Kydland, Finn E, and Edward C Prescott. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica: Journal of the Econometric Society*, 1345–70.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2016. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles." *arXiv Preprint arXiv:1612.01474*.

Lucas, JR. 1976. "Econometric Policy Evaluation: A Critique", in k. Brunner and a Meltzer, the Phillips Curve and Labor Markets, North Holland."

Lütkepohl, Helmut. 2005. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.

Masini, Ricardo P, Marcelo C Medeiros, and Eduardo F Mendes. 2021. "Machine Learning Advances for Time Series Forecasting." *Journal of Economic Surveys*.

McCracken, Michael W, and Serena Ng. 2016. "FRED-MD: A Monthly Database for Macroeconomic Research." *Journal of Business & Economic Statistics* 34 (4): 574–89.

McCulloch, Warren S, and Walter Pitts. 1990. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biology* 52 (1): 99–115.

Olah, Chris. 2015. "Understanding LSTM Networks." <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>.

Perry, George L, and James Tobin. 2010. *Economic Events, Ideas, and Policies: The 1960s and After*. Brookings Institution Press.

Romer, Christina D, and David H Romer. 1989. "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." *NBER Macroeconomics Annual* 4: 121–70.

Sims, Christopher A et al. 1986. "Are Forecasting Models Usable for Policy Analysis?" *Quarterly Review* 10 (Win): 2–16.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *The Journal of Machine Learning Research* 15 (1): 1929–58.

Verstyuk, Sergiy. 2020. "Modeling Multivariate Time Series in Economics: From Auto-Regressions to Recurrent Neural Networks." *Available at SSRN 3589337*.

Zhang, G Peter. 2003. "Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model." *Neurocomputing* 50: 159–75.

Zhang, Guoqiang, B Eddy Patuwo, and Michael Y Hu. 1998. "Forecasting with Artificial Neural Networks:: The State of the Art." *International Journal of Forecasting* 14 (1): 35–62.

Appendix

8 Tables

9 Figures

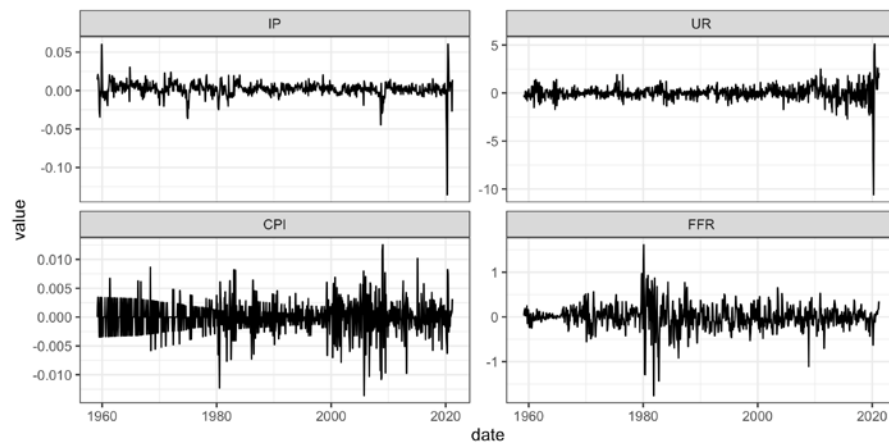


Figure 9.1: Time series

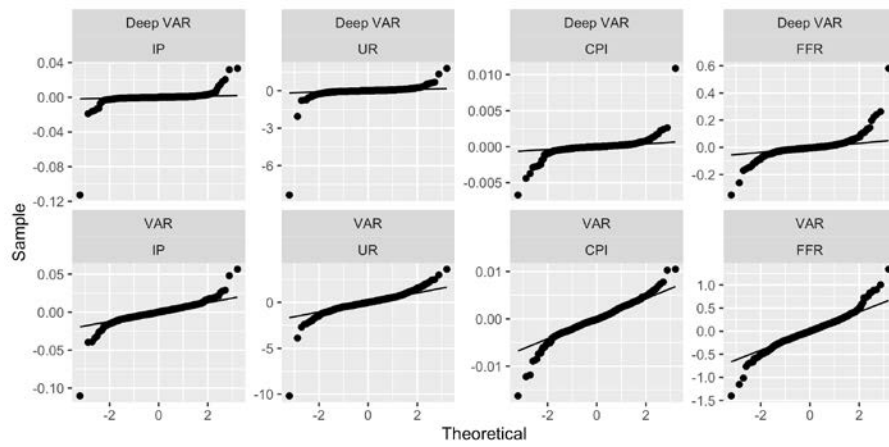


Figure 9.2: Quantile-quantile plots of full-sample residuals.

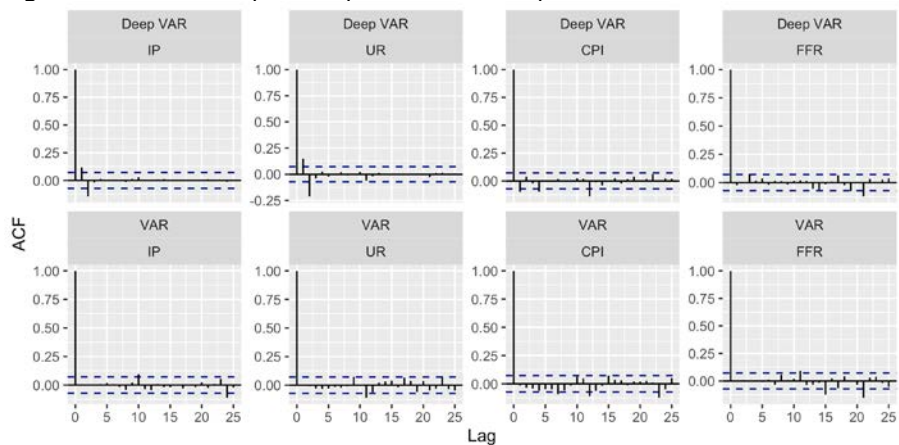


Figure 9.3: ACF plots of full-sample residuals.

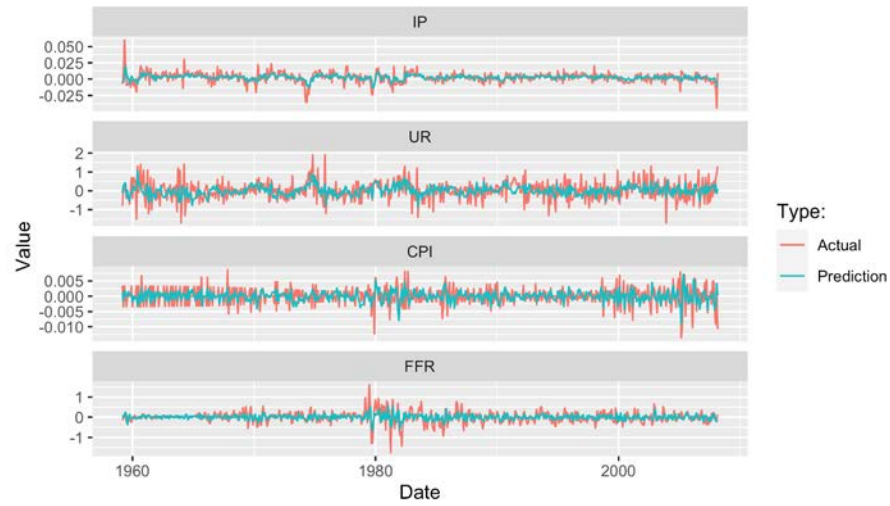


Figure 9.4: VAR fitted values plotted against observed values for the training sample.

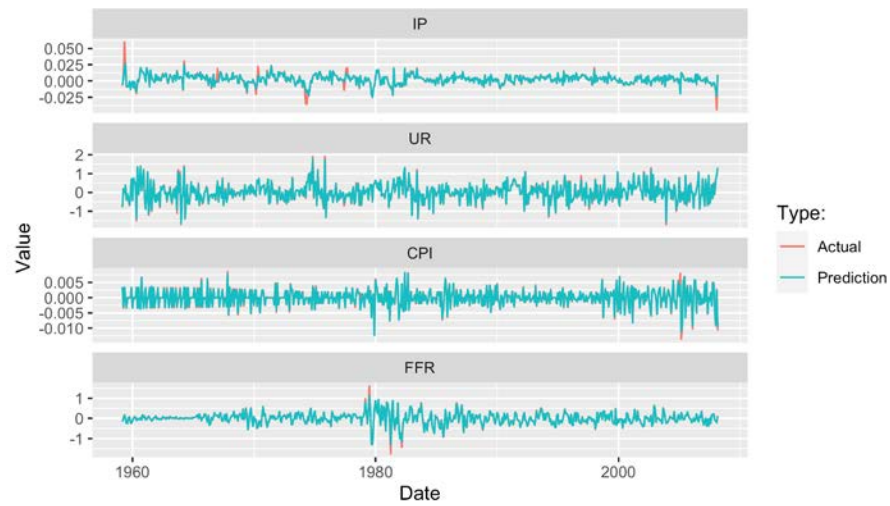


Figure 9.5: Deep VAR fitted values plotted against observed values for the training sample.

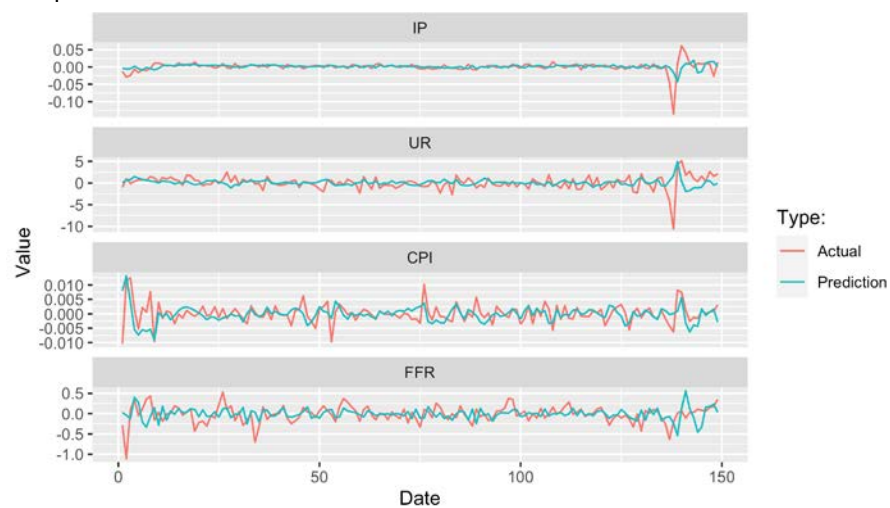


Figure 9.6: VAR fitted values plotted against observed values for the test sample.

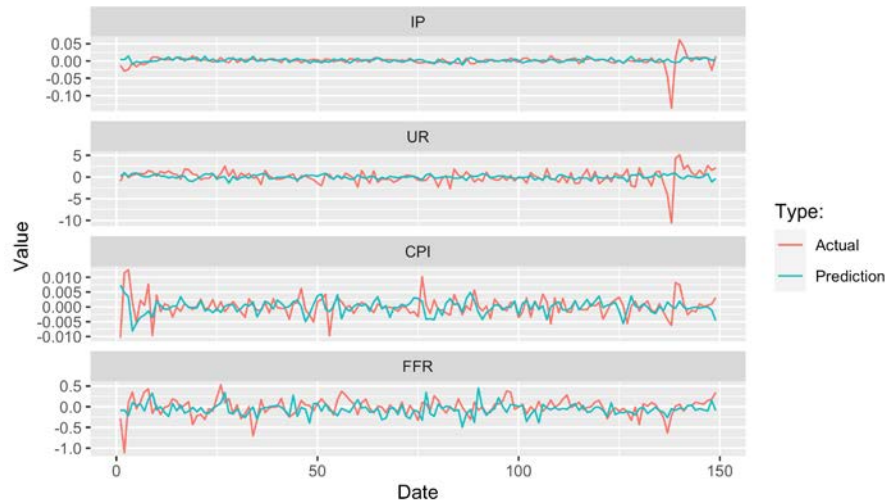


Figure 9.7: Deep VAR fitted values plotted against observed values for the test sample.

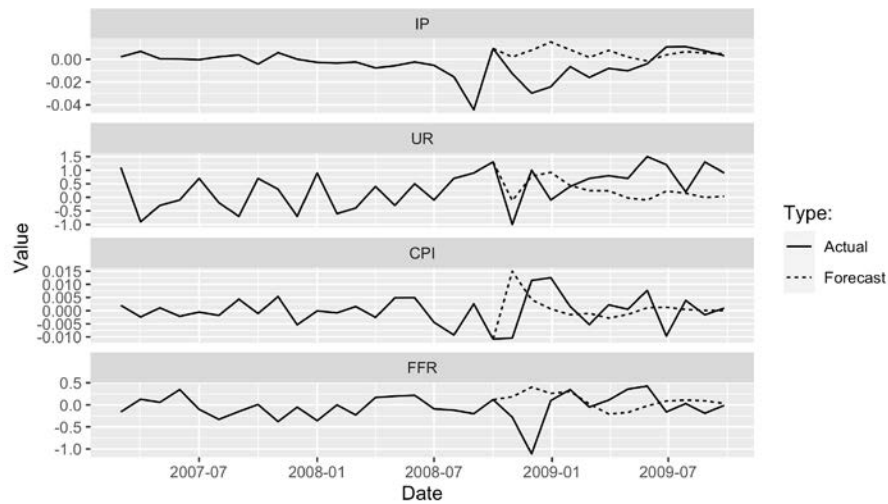


Figure 9.8: VAR n-step ahead forecasts plotted against observed values. Forecasts are for the first year of the test sample.

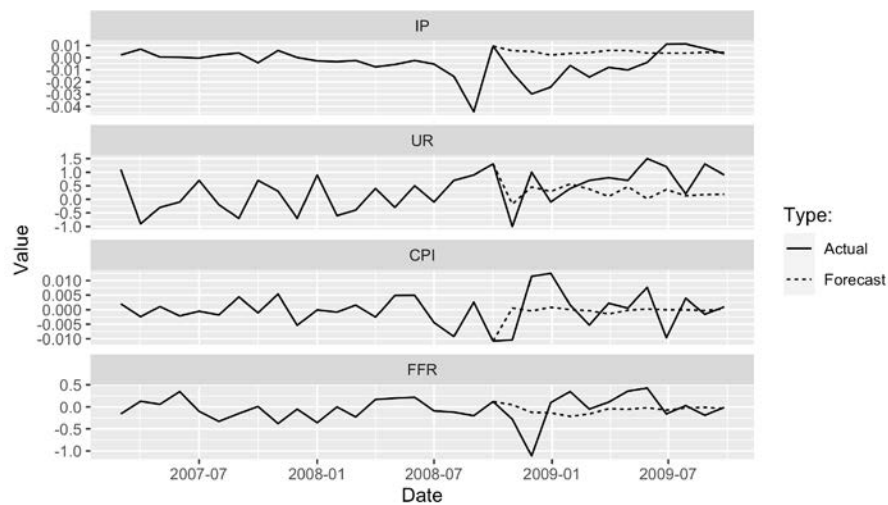


Figure 9.9: Deep VAR n-step ahead forecasts plotted against observed values. Forecasts are for the first year of the test sample.

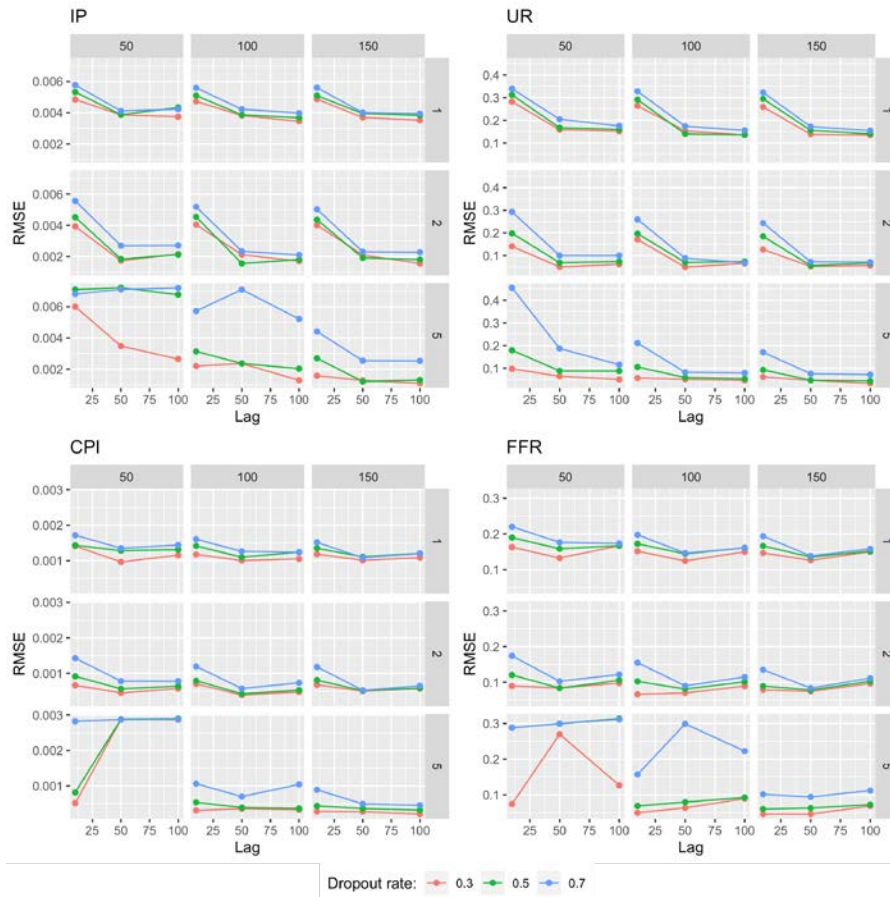


Figure 9.10: Train sample RMSE for Deep VAR for different variables.

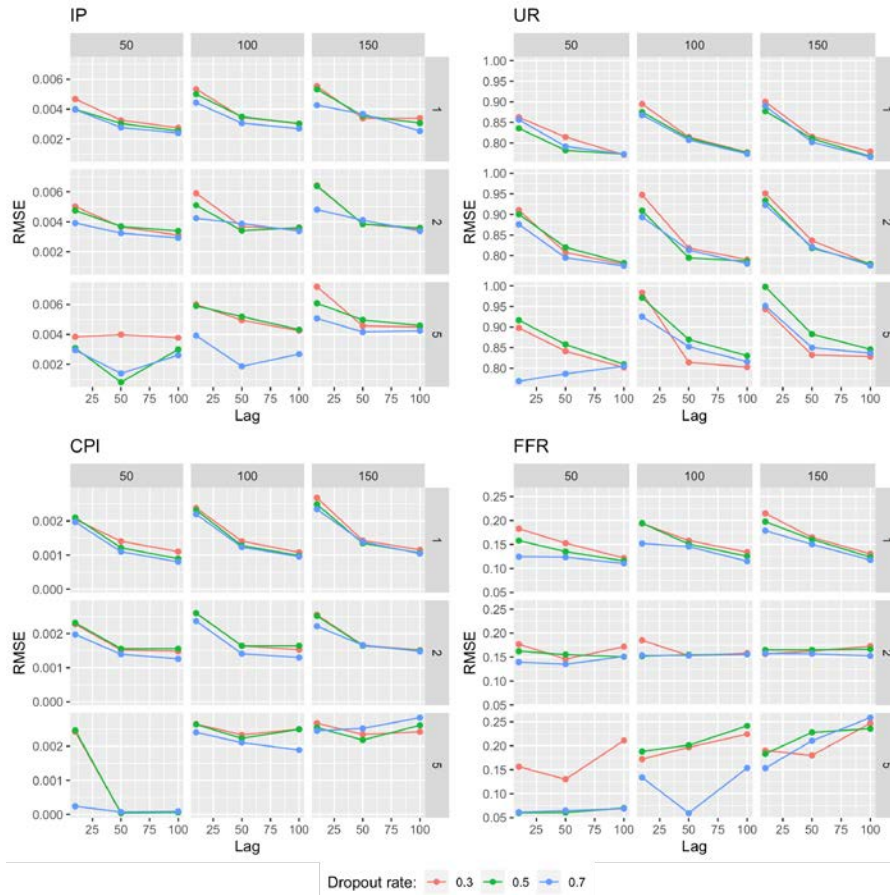


Figure 9.11: Test sample RMSE for Deep VAR for different variables.

