Irving Fisher Committee
on Central Bank Statistics

IFC Bulletin
No 57

Machine learning in central banking

November 2022

BANK FOR INTERNATIONAL SETTLEMENTS

Contributions in this volume were prepared for the proceedings of the IFC-Bank of Italy Workshop on "Data Science in Central Banking", Part 1: Machine learning applications, Rome (virtual event), 19-22 October 2021.

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

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

ISSN 1991-7511 (online)

ISBN 978-92-9259-607-1 (online)

Machine learning in central banking

IFC Bulletin No 57
November 2022

Proceedings of the IFC-Bank of Italy Workshop on "Data Science in Central Banking", Part 1: Machine learning applications

Rome (virtual event), 19-22 October 2021

Overview

Opening remarks

Keynote speech

1. Introduction: increased central bank use of ML techniques

2. Gathering better and more information

Irving Fisher Committee on
Central Bank Statistics

BIS

Machine learning applications in central banking

Douglas Araujo, Giuseppe Bruno, Juri Marcucci, Rafael Schmidt, Bruno Tissot[1]

Executive summary

On 18–22 October 2021, the Irving Fisher Committee on Central Bank Statistics (IFC) and the Bank of Italy co-organised, with the support of the European Central Bank (ECB) and the South African Reserve Bank (SARB), a **workshop on "Data science in central banking" that focused on machine learning (ML) applications**. This event was an opportunity to take stock of how central banks are deploying ML across a variety of use cases. It also illustrated the importance of these new techniques in improving the efficiency and effectiveness of their related operations, including by increasing their availability to deal with larger and new sources of information in a more automatised way.

Indeed, the workshop underlined the diversity and maturity of ML approaches already developed and used by central banks. **This reflects their potential and usefulness for central banks in dealing with the increasingly complex environment in which they operate**.

To start with, the new techniques can **help gather more and better information**, which is key for central banks that rely heavily on data. ML can help respond to this demand by enhancing the data quality, eg dealing with outliers, addressing the problems posed by missing values, limited frequency and/or timeliness, and by providing richer contextual insights.

In addition, **a key issue for central banks is to make sense of the wealth of data available to derive useful insights on specific economic and financial situations.** This needs to happen in a reasonably fast and largely automated fashion, considering the constantly changing environment. Coping with the often exponential growth of data and associated complexity of the statistical analysis is a challenge for central bank statisticians. Fortunately, ML can greatly help central banks in this

[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 & Head of the Secretariat of the Irving Fischer 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 ECB, the IFC, the SARB or any of those institutions represented at the workshop.

We thank Thomas Gottron, Robert Kirchner, Silke Stapel-Weber and Ulf von Kalckreuth for helpful comments and suggestions.

context by facilitating the modelling of economic and financial problems and supporting the related statistical exercises.

In turn, **the insights gained can effectively back the conduct of evidence-based central bank policies.** This is obviously the case regarding monetary stability, not least in terms of better understanding the drivers of monetary policy decisions that can be provided by ML. Similarly, applying ML in suptech can be instrumental in helping financial supervisors to perform their oversight tasks, including identifying and tackling micro-level fragilities and other emerging threats such as climate-related financial risks. Turning to the macroprudential perspective, central banks can benefit from the increased use of ML to interpret information from various, often unrelated, data sources to assess system-wide vulnerabilities and their evolution over time. Moreover, the new techniques can support other tasks that are also relevant from a financial stability perspective, including the functioning of the payment system, financial inclusion, consumer protection, anti-money laundering and the secure printing of money.

At a more practical level, the workshop provided useful benchmarking, feedback and training on ML models for the participants. **Several lessons and observations are worth noting for those in charge of deploying ML-based tools in their central banks.**

First, there is a wealth of alternative information sources that have barely been tapped by central banks and which can provide new, useful insights if explored with ML techniques. The ultimate goal is that policymakers have at their disposal better-quality, timelier and interpretable data when taking decisions, especially in uncertain times such as the Covid-19 pandemic. Second, complementarity is essential: ML methods can provide additional insights to traditional approaches but have to be blended with other types of exercises as well as with strong business expertise. Third, there are benefits to calibrating many ML tools, not just one, since combining different approaches can provide better results with usually limited additional effort. Particular emphasis needs to be placed on avoiding ML model overfitting, eg through cross-validation. Fourth, there is merit in following a pragmatic and gradual approach when implementing the new tools. A considerably varied set of ML methods can be considered, and it is important to carefully assess them before actual deployment, with due consideration of the available skill set and computing environment. Fifth, having more data is often better than increasing the sophistication of the ML model. Sixth, while ML can be instrumental in dealing with complexity, there is also a risk of developing black box solutions that would compound the challenges faced by users as their functionality is rarely intuitive. The focus should therefore be on the interpretability of the results obtained and on addressing well defined use cases. Lastly, ML exploratory work has only started, and substantial staff and IT investment as well as business adjustments will continue to be needed to make the most of the new techniques, computing equipment and data.

Addressing these issues will require **further modifications in central banks' current operational processes** – eg in developing software ("DevOps") and putting ML algorithms into production ("MLOps") – **and collaboration models** – with close cooperation between core IT experts, data scientists and business specialists. It also puts a premium on the IFC's mission to promote cooperation between central banks through the sharing of national use cases and to draw relevant lessons from the experiences observed outside the public community.

1. Introduction: increased central bank use of ML techniques

One of the IFC's raisons d'être is to foster cooperation between central banks on statistical issues based on showcasing projects and sharing national experiences. To this end, **the Committee has initiated recurrent workshops on "Data science in central banking"** aimed at a broad audience of practitioners and technicians, with the goal of reviewing the adoption of data analytics and business intelligence techniques and developments in the big data ecosystem. The first event, hosted by the Bank of Italy in October 2021 with the support of the ECB and the SARB, focused on the contribution of ML applications to central banking. This virtual event was attended by almost 500 participants, representing about 180 institutions from the public and private sectors.

ML can be defined as an algorithm – a method of designing a sequence of actions to solve a problem – that optimises automatically through experience (ie from data) and with limited or no human intervention (FSB (2017)). ML algorithms are a subset of Artificial Intelligence (AI) techniques and are typically divided into four main types: supervised, unsupervised, reinforcement and deep learning (Wibisono et al (2019)). They have been increasingly used in economic and financial academic and practitioner settings, and central banks are not far behind. One reason is that the **compilation of large and/or complex granular databases (Israël and Tissot (2021)) and the development of big data analytics (IFC (2019)) have spurred their ability to use ML tools** to support the conduct of their policies, especially in the areas of monetary and financial stability, including the associated statistical, analytical and communication tasks (Chakraborty and Joseph (2017), Doerr et al (2021) and Bruno and Marcucci (2021)).

Indeed, the IFC workshop highlighted the **diversity of ML approaches developed in central banking**. Authorities are exploring, and in some cases already deploying, ML techniques to support a wide range of use cases that encompass macroeconomic modelling, economic and inflation analysis, the support of monetary and financial stability policies (including microprudential tasks for those central banks in charge of financial supervision) and specific statistical work (eg detection of data anomalies). Moreover, the range of central banks involved in this exploratory work is broad and comprises most advanced economies as well as a growing number of emerging market economies.

The fact that diverse ML tools have been successfully applied across a wide spectrum of use cases underscores the great value of the analytical insights they can provide as well as the operational gains brought about by automatising and making more efficient various production processes. **One key benefit for central banks is, in particular, the ability to deal with the increasingly complex environment in which they operate**. As regards monetary policy for example, the communication of policy decisions has become increasingly important and multifaceted, especially after the Great Financial Crisis (GFC) of 2007–09 (Gros (2018), Cieslak and Schrimpf (2019) and Hansen and McMahon (2018)); it has in particular benefited from the use of natural language processing (NLP) techniques (Gentzkow et al (2019)) to facilitate dealing with textual information (Apel et al (2021), Ferreira (2021), Hansen et al (2018) and Ahrens and McMahon (2021)). Another example relates to financial supervisory tasks, for which new and complicated topics are constantly emerging, such as those related to climate-related financial risks (Hernández de Cos (2022)) or to the consequences of the Covid-19 pandemic (Casanova et al (2021)); it has in fact been argued that ML can increase central banks' efficiency in the supervisory area by helping them cover more ground with the same resources (Beerman et al (2021)).

To be successful, **ML projects require the availability of adequate staff and IT resources as well as good coordination with the business areas** (IFC (2020a)). Fortunately, the wide variety of the techniques already deployed by central banks suggests that they have been able to both rely on adequate human skills – including subject matter and IT experts, and data scientists – and address the associated complex IT requirements. They have benefited in particular from the fact that many ML computing frameworks are widely available in the public domain as "open source" – cf scikit-learn (Pedregosa et al (2011)), PyTorch (Paszke et al (2019)) and TensorFlow (Abadi et al (2015)). Moreover, new approaches such as transfer learning are further improving the accessibility to central banks of large, high-performance ML models (Box 1).

The above developments have allowed central banks to take advantage of the wide range of ML techniques to address increasingly sophisticated use cases, as illustrated in Professor Michael McMahon's keynote speech. For instance, traditional NLP techniques can be combined with algorithms that recognise the temporal dimension of texts (cf Chang and Manning (2012)) to assess the information content of monetary policy statements. Another interesting case is the use of NLP techniques calibrated according to specific macroeconomic variables to analyse central bank communication.

Yet the continuous development of ML algorithms and related areas, such as big data analytics and cloud computing, is likely to require further substantial investment in skilled staff (eg data scientists). As emphasised in Bank of Italy Deputy Governor Piero Cipollone's introductory speech, the necessary skills transcend the technical to include also the ability to recognise and address the challenges and risks inherent in data science, such as the presence of bias in big data sets and the imperative to consider data integrity, confidentiality and privacy.

Further investment in IT equipment remains necessary. In fact, many central banks are in the process of implementing, or have already implemented, a cloud adoption strategy to, inter alia, better enable ML/high-performance computing use cases and to facilitate related collaboration with external researchers (as observed eg in the case of the Bank of Canada). A key reason is that cloud computing offers more agility for data analysis and experimentation; access to computing power is easier to scale up, and, importantly, the responsibility for maintaining an updated hardware and software environment in a so-called "plug-and-play model" (ie requiring little involvement by the end user in the necessary IT setting) can be shifted from internal staff to external service providers.

To sum up, the availability of large sets of data from many different and unstructured sources along with the development of new, innovative analytics and IT tools are **changing the financial landscape in which central banks operate**. This provides a key opportunity to leverage (automatised) ML algorithms to strengthen their economic and financial modelling toolkits, analytical capabilities and risk management tools, in turn helping them to pursue their mandates more effectively.

The following sections elaborate on the various types of ML applications that can be used effectively in view of the projects presented at the workshop. Section 2 describes ways to support central banks in making the best use of the data available as a key input for their operations. Sections 3 to 6 outline ML use cases in the specific areas of macroeconomic analyses, monetary policy, micro-financial supervision and (macro-)financial stability.

Box 1

Facilitating the use of machine learning by central banks with transfer learning

Douglas Araujo

Some ML models aim to achieve or even exceed human-level performance based on large and complex information, such as big data sets, text or images. Using them typically requires powerful and sophisticated IT systems due to their training on vast amounts of data. Such models can only achieve the desired performance if they are trained with a correspondingly large amount of data. For example, GPT-3,[1] a model famous for its ability to create human-like text, comprises 175 billion parameters that were trained on petabytes of text from two broad internet text corpora, numerous books and the whole of English language Wikipedia. Thus, the development of models with the highest performance in certain fields is usually done only by a few organisations with sufficient resources and specialised computer engineering capabilities.

Transfer learning is a technique that facilitates the use of these large models. The starting point is to download a pre-trained version of them and then fine-tune their specifications to the particular use cases at hand. In practice, organisations such as big tech firms (eg Google, Meta) or boutique ML firms (eg HuggingFace, DeepMind) will typically train reference models on selected big data sets. Then, they usually make the models publicly available by storing them in so-called "model hubs",[2] where the general public can search, compare and download desired models. Once a suitable pre-trained model is selected, the user can run it off the shelf on more specific data sets, or more commonly, fine-tune the model to the data for that particular use case. Importantly, given all the extensive pre-training work already done, the fine-tuning step can be effective even with a limited volume of data. Thus, the user benefits from high performance without the need to spend considerable time and resources creating or refining the model. This technique also facilitates the use of several large models for comparison or as an ensemble.

Transfer learning can help advance the use of ML by central banks. First, there is a broad range of models available in the main hubs, so that central banks can easily find adequate models to address a wide range of specific use cases they might have. This supports central banks in jumpstarting ML-powered projects in different areas without having to invest excessive resources or time. Second, central banks can explore and combine different models depending on the circumstances and analytical needs. Third, comparing these external "state-of-the-art" models with their own models and algorithms can help central banks enhance the performance and accuracy of such internally developed applications.

However, there are important challenges associated with transfer learning. One relates to quality assurance: model hubs are typically administered by reputable institutions, but in many instances the models themselves are developed and posted in the hubs by third parties. Hence, proper care during model selection is warranted. For instance, it is important to correctly test models trained by third-party entities to avoid biases or limitations that might be important in the particular use case of interest. Another issue to bear in mind is that, because transfer learning entails downloading the original models, the application would need to be (often manually) updated in case a new version of the original model is made available.

In summary, transfer learning can facilitate the use of high-performance ML models to support a variety of central bank applications, in turn supporting their use of big data analytics and ML tools. Moreover, transfer learning is operationally much simpler than developing models from scratch and managing them, reducing the associated burden in terms of resources (eg necessary skill set, IT equipment, budget). Finally, the development phase of ML models may require considerable energy. It has, for instance, been reported that the CO_2 emissions from training a single large model can reach multiples of the emissions of an average car during its whole lifetime.[3] By enabling the use of pre-trained models instead of developing very similar models by multiple central bank users, transfer learning can help to limit the carbon footprint from ML usage.

[1] See T Brown, B Mann, N Ryder, M Subbiah, J D Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, A Herbert-Voss, G Krueger, T Henighan, R Child, A Ramesh, D Ziegler, J Wu, C Winter, C Hesse, M Chen, E Sigler, M Litwin, S Gray, B Chess, J Clark, C Berner, S McCandlish, A Radford, I Sutskever, D Amodei, "Language models are few-shot learners", *NeurIPS Proceedings*, 2020.

[2] Examples of widely used model hubs are TensorflowHub (https://tfhub.dev), PyTorchHub (https://pytorch.org/hub) and HuggingFace Models (https://huggingface.co/models).

[3] E Strubell, A Ganesh and A McCallum, "Energy and policy considerations for deep learning in NLP", *ACL*, 2019.

2. Gathering better and more information

Apart from being a key pillar of national statistical systems as producers of official statistics, central banks are heavy users of information to support the conduct of their policies, which are increasingly based on quantitative evidence. ML can help to respond to this appetite by enhancing the data quality, eg dealing with outliers and addressing the problems posed by missing values, limited frequency and/or timeliness, and by providing richer contextual insights.

Setting up adequate quality assurance frameworks

The growing availability of large and complex granular data sets ("financial big data"; IFC (2015)) obtained from statistical or supervisory reporting often at the transaction level and at very high frequencies (such as daily), has amplified **the need for better and faster data quality management frameworks** to ensure that the information collected can be reliably used for statistical production. In particular, many central banks have been leveraging on ML algorithms, sometimes in combination with traditional methods, to develop new data validation processes to better check the quality of the data at stake and correct them more effectively and/or efficiently.

One recent example is the ECB's new **anomaly detection project to support data quality checks in the production of statistics** on euro short-term interest rates. Their compilation is derived from a granular data set on individual transactions observed in money markets, ie the Money Market Statistical Reporting (MMSR), which involves 47 banks located in 10 countries and represents a total of around 50,000 daily transactions. There were important challenges related, in particular, to the presence of non-numerical variables, distribution skewness, the need for rapid data quality checks to support a daily production process, and the difficult interpretability of the results obtained for the users. These challenges were addressed with the use of various ML techniques to convert categorical variables into numerical ones, exploit the observed correlations, and detect anomalies through different models/algorithms, namely: standard regression analysis to compare observed data with model estimates; isolation forest and hierarchical clustering to isolate specific data points from the rest of the distribution; anomaly identification, based on the training on past data (supervised learning) using XGBoost (Chen and Guestrin (2016)); and the use of the Local Interpretable Model-Agnostic Explanations (LIME) algorithm

(Ribeiro et al (2016)) to facilitate the interpretation of the algorithm's results that could otherwise resemble a "black box" and provide users with a more interpretable model.

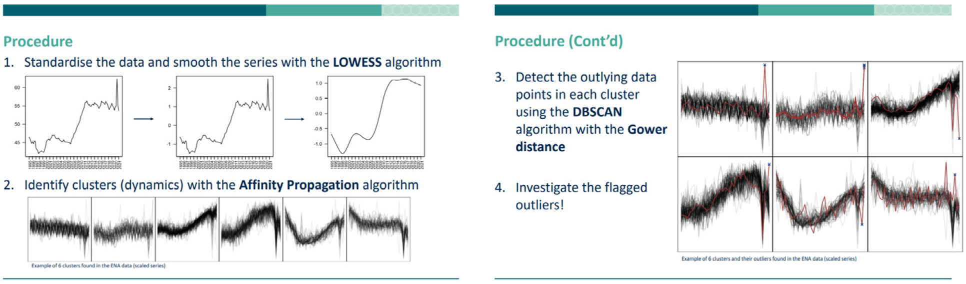
Another example is the BIS initiative to develop a **highly automated validation workflow relying on ML** tools and enhanced computer capacity. The data validation approach implemented is reported to be suitable for a large volume of indicators – about 3,000 daily time series of financial market data. It therefore clearly outperforms more traditional methods, such as graphical controls or threshold-based warnings. Moreover, the new solution appears better able to address the risk of errors that are "Type I" or "false positive" (ie mistaken rejection of the existence of an anomaly that in fact exists) as well as "Type II" or "false negative" (ie failure to reject the existence of an anomaly while the data are in fact correct). The workflow starts with the checking of potential data gaps, followed by a reduction in the dimensionality of the problem (by concentrating on a smaller set of series), and the use of a long short-term memory (LSTM) artificial neural network[2] that can process entire sequences of data (and not only single data points) to estimate prediction errors and detect possible anomalies.

In addition to facilitating the handling of a large number of series, the use of ML techniques can help to better deal with the fact that macroeconomic time series are often subject to sudden and unexpected shocks (eg the Covid-19 pandemic). **These changes imply that data quality monitoring procedures can be constantly challenged** as time passes. To address this issue, the approach developed jointly at the Bank of England and the ECB is based on a clustering procedure in order to automatically identify anomalies within an evolving database. This is done by analysing the correlations within the observations, representing 6,638 single time series from 31 countries in that particular case. The process involves the standardisation of the data, the smoothing of the series with a specific filter (the LOWESS algorithm), the identification of specific clusters using a dedicated ML algorithm (Affinity Propagation (AP); Frey and Dueck (2007)), and the detection of potential anomalies within each cluster through an algorithm grouping together similar observations – with the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) technique based on a specific metric that measures dissimilarities, the "Gower distance" (Graph 1). The method is data-driven, appears more robust to systemic shocks and allows for a high level of automation.

[2] Neural networks (or "artificial neural networks") are ML algorithms that transmit a signal from one processing node to another (loosely analogous to the interactions between brain neurons) to identify non-linear relationships in the data. Such models are deemed capable of capturing and representing complex relationships (Richardson et al (2019)).

Procedure to detect anomalies developed at the Bank of England and the ECB

Graph 1



Procedure

1. Standardise the data and smooth the series with the **LOWESS** algorithm

2. Identify clusters (dynamics) with the **Affinity Propagation** algorithm

Example of 6 clusters found in the ENA data (scaled series)

Procedure (Cont'd)

3. Detect the outlying data points in each cluster using the **DBSCAN** algorithm with the **Gower distance**

4. Investigate the flagged outliers!

Example of 6 clusters and their outliers found in the ENA data (scaled series)

Source: A Maurin and N Benatti, "Time series outlier detection, a data-driven approach", *IFC Bulletin*, no 57, November 2022.

Outlier detection tools

One important focus point of the new quality approaches that are leveraging ML techniques is to **detect outliers in the vast and increasing amount of observations now collected in real time** by big data repositories, which are making traditional manual actions performed by humans (eg use of spreadsheets and simple graphical tools) increasingly inefficient if not impossible to perform. To address these issues, the Bank of Israel has developed dashboards using a specific package (R Shiny app, which uses the R programming language for statistical computing and graphics) that can check all the daily transactions in derivatives markets reported by financial institutions. Once the data are uploaded into a dashboard, the users can choose the variables to analyse, add filters, explore the data graphically and practice specific outlier detection algorithms – including detection graphical tools (eg Bootlier Plot) based on density histograms, isolation forest, etc.

The Deutsche Bundesbank has also adopted **unsupervised ML algorithms to detect outliers for a wide range of voluminous financial data sets** – eg on interest rates, money market statistics, sectoral securities holdings, investment fund holdings that differ markedly in terms of size (from 25,000 to 5 million rows), features (from 12 to 150) and number of outliers (from 0.04% to 5%). The approach relies on various ML algorithms to group information in specific clusters (eg tree-based methods like the isolation forest), assess dissimilarities (eg distance/density-based classification methods like the K-nearest neighbours (KNN) algorithm), compressing the information to be analysed (eg reduction in the size of input data that can be reconstructed afterwards with greater details through the use of self-supervised ML tools such as autoencoders),[3] and generate explanations so that humans can understand the decisions or predictions made (eg use of explainable AI/ML (XAI/XML) techniques like the Shapley value approach).

[3] An autoencoder is a type of neural network that learns the main features of the input information by constructing a lower-dimensional representation of it (similar to transforming an original photo into a lower-resolution version) and then reconstructing it (Rubio et al (2020)). The objective is to facilitate analytical and computing work that is easier to conduct when the dimensions are small.

Imputing missing values and interpolating data series

Reflecting the importance put by central banks on having enough data at their disposal to support their decision-making processes, **an important stream of work is to augment the information available, especially in the case of missing data points or when the data are not timely enough and/or not available with sufficient frequency**.

The problem posed by missing values may arise for various reasons, eg the information had not been reported or was collected with significant quality problems and had to be disregarded. This can create serious challenges, for instance in terms of the reduction of the sample of the data available or the introduction of potential biases, in turn possibly undermining the validity of the information and hence the relevance of the actions taken on its basis. While many statistical methods have been traditionally mobilised to address these issues, **ML approaches have become more popular ways to facilitate the imputation of missing data points** (cf the review of ML-based data augmentation and related management methods by Kumar et al (2017)). Frequently used techniques include random forest–based learning methods (Stekhoven and Bühlmann (2012), Rahman and Islam (2013), Tang and Ishwaran (2017) and Ramosaj and Pauly (2019)); the automatic discovery of regular data patterns through "generative deep learning" methods (cf Yoon et al (2018), Nazábal et al (2020), Hou et al (2022) and Qian et al (2022)); approaches that make use of observed differences between specific parts in the data, eg those using "discriminative deep learning" methods (Biessmann et al (2018)); and algorithms for forecasting time series such as the Deep Autoregressive model (DeepAR) (Salinas et al (2020)). These examples represent a small number of ML-based imputation methods available, and Jäger et al (2021) provide a useful review of their performance by looking at a large number of data sets under realistic conditions in terms of missing values.

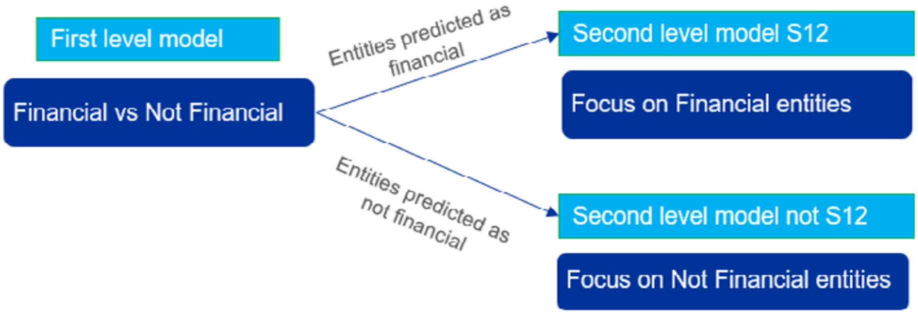
A similar problem occurs with data sets that have a low frequency (eg annual or semiannual) and are typically available with too-long time lags. As was clearly obvious when the Covid-19 pandemic struck, the usefulness of such information is limited for those central banks willing to take decisions on a timelier and/or more frequent basis. Here also, **ML techniques can be particularly useful in mitigating these problems by helping to interpolate low-frequency series into higher-frequency ones or by speeding up their release using additional information**. For instance, the Central Bank of the Russian Federation (CBRF) has developed specific tools to facilitate the quarterly compilation of financial accounts and balance sheets in the System of National Accounts despite the fact that certain non-bank financial firms report their financial statements at an annual frequency only. Traditional interpolation methods were compared with various ML-based techniques, namely random forest, that basically relates to classification algorithms (Breiman (2001)), gradient boosting trees decision models that are used in regression and classification tasks (Friedman (2002)), and neural network-based tools to generate new data consistent with the information observed (eg the Wasserstein generative adversarial network or "GAN" approach (Arjovsky et al (2017)). The first two types of techniques were fine-tuned on annual values and their lags and used to estimate quarterly figures. The third involved a learning phase based on the data patterns observed for those companies producing quarterly statements and the simulation of the corresponding data for those reporting only annual figures. Yet one issue was the fact that these various approaches can lead to considerably different outcomes; moreover, their results are difficult to interpret for users, representing an important drawback compared with more traditional statistical techniques.

Provision of contextual information

ML also supports a richer augmentation of the data sets available, by incorporating complementary public information or records derived from administrative registers. For example, the newly established statistical reporting of the firms' Legal Entity Identifier (LEI) does not comprise information on the institutional sector of these entities (ie whether they are banks, money market funds, insurance firms, households, non-financial corporations etc), which can be important for supervisory monitoring purposes. The ECB has accordingly developed an ML-based way to augment the LEI database to also include estimates of the institutional sector of the reporting entities. The approach uses a random forest classifier technique to, first, separately identify financial companies from all other firms and, second, estimate specific subsectors in these two main groups (Graph 2). This two-level classification technique is initially estimated ("trained') on a specific data set for which the institutional sector is known, and then applied to the observed database for which the information is missing.

Entity sectoral classification model used by the ECB

Graph 2


First level model
Financial vs Not Financial
Entities predicted as financial
Entities predicted as not financial
Second level model S12
Focus on Financial entities
Second level model not S12
Focus on Not Financial entities

S12 represents financial sector subclassifications of the European System of National and Regional Accounts.

Source: F Benevolo, T Gottron, I Febbo, and N Pegoraro, "Supervised machine learning for estimating the institutional sectors of legal entities on a large scale", *IFC Bulletin*, no 57, November 2022.

In practice, these approaches cannot rely on the simple running of algorithmic techniques and require significant subject matter expertise. The ECB project, for instance, benefited from extensive business area knowledge to detect the presence of specific words in the entity names at stake – eg words similar to "bank" or "manufacturing" had to be selected as relevant by statistical experts and were therefore included in the classification process. Further, an LSTM neural network (cf above) was applied to deal with the names of similar entities that can be expressed in multiple languages. Finally, the models were selected with due consideration of users' preferences; for instance, a key element was their ability to reduce the risk of wrongly classifying a firm as a non-financial company, reflecting the business need to focus on the monitoring of financial entities as a priority.

3. Macroeconomic and financial analytical tasks

With central banks' decisions becoming increasingly based on factual evidence, **a key issue for them is to make sense of the wealth of existing data to derive useful insights on the economic and financial situation** so that proper policies can be conducted. Fortunately, ML techniques can greatly support this task, by: (i) making sense of the economic and financial data available; (ii) facilitating the modelling of the economy; and (iii) supporting forecasting exercises.

Making sense of the data available

Central banks' policy decision-making hinges on thorough, continuous analyses of a large set of variables to estimate the current state and outlook for the economy. Because of their ability to deal rapidly with vast and complex sets of observations, ML techniques can facilitate these analytical tasks.

One example is the ECB project to **explore alternative sources of data to extract useful insights in almost real time**. These new sources have become increasingly relevant with the digitalisation of economic activities, as was particularly evident with the use of online platforms for shopping, trading and entertainment during the Covid-19 pandemic. This project uses credit card data for developing supplementary indicators in partnership with the Fable Data firm,[4] which has specialised in the European alternative market to provide real-time banking and credit card data. And a further source of useful information relates to the development of fintech firms, as it provides further opportunities for exploring and analysing new types of data as a complement to the more "traditional" supervisory reporting exercises organised by financial supervisors and monetary authorities.

These various initiatives have underscored the importance of continuously innovating in order to make progress and in particular: (i) to maximise the use of the data available; (ii) to explore untapped, alternative sources of information; and (iii) to enhance cooperation with the related new private sector entities that are increasingly producing vast amounts of data. Yet a key drawback for central banks is that **large numbers of data points are not sufficient to guarantee the veracity of the indicators compiled**. Indeed, big data sets may present important composition bias, hampering their accuracy (Bender et al (2021), IFC (2017)). For instance, the information collected by one or a few firms may not represent the whole underlying economic and financial reality – not everybody is paying with a credit card, or at least not in all circumstances.

The exploration of untapped alternative information sources can not only help to improve the data available in a specific area but also shed light on phenomena for which reliable data are notoriously difficult to find. A good example relates to how inflation is perceived by households, which can be affected by various psychological factors, may differ markedly from headline inflation figures (cf in Europe with the launch of the euro in the early 2000s), and is difficult to gauge –

[4] See www.fabledata.com/.

typically requiring ad hoc surveys that are complex to set up and depend on the type of reporters questioned (eg the ECB's Survey of Professional Forecasters, the household inflation survey by the French national statistical agency INSEE). To address these issues, the Bank of France has harnessed non-traditional indicators from social networks such as Twitter to estimate inflation perceptions. All the relevant tweets were analysed with a dictionary-based filter – word2vec, a neural network-based NLP algorithm that associates words out of a large corpus of text. This allowed them to be classified into topics, so as to produce a price perception indicator predicated on the difference between the number of inflation- and deflation-related tweets.

Modelling

Turning to **macroeconomic modelling exercises, central banks' experience shows that they can benefit from the availability of unconventional data sources and new ML-based methods** such as deep learning. This lesson is in line with the existing literature, especially when dealing with cases when the data/expertise is limited. For example, Chauvet and Guimarães (2021) have trained a tool on US data and proposed a transfer learning strategy to identify business cycle phases in Brazil and the euro area. One interest of this approach is to make use of the knowledge gained from one region's economic experts and apply it to other geographical areas, for instance in the absence of a well recognised business cycle dating committee.

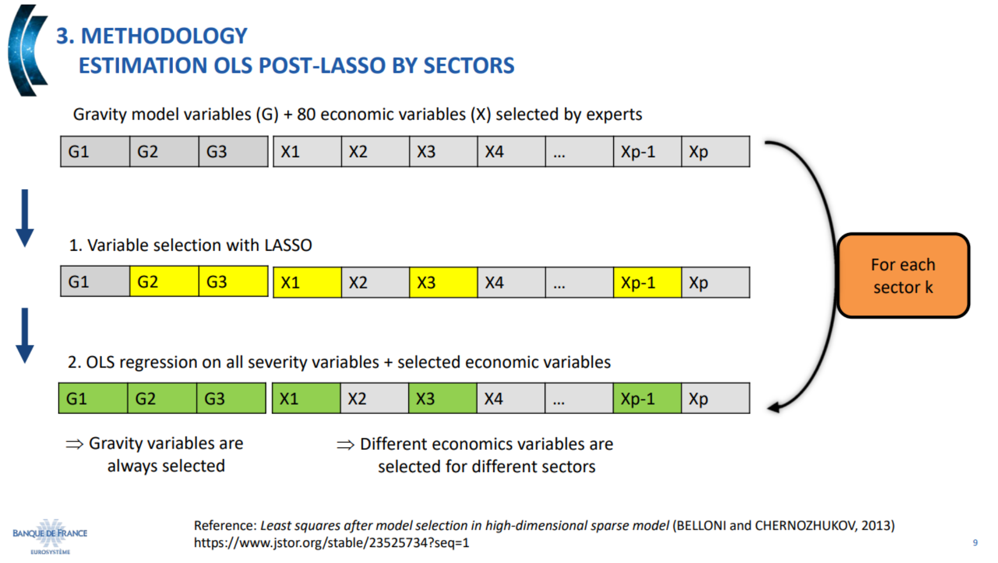
Another **important ML use case is to agnostically understand what the drivers of macroeconomic variables are** – that is, by following a pure data-driven approach instead of relying on ex ante assumptions. For instance, Kohlscheen (2021, 2022) has applied the random forest technique to analyse the drivers of inflation and in particular the role of financial factors that are typically disregarded in the toolkit of macroeconomic modellers.

More generally, **ML-based models appear particularly well suited to uncovering explanatory factors from a multitude of candidate variables**. One recent example is the Bank of France project, BIZMAP, to support the internationalisation efforts of French small and medium-sized enterprises (SMEs) (Graph 3). The aim is to make sense of the wide range of publicly available information[5] to help identify attractive EU regions in terms of exports or direct investment. The intelligence behind the tool is programmed as follows: missing data are imputed using ML tools (eg Kalman filter or missForests), relevant variables are selected to explain exports and foreign direct investment, and a gravity trade model is estimated using the least absolute shrinkage and selection operator (Lasso) methodology – a regression analysis method to select more accurate explanatory variables (Tibshirani (1996)). In a similar way, the CBRF has used ML-based methodologies to estimate financial flows in the economy to cope with the fact that a large number of unknown parameters would need to be considered if one followed a more traditional, deductive approach. The project relied on the Variational Bayes (VB) methodology, an ML-based inference technique for making the necessary

[5] Eighty-two publicly available variables from seven sources: Eurostat, the Organisation for Economic Co-operation and Development, the World Bank, the ECB, the European Investment Bank, the European Commission and the Centre for Research and Expertise on the World Economy.

approximations and that appears suitable for dealing with large data sets and complex models – both in terms of computational efficiency and estimation precision.

Use of ML to select covariates in the Bank of France's trade model to estimate exports and foreign direct investment Graph 3



3. METHODOLOGY
ESTIMATION OLS POST-LASSO BY SECTORS

Gravity model variables (G) + 80 economic variables (X) selected by experts

G1	G2	G3	X1	X2	X3	X4	...	Xp-1	Xp

1. Variable selection with LASSO

G1	G2	G3	X1	X2	X3	X4	...	Xp-1	Xp

2. OLS regression on all severity variables + selected economic variables

G1	G2	G3	X1	X2	X3	X4	...	Xp-1	Xp

For each sector k

⇒ Gravity variables are always selected

⇒ Different economics variables are selected for different sectors

Reference: *Least squares after model selection in high-dimensional sparse model* (BELLONI and CHERNOZHUKOV, 2013)
https://www.jstor.org/stable/23525734?seq=1

9

Source: C B L Kerhor, Y Houri, J-N Kien, E Kintzler and L Richardet, "Fostering European SME's internationalization using big data: the BIZMAP application", *IFC Bulletin*, no 57, November 2022.

Reflecting the above factors, **ML, and deep learning techniques in particular, are being increasingly utilised to support macroeconomic modelling exercises.** One example relates to the solving of convex optimisation problems and overcoming the "curse of dimensionality" (Bach (2017)) – that is, the problems faced when coping with an avalanche of data with increasing dimensions, including with respect to the computational efforts required for their processing and analysis. More generally, ML-based techniques are gradually and flexibly used for complex model estimations (Fernández-Villaverde et al (2020a,b), Maliar et al (2021) and Maliar and Maliar (2022)). A promising avenue relates to the area of dynamic stochastic general equilibrium (DSGE) models,[6] as argued by Fernández-Villaverde and Guerrón-Quintana (2020) and illustrated by the Bank of Canada project using deep learning methods to solve a neoclassical growth model. Lastly, neural networks are more and more popular among macroeconomic modelers, spurred by the availability of popular open source libraries, such as PyTorch and TensorFlow (cf above).

Forecasting

Given their growing contributions to economic and financial analysis and the modelling of agents' behaviour, it should not be surprising that **ML is increasingly called upon to support forecasting exercises covering the short-term – ie "nowcasting" exercises that try to predict the very recent past and the**

[6] See Tovar (2008) for a discussion of the main usage of DSGE models by central banks.

present – to the longer-term horizon – including risk scenarios. The focus has been primarily on enhancing the accuracy of "standard" central bank forecasting exercises that typically focus on real GDP and inflation as key variables influencing their policy decisions.

As regards **economic activity**, the Central Bank of Malaysia has shown the relevance of using ML techniques to extract sentiment indicators from newspaper text, which can in turn improve the forecasting accuracy of key macroeconomic indicators, ie GDP growth and its demand side components. The approach relied on building a corpus including over 720,000 business and financial news articles from 16 news portals. Interestingly, the positive results observed prior to the Covid-19 pandemic remained basically valid after this macroeconomic shock. However, the estimates also suggested that ML-based forecasts do not always outperform other models, as this can depend on the variable at stake. For instance, the computed news sentiment was deemed to improve the forecast of private investment compared with the benchmark autoregressive model, but not for the other components of economic activity.

Turning to **inflation**, the CBRF and the Higher School of Economics have analysed the contribution of various ML techniques to forecasts of consumer price inflation (CPI) using real-time versus adjusted data. To this end, a horse race was run among four popular ML algorithms: random forest, gradient boosting, the Bayesian neural network and regularised regression – ie a type of linear regression adapted to deal with a high number of variables to avoid overfitting, such as elastic net (Zou and Hastie (2005)). All of them were found to outperform an autoregressive model, providing further evidence of the usefulness of ML methods in forecasting. However, the selection of the best performing model was different depending on the forecasting horizon. For instance, gradient boosting and neural networks were found to perform better for one-month forecasts, while the elastic net had the top performance at a six-month horizon. Another important lesson was the need to assess the forecasting performance of these different models depending on the vintages of the data considered, for instance by using data available on a real-time basis or after successive statistical revisions.

Lastly, one benefit of ML techniques is to **allow the expansion of forecasting exercises to cover a wider range of potential variables of interest compared with more traditional approaches**. For instance, Bank Indonesia has been using news articles to enhance the forecasting of the situation in the labour market. The approach involved building a statistical index of employment vulnerability, computed from a corpus of around 27,000 monthly news texts covering a period of 23 years and based on NLP techniques. It facilitated the provision of forecasts on the weakening of the labour market and assessment of unemployment risks at a certain horizon and in specific sectors.

4. Monetary policy

As noted above, **ML techniques can be used to enhance the analysis and forecasts of economic output and inflation, two key variables of interest indirectly determining central banks' monetary policy reaction functions** (Taylor (1993)). In addition, these techniques also allow the integration of a much wider set of variables and contribute to a better understanding of the monetary policy decision process itself.

Assessing the influence of a wider range of factors

Monetary policy decisions can deviate from the "pure" influence of macroeconomic developments in terms of output and inflation because of **additional factors**. Yet the relationships involved are typically complex to analyse, not least because of non-linearity (eg the occurrence of an economic shock) and time-dependency issues (eg the different contributions of specific elements between expansionary and recessionary phases), hence representing an important potential use case for ML techniques.

For instance, Bank Indonesia has developed an ML-based approach to better take into consideration the **impact of foreign investors' behaviour (in terms of external capital flows into Indonesian government bonds) on exchange rate developments and, in turn, on monetary policy decisions**. The exercise involved analysing approximately 2,000 variables, derived from private data providers plus a supervisory data set of government bond transactions. Tree-based classification algorithms – the decision tree of Breiman et al (1984), random forest and XGBoost – were first used to select the most meaningful variables and prediction lags. Second, the more limited set of variables and lags obtained was kept to predict individual investors' daily investment amounts, again using a variety of ML techniques – logistic regression; support vector machine (SVM), a supervised learning algorithm used to predict discrete values (Boser et al (1992)); KNN; decision tree; random forest; XGBoost; and LSTM. Third, the LIME algorithm (cf section 2 above) was applied to explain the predictions of the best models for each investor and allow users to check the plausibility of the outcomes. The result showed that bond yields were important predictors of external investment flows, depending on the investor type (eg short-term versus long-term focus). Future work is planned to build similar ML models for analysing the stock and currency markets and to disseminate the results through a dashboard in order to facilitate feeding them into monetary policy operations.

Shedding light on the monetary policy decision process

In addition to facilitating the capture of a wider set of determinants, **ML tools appear to support a better understanding of the monetary policy decision process itself.** A study by the International Monetary Fund, also published as Edison and Carcel (2021), focused on the US monetary policy discussions and used a dedicated NLP technique – the Latent Dirichlet Allocation (LDA) of Blei et al (2003) – to analyse the topics discussed by the Federal Open Market Committee (FOMC) members over 2003–12. The meeting transcripts were divided into about 45,000 text entries, consisting of sentences or paragraphs said by the Governors. The algorithm was implemented with the goal of splitting the whole text data into eight topics: forecasting, economic modelling, statement language, risks, banking, voting decisions, economic activity and communication. This work showed the issues which were most discussed by the FOMC when taking policy decisions. For instance, it was found that discussions on economic modelling predominated during the GFC, but the main topic was on banking in the subsequent periods, and, later on, communication.

Relatedly, the joint work presented by the Swiss National Bank has been relying on a dedicated algorithm to **study the interlinkages between central bank independence and the evolution of inflation**, extending previous work by Baumann et al (2021). The related causal inference question – eg whether independence can help to reduce inflation – has been an issue of much interest but is difficult to answer using standard regression approaches. This provides an opportunity for using ML techniques, which are arguably better suited to dealing with complicated model specifications, non-linear relationships and a large number of potential explanatory variables compared with sample size. In this particular case, the application of the longitudinal targeted maximum likelihood estimation (LTMLE) of Tran et al (2019) showed that the role of central bank independence was complex and could vary depending on the historical path of inflation; it also highlighted the aptitude of ML for dealing with the causal inference problem.

5. Financial microsupervision

The field of financial supervision (principally of banking entities for those central banks tasked with their oversight, but also for other non-bank financial institutions when applicable) has seen a **marked rise in the use of ML, most notably to enable suptech**, that is, the use of new technologies and big data analytics to support supervision (Broeders and Prenio (2018), Beerman et al (2021)). These techniques can support supervisors' efficiency in: (i) covering traditional supervisory tasks (eg quality reporting, anomaly detection, sending of instructions); (ii) facilitating the assessment of micro-level fragilities; and (iii) identifying and tackling new emerging topics, such as climate-related financial risks, vulnerabilities from the Covid-19 pandemic, or the consequence of increased digitisation in finance (eg the development of fintechs).

Enhanced supervisory process

As regards the traditional micro financial supervisory tasks, the deployment of ML can strengthen the flow of information and communication between authorities and the entities they oversee.

On the one hand, the supervisory information flow starts with the reporting of individual records from monitored firms to the authorities. As for macro statistical exercises (cf Section 3), ML can make this reporting more efficient by strengthening the quality of the data in question. For instance, the Bank of Spain has developed, in collaboration with the Knowledge Engineering Institute, an ML-powered tool that imputes missing information and also detects outliers in non-financial firms' accounting statements. Out of 6.2 million statements, this tool facilitated the correction of those with insufficient data quality and those with missing data (0.5 million in both cases). Among the various methodologies tested to detect outliers – eg principal component analysis (PCA), the Mahalanobis distance and KNN – the one selected was a version of isolation forest ("missolation forest"). In parallel, missing values were imputed through regression analysis, so that information absent for one variable of a firm's report could be estimated from the values of its other variables. All in all, this project highlighted the importance of selecting the appropriate features of the model being contemplated, duly considering expert domain knowledge and factoring in the impact of computation costs during the training phase.

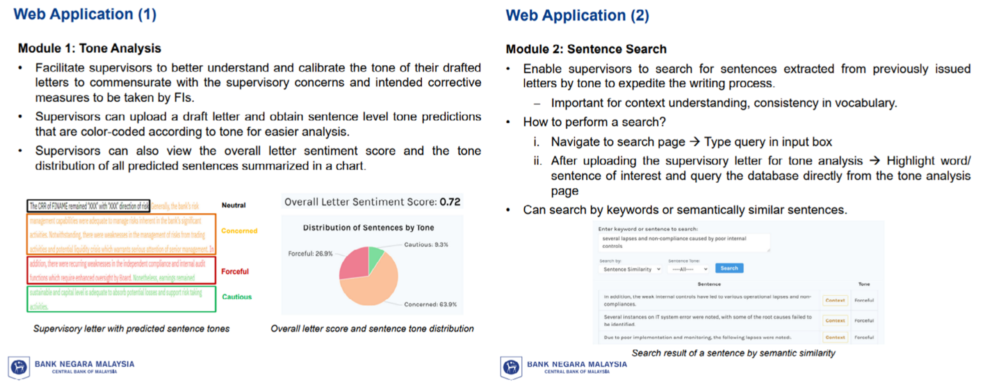
In a similar vein, the Bank of Canada has developed **a novel method based on ML to detect anomalies in data reported by financial institutions**. The objective was to enhance the efficiency and quality of the existing process that deals with millions of data points per month and that can be sensitive to support economic policy. The project relied on a two-step procedure, where financial institutions designated “similar” were clustered in a group and then analysed jointly using a supervised ML algorithm. By overcoming the traditional rule-based approach followed previously, the new method helped to detect anomalies not found before, while saving significant time. Moreover, the procedures were designed to be scalable, fully explainable and able to be run in either a cloud environment or a proprietary data lake.

On the other hand, **supervisory communication also includes the so-called drafting of supervisory letters.** This process is usually time-consuming and requires advanced analytical and communication skills. Moreover, maintaining consistency in the various messages prepared by the whole supervisory team and conveyed to a large number of firms can be challenging and burdensome. In view of these issues, the Central Bank of Malaysia has developed a suptech tool that supports communication with supervised entities, with the aim of enhancing both the efficiency of the process and the consistency of the messages conveyed. This tool had two main functionalities that complement each other: Tone Analysis and Sentence Search (Graph 4). Tone Analysis is based on a text classifier that can characterise any sentence from a supervisory letter as “neutral”, “cautious”, “concerned” or “forceful”. The training data consisted of confidential supervisory letters from 2013 to 2016, yielding 5,000 individual sentences anonymised by a specific tool – ie applying a Named Entity Recognition (NER) algorithm. The process relied on the manual intervention of experienced supervisors and the use of the deep learning model DistilBERT[7] to classify new sentences. The second feature, Sentence Search, is a tool that searched text by keyword or similar semantics. The writing of new supervisory letters with the desired tone was supported by the SentenceBERT tool, which fosters semantic similarity between a reference document and the new draft being queried (by minimising the distance between their representative vectors). Lastly, the project illustrated well the complete workflow supporting the implementation of an ML-based solution, from the development of the model, its deployment in production, and its subsequent retraining for application to a next cycle – for instance, users were able to provide feedback, allowing administrators to refine the model for subsequent estimations.

[7] DistilBERT – a general-purpose language representation model with a compression technique to reduce the number of parameters to be estimated – was the best performing model among the alternatives explored by the authors (including logistic regression, XGBoost etc).

Overview of ML-powered suptech apps used by the Central Bank of Malaysia

Graph 4



Web Application (1)

Module 1: Tone Analysis

- Facilitate supervisors to better understand and calibrate the tone of their drafted letters to commensurate with the supervisory concerns and intended corrective measures to be taken by FIs.
- Supervisors can upload a draft letter and obtain sentence level tone predictions that are color-coded according to tone for easier analysis.
- Supervisors can also view the overall letter sentiment score and the tone distribution of all predicted sentences summarized in a chart.

The CRR of FINAME remained "XXX" with "XXX" direction of risk. Generally, the bank's risk management capabilities were adequate to manage risks inherent in the bank's significant activities. Notwithstanding, there were weaknesses in the management of risks from trading activities and potential liquidity crisis which warrants serious attention of senior management. In addition, there were recurring weaknesses in the independent compliance and internal audit functions which require enhanced oversight by Board. Nonetheless, earnings remained sustainable and capital level is adequate to absorb potential losses and support risk taking activities.

Supervisory letter with predicted sentence tones

Overall Letter Sentiment Score: **0.72**

Overall letter score and sentence tone distribution

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

Web Application (2)

Module 2: Sentence Search

- Enable supervisors to search for sentences extracted from previously issued letters by tone to expedite the writing process.
  - Important for context understanding, consistency in vocabulary.
- How to perform a search?
  i. Navigate to search page → Type query in input box
  ii. After uploading the supervisory letter for tone analysis → Highlight word/ sentence of interest and query the database directly from the tone analysis page
- Can search by keywords or semantically similar sentences.

Enter keyword or sentence to search: several lapses and non-compliance caused by poor internal controls

Search by: Sentence Similarity

Sentence Tone: ----All----

Sentence		Tone
In addition, the weak internal controls have led to various operational lapses and non-compliances.	Context	Forceful
Several instances on IT system error were noted, with some of the root causes failed to be identified.	Context	Forceful
Due to poor implementation and monitoring, the following lapses were noted:.	Context	Forceful

Search result of a sentence by semantic similarity

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

Source: J Tan, C K Shum and M A Amri, "Supervisory letter writing app: expediting letter drafting and ensuring tone consistency", *IFC Bulletin*, no 57, November 2022.

Assessing micro-level fragilities

An important building block of the supervisory process is the assessment of micro-level fragilities to **identify the risks faced by a given financial institution and the potential actions warranted to mitigate these**. Given the large amount and complexity of granular data to be digested in this endeavour, ML has proved particularly well suited to addressing these tasks.