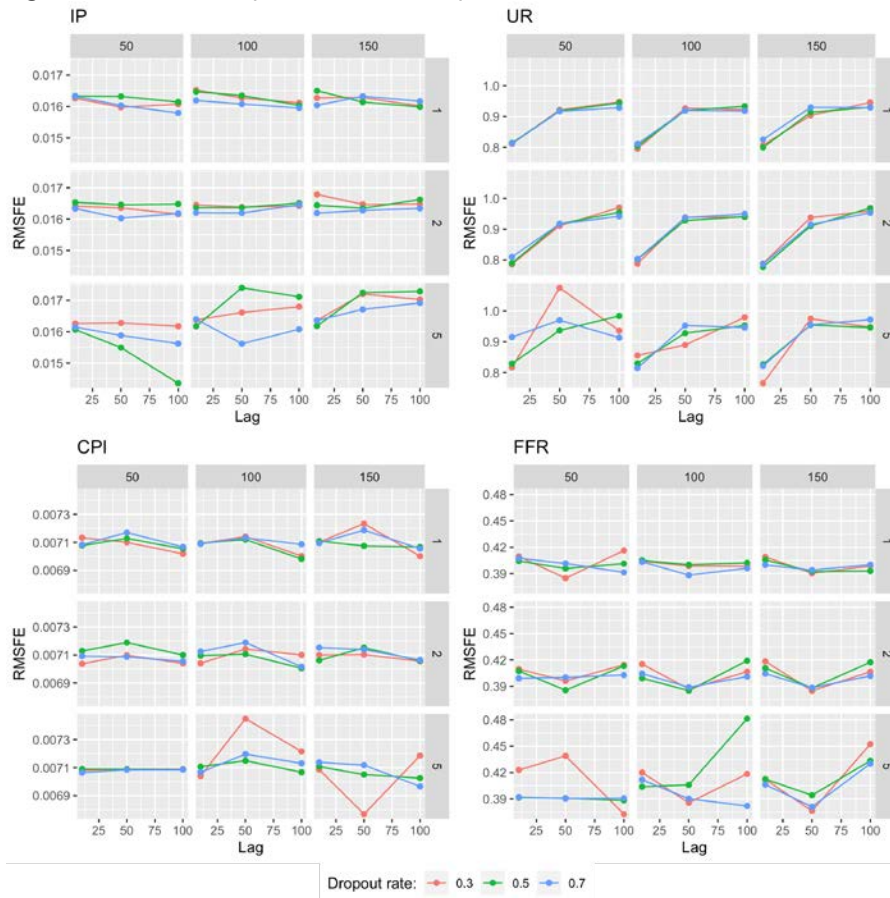


Figure 9.12: Pseudo out-of-sample RMSFE for Deep VAR for different variables.

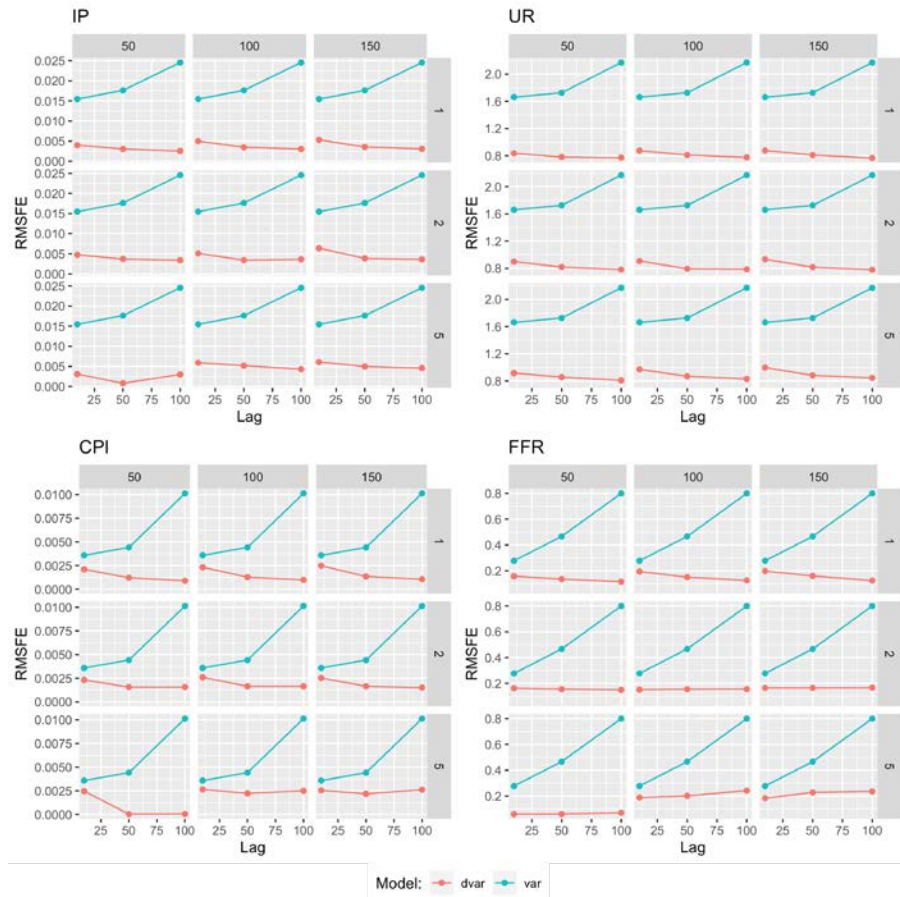


Figure 9.13: Comparison of out-of-sample RMSE for conventional VAR and Deep VAR for different variables.

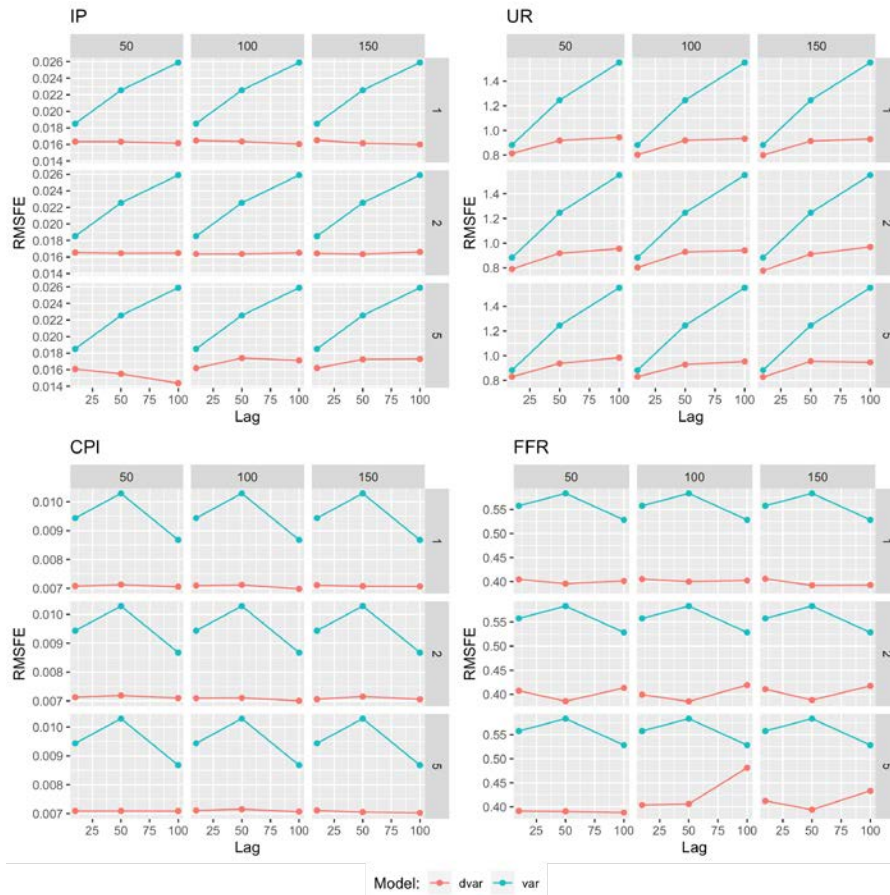


Figure 9.14: Comparison of pseudo out-of-sample RMSFE for conventional VAR and Deep VAR for different variables.

10 Code and Package

All code used for the empirical analysis presented in this article can be found on the corresponding GitHub repository. Researchers interested in using Deep VARs more generally for their own empirical work may find the R `deepvars` package useful which is being maintained by one of the authors. The package is still under development and as of now only available on GitHub. To install the package in R simply run:

```
devtools::install_github("pat-alt/deepvars", build_vignettes=TRUE)
```

Package vignettes will take you through the basic package functionality. Once the package has been installed simply run `utils::browseVignettes()` to access the documentation.

Deep Vector Autoregression for Macroeconomic Data

Patrick Altmeyer¹ Marc Agusti² Ignacio Vidal-Quadras
Costa³

31 January, 2022

¹Delft University of Technology, p.altmeyer@tudelft.nl

²European Central Bank, marc.agusti_i_torres@ecb.europa.eu

³European Central Bank, ignacio.vidalquadrascosta@barcelonagse.eu

Motivation

Can we leverage the power of deep learning in VAR models?

- ▶ We propose **Deep VAR**: a novel approach towards VAR that leverages the power of deep learning in order to model non-linear relationships.
- ▶ Worked under the following premise: **maximize performance** of an existing and trusted framework under **minimal intervention**.
- ▶ We maintain the additive structure of the VAR, but relax the assumption of linearity by modelling each equation of the VAR system as a recurrent neural network.
- ▶ By staying methodologically as close as possible to the original benchmark, we hope that our approach is more likely to find acceptance in the economics domain.

Key contributions

- ▶ Simple methodology close in spirit to conventional benchmark.
- ▶ Significant improvement in model fit and forecasting accuracy.
- ▶ Open source R package `deepvars` to facilitate reproducibility.

Work-in-progress:

- ▶ Master's thesis was selected for publication by Universitat Pompeu Fabra.
- ▶ Feedback rounds with Eddie Gerba (Bank of England, LSE) and Chiara Osbat (ECB).
- ▶ Presented an updated version of the paper at NeurIPS 2021 MLECON workshop in December.

Previous literature

- ▶ Non-linear dependencies are likely to form part of the data generating process of variables commonly used to model the monetary transmission mechanism (Brock et al. 1991).
- ▶ A range of machine learning models has previously been used in the context of time series forecasting Kihoro, Otieno, and Wafula (2004). Deep learning has been shown to be particularly successful at capturing non-linearities G. P. Zhang (2003).
- ▶ Joseph et al. (2021) review both machine learning and deep learning methods for forecasting inflation and find that neural networks in particular are useful for forecasting especially at a longer horizon.

Methodology

- Relax the assumption of linearity and instead model the process as system of potentially highly non-linear equations:

$$y_{it} = f_i(\mathbf{y}_{t-1:t-p}; \theta) + v_{it} \quad , \quad \forall i = 1, \dots, K \quad (1)$$

- Each single variable in model is modelled as a recurrent neural network:

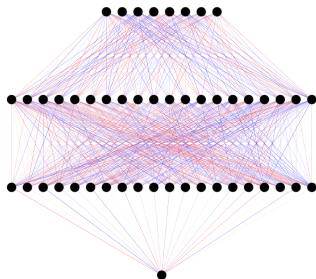


Figure 1: Neural Network Architecture.

Data

- ▶ To evaluate our proposed methodology empirically we use the FRED-MD data base to collect a sample of monthly US data on:
 - ▶ output (IP)
 - ▶ unemployment (UR)
 - ▶ inflation (CPI)
 - ▶ interest rates (FFR)
- ▶ Our sample spans the period from January 1959 through March 2021.

Model fit

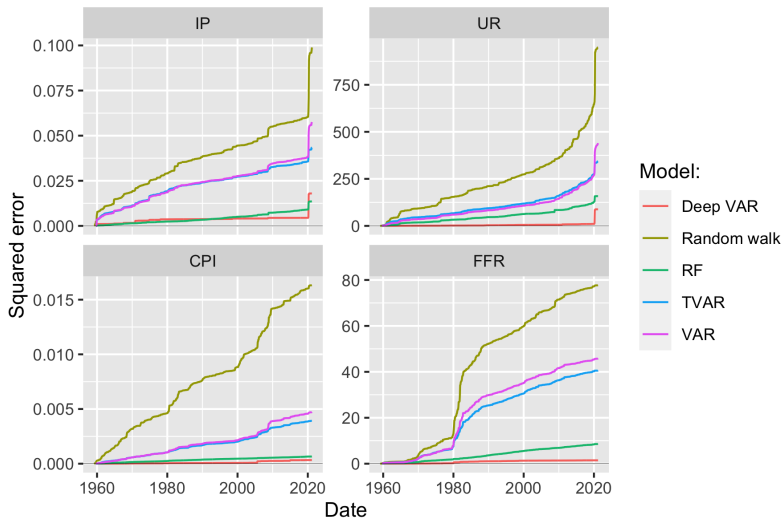


Figure 2: Comparison of cumulative loss over the entire sample period for Deep VAR and benchmarks.

Forecasting

Question: recursive forecasts like in conventional VAR or training on n outputs?

- ▶ Initially we opted for the former approach and provided anecdotal evidence that Deep VAR outperforms
- ▶ Have since tested this more rigorously using rolling window:
 - ▶ Deep VAR still outperforms VAR especially at short horizon
 - ▶ Currently investigating if training on n outputs provides additional edge.

Concluding remarks

- ▶ Simple framework that relies on the premise of minimal intervention in the conventional and trusted framework.
- ▶ Deep learning appears to do a good job at capturing non-linear dependencies.

But...

- ▶ Added complexity is (often) coupled with lack of interpretability:
 - ▶ No analytical expressions for impulse response functions and variance decompositions
 - ▶ Verstyuk (2020) manages to recover IRFs numerically; should be readily applicable to our Deep VAR framework.
- ▶ Uncertainty estimation can be done through Bayesian methods: deep ensemble, MC dropout, Variational Inference:
 - ▶ All of the above entail an added layer (layers really!) of computational complexity.
 - ▶ Laplace Redux for effortless Bayesian Deep Learning (Daxberger et al. 2021) holds promise, but not yet implemented.

Your questions and comments

References I

- Brock, William Allen, William A Brock, David Arthur Hsieh, Blake Dean LeBaron, and William E Brock. 1991. *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence*. MIT press.
- Daxberger, Erik, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. 2021. "Laplace Redux-Effortless Bayesian Deep Learning." *Advances in Neural Information Processing Systems* 34.
- Hamzaçebi, Coşkun. 2008. "Improving Artificial Neural Networks' Performance in Seasonal Time Series Forecasting." *Information Sciences* 178 (23): 4550–59.
- Joseph, Andreas, Eleni Kalamara, George Kapetanios, and Galina Potjagailo. 2021. "Forecasting Uk Inflation Bottom Up."
- Kihoro, J, RO Otieno, and C Wafula. 2004. "Seasonal Time Series Forecasting: A Comparative Study of ARIMA and ANN Models."

References II

- Olah, Chris. 2015. "Understanding LSTM Networks." <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>.
- Verstyuk, Sergiy. 2020. "Modeling Multivariate Time Series in Economics: From Auto-Regressions to Recurrent Neural Networks." *Available at SSRN 3589337*.
- Zhang, G Peter. 2003. "Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model." *Neurocomputing* 50: 159–75.
- Zhang, Guoqiang, B Eddy Patuwo, and Michael Y Hu. 1998. "Forecasting with Artificial Neural Networks:: The State of the Art." *International Journal of Forecasting* 14 (1): 35–62.

Hiddens

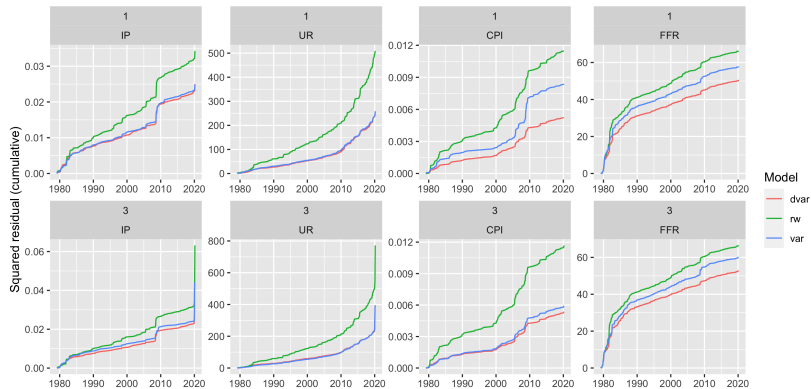
Long Short-Term Memory

- ▶ The most common choice of neural networks architectures for modelling persistent time series is the LSTM:

“The LSTM [has] the ability to remove or add information to the cell state, carefully regulated by structures called gates. Gates are a way to optionally let information through.” — Olah (2015)

$$\begin{aligned}\mathbf{f}_t &= \sigma(\mathbf{b}_f + \mathbf{W}_f \mathbf{h}_{t-1} + \mathbf{U}_f \mathbf{h}_{-1}) \\ \mathbf{i}_t &= \sigma(\mathbf{b}_i + \mathbf{W}_i \mathbf{h}_{t-1} + \mathbf{U}_i \mathbf{h}_{-1}) \\ \mathbf{o}_t &= \sigma(\mathbf{b}_o + \mathbf{W}_o \mathbf{h}_{t-1} + \mathbf{U}_o \mathbf{h}_{-1}) \\ \mathbf{C}_t &= \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{b}_C + \mathbf{W}_C \mathbf{h}_{t-1} + \mathbf{U}_C \mathbf{h}_{-1}) \\ \mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{C}_t) \\ \hat{y}_{it} &= c + \mathbf{v}^T \mathbf{h}_t\end{aligned}\tag{2}$$

Rolling window forecasts



IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

What should be the optimal financial structure of the FDI inflows to Poland in stimulating growth processes?¹

Aneta Kosztowniak,
Narodowy Bank Polski

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

What should be the optimal financial structure of the FDI inflows to Poland in stimulating growth processes - experience in the years 2004-2021

Aneta Kosztowniak, Economic Analysis and Research Department, Narodowy Bank Polski, SGH Warsaw School of Economics

Abstract

The article is aimed at answering a question what should be the optimal financial structure of the FDI inflows to Poland for stimulating growth processes. The study on the dependence between financial components of FDI inflows and GDP for Poland covers the period 2004:Q1-2021:Q3. Define the structure of FDI would be the added value of this work. Among the important questions are the following: 1) How the components of the FDI inflows influence the degree of explaining GDP changes? 2) Which components of the FDI inflows explain to the greatest extent changes in GDP? 3) Which components of the FDI inflows explain to the lesser extent changes in GDP? 4) What should be the optimal FDI inflows structure to Poland that maximizes the positive impact on growth processes? The following thesis has been put forward: The optimal structure of FDI inflows to Poland is a structure that maximizes equities and reinvestment of earnings with less participation of debt instruments and dividend payments outside the host country. Results of the VAR/VECM model and forecast error variance decomposition (FEVD) indicate that in the optimal (growth-enhancing) structure of the FDI inflows, the share of equities and the reinvestment of earnings should be maximized, while the shares of debt securities and dividend payments could be rather minimized.

Keywords: FDI inflows (equity, reinvestment of earnings, debt instruments), GDP, EU, Poland, VECM, the impulse function, the variance decomposition

JEL classification: C82, F21

Contents

| | |
|---|----|
| What should be the optimal financial structure of the FDI inflows to Poland in stimulating growth processes - experience in the years 2004-2021 | 1 |
| 1. Introduction | 3 |
| 2. Financial instruments of FDI inflows - methodology | 3 |
| 3. Empirical data – structure of FDI inflows in EU countries..... | 6 |
| 4. Data and research procedure of cause-and-effect relationships between components of FDI inflows and GDP in Poland | 8 |
| 5. Results – VECM model, the impulse response functions and the decomposition of variance | 9 |
| 5.1. The impulse response functions | 10 |
| 5.2. The variance decomposition | 10 |
| 6. Conclusion and way forward | 13 |
| References..... | 13 |

1. Introduction

Most economic theories emphasize the beneficial influence of FDI on economic growth, while empirical research by both international financial institutions and economists shows inconclusive results. International standards International Monetary Fund (IMF, 2009), Organisation for Economic Co-operation and Development (OECD, 2008) relate to the methodology of estimating the value of FDI (inflows/ outflows, stocks) and its components, qualification to the financial account in terms of the balance of payments or the calculation of the international investment position. This methodology addresses the important quantitative dimension of FDI statistics.

However, the expansion of FDI in the global economy has continued uninterruptedly since the 1950s.-1960s. XX-th century, that is for over 70 years. According to the author, it is worth extending the current analytical dimension of FDI – mainly quantitative and aggregate (a monolithic rather than multidimensional variables) – about a new pro-growth dimension of FDI at the level of calculations by international financial institutions. The added value of the analysis of the impact of financial structure of FDI inflows (equity, reinvestment of earnings and debt instruments) on stimulating economic growth may be the identification of the importance of individual components in terms of their impact strength and impact period. Thus, these results may, at least in part, explain the ambiguous effects of empirical research for FDI aggregates.

Estimating the pro-growth dimension of FDI is important for investor and host countries (which the analysis focuses on). The expectations of the FDI's pro-growth (efficiency / quality) effects, for example in the Central and Eastern Europe (CEE) countries, including Poland, result from many challenges, such as: filling the shortages of capital accumulation (gross fixed capital formation), increasing productivity, stimulating innovation and technology transfer or real convergence processes. The presence of foreign investors in the host market is associated with influencing the labor market, the structure of the economy or the tax incomes.

For countries such as Poland or other CEE countries, which are still "catching up" (with varying intensity) Western European countries, it's important not only value the FDI inflows (quantity) but also the possibility of forecasting the impact of a specific type of investment (material or financial) on economic growth (share of growth/ productivity of capital or per employee). Moreover, the ability to estimate this multidimensional nature of the FDI inflows for GDP dynamics is awaited.

2. Financial instruments of FDI inflows - methodology

The direct investment enterprise denotes an enterprise in which direct investor owns at least 10% of the voting power in the decision making body of the company. The direct investment capital comprises equity capital in the form of shares and other equity, reinvestment of earnings and assets and liabilities vis-à-vis debt instruments. According to OECD (2008) direct investment transactions are all transactions between direct investors, direct investment enterprises, and/or other fellow enterprises. For transactions, the asset/liability principle is schematically shown as follows (Table 1).

Equity, other than reinvestment of earnings comprises equity in branches, all shares in subsidiaries and associates (except non – participating, preferred shares that are treated as debt securities and included under direct investment, debt instruments) and other contributions of an equity nature.

Reinvestment of earnings denote the part of profits, accruing to a direct investor, which remains in the direct investment enterprise, and which is allocated to its further development. According OECD (2008) *reinvestment of earnings* of direct investment enterprises (items A1.2 and L1.2) reflects earnings accruing to direct investors (that is, proportionate to the ownership of equity) during the reference period less earnings declared for distribution in that period.

According OECD (2008) **reinvestment of earnings** of direct investment enterprises (items A1.2 and L1.2) reflects earnings accruing to direct investors (that is, proportionate to the ownership of equity) during the reference period less earnings declared for distribution in that period. Earnings are included in direct investment income because they are deemed to accrue to the direct investor, whether they are reinvested in the direct investment enterprise or remitted to the direct investor. However, reinvested earnings are not actually transferred to the direct investor but rather increase the direct investor's investment in its direct investment enterprise. Therefore, an entry that is equal to that made in the direct investment income account but that flows in the opposite direction is made in the direct investment financial transactions account. In the direct investment income account, this form of income is referred to as "reinvested earnings" (ECB, 2019). However, in the direct investment transactions account, "reinvestment of earnings" is the term that is used, to more clearly differentiate between the income and financial transactions. Moreover, in cases where the equity asset holder has less than 10% voting power (*reverse investment* and investment in fellow enterprises), reinvested earnings and reinvestment of earnings are not recorded.

Debt instruments mean all forms of investing other than the acquisition of shares or equity, or reinvestment of earnings associated with such shares or equities. Debt instruments include, among others, credits and loans, debt securities and other unsettled payments between entities in direct investment relationship (OECD 2018) (Table 1).

It is worth adding that the structure of the FDI inflow in host countries depends on many conditions. On the side of a foreign investor, this structure depends on the investment strategy and dividend policy (including superdividends) (Borga 2018). On the side of the host country, location conditions are important, including: the existing structure of the economy, absorption possibilities or the so-called broadly understood investment climate.

Paying attention to the financial structure of the FDI can be found, e.g. in studies of central bank employees in the context of FDI life cycle phases (Brada and Tomšík, 2003, Novotný, Podpira 2008; Novotný, 2015, 2018) or forecasting direct investment equity income for the balance of payment (Dagmaard et al., 2010, Knetsch and Nagengast 2016).

FDI transactions according to the asset/liability principle

Table 1

| Of direct investors in direct investment enterprises | Of direct investment enterprises to direct investors |
|--|--|
| A1. Equity | L1. Equity |
| A1.1. Equity transactions | L1.1. Equity transactions |
| A1.2. Reinvestment of earnings | L1.2. Reinvestment of earnings |
| A2. Debt instruments | L2. Debt instruments |
| Of direct investment enterprises in direct investors | Of direct investment enterprises in direct investors |
| -Reverse investment | - Reverse investment |
| A3. Equity | L3. Equity |
| A4. Debt instruments | L4. Debt instruments |
| In fellow enterprises | To fellow enterprises |
| A5. Equity | L5. Equity |
| A5.1. If ultimate controlling parent is resident | L5.1. If ultimate controlling parent is non-resident |
| A5.2. If ultimate controlling parent is non-resident | L5.2. If ultimate controlling parent is resident |
| A6. Debt instruments | L6. Debt instruments |
| A6.1. If ultimate controlling parent is resident | L6.1. If ultimate controlling parent is non-resident |
| A6.2. If ultimate controlling parent is non-resident | L6.2. If ultimate controlling parent is resident |

Sources: OECD (2008, p. 70-71).

In the theory of the FDI profitability life cycle, according to Brada and Tomšík (2003) shows:

- stage 1 (entry) is connected with expenditures of foreign investors in a host country and means negative profitability (increased equity);
- at stage 2 (growth), profit peaks at around the 6th year of the cycle (this means increased reinvested earnings and debt instruments);
- stage 3 (investment repatriation) relates to distribution of profits and dividend payment (a part of reinvested earnings can be transferred abroad and increase debt instruments, or disinvestments).

According to studies by Novotný and Podpiera (Novotný and Podpiera 2008, Novotný 2015), among the countries of Central and Eastern Europe, the full-fledged FDI life cycle usually covers 15 years, followed by projections toward zero (null) annual profitability.

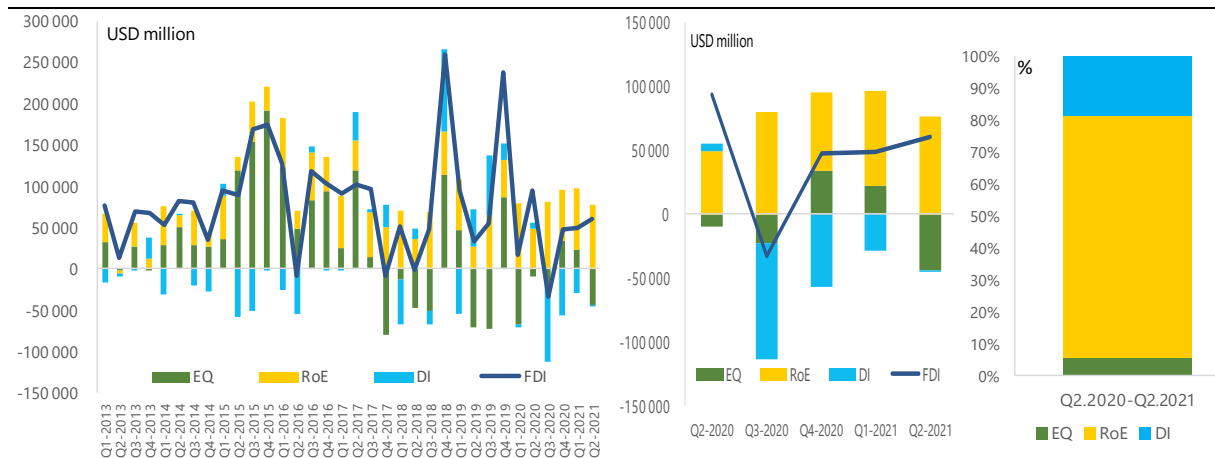
Moreover, many central bank analysts involved in forecasting the FDI components for the purposes of balance of payments estimates. Damgaard et al. (2010) from Danmarks Nationalbank for Denmark or Novotný (2015) from Czech National Bank for Czech Republic, are interested in FDI components. Knetsch and Nagengast (2016) from Deutsche Bundesbank forecasting FDI outflows in order to assess potential tax revenues of state budgets of countries in which transnational corporations have their headquarters.

3. Empirical data – structure of FDI inflows in EU countries

The financial structure of FDI inflows in the EU countries presented on a quarterly basis for the years 2013-2021, indicate their significant fluctuations. Nevertheless, the positive values were mainly due to equity and reinvestment of earnings (2013-2017), with negative value debt instruments. It can be added that debt instruments were used as highly liquid instruments (more mobile than others) to regulate liquidity between e.g. the parent company and subsidiaries. During the COVID-19 pandemic (Q2.2020-Q3.2022), the structure of FDI inflows was generally maintained.

The financial structure of FDI inflows in EU countries in the period Q1.2013-Q2.2021 (USD million, per cent of FDI inflows total)

Figure 1



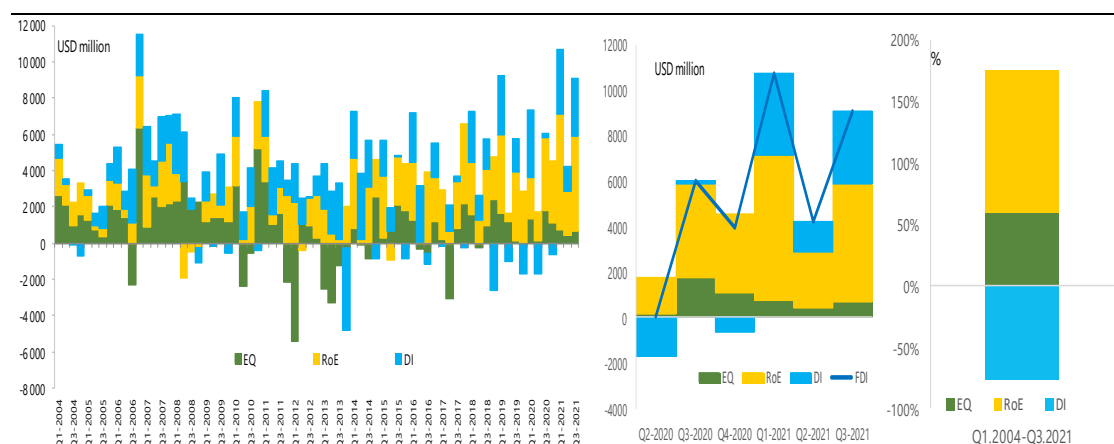
Sources: Author's compilation based on OECD (2022) and Eurostat (2022).

In the case of Poland, from the time of joining the structures of the European Union (EU) in 2004 to 2010, positive values, mainly equity and reinvestment of earnings, prevailed, while in the following years (2010-2021) reinvestment of earnings and negative debt instruments were dominant (2018-2021). During the pandemic, a similar structure was maintained in recent years. The average share of FDI inflows components in 2004-2021 was: equity (33.1%), reinvestment of earnings (139.0%) and debt instrument (-7.1%) (Figure 2).

It is worth noting that the financial structure of FDI (including FDI positions) differs significantly between countries, e.g. Central and Eastern Europe (V4). In 2020, Slovak Republic had the highest share in terms of equity including reinvestment earnings in inward FDI positions (91.5%) and the lowest in Poland (76.6%). The share of debt instrument was respectively for Hungary (3.3%) and for Poland (23.4%). This means that the presented differences in the financial structure of FDI could have determined the different impact of financial structure of FDI inflows on economic growth in individual V4 countries (Table 2).

The structure of FDI inflows in Poland in the period Q1.2004-Q3.2021
(USD million, per cent of FDI inflows total)

Figure 2



Sources: Author's compilation based on NBP (2022).

Inward FDI positions in Visegrad group (V4) in 2020 (per cent)

Table 2

| Specification | Czech Republic (CZ) | Hungary (HU) | Poland (PL) | Slovak Republic (SK) |
|--|---------------------|--------------|-------------|----------------------|
| Equity including reinvestment earnings | 89.8 | 96.7 | 76.6 | 91.5 |
| Debt | 10.2 | 3.3 | 23.4 | 8.5 |

Sources: Author's calculation based on OECD.Stat (2022).

Returning to FDI inflows, it is worth noting that there are significant differences between the V4 countries in the relation of these investments to gross fixed capital formation (GFCF) or GDP (Table 3).

FDI inflows as a percentage of GFCF, GDP and total world in Visegrad group (V4) in the year 2004-2020 (per cent)

Table 3

| Specification | Czech Republic (CZ) | Hungary (HU) | Poland (PL) | Slovak Republic (SK) | Visegrad group (V4) |
|-------------------------|---------------------|--------------|-------------|----------------------|---------------------|
| GFCF (2004-2019) | 12.8 | 12.4 | 14.2 | 13.8 | 13.3 |
| GDP (2004-2020) | 3.4 | 2.8 | 2.7 | 3.0 | 2.4 |
| Total world (2004-2020) | 0.5 | 0.3 | 0.9 | 0.2 | 0.5 |

Sources: Author's calculation based on UNCTAD (2022).

4. Data and research procedure of cause-and-effect relationships between components of FDI inflows and GDP in Poland

The author decided to analyze the impact of FDI on economic growth, using the FDI inflows, not FDI stocks. This decision resulted from the fact that FDI inflows better characterize the impact on the economic growth rate (e.g. Osei and Kim, 2020; Carbonell and Werner, 2018; Moudatsou, 2003), while FDI stocks can be used to assess economic development (changes in GDP per capita).

The research is based on statistics from the NBP (FDI components) and OECD Internet databases (GDP) for the period 2004:Q1–2021:Q3 (71 quarters). NBP compiles data on direct investment in compliance with the OECD definition (OECD 2008, IMF 2009). FDI components come from the balance of payments (BP) data calculated according to assets and liabilities presentation. The analysis uses FDI as direct investment comes from the financial account of BP (liabilities, net transactions). This means that the analysis refers to the impact of FDI financial instruments on GDP changes.

Moreover, such data was selected as it is published on a quarterly basis. Quarterly data is important from the point of view of econometric modelling. In contrast, the data on the FDI inflows (according to directional principle) are published only on an annual basis. This means that a short series of annual data makes modelling difficult.

In order to analyse the relationship between changes in GDP values and financial instruments (components) of FDI in Poland in the years 2004:Q1–2021:Q3 (by used the simple moving average), a final formula for the GDP function was developed:

$$GDP_t = \alpha_0 + \alpha_1 EQ_t + \alpha_2 RoE_t + \alpha_3 DI_t + \xi_i \quad (1)$$

The model used consists of the dependent variable (GDP) and three independent variables, where:

GDP – gross domestic product (USD million),

EQ – equity other than reinvestment of earnings (USD million),

RoE – reinvestment of earnings (USD million),

DI – debt instruments (USD million),

ξ_i – random component,

t – period.

In this study, methods are used known from literature on international economics and international finance and econometric methods like the VECM model (*Vector Error Correction Method*) including the impulse response functions and forecast error variance decomposition analysis.

The data verification procedure and the selection of the analysis method included: ADF test, KPSS stationary test, VAR inverse root, the Engle-Granger and Johanson test and lag order (AIC, BIC, HQC criteria). In order to verify correctness of the VECM model results: two tests were carried out verifying the

5. Results – VECM model, the impulse response functions and the decomposition of variance

Co-integration was verified by means of the Engle-Granger and Johansen tests which confirmed the occurrence of co-integration and thus justified the use of the VECM model for the lag order 8 and co-integration of order 1.

In accordance with the Granger representation theorem, if variables y_t and x_t are integrated to the order of 1 (1) and are co-integrated, the relationship between them can be represented as a vector error correction model (VECM) (Piłatowska 2003).

The general form of the VECM can be written as:

$$\begin{aligned}\Delta Y_t &= \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} + \dots + \Gamma_{k-1} \Delta Y_{t-k+1} + \pi Y_{t-k} + \varepsilon_t = \\ &= \sum_{i=1}^{k-1} \Gamma_i \Delta Y_{t-i} + \pi Y_{t-k} + \varepsilon_t,\end{aligned}\quad (2)$$

where:

$$\Gamma_i = \sum_{j=1}^i A_j - I, \quad i = 1, 2, \dots, k-1, \quad \Gamma_k = \pi = -\pi(1) = -\left(I - \sum_{i=1}^k A_i\right)$$

and I is a unit matrix.

The analysis of the VECM model allows us to draw the following conclusions: the levels of vector α parameters indicating the rate of GDP adjustments in successive VECM model equations show that the highest rate of these adjustments was noted for own changes in GDP.

The evaluation of the EC1 indicates that the strongest correction of the deviation from long-term equilibrium occurs in the case of the own GDP equation. Here, around 0.7% of the imbalance from the long-term growth path is corrected by a short-term adjustment process. Weaker deviations adjustments occur for DI (0.61%), EQ (0.29%), and the worst for RoE (0.17%). The values of the coefficient of determination R^2 reveal adjustment of the VECM model equations to empirical data, i.e., for DI (83.0%), GDP (80.5%), RoE (71.4%) and EQ (52.6%) (Table 4).