For instance, **ML techniques can help to enhance the quality of the firm-level data used in supervisory exercises**, as highlighted in a study by Hitotsubashi University Business School and the Bank of Japan. It compared predictions of the risk of a firm exiting the market in case of insolvency or of a voluntary exit made by humans (professional analysts of a Japanese credit bureau) with those made by ML algorithms (random forest). The results showed that the algorithms outperformed experts' predictions in general, although humans could perform better when assessing firms with less data available – possibly reflecting their greater ability to consider "soft information" compared with automatised algorithms. Hence, one key lesson was that ML techniques cannot fully replace expert judgment but should be used as a useful complement, depending on firm-level characteristics (eg the degree of information available) as well as on users' preferences for minimising the risk of errors of type I versus type II (cf above). Perhaps more importantly, the study also underscored the importance of systematically analysing the accuracy of the new tools in comparison with traditional methods, and in particular of analysing the causes of the respective errors observed.

Moreover, **ML techniques can be applied to enhance the models used for financial stability purposes by incorporating additional sources of information**. For instance, the CBRF has developed ML algorithms (eg logistic regression combined with random forest) that consider additional information on daily payments to improve the traditional default probability models that sit at the core of financial supervisory exercises and are typically based on firm-level accounting data. In that case too, it was found that the degree of accuracy of the respective techniques needed to be carefully analysed, with due consideration in particular of the differences observed across economic sectors.

Dealing with non-supervised entities

While micro supervisory tasks focus on the situation of the specific firms that have to be monitored by the authorities, **it is also important to consider other, less regulated, sectors, not least to prevent regulatory arbitrage** – that is, when non-regulated firms compete in the provision of services that are similar to the ones offered by regulated entities (cf discussion in Fleischer (2010)). On this front too, the use of ML techniques can provide useful insights to supervisors. It can also help supervisors to apply "proportionality" when considering new entrants in the financial system (BCBS and World Bank (2021)).

One telling example relates to the identification of entities involved in fintech, defined as technological innovation used to support or provide financial services (IFC (2020b)). The Bank of France and the Deutsche Bundesbank have created two complementary ML-based tools to identify and monitor these entities. The goal was to overcome the lack of sufficient information available about them due to their fast-paced development and churn. The projects, which are still in their early stages, required mostly public data and were designed to be replicable more broadly in other jurisdictions.

The tool developed by the **Bank of France** focused on classifying whether or not firms are potential fintechs, using publicly available data (covering 84 features) and the isolation forest outlier detection algorithm (Liu et al (2008)). Training and validation were conducted for 10,000 individual non-fintech firms, helping subject matter experts to identify around 350 firms as potential fintechs. The features found most relevant for supporting this identification exercise included newspaper articles about the firms, economic sectors, employees' job titles and the names of senior managers.

Turning to the tool of the **Deutsche Bundesbank**, it only needs an initial list of web addresses (belonging to already identified fintech and non-fintech firms)[8] as input for training and validation: the tool scrapes these websites and creates a graph database consisting of companies, named entities (persons, organisations and locations) and keywords as nodes. In setting up the graph, a large amount of information had to be processed – in this case 515,000 webpages with 1.1 million named entities. According to the location in the graph, a neural network algorithm will decide whether a new and hitherto unclassified company is a fintech or not. This approach appears particularly well suited for dealing with non-structured information.

6. Macro-financial stability policies

Independently of whether the central bank is in charge or not of micro-financial supervision, one of its key policy mandates relates to the macro dimension of financial stability (Crockett (2000)). The impact of the GFC has reinforced interest in developing

[8] The proof of concept uses a data set of 1,190 company web addresses, of which 390 are identified fintechs.

a **system-wide approach to monitoring financial risks, with a dual focus on the situation of different institutions together at a point in time and on the evolution of risks over time as the financial cycle evolves**. This duality calls for collecting and analysing huge amounts of data, covering a wide range of firms and over long periods. Hence, it should not be a surprise that the financial stability function of central banks can benefit from the increased use of ML – cf for instance Fouliard et al (2021), who document how it can enhance the ability to predict crises well before they take place. Two important elements which deserve to be highlighted from this perspective are: (i) the support of ML to match information from various, often unrelated corners that can help to identify system-wide vulnerabilities and their evolution over time; and (ii) the ability to support other policy tasks that are also relevant from a financial stability perspective.

Support of macroprudential exercises

Supporting the macro-financial function requires **collecting trustworthy statistics from various sectors of the economy, hence putting a premium on strong quality assurance processes** for dealing with databases that are not directly produced by the central bank alone. One example is the Bank of Portugal experience with the use of information from the Portuguese credit registry. This source is characterised by an extremely high level of granularity, resulting in a large number of complex observations (over 200 attributes) and calling for strong data quality controls to detect anomalies and identify subtle evolutions. To address the full range of potential anomalies, two automatic filters were created. The first was the Reporting Consistency test, to evaluate if all the financial instruments were reported in a consistent way until their maturity. A second test was the Concentration Check, to check the consistency of the reporting of categorical variables at the agent level. The detection of anomalies was then based on an isolation forest algorithm and was found to have facilitated the detection of reporting gaps, strange data patterns and structural breaks, in turn enhancing the quality of the information available to support macroprudential analyses.

Moreover, **the sheer scale of the detailed data sets of potential interest to policymakers is also a key factor supporting ML-based initiatives to develop a more structural framework**. The reason is that, even abstracting from data quality problems, analytical and computational limitations can prevent the use of these data for effective financial stability monitoring. To address this point, the ECB, Deloitte and Google have jointly developed a solution in the form of a dynamic multilayer network that helps supervisors to look at the available statistics in a comprehensive way and analyse them through various operations supported by data science tools – such as aggregation, filtering and bottom-up analyses from the individual data level. This solution is reported to have facilitated financial stability monitoring tasks in the face of systemic risk events, such as during the Covid-19-induced turmoil in financial markets in March 2020 (FSB (2020)).

The above approaches can be instrumental in **facilitating the analysis of interconnections observed at a given point in time across the various segments of the financial system** and that are a key source of attention for macroprudential authorities. One example relates to the relationships between the banking and housing sectors, as analysed by the Australian service provider Quant Property Solutions. In particular, an important feature mechanism is that lenders willing to foreclose on mortgaged real estate can trigger important developments in housing prices, with possibly severe financial stability implications because of imperfect available information. For instance, foreclosed houses are typically sold below market values, presumably reflecting a specific bias among those market participants willing to sell their collateral, not least because of banks' balance sheet considerations. ML techniques were used to support the market price discovery mechanism, by helping to disentangle the contributions of the multiple factors at play (eg the situation of the bank selling a property, geo-specific real estate features, the economic outlook).

Turning to the time dimension of systemic risk, ML methods can support dealing with large and complicated data sets that change over time. For instance, the CBRF has adopted this type of approach to facilitate work on (changing over time) micro-level databases on banking loans, with several benefits observed in terms of data quality assurance, scalability and automation of the operations, and higher interpretability of the results. Moreover, the solution also allowed for matching the database in question with other sources, namely the Federal Tax and the State Statistics services.

Additional financial stability dimensions

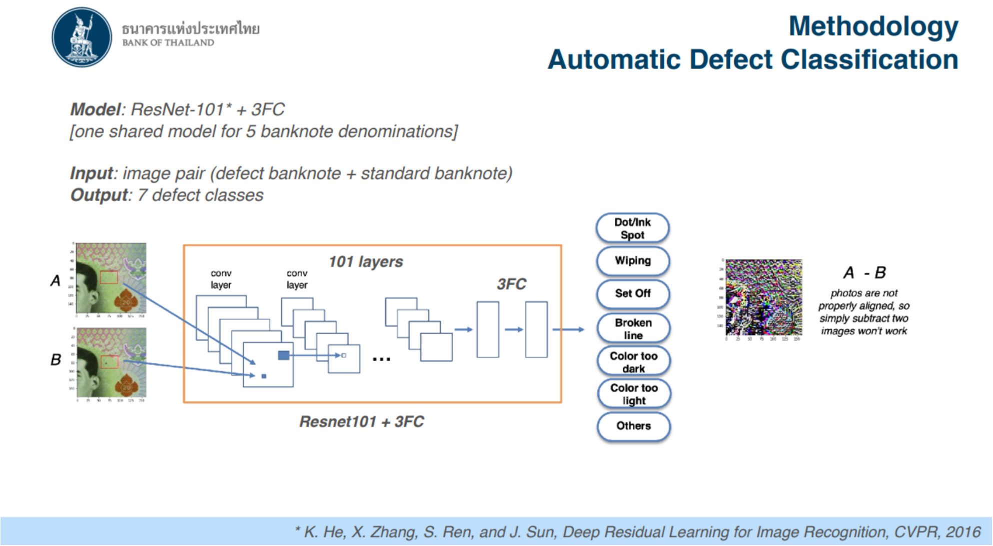
There are additional tasks performed by central banks that are dedicated to the service of society and that, by protecting prosperity and providing financial security and confidence, also play a role in supporting financial stability more generally. **Cases in point relate to financial inclusion, consumer protection and anti-money laundering**, three areas that are reported to have benefited significantly from the development of big data analytics in recent years.

Another important domain is **the safeguarding of the payment system**, whose monitoring sits at the core of a central bank's mandate to both ensure a smooth functioning of payments and prevent its misuse. ML techniques can be instrumental in coping with the large amount of individual transactions involved, as shown by the joint experience reported by the Central Bank of Ecuador in developing neural networks for outlier detection – in that case autoencoders (cf Section 2), with the goal of identifying abnormal transactions that might require closer scrutiny by the payment system's oversight team. As argued by Rubio et al (2020), this application has been able to identify a wide range of payment transaction anomalies. Their findings confirm the experience of previous similar projects, for instance at the Netherlands Bank (Triepels et al (2017)).

A final area where ML-based anomaly detection and classification techniques can support central banks' operations relates to their core mandate of printing money. Occasionally, some banknotes are produced with defects, which can happen at different steps of the production process. While the problematic notes are typically detected by cash machines, analysing the defects to identify their causes can be a laborious, time-consuming and repetitive task. To cope with these challenges, one can usefully deploy ML techniques in line with the example of the Bank of Thailand. This institution has implemented a convolutional neural network-based tool (ResNet-101), which is a type of artificial neural network commonly applied to analyse images (Graph 5). The experience so far is that the number of misprinted notes has fallen by more than half in Thailand.

Bank of Thailand's computer vision model for detecting banknote defects

Graph 5



ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Methodology
Automatic Defect Classification

Model: *ResNet-101* + 3FC*
[one shared model for 5 banknote denominations]

Input: *image pair (defect banknote + standard banknote)*
Output: *7 defect classes*

photos are not properly aligned, so simply subtract two images won't work

** K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, CVPR, 2016*

3FC refers to the three fully connected neural layers.

Source: J Kerdsri and P Treeratpituk, "Using deep learning technique to automate banknote defect classification", *IFC Bulletin*, no 57, November 2022.

References

Abadi, M, A Agarwal, P Barham, E Brevdo, Z Chen, C Citro, G S Corrado, A Davis, J Dean, M Devin, S Ghemafwat, I Goodfellow, A Harp, G Irving, M Isard, R Jozefowicz, Y Jia, L Kaiser, M Kudlur, J Levenberg, D Mané, M Schuster, R Monga, S Moore, D Murray, C Olah, J Shlens, B Steiner, I Sutskever, K Talwar, P Tucker, V Vanhoucke, V Vasudevan, F Viégas, O Vinyals, P Warden, M Wattenberg, M Wicke, Y Yu and X Zheng (2015): "TensorFlow: Large-scale machine learning on heterogeneous systems", static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf.

Ahrens, M and M McMahon (2021): "Extracting economic signals from central bank speeches", *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pp 93–114.

Apel, M, M Grimaldi and I Hull (2021): "How much information do Monetary Policy Committees disclose? Evidence from the FOMC's minutes and transcripts", *Journal of Money, Credit and Banking,* doi.org/10.1111/jmcb.12885.

Arjovsky, M, S Chintala and L Bottou (2017): "Wasserstein generative adversarial network", *Proceedings of the 34th International Conference on Machine* Learning, PMLR no 70, pp 214–23.

Bach, F (2017): "Breaking the curse of dimensionality with convex neural networks", *Journal of Machine Learning Research*, no 18, pp 1–53.

Basel Committee on Banking Supervision (BCBS) and World Bank (2021): *Proportionality in bank regulation and supervision – a joint global survey*.

Baumann, P, E Rossi and A Volkmann (2021): "What drives inflation and how? Evidence from additive models selected by cAIC", *Swiss National Bank Working Papers*, vol 12.

Beerman, K, J Prenio and R Zamil (2021): "Suptech tools for prudential supervision and their use during the pandemic", *FSI Insights on policy implementation*, no 37.

Bender, E, T Gebru, A McMillan-Major and M Mitchell (2021): "On the dangers of stochastic parrots: can language models be too big?", *Proceedings of the 2021 Association for Computing Machinery Conference on Fairness, Accountability, and Transparency*.

Biessmann, F, D Salinas, S Schelter, P Schmidt, and D Lange (2018): "Deep learning for missing value imputation in tables with non-numerical data", *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*.

Blei, D, A Ng and M Jordan (2003): "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, no 3, pp 993–1022.

Boser, B, I Guyon and V Vapnik (1992): "A training algorithm for optimal margin classifiers", *Proceedings of the Fifth Annual Workshop on Computational Learning Theory,* pp 144–52.

Breiman, L (2001): "Random forests", *Machine Learning*, no 45, pp 5–32.

Breiman, L, J Friedman, R Olshen and C Stone (1984): "Classification and regression trees", *Wadsworth Advanced Books and Software*.

Broeders, D and J Prenio (2018): "Innovative technology in financial supervision (suptech): the experience of early users", *FSI Insights on policy implementation*, no 9.

Bruno, G and J Marcucci (2021): "Data science and machine learning for a data-driven central bank", in P Nymand-Andersen (ed), *Data science in economics and finance for decision makers*, Chapter 10.

Casanova, C, B Hardy and M Onen (2021): "Covid-19 policy measures to support bank lending", *BIS Quarterly Review*, September, pp 45–59.

Chakraborty, C and A Joseph (2017): "Machine learning at central banks", *Bank of England Working Paper*, no 674, www.bankofengland.co.uk/working-paper/2017/machine-learning-at-central-banks.

Chang, A X and C Manning (2012): "SUTime: A library for recognizing and normalizing time expressions", *8th International Conference on Language Resources and Evaluation*.

Chauvet, M and R S Guimarães (2021): "Transfer learning for business cycle identification", *Banco Central do Brasil Working Paper*, no 545.

Chen, T and C Guestrin (2016): "XGBoost: a scalable tree boosting system", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp 785–94.

Cieslak, A and A Schrimpf (2019): "Non-monetary news in central bank communication", *Journal of International Economics*, no 118, pp 293–315.

Crockett, A (2000): "Marrying the micro- and macro-prudential dimensions of financial stability", remarks before the Eleventh International Conference of Banking Supervisors, Basel, 20–21 September.

Doerr, S, L Gambacorta and J M Serena (2021): "Big data and machine learning in central banking", *BIS Working Papers*, no 930.

Edison, H and H Carcel (2021): "Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts", *Applied Economics Letters,* vol 28, no 1, pp 38–42.

Fernández-Villaverde, J and P A Guerrón-Quintana (2020): "Estimating DSGE models: recent advances and future challenges", *NBER Working Paper*, no 27715, August.

Fernández-Villaverde, J, S Hurtado and G Nuno (2020a): "Financial frictions and the wealth distribution", *Banco de España Working Paper,* no 2013.

Fernández-Villaverde, J, G Nuno , G Sorg-Langhans and M Vogler (2020b): "Solving high-dimensional dynamic programming problems using deep learning", mimeo.

Ferreira, L N (2021): "Forecasting with VAR-teXt and DFM-teXt models: exploring the predictive power of central bank communication", *Banco Central do Brasil Working Paper*, no 559.

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.

——— (2020): *Holistic Review of the March Market Turmoil and COVID-19 pandemic: Financial stability impact and policy responses*, November.

Fleischer, V (2010): "Regulatory arbitrage", *Texas Law Review,* vol 89, no 7.

Fouliard, J, M Howell and H Rey (2021): "Answering the Queen: machine learning and financial crises", *National Bureau of Economic Research Working Paper,* no 28302.

Frey, B J and D Dueck (2007): "Clustering by passing messages between data points", *Science*, vol 315, no 5814, pp 972–76.

Friedman, J H (2002): "Stochastic gradient boosting", *Computational Statistics & Data Analysis*, vol 38, no 4, pp 367–78.

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

Gros, D (2018): "When communication becomes the policy", *Monetary dialogue*, European Parliament, September.

Hansen, S and M McMahon (2018): "How central bank communication generates market news", *Handbook of Macroeconomics*, Chapter 15.

Hansen, N, M McMahon and A Prat (2018): "Transparency and deliberation within the FOMC: A computational linguistics approach", *Quarterly Journal of Economics*, vol 133, no 2, pp 801–70.

Hernández de Cos, P (2022): "Old risks, news challenges, same objective: the work programme of the Basel Committee in 2022", speech, www.bis.org/speeches/sp220225.htm.

Hou, J, H Jiang, C Wan, L Yi, S Gao, Y Ding and S Xue (2022): "Deep learning and data augmentation based data imputation for structural health monitoring system in multi-sensor damaged state", *Measurement*, vol 196, 111206, doi.org/10.1016/j.measurement.2022.111206.

Irving Fisher Committee (IFC) (2015): "Central banks' use of and interest in 'big data", *IFC Report*, no 3, October.

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

——— (2018): "IFC report on central banks and trade repositories derivatives data", *IFC Report*, no 7, October.

——— (2019): "The use of big data analytics and artificial intelligence in central banking", *IFC Bulletin*, no 50.

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

——— (2020b): "Towards monitoring financial innovation in central bank statistics", *IFC Report*, no 12.

Israël, J-M and B Tissot (2021): "Incorporating micro data into macro policy decision-making", *IFC Bulletin*, no 53.

Jäger, S, Allhorn A and F Biessmann (2021): "A benchmark for data imputation methods", *Front. Big Data*, 4:693674.

Kohlscheen, E (2021): "What does machine learning say about the drivers of inflation?", *BIS Working Papers*, no 980.

Kohlscheen, E (2022): "Quantifying the role of interest rates, the dollar and Covid in oil prices", *BIS Working Papers,* no 1040.

Kumar, A, M Boehm and J Yang (2017): "Data management in machine learning: Challenges, techniques, and systems", *Proceedings of the 2017 ACM International Conference on Management of Data*, pp 1717–22.

Liu, F, K Ting and Z Zhou (2008): "Isolation forest", *Eighth IEEE International Conference on Data Mining*.

Maliar, L, S Maliar and P Winant (2021): "Deep learning for solving dynamic economic models", *Journal of Monetary Economics*, vol 122, pp 76–101.

Maliar, L and S Maliar (2022): "Deep learning classification: modelling discrete labor choice", *Journal of Economic Dynamics and Control*, vol 135.

Nazábal, A, P Olmos, Z Ghahramani and I Valera (2020): "Handling incomplete heterogeneous data using VAEs", *Pattern Recognition*, vol 107:107501.

Paszke, A, S Gross, F Massa, A Lerer, J Bradbury, G Chanan, T Killeen, Z Lin, N Gimelshein, L Antiga, A Desmaison, A Kopf, E Yang, Z DeVito, M Raison, A Tejani, S Chilamkurthy, B Steiner, L Fang, J Bai and S Chintala (2019): "PyTorch: an imperative style, high-performance deep learning library", *NeurIPS proceedings*, pp 8024–35, papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot and E Duchesnay (2011): "Scikit-learn: machine learning in Python", *Journal of Machine Learning Research*, no 12, pp 2825–30.

Qian, Y, L Tian, B Zhai, S Zhang and R Wu (2022): "Informer-WGAN: high missing rate time series imputation based on adversarial training and a self-attention mechanism", *Algorithms*, vol 15, no 252, www.doi.org/10.3390/a15070252.

Rahman, M G and M Islam (2013): "Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques", *Knowledge-Based Systems*, vol 53, pp 51–65.

Ramosaj, B and M Pauly (2019): "Predicting missing values: A comparative study on non-parametric approaches for imputation", *Comput Stat*, vol 34, no 4, pp 1741–64.

Ribeiro, M, S Singh and C Guestrin (2016): "'Why should I trust you?' Explaining the iredictions of any classifier", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 1135–44.

Richardson, A, T van Florenstein Mulder and T Vehbi (2019): "Nowcasting New Zealand GDP using machine learning algorithms", *IFC Bulletin*, no 50, May.

Rubio, J, P Barucca, G Gage, J Arroyo and R Morales-Resendiz (2020): "Classifying payment patterns with artificial neural networks: An autoencoder approach", *Latin American Journal of Central Banking*, vol 1, no1.

Salinas, D, V Flunkert, J Gasthaus and T Januschowski (2020): "DeepAR: Probabilistic forecasting with autoregressive recurrent networks", *International Journal of Forecasting*, 36, pp 1181–91.

Stekhoven, D J and P Bühlmann (2012): "MissForest – non-parametric missing value imputation for mixed-type data", *Bioinformatics*, no 28, pp 112–8. doi:10.1093/bioinformatics/btr597.

Tang, F and H Ishwaran (2017): "Random Forest missing data algorithms", *Stat Analysis Data Mining*, vol 10, no 6, pp 363–77.

Taylor, J (1993): "Discretion versus policy ules in practice", *Carnegie-Rochester Conference Series on Public Policy*, vol 39.

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.

Tovar, C (2008): "DSGE models and central banks", *BIS Working Papers*, no 258, September.

Tran, L, C Yiannoutsos, K Wools-Kaloustian, A Siika, M Van Der Laan, M Petersen (2019): "Double robust efficient estimators of longitudinal treatment effects: Comparative performance in simulations and a case study", *The International Journal of Biostatistics*, vol 15, no 2.

Triepels, R, H Daniels and R Heijmans (2017): "Anomaly detection in real-time gross settlement systems", *ICEIS*, vol 1 .

Wibisono, O, H Ari, A Widjanarti, A Zulen and B Tissot (2019): "Using big data analytics and artificial intelligence: a central banking perspective", in "Data Analytics", *Capco Institute Journal of Financial Transformation*, 50th edition, pp 70–83.

Yoon, J, J Jordon and M van der Schaar (2018): "GAIN: missing data imputation using generative adversarial nets," *Proceedings of the 35th International Conference on Machine Learning*, pp 5675–84.

Zou, H and T Hastie (2005): "Regularisation and variable selection via the elastic net", *Journal of the Royal Statistical Society*, Series B (Methodological), vol 67, no 2, pp 301–20.

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Opening remarks[1]

Piero Cipollone, Deputy Governor, Bank of Italy

[1] These opening remarks were 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 central banks and other institutions represented at the event.

BANCA D'ITALIA
EUROSISTEMA

Data Science in Central Banking

Welcome address by Piero Cipollone
Deputy Governor of the Bank of Italy

Rome, 19 October 2021

Ladies and Gentlemen,

I am delighted to open this virtual conference on **Data Science in Central Banking**, jointly organized by Banca d'Italia and the Irving Fischer Committee of the Bank for International Settlements. I would like to welcome all the participants joining us today from some sixty countries.

1. Introduction

For the last two years, we have lived through a dramatic period as the COVID-19 pandemic has swept the globe. Never before as in these dark days has Data Science in the form of big data and machine learning (ML) algorithms proved so helpful in the war against Coronavirus. The lack of biological knowledge on this virus has spurred the data science community to step up and contribute to the fight against COVID-19. Scientists from many different disciplines and public organizations have acknowledged the importance of data analytics by open sourcing the virus genome and other datasets in the hope of a swift data-driven solution.

In this four-day conference we will endeavour to share among institutions and academia the newest and most interesting applications of Data Science and machine learning to sharpen our analytical capacity to cope with new and rapidly evolving economic equilibria.

2. The role of Data Science in central banking activities

Data Science is an interdisciplinary field that combines computer science, statistics and business domain knowledge aimed at generating insights from noisy and often unstructured data. It integrates mathematics with scientific methods and computing platforms. Still a young field, it has quickly developed over the last few years. Its main driver is the astounding volume of data stored by private companies and public authorities, which can now be treated more easily with new algorithms to extract the information hidden among them.

Nonetheless, at the Bank of Italy, Data Science is not completely new. We do have a sound history of basing our decisions on data. In 2016, we established a multidisciplinary team to address the potential benefits and hidden risks of embracing the technological challenges of artificial intelligence (AI) and machine learning (ML) fuelled by the advances in big data, which continue to evolve at an incredible speed.

A gargantuan amount of digital activity is occurring every day. In fact, data are constantly generated by our internet activities. This explosion stems from the aggregate actions of about 4.7 billion active internet users worldwide.[1] These numbers are projected to rise even further in the coming years. According to SeedScientific,[2] at the dawn of 2020 the total amount of data in the world was around 44 zettabytes, tantamount to $44 \cdot 10^{21}$ bytes, which is the number of cells we could count in more than 1,400 human beings.

This astonishing amount of data can give us a better understanding of the state of the economy at both the micro and the macro level, provided that we able to extract the signal from the noise. Banca d'Italia has constantly striven to be at the cutting edge in developing software and hardware platforms, enabling big data analytics[3] for statistical and economic applications.

The rise of Data Science started at the beginning of 2010 when high-quality models for image recognition were created, computational power achieved sufficient growth, and people in many scientific areas realized the full potential of such an approach.

Data Science glues together machine learning and data processing. The former is a collection of tools, which allows us to learn from the given data and to extract patterns and interactions between series and values. The latter describes the possible set of actions in relation to the data itself: collection, manipulation, preparation, and visualization.

It is important to flag a few differences between Data Science and the classical Econometrics we study at University. Unlike Econometrics, which concentrates on solving non-linearity bias problems in a typical linear framework, Data Science is about improving our ability to work with non-linear relationships in the system. Another difference is that Econometrics concentrates on methods such as robustness, while machine learning algorithms became popular for their outstanding predictive performance.[4]

3. Data Science at the Bank of Italy

Banca d'Italia is organizing, along with the Federal Reserve Board, the Sveriges Riksbank, the University of Pennsylvania and the Imperial College, a series of webinars

[1] As of January 2021.

[2] See https://seedscientific.com/how-much-data-is-created-every-day/

[3] See, for example, 'Big data processing: Is there a framework suitable for economists and statisticians?', 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 2804-2811, doi: 10.1109/BigData.2017.8258247. 'Weaving Enterprise Knowledge Graphs: The Case of Company Ownership Graphs.'

[4] I am thinking of Extreme Gradient Boosting (XGBoost), which owes its wide popularity to its dominance over several machine learning algorithms rather than to its mathematical properties.

on ‘Applied Machine Learning, Economics, and Data Science’ (AMLEDS).[5] The aim of these webinars is to foster the integration of data science tools into economics and policy-related issues and to promote closer cooperation on issues related to data science, big data and machine learning techniques applied to policy questions. Such cooperation has intensified during the pandemic, leading to many important initiatives such as a series of conferences on non-traditional data organized by the Federal Reserve Board and the Bank of Italy (for example, this year’s conference is going to be hosted by the Bank of Canada in November). This new world has, of course, set many challenges for central banks and public institutions: on one level, central banks have had to devise specific organizational structures to have a consistent and more efficient approach to the big data and machine learning tools used by each institution. At the same time, they have had to increase their data production, leveraging on non-traditional data to get more timely and high-frequency indicators of economic activity, which are important during unprecedented shocks like the COVID-19 pandemic.

4. The challenges and risks of Data Science

We need to exert extreme caution when employing these new tools.

Let me briefly go over some of them.

First, big/web data might lose their statistical relevance when they are employed in an unsound way. Such data typically entail selection bias because of the features of the population; increasing the sample size will not shrink the sampling error if the estimation algorithm does not correct for this kind of distortion.

Second, the availability of a huge amount of data raises the importance of its integrity, confidentiality and privacy. Personal and company data protection is central to our societies.

The sheer amount of personal data now available, and the growing ease with which individual information can be merged across databases, have far-reaching implications for privacy, competition and freedom. Indeed, this has prompted the development of regulations concerning the treatment of digital data (think of the GDPR or the CCPA in California).

Rules on data management often differ across jurisdictions and data domains. Therefore, international cooperation is to be encouraged as far as possible.

Technologies enabling the processing of personal data are already available. In 2019, Google made available its differential privacy library, which allows sensitive data to be processed privately.

To reach these welfare-improving goals, further investment from both the public and the private sector are required. Close cooperation between the private sector,

[5] AMLEDS are a series of webinars open to all those in the world who are interested in applied machine learning, big data, and natural language processing for economics and in how these techniques and data science can be applied to social science.

which typically owns most of the new data, and the public sector, which uses (or would like to use) such data for policy reasons and for the common good, will also be essential.

This is why central banks have always made great efforts when it comes to collecting and analysing data. Throughout its history, Banca d'Italia has drawn extensively on data published by the National Statistical Institute and other national and international agencies. It has also been an active producer of statistics, not only on banking, financial and fiscal variables, but also on firms and households.

Banca d'Italia has been collecting micro-level statistical information on companies since the early 1950s. These data are now enriched with non-traditional sources such as social media, blogs, newspapers, and private company datasets.[6]

5. Conclusions

Let me conclude my talk by thanking, once again, all the speakers and participants for joining us today, even in this virtual fashion. We hope to welcome you here in Rome in person in the near future. Special thanks go to those who have helped to organize this workshop, which brings together leading economists, statisticians, artificial intelligence and machine-learning specialists, data scientists from about fifty-five central banks and fifteen participants from Universities and government agencies. This guarantees a broad variety of perspectives and a lively discussion.

I am sure that you are going to have a very interesting and productive workshop.

[6] Such as the real estate website, immobiliare.it, and the mortgage website, mutuionline.it

Designed by the Printing and Publishing Division of the Bank of Italy

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome, virtual event

Keynote speech

Monetary economics and communication: new data, new tools, new and old questions[1]

Michael McMahon, Professor of Economics,
University of Oxford

[1] This presentation was prepared for the conference. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the event.

Monetary Economics and Communication: New Data, New Tools, New and Old Questions

Michael McMahon

University of Oxford, CEPR & Irish Fiscal Advisory Council

Oct 19, 2021

This talk represents my views and not necessarily those others including the Irish Fiscal Advisory Council, Central Bank of Ireland (co-authors), or anyone else including co-authors!

What is Central Bank Communication?

Central Bank Communication

Central bank communication broadly defined as the information that the central bank makes available about its current and future policy objectives, the current economic outlook, and the likely path for future monetary policy decisions (Blinder et al, 2008).

- **Blinder (1998):**

  "expectations about future central bank behavior provide the essential link between short rates and long rates."

- **Bernanke (2003):**

  "A given [monetary] policy action... can have very different effects on the economy, depending (for example) on what the private sector infers... about the information that may have induced the policymaker to act, about the policymaker's objectives in taking the action..."

The Nature of Central Bank Transparency has Changed

Pre-1994 in US: No policy announcements, maximum opacity

- **Montagu Norman (1920-44):** *'Never apologise, never explain.'*
- **Alan Greenspan (1987):** *'If I seem unduly clear to you, you must have misunderstood what I said.'*

Fed since 1994: Announcements of policy & increasingly regular speeches

- **Blinder (1996):** *'Greater openness might actually improve the efficiency of monetary policy... [because] expectations about future central bank behavior provide the essential link between short rates & long rates.'*

More recently (Effective Lower Bound): Forward guidance

- ⇒ Move to Open **Mouth** Operations

Why Does This New Era of Communication Matter?

What we need:

1. Answers to new questions, and new answers to old questions
2. Creation and utilisation of new data resources
3. Application and development of new methodologies

⇒ New avenues for future research

My work: 3 Related Themes

Theme I Understanding the Transmission of Monetary Policy

Theme II Monetary Policy and Expectations Management

Theme III Monetary Policy and Uncertainty

Project I: Why does CB communication move markets?

New Research with David Byrne, Robert Goodhead & Conor Parle

The Central Bank Crystal Ball: Understanding temporal information in monetary policy communication

Project I: Why does CB communication move markets?

New Research with David Byrne, Robert Goodhead & Conor Parle

The Central Bank Crystal Ball: Understanding temporal information in monetary policy communication

Key Concept

Information Deficit: the event must provide new, relevant information to market in order to cause them to update their beliefs.

Project I: Why does CB communication move markets?

New Research with David Byrne, Robert Goodhead & Conor Parle

The Central Bank Crystal Ball: Understanding temporal information in monetary policy communication

Key Concept

Information Deficit: the event must provide new, relevant information to market in order to cause them to update their beliefs.

Key Question

What is the nature of the central bank information deficit?

Project I: Why does CB communication move markets?

New Research with David Byrne, Robert Goodhead & Conor Parle

The Central Bank Crystal Ball: Understanding temporal information in monetary policy communication

Key Concept

Information Deficit: the event must provide new, relevant information to market in order to cause them to update their beliefs.

Key Question

What is the nature of the central bank information deficit?

Our Approach

Measure a new dimension of communication.

Contributions

Contribution 1: Methodological Breakthrough

Measuring the *temporal* dimension of text - 3rd T.

Contribution 2: Importance of Temporal Dimension in Information Deficit

Conjunctural context is key for policy making. Even if forward-looking aspect of policy is key for expectations, communication about context and processing of data seems to be as important as forward-looking communication.

Contribution 3: Questions identify the deficit

Speeches that fill the deficit are more likely to give rise to more market news.

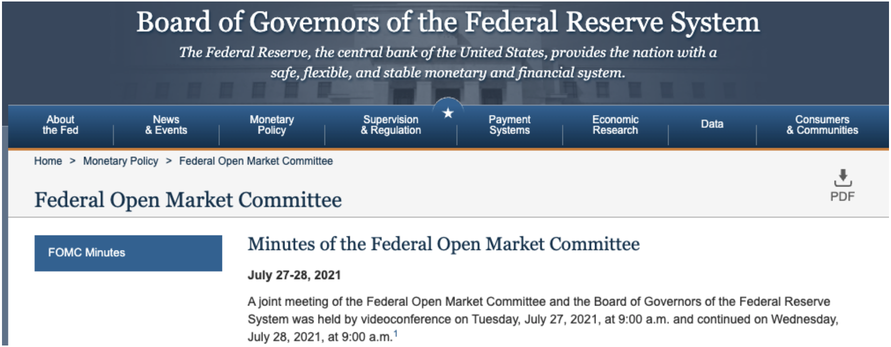
Monetary Policy Decision-making Process

1. $\Omega_m^{CB} = g_m\left(X_m^{CB}\right)$ - *Assessment Function*
   Map data into a vector of beliefs about the state of the economy.
   - Macro models typically don't focus on $g_m(.)$.
   - Captures two important analytical steps:
     - *Evaluation*
     - *Projection*
   - Both could be the source of $g_m(.) \neq g_{m-1}(.)$

Monetary Policy Decision-making Process

1. $\Omega_m^{CB} = g_m\left(X_m^{CB}\right)$ - *Assessment Function*
   Map data into a vector of beliefs about the state of the economy.
   - Macro models typically don't focus on $g_m(.)$.
   - Captures two important analytical steps:
     - *Evaluation*
     - *Projection*
   - Both could be the source of $g_m(.) \neq g_{m-1}(.)$
2. $i_m = f_m\left(\Omega_m^{CB}\right)$ - *Reaction Function*
   Select the appropriate interest rate as a function of this state.
   - Time variation in these functions
   - No MP shock - see also McMahon and Munday (2021)

Consider the FOMC in July 2021



Board of Governors of the Federal Reserve System

The Federal Reserve, the central bank of the United States, provides the nation with a safe, flexible, and stable monetary and financial system.

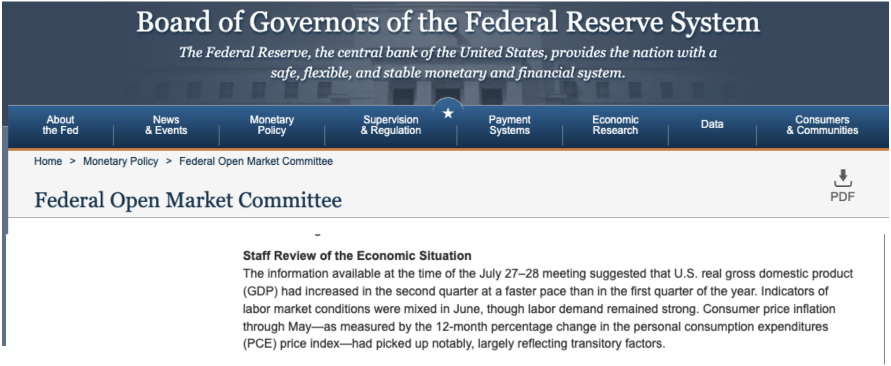
Federal Open Market Committee

Minutes of the Federal Open Market Committee

July 27-28, 2021

A joint meeting of the Federal Open Market Committee and the Board of Governors of the Federal Reserve System was held by videoconference on Tuesday, July 27, 2021, at 9:00 a.m. and continued on Wednesday, July 28, 2021, at 9:00 a.m.[1]

Consider the FOMC in July 2021

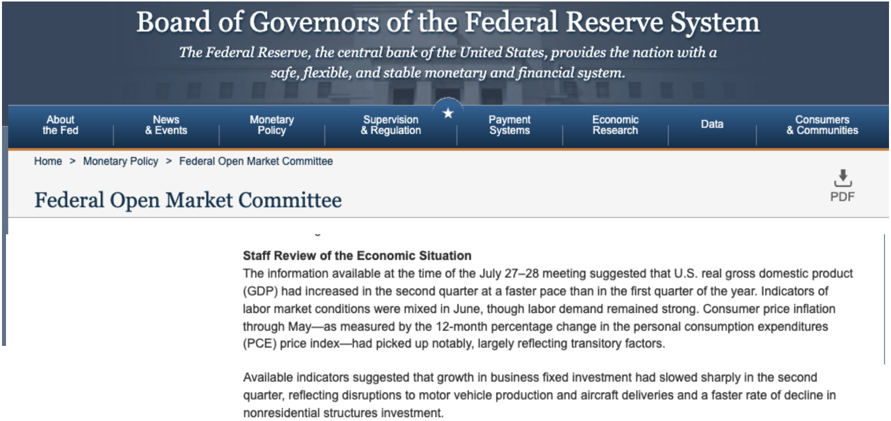

Board of Governors of the Federal Reserve System
The Federal Reserve, the central bank of the United States, provides the nation with a safe, flexible, and stable monetary and financial system.
About the Fed
News & Events
Monetary Policy
Supervision & Regulation
Payment Systems
Economic Research
Data
Consumers & Communities
Home > Monetary Policy > Federal Open Market Committee
PDF

Federal Open Market Committee

Staff Review of the Economic Situation

The information available at the time of the July 27–28 meeting suggested that U.S. real gross domestic product (GDP) had increased in the second quarter at a faster pace than in the first quarter of the year. Indicators of labor market conditions were mixed in June, though labor demand remained strong. Consumer price inflation through May—as measured by the 12-month percentage change in the personal consumption expenditures (PCE) price index—had picked up notably, largely reflecting transitory factors.

Consider the FOMC in July 2021



Board of Governors of the Federal Reserve System

The Federal Reserve, the central bank of the United States, provides the nation with a safe, flexible, and stable monetary and financial system.

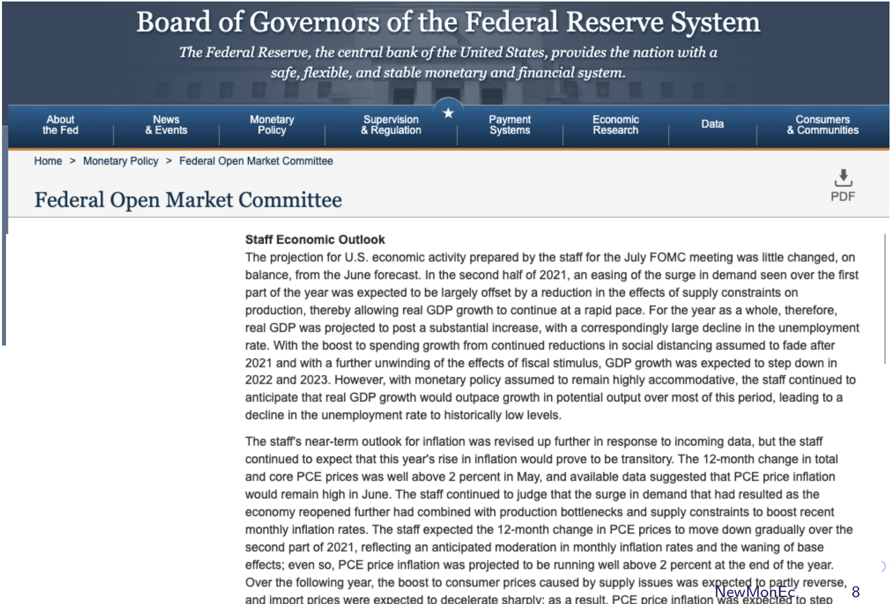
Home > Monetary Policy > Federal Open Market Committee

Federal Open Market Committee

Staff Review of the Economic Situation
The information available at the time of the July 27–28 meeting suggested that U.S. real gross domestic product (GDP) had increased in the second quarter at a faster pace than in the first quarter of the year. Indicators of labor market conditions were mixed in June, though labor demand remained strong. Consumer price inflation through May—as measured by the 12-month percentage change in the personal consumption expenditures (PCE) price index—had picked up notably, largely reflecting transitory factors.

Available indicators suggested that growth in business fixed investment had slowed sharply in the second quarter, reflecting disruptions to motor vehicle production and aircraft deliveries and a faster rate of decline in nonresidential structures investment.

Consider the FOMC in July 2021



Board of Governors of the Federal Reserve System

The Federal Reserve, the central bank of the United States, provides the nation with a safe, flexible, and stable monetary and financial system.

Federal Open Market Committee

Staff Economic Outlook

The projection for U.S. economic activity prepared by the staff for the July FOMC meeting was little changed, on balance, from the June forecast. In the second half of 2021, an easing of the surge in demand seen over the first part of the year was expected to be largely offset by a reduction in the effects of supply constraints on production, thereby allowing real GDP growth to continue at a rapid pace. For the year as a whole, therefore, real GDP was projected to post a substantial increase, with a correspondingly large decline in the unemployment rate. With the boost to spending growth from continued reductions in social distancing assumed to fade after 2021 and with a further unwinding of the effects of fiscal stimulus, GDP growth was expected to step down in 2022 and 2023. However, with monetary policy assumed to remain highly accommodative, the staff continued to anticipate that real GDP growth would outpace growth in potential output over most of this period, leading to a decline in the unemployment rate to historically low levels.

The staff's near-term outlook for inflation was revised up further in response to incoming data, but the staff continued to expect that this year's rise in inflation would prove to be transitory. The 12-month change in total and core PCE prices was well above 2 percent in May, and available data suggested that PCE price inflation would remain high in June. The staff continued to judge that the surge in demand that had resulted as the economy reopened further had combined with production bottlenecks and supply constraints to boost recent monthly inflation rates. The staff expected the 12-month change in PCE prices to move down gradually over the second part of 2021, reflecting an anticipated moderation in monthly inflation rates and the waning of base effects; even so, PCE price inflation was projected to be running well above 2 percent at the end of the year. Over the following year, the boost to consumer prices caused by supply issues was expected to partly reverse, and import prices were expected to decelerate sharply; as a result, PCE price inflation was expected to step

Market News

- Market Surprise: $\varepsilon_m^i = \mathbb{E}\left[\, i_m \,\middle|\, \mathcal{I}_m^{mkt} \right] - \mathbb{E}\left[\, i_{m-} \,\middle|\, \mathcal{I}_{m-}^{mkt} \right]$

Market News

- Market Surprise: $\varepsilon_m^i = \mathbb{E}\left[i_m \,\middle|\, \mathcal{I}_m^{mkt} \right] - \mathbb{E}\left[i_{m-} \,\middle|\, \mathcal{I}_{m-}^{mkt} \right]$

$$\varepsilon_m \approx \underbrace{\tilde{f}_m(\Omega_m^{mkt}) - \tilde{f}_{m-}(\Omega_m^{mkt})}_{\text{Updated Reaction function}} - \underbrace{\left[\tilde{g}_m(X_{m-}^{mkt}) - \tilde{g}_{m-}(X_{m-}^{mkt})\right] \tilde{f}'_{m-}(\Omega_m^{mkt})}_{\text{Reassessment}}$$

$$- \underbrace{(X_m^{mkt} - X_{m-}^{mkt})\tilde{g}'_m(X_m^{mkt})\tilde{f}'_{m-}(\Omega_m^{mkt})}_{\text{Effect of New Info}}$$

1. CB may choose to react to a given state of economy more or less aggressively than previously; $f_m(.) \neq \tilde{f}_{m-}(.)$.
2. CB could provide more details about $g_m(.) \neq \tilde{g}_{m-}(.)$
3. CB could reveal new information not in X_m^{mkt}

Measurement: The 3 Ts of Text Analysis

- The 3 Ts
  1. *Topic* – Often measured with Latent Dirichlet Allocation (LDA)
  2. *Tone* – e.g. Dictionary Methods, VADER
  3. *Time* – Rarely captured explicitly

Measurement: The 3 Ts of Text Analysis

- The 3 Ts
  1. *Topic* – Often measured with Latent Dirichlet Allocation (LDA)
  2. *Tone* – e.g. Dictionary Methods, VADER
  3. *Time* – Rarely captured explicitly

Temporal dimension in our analysis

Does the 3rd T helps with understanding the nature of what drives the information deficit?

- To distinguish the temporal dimensions of communication, we use 2 approaches to time tagging:
  1. Temporal tagging via SUTime
  2. Tense tagging via TMV

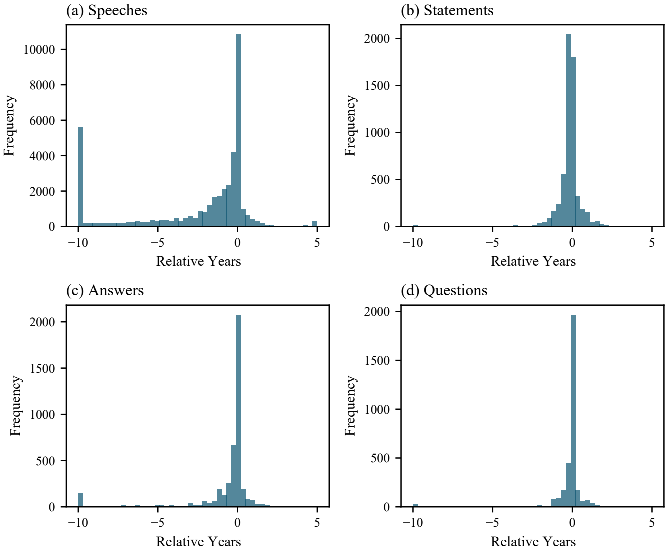
Example of SUTime Output

- With todays comprehensive package of monetary policy decisions, we are providing substantial monetary stimulus to ensure that financial conditions remain very favourable and support the euro area expansion, the ongoing build-up of domestic price pressures and, thus, the sustained convergence of inflation to our medium-term inflation aim.
- Let me now explain our assessment in greater detail, starting with the economic analysis.
- Euro area real GDP increased by 0.2%, quarter on quarter, in the second quarter of 2019, following a rise of 0.4% in the previous quarter.
- Incoming economic data and survey information continue to point to moderate but positive growth in the third quarter of this year.
- At the same time , the services and construction sectors show ongoing resilience and the euro area expansion is also supported by favourable financing conditions, further employment gains and rising wages, the mildly expansionary euro area fiscal stance and the ongoing albeit somewhat slower growth in global activity.

Draghi TMV Example

Parsed Text Data	Verbal Complex	Finite	Tense	Mood	Voice	Negation
Within our mandate ,	takes	yes	present	indicative	active	no
the ECB is ready to do	is	yes	present	indicative	active	no
whatever it takes	to preserve	no	-	-	-	-
to preserve the euro .	to do	no	-	-	-	-
And believe me ,	will be	yes	futurel	indicative	active	no
it will be enough .						

Allocations



Temporal Tagging: SUTime Measures

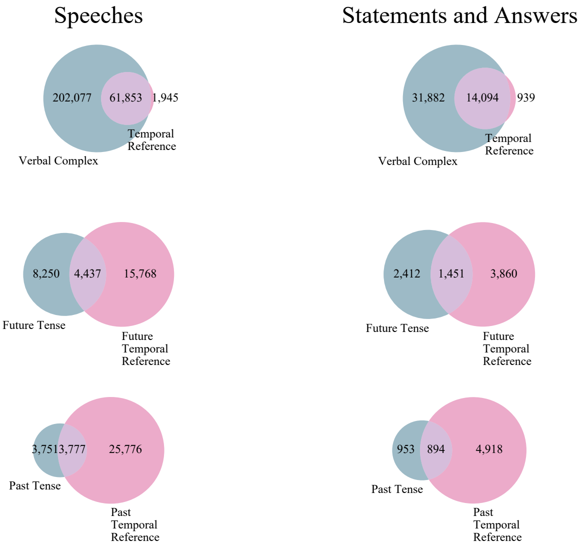
- We construct 2 measures of future (*past*) orientation using SUTime:

$$p_s^{f,CAT} = \frac{\sum_{i=1}^{i=N_s^{CAT}} \mathbf{1}\{CAT_{i,s} \ = \ Future(Past)\}}{N_s^{CAT}},$$

$$p_s^{f,NUM} = \frac{\sum_{i=1}^{i=N_s^{NUM}} \mathbf{1}\{NUM_{i,s} \ = \ Future(Past)\}}{N_s^{NUM}},$$

where N_s^{CAT} is the number of SUTime categorical references in speech s, N_s^{NUM} is the number of SUTime numerical references in speech s, and $\mathbf{1}$ is a dummy variable indicating when a numerical reference $NUM_{i,s}$ or a categorical reference $CAT_{i,s}$ is future (*past*).