The main research results for VECM model

VECM system, lag order 8. Maximum likelihood estimates, observations 2006:4-2021:3 (T = 60)

Cointegration rank = 1, Case 3: Unrestricted constant

Table 4

| β (cointegrating vectors, standard errors in parentheses) | | | α adjustment vectors) | |
|--|--------------------------|--------------------------|---------------------------------|---------------------------|
| ma_GDP | 1.0000 | (0.00000) | d_GDP | 0.0071702 |
| ma_EQ | -0.81884 | (9.1276) | d_EQ | -0.0029978 |
| ma_RoE | -167.94 | (20.255) | d_RoE | -0.0017278 |
| ma_DI | 167.11 | (18.882) | d_DI | -0.0061515 |
| Specification | ma_GDP | ma_EQ | ma_RoE | ma_DI |
| | Coefficient (p-value) | Coefficient (p-value) | Coefficient (p-value) | Coefficient (p-value) |
| EC1 | 0.00717025 (0.7496) | -0.00299777 (0.3291) | -0.00172779 (0.1930) | -0.00615147 (3.89e-05) |
| R2 | 0.805401 | 0.526401 | 0.714782 | 0.830466 |
| DW | 1.937766 | 1.776624 | 2.063881 | 1.903655 |

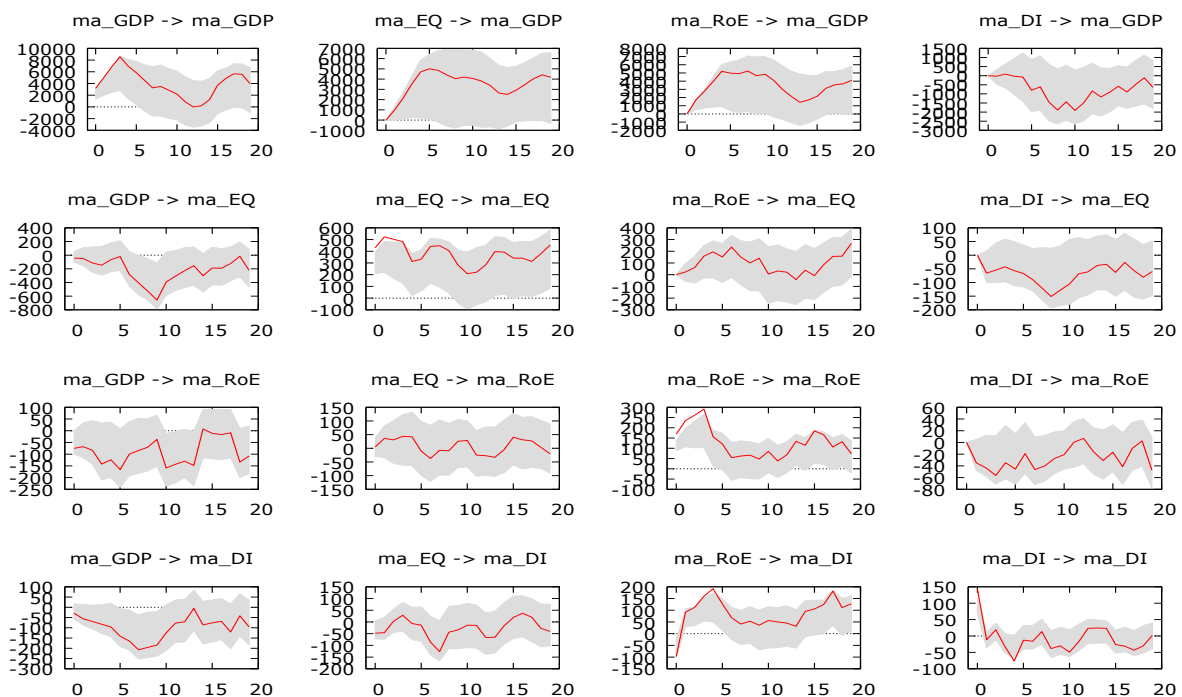
Sources: *Source*: author's own calculations.

5.1. The impulse response functions

Analysis of GDP responses to shocks derived from FDI components reveal that GDP responses are the strongest to impulses from reinvestment of earnings (RoE) and equity (EQ). The strongest GDP responses occur in the periods (quarters) 1–8 (RoE) and 1–6 (EQ). Next periods are characterized by falling fluctuations in GDP responses and growing in subsequent periods (15–20) (Figure 3).

The impulse response functions (summary statement), forecast horizon 20q, include bootstrap confidence interval 1- α =0.90 (shaded area)

Figure 3

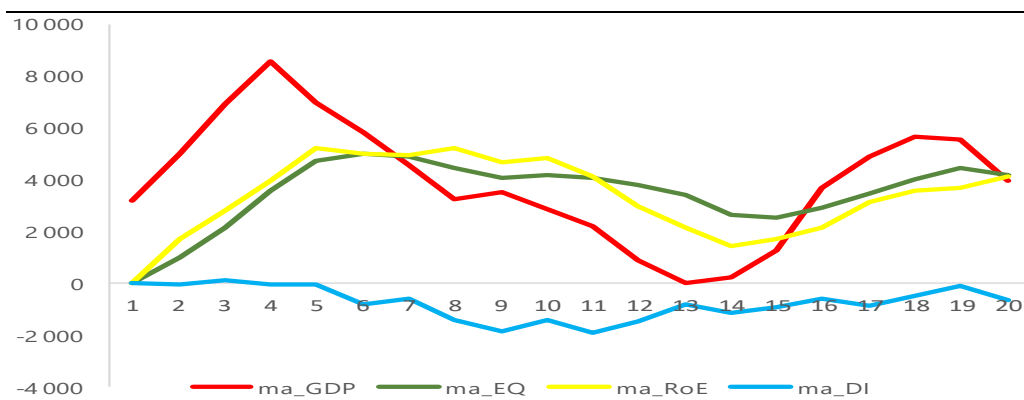


Sources: Source: author's own calculations.

Presentation of GDP responses to impulses shows distinctly that GDP responds most strongly to its own standard deviations in the period 1–5. However, in the periods 6–13, GDP responses to their own errors in forecasts indicate fading/weakening tendencies and next growing stronger over the few quarters (15–20) than the FDI components. Reassuring, in the case of impulses derived from FDI components, the GDP is the strongest to positive impulses from reinvestment of earnings and equity and negative stimuli from debt instruments (the weakest response) (Figure 4).

The response of GDP to a standard shock in own GDP and components of FDI inflows (quarters)

Figure 4



Sources: Source: author's own calculations.

5.2. The variance decomposition

GDP and all FDI components were analysed by means of decomposition of variance in the forecast horizon of 20 quarters (Table 5, Figure 5). The results of GDP decomposition indicate that in period 1 these changes are in 100% accounted for by their own forecast errors. In period 2, their own changes lose in significance (90.2%), and such FDI components as RoE (7.2%), equity (2.4%) and DI (0.2%) grow in significance. In the following periods, GDP's own changes fall in period 20th (42.5%), whereas RoE grow (27.5%, by maximum 29.3 in 12nd), similarly EQ (27.9%) and DI (about 2.0%). Thus, we can conclude that FDI significance in forecasting the degree of explanation of GDP amounts jointly to ca. 57.5% in the 20th quarter, that is 5 years (Table 5, Figure 5).

Decomposition of variance for ma_GDP (per cent)

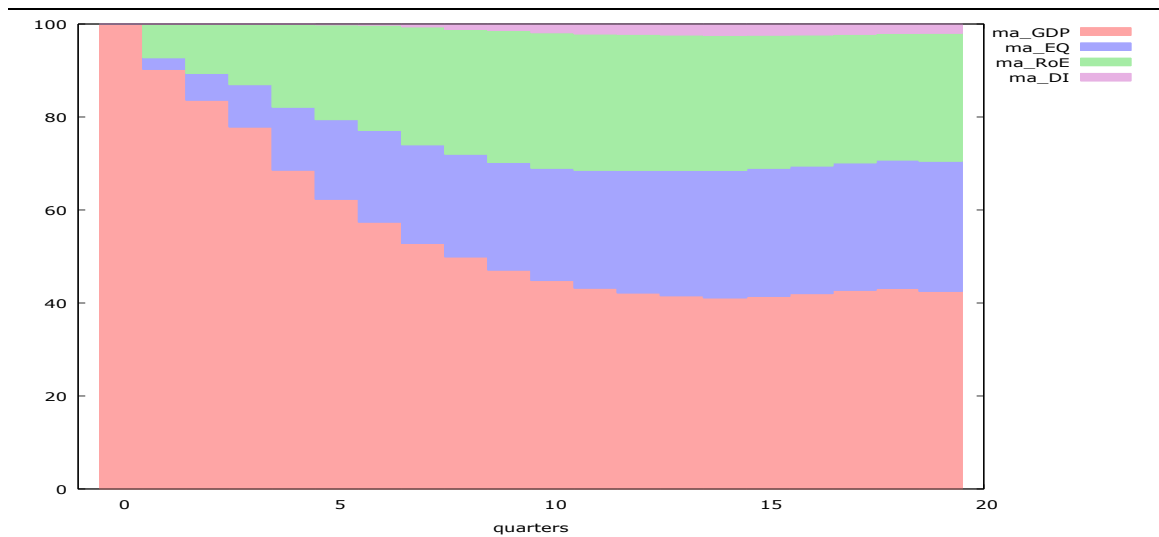
Table 5

| period | ma_GDP | ma_EQ | ma_RoE | ma_DI |
|--------|---------|---------|---------|--------|
| 1 | 100 | 0 | 0 | 0 |
| 2 | 90,2598 | 2,4795 | 7,2587 | 0,0019 |
| 3 | 83,6233 | 5,7306 | 10,6377 | 0,0084 |
| 4 | 77,8407 | 9,1040 | 13,0507 | 0,0046 |
| 5 | 68,5739 | 13,5497 | 17,8715 | 0,0049 |
| 6 | 62,3186 | 17,1010 | 20,4084 | 0,1721 |
| 7 | 57,3975 | 19,7228 | 22,6504 | 0,2294 |
| 8 | 52,7771 | 21,2409 | 25,3707 | 0,6113 |
| 9 | 49,9027 | 22,1080 | 26,8059 | 1,1835 |
| 10 | 47,0728 | 23,1272 | 28,3794 | 1,4207 |
| 11 | 44,8134 | 24,1722 | 29,1301 | 1,8843 |
| 12 | 43,1929 | 25,3596 | 29,3069 | 2,1406 |
| 13 | 42,1546 | 26,3888 | 29,2665 | 2,1900 |
| 14 | 41,5409 | 26,9885 | 29,1208 | 2,3498 |
| 15 | 41,0897 | 27,4240 | 29,0585 | 2,4278 |
| 16 | 41,4267 | 27,5620 | 28,6266 | 2,3847 |
| 17 | 41,9984 | 27,4622 | 28,1923 | 2,3471 |
| 18 | 42,7385 | 27,4005 | 27,6539 | 2,2071 |
| 19 | 43,1311 | 27,6194 | 27,1942 | 2,0554 |
| 20 | 42,5052 | 27,9820 | 27,5218 | 1,9910 |

Sources: Source: author's own calculations.

Decomposition of variance for ma_GDP (per cent)

Figure 5



Sources: Source: author's own calculations.

6. Conclusion and way forward

Based on the research, the following conclusions were formulated.

- The experience of several decades of expansion of FDI flows and changes in their structure indicates the need for research on the impact of the financial structure of FDI inflows on GDP, i.e. as a multidimensional variable and not only as a monolithic variable.
- In the structure of FDI inflows in Europe and in Poland, the share of reinvestment of earnings has grown in the last 10 years. The period of the COVID-19 pandemic did not change that either.
- The research on the impact of FDI components on GDP in Poland using the VECM model, the impulse response functions and the decomposition of variance, confirmed the increasing over time, positive impact of mainly capital shares (equity) and reinvestment of earnings and the weakest impact of debt instruments. The impulse response function indicates that the impact of equity and reinvestment of earnings grows strongly on GDP in the first 2 years (about 8 quarters), slightly weakening the growth rate in the following years. According to decomposition variance financial structure of FDI inflows grow significance in forecasting the degree of explanation of GDP amounts jointly to ca. 57.5% in the 20th quarter, that is 5 years.
- The results of the research confirm the importance of pursuing an investment policy focused on attracting new investments (new equity), including the so-called greenfield and on maintaining the existing ones (reinvestment of earnings).
- Considering that the financial structure of FDI inflows also depends on the industry structure of the host economy, further research should focus on the diagnosis of these relationships and their impact on economic development.

Reassuring, the results of the research allowed for a positive verification of the formulated thesis that: The optimal structure of FDI inflows to Poland is a structure that maximizes equities and reinvestment of earnings with less participation of debt instruments and dividend payments outside the host country. These findings seem plausible and important because their implications can find practical applications and can become the basis of recommendations for economic policy in Poland. In the future, these results may have methodological application for international FDI statistics and measures of economic growth.

References

- Borga M. (2018), "Recording of FDI Income, reinvested Earnings, and Dividends: The Case of Superdividends", Working Group on International Investments Statistics, WGIIIS Meetings, October 2-4, Paris, France, JT03435386.
- Brada J.C., Tomšík V. (2003), "Reinvested Earnings Bias, The 'Five Percent' Rule and the Interpretation of the Balance of Payments – with an Application to Transition Economies", William Davidson Institute, Working Paper, no. 543.

- Carbonell J.B., and Werner R., (2018), "Does Foreign Direct Investment Generate Economic Growth? A New Empirical Approach Applied to Spain", *Economic Geography*, no. 94:4, pp. 425-456.
- Damgaard M.L. – Laursen F. – Wederinck R. (2010), "Forecasting direct investment equity income for the Danish Balance Payments", *Danmarks National Bank, Working Papers*, no. 65, pp. 1-32.
- ECB (2019), European Commission, "Guidance on the estimation of Reinvested Earnings on Foreign direct investment", 08 April.
- Eurostat (2022), <https://ec.europa.eu/eurostat/web/economic-globalisation/globalisation-in-business-statistics/foreign-direct-investments>.
- IMF (2009), "Balance of Payments and International Investment Position Manual", 6th edition.
- Knetsch T.A., Nagengast A.J. (2016), "On the Dynamics of the Investment Income Balance", *Deutsche Bundesbank, Discussion Paper*, 2016, no. 21.
- Moudatsou A., (2003), "Foreign Direct Investment and Economic Growth in European Union", *Journal of Economic Integration*, no. 18(4), pp. 689-707.
- NBP (2022), "Balance of Payments, Financial Account – Direct Investment", http://www.nbp.pl/home.aspx?f=/statystyka/bilans_platniczy/bilansplatniczy_kw.html.
- Novotný F. – J. Podpiera (2008), "The profitability Life-Cycle of Direct Investment: An International Panel Study", *Economic Change and Restructuring*, no. 41(2), pp.143-153.
- Novotný F. (2015), "Profitability Life Cycle of Foreign Direct Investment and Its Application to the Czech Republic", *Czech National Bank, Working Paper Series*, no. 11, pp. 1-25.
- Novotný F. (2018), "Profitability Life Cycle of Foreign Direct Investment: Application to the Czech Republic", *Emerging Markets Finance and Trade*, no. 54 (7), pp. 1623-1634, <https://doi.org/10.1080/1540496X.2017.1316259>.
- OECD (2008), "Benchmark Definition of Foreign Direct Investment", 4th edition, OECD Publishing Paris, pp. 70-72.
- OECD (2022), *OECD.Stat.*, <http://stats.oecd.org/>.
- Osei M.J., and Kim J. (2020), "Foreign Direct Investment and Economic Growth: Is more Financial Development better?", *Economic Modelling* 93, pp. 154-161.
- UNCTADStat (2022), *Statistics*, <https://unctadstat.unctad.org>.



NARODOWY
BANK POLSKI

Irving Fisher Committee on
Central Bank Statistics



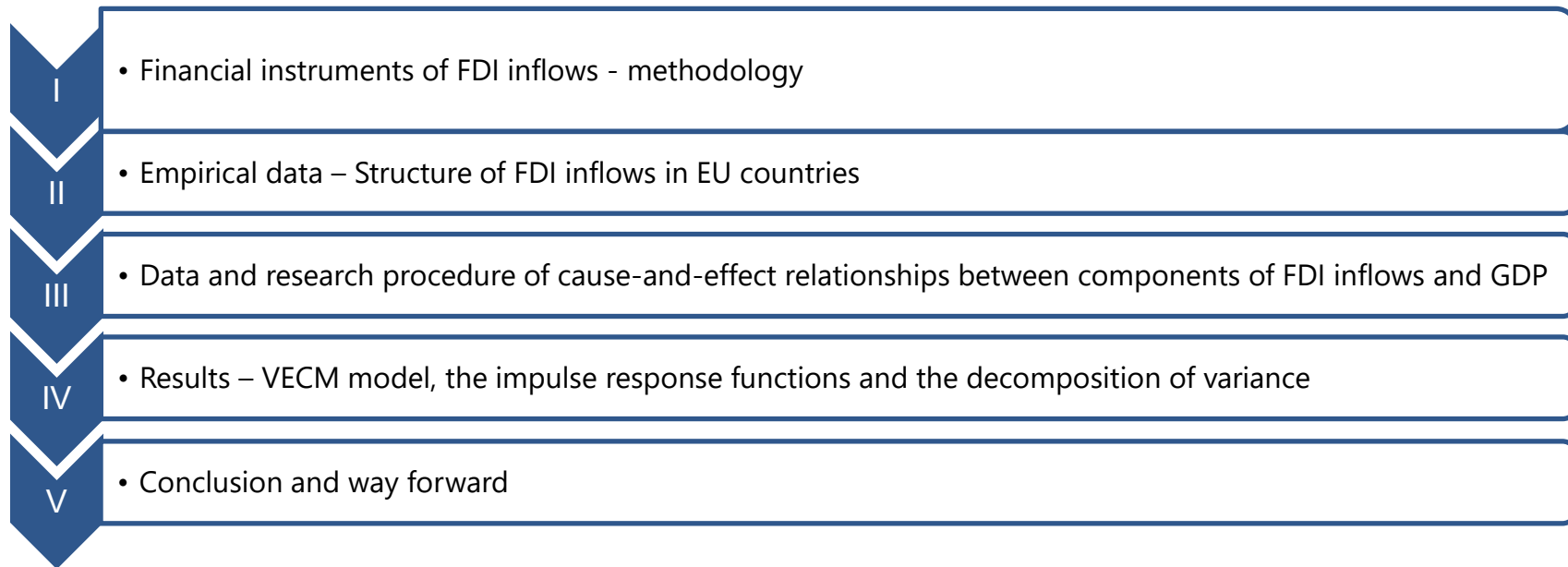
What should be the optimal financial structure of the FDI inflows to Poland in stimulating growth processes?

Aneta Kosztowniak, NBP, Economic Analysis and Research Department

- This presentation should not be reported as representing the views of the National Bank of Poland.
- The views expressed are those of the authors and do not necessarily reflect those of the NBP.



Overview



Financial instruments of FDI inflows - methodology

Table 1. FDI transactions according to the asset/liability principle

| Transactions in assets | Transactions in liabilities |
|---|---|
| Of direct investors in direct investment enterprises | Of direct investment enterprises to direct investors |
| A1. Equity | L1. Equity |
| A1.1. Equity transactions | L1.1. Equity transactions |
| A1.2. Reinvestment of earnings | L1.2. Reinvestment of earnings |
| A2. Debt instruments | L2. Debt instruments |
| Of direct investment enterprises in direct investors- Reverse investment | Of direct investment enterprises in direct investors- Reverse investment |
| A3. Equity | L3. Equity |
| A4. Debt instruments | L4. Debt instruments |
| In fellow enterprises | To fellow enterprises |
| A5. Equity | L5. Equity |
| A5.1. If ultimate controlling parent is resident | L5.1. If ultimate controlling parent is non- resident |
| A5.2. If ultimate controlling parent is non- resident | L5.2. If ultimate controlling parent is resident |
| A6. Debt instruments | L6. Debt instruments |
| A6.1. If ultimate controlling parent is resident | L6.1. If ultimate controlling parent is non- resident |
| A6.2. If ultimate controlling parent is non- resident | L6.2. If ultimate controlling parent is resident |

Source: OECD (2008).

According to OECD (2008) direct investment transactions are all transactions between direct investors, direct investment enterprises, and/or other fellow enterprises. For transactions, the asset/liability principle is schematically shown as follows (*Table 1*).