Venn Diagram of Classifications



Measuring Temporal Topics

In order to measure the topic content of the future statements, we construct topic-specific future orientation measures:

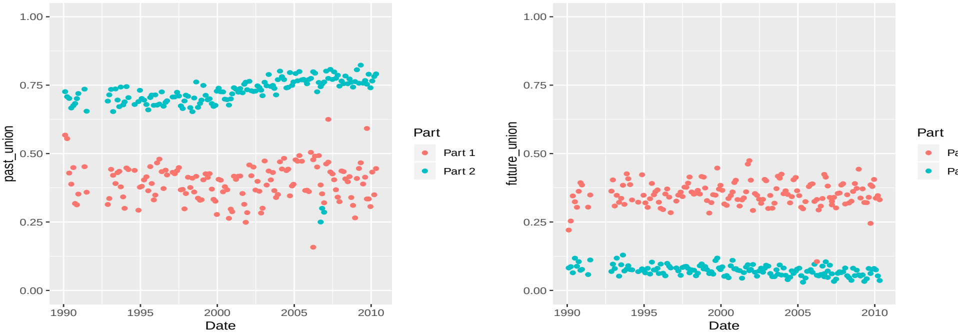
1. Identify future (*past*) statements using one/all of the measures:
   - SUTime Categorical
   - SUTime Numerical
   - TMV
2. Calculate the average topic share within this subset:

$$p_s^{f(p),X}(T_k) = \frac{1}{N_s^{X=Fut(Past)}} \sum_{i=1}^{i=N_s^X} \mathbf{1}\{X_{i,s} \ = \ Future(Past)\}\, \phi_{k,m(i)}$$

Essentially the measures are the weighted average of topic-shares from the sub-set of sentences associated with the future, according to three different metrics.

Temporal Dimensions of Monetary Policy

- FOMC Greenbooks (Tealbooks since June 2010):
  Part 1: Summary and Outlook $\equiv$ *Projection* $=$ Future & Past
  Part 2: Recent Developments $\equiv$ *Evaluation* $=$ Past



(a) Past

(b) Future

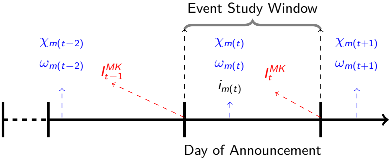
Temporal Info Effects

How important is temporal information?

- Regress event news on “narrative signals” - LASSO regression
- Focus on ECB Governing Council Decisions and Press Conferences
- Use a variety of data
- Key findings:
  - Communication is best captured by multi-dimensional signal vectors
  - Past information is about as equally as informative as future information in explaining the news.

Event Study

- Measuring beliefs via event study of asset news

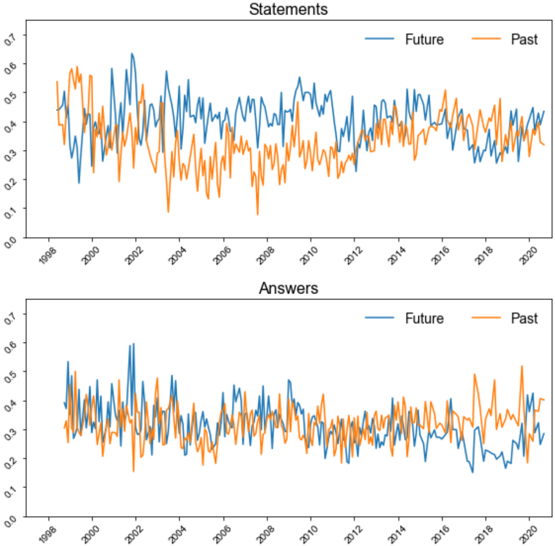


- Regress event news on "narrative signals" - LASSO regression

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-\mathbf{X}_i\beta)^2+\lambda\left[\frac{1}{2}(1-\alpha)||\beta||_2^2+\alpha||\beta||_1\right]\right\}$$

  - $\alpha = 0.99$
  - Estimate λ by 10-fold cross-validation
  - LASSO estimation via a non-parametric bootstrap - 5000 draws
  - Adjusted R^2 from OLS on the subset of LASSO-selected variables

Temporal Variation in ECB Statements and Answers



Expected Interest Rates

Table: Adjusted R^2 of yield curve by specification (union)

Specification	OIS	OIS	OIS	OIS	DE	DE
	1M	1Y	2Y	3Y	5Y	10Y
Topics Only	0.25	0.29	0.28	0.23	0.19	0.15
Topics and Future	0.32	0.36	0.37	0.31	0.25	0.19
Topics and Past	0.35	0.37	0.36	0.29	0.22	0.20
Topics, Future and Past	0.41	0.43	0.43	0.36	0.27	0.24
Topics, Future and Past*	0.59	0.62	0.63	0.59	0.56	0.47
Forecasts & Revisions	Yes	Yes	Yes	Yes	Yes	Yes

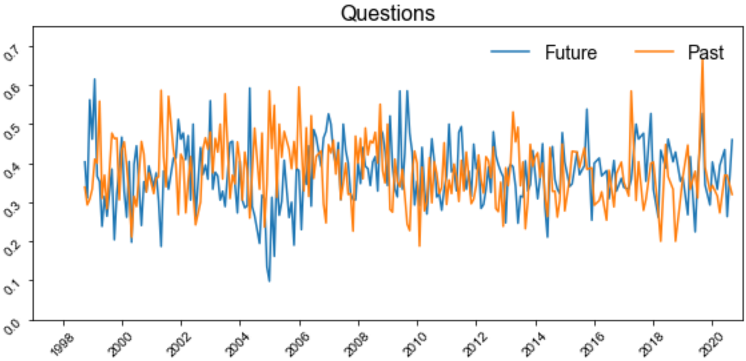
ABGMR factors

Table: Adjusted R^2 of ABGMR (2019) surprises by specification

Specification	Target	Timing	Forward Guidance	Quantitative Easing
Topics Only	0.23	0.28	0.29	0.21
Topics and Past	0.29	0.36	0.33	0.26
Topics and Future	0.37	0.36	0.34	0.27
Topics, Future and Past	0.42	0.42	0.37	0.32
Topics, Future and Past*	0.59	0.62	0.63	0.59
Forecasts & Revisions	Yes	Yes	Yes	Yes

Questions identifying information deficits

- It is hard to measure the information deficit
- One source is the journalists at the press conference
    - Their business model is to extract information that the market demands
- First Evidence: Questions are mix of past and future



Measuring similarity to question

- Idea: Typically answering questions involves talking about the same material as the question
- A measure of similarity:

$$Sim_{i,j} \equiv \frac{\sum_{k=1}^{n} (dtm_{ik} \times dtm_{jk})}{\sqrt{\sum_{k=1}^{n} dtm_{ik}^2} \times \sqrt{\sum_{k=1}^{n} dtm_{jk}^2}}$$

$n =$ number of terms in the document term matrix
dtm_{ik} is the kth term in the vector corresponding to document i.

Similarity

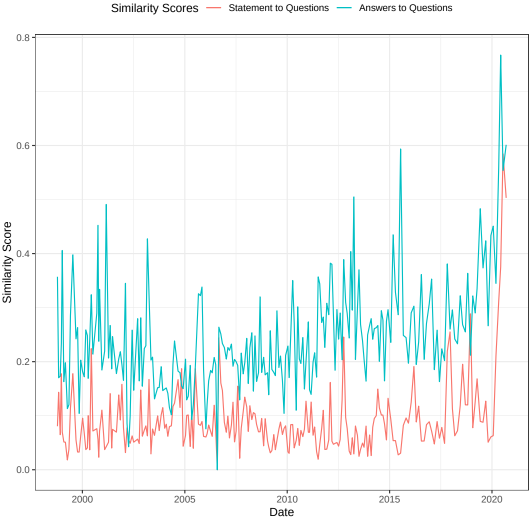


Figure: Similarity Scores

Similarity

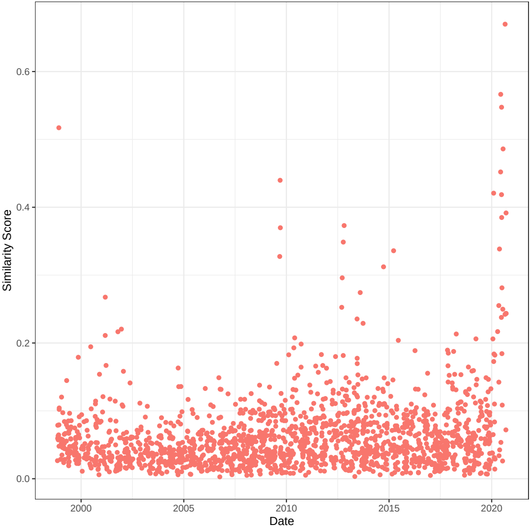


Figure: Similarity Scores

The right information constitutes news

Table: Speech-Question similarity and Information Deficit

	OIS 1M b/se	OIS 1Y b/se	OIS 3Y b/se	DE 5Y b/se	DE 10Y b/se
$Sim_{Sp,Q}$	-0.23	13.36***	13.66*	8.79	6.90
	(4.98)	(5.14)	(7.76)	(9.10)	(8.82)
$Sim_{A,Q}$	0.07	1.65	0.77	2.84	2.77
	(1.48)	(1.53)	(2.39)	(2.71)	(2.63)
$Sim_{Sp,Q} \times Sim_{A,Q}$	-8.8	-40.3**	-39.4	-31.7	-24.6
	(17.14)	(17.71)	(26.51)	(31.31)	(30.35)
Constant	-0.37	13.35*	7.10	14.18	21.32*
	(7.03)	(7.26)	(10.77)	(12.85)	(12.45)
Speaker FE	Yes	Yes	Yes	Yes	Yes
Macro controls	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Topics, Future and Past	Yes	Yes	Yes	Yes	Yes
N	1120.00	1120.00	1007.00	1122.00	1122.00
r2	0.26	0.41	0.34	0.22	0.15
Adj. R-Squared	0.19	0.35	0.27	0.15	0.07

Project 2

Cacophany of Voices

Extracting Economic Signals from Central Bank Speeches
joint with Maximilian Ahrens (University of Oxford)

- Blinder 2004:

> *"A central bank that speaks with too many voices may have no voice at all."*

Our key contributions

1. We provide to the research community a novel, monetary policy shock series based on central bank speeches
2. We construct a monetary policy signal dispersion index along three key economic dimensions: GDP, CPI and unemployment
    - More frequent than FOMC meeting series
    - Opens possibilities of answering new questions regarding CB communication
3. For example, do markets form different expectations when facing a ”cacophony of policy voices”? Our initial estimates suggest there might be evidence for it

08/08/2006 GB pt2:

Prices ⇒ Ec.Prices

Consumer price inflation has continued to move up, on balance, in recent months. The overall PCE price index rose at an annual rate of 3.9 percent during the first six months of the year, 1 percentage point faster than in the twelve months of last year. While rising energy prices have been a major source of the increase in overall consumer price inflation, the prices of core goods and services–particularly housing rents but also other core prices–have accelerated as well. Core PCE prices increased at an annual rate of 2.7 percent during the first half of this year, which is more than 1.0 percentage point higher than during 2005.

Although consumer energy prices declined 0.9 percent in June, PCE energy prices surged at an annual rate of about 25 percent over the first half of this year, up from the 17 percent increase posted in 2005. Furthermore, the downturn in energy prices in June was transitory: Spot prices for crude oil moved back up during July to a level near the peak for the year; weekly survey data on gasoline point to an increase of about 5 percent in July in the PCE price index for gasoline, a rise that would more than reverse the decline in June. Although most of the recent increase in gasoline prices reflected the higher cost of crude oil, the margin between retail gasoline prices and crude prices has moved up somewhat from an already high level despite a 35 percent decline since late June in the price of ethanol used for reformulated gasoline.

....

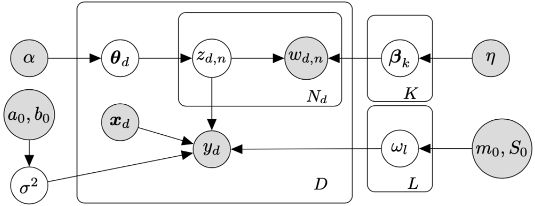
Text modelling

- Different text representation approaches possible (topics, word embeddings,...)
- We use a supervised topic modelling approach that learns a domain specific text representation that is optimized to predict the target variable together with other numerical covariates
  - Topic models proved to work well in relatively 'small' datasets (wide application in economics and other social sciences)
  - rSCHOLAR model (Card et al., 2018; Ahrens et al., 2021) provides ideal topic regression model for the task
  - Worth exploring different approaches for further work

Multimodal Bayesian Topic Regression

Maximilian Ahrens, Julian Ashwin, Jan-Peter Calliess and Vu Nguyen

1. The joint supervised learning approach of text representation and regression parameters allows for rigorous statistical inference when text as well as numerical features are relevant.
2. Allows us control for potential confounding factors, addressing the problem of omitted variables bias in supervised topic modelling for economic analysis.
3. Incorporating information from numerical features into the topic learning process can yield improved out-of-sample prediction performance.



Economic NLP modelling

The model estimation process is broken down into two steps:

1. Learn the mapping from central bank language to economic conditions based on Greenbook data
2. Apply the learned mapping to central bank speeches

Estimate mapping for 3 distinct economic signals:

1. GDP
2. CPI
3. Unemployment

Further work opportunities:

1. Possible to extend to additional economic dimensions
2. Possible to deploy different topic/language models to learn mapping

Example - mapping equation for CPI

Variables:

- $\Delta\pi_{4:0,m}$: change in the CPI forecast π over the next year at FOMC meeting timestamp m (target variable for CPI)
- θ_π: topic mixtures for the CPI corpus. $\theta_{\{\pi,g,u\},k}$ represents the k^{th} topic feature for the respective corpus
- $\pi_{0,m-1}, g_{0,m-1}, u_{0,m-1}$: controlling for lagged variables of GDP (g), CPI (π) and unemployment (u), both in levels and in differences

Full CPI language-to-forecast mapping equation:

$$\begin{aligned}\Delta\pi_{4:0,m} =& \rho_u u_{0,m-1} + \rho_\pi \pi_{0,m-1} + \rho_g g_{0,m-1} + \rho_{\Delta_u}\Delta u_{4:0,m-1} \\ &+ \rho_{\Delta_\pi}\Delta\pi_{4:0,m-1} + \rho_{\Delta_g}\Delta g_{4:0,m-1} + \sum_{k=1}^{K}\omega_k\theta_{\pi,k} + \epsilon_m \qquad (1)\end{aligned}$$

ρs and ωs: regression weights, ϵ: measurement error

Results - Estimating Implied Signals in Speeches

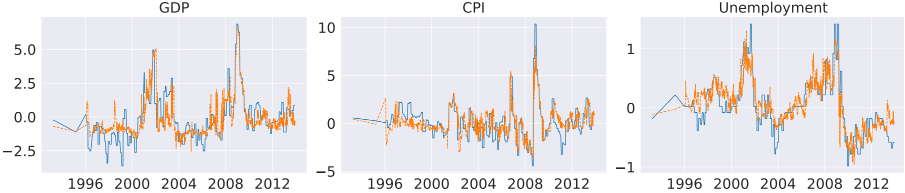


Figure: Out of sample implied policy signals: realised value, topic model estimation (K=20)

Results - Estimating Implied Signals in Speeches

predictive R^2	numeric	numeric + text
Speeches - GDP signal	0.524	**0.577 (0.016)**
Speeches - CPI signal	0.346	**0.575 (0.039)**
Speeches - Unempl. signal	0.630	**0.681 (0.019)**
Greenbook - GDP training	0.502	**0.766 (0.080)**
Greenbook - CPI training	0.295	**0.790 (0.147)**
Greenbook - Unempl. training	0.458	**0.657 (0.011)**

Table: Predictive R^2. Models trained on Greenbook dataset, tested on speeches dataset. Best model in bold. Reported means across 50 model runs, standard errors in brackets. Numeric (OLS) has analytical solution.

Results - Estimating Speech Dispersion (CPI example)

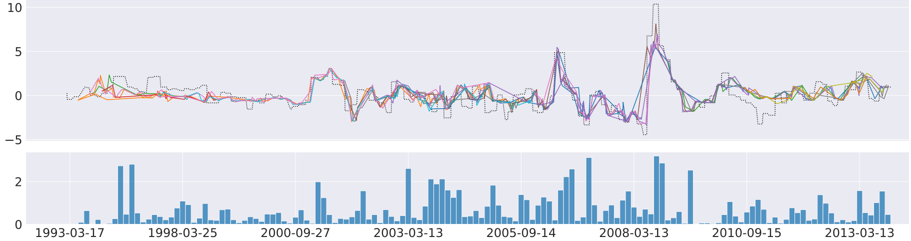


Figure: Out of sample estimation of monetary policy signals on CPI. Top figure: signal by individual central banker speaker. Bottom figure: derived dispersion measure (grouping window: inter-FOMC-meeting periods).

Results - Estimating Speech Dispersion

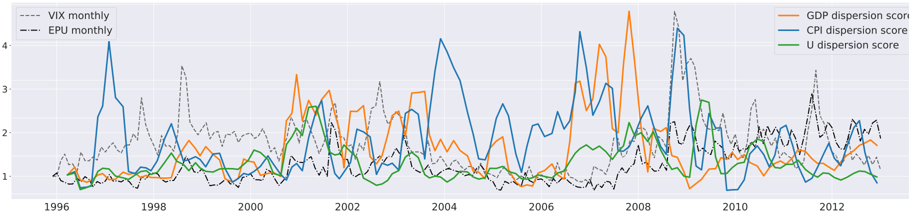


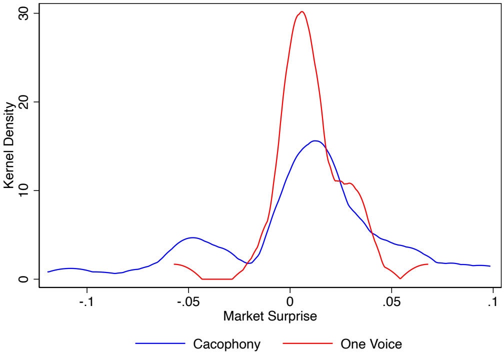
Figure: Dispersion scores for GDP, CPI and unemployment compared to VIX and Economic Policy Uncertainty (EPU) index. All indices re-indexed to beginning of displayed time-series.

Estimating Dispersion Effects

- We create a single indicator:
  1. Use the interquartile range (IQR) for each of the GDP, CPI and unemployment series.
     - Simple, and protects against outliers
  2. Average these 3 dispersion series
     - Cacophony could be driven by different signals on only a subset of the indicators.
     - Therefore, alternative measure: average the two most-dispersed series
  3. This is the *Dispersion Index* variable

- Alternative is to create dummy indicating an intermeeting period as one of *Cacophony* when the *Dispersion Index* variable is above the median.
  - Results are similar

Kernel Density of Market Surprises



- Market surprises are calculated using a narrow, 30-minute window around the FOMC announcement

Effect of Cacophony on subsequent Market News

- $|MktNews|_t = \alpha + \beta X_t + \gamma Dispersion\ Index_{t-1} + \epsilon_t$

Regressors	(1) Mkt News	(2) Mkt News	(3) Mkt News
Lagged Dispersion Index		0.019*** [0.002]	
Lagged Dispersion Index (alt)			0.014*** [0.004]
Controls	YES	YES	YES
R-squared	0.180	0.262	0.261

- Controls: NBER recession indicator, number of speeches, market volatility (VIX), policy uncertainty (BBD) and the average signal for each indicator.

Why Does This New Era of Communication Matter?

What we need:

1. Answers to new questions, and new answers to old questions
2. Creation and utilisation of new data resources
3. Application and development of new methodologies

⇒ New avenues for future research

END

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Cloud computing research collaboration: an application to access to cash and financial services[1]

Danielle V Handel (Stanford Institute for Economic Policy Research, Stanford University),
Anson T Y Ho (Ted Rogers School of Management, Toronto Metropolitan University),
Kim P Huynh (Bank of Canada), David T Jacho-Chavez and Carson Rea (Emory University)

[1] This presentation 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 central banks and other institutions represented at the event.

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-Chávez[¶]
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.

JEL codes: A11; A22; A23; C87; C88.

[*] We are grateful to Giuseppe Bruno for his helpful comments and suggestions. We also thank Brian Nadon, Phil Riopelle and the Analytical Environment Business Systems team at the Bank of Canada for their excellent assistance in the Digital Analytical Zone. The views expressed in this article are those of the authors. No responsibility for them should be attributed to the Bank of Canada. All remaining errors are the responsibility of the authors.

[†] Stanford Institute for Economic Policy Research, Stanford University, John A. and Cynthia Fry Gunn Building, 366 Galvez St, Stanford, CA 94305, USA. E-mail: dvhandel@stanford.edu

[‡] Ted Rogers School of Management, Toronto Metropolitan University, 55 Dundas Street West, Toronto ON, M5G 2C3, Canada. E-mail: atyho@ryerson.ca

[§] Corresponding Author: Bank of Canada, 234 Wellington Ave., Ottawa ON, K1A 0G9, Canada. E-mail: khuynh@bankofcanada.ca.

[¶] Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr., Atlanta,GA 30322-2240, USA. E-mail: djachocha@emory.edu

[‖] Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: chrea@emory.edu

1 Introduction

Demand for computational resources have been rapidly expanding in recent years, driven by the increasing interest in big data, new analytical methods in data science, and a slowdown in technological progress. In response, cloud computing has become a popular solution for institutions to meet their computational needs. At the Bank of Canada, the *Digital Analytical Zone* (DAZ) is a cloud computing platform implemented through Microsoft® Azure, their contracted vendor. It allows for an *on-demand* computing service that is agile and highly scalable to meet the resource requirement of different projects, aligning resources with needs and making high performance computing more accessible.

This case study illustrates the use of the Bank of Canada's DAZ to access big data tools for research collaboration with external academic researchers. Our data analysis is conducted on Azure Databricks, an Apache Spark-based data analytics service. As shown in Handel et al. (2021), cloud computing is convenient as it removes infrastructure constraints. However, depending on the application complexity, setting up virtual machines on the cloud may still require considerable initial cost and expertise in cloud computing. Managed cloud service platform simplifies the use of cloud computing by providing a pre-configured computational service. In our case, Azure Databricks provides a fully managed Spark cluster that enable researchers to easily harness the power of data parallelism for large scale data processing.

The Bank of Canada's DAZ responds to the challenges and options highlighted in Bruno et al. (2020). It shows a promising path to efficiently employ big data analysis. This cloud-based platform provides the opportunity to easily leverage the increasingly complex "financial big data sets" and work with innovative data to yield new insights important to the functions of central banks. For example, Canadian consumer credit data is used to analyze the effect of COVID-19 on consumer finance (Ho, 2020) and the interdependence of financial institutions in the consumer credit markets (Ho et al., 2021). For the rest of the paper, Section 2 describes the cloud computing service at the Bank of Canada. Section 3 provides a use case of cloud computing, and Section 4 concludes.

2 Cloud Computing at the Bank of Canada

The Bank of Canada currently offers computational resources through several different channels. Researchers can access an on-premise high-performance-computing (HPC) cluster named *Edith2*.[1] In addition, the Bank of Canada also provides a cloud computing environment called the *Digital Analytical Zone (DAZ)* to support scientific research (Elsey et al., 2021). In general, the DAZ is designed to be completely separate from the Bank of Canada's network. It provides an environment for users to experiment with different ideas. The DAZ is supported by Microsoft Azure®, the Bank of Canada's service vendor, which provides various types of cloud computing services. *Research Services* is a managed cloud computing platform, where users can launch their virtual machines with pre-configured specifications. Based on business needs, the DAZ also provide a

[1] For more information, see Collignon (2019).

more flexible *Research Lab* platform, closer to full-fledged Microsoft Azure®, where users can configure their virtual machines.

The DAZ offers several advantages over an on-premise HPC cluster. First and foremost, it provides users with *on-demand* computing, with which users do not have to wait in queue due to Edith2's capacity limit. Timely access to a computational resource increases users' productivity, particularly when there is expanding demand from big data and more complex data science methods. It also allows users to access a computational resource on the internet without the need of pre-configured physical device. Second, the DAZ is more flexible in providing up-to-date services than an on-premise HPC cluster. While most of the cutting-edge data science methods come from community-contributed libraries, installing them on an on-premise HPC cluster often requires extensive testing and system configuration. On the other hand, the DAZ is maintained by the service vendor, which benefit from the economies of scale and affords users to have a higher level of administrative rights. This flexibility encourages users to explore new techniques and promotes innovation. Third, the DAZ can easily extend the Bank of Canada's computational resource to external partners for project collaboration. DAZ administrators can simply create accounts for external partners on the cloud platform for instant collaboration, instead of granting external partners access to the on-premise HPC cluster that may involve costly equipment and lengthy security clearance.

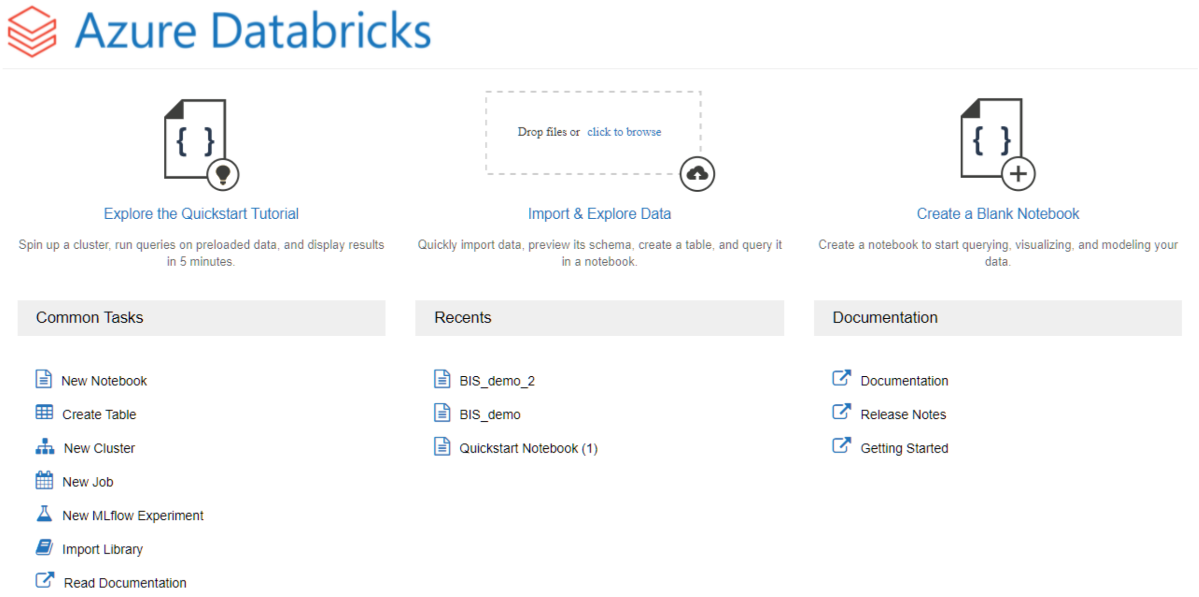
In this paper, we demonstrate how the DAZ can facilitate collaborations between the Bank of Canada researchers and external partners. While the interface and the functions available on the DAZ are identical to a typical Microsoft Azure® portal, some functionality requires prior approval from the DAZ administrator. To initialize a project, a DAZ administrator sets up a resource group on Azure according to the user's business case. If external collaborators are involved, additional accounts are created and assigned to that specific resource group.[2] All users within a resource group shares the same pool of computational resource. For big data analysis and machine learning, we further focus on Azure Databricks, among other services available.

2.1 Azure Databricks

Azure Databricks is an Apache Spark-based data analytics service. It supports multiple programming languages, including Python, Scala, R and SQL. It also supports popular machine learning libraries such as Apache Spark MLlib, Tensorflow, Pytorch, allowing for the use of both advanced statistical and machine learning techniques. Its markdown-compatible notebook environment also provides the opportunity for clear documentation, efficient debugging, and promoting reproducible research.

After entering the Azure portal, utilizing these tools involves first creating and naming a Databricks resource. Researchers can then launch a Databricks workspace, which is an interactive interface for managing all of the Databricks tools. Figure 1 shows the layout of Azure Databricks. Within the workspace, there are various options for launching a project-specific cluster, which is a collection of servers that will provide the

[2] The DAZ's Azure account is independent of other Microsoft Azure accounts that a user may have.


Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Quickly import data, preview its schema, create a table, and query it in a notebook.

Create a notebook to start querying, visualizing, and modeling your data.

Figure 1: Screenshot of Microsoft Azure Databricks

computing power needed to complete big data tasks. For the demonstration in this paper, we provision a standard cluster with 14 GB of memory and 4 cores. Memory size and parallel computing specifications are highly customizable, with options for additional memory and cores and different runtimes. Also available is the serverless option that allows for auto-scaling, meaning that the resources employed will be used efficiently and adjusted upon need. As opposed to shared on-premise HPC cluster, these on-demand project-specific clusters can be deployed without going through a job scheduler. After initiating a cluster, any project data can be uploaded into the Databricks File System (DBFS), converted into a table using the user interface, and subsequently used in any projects in the workspace. Researchers may refer to any data stored in DBFS using absolute file paths as if using data stored on a local machine.

In order to execute analyses using big data tools and interact with Apache Spark, researchers may use the interactive Databricks notebooks. After either uploading an existing notebook or creating a new Databricks notebook, attaching the notebook to a running cluster will allow for interactive management of Spark resources within the Jupyter-style interface. For the demonstration, we interact with Spark within a Python script using the PySpark interface. The Databricks notebooks provide detailed metrics regarding Spark performance and runtime, and researchers can interact with tables and other objects to quickly produce plots and summary statistics or save output to their local machine. Researchers may choose to link Databricks notebooks to Github repositories using built-in Git integration, which allows for version control working with notebook revision histories and enhanced capability for collaboration. The built-in Git integration also includes the capability to create branches and pull requests in the relevant repository from within the Databricks UI.

3 Application – Canadian Access to Financial Services

In our empirical application, we measure Canadians' access to financial services by computing the distances between their residential locations and the nearest branch of a financial institution (FI) in 2017. Physical proximity as a measurement for access to financial services is supported by Mintel's (2018) survey findings that consumers rely on access to physical branches for the purchase of complex financial products or first-time banking interactions. Linking a spatial network of FIs with methods-of-payment survey further yields a comprehensive analysis on the role of banking in influencing consumer payment choices (Henry et al., 2018) and their cash withdrawal (Chen et al., 2021a).

To measure the physical proximity to local branches, we compute the straight-line distances from the population centroid of each postal code to all FIs, and then identify the distance to the closest branch. The same methodology is used in Tischer et al. (2020) and Chen et al. (2021b). The postal code data set comes from Statistics Canada's Postal Code Conversion File (PCCF) and the addresses of financial institutions are reported in Payment Canada's Financial Institutions File (FIF). We further appended the postal codes data set with additional demographic information at census-dissemination-area level from the 2016 Canadian Census.

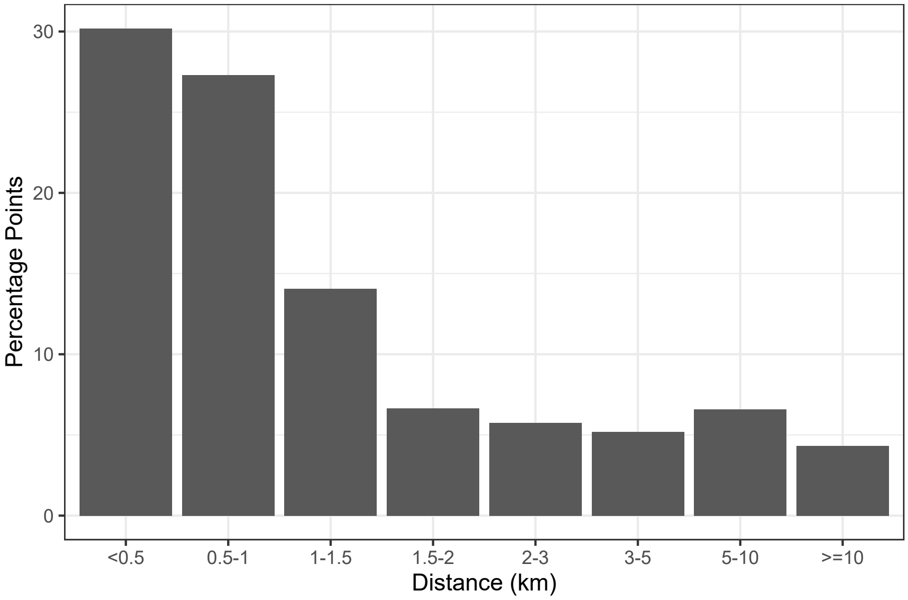
This application demonstrates the value of a cloud computing platform, by expediting a process which typically requires high local computing power and extensive time. In 2017, FIs operated a total of 11,029 local branches across Canada, serving Canadians resided in 765,723 different postal codes. To address the volume of data and complex spatial calculations, we employ Apache Spark's dimension reduction algorithms in Azure Databricks.

Our estimates on the access to financial services are illustrated in Figure 2. Overall, 57.4% of credit active Canadian residents have access a FI branch within 1 km of their residential location. The most common distance is less than 0.5 km, which contains about 30.1% of all residents. This suggests that most residents have convenient access to FI branches for financial services. While the fraction of people farther away from FI branches drops quickly, the distribution also exhibits a long tail with 4.3% residents having the nearest branch being more than 10 km away.

Canadian residents' proximity to financial services is related to FIs strategically locating their branches (Allen et al., 2008; Chen and Strathearn, 2020). As shown in Figure 3, vast majority of residents in the highest density quintile has access to a branch with 1 km. Distance to FI branches increases gradually for people living in areas with lower population density. Notably, majority of people residing in the lowest density quintile have to travel 3 km or more to visit a FI for financial service.

While we observed distinct patterns in the access to financial service, we did not find lower income groups are disadvantage in access to FI branches. Indeed, we observed the opposite. In general, residents in the lowest income quintile neighborhoods have closest access to local branches, with almost 50% of the residents having access in less than 0.5 km. Proximity to FI branches decreases steadily in higher income neighborhoods, and the distribution also becomes more skewed. Nonetheless, in the second to the fifth income

Figure 2: Overall Access to Financial Service



quintile neighborhoods, about 70% of residents have the nearest branches within 1.5 km.

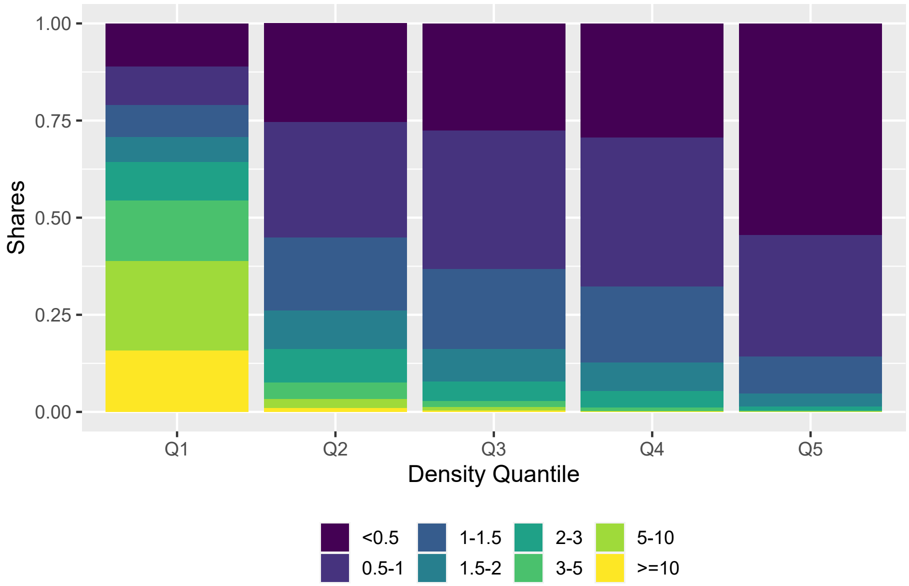
In terms of spatial pattern, we categorize postal codes into regions using the Statistical Area Classification (SAC) provided by Statistics Canada. Specifically, the land mass of Canada is divided into census metropolitan areas (CMAs), census agglomerations (CAs), and census metropolitan influenced zones (MIZs) based on population size and commuting patterns. A summary of SAC and their shares of population is reported in Table 1. Graphically, the SAC for Ottawa-Montreal-Quebec City region is illustrated in Figure 5.

Table 1: Statistical Area Classification

Classification	Description	Pop. Share
CMA	Total population ≥100,000; core population ≥50,000	0.714
CA	Total population <100,000; core population ≥10,000	0.123
Strong MIZ	≥30% of employed residents commute to work in CMA or CA	0.055
Moderate MIZ	5% - 30% of employed residents commute to work in CMA or CA	0.066
Weak MIZ	0% - 5% of employed residents commute to work in CMA or CA	0.036
No MIZ	No employed residents commute to work in CMA or CA	0.005

Access to financial service exhibits a unique spatial pattern. As shown in Figure 6, residents in CMAs have the shortest distance to FI branches, since these areas have higher levels of urban development and business activities. Distance to branches increases with areas farther away from CMAs, potentially due to smaller population size. It also shows a more skewed distribution, in which some residents are very far away from any branches. For instance, about 50% of the strong MIZ residents have to travel at least 5 km for visiting a FI.

Figure 3: Access to Financial Service by Population Density



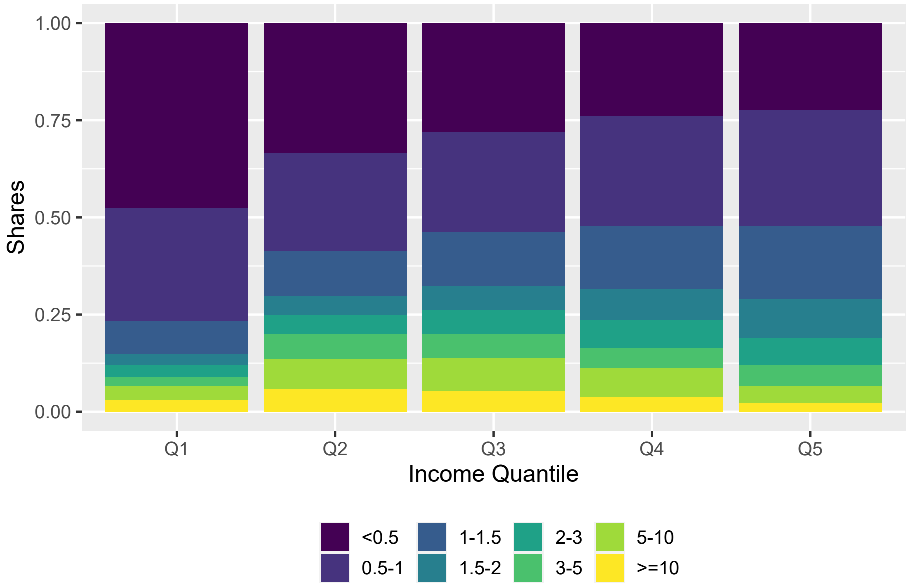
Most interestingly, distance to bank branches *decreases* from strong MIZ to weak MIZ. The fraction of residents with access in less than 1 km increases significantly from strong MIZ to weak MIZ, while that with access between 3-10 km decreases. It implies that people living outside of sizable cities have closer access to banks as the influence of metropolitan area fades. The economic explanation is that people can also access financial services via their daily trips to work. With smaller fractions of residents in moderate (5% to 30%) and weak MIZ (less than 5%) commuting to CMA or CA for work, there are stronger local banking needs that warrants the operation of a local branch. Such clustering is even more obvious in no MIZ areas, where the distribution of banking access shows a bimodal distribution with about 24% of residents having access in less than 0.5 km and about 50% of them having to travel for more than 10 km.

4 Conclusion and Considerations

We illustrate an example of how cloud computing is used at the Bank of Canada for research collaboration with external partners. Our use case shows that cloud computing is scalable and capable to handle computationally intensive data analysis. Most importantly, the cloud computing platform at the Bank of Canada allows for easy resource sharing among collaborators in different institutions, different regions, and different time zones. It abstracts out institutional-specific infrastructure configurations, providing a common platform for users to collaborate on their data analysis.

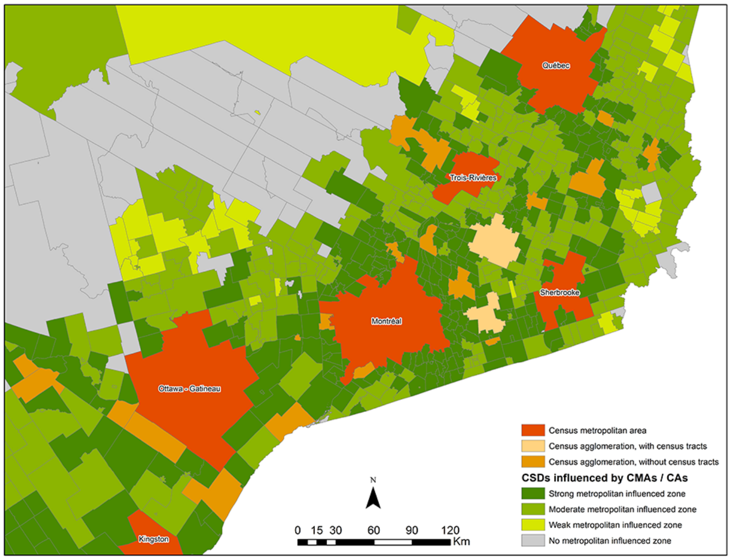
While cloud computing brings resource flexibility to institutions, implementing a cloud computing platform also involve various considerations. Migrating to a cloud platform involves training and additional support for users to adapt to a new infrastructure and a new workflow. The extra cost of time and support

Figure 4: Access to Financial Service by Income Quintile



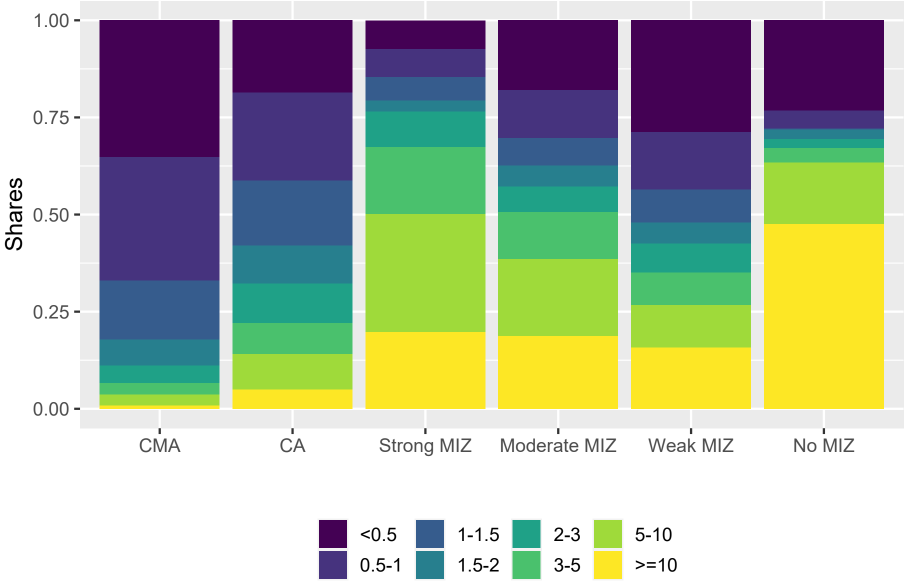
personnel should be included when comparing different solutions for fulfilling users' computational needs, at least in the short run. Furthermore, a usage policy should be set up for the implementation of cloud computing. This may entail the amount of resource and computation time budgeted for a project, as well as the type of data that can be stored on the cloud.

Figure 5: Example of Statistical Area Classification


Québec
Trois-Rivières
Sherbrooke
Montréal
Ottawa - Gatineau
Kingston
N
0 15 30 60 90 120
Km
Census metropolitan area
Census agglomeration, with census tracts
Census agglomeration, without census tracts
CSDs influenced by CMAs / CAs
Strong metropolitan influenced zone
Moderate metropolitan influenced zone
Weak metropolitan influenced zone
No metropolitan influenced zone

Source: Statistics Canada, 2016 Census of Population.

Figure 6: Access to Financial Service by Area Type



References

Allen, Jason, Robert Clark, and Jean-François Houde, "Market Structure and the Diffusion of E-Commerce: Evidence from the Retail Banking Industry," Technical Report, Bank of Canada Staff Working Paper 2008-32 2008.

Bruno, Giuseppe, Hiren Jani, Rafael Schmidt, Bruno Tissot, Bank für Internationalen Zahlungsausgleich, and Irving Fisher Committee on Central Bank Statistics, *Computing platforms for big data analytics and artificial intelligence* 2020. OCLC: 1187922905.

Chen, Heng and Matthew Strathearn, "A Spatial Model of Bank Branches in Canada," Staff Working Paper 2020-4, Bank of Canada February 2020.

_ , _ , and Marcel Voia, "Consumer Cash Withdrawal Behaviour: Branch Networks and Online Financial Innovation," Technical Report, Bank of Canada Staff Working Paper 2021-28 2021.

_ , Walter Engert, Kim P. Huynh, and Daneal O'Habib, "An Exploration of First Nations Reserves and Access to Cash," Staff Discussion Paper 2021-8, Bank of Canada 2021.

Collignon, Barbara, "BOC's Analytic Environment: A Leap Into The Future," in "Bank of Italy and BIS Workshop on 'Computing Platforms for Big Data and Machine Learning"' 2019.

Elsey, Rob, Richard Harmon, Ulrika Pilestl, Ben Sorensen, and Dirk Robijns, "Central bankers at the frontier: the state of the art in advanced analytics and AI," in "BIS Innovation Summit 2021" 2021.

Handel, D.V., A.T.Y. Ho, K.P. Huynh, D.T. Jacho-Chavez, and C. Rea, "Econometrics Pedagogy and Cloud Computing: Training the Next Generation of Economists and Data Scientists," *Journal of Econometric Methods*, 2021, *10* (1), 89–102.

Henry, Christopher, Kim Huynh, and Angelika Welte, "2017 Methods-of-Payment Survey Report," *Bank of Canada Staff Discussion Papers*, 2018, (18-17).

Ho, Anson T. Y., "Interconnectedness through the Lens of Consumer Credit Markets," in Á de Paula, E Tamer, and M C Voia, eds., *The Econometrics of Networks (Advances in Econometrics, Vol. 42)*, Emerald Publishing Limited, oct 2020, pp. 315–333.

_ , Lealand Morin, Harry J. Paarsch, and Kim P. Huynh, "Consumer Credit Usage in Canada during the Coronavirus Pandemic," *Canadian Journal of Economics*, 2021, *Special issue: The COVID-19 Pandemic* (54). forthcoming.

Mintel, "The Branch Banking Experience - Canada - February 2018," Technical Report, Mintel February 2018.

Tischer, Daniel, Isobel Oxley, Jamie Evans, and Richard Scott, "Where to Withdraw: National Mapping of Access to Cash," December 2020.

BANK OF CANADA
BANQUE DU CANADA

EMORY UNIVERSITY

Toronto Metropolitan University

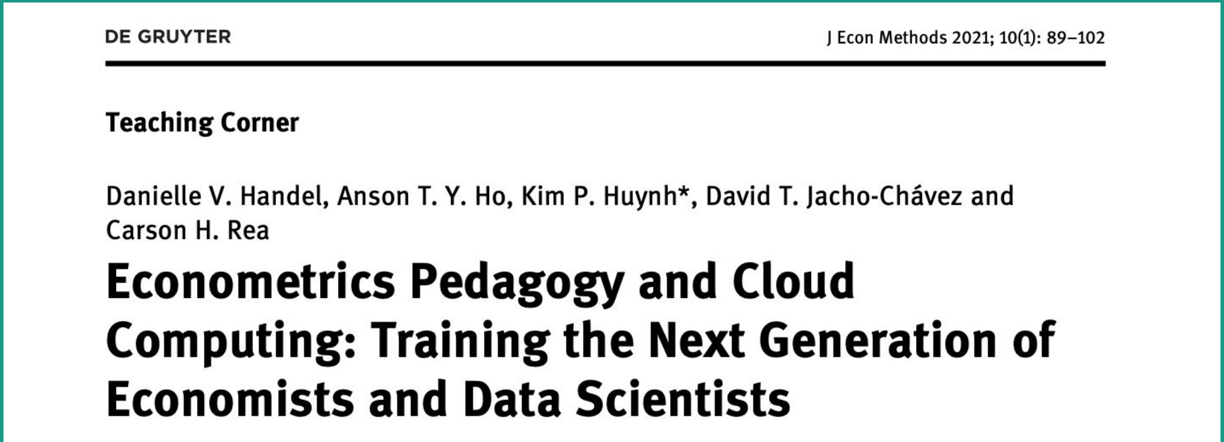
Stanford University

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

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

THE VIEWS EXPRESSED IN THIS PRESENTATION ARE THE AUTHORS' AND DO NOT REPRESENT THE OFFICIAL VIEWS OF THE BANK OF CANADA.

Harnessing the power of on-demand cloud computing for collaborative research at the Bank of Canada



DE GRUYTER J Econ Methods 2021; 10(1): 89–102

Teaching Corner

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

Econometrics Pedagogy and Cloud Computing: Training the Next Generation of Economists and Data Scientists

High Performance Computing (HPC) Resources at the Bank of Canada

- Edith2
  - On-premise HPC appliance
  - For use with sensitive data

- Digital Analytical Zone (DAZ)
  - Cloud computing platform
  - Supported by Microsoft Azure

Photo: Edith Whyte. ca 1966. Bank of Canada Archives (PC223-17) Credit: Unknown.