Equity, other than reinvestment of earnings comprises equity in branches, all shares in subsidiaries and associates (except non – participating, preferred shares that are treated as debt securities and included under direct investment, debt instruments) and other contributions of an equity nature.

Reinvestment of earnings denote the part of profits, accruing to a direct investor, which remains in the direct investment enterprise, and which is allocated to its further development. According OECD (2008) *reinvestment of earnings* of direct investment enterprises (items A1.2 and L1.2) reflects earnings accruing to direct investors (that is, proportionate to the ownership of equity) during the reference period less earnings declared for distribution in that period.

Debt instruments mean all forms of investing other than the acquisition of shares or equity, or reinvestment of earnings associated with such shares or equities. Debt instruments include, among others, credits and loans, debt securities and other unsettled payments between entities in direct investment relationship (OECD 2018; NBP 2017).

Empirical data – Structure of FDI inflows in EU countries and Poland

Figure 1. The structure of FDI inflows in EU countries (USD million, per cent of FDI inflows total)

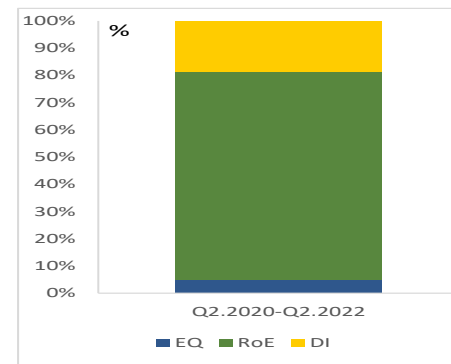
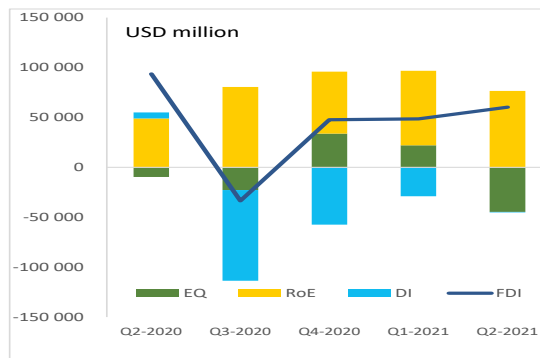
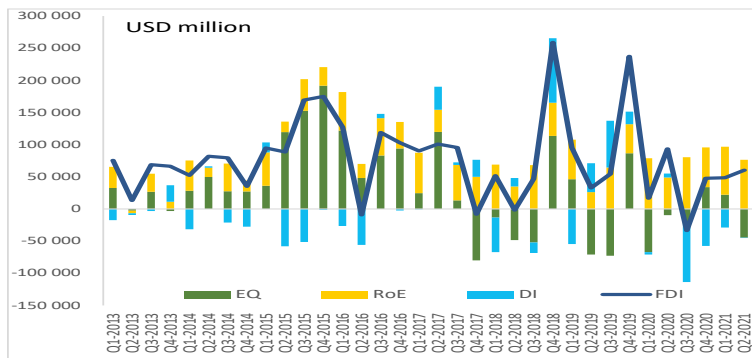
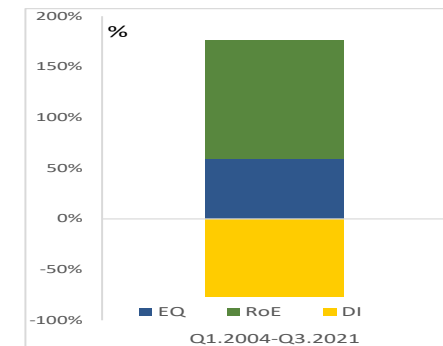
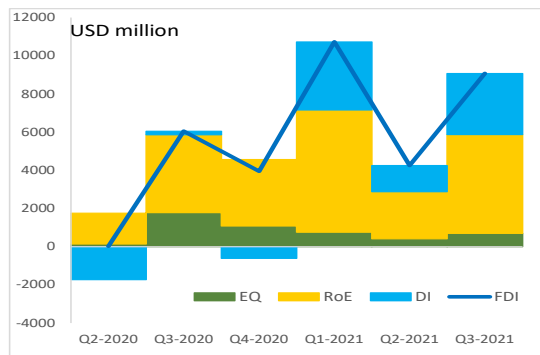
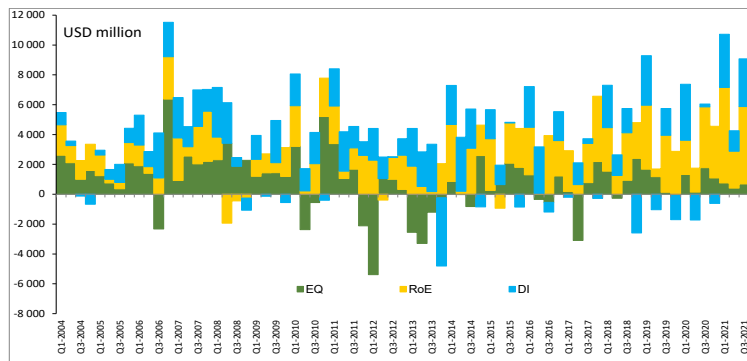


Figure 2. The structure of FDI inflows in Poland (USD million, per cent of FDI inflows total)



Source: Author's compilation based on OECD (2022) and Eurostat (2022).

Data and research procedure of cause-and-effect relationships between components of FDI inflows and GDP

Data: the quarters time-series data covering the period 2004:Q1-2021:Q3 (71 quarters; by used the simple moving average);

Sources: FDI inflows (NBP, BoP, 2022) and GDP (OECD Internet databases (2022)).

The data verification procedure and the selection of the analysis method included: ADF test, KPSS stationary test, VAR inverse root, the Engle-Granger and Johanson test and lag order (AIC, BIC, HQC criteria).

In order to verify correctness of the VECM model results: two tests were carried out verifying the occurrence of autocorrelation, i.e.: Autocorrelation Ljung-Box Q' test, and ARCH test.

The final formula for the GDP function:

$$GDP_t = \alpha_0 + \alpha_1 EQ_t + \alpha_2 RoFE_t + \alpha_3 DI_t + \xi_t$$

The general form of the VECM model:

$$\begin{aligned} \Delta Y_t &= \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} + \dots + \Gamma_{k-1} \Delta Y_{t-k+1} + \pi Y_{t-k} + \varepsilon_t = \\ &= \sum_{i=1}^{k-1} \Gamma_i \Delta Y_{t-i} + \pi Y_{t-k} + \varepsilon_t, \end{aligned} \quad (2)$$

where:

$$\Gamma_i = \sum_{j=1}^i A_j - I, \quad i = 1, 2, \dots, k-1, \quad \Gamma_k = \pi = -\pi(1) = -\left(I - \sum_{i=1}^k A_i\right)$$

and I is a unit matrix.

Table 2. The main research results for VECM model

| VECM system, lag order 8 | | | | |
|---|---|-------------|----------------------------------|-------------|
| Maximum likelihood estimates, observations 2006:4-2021:3 (T = 60) | | | | |
| Cointegration rank = 1, Case 3: Unrestricted constant | | | | |
| | β (cointegrating vectors, standard errors in parenthes) | | α (adjustment vectors) | |
| ma_GDP | 1.0000 | (0.00000) | d_GDP | 0.0071702 |
| ma_EQ | -0.81884 | (9.1276) | d_EQ | -0.0029978 |
| ma_RoE | -167.94 | (20.255) | d_RoE | -0.0017278 |
| ma_DI | 167.11 | (18.882) | d_DI | -0.0061515 |
| | ma_GDP | ma_EQ | ma_RoE | ma_DI |
| Specification | Coefficient | Coefficient | Coefficient | Coefficient |
| | (p-value) | (p-value) | (p-value) | (p-value) |
| EC1 | 0.00717025 | -0.00299777 | -0.00172779 | -0.00615147 |
| | (0.7496) | (0.3291) | (0.1930) | (3.89e-05) |
| R2 | 0.805401 | 0.526401 | 0.714782 | 0.830466 |
| DW | 1.937766 | 1.776624 | 2.063881 | 1.903655 |

Source: author's own calculations.

The impulse response functions

Figure 3. Forecast horizon 20q, include bootstrap confidence interval $1-\alpha=0.90$ (shaded area)

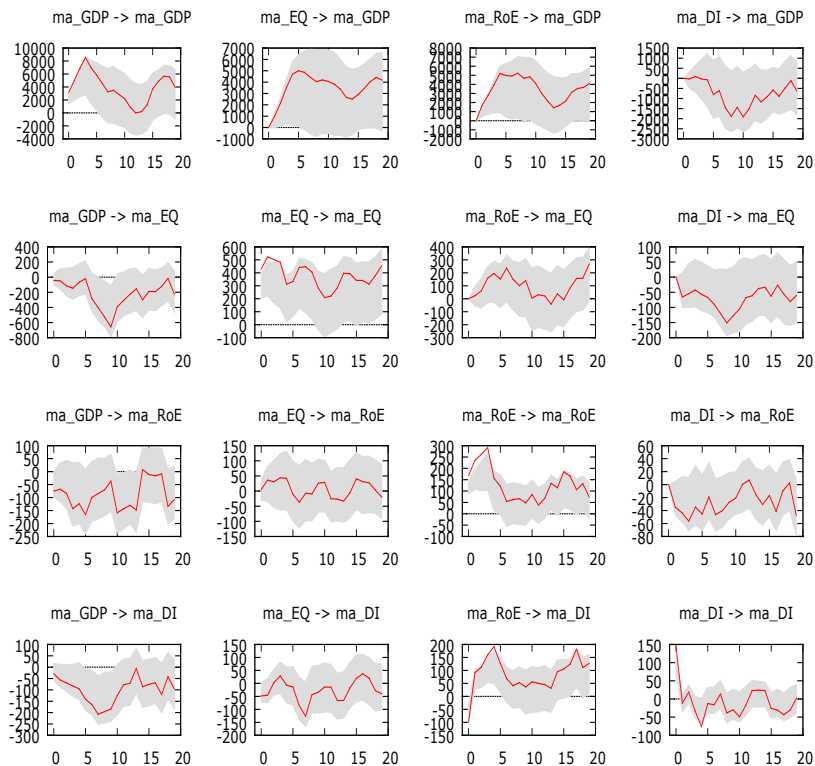
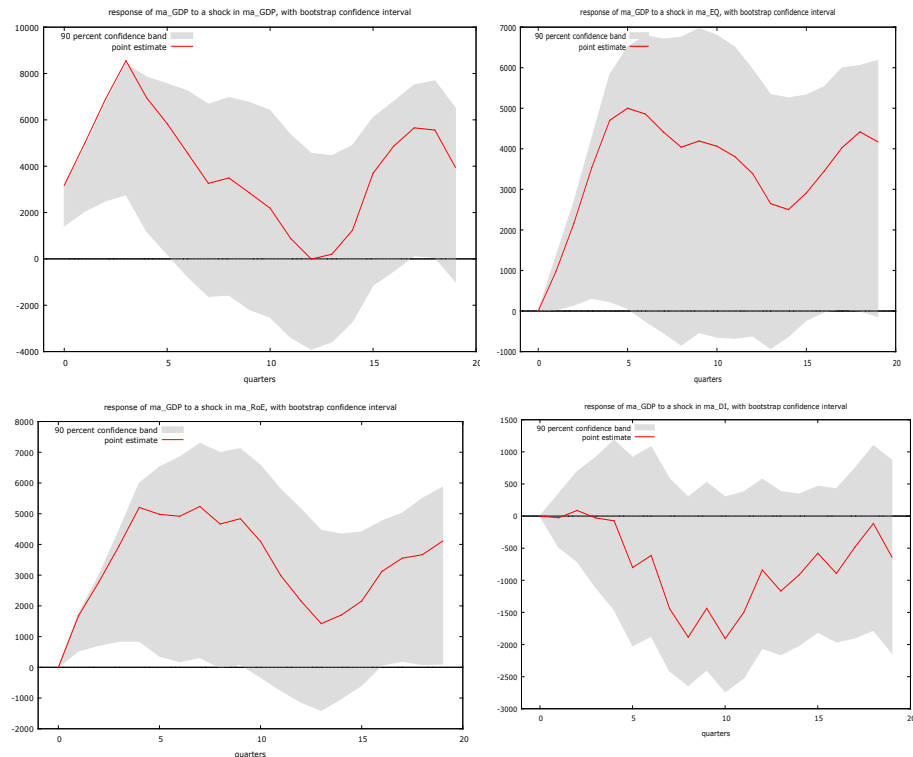


Figure 4. Response of GDP to a standard shock in own GDP and components of FDI inflows (quarters)



The variance decompositions

Figure 5. Forecast variance decomposition for GDP and EQ, RoE, DI (quarters)

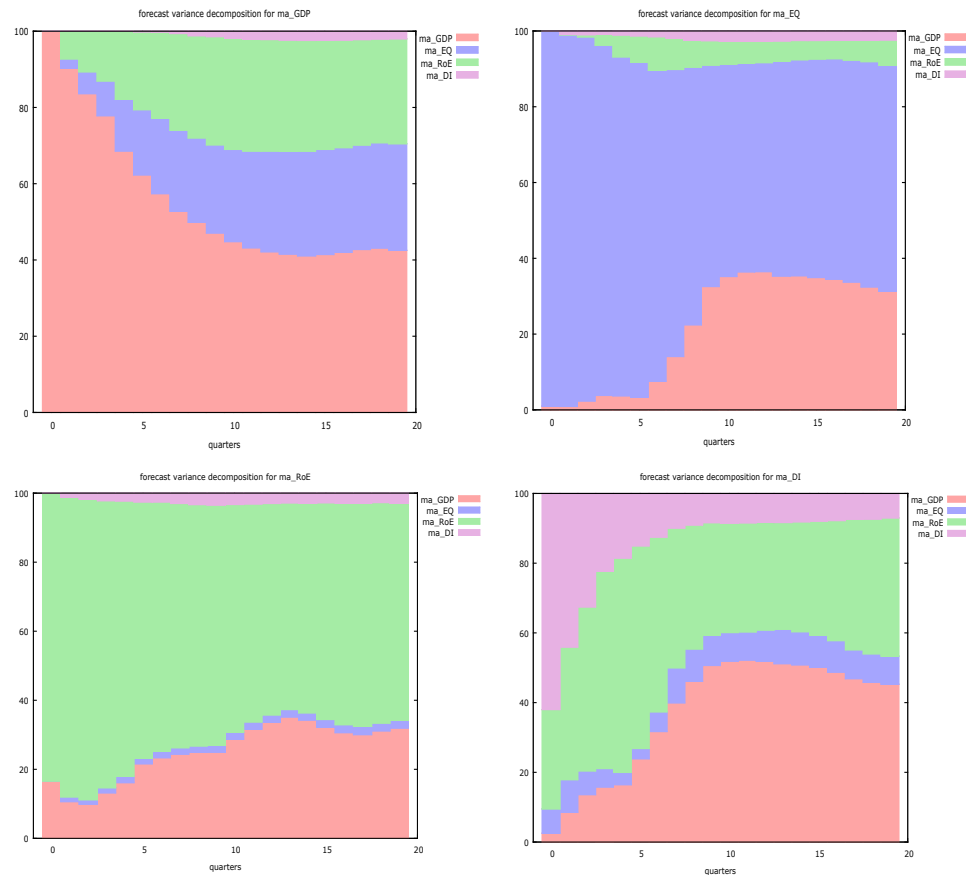


Table 3. Decomposition of variance for ma_GDP (per cent)

| period | ma_GDP | ma_EQ | ma_RoE | ma_DI |
|--------|---------|---------|---------|--------|
| 1 | 100 | 0 | 0 | 0 |
| 2 | 90,2598 | 2,4795 | 7,2587 | 0,0019 |
| 3 | 83,6233 | 5,7306 | 10,6377 | 0,0084 |
| 4 | 77,8407 | 9,1040 | 13,0507 | 0,0046 |
| 5 | 68,5739 | 13,5497 | 17,8715 | 0,0049 |
| 6 | 62,3186 | 17,101 | 20,4084 | 0,1721 |
| 7 | 57,3975 | 19,7228 | 22,6504 | 0,2294 |
| 8 | 52,7771 | 21,2409 | 25,3707 | 0,6113 |
| 9 | 49,9027 | 22,108 | 26,8059 | 1,1835 |
| 10 | 47,0728 | 23,1272 | 28,3794 | 1,4207 |
| 11 | 44,8134 | 24,1722 | 29,1301 | 1,8843 |
| 12 | 43,1929 | 25,3596 | 29,3069 | 2,1406 |
| 13 | 42,1546 | 26,3888 | 29,2665 | 2,1900 |
| 14 | 41,5409 | 26,9885 | 29,1208 | 2,3498 |
| 15 | 41,0897 | 27,4240 | 29,0585 | 2,4278 |
| 16 | 41,4267 | 27,5620 | 28,6266 | 2,3847 |
| 17 | 41,9984 | 27,4622 | 28,1923 | 2,3471 |
| 18 | 42,7385 | 27,4005 | 27,6539 | 2,2071 |
| 19 | 43,1311 | 27,6194 | 27,1942 | 2,0554 |
| 20 | 42,5052 | 27,9820 | 27,5218 | 1,9910 |

Conclusion and way forward

I

- The experience of several decades of expansion of FDI flows and changes in their structure indicates that research on the impact of the financial structure of FDI inflows on GDP, i.e. as a multidimensional variable and not only as a monolithic variable.

II

- In the structure of FDI inflows in Europe and in Poland, the share of reinvestment of earnings is growing.
- The period of the COVID-19 pandemic did not change that either.

III

- The research on the impact of FDI components on GDP in Poland using the VECM model, the impulse response functions and the decomposition of variance, confirmed the increasing over time, positive impact of mainly capital shares (equity) and reinvestment of earnings.
- The impact of these variables grows strongly in the first 2 years, slightly weakening the growth rate in the following years.

IV

- The results of the research confirm the importance of pursuing an investment policy focused on attracting new investments (new equity), including the so-called greenfield and on maintaining the existing ones (reinvestment of earnings).

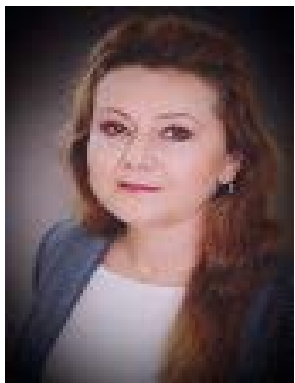
V

- Considering that the financial structure of FDI inflow also depends on the industry structure of the host economy, further research should focus on the diagnosis of these relationships and their impact on economic development.



NARODOWY
BANK POLSKI

Thank for your attention



Aneta Kosztowniak

Economic Expert

Economic Analysis and Research Department

phone: +48 22 185 15 07

mobile: +48 691 034 170

e-mail: Aneta.Kosztowniak@nbp.pl; aneta.kosztowniak@wp.pl

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022

Big data analytics on payment system data for measuring household consumption in Indonesia¹

Muhammad Abdul Jabbar, Mohammad Khoyrul Hidayat and Alvin Andhika Zulen,
Bank Indonesia

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Big Data Analytics on Payment System Data for Measuring Household Consumption in Indonesia

Alvin Andhika Zulen¹, Mohammad Khoyrul Hidayat², Muhammad Abdul Jabbar³

Abstract

Consumer spending is one of the main indicators to measure state of the economy in Indonesia. However, data related to household consumption expenditure (as part of Gross Domestic Product (GDP)) are published and available on a quarterly basis with a publication lag of one month. On the other hand, technological advancement and the widespread use of cashless payment systems in the digital era have opened up the opportunities to explore large dataset of payments data for monitoring economic activity. This study aims to examine the use of retail payment system data, particularly from the Bank Indonesia National Clearing System (SKNBI), as a proxy for household consumption indicators in Indonesia. By utilizing the Big Data Analytics methodology on granular payment system data, we are able to construct more timely household consumption indicators, which are available within a few days after the end of the reference period. This indicator can server as initial proxy for household consumption in Indonesia, which is also indicated by a good correlation with the official data.