See: “Central bankers at the frontier: the state of the art in advanced analytics and AI” for a detailed description of this infrastructure

Cloud Computing for Research Collaboration

Advantages

- On-demand
- Scalable
- Easy collaboration with external researchers
- Low startup costs

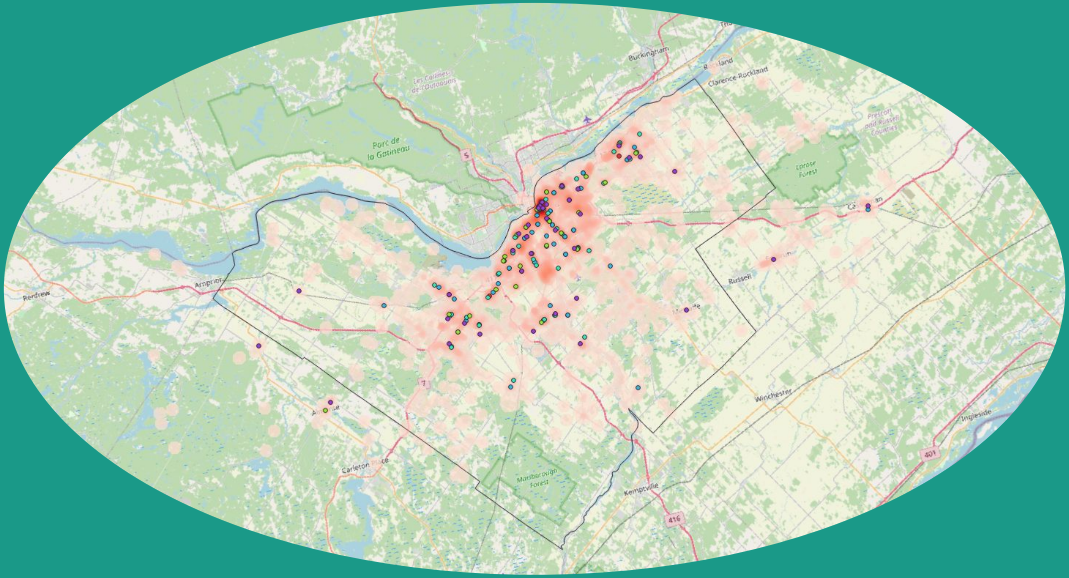
Considerations

- Latency
- Training costs
- Budget restrictions

Microsoft Azure Databricks for Research Collaboration

- Jupyter-style notebook interface
- Fully managed clusters
- Machine learning tools
- Integrated version control

Empirical Use Case



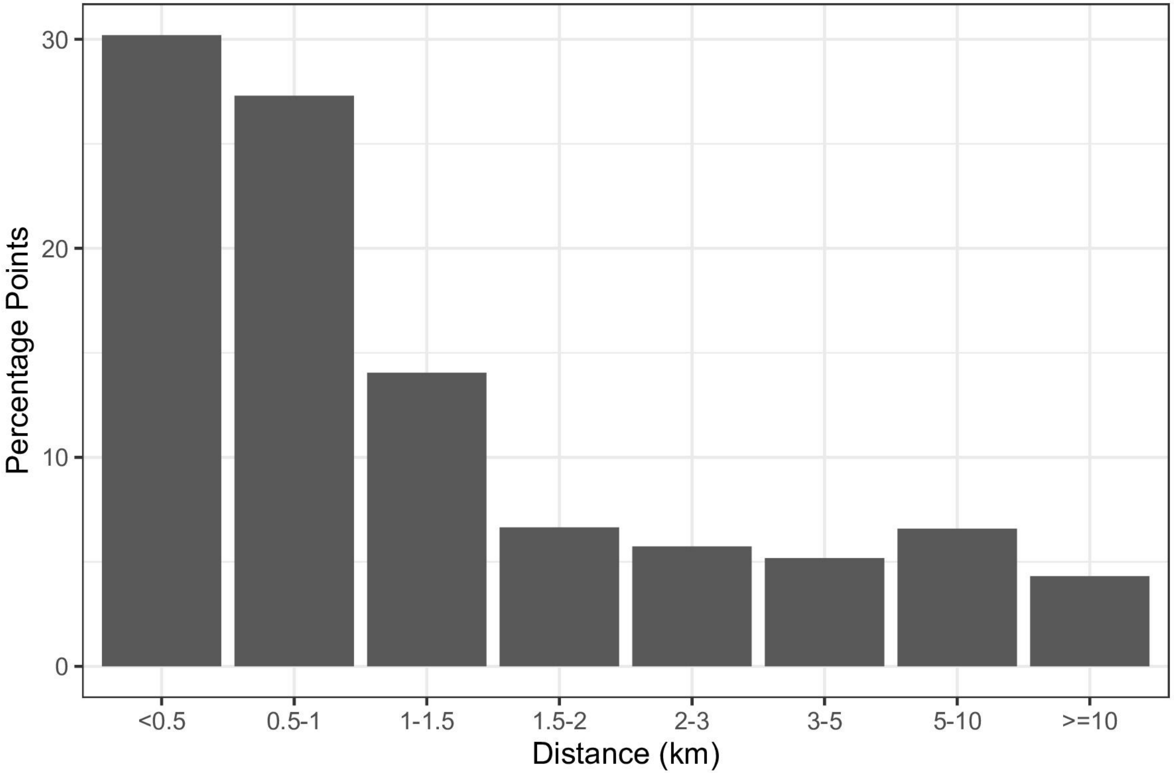
Access to Financial Services in Canada

Mapping consumers and their nearest bank branch

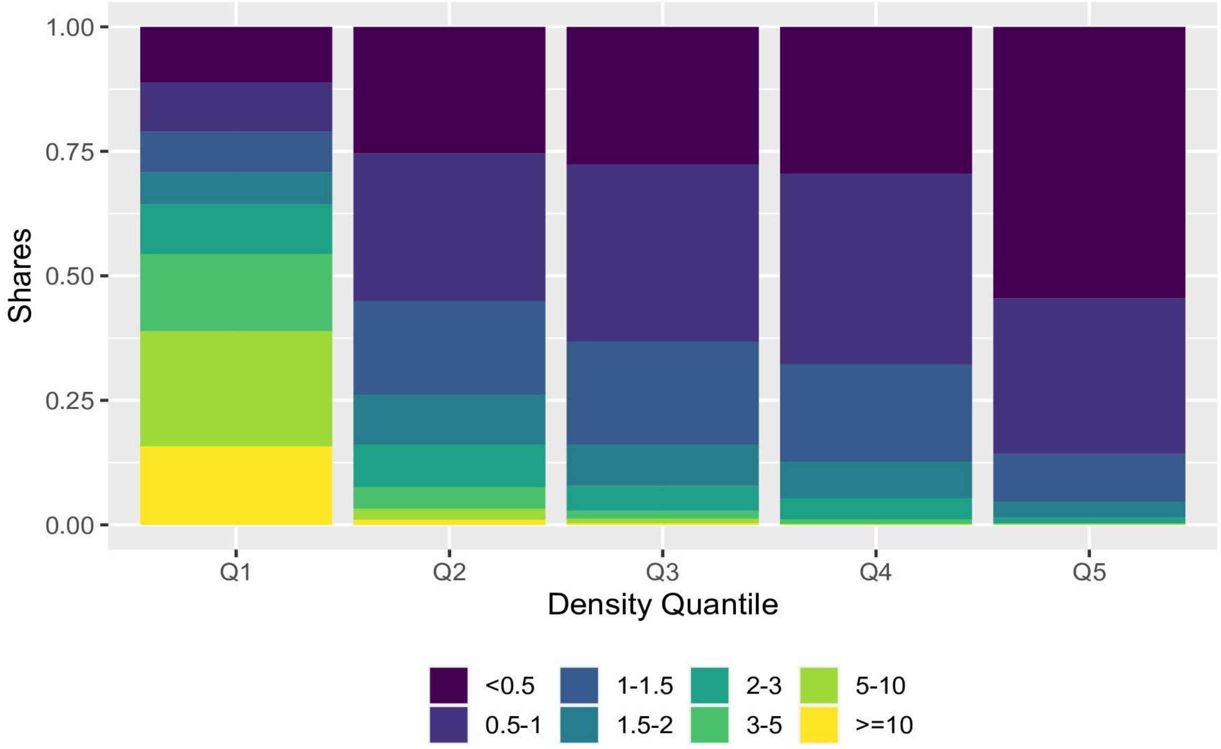
- Challenge: **24,000+** postal codes
- Computed straight line distances from population centroids to financial institutions
- Use Jupyter/PySpark docker image to manage computational needs

See: “A Spatial Model of Bank Branches in Canada” Staff Working Paper 2020-4 (English) by Heng Chen, Matthew Strathearn

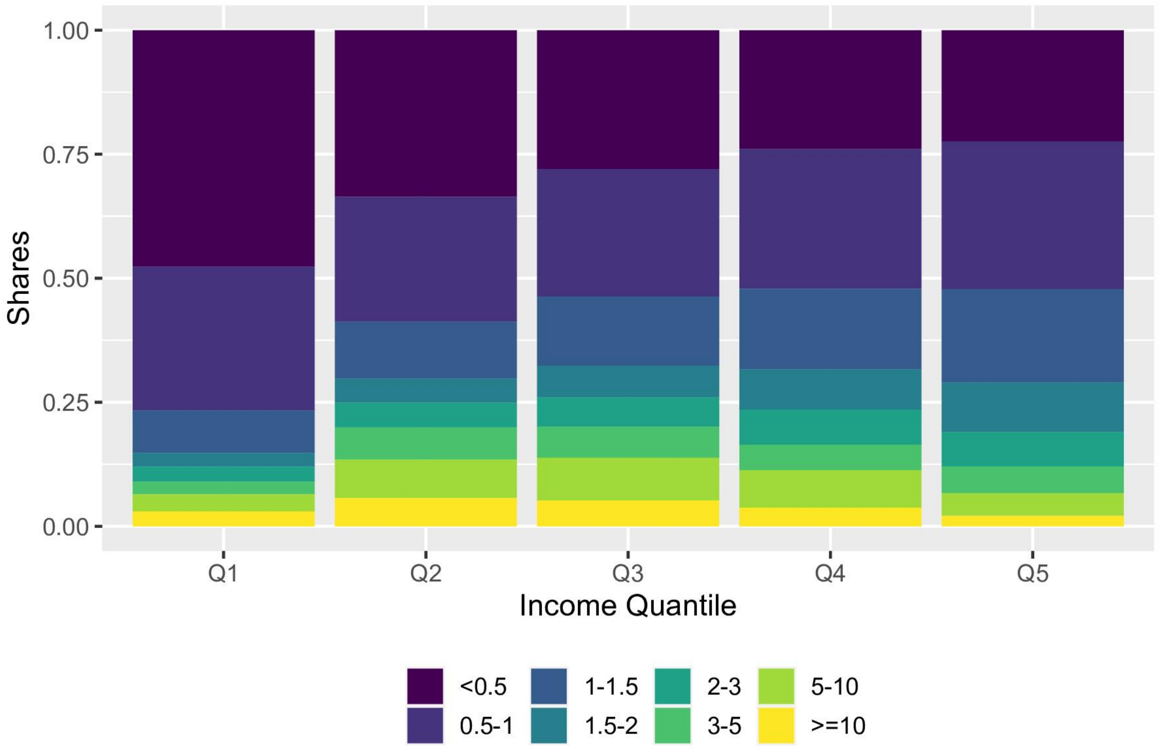
Over 50% of Canadians have access to a bank branch within 1 km



Banks are located in the most densely populated areas



Bank branches are distributed across diverse income levels



The Added Value of Cloud Computing

- Custom, scalable resources
- Streamlined collaboration
- Low startup costs

10
CANADA
SPECIMEN
VIOLA DESMOND
BANK OF CANADA • BANQUE DU CANADA
DIX • TEN DOLLARS
10
10
JHS2912607
JHS2912607
SPECIMEN
CANADIAN MUSEUM FOR HUMAN RIGHTS
MUSÉE CANADIEN POUR LES DROITS DE LA PERSONNE
10
CANADA

Thanks/Merci

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 "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Machine learning for anomaly detection in datasets with categorical variables and skewed distributions[1]

Matteo Accornero and Gianluca Boscariol,
European Central Bank

[1] This presentation 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 central banks and other institutions represented at the event.

EUROPEAN CENTRAL BANK

EUROSYSTEM

Machine learning for anomaly detection in datasets with categorical variables and skewed distributions

IFC-BIS Banca d’Italia
22/10/2021

Matteo Accornero & Gianluca Boscariol
European Central Bank – DG Statistics

Overview

1 Introduction and motivation

2 Anomaly detection algorithms used

3 Workflow and pipeline plumbing interventions

1

Introduction and motivation

Background

- MMSR (Money Markets Statistical Reporting) is a granular (transaction by transaction) dataset collecting data on money markets with daily frequency in the euro area
- In 2018 the ECB embarked in a new anomaly detection project to support the data quality checks connected with the euro short-term rate (€STR) production.
- In development phase, several challenges were identified, in particular with regards to:
  - Workflow: feedback loop needs an effective and sustainable flow of information among involved parties
  - Performance: daily data requires timely data quality information flows, quick reaction times, a lean process
  - Categorical variables: MMSR has few numerical variables, which poses challenges to analysis
  - Skewness: similarly to other granular financial datasets, the *exceptions* are the *rule*
  - Interpretability: how to guide users through the results, especially when using multitude of categorical values
  - Ensemble: results from multiple algorithms need to enter a single data quality process pipeline
- In 2019-20 the MMSR team tried to systematically tackle the issues encountered by extending and modifying existing functions, generalising the solutions identified

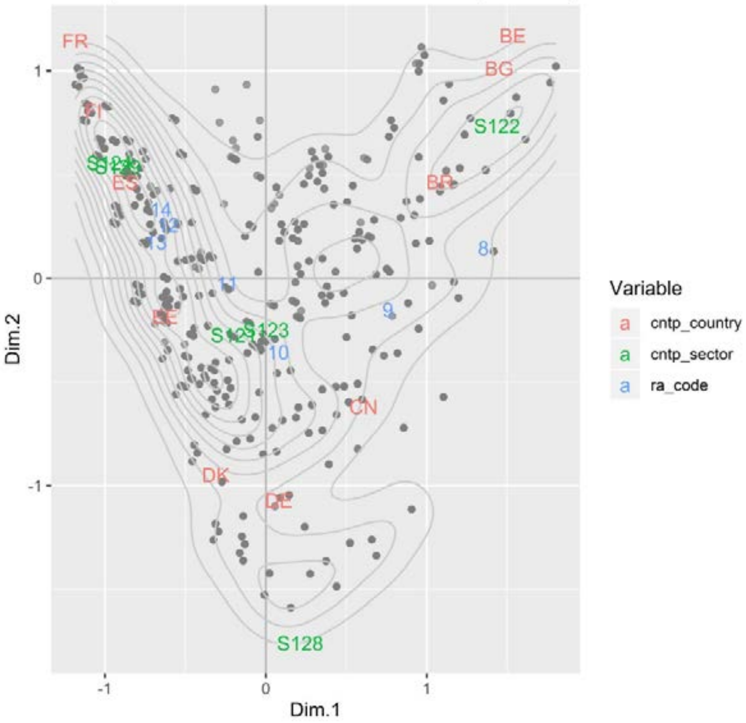
MMSR data quality management – an overview

- The Money Market Statistical Reporting (MMSR) is a daily data collection involving 47 banks, located in 10 different euro area countries
- Reporting agents report transaction-by-transaction data of their euro money market activity
- ~50,000 total transactional records received on a daily basis: ~30,000 Secured, ~15,000 Unsecured, ~5,000 FX Swap & Overnight Index Swap
- MMSR data are enriched with reference data (extra categorical values)
- 4 national central banks + ECB participate in data quality management
- Daily workflow implies limited budget of data quality inquiries
- Structured feedback: labelled anomalies datasets for training available

2

Anomaly detection algorithms used

Multiple correspondence analysis (MCA)



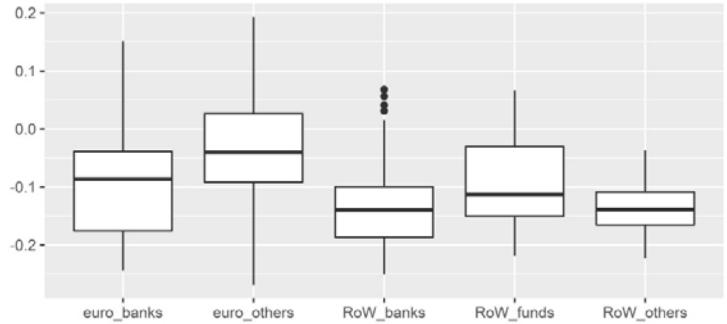
MCA plot of selected variables and RAs (~3k obs)

Note: Illustrative analysis performed using synthetic data

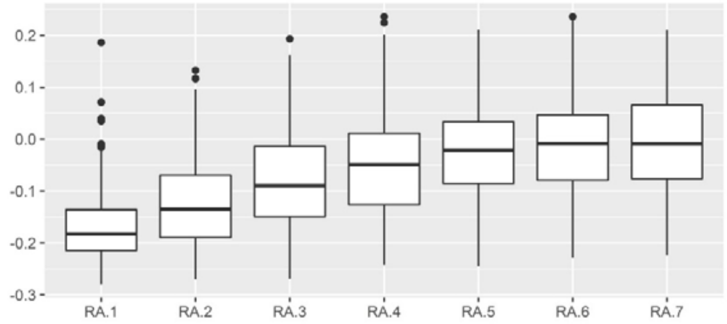
- A data transformation previous to the application of ML techniques
- Categorical variables raise problems for ML algorithms
- MCA is used to convert categorical variables into numerical values
- MCA exploits the “correlation” between features represented in different categorical variables
- The obtained numerical variables represent observations in a multidimensional space where frequently associated features appear clustered together

Anomaly detection – regression analysis

Uns. O/N borr. - spreads by cntp geo-sector



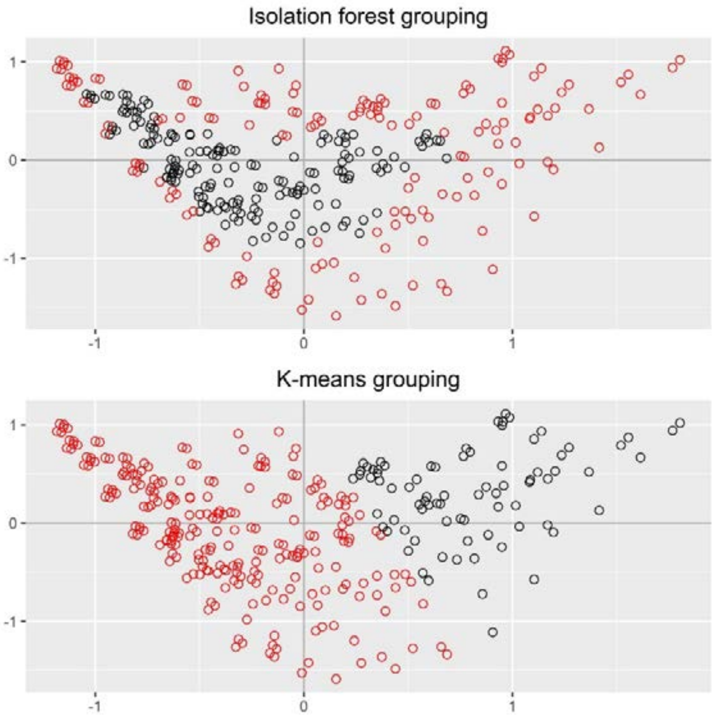
Uns. O/N borr. - spreads by RA name



Note: Illustrative analysis performed using synthetic data

- Model based: anomalies are defined as transactions far off the prediction of a model
- Model: $s_i = \alpha + \boldsymbol{g}_i'\boldsymbol{\beta} + \boldsymbol{m}_i'\boldsymbol{\gamma} + \delta \log(vol_i) + \varepsilon_i$
- Dependent variable: spread between deal rate and benchmark rate
- Explanatory variables: vectors of dummies for RA-geo-sector (*g*) and maturity (*m*), transactional nominal amount (*vol*)
- Model defined on the basis of descriptive evidence on typical trading patterns
- Estimation: weighted least squares
- Anomalies: transactions having the highest studentized residuals

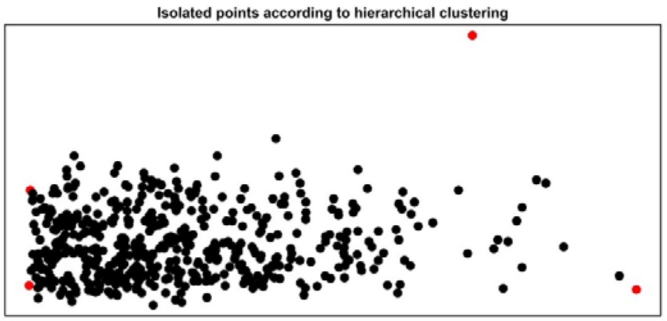
Anomaly detection – isolation forest

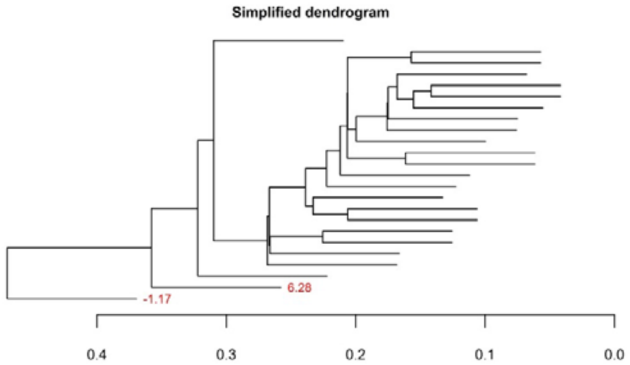


- Isolation forest only works with numerical variables
- It consists in a repeated random partition of the data until all data points in the sample are isolated
- Data points are considered anomalies when the number of partitions required for their isolation is small
- Advantages:
  - It has low linear time complexity and a small memory requirement (it samples)
  - Identifies both scattered and clustered anomalies
  - It is robust to “swamping” and “masking”

Note: Illustrative analysis performed using synthetic data

Anomaly detection – hierarchical clustering





Note: Illustrative analysis performed using synthetic data

- Hierarchical clustering identifies data points isolated and poorly connected to other data points
- Algorithm used: HDBSCAN - Hierarchical Density-Based Spatial Clustering of Applications with Noise
- Advantages:
  - Performance (limited complexity)
  - Parsimony in parameters (minimum cluster size is intuitive)
  - Robust to “chaining phenomenon” and other drawbacks of single-linkage
  - Association with GLOSH (Global-Local Outlier Score from Hierarchies)

Anomaly detection – XGBoost

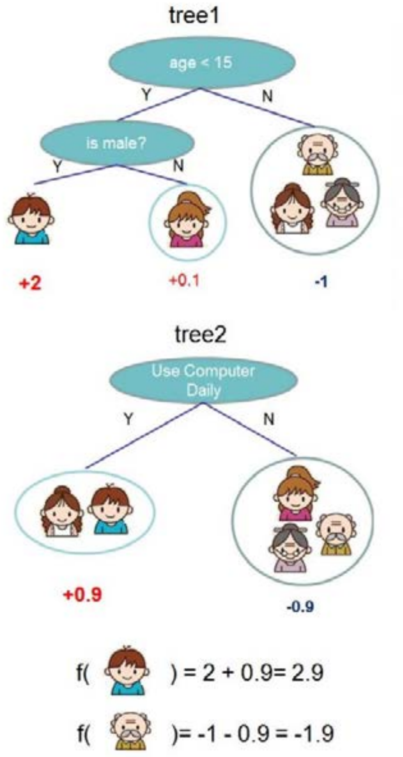
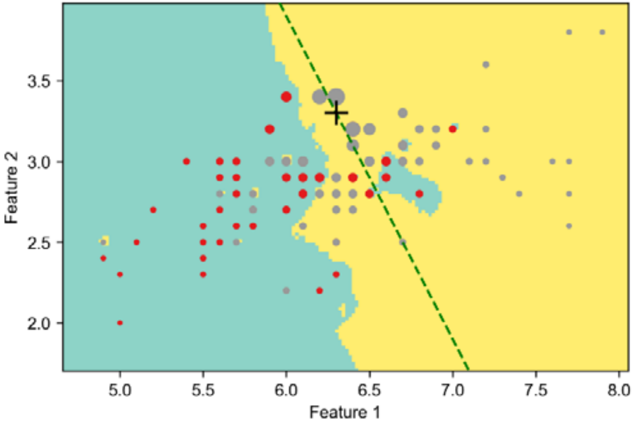


Figure from XGBoost: Chen Guestrin 2016, A Scalable Tree Boosting System. (Does the person like computer games?)

- In XGBoost weak predictors are employed to solve classification problems
- Anomalies are identified on the basis of a training on past data (supervised learning)
- Being a supervised learning algorithm, it requires a dataset of labelled anomalies to be trained with
- Advantages: award-winning algorithm, excelling in both efficiency and accuracy
- Success of the algorithm relies in the quality of the labelled anomalies dataset
- Necessity of structured and ordinate feedback from RAs for enquiries regarding outlying observations
- The verification workflow, integrated with the MMSR DQM aims at “automatizing” and simplifying the internal communication and the storage of information (including pre-defined set of feedback options)

Explaining detected anomalies: LIME algorithm

- LIME (Local Interpretable Model-agnostic Explanations) is an algorithm for the explanation of black-box algorithms results
- LIME is a model-agnostic method: it is equally applicable to every model.
- LIME provides approximated results: for this reason it is reasonably quick, but also somewhat volatile.



Graphical intuition of how LIME works

The black-box model f (unknown to LIME) is represented by the green/yellow background.

The bold black cross is the instance being explained.

LIME samples instances, gets predictions using f and weighs them by the proximity to the instance being explained (represented here by size of dots).

The dashed line is the learned explanation that is locally (but not globally) faithful.

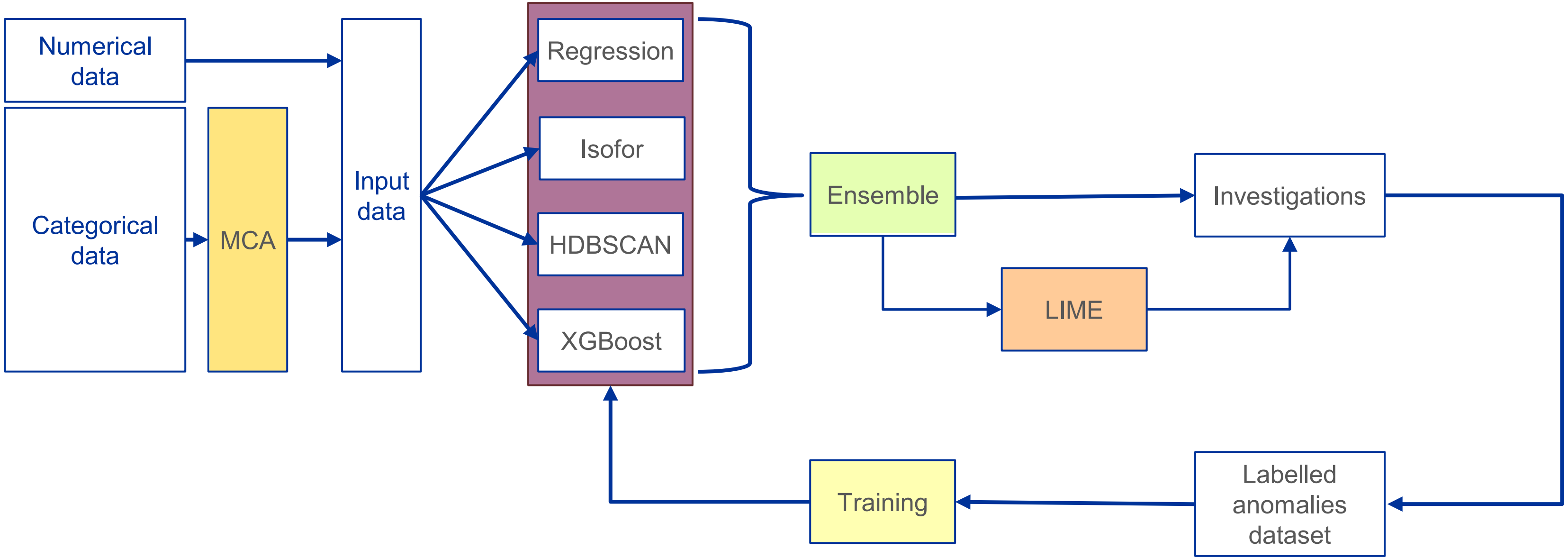
Note: Illustrative analysis performed using synthetic data

3

Workflow and pipeline plumbing interventions

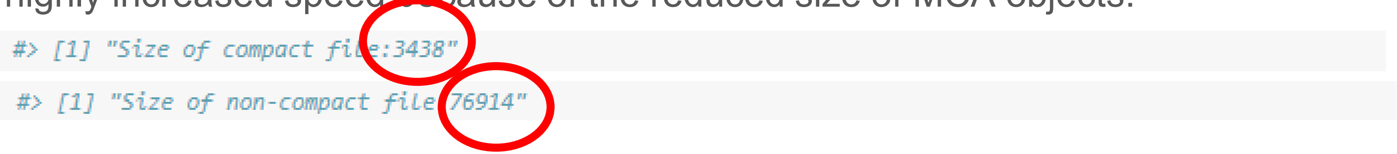
Workflow definition

- Workflow defined to accommodate feedback loop and ensemble of algorithms



Improving MCA performance

- Reduced weight of MCA objects by exploitation of *conversion formula*[(1)]:
  - Available MCA functions applied to *n* x *k* matrices produce MCA objects including elements having *n* observations
  - This increases extremely the size of these objects when MCA is applied to large datasets
  - MCA in our setting is used to convert *rows* into *row factors*, so there is no interest in column factors
  - Consequently, in the solution adopted all n-dimensional elements are dropped from the stored object
  - This speeds up the saving and the loading in memory of MCA objects
  - The prediction of row factors related to the daily serving input dataset can be then obtained with highly increased speed because of the reduced size of MCA objects:

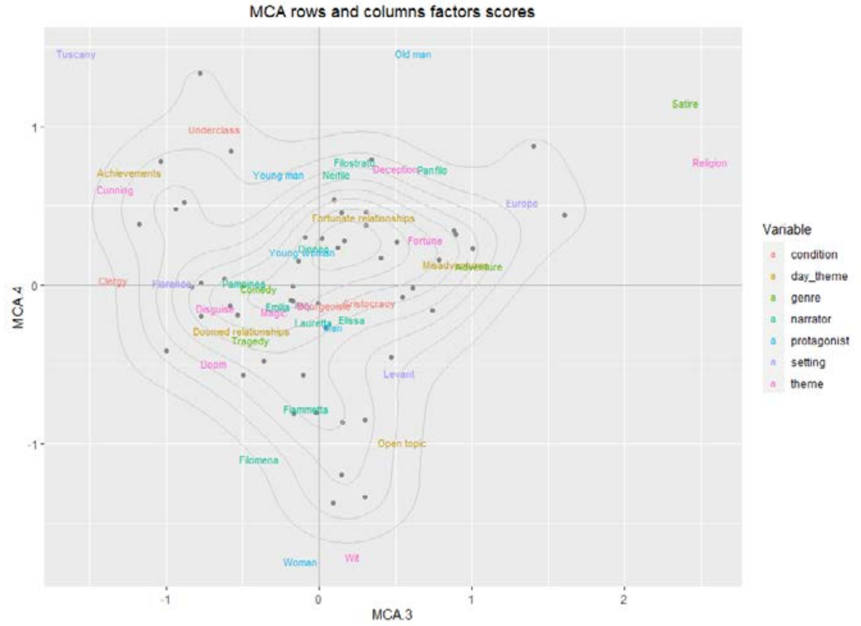


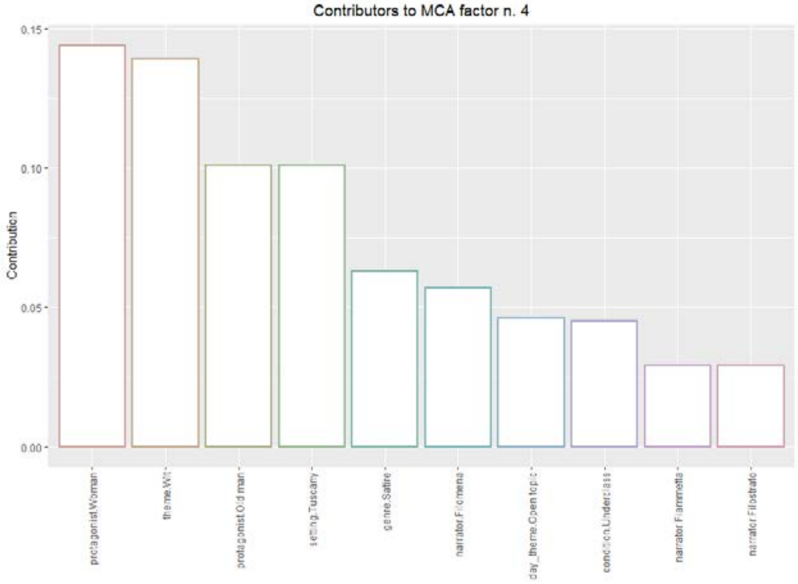
```
#> [1] "Size of compact file:3438"
#> [1] "Size of non-compact file:76914"
```

(1) See H. Abdi, J. Williams “Correspondence Analysis” in N. Salkind (Ed.) *Encyclopaedia of Research Design*

Improving MCA visualization

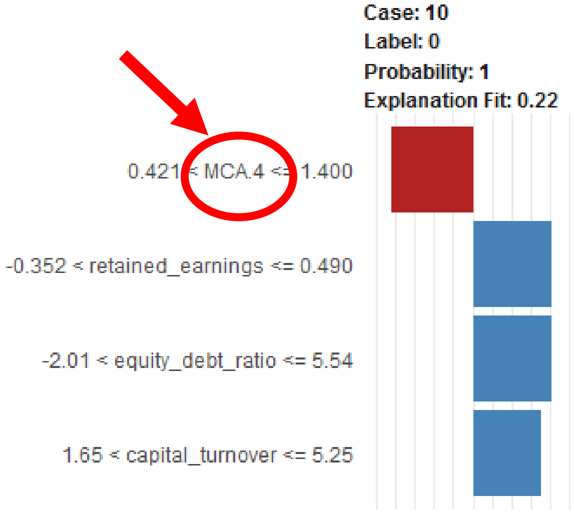
- Improved charting facilities aimed at better linking original variables and factors (focus on the contributors to the factors)



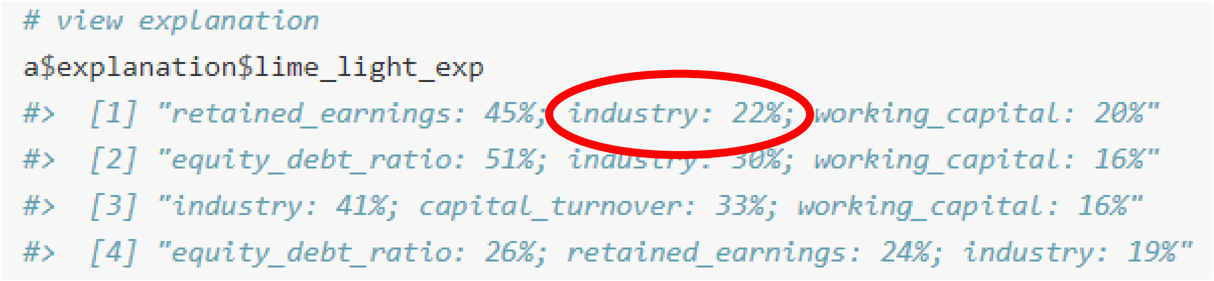


Improving interpretation of results

- LIME adopted for the explanation of results obtained via ML algorithms
- Improved readability of LIME results by means of:
  - Conversion of LIME results to textual explanations, to be easily integrated within the normal communication channels
  - Conversion of LIME results involving MCA factors to assessments related to the original categorical variables



Case: 10
Label: 0
Probability: 1
Explanation Fit: 0.22



```
# view explanation
a$explanation$lime_light_exp
#>  [1] "retained_earnings: 45%; industry: 22%; working_capital: 20%"
#>  [2] "equity_debt_ratio: 51%; industry: 30%; working_capital: 16%"
#>  [3] "industry: 41%; capital_turnover: 33%; working_capital: 16%"
#>  [4] "equity_debt_ratio: 26%; retained_earnings: 24%; industry: 19%"
```

Improving use of ensemble of algorithms

- An ensemble of algorithms is employed
- A comparison among heterogeneous approach is required to make sense of the results (scores)
- The ranking of the scores obtained from heterogeneous algorithms is obtained as the weighted average of the provided scores.
- Comparability among algorithms is pursued through robust standardisation
- Feasibility is ensured by means of a cap on the overall workload
- Relevance is tested against labelled anomalies dataset

Thank you for your attention

For any question or comment please feel free to contact us:

matteo.accornero@ecb.europa.eu

gianluca.boscariol@ecb.europa.eu

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

A novel machine learning-based validation workflow for financial market time series [1]

Magdalena Erdem and Taejin Park,
Bank for International Settlements

[1] This presentation 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 central banks and other institutions represented at the event.

A novel machine learning-based validation workflow for financial market time series[1]

Magdalena Erdem and Taejin Park[2]

Abstract

The size and complexity of data managed by central banks have been increasing providing them with opportunities to make evidence-based policy decisions. As a result, data validation has become a more challenging task. In this paper, we propose a highly automated validation workflow that outperforms traditional approaches and is suitable for a large volume of financial market time series, based on machine learning algorithms. Using some real-life examples, we illustrate how machine learning (ML) algorithms can help address key challenges, such as understanding the context of various financial instruments and dynamically coping with constantly evolving market environments.

Keywords: machine learning, data science, data validation; financial market time series

JEL classification: C55, C80, G17, O33

[1] The views expressed in this article are those of the authors and do not necessarily reflect those of the Bank for International Settlements.

[2] Respectively, Head of Departmental Research Support, BIS (Magdalena.Erdem@bis.org); and Head of Financial Markets and Research Support, BIS (Taejin.Park@bis.org)

Introduction

Data validation is a key accountability of statistics teams in central banks. With more data available for central banks to make evidence-based policy decisions, the amount of data managed by central banks have been increasing accordingly. As a result, data validation has become a more challenging task, especially with large volumes of data, such as financial market series that are usually available at high frequencies – daily or intraday. Validating such large financial market data can requires significant resources so it is often outsourced to data providers which requires trust in the data they provide. Accepting third-party data without due verification could lead to high operational risks.

While the need to guarantee high quality data for policy makers is being given more importance, there has been little research focused to central banks' (or financial authorities') toolkits for checking financial market time series. Eurostat publishes a comprehensive data validation framework focusing on official statistical agencies (Eurostat, 2018). Among central banks, the European Central Bank (ECB) provides a framework, mainly covering governance and data quality dimensions and metrics, and an additional list of data quality checks on supervisory reporting data (ECB; Hogan, 2017). The US Board of Governors of the Federal Reserve System (Federal Reserve Board) has developed Information Quality Guidelines for data quality, objectivity, utility and integrity as required by the US Office of Management and Budget. The Bank of England offers its data quality framework to enable data users to be informed about the quality of the data they use (Bank of England, 2014). Such frameworks can be very useful to understand the various aspects of data validation principles. However, complementing them with more detailed practices and examples can further help statisticians address their day-to-day challenges.

Meanwhile, in the machine learning community, research has been conducted in detecting anomalies and outliers in financial market data (Au Yeung et al, 2020; Ahmed et al, 2016; Golmohammadi and Zaiane, 2015; Ferdousi and Maeda, 2006). Such research aims mostly at detecting fraud or informing investment decisions. Whilst this can provide useful insights for developing data validation techniques specific to financial market data, it is not directly applicable to *error* detection or confirming *true outliers* (ie seemingly suspicious but actually correct data points).

To fill the research gap in practical guidance for validating large volumes of high-frequency financial market series, this paper proposes a solid data validation workflow that would allow for full automation requiring low maintenance costs.

Recent developments in data science provide great opportunities for central banks. Among many, machine learning techniques have been developing quickly in recent years. Thanks to its popularity, machine learning is also now much more accessible. There are free software packages for machine learning analysis (eg Python and R), active online forums (eg Stack Overflow) and open source off-the-shelf code libraries (eg Scikit-learn, TensorFlow and Keras). In addition, advancing computing capacity enables statisticians and data scientists to solve complex algorithms. In particular, big data platforms, GPU units and cloud computing can significantly boost efficiency and broaden the scope of machine learning analysis.

Leveraging on these opportunities, this paper proposes a highly *automated* validation workflow that *outperforms* traditional approaches and is suitable for a large volume of *financial market time series*. The main objective of our analysis is to develop an end-to-end workflow for data validation, with examples of step-by-step machine learning applications. We do not intend to propose any single specific machine learning model as individual circumstances will predominantly determine model selection.

Challenges with traditional validation approaches

To better understand the requirements of a good validation tool, we first reviewed the most commonly used traditional validation methods – graphical method, conditional controls, threshold-based warnings and cross-referencing. These traditional methods enable some degree of automation but still require frequent human intervention, making them unsuitable for large scale validation processes (Table 1). Validation of even a small number of financial market series, taking account of the context of various financial instruments in different market segments, can be time-consuming. Furthermore, financial market environments are continuously evolving and often entail structural changes in the market due to central bank policy actions (eg policy rate changes or quantitative easing), financial conditions and new or outdated instruments. Keeping up with such changes in diverse financial market segments can be very labour-intensive.

Overview of common traditional validation approaches

Table 1

	Description	Limitations
Graphical method	Visual inspection of time series graphs to detect any anomalies	Time consuming; vulnerable to oversight; difficult to validate plausible but erroneous data points or true outliers
Conditional controls	Implementing pre-defined "If-Then" controls	Requires a good understanding of market contexts; setting appropriate parameters for controls/thresholds to many heterogeneous series is challenging; becomes ineffective in case of any structural changes or turmoil in the market
Threshold-based warnings	Setting certain thresholds to give warnings, for example, based on percentage changes or z-scores. Usually used together with conditional controls.	
Cross-referencing	Cross checking with the same series from another source or other related series to see there's any divergence in their relationship	Requires a good understanding of market contexts to determine appropriate reference series to validate the target series; alternative source might not be available; especially challenging when validating many series in different market segments.

Source: Authors' elaboration.

Graph 1 illustrates some of the challenges in validating financial market time series in practice. The daily time series of the preliminary euro short-term rate (Pre-€STR) exhibits various anomalies at a glance (left panel). If the series is reviewed by any graphical methods and/or threshold-based controls, it is likely giving warnings for the spike each quarter and the sudden break in September 2019. However, the real error in this series is the repeating values at the end of the series when the instrument is discontinued on the official launching of €STR on 2 Oct 2019. Since the legacy preliminary series was treated as a separate instrument by its data provider, instead of being replaced by the official one, the last quoted value kept appearing in the following days. That repeated values is not unusual in some illiquid financial markets can make a statistician consider the data plausible, potentially leading to a false positive error (ie Type I error).

A similar pattern is observed in the Shanghai stock exchange (SSE) A share index where a value is repeated for eight straight days in mid-February 2021 (middle panel). As opposed to the previous case, however, the long repetition is what happens in the market during Chinese (Lunar) New Year holidays, which fall on different dates in January or February of the Gregorian calendar each year.

These two examples show similar data patterns in appearance but require different actions, indicating that data validation necessitates a good understanding of the market context. Those who have sufficient knowledge about euro money markets would know that such quarterly spikes and a break in series due to a policy rate change are common. Similarly, those who are familiar with Chinese markets would not regard the long repetition as an error. However, keeping up with many financial instruments in diverse market segments is challenging.

The right panel of Graph 1 shows the US money market rates in 2019. Immediately noticeable is the striking outlier of the Secured Overnight Financing Rate (SOFR) in Q3 2019. Again, any outlier detection controls in place will likely give warnings given the serious magnitude of change in a day (ie Type II error, (false negative error)). Even for an expert in money market instruments, such a spike in a secured interbank overnight interest rate so much above other unsecured rates would be perceived as an extremely rare event. To validate this outlier, one can compare it with other US secured interbank rates that usually follow similar trends, such as the Broad General Collateral Rate (BGCR) or the Tri-Party General Collateral Rate (TGCR). Since the other secured rates appear to have similar shocks on the same date, the outlier can eventually be confirmed as a true value.

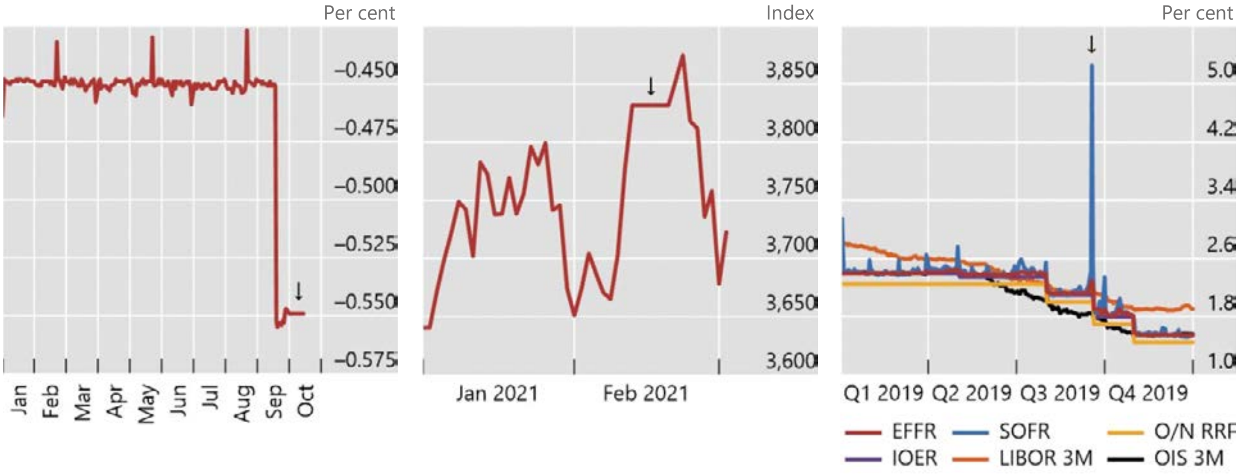
Traditional validation methods

Graph 1

Preliminary euro short-term rate in 2019

Shanghai stock exchange (SSE) A share

US money market rates



Sources: Bloomberg; Refinitiv.

One might argue that such false negative errors are still useful to maintain alertness and in fact, receiving false negative errors for suspicious data points is better than missing any true errors. However, in the long term frequent false negative errors can make statisticians become accustomed to the warning messages and thus ignore them without investigation. Reducing false negative errors is also therefore highly important to manage operational risks. To summarise, a good data validation system should set a precise threshold that is free from both false positive and false negative errors.

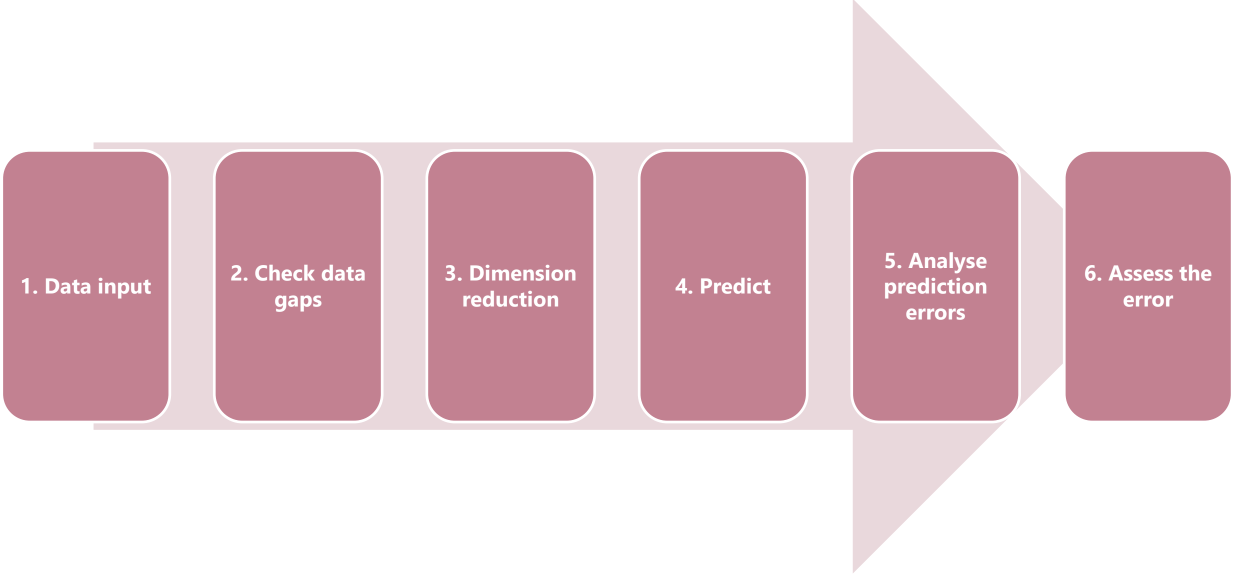
Proposed data validation workflow using machine learning models

The challenges illustrated in the previous section highlight the main requirements of a good data validation model for financial market series. Without a good understanding of the context of each instrument, those common traditional methods can produce frequent Type I or II errors. An effective data validation model should imitate the behaviours of a subject matter expert who is familiar with the unique characteristics of the instrument and keeps up with recent market development. Therefore, analysing the underlying processes of a human expert validating a financial market series can provide insight for developing an optimal validation model.