Keywords: GDP; household consumption; payment system; big data

JEL classification: B22, C55, E21

¹ Statistics Department – Bank Indonesia; e-mail: alvin_az@bi.go.id

² Statistics Department – Bank Indonesia; e-mail: moh_khoyrul@bi.go.id

³ Statistics Department – Bank Indonesia; e-mail: muhammad_abdul@bi.go.id

Contents

| | |
|---|----|
| 1. Background..... | 3 |
| 2. Literature Review..... | 4 |
| 2.1 Utilization of Big Data for Macroeconomic Indicators..... | 4 |
| 2.2 Utilization of Payment System Data..... | 5 |
| 3. Methodology..... | 5 |
| 3.1 Data | 5 |
| 3.2 Workflow..... | 7 |
| 3.2.1 Data Preprocessing..... | 7 |
| 3.2.2 Data Extraction | 8 |
| 3.2.3 Data Validation..... | 8 |
| 4. Result and Analysis..... | 8 |
| 4.1 Evaluation of Classification Model..... | 8 |
| 4.2 Result Validation | 9 |
| 5. Conclusion and Future Work..... | 11 |
| 5.1 Conclusion | 11 |
| 5.2 Future Work | 11 |
| References..... | 12 |

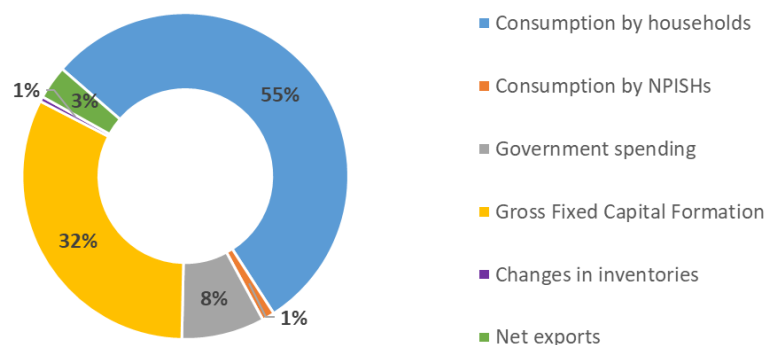
1. Background

The growth of an economy can be measured by Gross Domestic Product (GDP) data. GDP can be calculated through three approaches, i.e. production, income, and expenditure approach. The expenditure approach is a crucial aspect for Bank Indonesia in carrying out its mandate as the monetary, macroprudential, and payment system authority.

The expenditure approach can be analysed through several indicators. One of the indicators is the household consumption expenditure indicator, which represents the spending on goods and services by resident households for final consumption purposes. Household consumption is the most significant contributor ($\pm 55\%$, Figure 1) to Indonesia's GDP based on the expenditure in 2020.

Contribution of Indonesia's GDP Components by Expenditure in 2020

Figure 1



Source: BPS

Bank Indonesia, in synergy with the government, constantly strives to formulate the appropriate policies in its 3 (three) main objectives of Bank Indonesia, which are monetary, financial system stability, and payment system. For the central bank, it is significant to know the current economic condition (state of the economy) which will be used as the basis predicting the economy's growth in the future. Considering Bank Indonesia's policies are aimed at influencing the expenditure sides, Bank Indonesia needs to observe the movement of household consumption as the most significant expenditure component in Indonesia's GDP as early as possible. However, data related to household consumption indicators in GDP are published and available on a quarterly basis with a publication lag of one month.

Nowcasting is a method used to predict the direction of the economic movement. Through nowcasting, policymakers can assess the direction of the economic movement by using representative high-frequency data to capture the dynamics of the reference indicators (i.e. GDP). Many researches related to nowcasting have been published, and several nowcasting models have been widely used to estimate various indicators (Tarsidin, Idham, & Rakhman, 2016). In addition, this method may help policymakers to formulate policy responses while waiting for the official release of macroeconomic indicators.

The Covid-19 pandemic since the beginning of 2020 has directly impacted the world economy, and Indonesia is no exception. Indonesia has fallen into its first recession in 22 years as the Covid-19 pandemic continues to take its toll. In response to this situation, policymakers need to project macroeconomic indicators to formulate appropriate policies. During this pandemic, the analysis of household consumption indicators as one of the macroeconomic indicators is very fundamental, especially in helping the central bank and government to see the growth and predict household consumption behavior.

On the other hand, technological advancement and the widespread use of cashless payment systems in the digital era have opened up the opportunities to explore large dataset of payments data for monitoring economic activity. For example, current technological advancement allows us to utilize Big Data Analytics for processing large dataset and estimating economic indicators in advance (Buono et al., 2018). This study aims to examine the use of retail payment system data, particularly from the Bank Indonesia National Clearing System (SKNBI), which has a high availability frequency, as a proxy for household consumption indicators in Indonesia.

2. Literature Review

2.1 Utilization of Big Data for Macroeconomic Indicators

Along with the technological advancement, various parties have taken the advantage of Big Data Analytics more broadly. To be more specific, near real-time and faster data processing is urgently needed, especially for supporting policy formulation during the current Covid-19 pandemic. By collecting and processing data on a large scale and high frequency, central bank can determine the current state of the economic in advance as a basis for policy formulation. In addition, literature studies related to the use of Big Data Analytics in the economic field have been developed with various methodologies.

Kapetanios & Papailias (2018) discusses the potential use of Big Data Analytics in nowcasting GDP and other macroeconomic indicators in the UK. In this study, the authors describe various initiatives related to Big Data Analytics in nowcasting macroeconomic indicators. The research also describes the benefit of using Big Data Analytics, which makes it possible to process monthly, weekly, daily, or higher frequency data on a large scale.

In another study, Buono et al. (2018) discusses GDP projection by utilizing various type of data: (i) macroeconomic data with monthly frequency, i.e., core consumer prices, consumer price index, house prices, job vacancies index; (ii) financial data with weekly frequency, i.e., Interest rates, equity indexes, and (iii) uncertainty indicators based on keyword searches in Google. In this study, the author concludes that the results of the uncertainty indicator from Big Data contribute in reducing the RMSFE (root mean squared forecast error) between the nowcasting results and the actual value.

2.2 Utilization of Payment System Data

Several studies have proven that high-frequency data, i.e., data from the payment system, can be used to estimate macroeconomic indicators. By utilizing high-frequency data, which is available faster, we can produce an earlier estimate of current economic conditions. Through this approach, the policy-making authorities benefit by being able to obtain prompt indicators for supporting policy assessment and formulation.

For example, Galbraith & Tkacz (2015) conducted an assessment in using large-scale datasets from the payment system, i.e., debit card, credit card, and cheque transactions, as a proxy for GDP growth in Canada. This study found that by using payment system data as one of the input variables can reduce the nowcasting error by 65%, compared to only using macroeconomic indicators as the input variables.

Recent research was also conducted by Dunn et al. (2020) to measure the impact of the Covid-19 pandemic on consumer spending by utilizing payment transaction data. This study shows a high correlation between official survey data and payment transaction data, especially for retail, accommodation, and restaurant sectors. In terms of data availability, the payment transaction data can be available daily with a lag of 3 (three) days, much higher in frequency when compared to data from monthly surveys that have a publication time lag of 1 (one) month. This study concludes that payment transaction data can be used as alternative data and initial proxy for consumption indicators.

Thus, this study is expected to answer the following research questions:

1. Can payment system data be used as a data source to measure household consumption indicators?
2. Can the resulting household consumption indicators complement the existing indicators?

3. Methodology

3.1 Data

The data source used in this study was obtained from the National Clearing System of Bank Indonesia (SKNBI). This National Clearing System of Bank Indonesia (SKNBI) is a Retail Value Payment System (RVPS) infrastructure operated by Bank Indonesia to process electronic financial data for fund transfer services, debit clearing services, regular payment services, and regular billing services (Regulation of Member of Board of Governors Number 21/12/PADG/2019). Since September 2019, SKNBI has been able to process payment transactions with less than Rp. 1,000,000,000.00 (one billion rupiah) in amount.

The scope of data used in this study is SKNBI fund transfer transaction data from July 2015 to November 2021. Fund transfer service is a service within SKNBI that facilitates the transfer of funds between participant banks, with an average number of transactions reaching + 13 million transactions per month. In the fund transfer service, there are several types of transaction:

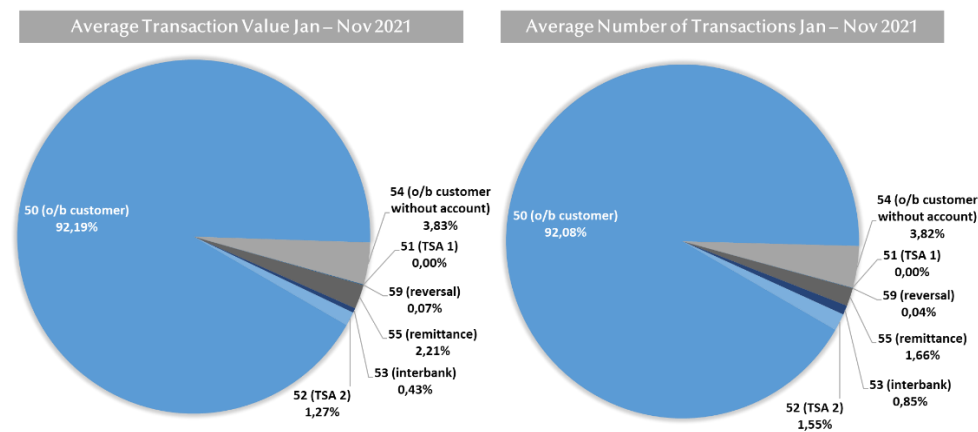
1. 50: Transfer of funds between participants on behalf of the customer;

2. 51: Transfer of funds between participants related to Government's Treasury Single Account (TSA);
3. 52: Transfer of funds between participants and Bank Indonesia's Treasury Single Account;
4. 53: Transfer of funds between participants that is not for the customers' needs;
5. 54: Transfer of funds between participants on behalf of the customer without accounts;
6. 55: Transfer of funds between participants on behalf of the customer related to money remittance; and
7. 59: Refund of fund transfers and payments (reversal).

The share for each transaction type is shown in Figure 2, which shows that the largest share of transaction ($\pm 92\%$) is fund transfer transactions between participants on behalf of the customer (code 50).

Nominal and Frequency of SKNBI Transactions Based on Transaction Type

Figure 2



Source: SKNBI (processed)

The data structure obtained from the fund transfer service transactions is as shown in Table 1. Although the data structure of the SKNBI fund transfer is quite comprehensive, there are still issues related to the data validity. For example, we found that the customer's type code does not always match the customer's name category, both in the sender and recipient fields.

SKNBI Fund Transfer Data Structure

Table 1

| No. | Data Field |
|-----|----------------------------------|
| 1 | DKE ID |
| 2 | BATCH ID |
| 3 | TRANSACTION DATE |
| 4 | SENDER BANK CODE |
| 5 | SENDER LOCATION |
| 6 | BENEFICIARY BANK CODE |
| 7 | BENEFICIARY LOCATION |
| 8 | AMOUNT |
| 9 | TRANSACTION TYPE CODE |
| 10 | SENDER CUSTOMER'S NAME |
| 11 | SENDER CUSTOMER'S ACCOUNT NUMBER |

| No. | Data Field |
|-----|---------------------------------------|
| 12 | SENDER CUSTOMER'S ADDRESS |
| 13 | SENDER CUSTOMER'S ID NUMBER |
| 14 | SENDER CUSTOMER'S TYPE |
| 15 | BENEFICIARY CUSTOMER'S NAME |
| 16 | BENEFICIARY CUSTOMER'S ACCOUNT NUMBER |
| 17 | BENEFICIARY CUSTOMER'S ADDRESS |
| 18 | BENEFICIARY CUSTOMER'S ID NUMBER |
| 19 | BENEFICIARY CUSTOMER'S TYPE |
| 20 | DESCRIPTION |

3.2 Workflow

In constructing household consumption indicators from SKNBI fund transfer data, we develop a text mining model with a rule-based approach for processing unstructured information, backed up by parallel computing technology in Apache Spark – Hadoop. We use Python as the programming language. In general, the workflow in this study consists of data preprocessing, data extraction, and data validation.

3.2.1 Data Preprocessing

The data preprocessing stage is carried out to prepare the raw SKNBI fund transfer transaction data so that they can be further processed at the next stage. The process is as follows:

1. Filter transactions that are not on behalf of customers.

The data that will be further processed is only data with transaction type code 50 (transfer of funds between participants on behalf of the customer) and 54 (transfers of funds on behalf of the customer without accounts).

2. Classification of customers' categories.

For each transaction, we classify sender and beneficiary customers into business entities, governments, and others. This process is critical since customers' type code has validity issues. Classification is conducted using a rule-based approach with rules as shown in Table 3.

3. Filter transactions that do not have a description.

Rules for Classification of Customers' Categories Table 3

| CUSTOMER CLASSIFICATION | SAMPLE KEYWORDS |
|----------------------------|---|
| Business Entities | 'PT.', 'CV.', 'UD.', 'PD.', 'BANK', 'TBK', 'PERUSAHAAN', 'INTERNATIONAL', 'INTERNASIONAL', 'FINANCE', 'LIFE', 'INSURANCE'. |
| Government | 'OTORITAS', 'BADAN', 'BPJS', 'DINAS', 'KAB.', 'KABUPATEN', 'KECAMATAN', 'KELURAHAN', 'KPU', 'PEMERINTAH', 'PEMKAB', 'PEMKOT', 'PROVINSI', 'PROV.', 'PEMPROV'. |
| Others (individual) | Does not contain keywords business entities and governments. |

3.2.2 Data Extraction

There is limited information on the description of fund transfer transactions in SKNBI, in which there is no standard format/reference code for the information written in the description field (free text). We develop a text mining model to analyze the transaction data to handle this issue. The resulting data from the previous stage (section 3.2.1) are used as input for this stage with the following process:

1. Classification of the purposes of SKNBI fund transfer transactions.

Classification is done using a rule-based approach based on predefined keywords, as shown in Table 4.

2. Data aggregation.

After classifying the purpose of the transaction, data can be aggregated as indicators for each transaction purposes, e.g. household consumption, household income, and business. As for transactions with transfer purposes other than those three categories are classified as "Others".

Rules for Classification of Transaction Purposes

Table 4

| SENDER | BENEFICIARY | KEYWORD IN TRANSACTION DESCRIPTION | CLASSIFICATION |
|---|---|--|-----------------------|
| Other than Business Entities and Government | - | 'pembayaran', 'konsumsi', 'belanja', 'pelunasan', 'kos', 'cicilan', 'dp', 'payment', 'dana', 'pelunasan', 'setoran', 'angsuran', 'jasa', 'invoice', 'listrik', 'service', 'sewa rumah', 'kasbon', dsb. | Household Consumption |
| - | Other than Business Entities and Government | 'gaji', 'honor', 'upah', 'payroll', 'salary', 'remunerasi', 'insentif', 'lembur', 'overtime', "kompensasi", 'bagi hasil', 'bonus', 'komisi', 'tukin', 'uang makan', dsb. | Household Income |
| Business Entities | Business Entities | | Business |

3.2.3 Data Validation

The resulting indicators from the previous stage (section 3.2.2) are then validated with the GDP data - Household Consumption (current prices) as the reference indicator. The monthly indicators from the SKNBI are converted into quarterly data by accumulating the nominal amount of transactions in each quarter. This step is required to obtain indicators with the same frequency as household consumption data in GDP. After that, validation is conducted by calculating the correlation value between those two data.

4. Result and Analysis

4.1 Evaluation of Classification Model

After we develop the classification model with the rule-based approach in the previous section, we need to evaluate our model to find out how accurate the model

is in classifying the transaction purposes, particularly for consumption. In this study, we use F1-score⁴ as an evaluation metric. The evaluation was carried out on +3,000 transactions (random sampling) during the period of 2016 to 2019.

The evaluation results in Table 5 show us a good value of the overall F1-score (80%). However, the F1-score for predicting consumption transaction is still relatively low compared to the results for other categories. If we analyze using the confusion matrix in Table 6, there are still many false positive cases, i.e. transactions predicted to be "consumption" but should be included in other categories. We suspect that the prediction error could be caused by the accuracy of the customer categories classification, which still need to be improved.

Evaluation of Transaction Purposes Classification Model

Table 5

| TRANSACTION PURPOSES | RECALL | PRECISION | F1-SCORE | AVERAGE OF F1-SCORE | ACCURACY |
|----------------------|--------|-----------|----------|---------------------|----------|
| Consumption | 92% | 37% | 53% | 80% | 87% |
| Income | 85% | 99% | 91% | | |
| Business | 82% | 92% | 87% | | |
| Others | 90% | 90% | 90% | | |

Confusion Matrix of Transaction Purposes Classification Model

Table 6

| | | PREDICTION | | | |
|--------|----------------------|-------------|--------|----------|--------|
| | | Consumption | Income | Business | Others |
| ACTUAL | Transaction Purposes | | | | |
| | Consumption | 107 | 0 | 0 | 9 |
| | Income | 44 | 497 | 8 | 38 |
| | Business | 24 | 1 | 722 | 134 |
| | Others | 116 | 6 | 51 | 1571 |

4.2 Result Validation

As previously explained in section 3.2.3, the household consumption indicator from the SKNBI (growth, y.o.y) is validated with the GDP-household consumption indicator at current prices (growth, y.o.y). The correlation of the two indicators can be seen in Table 7 and the graph visualization in Figure 4.

The validation results show a high correlation between the two indicators since 1st quarter of 2019, including during the Covid-19 pandemic. These results indicate that consumption indicators from SKNBI transaction data can be used as a proxy for household consumption. Moreover, it can be available earlier, i.e. 2 (two) days lag after the end of the period, both weekly and monthly. The availability of this indicator

⁴ F1 : $2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$

Precision : $(\text{true positive}) / (\text{true positive} + \text{false positive})$

Recall : $(\text{true positive}) / (\text{true positive} + \text{false negative})$

is much faster than the publication of GDP data which has a time lag of more than 1 (one) month.

Correlation Table

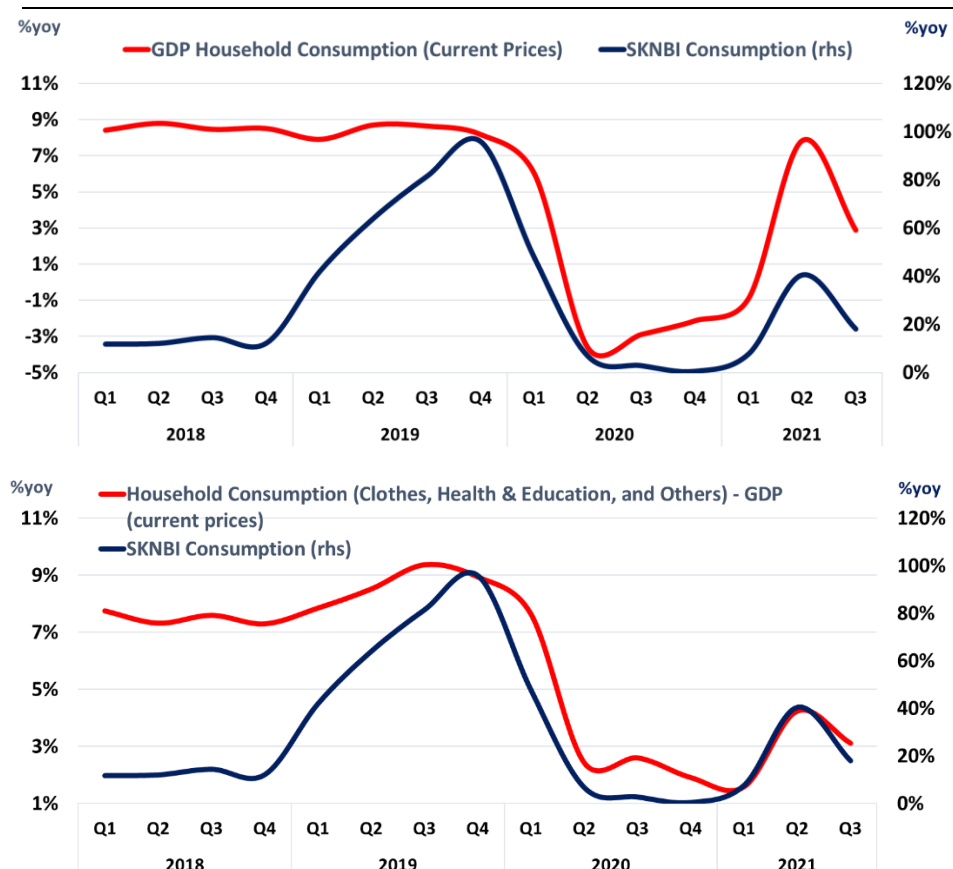
Table 7

| INDICATORS | CORRELATION OF SKNBI CONSUMPTION WITH GDP HOUSEHOLD CONSUMPTION (GROWTH, YOY) | | |
|--|---|-----------------------|-----------------------|
| | Q1-2018 to Q3-2021 | Q1-2019 to Q3-2021 | Q1-2020 to Q3-2021 |
| Total of Household Consumption | 55,6% | 87,8% | 93,9% |
| Clothes, Health & Education, Restaurant & Hotel, and Others | 61,2% | 85,9% | 90,5% |
| Clothes, Health & Education, and Others | 67,7% | 93,2% | 90,1% |
| Health & Education | 66,0% | 83,2% | 60,4% |
| Restaurant & Hotel | 51,0% | 70,7% | 80,0% |

Source: SKNBI, BPS (processed)

SKNBI Consumption Growth and Household Consumption
Growth – GDP (percent, y.o.y)

Figure 4



Source: SKNBI, BPS (processed)

5. Conclusion and Future Work

5.1 Conclusion

In this study, we have proposed a new approach in utilizing payment system transaction data as a proxy for household consumption indicators. Using text mining methodology with rule-based model, we can classify customer categories and transaction purposes from the SKNBI fund transfer data. Based on the evaluation of the model, the average F1-score of the model is 80%.

Using this methodology, we can obtain a proxy for household consumption indicators from high-frequency payment system data more quickly, compared to the official publication of GDP data. The validation results show a high correlation between these two indicators, which indicates that the household consumption indicator from the SKNBI fund transfer data can be used as a proxy for household consumption indicators.

5.2 Future Work

There are several improvements in the methodology that can be applied for future works.

1. Improving the methodology for classifying customers' categories and transaction purposes, including the use of machine learning algorithms.
2. Using consumption indicators from the payment system, e.g. fund transfers from SKNBI and payment transactions via cards, with other macroeconomic variables, to construct the nowcasting model of household consumption.

References

- Buono, D., Kapetanios, G., Macellino, M., Mazzi, G., & Papailias, F. (2018). *Big Data Econometrics: Now Casting and Early Estimates*. Universita Bocconi.
- Dunn, A., Hood, K., & Driessen, A. (2020). *Measuring the Effects of the COVID-19 Pandemic on Consumer Spending Using Card Transaction Data*. Bureau of Economic Analysis.
- Galbraith, J. W., & Tkacz, G. (2015). *Nowcasting GDP with Electronic Payments Data*. European Central Bank.
- Bank Indonesia. *Peraturan Anggota Dewan Gubernur Bank Indonesia Nomor 21/12/PADG/2019 tentang Penyelenggaraan Transfer Dana dan Kliring Berjadwal oleh Bank Indonesia*.
- Kapetanios, G., & Papailias, F. (2018). *Big Data & Macroeconomic Nowcasting: Methodological Review*. The Economic Statistics Centre of Excellence.
- Tarsidin, Idham, & Rakhman, R. N. (2016). *Nowcasting Konsumsi Rumah Tangga dan Investasi*. Bank Indonesia.



BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA



Big Data Analytics on Payment System Data for Measuring Household Consumption

Mohammad Khoyrul Hidayat, Muhammad Abdul Jabbar, Alvin Andhika Zulen
Statistics Department – Bank Indonesia
Email: moh_khoyrul@bi.go.id, muhammad_abdul@bi.go.id, alvin_az@bi.go.id

February 2022

*The views expressed here are those of the authors and
do not necessarily reflect the views of Bank Indonesia*

OUTLINE

1 Background

2 Data Source

3 Methodology

4 Result & Analysis

5 Conclusion





BACKGROUND

- Household consumption is one of the main indicators to measure state of the economy in Indonesia (largest contributor, 55%, in Indonesia's GDP). However, GDP data (incl. household final consumption expenditure) are published and available on a quarterly basis with a publication lag of one month.
- Bank Indonesia provides retail value payment system, i.e. SKNBI (The National Clearing System), that can generate data related to fund transfers, including household transactions.
- Advancements of technology and widespread use of payment systems have opened the opportunity to explore large dataset of payment data for monitoring economic activity.

OBJECTIVE

Developing a high frequency measure of household consumption in Indonesia from retail value payment system data (SKNBI), by utilizing Big Data Analytics methodology, particularly text mining.

Fund Transfer of SKNBI:

Credit transfer transaction between participants (banks) on behalf of the customers.



- ✓ **Total Transactions** \cong 12 mio trx/month
- ✓ **Nominal Transactions** \leq Rp. 1 Billion/trx
- ✓ **Availability Period** : July 2015 s.d. December 2021

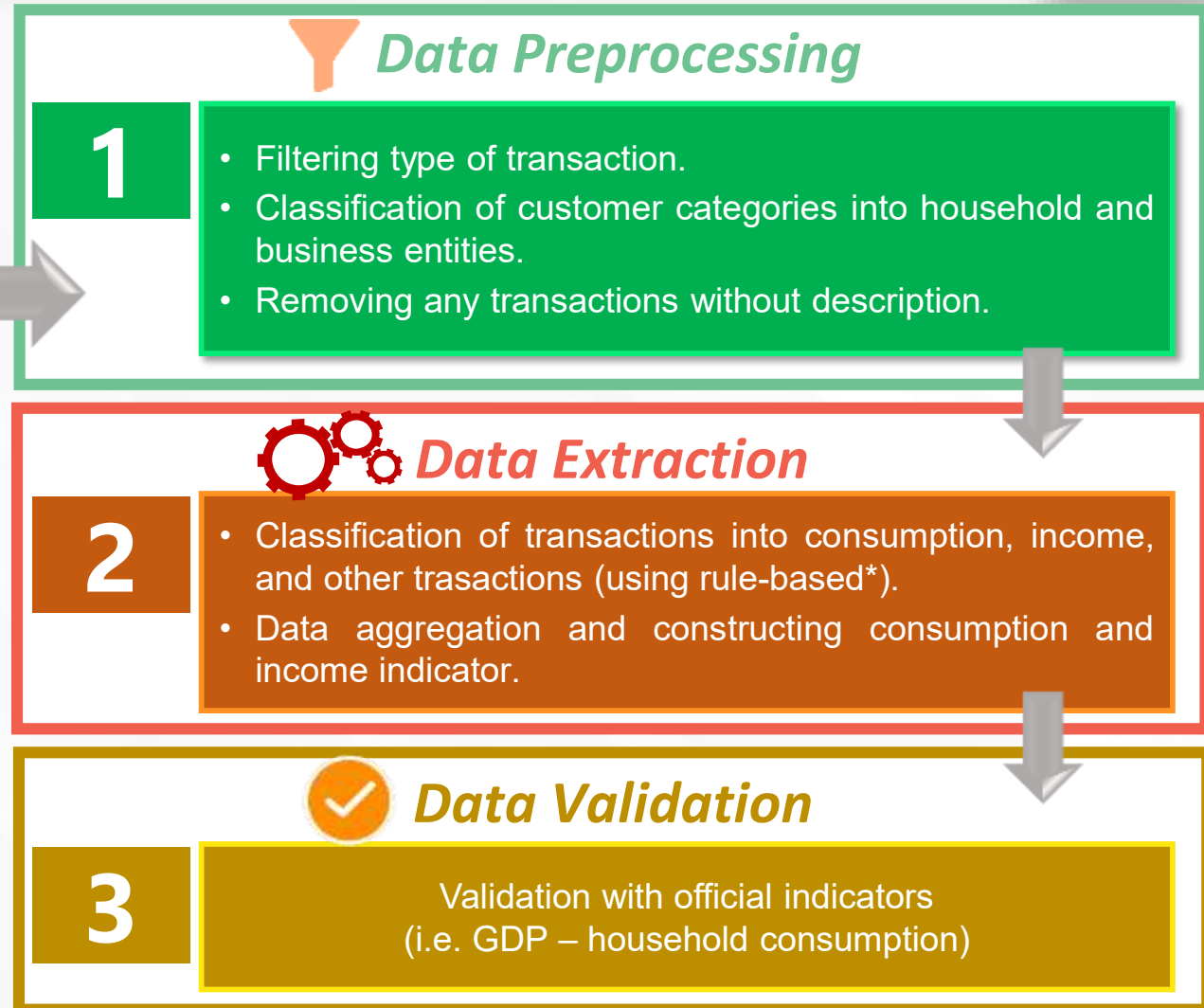
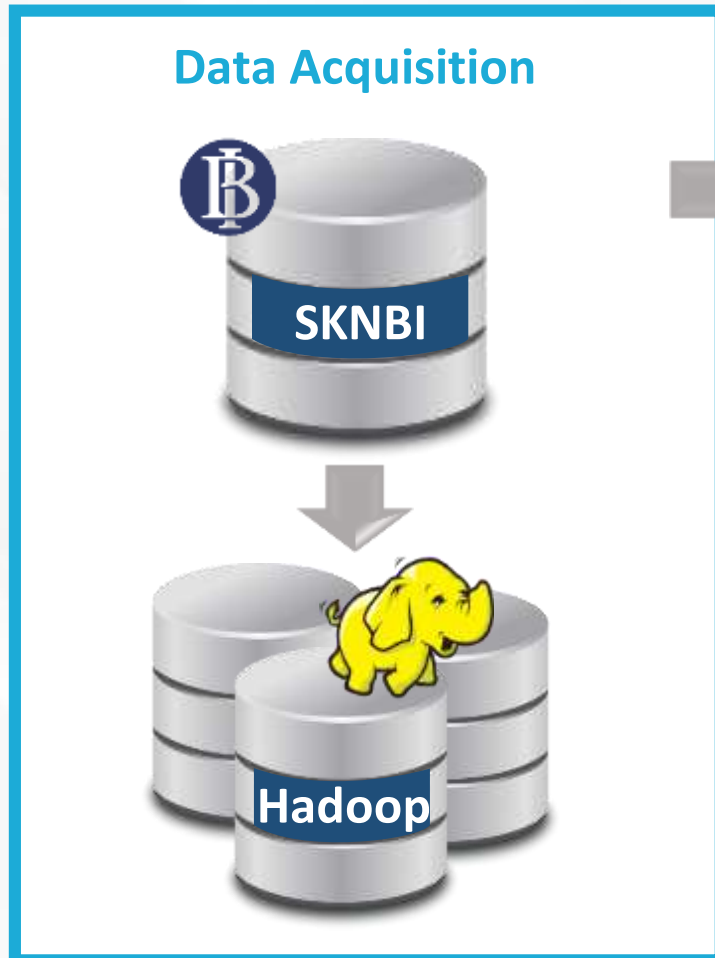
DATA SOURCE

DATA STRUCTURES

| COLUMN NAME | |
|-------------|-----------------------|
| 1 | DKE ID |
| 2 | BATCH ID |
| 3 | TRANSACTION DATE |
| 4 | ORIGINATING BANK CODE |
| 5 | SENDER LOCATION |
| 6 | BENIFICIARY BANK CODE |
| 7 | RECEIVER LOCATION |
| 8 | AMOUNT |
| 9 | TRANSACTION TYPE CODE |

| COLUMN NAME | |
|-------------|-----------------------------------|
| 10 | SENDER CUSTOMER'S NAME |
| 11 | SENDER CUSTOMER'S ACC NUMBER |
| 12 | SENDER CUSTOMER'S ADDRESS |
| 13 | SENDER CUSTOMER'S ID NUMBER |
| 14 | SENDER CUSTOMER'S TYPE CODE |
| 15 | BENEFICIARY CUSTOMER'S NAME |
| 16 | BENEFICIARY CUSTOMER'S ACC NUMBER |
| 17 | BENEFICIARY CUSTOMER'S ADDRESS |
| 18 | BENEFICIARY CUSTOMER'S ID NUMBER |
| 19 | BENEFICIARY CUSTOMER'S TYPE CODE |
| 20 | DESCRIPTION |

OVERALL WORKFLOW



*) e.g. : Consumption is SKNBI Fund Transfer with transaction detail containing keywords related to household consumption, e.g. : 'buy', 'shop', 'pay', 'paid off', 'installments', etc.

We use rule-based (keyword) approach for classifying customer categories.

'PT.', 'CV.', 'UD.', 'PD.', 'BANK', 'TBK', 'PERUSAHAAN', 'SDN BHD', 'INTERNATIONAL', 'INTERNASIONAL', 'FINANCE', 'LIFE', 'UNITED', 'RESTAURANT', 'ASURANSI', 'INSURANCE', 'AUTOMOTIVE', 'MANUFACTURING', 'MOTOR', 'LOGISTIC', 'LOGISTIK', 'PRODUCT', 'DEVELOPMENT', 'INDUSTRY', 'INDUSTRIES', 'PHARMA', 'CORP', 'PROD', 'DAIRY', 'LTD', 'GRAND', 'HOTEL', 'GREEN', 'CONSULTING', 'GROUP', 'GLOBAL', 'INTER', 'AGRO', 'RESORTS', 'TRANS', 'JASA', 'AJB', 'SECURITY', 'TEXTILE', 'INDO', 'SUKSES', 'RUMKIT', 'SUMBER', 'PERTAMINA', 'PLN', 'JASAMARGA', 'ELECTRONIC', 'INTL', 'POSCO', 'PTPN', '.CO', 'KANTOR PUSAT', 'PERUM', 'PROPERTY', 'RESEARCH', 'LIMITED', 'MEDICAL', 'ASTRA', 'HUSADA', 'TRADE', 'AGRICULTURE', 'FOOD', 'OPERATION', 'MAINTENANCE', 'SERVICES', 'STEEL', 'SUPPLIES'

**Business
Entities**

'OTORITAS', 'RSUD', 'BADAN', 'BPJS', 'PUSKESMAS', 'DINAS', 'KAB.', 'KABUPATEN', 'KEC.', 'KECAMATAN', 'KEL.', 'KELURAHAN', 'KPU', 'PEMERINTAH', 'PEMKAB', 'PEMKOT', 'PROVINSI', 'PROV.', 'PEMPROV', 'RKUD', 'EMBASSY', 'DITJEN', 'DIKBUD', 'CAMAT', 'DAERAH', 'PDAM', 'KERETA', 'KAS UMUM', 'RPKBUN', 'RPK-BUN', 'SPAN', 'SETDA', 'SEKDA', 'POLRES', 'POLDA', 'POLSEK'

Government

Example:

| Sender | Sender Category | Beneficiary | Beneficiary Category |
|--------------------|-----------------|------------------------|----------------------|
| INDO PRIMA SEMESTA | Business Entity | PT. UNITED FAMILY FOOD | Business Entity |
| RPKBUNP. SPAN BNI | Government | CV. TANJUNG AGUNG | Business Entity |
| TRI AMALIA | Others | PT. MAJU MOBILINDO | Business Entity |

METHODOLOGY – DATA EXTRACTION

Fund Transfer of SKNBI

Business

NULL

OTHERS

| | |
|-----------|-----------------|
| Sender | Business Entity |
| Recipient | Business Entity |

| | |
|-------------|--|
| Description | NULL, only number (123187937) or punctuation (-,/,:) |
|-------------|--|

| | |
|-------------|-----------------------------------|
| Description | Not related to consumption/income |
|-------------|-----------------------------------|

| | |
|-------------|-------------------|
| Recipient | others |
| Description | Related to Income |

| | |
|-------------|------------------------|
| Sender | others |
| Description | Related to Consumption |

We use rule-based (keyword) approach for classifying transaction purpose (consumption, income, business)

Household Income Keywords

'gaji', 'honor', 'upah', 'payroll', 'salary', 'remunerasi', 'insentif', 'wage', 'sales', 'pensiun', 'lembur', 'overtime', 'dividen', 'kompensasi', 'bagi hasil', 'bonus', 'claim', 'klaim', 'payoneer', 'komisi', 'tukin', 'uang makan'

Household Consumption Keywords

'pembayaran', 'konsumsi', 'belanja', 'pelunasan', 'kos', 'cicilan', 'dp', 'payment', 'dana', 'pelunasan', 'setoran', 'angsuran', 'jasa', 'invoice', 'listrik', 'service', 'sewa rumah', 'kasbon', 'catering', 'premi', 'asuransi', 'umroh', 'haji', 'tiket', 'spp', 'donasi', 'air', 'kuliah', 'obat', dsb.