A human expert can access all readily available information and identify only the relevant information for a specific context. They can then make a best judgement to assess if a given value is within a reasonable range according to the key information collected. In statistical terms, the process can be translated into: input data collection; filter useful explanatory data (or reduce dimensions); and then assess the target value if it is within the range predicted by the explanatory data. Based on this intuitive approach to the validation, we propose a model workflow of financial market data validation (Graph 2). We predict any data point to validate as if it is invisible to us for the date. The prediction fully makes use of many other financial market time series including the date where the target series is assumed invisible. This workflow uses machine learning models in various steps to minimise human intervention. In other words, the process is highly automated for scalability and dynamic adaptation to evolving market environments. Once implemented, the workflow can continuously run every day to detect anomalies. In the remainder of this section, we will provide more details about each of the six steps with example machine learning applications for the anomaly cases shown in the previous section.

Stylised workflow of data validation process using machine learning models

Graph 2



Source: Authors' elaboration.

1. Data input

For the purpose of the illustration, we use about 3,000 daily incoming financial market time series, covering a variety of market segments as input data. The aim is to develop a workflow that is suitable for validating any or all of them. Additionally, the 3,000 series can be used as input to validate any specific series within the sample.

These input data have the typical financial market time series characteristics that can pose challenges to effective data validation. First, the size of dataset is so big that manual validation processes can't be applied. Second, as previously illustrated, anomalies are not always easily detectable when analysing them in isolation. Third, they show frequent structural changes within a data series, which makes it difficult to fit any specific model that can reflect such changes dynamically. Lastly, the dataset covers diverse market segments (eg equity, interest rates, FX, credit, commodities, etc) for around 70 geographical markets, making it resource intensive to track and monitor all market developments.

However, having such a large financial market dataset can also provide an important opportunity for data validation. A financial market series usually has strong explanatory series. For example, there can be several equity indices that are tracking the same market. Yields of instruments with similar maturities in the same market show similar trends. Exchange rates that are pegged to the same currency also move together. This is a valuable feature of financial market time series that can enable precise predictions.

2. Checking data gaps

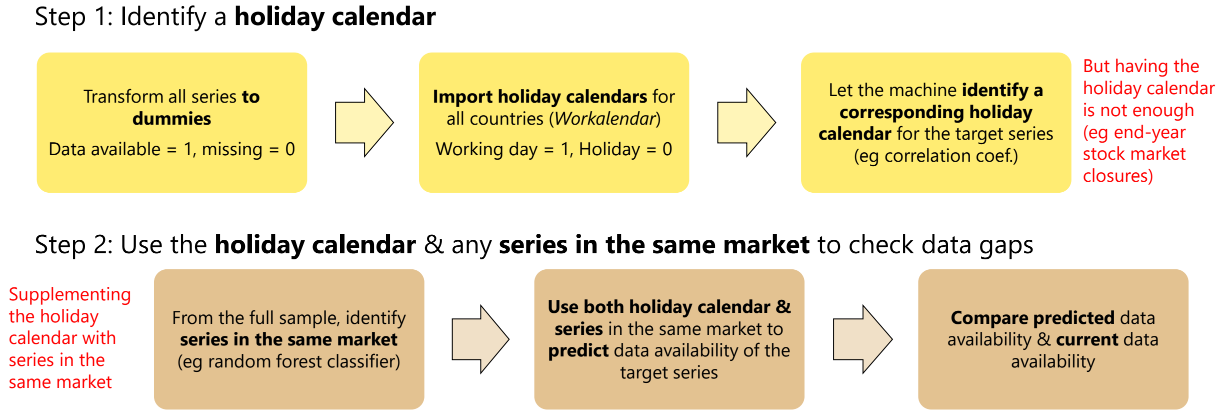
Data input should be followed with a check of data availability in the target series for a recent date (or any other date of one's interest). Missing data in the financial market series is usually due to market closure. Most financial markets follow the holiday calendar of the country in operation. Therefore, detecting any erroneously missing data can start with cross-checking the data with its corresponding holiday calendar. However, checking against holiday calendars alone is insufficient because of exceptions. For example, some FX markets are open every day including weekends and holidays. Some stock exchanges are exceptionally closed on a few non-holidays. For more accurate validation, supplementing holiday calendars with additional information is essential. Financial instruments traded in the same exchange usually follow the same opening and closing schedule. Hence data availability of a financial market series can also be assessed by cross-checking with another financial market series traded in the same market (eg Nasdaq & S&P500). If two financial market series historically show similar availability patterns, data availability of one series can be predicted by reference to the other series. Based on the two datasets, holiday calendars and series in the same market, this paper proposes a two step approach for checking data gaps in the target series.

The first step automatically identifies a holiday calendar for the target time series. As summarised in Graph 3, all input series first need to be re-coded as binomial variables to only indicate data availability. Holiday calendars are imported also as binomial variables. The holiday calendars for most countries are retrieved from a Python module (*workcalenda*[3]). Then, a simple algorithm (eg correlation coefficients) can identify the most relevant holiday calendar for the target series.

The second step is to identify any series in the same market. From the 3,000 financial market series in binomial terms, a machine learning algorithm (eg random forest classifier) can identify any financial market series that are likely traded in the same market as the target series. Once they are identified, together with the holiday calendar, they can collectively be used to predict today's data availability of the target series. For the prediction, a similar algorithm can be applied as in the identification of the series in the same market. The predicted data availability then can be used to evaluate the recorded data availability.

Two step approaches to check data gaps

Graph 3


Step 1: Identify a **holiday calendar**
Transform all series **to dummies**
Data available = 1, missing = 0
Import holiday calendars for all countries (*Workalendar*)
Working day = 1, Holiday = 0
Let the machine **identify a corresponding holiday calendar** for the target series (eg correlation coef.)
But having the holiday calendar is not enough (eg end-year stock market closures)
Step 2: Use the **holiday calendar** & any **series in the same market** to check data gaps
Supplementing the holiday calendar with series in the same market
From the full sample, identify **series in the same market** (eg random forest classifier)
Use both holiday calendar & series in the same market to **predict** data availability of the target series
Compare predicted data availability & **current** data availability

Source: Authors' elaboration.

[3] Workalendar Maintainers (2021), https://github.com/workalendar/workalendar

3. Dimension reduction

Dimension reduction is considered to be a common step in machine learning analysis to better fit a model to improve prediction power and to estimate the model more efficiently. Since our analysis used around 3,000 input variables, filtering irrelevant variables and focusing only on a small set of series that best explain the target series is imperative.

In this paper, dimension reduction is done in two steps. First, largely irrelevant variables are filtered following a simple traditional algorithm, that is, correlation coefficients. After this filtering, only dozens of series that are somewhat related to the target series can be taken forward for further dimension reduction (or feature selection). In the second step, we apply a random forest regression model to leave few key features that will become explanatory variables for a prediction model in the next step of the validation workflow.

Graph 4 shows results of an application of the dimension reduction processes to the challenging cases discussed at the beginning of the paper. For the SSE A share index, the model returned the SSE composite index as a dominant feature to explain the target series (left panel). The SSE composite index consists of both A shares and B shares that are traded in the SSE. Since market capitalisation of B shares takes only 0.2%[4] of that of A shares, the composite index should move together with the SSE A share index in a highly synchronised way. Therefore, one can precisely predict movement of the SSE A share index if the SSE composite index is known. For the US SOFR, the most important features turned out to be other US secured overnight money market rates, namely the TGCR and the BGCR, followed by unsecured overnight money market rates (right panel). For both the Chinese and US cases, it is interesting to note that the model automatically learns market contexts from the data and returns useful series. If a subject matter expert manually conducted the same exercise, the results would be very similar to what the model selected.

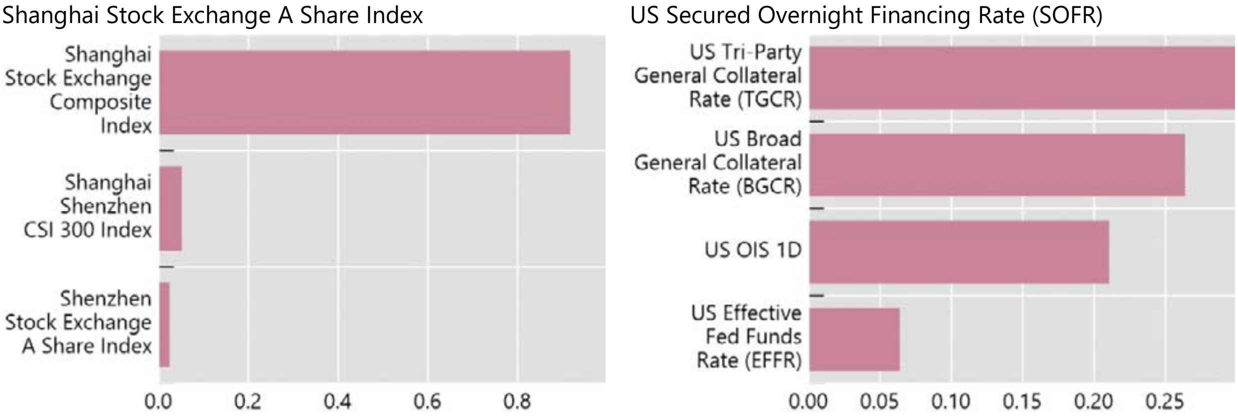
The number of features to be used for prediction can be determined based on the prediction model specifications and specific use cases. A minimum Gini importance cut-off value can be applied to individual series or to top-N series combined. Another approach is to develop any feature selection algorithm, such as backward/forward/recursive eliminations or exhaustive selection. The algorithm tries to find the best combination of features that can maximise the prediction performance of the exact model that will be used for prediction. In this approach, computing capacity is an important consideration.

[4] As of 11 October 2021.

Key features that can best explain the target series

Gini importance

Graph 4



Sources: Bloomberg; Refinitiv; authors' calculations.

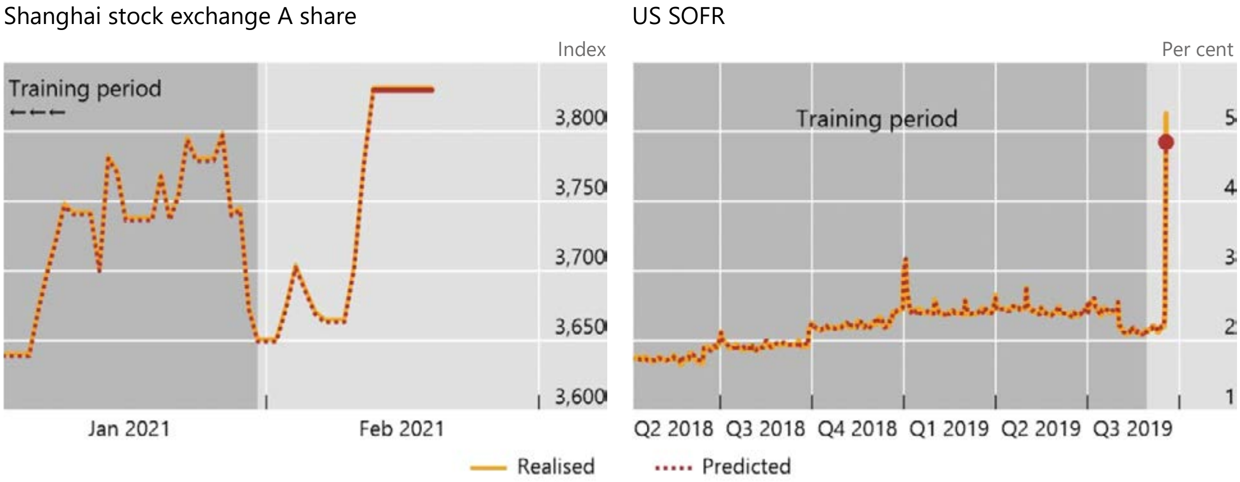
4. Prediction

Using the small set of features selected above, we can fit a machine learning model to predict a value for today or any date of interest for the target series as if the value is unknown. For model selection, we opted for a recurrent neural network (RNN), a deep learning model suitable for sequential data such as time series. More specifically, we used the Long Short-Term Memory (LSTM) model to capture long-term contexts without vanishing gradients and exploding gradients problems (Hochreiter and Schmidhuber, 1997). Once again, the main purpose of this paper is not to propose any specific model that can fit all different cases but rather to illustrate the proposed workflow.

Based on the two anomaly examples of financial market time series – the SSE A share index and US SOFR – we fit the LSTM model to predict values for the suspicious data points that were discussed in the previous section. The out-of-sample prediction results appear to be precise in both cases (Graph 5). Both predictions are based on the same model specifications, except for input data. This is one of the main advantages of using a machine learning algorithm as the machine itself can find a model to best fit input data with minimum human intervention.

Comparison of predicted and realised values

Graph 5


Shanghai stock exchange A share

US SOFR

The prediction is based on the long short-term memory (LSTM) model.

Sources: Bloomberg; Refinitiv; authors' calculations.

5. Analyse prediction errors

Once a predicted value is available, the realised value in question needs to be compared to the predicted value. The difference between the predicted and the realised value (ie prediction error) would indicate how much the realised value in question deviates from a reasonable expectation. When the difference is *significantly large* it would be worth investigating further.

An important question is then how to determine whether a prediction error is large enough to suspect erroneous data. One possibility is to set certain thresholds to evaluate a prediction error based on percentage deviations or standard deviations. However, if a series is highly volatile by nature or difficult to predict due to limited feature availability, such a static threshold-based approach can signal frequent Type I or II errors. Therefore, this paper applies an unsupervised machine learning algorithm to decide whether a prediction error is acceptable or not, based on historical patterns of prediction errors in the target series.

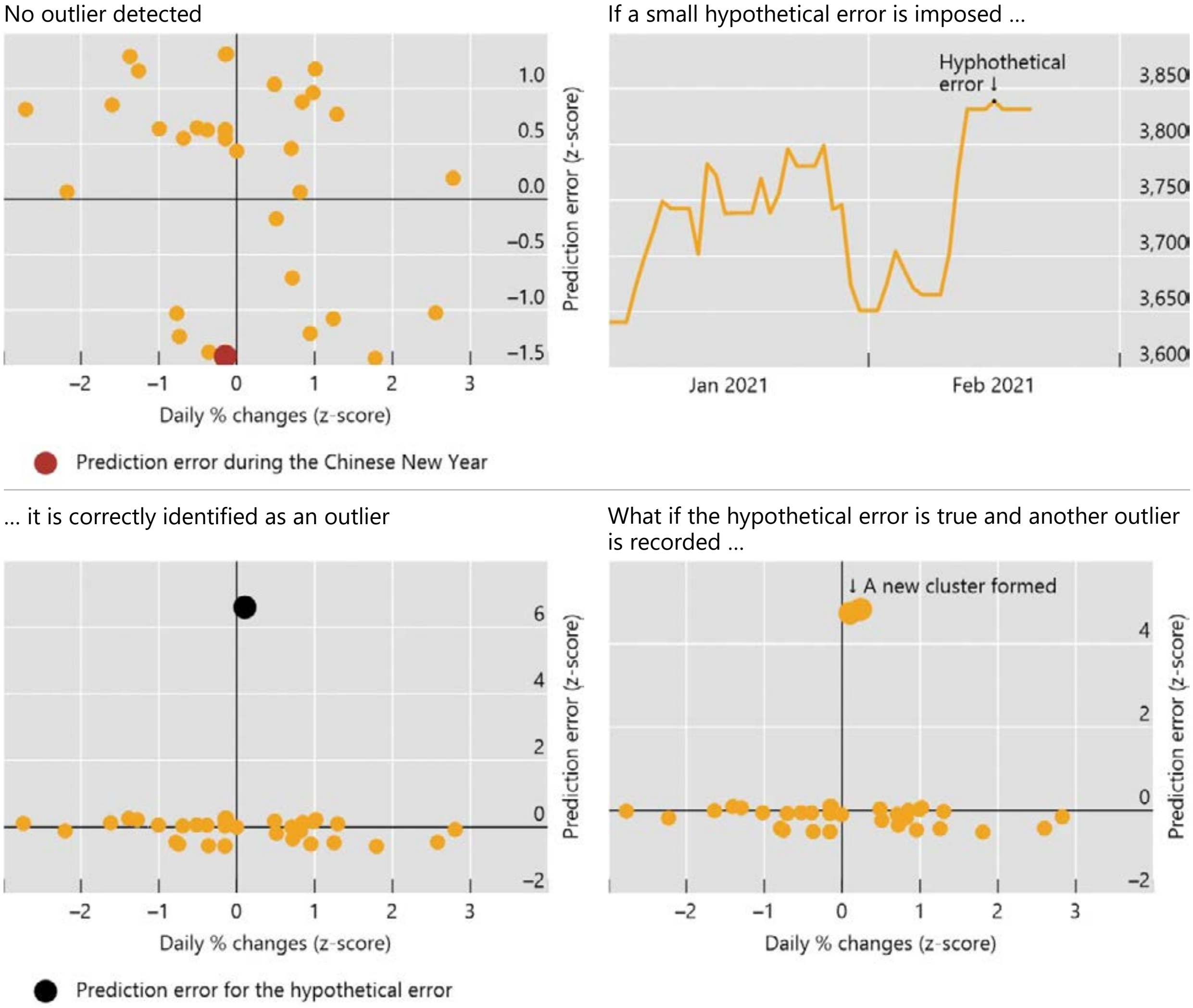
6. Assess the errors

To assess the prediction errors, we applied an unsupervised clustering algorithm, Density-Based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN requires minimum input parameters and discovers clusters with arbitrary shape, which makes it suitable for a large dataset (Ester et al, 1996).

When DBSCAN is applied to assess prediction errors for the SSE A share index case, the repeating values during the Chinese New Year are identified as non-outliers (Graph 6, top left). To check the robustness of the algorithm, we intentionally recorded a small hypothetical error equal to the average daily percentage change of the original series (top right). Since the time series data were precisely predicted in the previous section, such a small error that looks insignificant in the graph is identified as an obvious error in this algorithm (bottom left). To highlight how this algorithm can cope with a structural change in the market, we assume that the small error turns out to be true and a similar outlier is recorded again on the next date. In this case, the second outlier is not considered as an error anymore because the minimum number of samples to form a new cluster is set to 2 (bottom right). This implies that the model returns a false negative error message when it first encounters a significant structural change in the market, but it can correctly validate any similar types of outlier for the coming days. This is an advantage of machine learning models as they can dynamically learn contexts from data and reflect such structural changes in the market.

Outlier detection based on an unsupervised machine learning algorithm

Graph 6



No outlier detected

If a small hypothetical error is imposed ...

... it is correctly identified as an outlier

What if the hypothetical error is true and another outlier is recorded ...

The analysis is based on density-based spatial clustering of applications with noise (DBSCAN). The maximum distance between two samples for one to be considered as in the neighbourhood of the other is set to 2 Euclidian distance. The minimum number of samples to form a new cluster is set to 2.

Sources: Bloomberg; Refinitiv; authors' calculations.

Conclusion

While financial market time series data are key inputs to important policy decisions by central banks, little research focuses specifically on data validation processes for them. This paper first reviewed common practices used for time series data validation and their limitations in this emerging data-intensive environment. It then proposed an end-to-end workflow of daily data validation routines to overcome key challenges in ensuring high quality for large financial market time series datasets. While describing each step, we illustrated how machine learning algorithms can help address the key challenges, such as understanding the context of many financial instruments and dynamically coping with constantly evolving market environments. During our analysis, we intentionally focused on a few carefully selected examples to best illustrate key challenges and solutions in each step from central bank practitioners' perspectives.

We would like to reiterate that machine learning techniques are now more accessible than ever, even to non-experts. This provides a great opportunity for data validation work. At the same time, the abundance of models, code libraries and references available can create additional challenges, especially without a clear overview of one's unique business requirements. For this reason, we did not recommend which machine learning model would work best in each situation as such discussion would be less meaningful without a more detailed context of characteristics and materiality of datasets, infrastructure, business environments, etc.

Future work and investigation should focus precisely on describing the suggested methods and their application to specific cases. We hope this paper will provide a stimulating basis for emerging research on this topic to the benefit of the central banking community. The ultimate aim is to build knowledge blocks for more efficient quality assurance of data.

References

Au Yeung, J.F.K., Wei, Zk., Chan, K.Y. et al (2020): “Jump detection in financial time series using machine learning algorithms”, *Soft Computation*, no 24, pp 1789–1801

Bank of England Statistics and Regulatory Data Division (2014): *Data Quality Framework*

Board of Governors of the Federal Reserve System: *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Federal Reserve Board*, https://www.federalreserve.gov/iq_guidelines.htm

Chollet, F. et al (2015): “Keras”, https://github.com/fchollet/keras

European Central Bank: *Additional supervisory data quality checks*, https://www.bankingsupervision.europa.eu/banking/approach/dataqualitychecks/html/index.en.html

Eurostat (2018), *Methodology for data validation 2.0*

K. Golmohammadi and O. R. Zaiane (2015): "Time series contextual anomaly detection for detecting market manipulation in stock market," *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp 1–10

Mohiuddin Ahmed, Abdun Naser Mahmood, Md. Rafiqul Islam (2016): “A survey of anomaly detection techniques in financial domain”, *Future Generation Computer Systems*, Volume 55, pp 278–288

Pedregosa et al (2011): “Scikit-learn: Machine Learning in Python”, *JMLR,* no 12, pp. 2825–2830

P Hogan (2017): “ECB Supervisory Data Quality Framework, Tools and Products”, presented at the Supervisory Reporting Conference, ttps://www.bankingsupervision.europa.eu/press/conferences/shared/pdf/sup_rep_conf/2017/Data_quality_framework_tools_and_products.pdf

Workalendar Maintainers (2021): https://github.com/workalendar/workalendar

Z. Ferdousi and A. Maeda (2006): "Unsupervised Outlier Detection in Time Series Data", *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pp x121–x121

Irving Fisher Committee on
Central Bank Statistics

BIS

Deep learning as a novel validation tool for financial market time series

IFC and Bank of Italy Workshop on "Data science in central banking", part 1, 19-22 Oct 2021

Taejin Park, Head of Financial Markets and Research Support (FMRS), BIS (presenter)

Magdalena Erdem, Head of Departmental Research Support (DRS), BIS

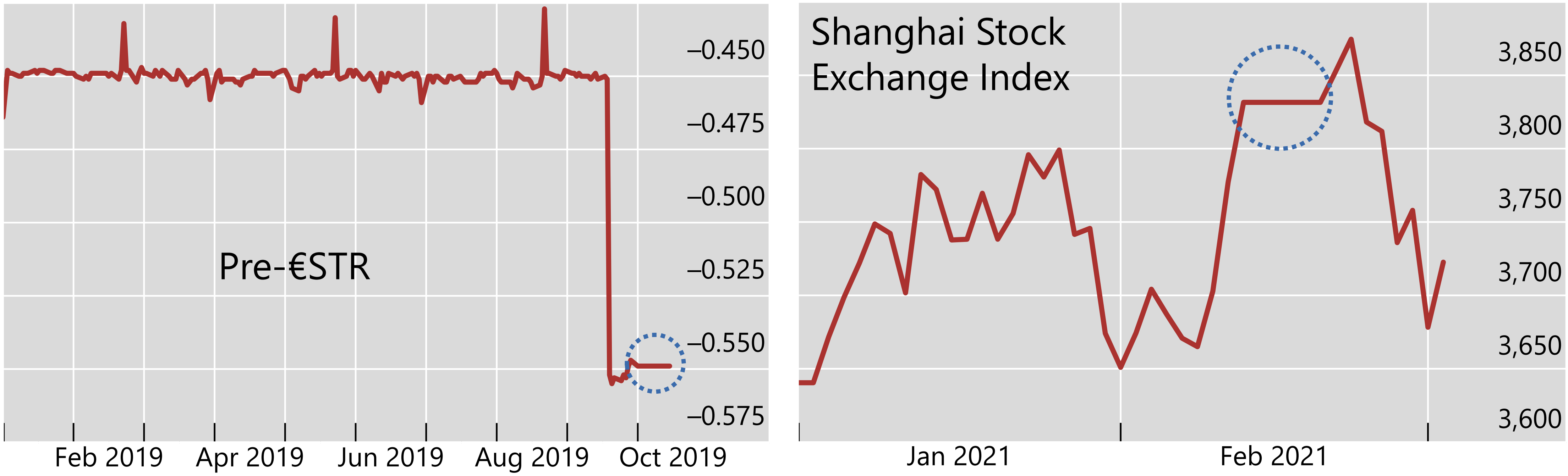
Introduction

- Data validation has been becoming **challenging**

- With machine learning & enhanced computing capacity, we propose:
  - A highly **automated** validation **work flow**
  - that **outperforms** traditional approaches
  - suitable for **a large volume of financial market** data

Challenges with the traditional approaches

- Are there any issues with these series?

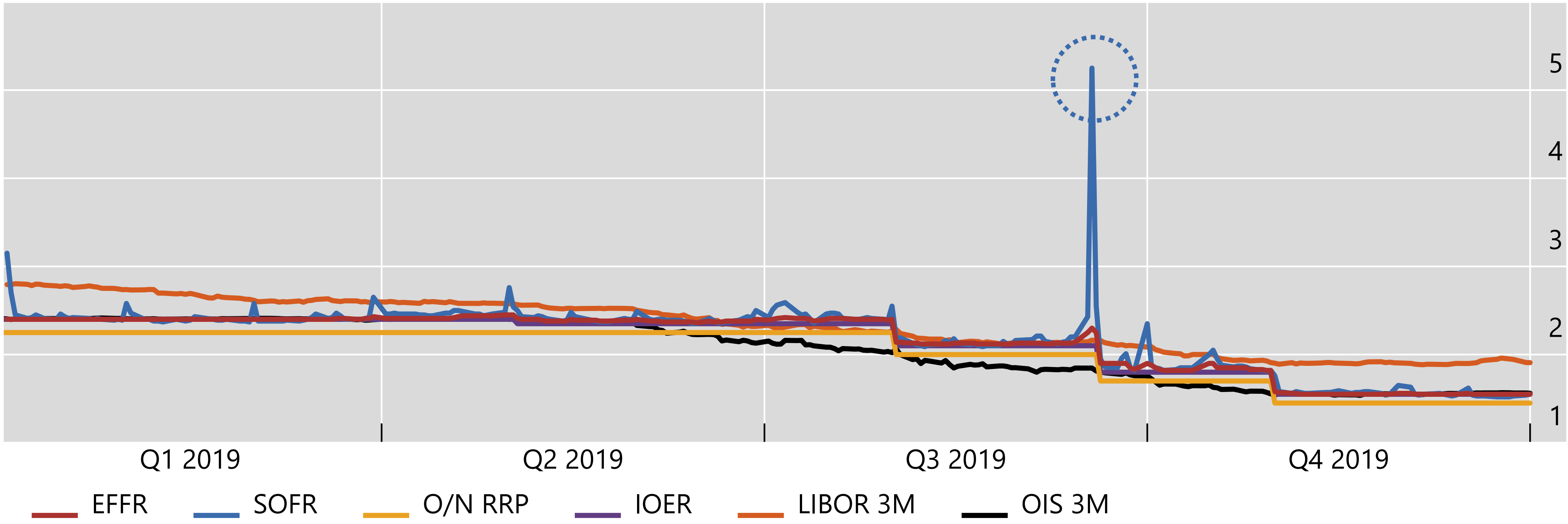


Sources: ECB; Refinitiv.

<u>Plausible but wrong data</u> (eg missing values, repeating values, ticker changes) are difficult to detect (False positive error (ie Type I error))

Challenges with the traditional approaches

- Is there any issue with these series?

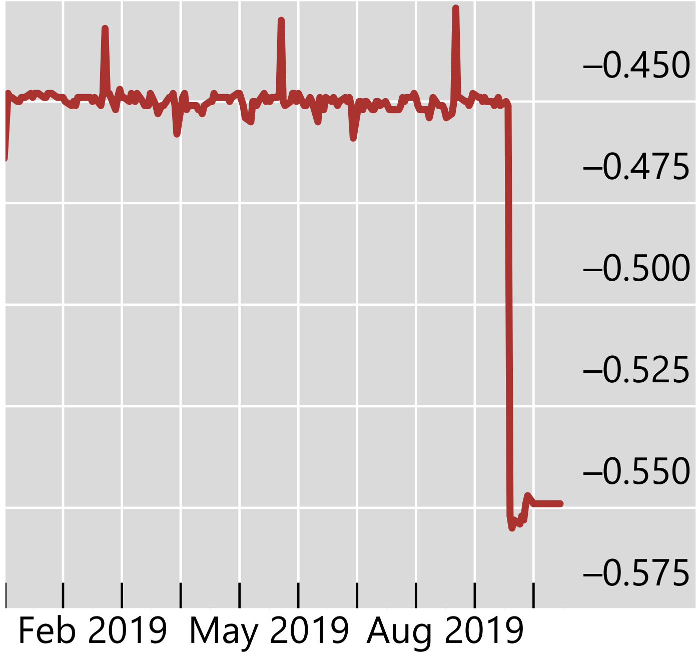


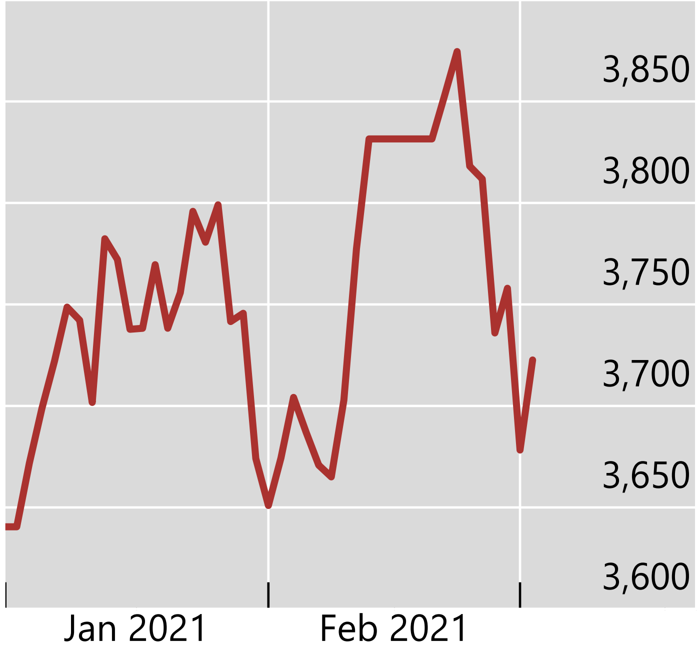
Sources: Bloomberg; Refinitiv.

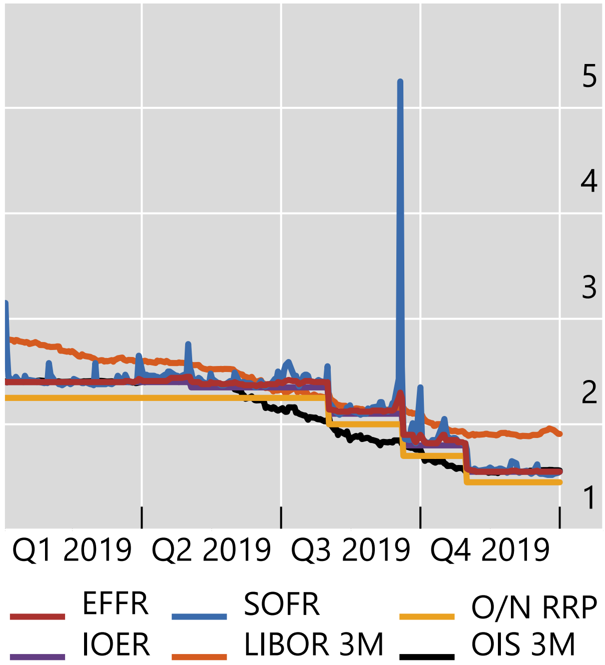
- Suspicious but correct data are also difficult to detect (False negative error (ie Type II error))
- What can be done to minimise such false alarms?

Overview of common traditional data validation techniques

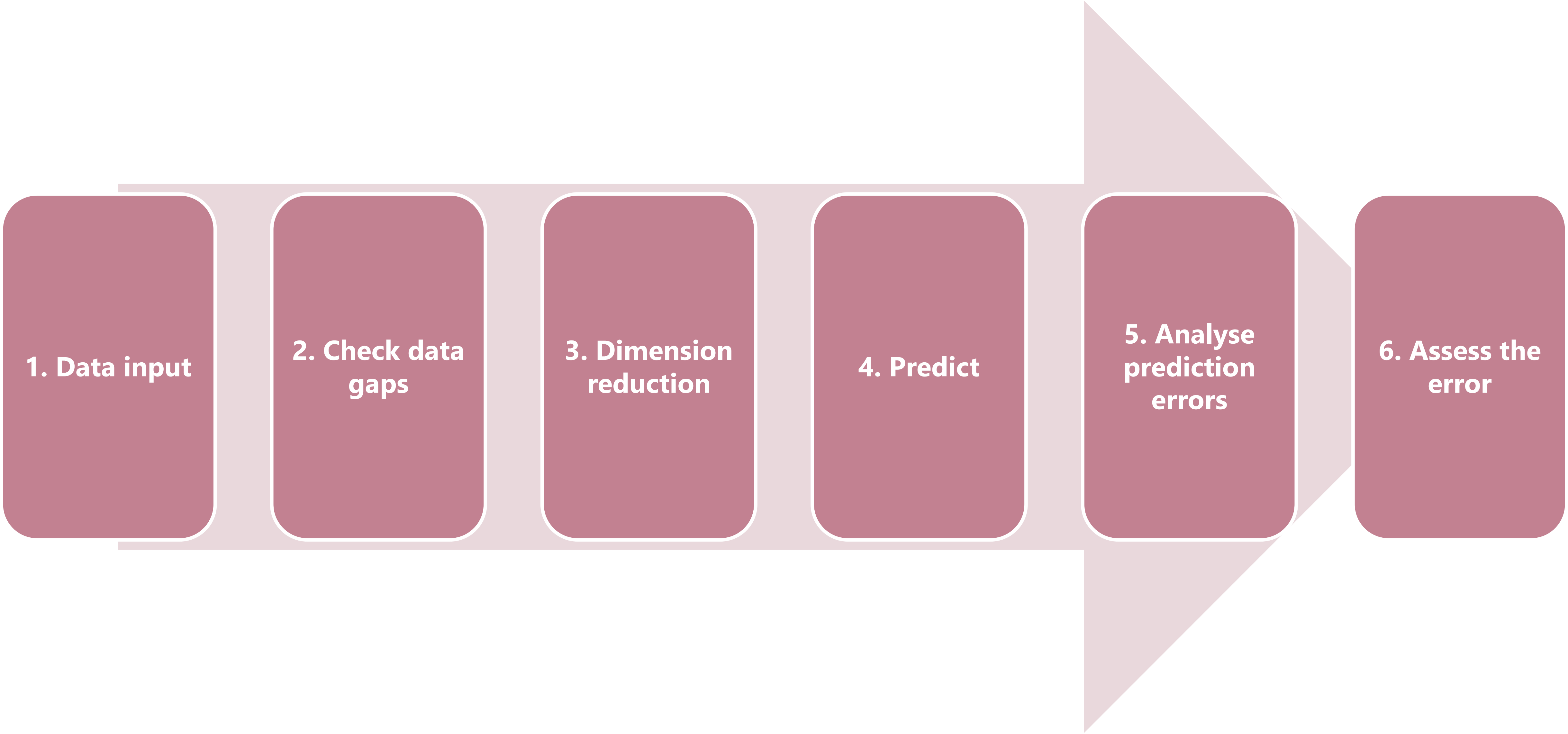
- Alone or combination of
  - Graphical method
  - Conditional controls (eg if … then …)
  - Threshold-based warnings
  - Cross-referencing
- Are they enough to validate the previous cases?







Stylized work flow of data validation process using machine learning models



1. Data input

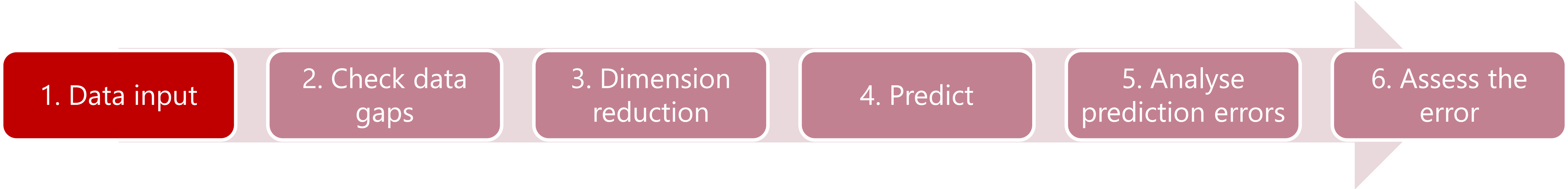
2. Check data gaps

3. Dimension reduction

4. Predict

5. Analyse prediction errors

6. Assess the error



1. Data input

2. Check data gaps

3. Dimension reduction

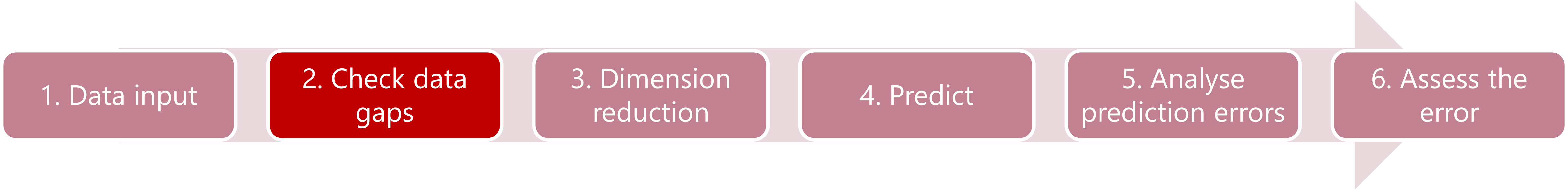
4. Predict

5. Analyse prediction errors

6. Assess the error

- About 3,000 daily incoming FM data from Bloomberg
- Characteristics of the financial market time series
  - **High-frequency big** data
  - Anomalies are **not easily visible** to human eyes
  - Frequent market **structural changes** due to market conditions, policy changes, new & outdated instruments
  - In the context of BIS – global coverage makes it difficult to understand the contexts

- **Most series have highly correlated and/or good explanatory series, for example:**
  - NASDAQ Index, NASDAQ 100, MSCI US, S&P500, ..
  - Yields of similar maturities
  - Pegged FX rates



1. Data input

2. Check data gaps

3. Dimension reduction

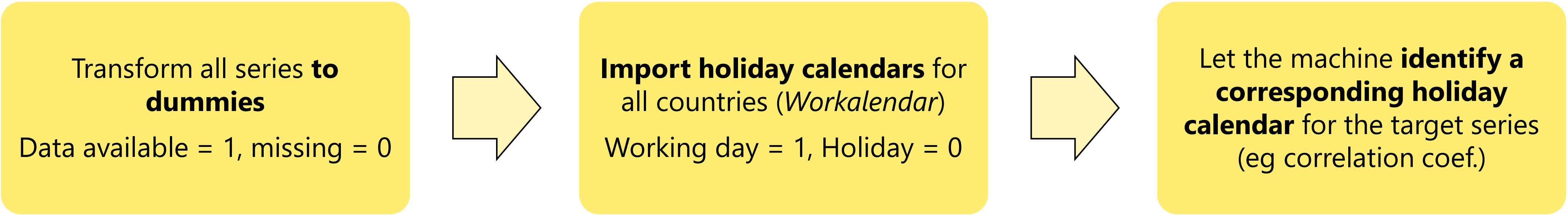
4. Predict

5. Analyse prediction errors

6. Assess the error

- How to verify if a data gap of a target series is due to market closure?

Step 1: Identify a **holiday calendar**



Transform all series **to dummies**

Data available = 1, missing = 0

Import holiday calendars for all countries (*Workalendar*)

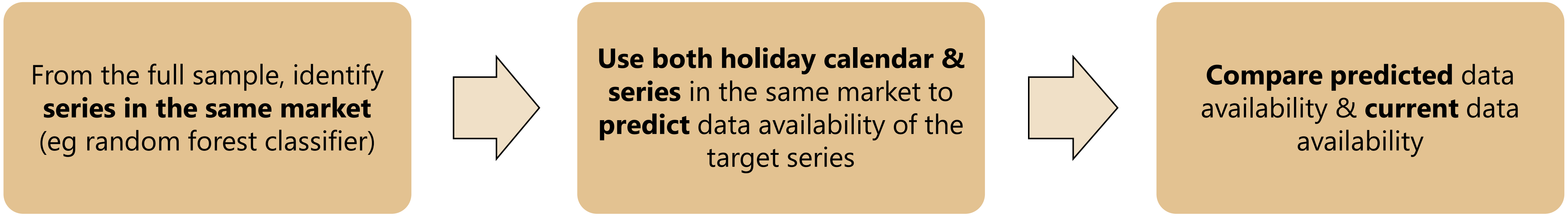
Working day = 1, Holiday = 0

Let the machine **identify a corresponding holiday calendar** for the target series (eg correlation coef.)

But having the holiday calendar is not enough (eg end-year stock market closures)

Step 2: Use the **holiday calendar** & any **series in the same market** to check data gaps

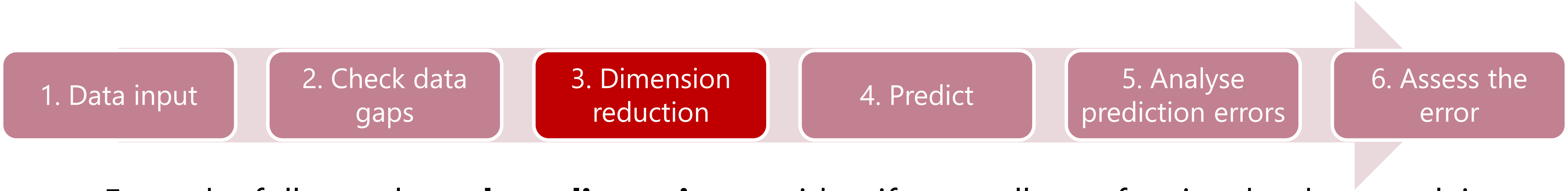
Supplementing the holiday calendar with series in the same market



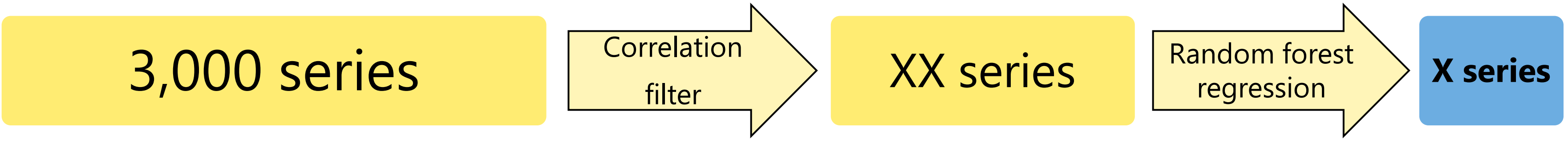
From the full sample, identify **series in the same market** (eg random forest classifier)

Use both holiday calendar & series in the same market to **predict** data availability of the target series

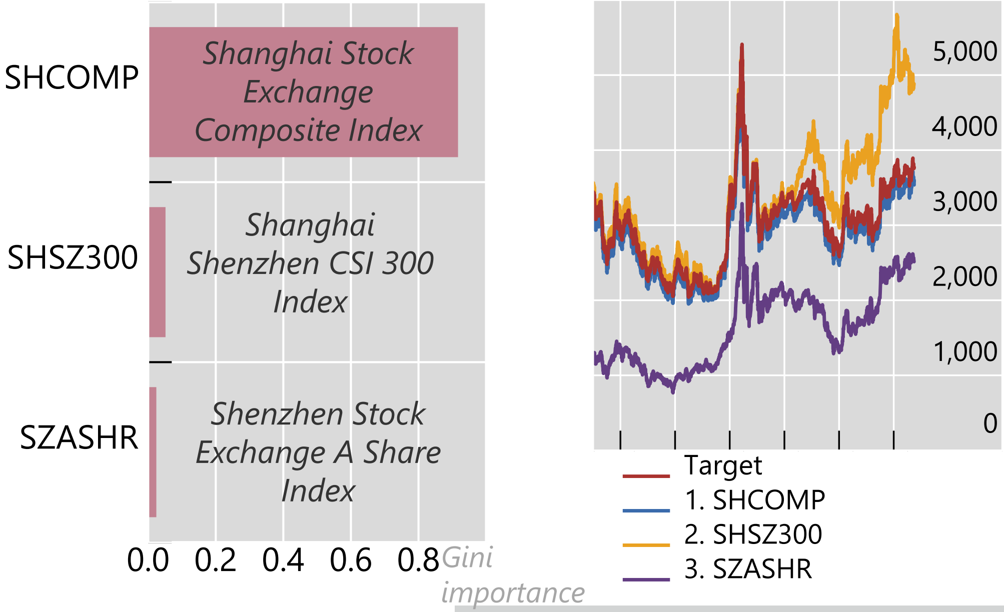
Compare predicted data availability & **current** data availability



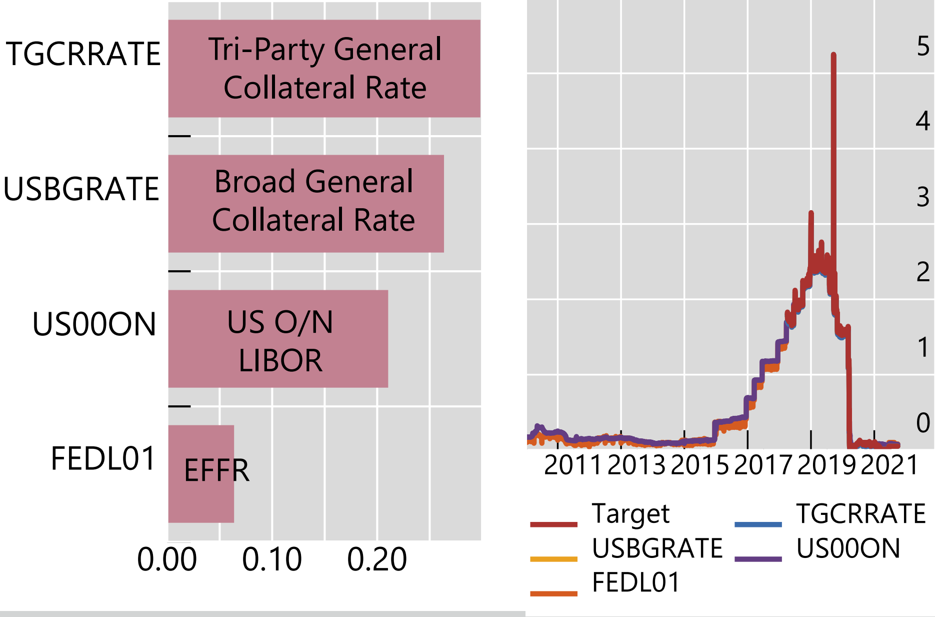
- From the full sample, **reduce dimensions** to identify a small set of series that best explain the target series (ie the most important features)



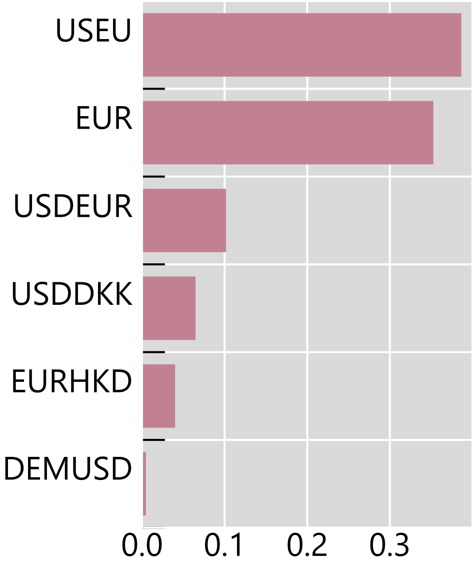
Shanghai Stock Exchange A Share Index

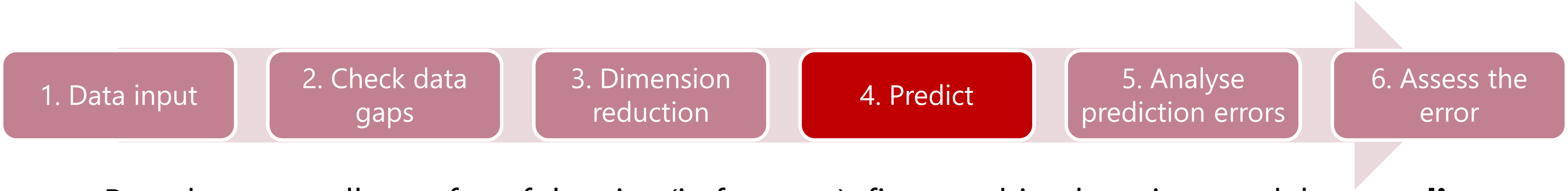


US Secured Overnight Financing Rate (SOFR)



EUR/USD

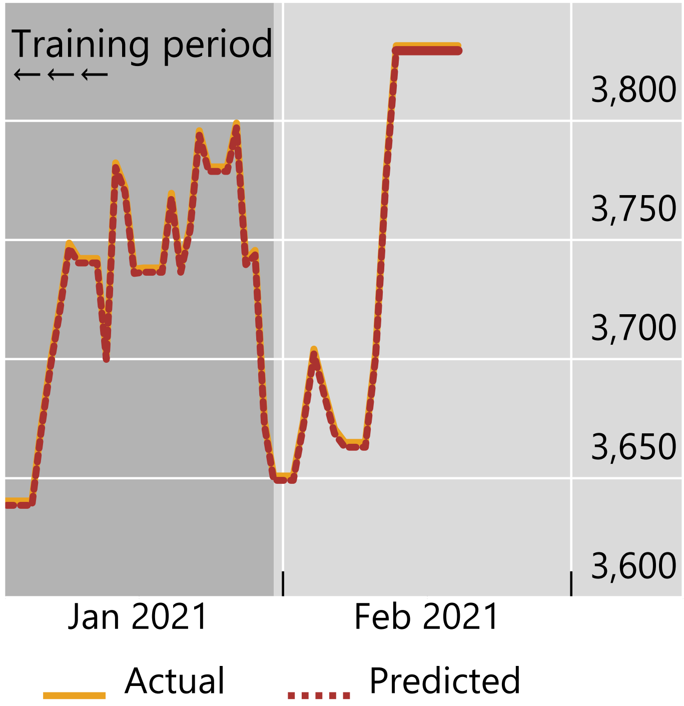




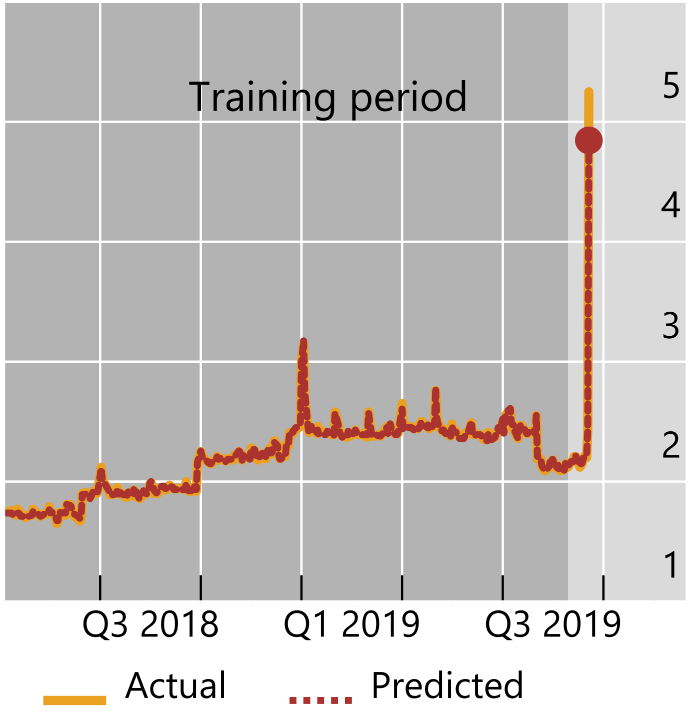
1. Data input
2. Check data gaps
3. Dimension reduction
4. Predict
5. Analyse prediction errors
6. Assess the error

- Based on a small set of useful series (ie features), fit a machine learning model to **predict today’s value** of the target series as if today’s value is unavailable.