Keywords that are not related to consumption/income

'retur', 'return', 'tabungan', 'refund', 'saving', 'reimburse', 'pemindahbukuan', 'nabung', 'tsa', 'span', 'pemerintah', 'pajak', 'sp2d', 'sppd', 'pendes', 'dana bos'

HOUSEHOLD INCOME

CONSUMPTION

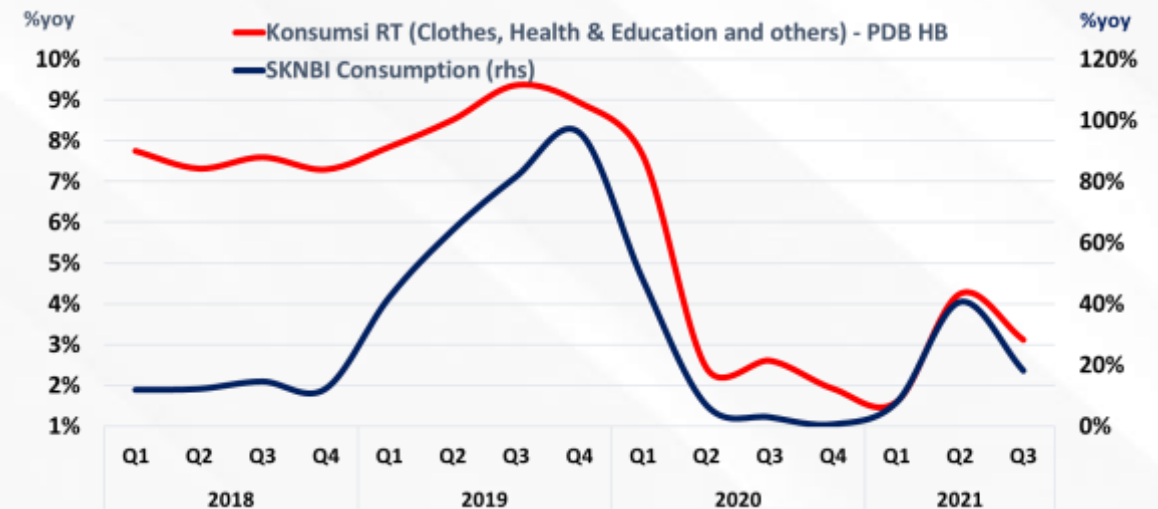
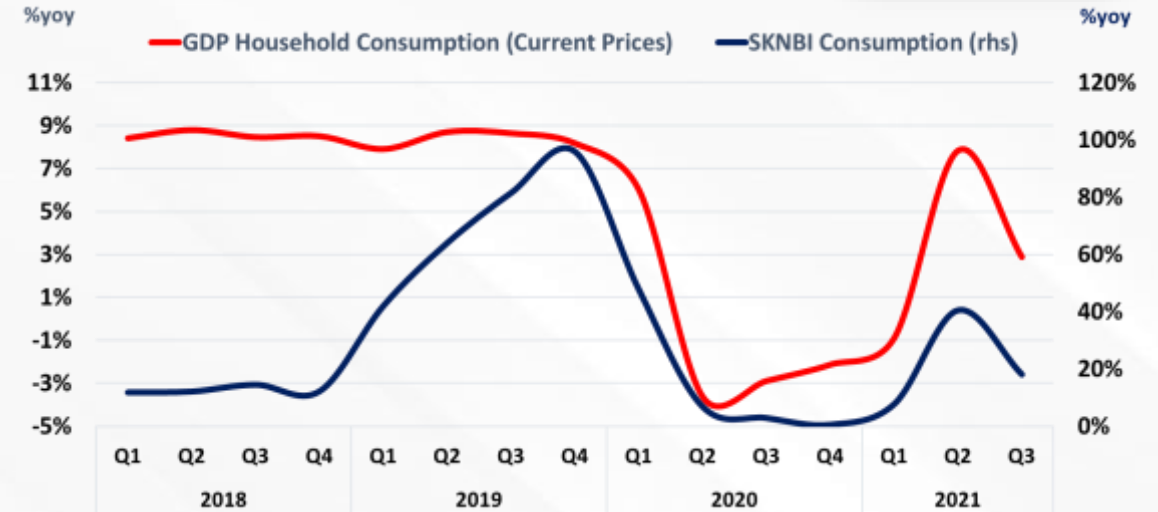
CONFUSION MATRIX EVALUATION RESULT

| | | PREDICTION | | | |
|----------------------------|-------------|-------------|--------|----------|--------|
| Purpose of Transaction | | Consumption | Income | Business | Others |
| A C T U A L | Consumption | 107 | 0 | 0 | 9 |
| | Income | 44 | 497 | 8 | 38 |
| | Business | 24 | 1 | 722 | 134 |
| | Others | 116 | 6 | 51 | 1.571 |

| Purpose of Transaction | <i>Recall</i> | <i>Precision</i> | <i>F1-score</i> | <i>F1-score (average)</i> | Accuracy |
|------------------------|---------------|------------------|-----------------|---------------------------|----------|
| Consumption | 92% | 37% | 53% | 80% | 87% |
| Income | 85% | 99% | 91% | | |
| Business | 82% | 92% | 87% | | |
| Others | 90% | 90% | 90% | | |

The validation results show a high correlation between the two indicators since quarter 1-2019, including during the Covid-19 pandemic.

| Indicators | Correlation of SKNBI Consumption Growth Rate with GDP | | |
|---|---|----------------------|----------------------|
| | Q1-2018 s.d. Q3-2021 | Q1-2019 s.d. Q3-2021 | Q1-2020 s.d. Q3-2021 |
| Total of Household Consumption | 55,6% | 87,8% | 93,9% |
| Clothes, Health & Education, Restaurant & Hotel, and others | 61,2% | 85,9% | 90,5% |
| Clothes, Health & Education and others | 67,7% | 93,2% | 90,1% |
| Health & Education | 66,0% | 83,2% | 60,4% |
| Restaurant & Hotel | 51,0% | 70,7% | 80,0% |



Conclusion

1. We have **proposed a new approach** in utilizing high-frequency payment system transaction data, **using text mining methodology through a rule-based model**, to construct a proxy indicator for household consumption. Based on the evaluation of the model, the average F1-score of the model is 80%.
2. Our consumption indicator can be **generated from payment system data more quickly** compared to household consumption indicators in GDP publications. The validation results show a **high correlation** between our consumption indicator from payment system and publication of GDP data, which indicates that the indicator from payment system can be used as **a proxy for household consumption indicators**.

Future Works

1. Improving the methodology for classifying customer categories and transaction purposes, including the use of machine learning algorithms.
2. Using consumption indicators from payment system, e.g. funds transfer from SKNBI customers or payment transactions via cards, with other macroeconomic variables, to construct the nowcasting model of household consumption.



BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

THANK YOU TERIMA KASIH

Mohammad Khoyrul Hidayat, Muhammad Abdul Jabbar, Alvin Andhika Zulen
Statistics Department – Bank Indonesia
Email: moh_khoyrul@bi.go.id, muhammad_abdul@bi.go.id, alvin_az@bi.go.id

The views expressed here are those of the authors and do not necessarily reflect the views of Bank Indonesia

IFC-Bank of Italy Workshop on “Data Science in Central Banking: Applications and tools”

14-17 February 2022


Tracking the economy during the Covid-19 pandemic: the contribution of high frequency indicators¹

Jérôme Coffinet, Jean-Brieux Delbos, Jean-Noël Kien, Étienne Kintzler,
Ariane Lestrade, Michel Mouliom, Théo Nicolas and Vojtech Kaiser,
Bank of France

¹ This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

TRACKING THE ECONOMY DURING THE COVID-19 PANDEMIC: THE CONTRIBUTION OF HIGH-FREQUENCY INDICATORS

IFC-BANK OF ITALY WORKSHOP “DATA SCIENCE IN CENTRAL BANKING: APPLICATIONS AND TOOLS”



The views expressed in the presentation do not necessarily represent those of the Banque de France, the ACPR or the Eurosystem

JEROME COFFINET
BANQUE DE FRANCE/ACPR



INTRODUCTION

- The so-called “open” data used in this presentation have three main characteristics:
 - they are not included in the official statistics published by institutions and authorities;
 - they can be accessed easily and free of charge via the internet;
 - they are available at a high frequency, generally at least daily.
- Open data are essentially statistics derived from social media platforms (Twitter), the internet (Google searches, webscraping), granular open data databases (e.g. statistics linked to energy and transportation).
- Open data have proved extremely useful during the Covid-19 crisis for tracking fluctuations in economic activity:
 - bank card data (Carvalho et al., 2020);
 - electricity consumption (Cicala, 2020);
 - weekly unemployment figures (Coibion et al., 2020);
 - or the real-time location of global cargo ships (Cerdeiro et al., 2020).

COVID-19: A SOURCE OF CONCERN FOR INTERNET USERS

- Google searches for the word “overindebtedness” point to another way in which the crisis has affected households, in two phases:
 - First, a turning point after the introduction of the lockdown (17 March 2020), with searches for overindebtedness dropping to very low levels; then from mid-April onwards searches begin to rise again.
 - The physical restrictions imposed during the lockdown appear to have discouraged households from seeking information about overindebtedness – at least initially. However, the length of the lockdown appears to have weighed heavily on their finances, especially for the most vulnerable households, a month after the start of the lockdown, the issue of overindebtedness resurfaces.
- Very interestingly, these data are consistent with the massive fall in the number of overindebtedness applications submitted to the Banque de France between February and May 2020 (fall of more than 60%), and the marked upturn as of June (rise of more than 70% between May and June 2020).

C5 Google searches for the word “overindebtedness”

(2 February-11 May 2020)



Source: Google.

- ACPR
BANQUE DE FRANCE

Source: Twitter.

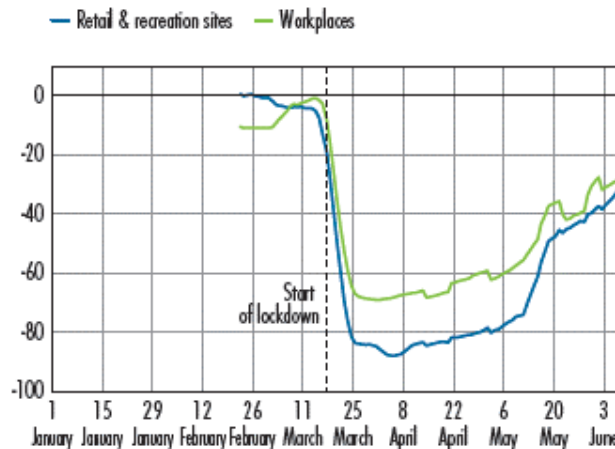
4

THE EFFECTS OF THE LOCKDOWN ARE REAL AND CAN BE MEASURED PHYSICALLY: MOBILITY AND POLLUTION

- Google provides a Google Mobility Index, which is based on Google Maps data on the number and length of visits to certain locations. In France, movement to leisure sites and places of work fell by 85% during the lockdown. After the lockdown was lifted mobility increased, but in June it was still 30% below its baseline level.
- Air pollution (corrected for the impact of temperatures, atmospheric pressure, wind speed and air humidity) also declined with the introduction of the lockdown. Nitrogen dioxide pollution dropped well below historical levels, reflecting the stoppage of industrial sites and of a portion of transportation activities. The decline can be seen in all French towns and Paris.

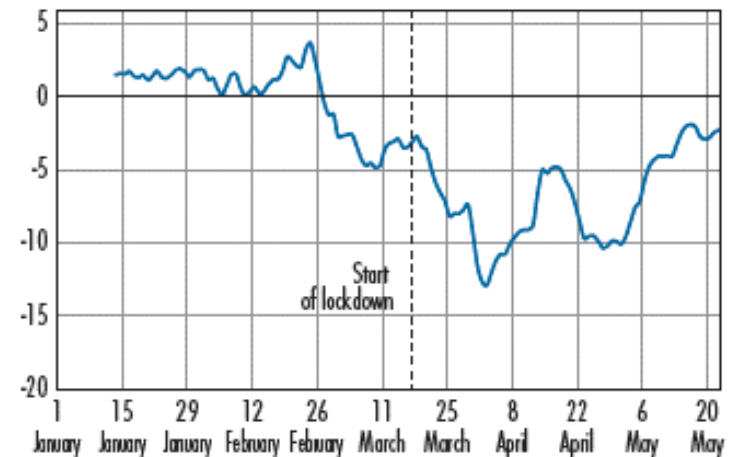
b) Mobility in France according to Google Mobility Indices

(% change versus baseline = average from 3 to 6 February 2020, 7-day moving average)



c) NO₂ pollution in Paris, corrected for meteorological data

(difference versus 2019, 14-day moving average)

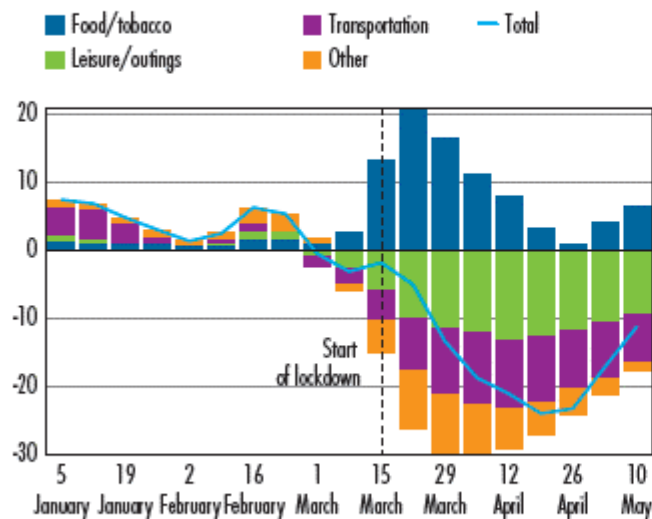


THE EFFECTS OF THE LOCKDOWN ARE REAL AND CAN BE MEASURED PHYSICALLY: TOURISM

- Google searches linked to consumer spending, dropped sharply following the lockdown in the categories “leisure” and “transportation”. Conversely, the category “food and tobacco” hit a peak during the lockdown.
- Tourism collapsed. The number of new reviews posted on Airbnb plunged by 99%, meaning that activity evaporated. Unlike the other indicators, which increased again once the lockdown was lifted, this one remained flat at end-April due to the continuing shutdown of the tourist industry.

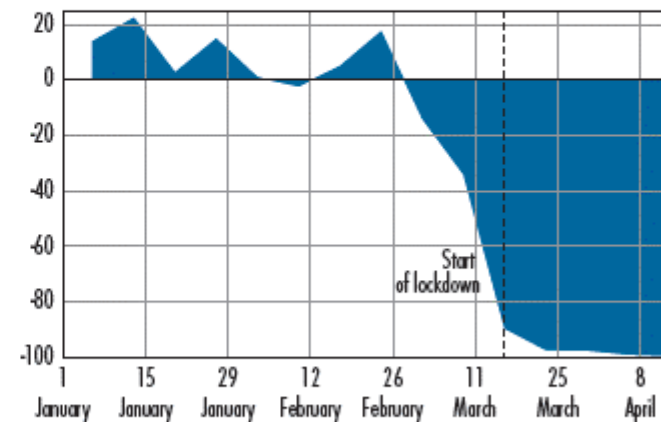
e) Popularity of searches linked to consumer spending according to Google Trends Indices

(% change versus 2019, 2-week moving average)



f) Number of new reviews on the Airbnb platform in Paris

(% change versus 2019)

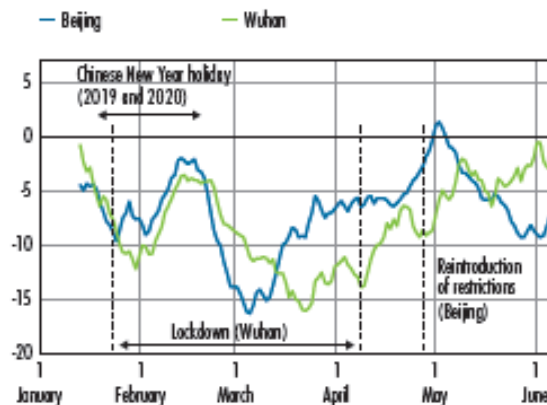


Sources: Google Covid-19 Community Mobility Reports, OAG Aviation Limited, World Air Quality Index, Google Trends, InsideAirbnb, authors' calculations.

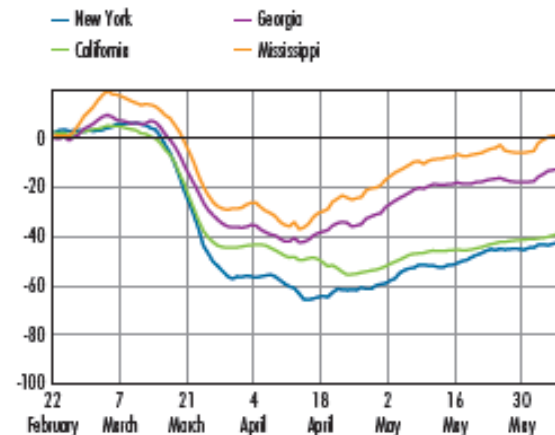
THE EFFECTS OF THE LOCKDOWN ARE REAL AND CAN BE MEASURED PHYSICALLY: GEOGRAPHICAL DISPARITY

- The indicators also make it possible to analyze the different stages of the lockdown lifting in each country:
 - Some indicators highlight regional disparities within countries. This is notably the case for the NO₂ pollution indicator in China. The situation differs in Beijing: the faster lifting of the lockdown meant that pollution returned more quickly to its 2019 levels; however, the reintroduction of certain lockdown measures at the start of April could be behind the renewed decline in pollution – albeit a smaller one than during the first lockdown;
 - Similarly, with the Google Mobility Indices in the United States, there is a striking contrast between the first federal states to be affected by the lockdown (New York and California) and others where the lockdown was introduced later, for a shorter duration and with less stringent rules (Georgia or Mississippi).

c) NO₂ pollution, corrected for meteorological data
(difference versus 2019, 14-day moving average)



d) Google Mobility Indices
(% change versus period from 3 January to 6 February 2020, 7-day moving average)



Sources: Google Covid-19 Community Mobility Reports, OAG Aviation Limited, World Air Quality Index, Google Trends, InsideAirbnb, authors' calculations.



CONCLUSION

- High-frequency indicators can provide information on their own, supplement existing indicators with additional or more immediate data, or be used to estimate other variables (production, consumption, GDP, etc.) using econometric methods.

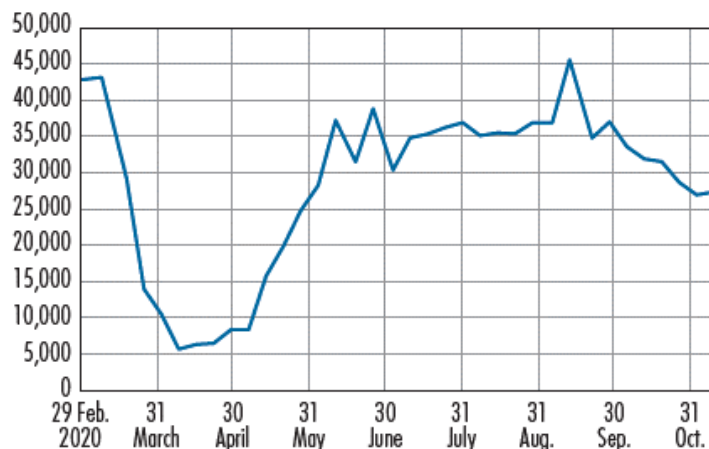
- Ongoing issues:
 - Assess the added value in normal times and link these high-frequency data with more conventional economic variables;
 - IT and legal matters for collecting, maintaining and releasing high-frequency data;
 - New avenues:
 - Use of satellite data to assess climate risk exposure and building activity;
 - Social media sentiment to gauge inflation expectations;
 - Use of webscraping data to gauge the risk on financial actors (including real estate transaction negotiation margins, reputational risk, etc.).

PERSPECTIVE: COVID-19 AND HOUSE PRICES, WHAT CAN BE LEARNED FROM WEB-SCRAPING DATA?

- Daily automatic upload of data on the Internet from the major real estate classified ad sites in the UK: Rightmove, Zoopla, OnTheMarket, PropertyPal.
- 1.5 million real estate listings are downloaded on average every day, corresponding to the total stock of available listings. These data confirm that during the first lockdown period, activity stalled and sellers adopted a wait-and-see attitude.
- The granular approach also reveals significant regional differences: while advertised prices were stable or even increased after the first lockdown in more rural areas, they declined continuously in London.

C1 New weekly property listings

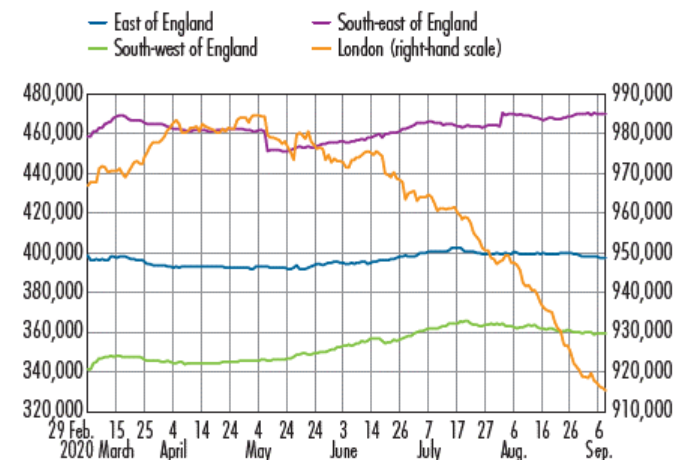
(number)



Sources: Zoopla and authors' calculations.

C3 Average listing price by region

(in GBP by property)



Sources: Zoopla and authors' calculations.