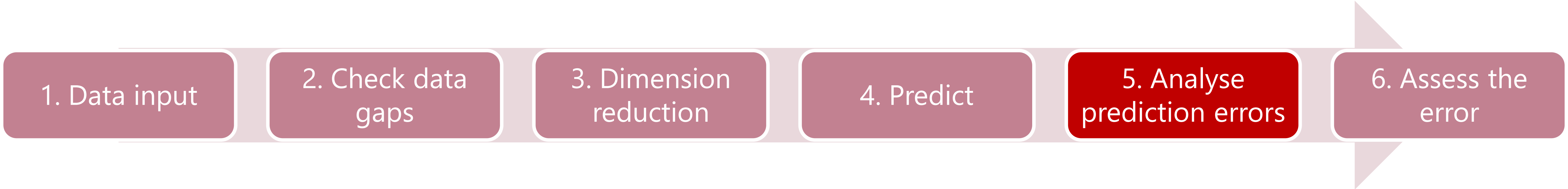
Shanghai Stock Index



US SOFR



Both were predicted based on the **exactly same LSTM specifications**.



1. Data input

2. Check data gaps

3. Dimension reduction

4. Predict

5. Analyse prediction errors

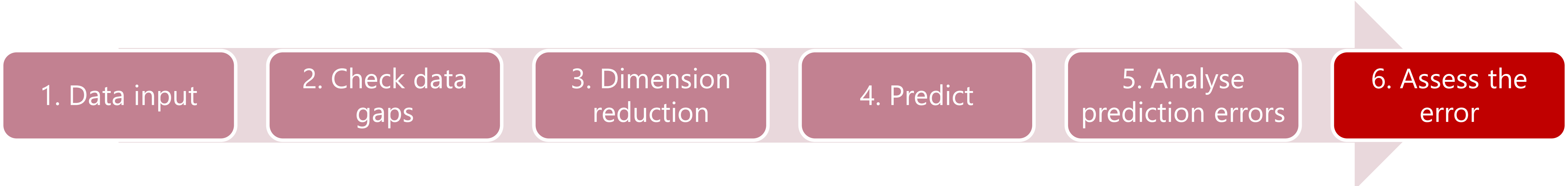
6. Assess the error

- Prediction error = Predicted value – actual value

- If a prediction error is significantly *larger than usual (ie outlier)*, it is worth investigating

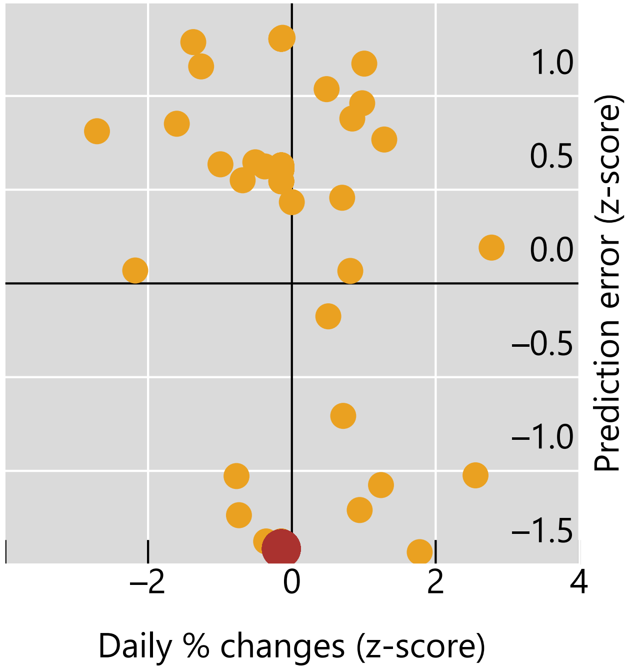
- Then, how can we decide whether a prediction error is an outlier?

➢ **Unsupervised learning algorithm** can help

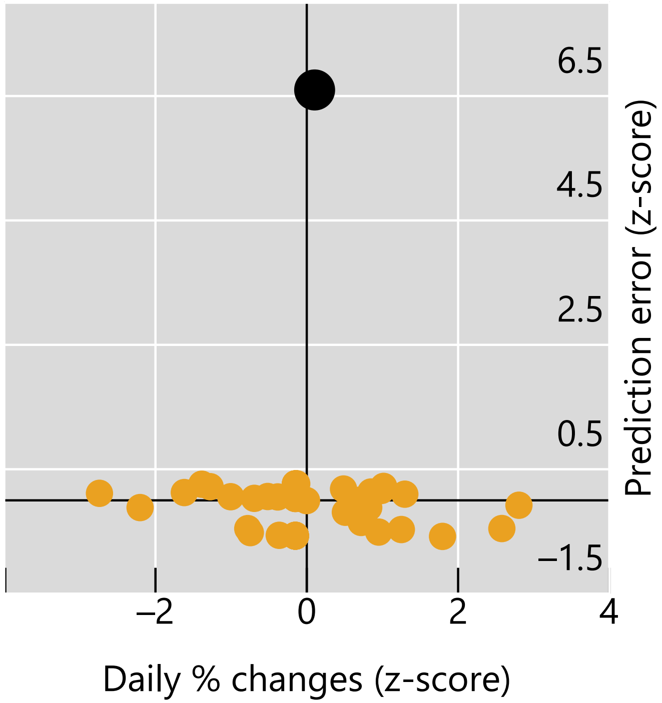


- A prediction error can be assessed based on unsupervised clustering algorithm (eg DBSCAN)
- With Shanghai Stock Index example:

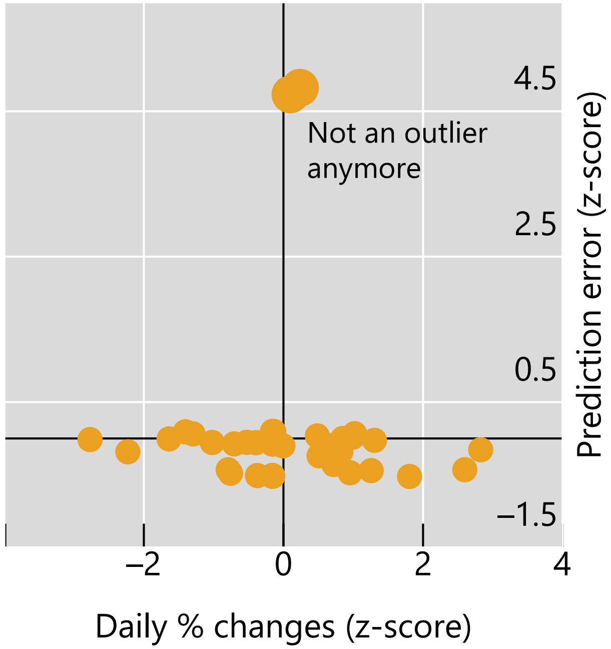
No outlier identified (eps=2, min sample=2)



What if we impose a small error in the data, will it be identified?



What if that was not an error but a *structural change* in the market



Thank you

- Select tools and models used in our analysis:
  - **Workalendar**: https://github.com/workalendar/workalendar
  - **Random forest classifier/regressor**: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html; https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
  - **Long short-term memory (LSTM):** https://keras.io/api/layers/recurrent_layers/lstm/
  - **Density-based spatial clustering of applications with noise (DBSCAN)**: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Time series outlier detection, a data-driven approach[1]

Nicola Benatti, European Central Bank,
and Alexis Maurin, Bank of England

[1] This presentation 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 central banks and other institutions represented at the event.

Time series outlier detection, a data-driven approach

Alexis Maurin, Nicola Benatti

The COVID-19 pandemic has severely impacted the world economy, leading to abrupt changes in collected statistics. It raised the need for new appropriate methodologies to ensure the production of accurate indicators. In this paper, we propose a methodology of macro-economic time series outlier detection, robust to breaks and spikes. By applying unsupervised machine learning techniques, we explore novel ways of identifying abnormal observations in the broad sense, focusing on the dynamic of time series. Clustering algorithms, which aim to group series with similar dynamics, can reveal exogenous information and help us to better detect outliers to be investigated.

The views expressed here are the sole responsibilities of the authors and should not be interpreted to reflect the views of the Bank of England nor the European Central Bank.

Keywords: outlier detection, unsupervised learning, clustering, dynamics finding

JEL classification: C14, C38, C82

Contents

1. Introduction

Motivation and needs

Macro-economic indicators are widely used by researchers and economists on a plethora of topics (e.g. financial accounts, consumer price index, business demographic, labour market) and at different frequencies (monthly, quarterly and yearly). These data are subject to large and unexpected shocks (e.g. economic crises, political or social reforms) which can cause abrupt movements from one period to another. The current COVID-19 pandemic has heavily impacted them and acted as stressor to the classical data quality monitoring procedures such as outlier detection, which would flag considerably more outlying data points than usual.

The classical methodologies to detect outlying data points often use a specified threshold on the growth rate, or forecast a value (e.g. ARIMA models) and verify whether the real value lies within the estimated confidence interval or not. However, it is quite intricate to define the proper threshold on growth rates above which a data point should be defined as an outlier, while the forecast-based approach can only be applied to the most recent data points and does not use the information from the latest data points. Beyond that, these methods are solely considering a univariate approach. They are not using the possible relations between the series, which can add valuable information with regards to the identification of outliers. There exist advanced time series forecasting algorithms which include the modelling of exogenous variables (ARIMAX model) or operate in a multivariate manner (VAR/VARMA models). The VARMAX model even combines the two approaches. These techniques can be efficient for specific cases but they require an advanced data treatment, thus restraining their generalised implementation, especially for automated procedures.

We explored data-driven ways to create an outlier detection procedure, robust to systemic breaks and spikes, applicable to any macro-economic time series data, with the aim of better detecting series with abnormal behaviour which might either come from reporting error or strong temporary factors.

Testing of classical methodologies

In this paper, we focus our interest on employment data from the National Accounts, which is described in Section 2.

Growth rate

Figure 1 displays two annual series from Spain: the left graph depicts the evolution of self-employed jobs in the G[1], H[2] and I[3] sectors and the right one depicts the evolution of the number of employees (as persons) in the F[4] sector. The complete description

[1] Wholesale and retail trade; repair of motor vehicles and motorcycles

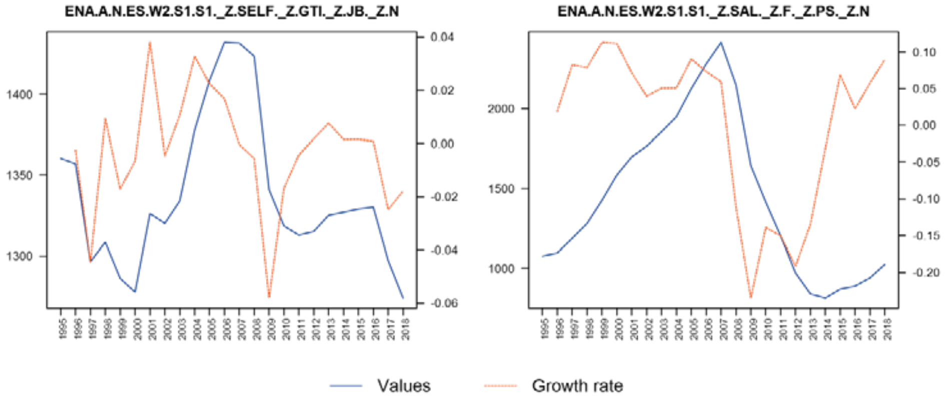
[2] Accommodation and food service activities

[3] Transportation and storage

[4] Construction

list of sectors defined following the NACE rev. 2 classification is available in Annex A. For illustration purposes, we processed annual data from 1995 to 2018. The values plotted as a blue solid line are reflected on the left y-axis. The growth rates plotted as an orange dashed line are reflected on the right y-axis.

Figure 1. Examples of annual series with abrupt decrease



The order of magnitude of the growth rates can differ between series: from -0.06 to 0.04 for the series on the left-hand side, from -0.22 to 0.10 for the series on the right-hand side. This can result in challenges with regards to the definition of the appropriate threshold above which a certain data point should be defined as an outlier. We need more information (e.g. what the macro-economic environment in this country for this sector is) to correctly identify an unusual change, and thus cannot rely on a univariate method.

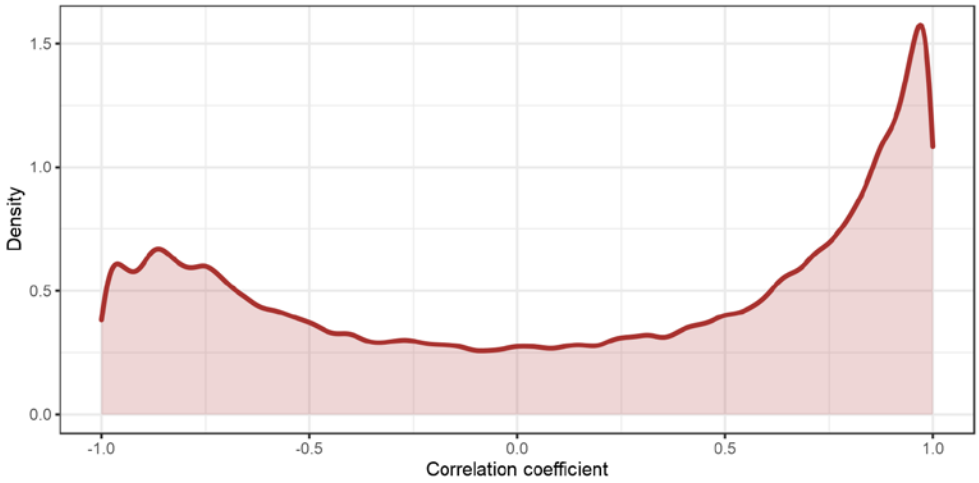
Time series modelling

We fitted ARIMA models on quarterly data, from 1995Q1 to 2020Q1, and forecasted the value for 2020Q2, a quarter that was heavily impacted by the COVID-19 pandemic. We then checked if the real value lies in the estimated confidence interval which was set to 99.5%, to relax the classical outlier rule. Among the 6638 series, 2957 were flagged as series containing outlying data points for 2020Q2. In comparison, when fitting ARIMA models from 1995Q1 to 2019Q3 and forecasting 2019Q4, only 149 series were flagged. The ARIMA models behave quite well in general but are overwhelmed in cases where external shocks impact and drive the series' movement, therefore requiring a procedure which can overcome this situation.

Our approach

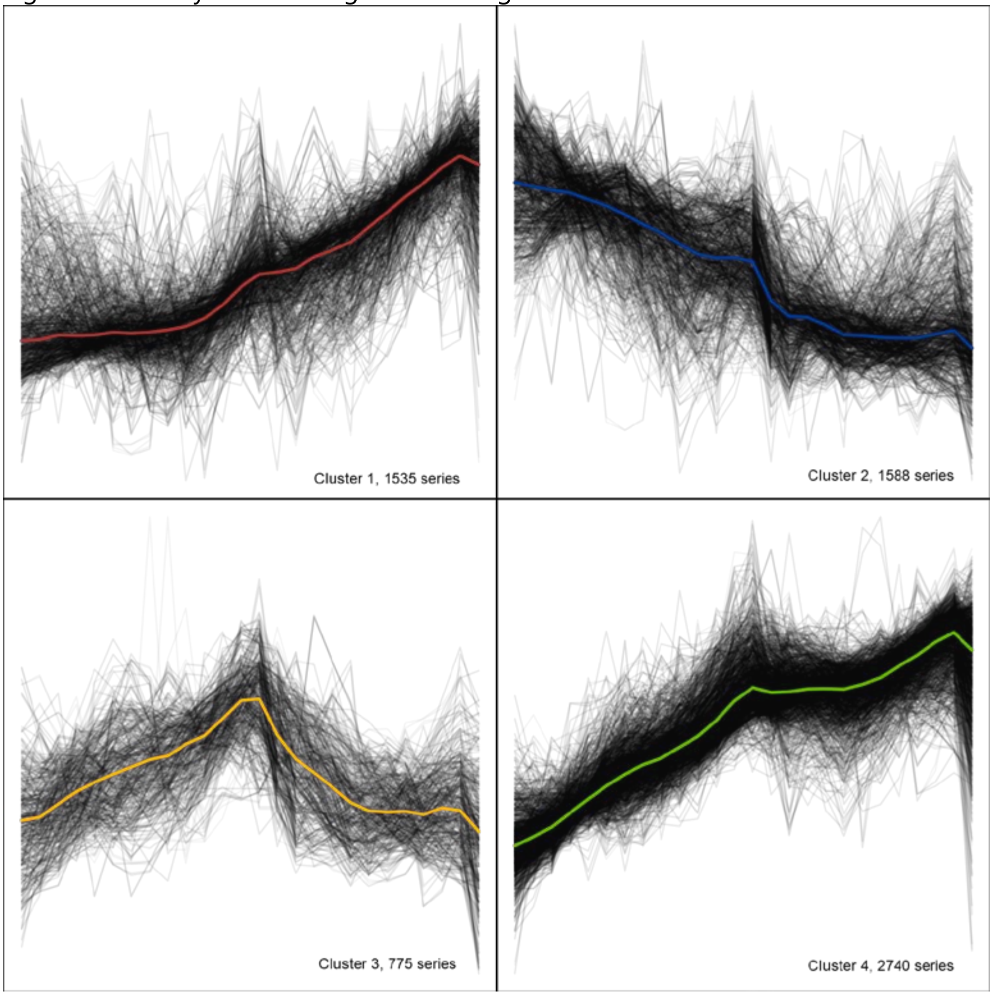
Macro-economic datasets cover topics broken down into multiple dimensions and therefore often exhibit relations between the time series that compose each dataset. By using the Spearman correlation, we can measure the rank correlation and observe monotonicity between the series. Figure 2 plots the density function of the correlation coefficients across the quarterly employment series of the National Accounts, from 1995Q1 to 2021Q1, smoothed with the LOWESS algorithm described in Section 3. We can observe higher masses for strongly correlated series, implying strong monotonic relationships. The lower mass around 0 indicates low correlation, i.e. of the uniqueness of one series' dynamic. We have a first evidence of groupable series.

Figure 2. Spearman correlation density curve



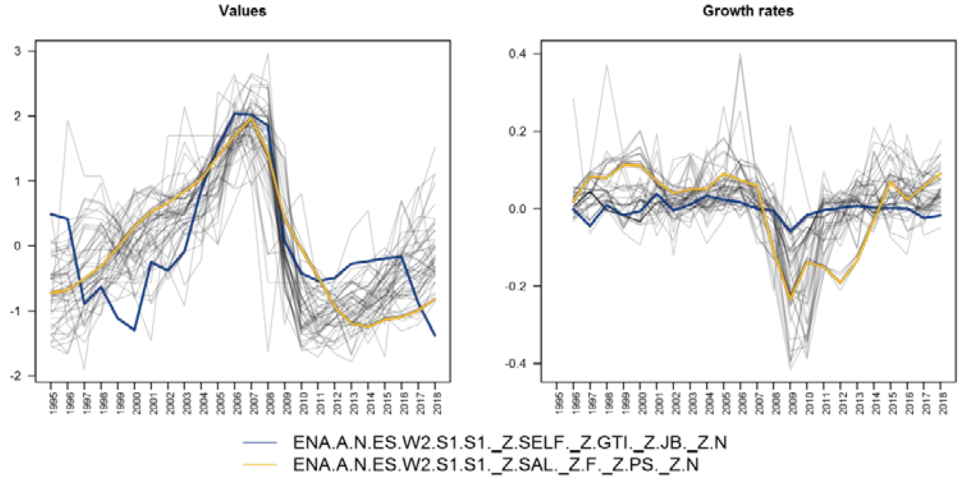
Going further into this inspection, we can identify the main dynamics. The K-means algorithm is one of the simplest and most popular clustering algorithms which enables the grouping of similar observations, series in our case. Figure 3 plots the four main patterns found in the annual data from 1995 to 2020, extracted from the quarterly data, which corroborate the presence of dynamics (median series coloured).

Figure 3. Main dynamics using K-means algorithm



The evidence of relationships between the time series can be used to answer questions the classical methodologies cannot easily solve: "Is a large movement an actual outlier or a response to an externality?". Basing our decision-making on movements for each series requires advanced expertise and would involve a tremendous amount of time as seen with the two series shown previously. Therefore, to overcome this issue in a generalised fashion, our approach is to cluster series with similar dynamics. The series from Figure 1 lie in the same cluster, hence exhibiting similar dynamics despite different values and growth rates. Figure 4 plots all series which lie within this cluster. We can see that, by focusing on the scaled series (i.e. dynamics) instead of inspecting the large growth rates' range, we can target the series with an abnormal behaviour better. For these specific series, the Great Financial Crisis is causing the drop in growth rates in 2009.

Figure 4. Example of a cluster of series with similar dynamics



The left graph displays the scaled series. The right graph depicts their respective growth rates. The series with an absolute growth rate greater than 0.5 (50%) have been removed for visibility. The cluster comes from the procedure described in Section 4.

Objectives

The objective is to create an outlier detection tool applied to macro-economic data, as automated as possible, in order to help narrow down data points that would be identified as outliers and would consequently require further investigation. The approach is completely data-driven, in other words, we let the data speak for itself. The initial and only assumption made is that the series can be grouped with respect to their dynamic. We only use historical data and do not use any additional information with regards to the type of outlier we are seeking prior to the application of the methodology. We try to mimic the human experience in the identification process and rather nowcast than forecast. Subsequently, we look for the data points that differ the most from the cluster to which they have been allocated. They might present some peculiarity as they are not where they are expected to be.

The data is presented in Section 2, the statistical tools used are defined in Section 3, the procedure is described in Section 4 and the results are reported in Section 5.

2. Data

The National Accounts are compiled according to the accounting definitions and methodologies set out in the ESA 2010 Regulation. Each quarter, Eurostat publishes several macro-economic aggregates, including employment data with industry breakdowns, what we call Employment National Accounts (ENA). The ENA data covers 37 countries[5] on annual and quarterly frequencies, broken down into three dimensions:

- Branches of economic activities (NACE rev. 2 classification, description available in Annex A)
- Employment status (employees and self-employed) and totals (total employment and total population)
- Unit of measure (jobs, persons, hours worked and full-time equivalent)

Due to differences in data availability between countries, all dimensions described above are not evenly represented across countries. We focused on quarterly data from 2011Q1 to 2021Q1, and did not use European aggregates. In total, we included 6638 series covering 31 countries.

3. Statistical tools

This project was undertaken with the R software and the different statistical tools used are presented and described below.

Smoothing method – LOWESS

The Locally Weighted Scatterplot Smoothing (LOWESS) is a non-parametric regression tool that fits a smooth curve to data points. It is an iterative process which fits local polynomials on a sliding window using weighted least squares (WLS).

First, it applies weights with respect to the abscissa closeness, giving more weight (or influence) to the points that are closest to the estimated one. Second, it adjusts these fitted values based on their distance from the actual ones, adding additional weights to the WLS. This second step can be repeated multiple times, until the curve is sufficiently smoothed.

The LOWESS algorithm requires setting different parameters: the polynomial degree, a weight function, the number of iterations and the window size (i.e. smoothing parameter). The smoothing parameter defines the proportion of data points to be used within the windows in order to fit each polynomial. Larger values lead to more smoothness.

[5] AL, AT, BE, BG, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LI, LT, LU, LV, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, TR

We use the lowess() function from the stats library which fits polynomials of degree 1 (i.e. lines) and defines weights with Tukey's biweight function (with a cut-off set at 6 times the median absolute deviation of the residuals). The default number of iterations is set to 3. We use this tool to smooth the time series prior to the clustering algorithm with the intention of reducing the noise (e.g. removing seasonality) and catching only their main dynamic, as this methodology gives a robust estimation of the outliers through the weighting and iteration process.

Metrics

The clustering algorithms require a distance metric in order to compute the dissimilarities between the observations and thereby allowing to group the similar ones together. Contingent upon the type of data we are using, we have selected two distances: the Minkowski distance for cases with only quantitative variables and the Gower distance for cases with both quantitative and qualitative variables.

Minkowski distance

In cases for which only quantitative data needs to be processed, we apply the Minkowski distance which is defined as follows:

$$D_{Minkowski}(x_i, x_j, p) = \left(\sum_{v=1}^{V} |x_i^v - x_j^v|^p \right)^{\frac{1}{p}}, for\ V\ variables.$$

For our methodology, we apply this distance with p=2 (Euclidean distance). We use the dist() function from the stats library to compute this distance.

Gower distance

In cases for which both quantitative and qualitative variables compose our dataset, we have to use a different metric that can handle these two types of data simultaneously: the Gower distance. It is defined as follows:

$$D_{Gower}(x_i, x_j) = \frac{1}{V} \sum_{v=1}^{V} d_{ij}^v, for\ V\ variables,$$

where $d_{ij}^v \in [0,1]$ is the partial dissimilarity between observations i and j for the variable v.

Depending on the type of the variable (qualitative or quantitative), the partial dissimilarity is computed differently. For the quantitative variables, it is defined as the ratio of the absolute difference and the maximum range observed:

$$d_{ij}^v = \frac{|x_i^v - x_j^v|}{|\max(x^v) - \min(x^v)|}$$

For the qualitative variables, the dissimilarity takes the value 0 if the observations are the same and 1 otherwise:

$$d_{ij}^{v} = \begin{cases} 0 \ if \ x_i^v = x_j^v \\ 1 \ otherwise \end{cases}$$

We use the daisy() function from the cluster library to compute this distance.

Pre-processing

To avoid the scaling effect between features, we need to standardise our data. The Gower distance does not need a preliminary transformation as it is already integrated in the formula with the min-max scaling at the denominator. However, the Minkowski distance needs to be processed. We use the Z-score normalisation:

$$x_i^{k'} = \frac{x_i^k - \mu^k}{\sigma^k}$$

where μ^k is the mean value of the feature k and σ^k the standard deviation of the feature k.

Clustering algorithms

Clustering is an unsupervised machine learning technique with the goal of grouping observations with similar characteristics. We use two clustering algorithms in accordance with the objective we are trying to reach:

- The Affinity Propagation algorithm to identify the main dynamics of the series;
- The DBSCAN algorithm to detect the outlying data points.

Affinity Propagation

The Affinity Propagation algorithm, proposed by Frey and Dueck in 2007, is a clustering method based on the concept of "message passing". The algorithm has a graph-based approach, in other words, it considers each observation as nodes of a network between which “messages” are being exchanged. For each observation, the goal is to find the one that is the most representative of its cluster: its exemplar. Let us note that one observation’s exemplar can be itself. To exchange these "messages", the algorithm uses 3 matrices:

- The Similarity matrix $s(i,j)$ which measures the similarity between the observation i and j. We use the negative squared Euclidean distance:

  $$s(i,j) = -\left\|x_i - x_j\right\|^2$$

- The Responsibility matrix $r(i,j)$ which quantifies the extent to which the observation j is suited to be the exemplar of the observation i, taking into account other potential exemplars.
- The Availability matrix $a(i,j)$ which quantifies the extent to which the observation i should choose j as its exemplar, taking into account other potential exemplars.

The algorithm is initialised by considering all the data points as a potential exemplar and takes two main parameters as input: the preference vector and the damping factor. The number of clusters does not need to be specified which yield the Affinity Propagation more favourable to the other classical algorithms.

The preferences vector $s(i,i)$ represents the a priori suitability of a data point to be an exemplar. Its value directly influences the number of clusters and a high value per observation will increase the propensity to be chosen as an exemplar. In case of no a priori knowledge, the vector can be set as any quantile of input similarities, typically the median.

During the message-passing procedure, numerical oscillations may occur in some cases, leading the algorithm not to converge. The damping factor (λ) comes into play. At each iteration t, the messages are damped by this factor as follows:

$$r_t(i,j) = \lambda \cdot r_{t-1}(i,j) + (1-\lambda) \cdot r_t(i,j)$$

$$a_t(i,j) = \lambda \cdot a_{t-1}(i,j) + (1-\lambda) \cdot a_t(i,j)$$

λ takes values between 0 and 1.

After having defined these parameters, the algorithm procedure can begin:

1. Set the responsibility and availability matrices to 0

2. Repeat until convergence (number of iterations reached or exemplar remains unchanged for a defined number of iterations):

   - Update responsibility

   $$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\}$$

   - Update availability

   $$a(i,k) \leftarrow \min_{i' \neq k}\left\{0, r(k,k) + \sum_{i' \notin \{\mathrm{i,k}\}} \max\{0, r(i',k)\}\right\} \, for\ i \neq k$$

   $$a(k,k) \leftarrow \sum_{i' \neq k} \max\{0, r(i',k)\}$$

3. Extract the exemplar of each observation to get the clusters:

$$examplar_i = \operatorname*{argmax}_k\{a(i,k) + r(i,k)\}$$

We use the apcluster() function from the apcluster library.

DBSCAN

The DBSCAN (Density-Based Spatial Clustering and Application with Noise), proposed by Ester, Kriegel, Sander and Xu in 1996, is a density-based clustering algorithm. Its goal is to seek high density areas, which are then defined as clusters. Similar to the Affinity Propagation algorithm, the number of clusters does not need to be specified and has two main parameters to set:

- The epsilon (eps) defines the radius around an observation in which other observations will be defined as neighbours.
- The minimum points (MinPts) defines the minimum number of neighbours required within an observation's radius to form a dense region.

A distance metric must be defined to measure the closeness between the observations, then the algorithm works as follows:

- For each observation, it computes its distance to all the other observations and counts how many are falling into the epsilon radius. If the count is inferior to the MinPts parameter (but not zero), the observation is marked as a border point. If the count is superior or equal, it is marked as a core point. Finally, if it has no neighbour, it is defined as noise.
- If a core point is not assigned to a cluster, a new one is created. Through a chaining process, all the connected core points are found and assigned to this cluster.
- Finally, it allocates each border point to the closest connected cluster.

We used the dbscan() function from the dbscan library.

4. Procedure

Dynamics finding

We first use a clustering algorithm to catch the unobservable information represented by the different dynamics. The main challenge here is to define the optimal number of patterns (i.e. clusters) as the partitioning algorithms require to set the number of clusters as a parameter and the hierarchical algorithms require to define a threshold from the dendrogram. Our need for an automated procedure led us to look for other types of clustering approaches as the previously cited ones were not efficient nor easily optimisable. We compared several algorithms[6] and based on metrics (GAP value, silhouette coefficient and Dunn Index) and pertinence of the clusters, we decided to opt for the Affinity Propagation algorithm which had demonstrated simplicity, applicability (i.e. automation) and performance.

Before running the clustering algorithm, we normalise the data and apply the LOWESS algorithm with a window size of 20%. This smoothing aims to get rid of the short-term variations and only grasps the general dynamic over the whole time span.

In order to get the minimum number of clusters, the "preference" parameter is set to the minimum (0), giving the minimum distance between points to the preference vector. We set the damping factor λ to 0.5, to control oscillations and ensure convergence of the algorithm.

[6] K-means, K-medoid, Fuzzy C-means, HDBSCAN, DBSCAN, OPTICS, MeanShift, SOM, EM

Outliers identification

The rationale behind our approach is the following: within each cluster found by the Affinity Propagation algorithm, we apply the DBSCAN algorithm over all series, and we expect that for each period, the data points should lie in the same density area. Therefore, any inconsistency in the allocation of a data point will set it as outlying. It also allows us to track down outliers over the entire series, or to focus on a certain segment (e.g. latest data points).

To do so, for each cluster, we define as variables the scaled value of a data point for a given period and the period in which it falls. By computing the Gower distances on these two variables, two data points lying in the same period will have a distance between 0 and 0.5 (as the distance for the period will be equal to 0) and two data points in different periods will have a distance between 0.5 and 1. Therefore, a data point far from the main density area, where the data points of the same period lie, will be defined as an outlier.

We look for the most outlying data points for each cluster and designed the above operation as follows:

1. Convert wide to long format, adding the period
2. Compute the Gower distance between each data point.
   In case the number of data points is too high, a sliding window is set and the distances are computed over the defined portion.
3. Run a DBSCAN algorithm with parameters set as:
   - eps = largest minimum distance between two points
   - MinPts = 1
4. Get the number of outlying points
5. While no data points are spotted as outliers
   - Set: $eps \ = \ eps * 0.95$
   - Re-run a DBSCAN
6. We proceed up to 5 iterations

5. Results

We focused on the data of the last 10 years, from 20211Q1 to 2021Q1, as we only wanted to depict the short/medium term dynamics of the series. The Affinity Propagation algorithm identified 39 clusters. They are displayed in Annex B. Within each cluster, we ran the outlier identification from 2019Q4 to 2021Q1 and found 126 outlying data points. They are displayed in Annex C. For comparison, we fitted ARIMA models and flagged outliers using the 95% confidence interval rule. A total of 10716 data points were flagged. Our procedure shows a more efficient way to reduce the time an expert will investigate into these outliers at a granular level.

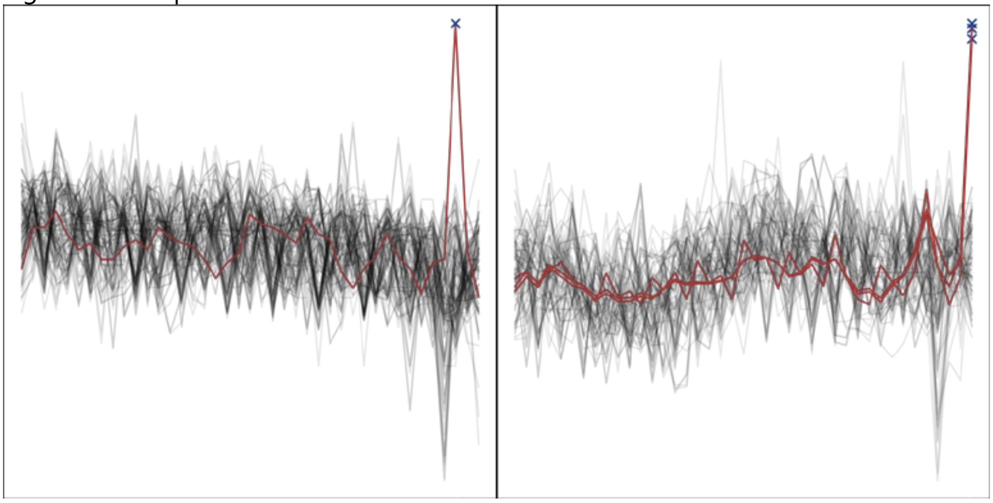
Table 1. Number of outliers found per period

	2019Q4	2020Q1	2020Q2	2020Q3	2020Q4	2021Q1	Computation time
ARIMA model	430	1563	3356	2557	1241	1369	5 hours 30 mins
Our procedure	7	7	52	6	19	35	13 mins

Examples of outliers

Figure 5 plots two examples of clusters in which the outliers flagged are showing a peculiar pattern compared to the cluster they lie in.

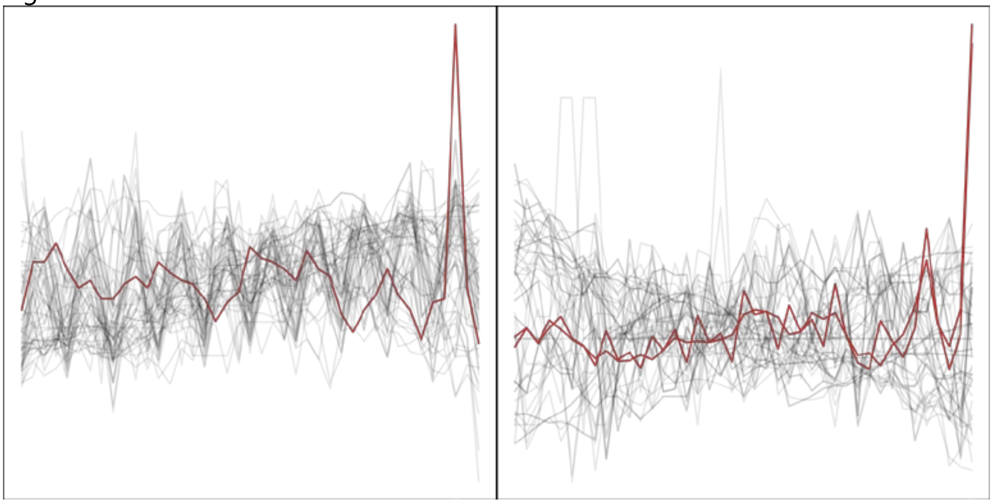
Figure 5. Example of outliers identified



The left cluster groups series which exhibit a slowly decreasing trend over time. While some series are showing an important decrease in 2020Q2, they recover quickly to the pre-shock level. One outlier is flagged in 2020Q3 as it is the only one displaying an abrupt increase. The outlier comes from the series depicting the number of employed persons in the A sector in a specific country. The right cluster groups series which exhibit a slow decrease from 2011Q1 to 2014Q1, followed by an increase until 2015Q2 and a stable trend until 2021Q1. While some series are showing a significant decrease in 2020Q2, they recover quickly to the pre-shock level. Three outliers are flagged in 2021Q1 as they are the only ones displaying a significant increase. The outliers come from the series depicting the total number of self-employed persons in the C sector (both seasonally adjusted and non-adjusted) as well as in the B, C, D and E sectors combined in a specific country.

We can further examine these series by comparing them to other series compiling the same statistic but from the other countries. Figure 6 plots sector A series on the left and the series of sector C on the right, with the outliers coloured. For sector A, only the spotted series shows the increase in 2020Q3. The rationale behind this increase for this country might be the compensation for the temporary suspension of fish activities to support commercial fishers who have been affected by the COVID-19 pandemic, coupled with the temporary easing of regulations covering aquaculture licences. For sector C, only the spotted series show the increase in 2021Q1. The increase for this country might be driven by the ongoing reopening of the economies and continued strength in demand from both domestic and European countries, combined with measures in response to COVID-19. The main measures to preserve employment and support household income for this country are special short-time work schemes and the extraordinary allowance for the self-employed (e.g. moratorium on tax debt and social security contributions, the deferral of tax payments).

Figure 6. A and C sectors scaled series



Our procedure solely uses historical data so that it can be run with any macro-economic data. On top of that, it allows to be less restrictive when it comes to the identification of outliers and rather flag too many data points than missing potential outliers. The complementary investigation carried out above could be implemented to enhance the outlier identification for our dataset, as well as the incorporation of additional information into the procedure (e.g. sectoral information).

6. Applications

The model we proposed has the characteristics of being completely data-driven, of considering long term dynamics of correlated series and, more important, of allowing for a break in the linearity of these correlations if this happens for a specific period as it was the case at the beginning of the COVID-19 pandemic. Given its wide range of applications on any type of time-series, the model can be considered to work independently or to be integrated into other tools. For example, in order to obtain more detailed explanations of why one observation is considered to be an outlier, we apply the feature-additive ranking technique on the observations spotted as outliers. This technique consists in running a XGBoost model using the outlying series spotted by the model presented above as dependent variable and using the remaining series in our dataset as features. A HDBSCAN model is then run on the residuals of the estimation to confirm outliers, Shapley values are calculated and aggregated to explain the model. The use of the time series clustering approach proposed in this paper allows us to filter a-priori the series to use as dependent variable and therefore maintain the process as computationally lightweight as possible.

7. Conclusion

This paper presents an approach for outlier detection for macro-economic datasets using unsupervised machine learning, robust to external shocks. Our procedure enables the reduction of the list of potential outliers, the ones with dynamics that differ the most from the general trends. It shows more efficiency compared to classical methodologies, especially amidst unstable periods such as the current COVID-19 pandemic. It can also be used as an investigation tool to identify time series with unusual movements in the broad sense, rather than looking for potential errors. The source code is available at *https://github.com/alexismaurin/ODMS*.

This approach is very generic as it only uses historical data and makes it applicable to any macro-economic data. The clustering process as well as the outlier identification can be adapted with respect to the data processed. This can be enhanced by adding exogenous data, in order to get more representative clusters and better target the data points that should be flagged as outliers.

Annexes

Annex A

Countries abbreviations

BE	Belgium	**HR**	Croatia	**PL**	Poland
BG	Bulgaria	**IT**	Italy	**PT**	Portugal
CH	Switzerland	**IS**	Iceland	**RO**	Romania
ME	Macedonia	**CY**	Cyprus	**AL**	Albania
CZ	Czech Republic	**LI**	Liechtenstein	**SI**	Slovenia
DK	Denmark	**LV**	Latvia	**SK**	Slovakia
DE	Germany	**LT**	Lithuania	**RS**	Serbia
EE	Estonia	**LU**	Luxembourg	**MK**	North Macedonia
IE	Ireland	**HU**	Hungary	**FI**	Finland
GR	Greece	**MT**	Malta	**SE**	Sweden
ES	Spain	**NL**	Netherlands	**TR**	Turkey
FR	France	**NO**	Norway		
GB	Great Britain	**AT**	Austria		

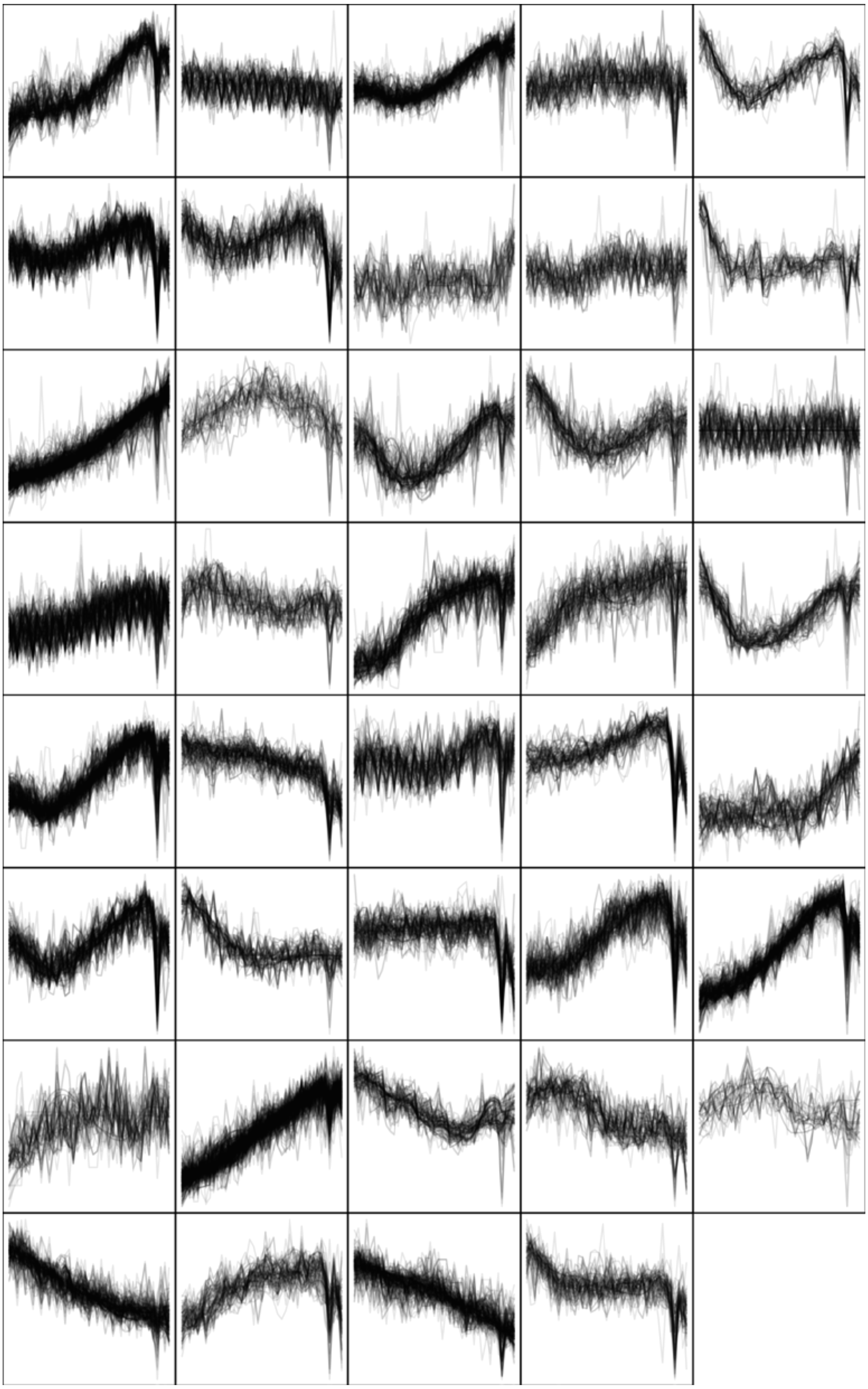
In accordance with EU practice, the EU Member States are listed in this report using the alphabetical order of the country names in the national languages.

NACE rev. 2 classification

Section	Description
A	Agriculture, forestry and fishing
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam and air conditioning supply
E	Water supply, sewerage, waste management and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Accommodation and food service activities
I	Transportation and storage
J	Information and communication
K	Financial and insurance activities
L	Real estate activities
M	Professional, scientific and technical activities
N	Administrative and support service activities
O	Public administration and defence; compulsory social security
P	Education
Q	Human health and social work activities
R	Arts, entertainment and recreation
S	Other service activities
T	Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use
U	Activities of extraterritorial organisations and bodies

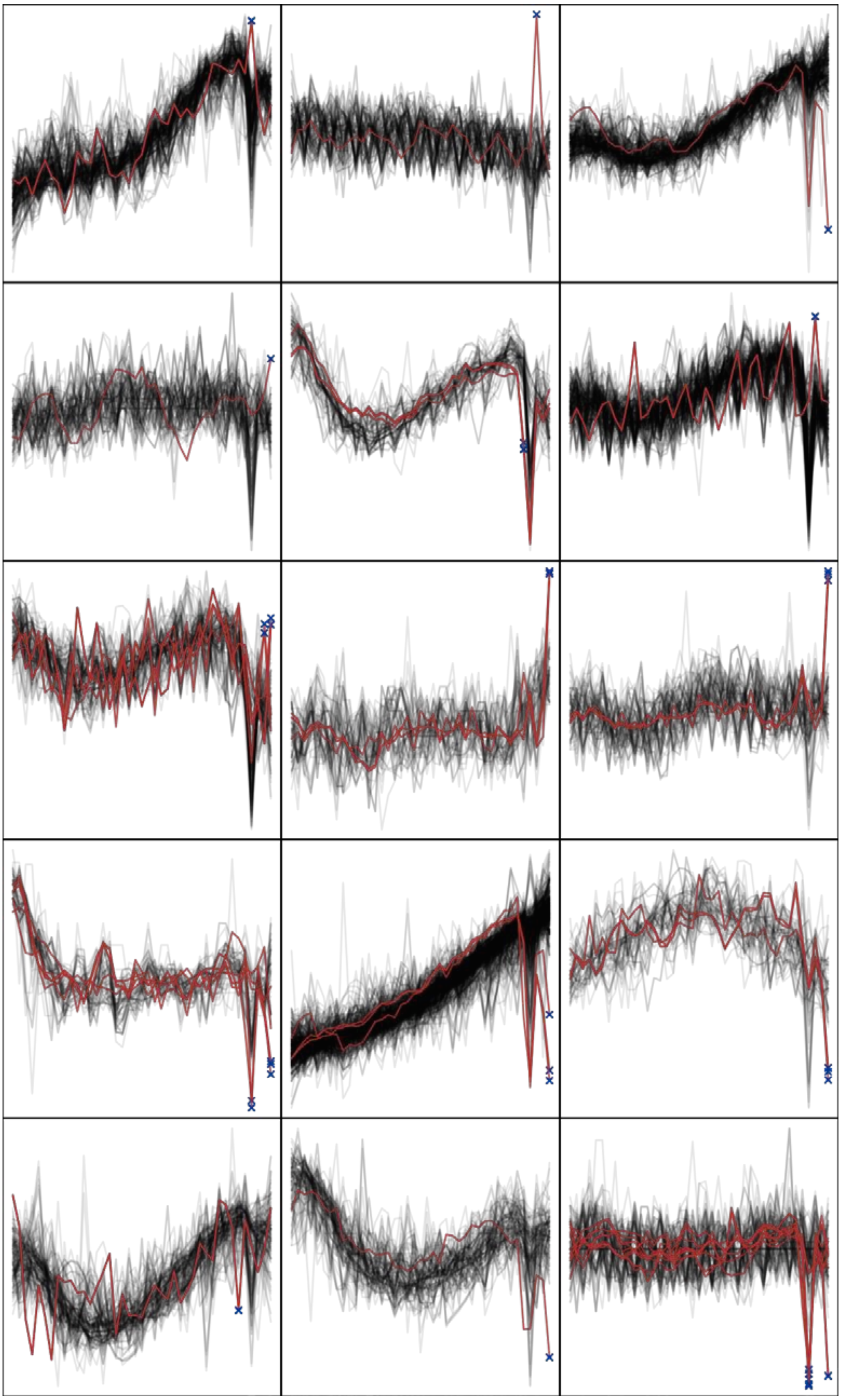
Annex B

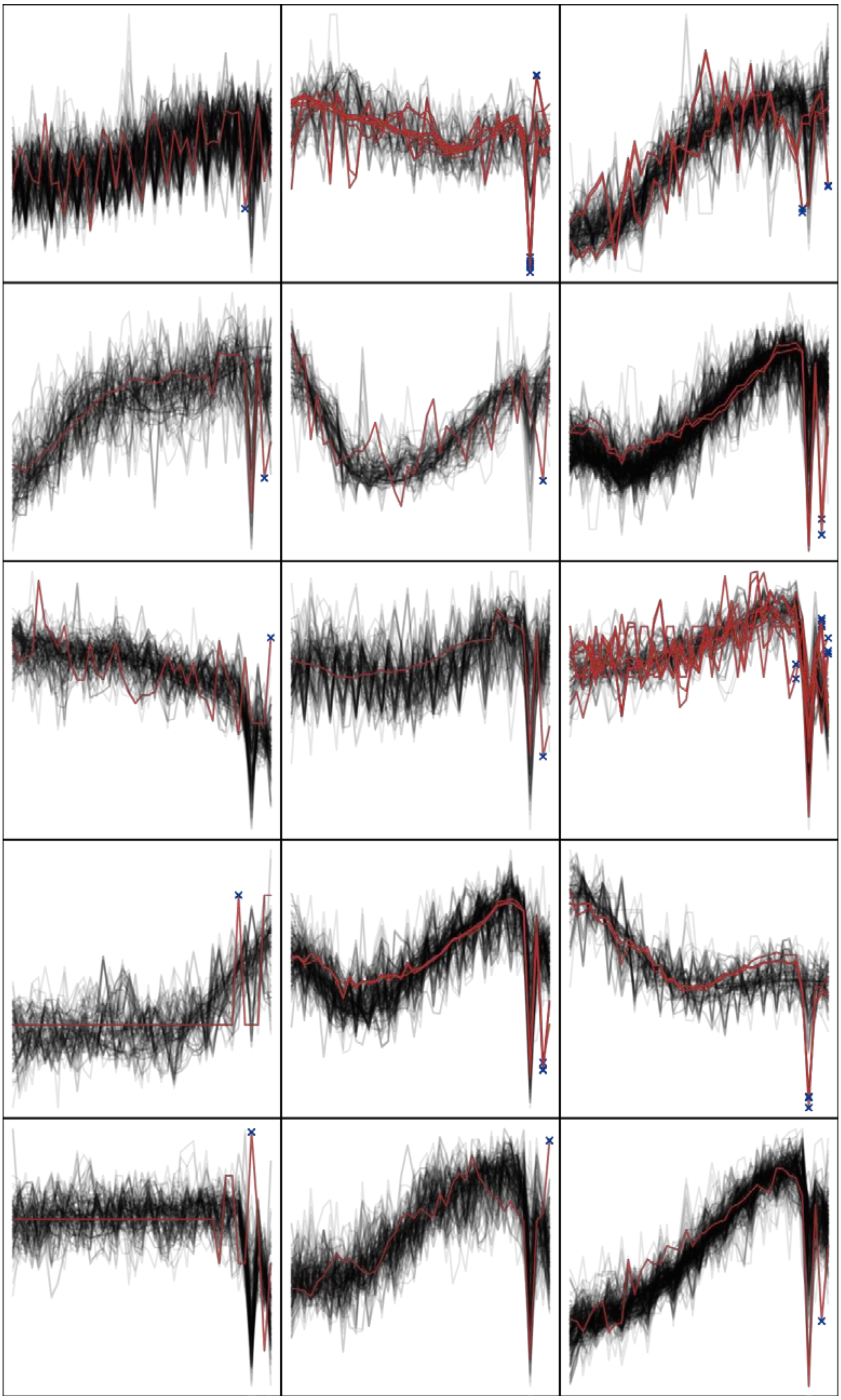
Clusters found by the Affinity Propagation algorithm from 2011Q1 to 2021Q1

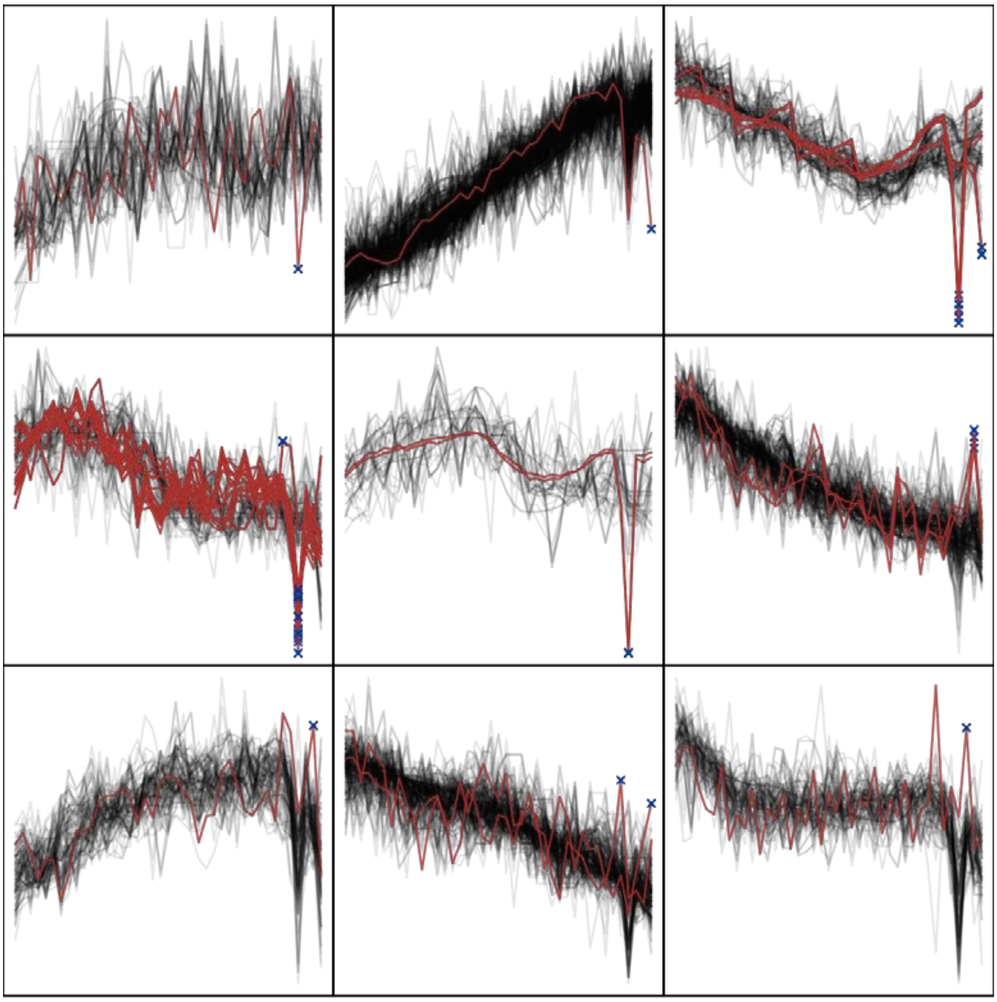


Annex C

Outliers found in each cluster from 2019Q4 to 2021Q1







References

Atkinson A.C., Koopman S.J. and Shepard N. (1997) "Detecting shocks: outliers and breaks in time series", Journal of Econometrics 80, 387-422

Benatti N. (2019) "A machine learning approach to outlier detection and imputation of missing data", IFC Bulletins chapters, Bank for International Settlements (ed.), Are post-crisis statistical initiatives completed?, volume 49, Bank for international Settlements.

Brendan J.F. and Delbert D. (2007) "Clustering by Passing Messages Between Data Points", Science 315, 972-976.

Cleveland W.S. (1979) "Robust locally weighted regression and smoothing scatterplots", Journal of the American Statistical Association 74, 829-836.

Ester M., Kriegel H., Sander J. and Xu X. (1996) "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Institute for Computer Science, University of Munich. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226-231.

Goin D.E. and Ahern J. (2019) "Identification of Spikes in Time Series" Epidemiologic Methods 8

Gower J.C. (1971) "A general coefficient of similarity and some of its properties", Biometrics 27, 857-874.

Hyndman R.J. and Khandakar Y. (2008) "Automatic time series forecasting: The forecast package for R", Journal of Statistical Software, 26(3).

Kaufman L. and Rousseeuw P.J. (1990) "Finding Groups in Data: An Introduction to Cluster Analysis", Wiley, New York.

MacQueen J. (1967) "Some methods for classification and analysis of multivariate observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and* Probability", Berkeley, University of California Press, 1:281-297.

Struyf A., Hubert M. and Rousseeuw P.J. (1997) "Integrating Robust Clustering Techniques in S-PLUS", Computational Statistics and Data Analysis 26, 17-37.

BANK OF ENGLAND

IFC Conference 19th-22nd Oct. 2021

Time series outlier detection, a data-driven approach

The project was carried out when both authors were working at the European Central Bank

Alexis Maurin - Bank of England
Nicola Benatti - European Central Bank

The views expressed here are the sole responsibilities of the authors and should not be interpreted to reflect the views of the Bank of England nor the European Central Bank.

Overview

- Data
- Motivation & Needs
- Approach & Goal
- Procedure
- Results & Application
- Conclusion

Employment National Accounts (ENA)

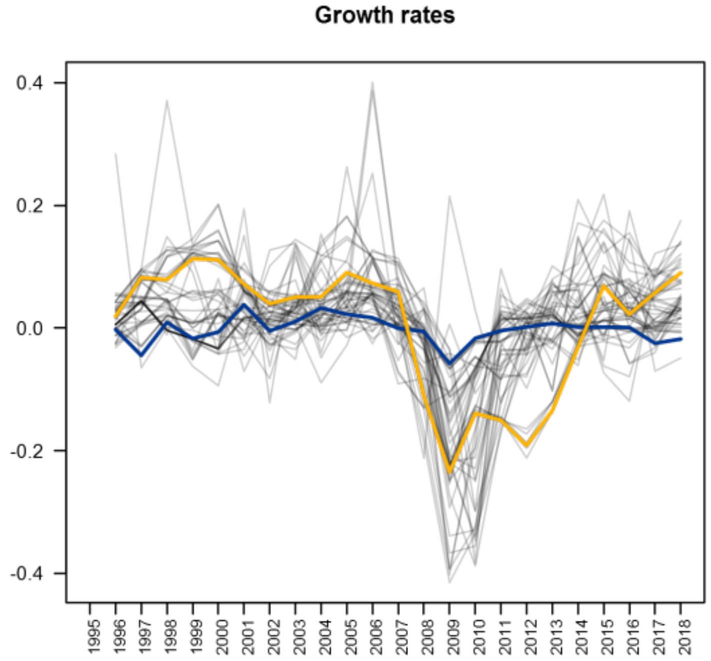
Quarterly data, broken down into three dimensions:

- Branches of economic activities
- Employment status (employees and self-employed) and totals (total employment and total population)
- Unit of measure (jobs, persons, hours worked and full-time equivalent)

- Total of 6638 series covering 31 countries

Motivation and needs

- Macro-economic indicators subject to unexpected shocks
- COVID-19 pandemic heavily impacted their movement
- Data quality monitoring procedures challenged
- Classical Methodologies :
  - Threshold on growth rate:
    <u>Complex to define</u>
  - Univariate Time Series Forecasting:
    <u>20x more outliers flagged during shocks</u>

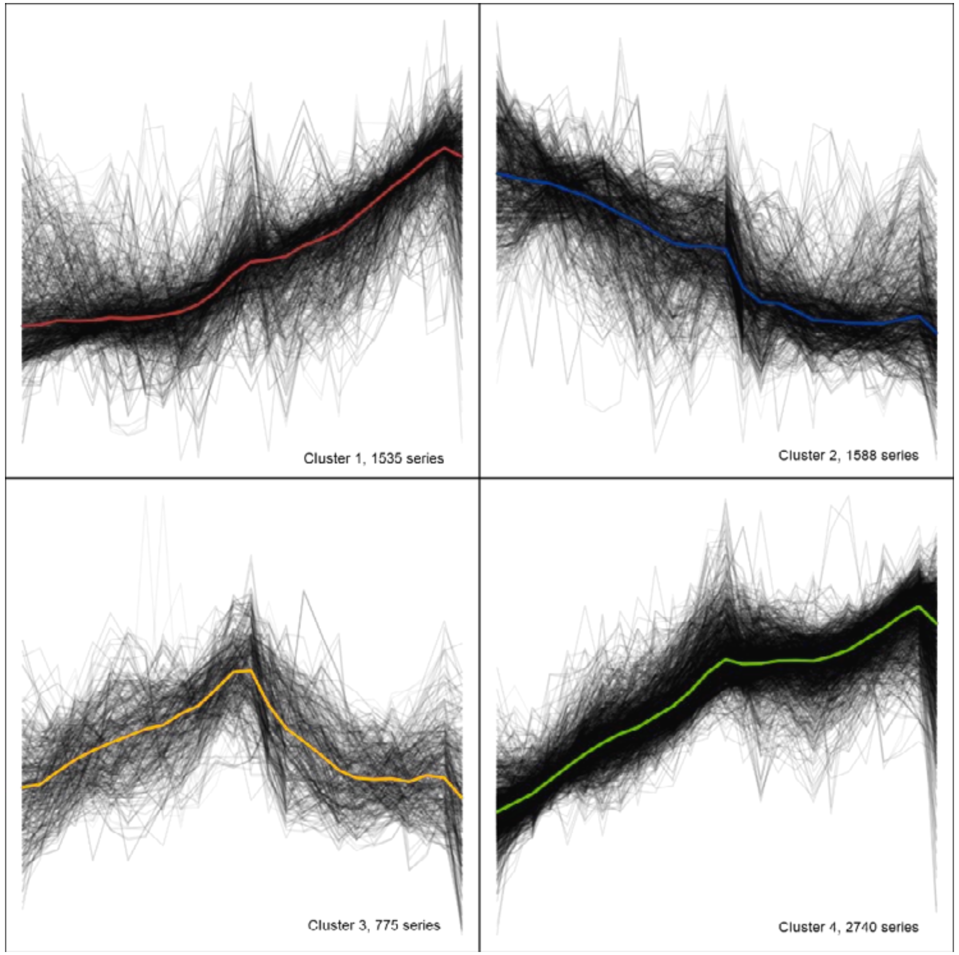


Approach and Goal

- Correlations within dataset
- Identify the main dynamics → **Clustering**
- Detect outliers within each cluster

Goal:

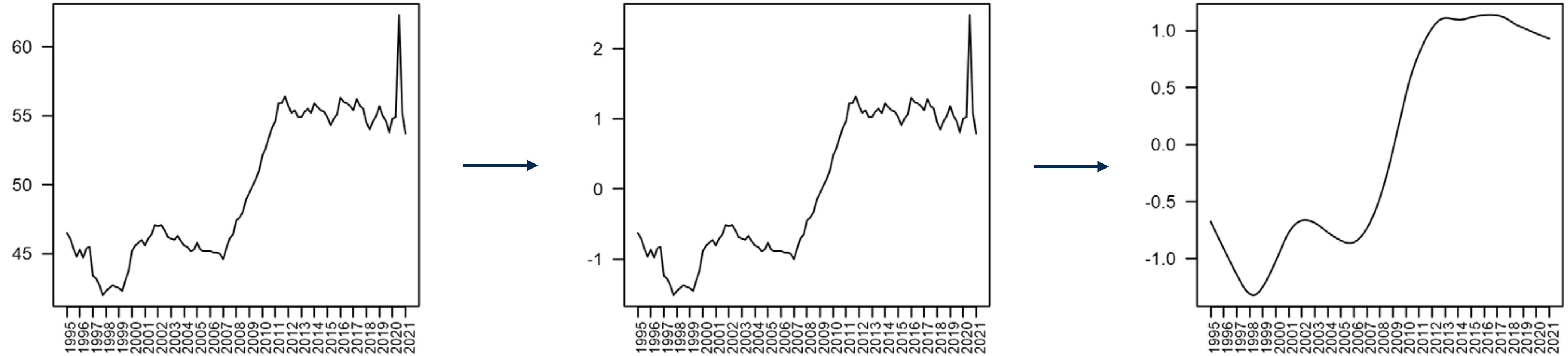
- Robust to systemic spikes and breaks
- Applicable to any macro-economic dataset
- High level automation



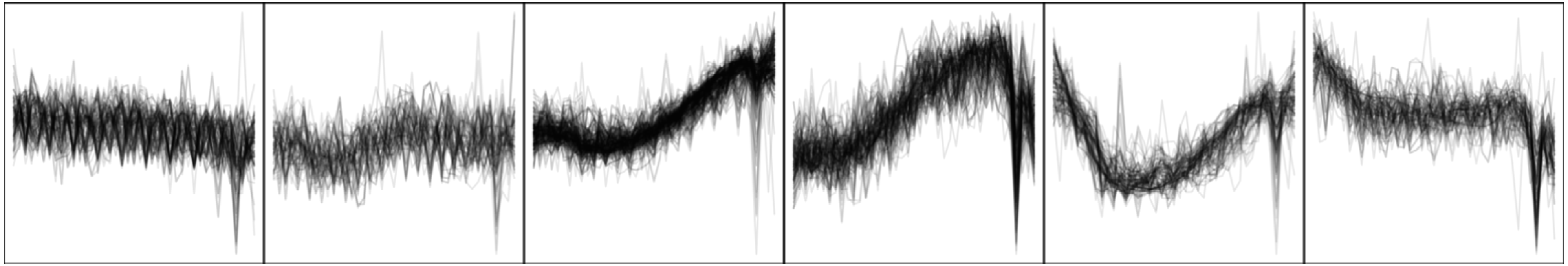
Main dynamics using K-means algorithm

Procedure

1. Standardise the data and smooth the series with the **LOWESS** algorithm



2. Identify clusters (dynamics) with the **Affinity Propagation** algorithm

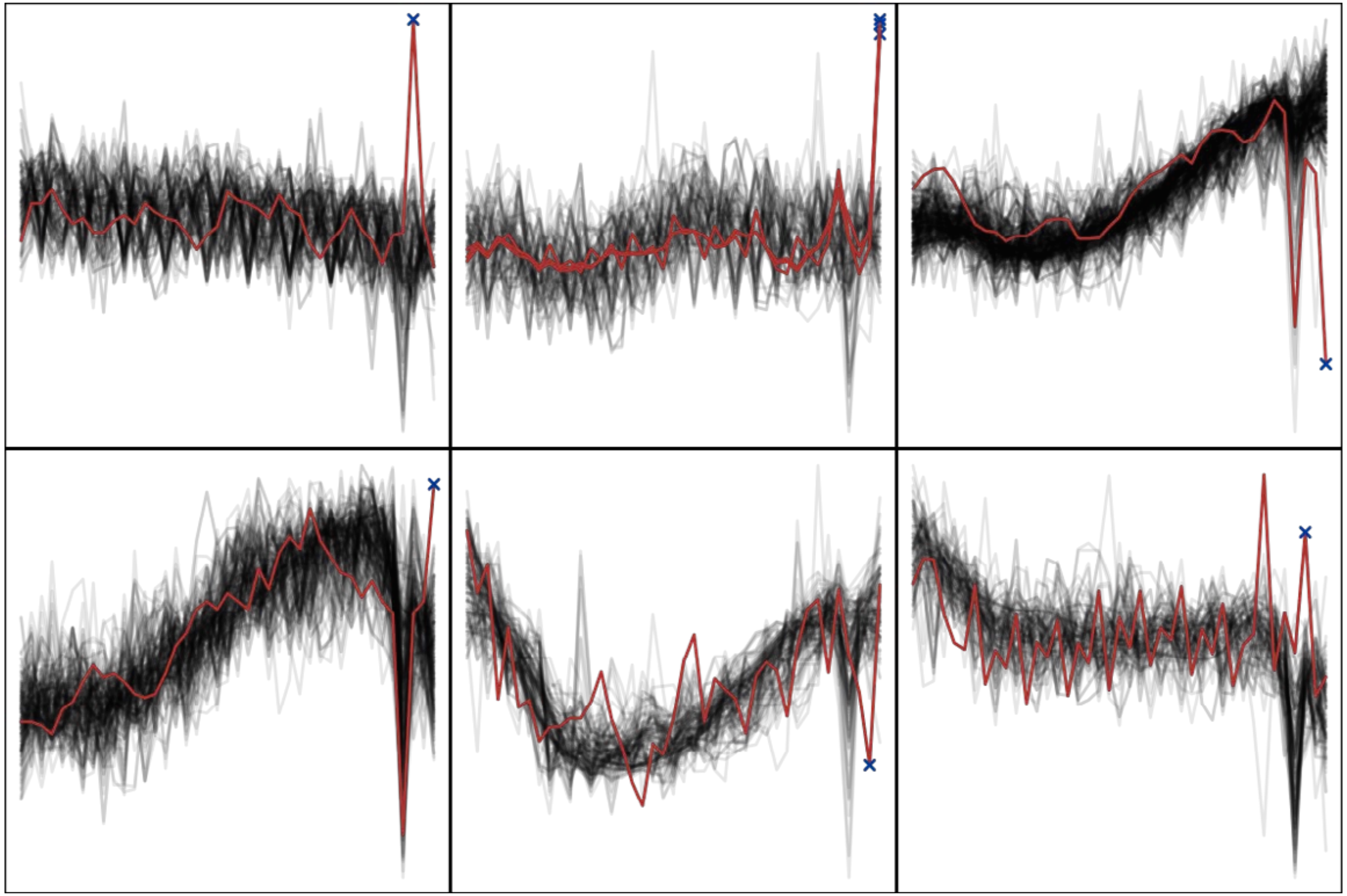


Example of 6 clusters found in the ENA data (scaled series)

Procedure (Cont'd)

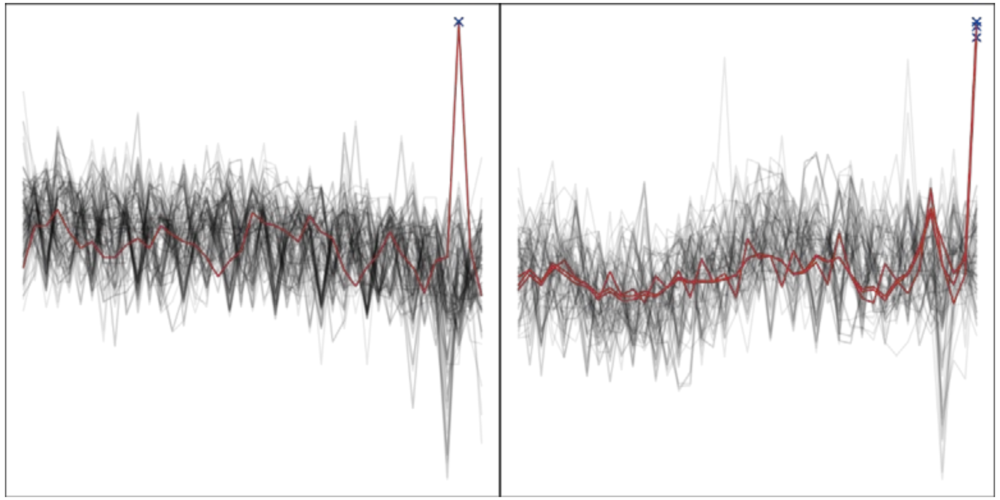
3. Detect the outlying data points in each cluster using the **DBSCAN** algorithm with the **Gower distance**

4. Investigate the flagged outliers!

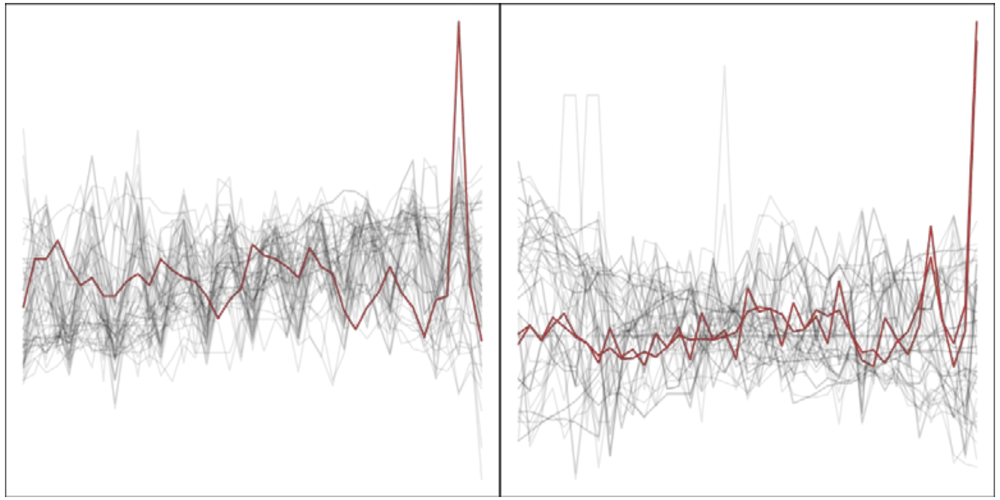


Example of 6 clusters and their outliers found in the ENA data (scaled series)

Results and Application



Example of 2 clusters and their outliers found in the ENA data (scaled series)



A and C sectors scaled series

- Used the data of the last 10 years
- 39 Clusters found
- 126 outliers found
- Spotted outliers are then investigated further and signalled to the data provider

Currently combined with a feature-additive ranking technique on the observations spotted as outliers.

Conclusion

- Outliers as observations with dynamics that differ the most from the general trend
- Robust to systemic spikes and breaks
- Data-driven, automated with few and adaptable parameters

Further improvement:

- Define the best period range to use
- Include different length time series
- Distance computation burden for Big Data cases

Thank you for your attention

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Anomaly detection methods and tools for big data[1]

Shir Kamenetsky Yadan,
Bank of Israel

[1] This presentation 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 central banks and other institutions represented at the event.

Anomaly Detection Methods and Tools for Big Data

Shir Kamenetsky Yadan,
Bank of Israel

October 2021

Abstract

Anomaly detection is the process of identifying observations in a dataset which deviate from the norm. In central banking, anomaly detection plays an essential role in managing, monitoring, and analysing data repositories. While traditional anomaly detection is manual, detecting anomalies in big data is humanely impossible. Consequently, the development of tools for mechanized and efficient detection of anomalous observations is becoming increasingly crucial for the ongoing work of database managers and researchers as real time big data repositories continue to expand at an accelerated rate.

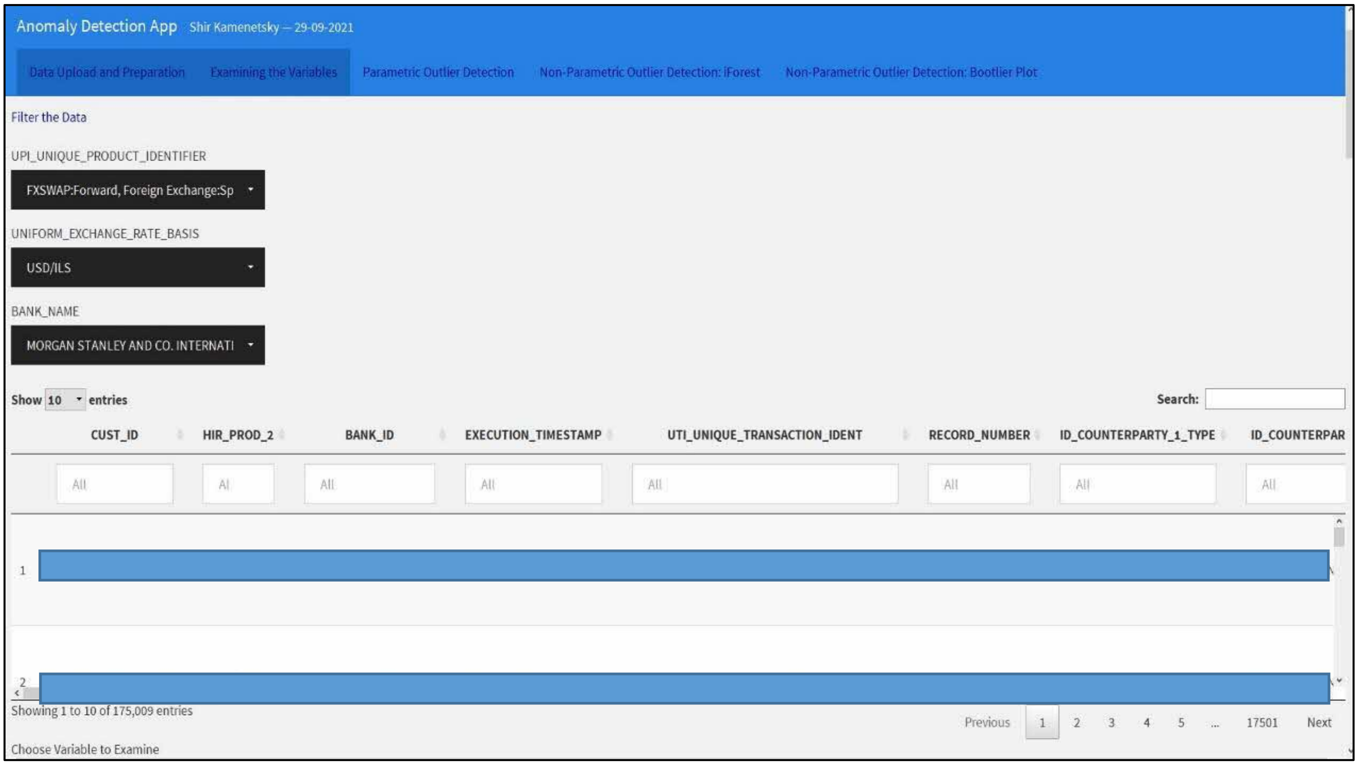
This paper presents a customized RShiny dashboard built using R's Flexdashboard format in Rmarkdown for user defined anomaly detection. The dashboard uses the Forex Data Repository which consists of daily transactions in foreign exchange derivatives and interest rates executed in the OTC market by financial intermediaries in Israel and abroad. These daily transactions accumulate to millions of records a year across 40 variables. In order to tackle the challenge of conducting quality control as well as analysing and extracting useful insights from a database of this size, we developed a tool for detecting anomalies which includes three main features; 1) Data upload and pre-processing, 2) Traditional anomaly detection, 3) Univariate and multivariate non-parametric anomaly detection .In the paper we expand on each one of these features and demonstrate their application. Some of the traditional methods we include are a visual examination of the data distribution and implementation of variance stabilizing and normalizing techniques such as Box-Cox transformation, subsequently applying standard deviation, median absolute deviation and interquartile range for detecting outliers. Economic series such as Forex transaction amounts are often characterized by highly right-skewed distributions making it difficult to use such traditional techniques.

Applying transformations to the data is not always enough to meet the symmetry or other parametric assumptions required by many of the traditional methods. In such circumstances a distribution-free test for outliers in data drawn from an unknown data generating process may give more reliable results. We apply two innovative non-parametric methods integrated into the third feature of the tool; a bootstrapping procedure for outlier detection Bootlier Plot (Singh and Xie, 2003) and Isolation Forests (Liu, Ting, and Zhou, 2009), and show their implementation on Forex data. Finally, we discuss the potential use of similar anomaly detection tools in different big data repositories such as the Central Credit Register and Payment Systems repository.

Contents

1. Data Upload, Pre-processing, and Exploration

Throughout the following sections we will analyse a dataset of about 290,000 Forex transactions from the second half of June 2020. The first step after uploading the dataset is choosing a variable to analyse and choosing desired filters as seen in Figure 1. Initial exploratory analysis as seen in Figures 2, 3, and 4 is important before proceeding to anomaly detection. This window of the tool offers tabular and visual exploration of the data which allows for familiarization with data distribution characteristics. This is necessary for choosing an appropriate anomaly detection method.


Anomaly Detection App
Shir Kamenetsky — 29-09-2021
Data Upload and Preparation
Examining the Variables
Parametric Outlier Detection
Non-Parametric Outlier Detection: iForest
Non-Parametric Outlier Detection: Bootlier Plot

Filter the Data

UPI_UNIQUE_PRODUCT_IDENTIFIER

FXSWAP:Forward, Foreign Exchange:Sp

UNIFORM_EXCHANGE_RATE_BASIS

USD/ILS

BANK_NAME

MORGAN STANLEY AND CO. INTERNATI

Show 10 entries

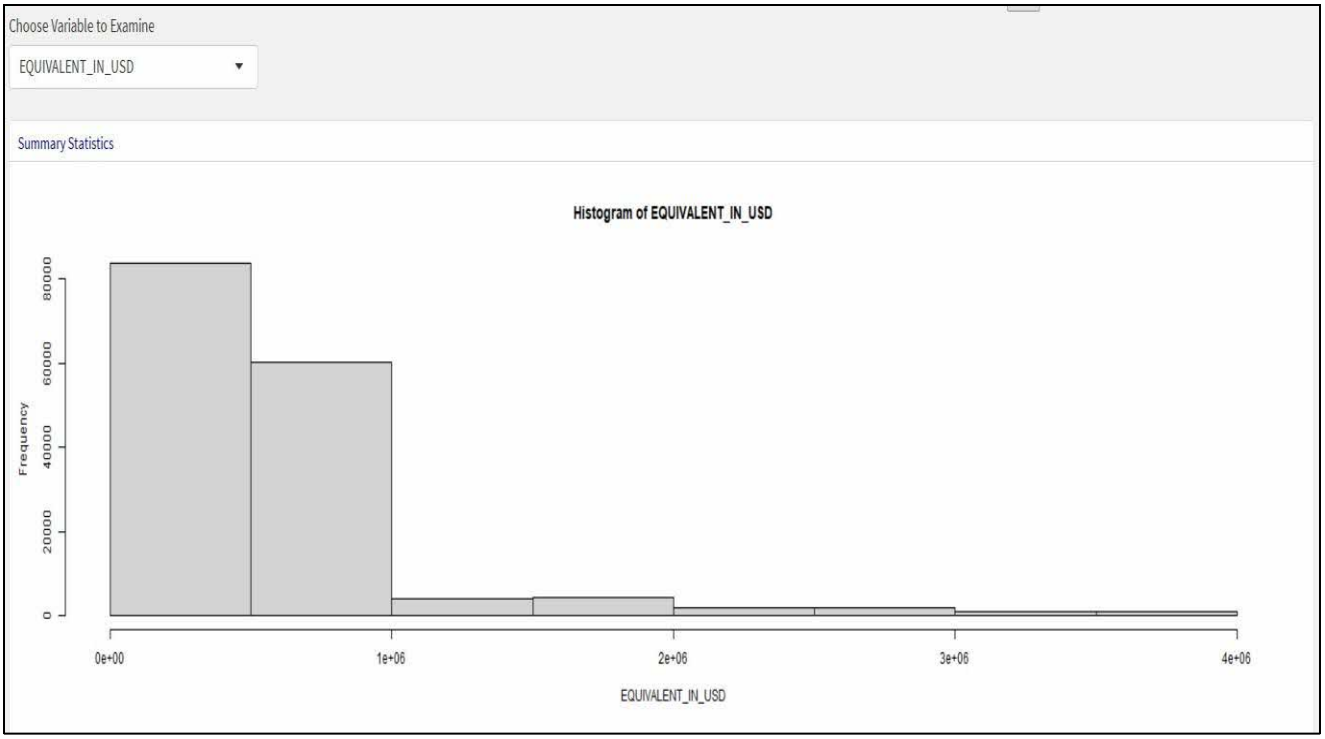
Search:

	CUST_ID	HIR_PROD_2	BANK_ID	EXECUTION_TIMESTAMP	UTI_UNIQUE_TRANSACTION_IDENT	RECORD_NUMBER	ID_COUNTERPARTY_1_TYPE	ID_COUNTERPAR
	All	Al	All	All	All	All	All	All
1								
2								

Showing 1 to 10 of 175,009 entries

Choose Variable to Examine

Figure 1: Filtering Window

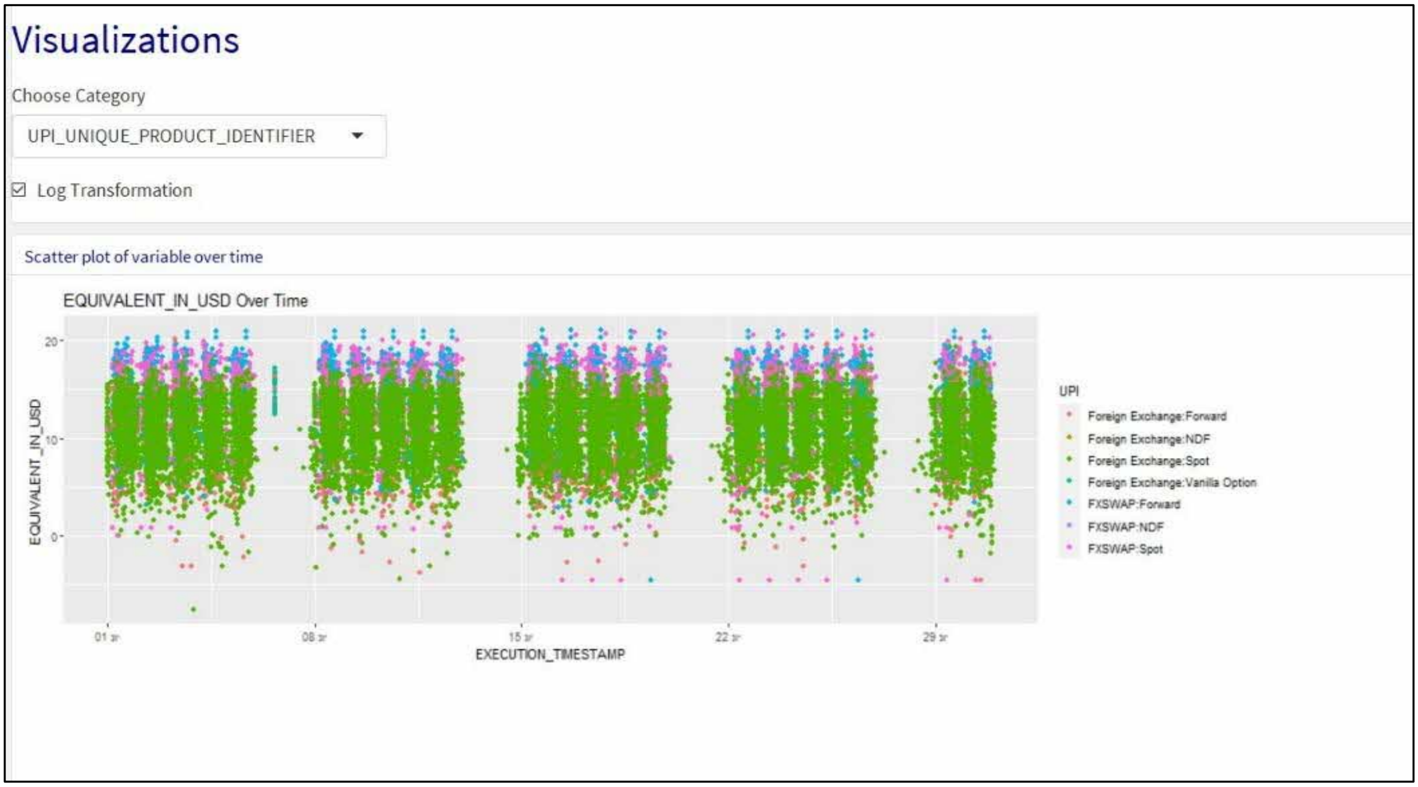


Choose Variable to Examine

EQUIVALENT_IN_USD

Figure 2: Histogram

In this example we choose to examine the variable "EQUIVALENT_IN_USD" which is the foreign exchange transaction amount in US dollars. We filter only the exchange basis, choosing transactions between US dollars and Israeli Shekel. We choose all UPI's (unique product identifiers) and all banks leaving us with about 175,000 entries. On the choice of variable window shown in Figure 2, a histogram appears giving us an idea of the distribution characteristics. Like in many financial data, the distribution has a strong right skew. This observation is important for further analysis. Figures 3 and 4 show additional visualizations of the data including a scatterplot of the chosen variable over time, with the option to choose a categorical variable for the colour of the points making the plot 3-dimentional. In this example each point is a single transaction sum with transaction time on the x-axis and colour by UPI. This plot can point out potential anomalies in the context of time. An additional plot called a ridge plot is shown in Figure 4, once again with the option of choosing a categorical variable to filter by. This time categories are separated on the y-axis while the x-axis shows the numerical variable of interest, in our case "EQUIVALENT _IN USD". What we learn from this plot is the distribution shape of "EQUIVALENT IN USD" for each of the different UPI's making it visually easy to compare between distributions of different UPI's. We notice, for example, that some UPI's have unimodal distributions while others are multi-modal.



Visualizations

Choose Category

UPI_UNIQUE_PRODUCT_IDENTIFIER ▾

☑ Log Transformation

Scatter plot of variable over time

Figure 3: Scatterplot

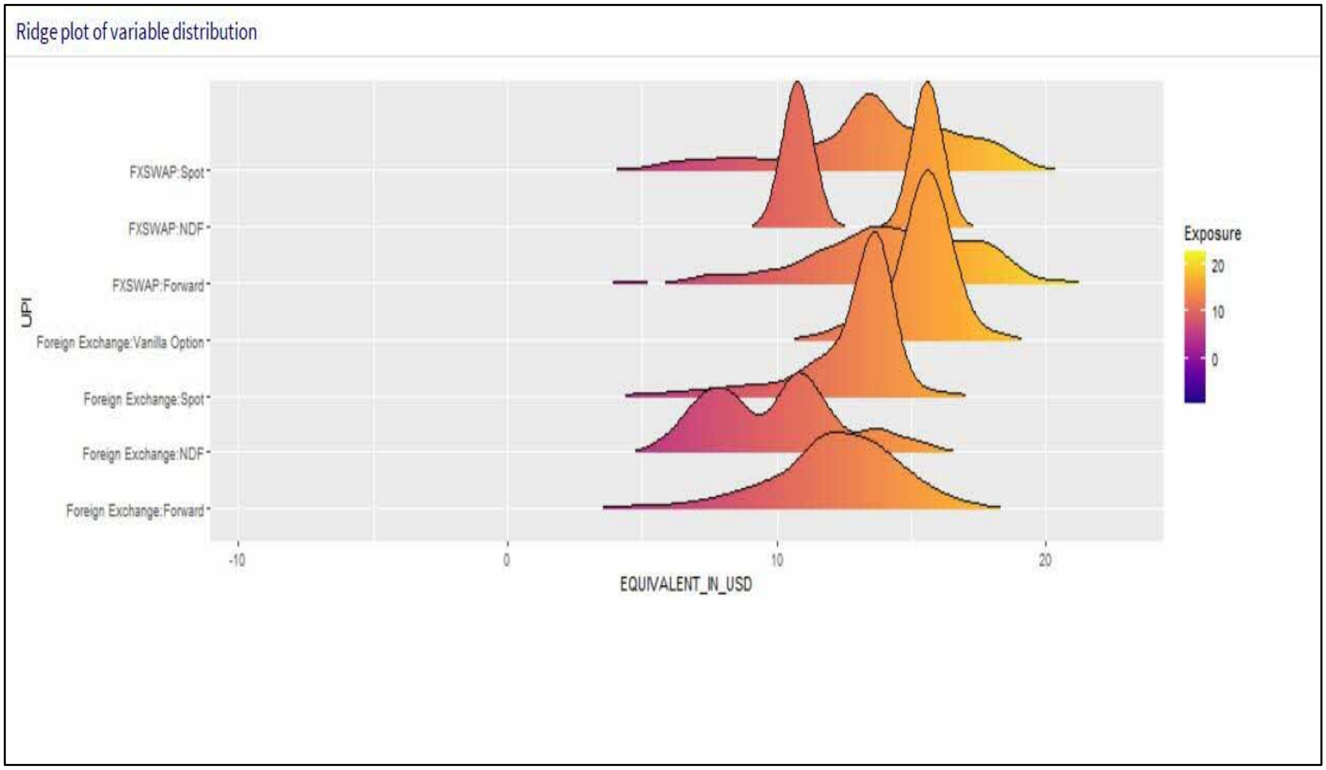


Figure 4: Ridge Plot

2. Traditional Methods

Traditional methods for identifying anomalies such as setting thresholds by taking a number of standard deviations from the mean are often most reliable when data meets certain parametric assumptions such as symmetry or normality of the distribution. We have seen in the previous section that our data, much like any financial data, is neither normal nor symmetric. In order to use traditional methods for anomaly detection on such data we begin by applying transformations to attempt to bring the data distribution to a more symmetric and normal shape. Two transformations are offered in this tool; Natural Log transformation and Box-Cox transformation.

Box-Cox Transformation

Box-Cox attempts to approximate the normal distribution.

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0 \\ \ln y_i \text{ if } \lambda = 0 \end{cases}$$

Where $y_i^{(\lambda)}$ is the Box-Cox transformed data and the optimal λ is one which results in best approximation of normal distribution curve.

After transforming the data, threshold values for anomalous data are calculated using 4 different choices of center and variability metrics; 1) Mean and Standard Deviation, 2) Median and Median Absolute Deviation (MAD), 3) Median and Double MAD, 4) Inter-Quartile Range.

Mean and Standard Deviation

Mean and Standard deviation is common practice and the most parametric as well as non-robust method. It is most reliable with normally distributed data. In the case of normally distributed data, 3 standard deviations taken from both sides of the mean will cover 99% of the data and any observations outside of these thresholds can be considered anomalies.

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

$$Anomaly\ Threshold = Mean\ \pm\ \alpha * SD$$

Median and MAD

The Median and MAD method is a non-parametric method and robust in two main aspects; the first being the use of the median as a measure of centrality which is robust to outliers in itself, and once again the use of the median for aggregation in the MAD measure of variability. Additionally, unlike standard deviation which squares the deviations of each observation from the mean causing larger deviations to explode, MAD uses absolute value which minimizes the effect of large deviations and thus adds to robustness.

$$MAD = k * median(|Y_i - median(Y\,)|)$$

$$Anomaly\ Threshold = Median \pm \alpha * MAD$$

Where k is called the scale factor and is taken to be 1.4826 if data is normally distributed. In this case MAD can be used as a consistent estimator for the estimation of the standard deviation. Since the distribution of our data is unknown we take k to be 1.

Median and Double MAD

The classic Median and MAD method defines a symmetric interval of anomaly thresholds around the median. This works best when the distribution is indeed symmetric. In cases like ours where the distribution is heavily skewed, often even after applying a transformation, there is the Double MAD measure which calculates two separate MAD values for the left and right sides of the distribution (using median as the center). [3]

$$\tilde{Y} = median(Y)$$

$$Y^{(u)} = \left\{y|y \in Y \bigcap y \geq \tilde{Y}\right\}$$

$$MAD^{(u)}(Y) = k * median(\left|Y_i^{(u)} - \tilde{Y}\right|)$$

Where one again k is the scale factor. u indicates "upper" distribution observations -those which are greater or equal to the median. The same calculations are done using lower observations for $MAD^{(l)}(Y)$. We now take α upper deviations and α lower deviations from the median to get upper and lower anomaly thresholds.

$$Lower\ Threshold = median - \alpha * MAD^{(l)}(Y)$$

$$Upper\ Threshold = median + \alpha * MAD^{(u)}(Y)$$

IQR and Tukey's Fences

Lastly, the anomaly detection tool gives the option of using IQR as a measure of variability with Tukeys' Fences to calculate anomaly thresholds. This method is nonparametric and robust.

$$IQR = Q_3 - Q_1$$

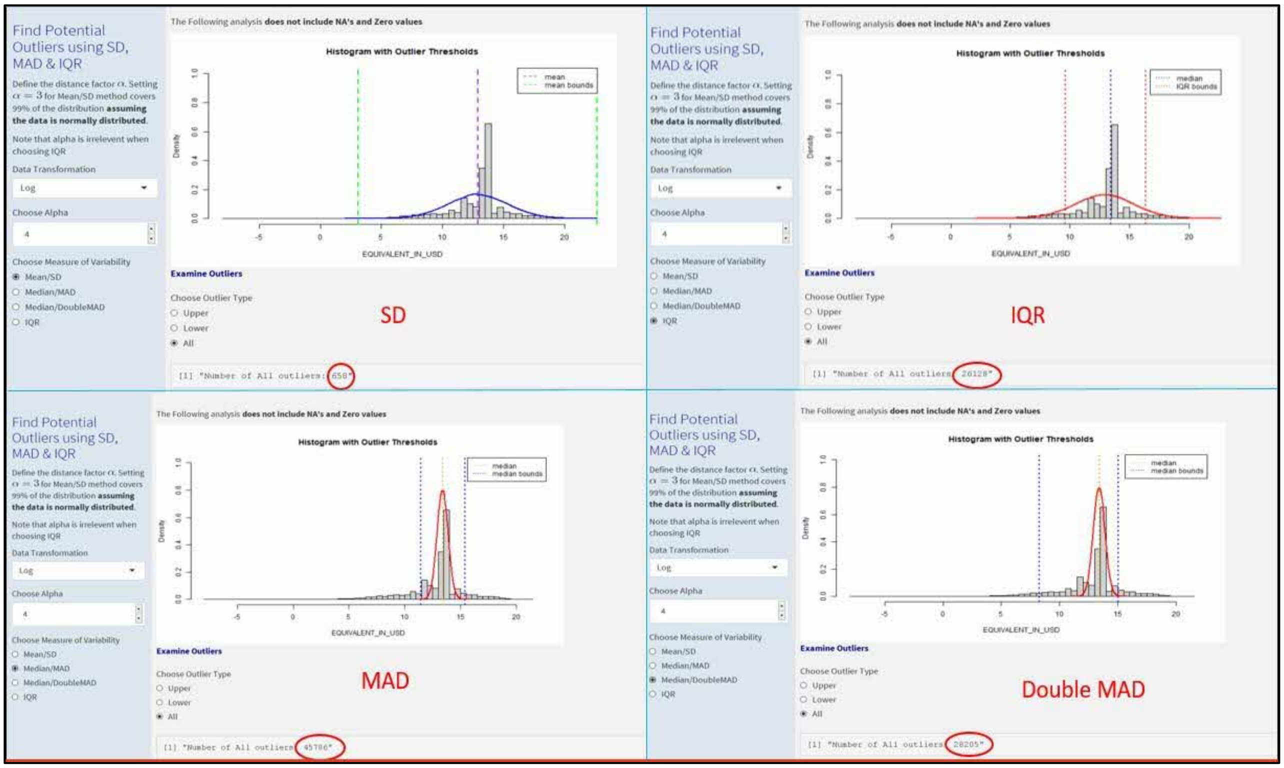
Where Q3 is the value that holds 25% of the values above it and Q1 is the value that holds 25% of the values below it.

$$Lower\ Threshold = Q_1 - \alpha * IQR$$

$$Upper\ Threshold = Q_3 + \alpha * IQR$$

In this case, α is taken to be 1.5 and the choice of alpha in the anomaly detection tool is disabled.

Figure 5 shows a comparison of all four methods on our example dataset. Notice that in this window of the anomaly detection tool the user chooses a transformation method, measure of centrality and variability pair, number of deviations to take from the center (alpha), and whether they want to consider upper, lower, or all outliers. A histogram is then plotted with dashed vertical lines showing the center value and anomaly thresholds and a kernel density curve of the normal distribution using either the mean and standard deviation or the median and MAD to generate observations from the normal distribution. In this example we choose alpha to be 4. We see how standard deviation thresholds are placed much further from the center than thresholds of the other measures. This occurs due to the lack of robustness of the standard deviation. The variability is affected by the values at the tails of the distribution more so than the robust measures. Notice the symmetric thresholds of the standard deviation and MAD methods versus the non-symmetric thresholds of Double MAD and IQR which are pulled further away from the center on the left side of the distribution due to the longer left tail. Choosing the method and size of alpha are ultimately up to the user. It is important that the user has an expertise in the field and good familiarity and knowledge of the data. The user can then decide which method gives the most accurate results and whether the flagged observations are indeed anomalous points.



Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability

● Mean/SD
○ Median/MAD
○ Median/DoubleMAD
○ IQR

The Following analysis **does not include NA's and Zero values**

Examine Outliers

Choose Outlier Type

○ Upper
○ Lower
● All

```
[1] "Number of All outliers: 658"
```

SD

Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability

○ Mean/SD
○ Median/MAD
○ Median/DoubleMAD
● IQR

The Following analysis **does not include NA's and Zero values**

Examine Outliers

Choose Outlier Type

○ Upper
○ Lower
● All

```
[1] "Number of All outliers: 26128"
```

IQR

Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability

○ Mean/SD
● Median/MAD
○ Median/DoubleMAD
○ IQR

The Following analysis **does not include NA's and Zero values**

Examine Outliers

Choose Outlier Type

○ Upper
○ Lower
● All

```
[1] "Number of All outliers: 45786"
```

MAD

Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability

○ Mean/SD
○ Median/MAD
● Median/DoubleMAD
○ IQR

The Following analysis **does not include NA's and Zero values**

Examine Outliers

Choose Outlier Type

○ Upper
○ Lower
● All

```
[1] "Number of All outliers: 28205"
```

Double MAD

Figure 5: Anomaly Detection

3. Nonparametric Univariate and Multi-Variate Methods

We may prefer to use nonparametric methods on data drawn from an unknown data generating process rather than traditional methods on transformed data. We offer two distribution-free methods for anomaly detection; 1) Isolation Forest and 2) Bootlier Plot. Both methods have multivariate implementation; however, we do not currently offer multivariate Bootlier Plot in our anomaly detection tool. In this section we will explain the methods and demonstrate their use in the anomaly detection tool.

3.1. Isolation Forest (Liu, Ting, and Zhou, 2009)

The basic idea of this method is to isolate anomalies rather than profiling normal instances. [2] Anomalies are “few” and “different” making them more susceptible to isolation than normal points. Isolation is conducted via construction of tree structure. Partitions are generated by randomly selecting a variable and then randomly selecting a split value between the maximum and minimum values of selected variable. This is done iteratively until each observation is isolated to its own node. Path length is equivalent to the number of partitions required to isolate a point, or in tree structure the number of edges from root node to external terminating node. Observations that are quicker to isolate and have shorter path lengths are ones considered to be anomalies. We see an example of this in Figure 6 where x_0 is an obvious anomaly placed far from the other points and is therefore isolated with fewer partitions.

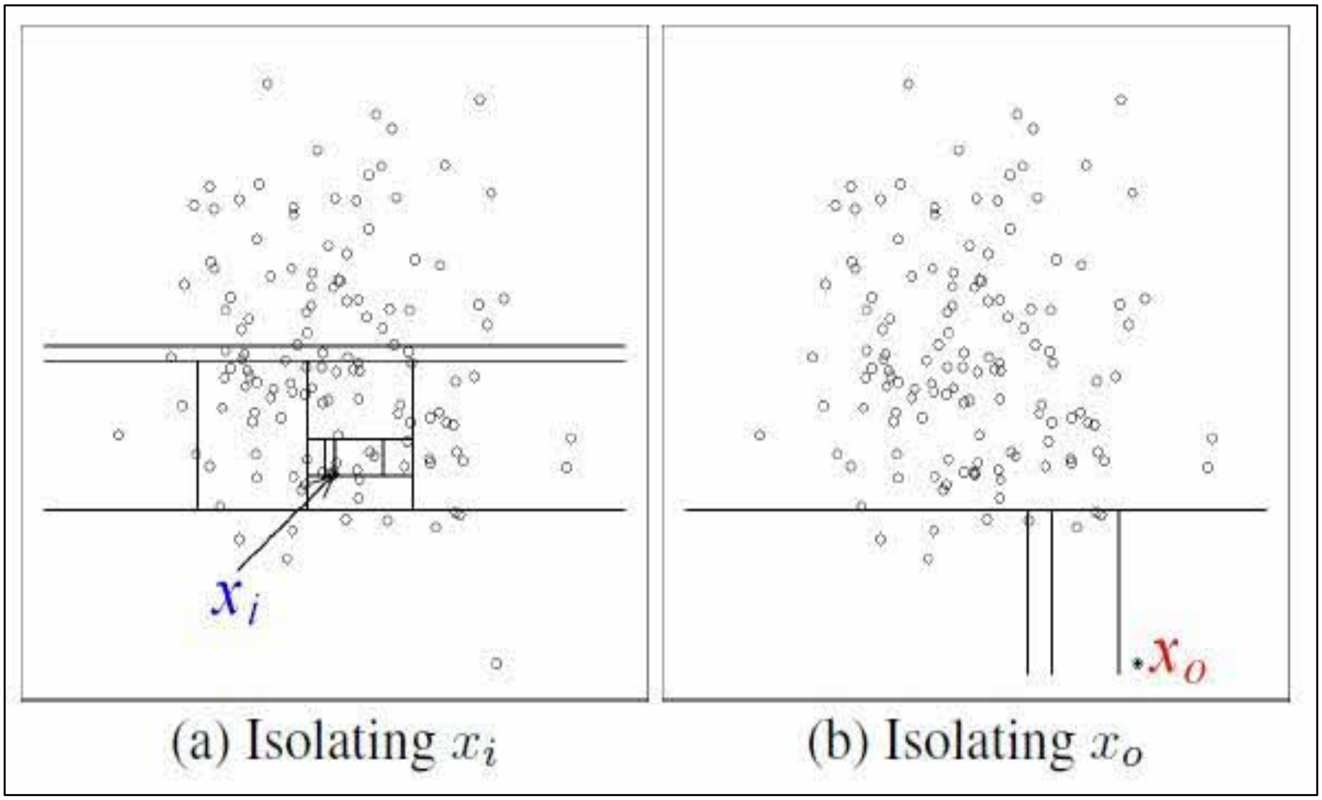


Figure 6: Isolation Forest

These steps are applied to subsets of the data on an ensemble of trees called an Isolation Forest. Path lengths for each observation are aggregated over the trees and a final anomaly score is given to each observation. This score is between 0 and 1 where scores close to 1 are considered anomalies and scores smaller than 0.5 can be considered regular instances. In figure 7 we see a visual example of an isolation forest with anomalous points closer to the root of the tree.

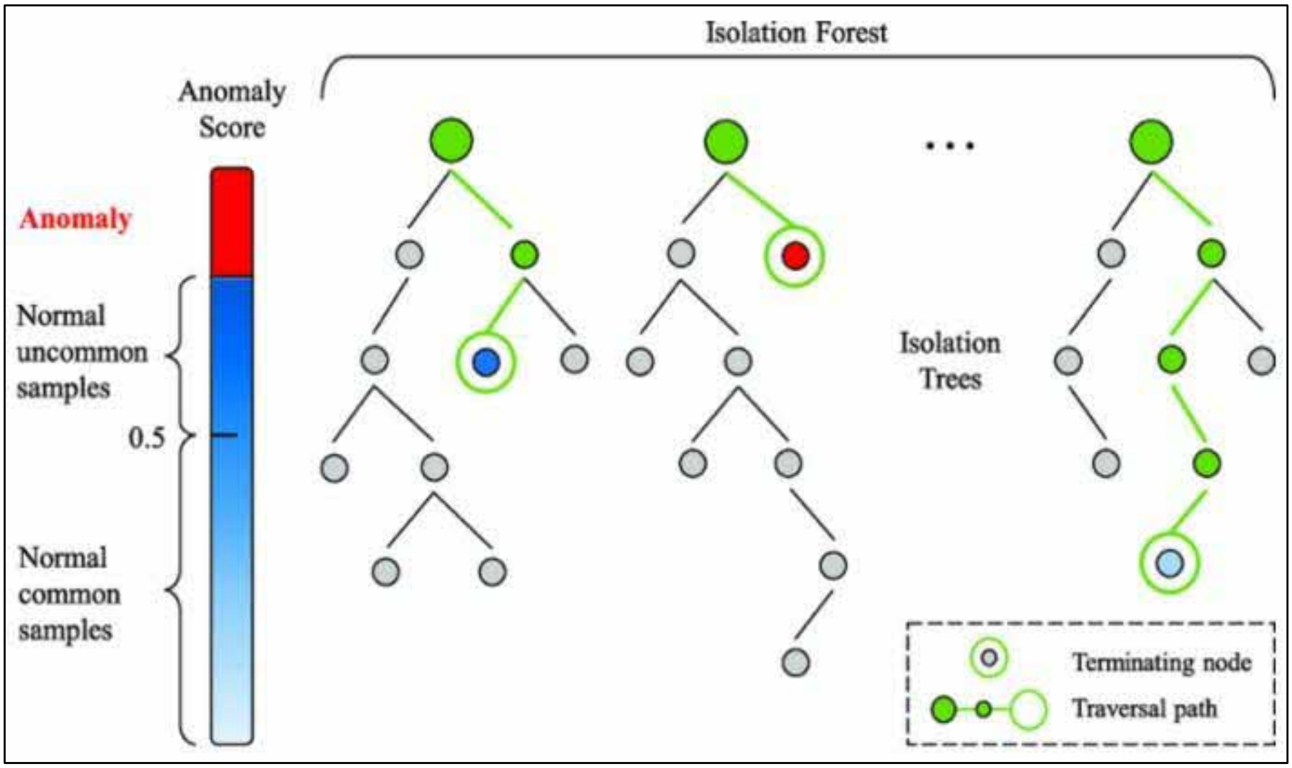
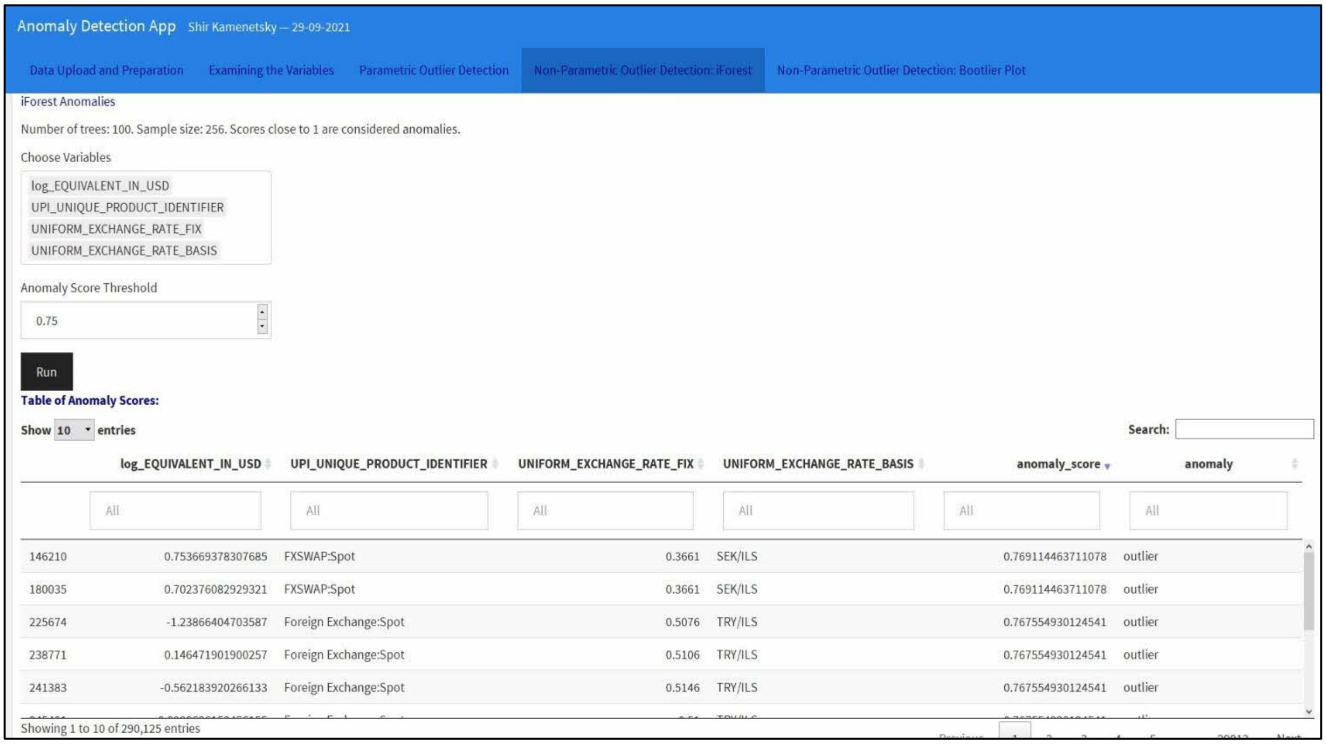


Figure 7: Isolation Forest

A major benefit of Iforest over other unsupervised anomaly detection methods resides in the use of sub-sampling with relatively small samples for each tree. Sub-sampling is conducted by random selection of instances without replacement. Building Isolation trees on smaller samples of data reduces the swamping and masking effects which are common in other anomaly detection methods[1]. Swamping occurs when normal observations are too close to anomalous points and are wrongly classified as anomalies. In other words the normal points are ”swamped” by the anomalies. In masking, a group of anomalous observations close together ”mask” their own presence and they are classified as normal points. These effects are common especially in very large datasets. For this reason Iforest is an especially suitable method for anomaly detection in our large data repositories.

Figure 8 demonstrates the use of Iforest in our anomaly detection tool on our example data. The user may choose multiple variables for analysis, in this example transaction amount, UPI, exchange rate, and coin are chosen. The user then chooses an anomaly threshold - the anomaly score for which any value above this score will be considered an anomaly. In this example an observation is classified as an anomaly if it receives an anomaly score over 0.75. The user can then run the algorithm and view a table of all points and their classification as anomalous or normal, a quantile table of the anomaly scores, and the final number of anomalies in the data.

[1] For extended explanation of how Iforest handles swamping and masking see Liu, Ting, and Zhou 2008



Anomaly Detection App
Shir Kamenetsky – 29-09-2021
Data Upload and Preparation
Examining the Variables
Parametric Outlier Detection
Non-Parametric Outlier Detection: iForest
Non-Parametric Outlier Detection: Bootlier Plot
Run
Search:

iForest Anomalies

Number of trees: 100. Sample size: 256. Scores close to 1 are considered anomalies.

Choose Variables

log_EQUIVALENT_IN_USD
UPI_UNIQUE_PRODUCT_IDENTIFIER
UNIFORM_EXCHANGE_RATE_FIX
UNIFORM_EXCHANGE_RATE_BASIS

Anomaly Score Threshold

0.75

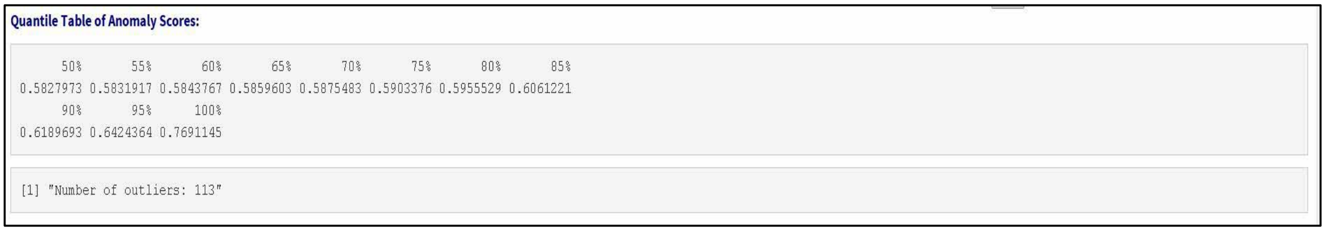
Table of Anomaly Scores:

Show 10 entries

	log_EQUIVALENT_IN_USD	UPI_UNIQUE_PRODUCT_IDENTIFIER	UNIFORM_EXCHANGE_RATE_FIX	UNIFORM_EXCHANGE_RATE_BASIS	anomaly_score	anomaly
	All	All	All	All	All	All
146210	0.753669378307685	FXSWAP:Spot	0.3661	SEK/ILS	0.769114463711078	outlier
180035	0.702376082929321	FXSWAP:Spot	0.3661	SEK/ILS	0.769114463711078	outlier
225674	-1.23866404703587	Foreign Exchange:Spot	0.5076	TRY/ILS	0.767554930124541	outlier
238771	0.146471901900257	Foreign Exchange:Spot	0.5106	TRY/ILS	0.767554930124541	outlier
241383	-0.562183920266133	Foreign Exchange:Spot	0.5146	TRY/ILS	0.767554930124541	outlier

Showing 1 to 10 of 290,125 entries

Figure 8: Isolation Forest



Quantile Table of Anomaly Scores:

```
      50%       55%       60%       65%       70%       75%       80%       85% 
0.5827973 0.5831917 0.5843767 0.5859603 0.5875483 0.5903376 0.5955529 0.6061221 
      90%       95%      100% 
0.6189693 0.6424364 0.7691145 
```

```
[1] "Number of outliers: 113"
```

Figure 9: Isolation Forest - Quantile Table

3.2. Bootlier Plot (Singh and Xie, 2003)

The Bootler Plot method [4] is based on bootstrapping. When an outlier exists in a dataset, some bootstrap samples will contain the outlier while others will not. The presence of an outlier is expected to cause a significant increase or decrease in the bootstrap mean, and make the bootstrap distribution of the sample mean a mixture distribution. Therefore, we expect the histogram of the sample mean to be multimodal. In order to make the bootstrap histogram more sensitive to a potential outlier, the chosen bootstrap statistic is the "mean – trimmed mean" [2]. Where the trimmed mean is the mean of the bootstrap sample after trimming k observations from each side of the sorted sample. **Mean-Trimmed Mean Statistic:**

$$T(Y^*) = \frac{1}{n}\sum_{1}^{n} Y_i^* - \frac{1}{n-2k}\sum_{k+1}^{n-k} Y_{(i)}^*$$

Where $T(Y^*)$ is the mean-trimmed mean statistic, $Y_1^*, Y_2^*, ..., Y_n^*$ denote bootstrap draws from a certain bootstrap, and Y_i^* the corresponding order statistics.

In Figures 10 and 11 we see the Bootlier Plots of data, and the same data with an additional anomalous observation. Figure 11, with the anomalous point has an obvious additional "bump". Bootlier plot becomes less practical when data is large and many potential outliers exist. Singh and Xie refer to this issue in their paper. The probability of having a bootstrap sample of a large size, *n*, free from potential outliers is very small and we may not see a clear bump in Bootlier plot. The solution is to reduce the bootstrap sample size to a fraction of *n*, i.e., [αn], α∈(0,1]. A practical

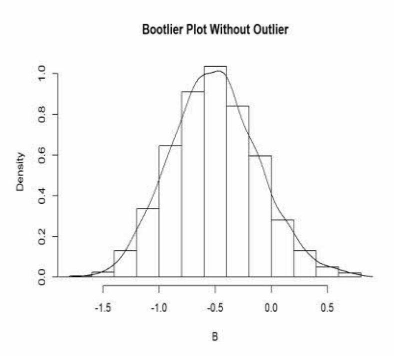


Figure 10:

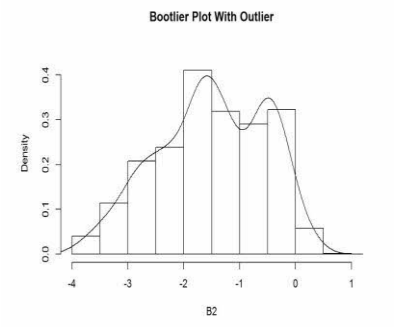


Figure 11:

[2] The reasoning for greater sensitivity of "mean-trimmed mean" statistic to potential outliers is explained in further detail in Singh and Xie, 2003

recommendation is to look at several Bootlier plots at different bootstrap sample sizes – if any one is found bumpy that would be indicative of the presence of outliers. We find large data size to be an issue when using Bootlier Plot on our data. Even when following the recommendations of the article for large amounts of data we typically get unimodal histograms. Since we do not have a set of ″real″ anomalies in our data to compare with, it is difficult to know whether this outcome stems from the fact that there really are no anomalies in the data or is a technical outcome of the algorithm such as the one discussed regarding very large data sets. In any case we offer the user the flexibility of choosing the number of desired bootstraps, sample size, and trimming amount[3] and view Bootlier Plots for each parameter trio. Another problem with large datasets in this method is a slower running time. It may be time-impractical to try very many parameter combinations especially when choosing large numbers of bootstraps.

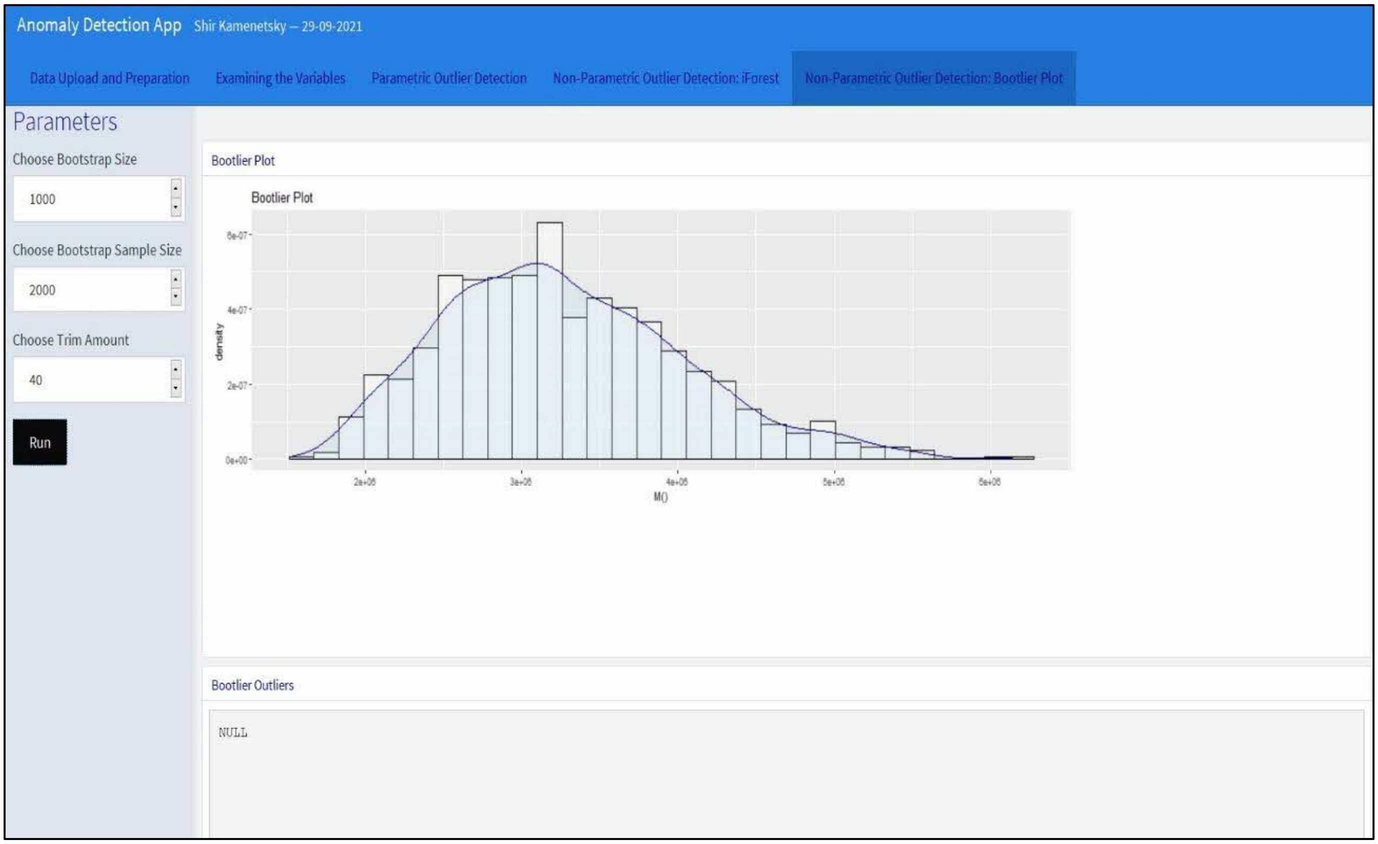
The Bootlier Plot indicates whether a dataset holds anomalies or not, however; it does not identify and give the values of these outliers. Candelon and Metiu, 2013 [1] extend on the Bootlier Plot in the Deutsche Bundesbank Discussion Paper – "A distribution-free test for outliers" developing a method for identifying outliers from the Bootlier Plot. They term this method the ″Bootlier Test". The method uses a two-step process:

1. Test for multimodality: **H0** – Bootlier plot has precisely one mode (and no local minimum), **H1** – Bootlier plot has more than one mode. Test hypothesis using "Bootlier Test" - Bootlier plot coupled with distribution free test for multimodality proposed by Silverman (1981)

2. Identify Outliers: 1) Build subsamples by sequentially cancelling observations from the tails of the original sample ordered in ascending order. 2) Perform Bootlier test on each ordered subsample until the null hypothesis of unimodality cannot be rejected for a particular subset of observations. 3) Data points not contained in this subset are the anomalies.

In figures 12 and 13 we demonstrate the use of Bootlier Plot and Bootlier Test on our example data. In the first example we choose 1000 bootstraps of sample size 2000 (recall that our data size is 175,000 so this is about 1% of the data) and trim amount of 10. With these parameters the Bootlier Plot is

[3] The writers mention that the optimal choice of trim amount is not given theoretical groundwork in the paper, we therefore leave this open for the users to try different sizes



Anomaly Detection App
Shir Kamenetsky — 29-09-2021
Data Upload and Preparation
Examining the Variables
Parametric Outlier Detection
Non-Parametric Outlier Detection: iForest
Non-Parametric Outlier Detection: Bootlier Plot
Parameters
Run
Bootlier Plot
Bootlier Plot
density
6e-07
4e-07
2e-07
0e+00
2e+06
3e+06
4e+06
5e+06
6e+06
M()
Bootlier Outliers

Choose Bootstrap Size: 1000

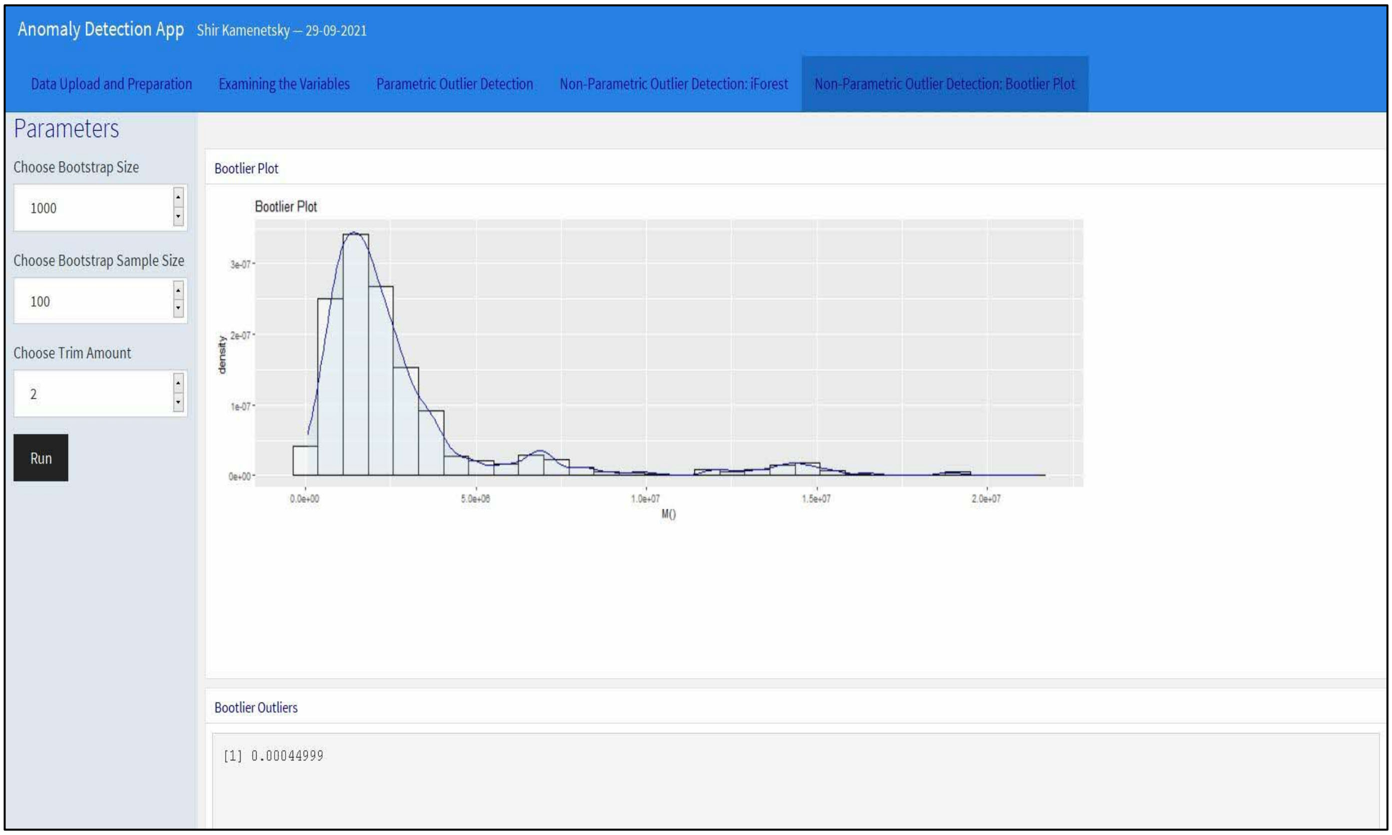
Choose Bootstrap Sample Size: 2000

Choose Trim Amount: 40

```
NULL
```

Figure 12: Bootlier Plot

unimodal and there is no evidence of anomalies. Likewise, no anomalies are found using the Bootlier Test. In the second example, we change thee sample size to 100 (about .06% of the data) and trim amount to 2. Now we see a multimodal Bootlier Plot along with one anomaly found by the Bootlier Test. In comparison to the Iforest method as well as the traditional methods we implemented it is evident that the Bootlier Plot and Test find dramatically less anomalies in the data. We must ”force” anomalies out of the method by using extreme parameters. This could indicate that we do not have anomalous points in our data, or be caused by a technicality of the method - perhaps because of our data’s large size as we mentioned before. For this reason, as well as greater processing power required for running the Bootlier Plot especially for larger bootstrap sizes, we find this method to be less practical and less reliable for our data than the other methods offered in the anomaly detection tool.


Anomaly Detection App
Shir Kamenetsky – 29-09-2021
Data Upload and Preparation
Examining the Variables
Parametric Outlier Detection
Non-Parametric Outlier Detection: iForest
Non-Parametric Outlier Detection: Bootlier Plot
Parameters
Run
Bootlier Plot
Bootlier Plot
3e-07
2e-07
1e-07
0e+00
density
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
M()
Bootlier Outliers

Choose Bootstrap Size: 1000

Choose Bootstrap Sample Size: 100

Choose Trim Amount: 2

```
[1] 0.00044999
```

Figure 13: Bootlier Plot

4. Concluding Remarks

The anomaly detection tool for big data enhances the efficiency of the ongoing work of database managers. It gives database managers the ability to filter large amounts of data, study data characteristics and distributions via tables and graphs, and signal suspicious observations with the flexibility of choosing between multiple anomaly detection methods, rather than scroll through enormous excel spreadsheets and eyeball data values. It is important to keep in mind and to emphasize to database managers that these methods alone cannot tell them whether an observation is an anomaly or not, but rather point their attention to suspicious observations. The expertise of the database manager is in the field and experience and knowledge with the data is crucial in deciding whether an observation is indeed anomalous, or not. This tool is made relatively generic and can be implemented on other data repositories with few adjustments. Besides the Forex repository, we are currently working on adapting the anomaly detection tool to the Central Credit Register and the Payment Systems repository.

References

[1] Bertrand Candelon and Norbert Metiu. "A distribution-free test for outliers". In: (2013). url: http://hdl.handle.net/10419/68604.

[2] Fei Tony Liu, Kai Ting, and Zhi-Hua Zhou. "Isolation Forest". In: (Jan. 2009), pp. 413–422. doi: 10.1109/ICDM.2008.17.

[3] Peter Rosenmai. *Using the Median Absolute Deviation to Find Outliers*. url: https://eurekastatistics.com/using-the-median-absolutedeviation-to-find-outliers/. (accessed: 10.11.2021).

[4] Kesar Singh and Minge Xie. "Bootlier-Plot: Bootstrap Based Outlier Detection Plot". In: *Sankhy¯a: The Indian Journal of Statistics (2003-2007)* 65 (2003), pp. 532–559. doi: 10.2307/25053287.

Anomaly Detection Methods and Tools for Big Data

SHIR KAMENETSKY

BANK OF ISRAEL

OCTOBER 22, 2021

Road Map

1. Motivation
2. Data exploration: pre-anomaly detection
3. Traditional Methods
4. Non-Parametric Univariate and Multivariate Methods

Motivation

Role of anomaly detection in central banking:

Role 1

- Managing and monitoring data repositories
  - Quality assurance – find erroneous data observations
  - Alert for sudden changes in economy, deviations from trends

Role 2

- Analyzing data repositories
  - Gaining greater familiarity and deeper understanding of data

Motivation

PROBLEM

Real time **big data** repositories continue to expand at an **accelerated rate**.

◆

Traditional **manual** anomaly detection is **humanely impossible**.

Motivation

SOLUTION

Development of **mechanized** and **efficient** tools for anomaly detection.

No more spreadsheets!

Motivation

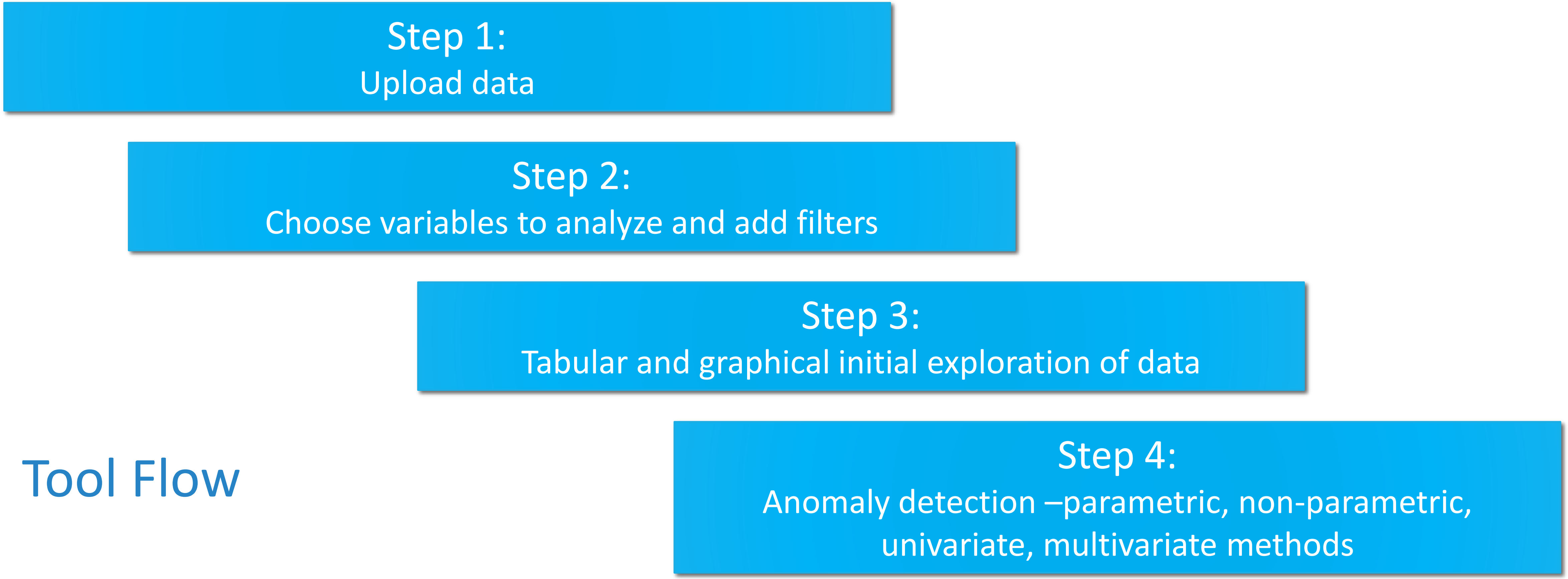
Tool: Anomaly Detection Dashboard App

Data: Forex data repository:

- ○ Daily transactions in foreign exchange derivatives and interest rates executed in OTC market by financial intermediaries in Israel and abroad.
- ○ Millions of records a year across 40 variables

Technologies: Shiny from R Studio flexdashboard R Markdown

Motivation



Step 1:
Upload data

Step 2:
Choose variables to analyze and add filters

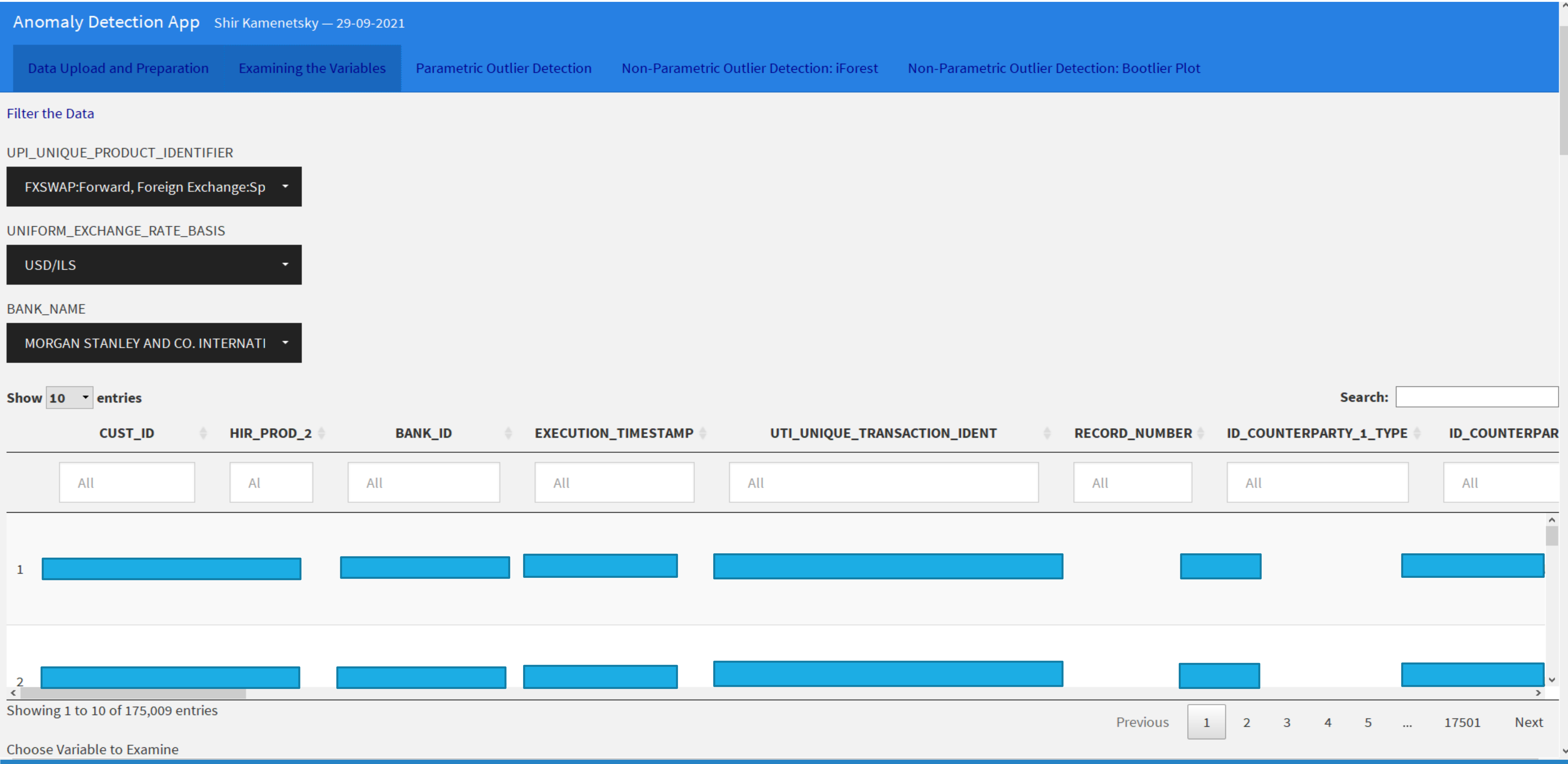
Step 3:
Tabular and graphical initial exploration of data

Step 4:
Anomaly detection –parametric, non-parametric, univariate, multivariate methods

Tool Flow

Data exploration: pre-anomaly detection

Data exploration: pre-anomaly detection



Anomaly Detection App   Shir Kamenetsky — 29-09-2021

Data Upload and Preparation
Examining the Variables
Parametric Outlier Detection
Non-Parametric Outlier Detection: iForest
Non-Parametric Outlier Detection: Bootlier Plot

Filter the Data

UPI_UNIQUE_PRODUCT_IDENTIFIER: FXSWAP:Forward, Foreign Exchange:Sp

UNIFORM_EXCHANGE_RATE_BASIS: USD/ILS

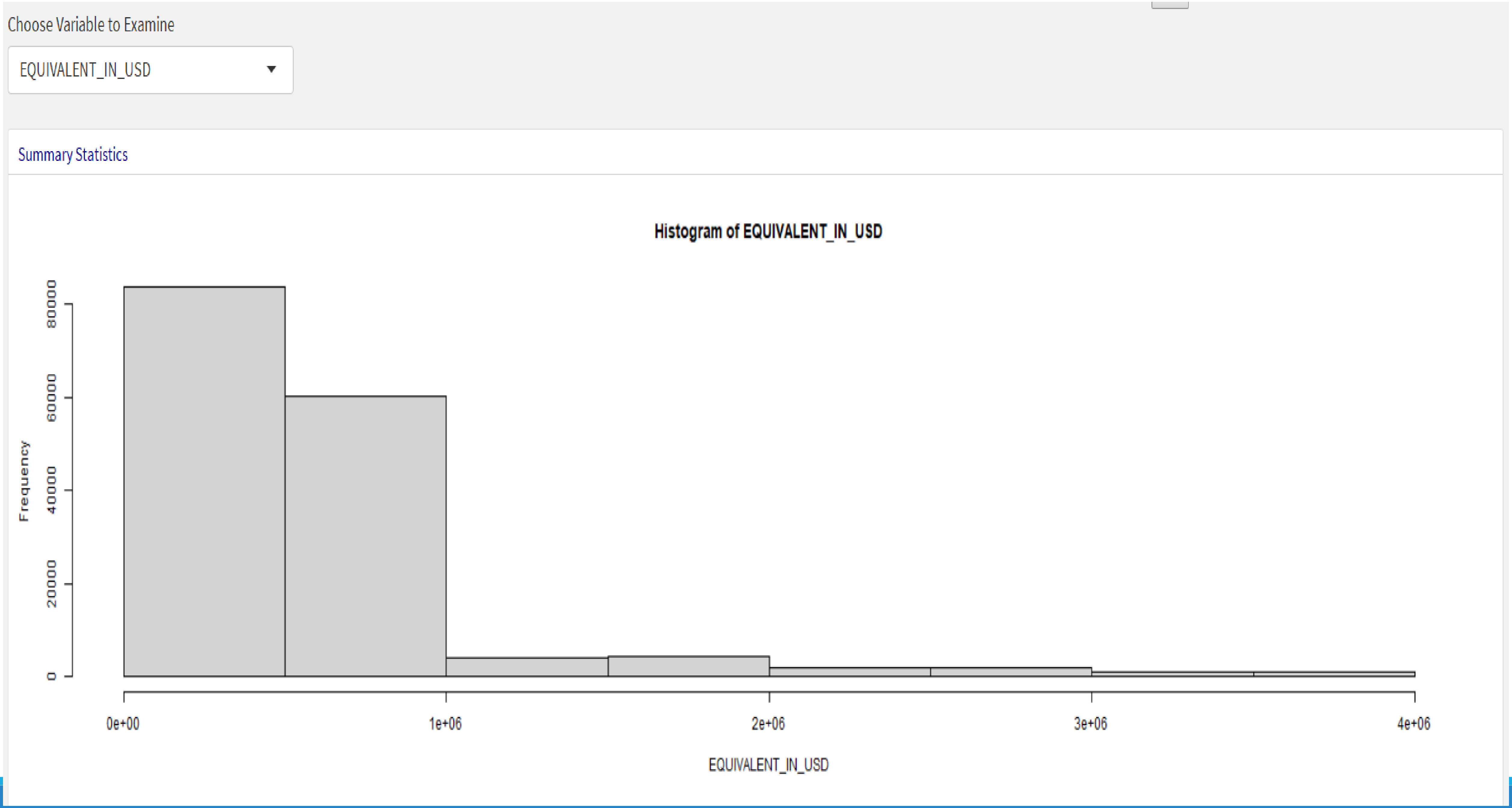
BANK_NAME: MORGAN STANLEY AND CO. INTERNATI

Show 10 entries   Search:

	CUST_ID	HIR_PROD_2	BANK_ID	EXECUTION_TIMESTAMP	UTI_UNIQUE_TRANSACTION_IDENT	RECORD_NUMBER	ID_COUNTERPARTY_1_TYPE	ID_COUNTERPAR
	All	Al	All	All	All	All	All	All
1	[redacted]		[redacted]	[redacted]	[redacted]	[redacted]		[redacted]
2	[redacted]		[redacted]	[redacted]	[redacted]	[redacted]		[redacted]

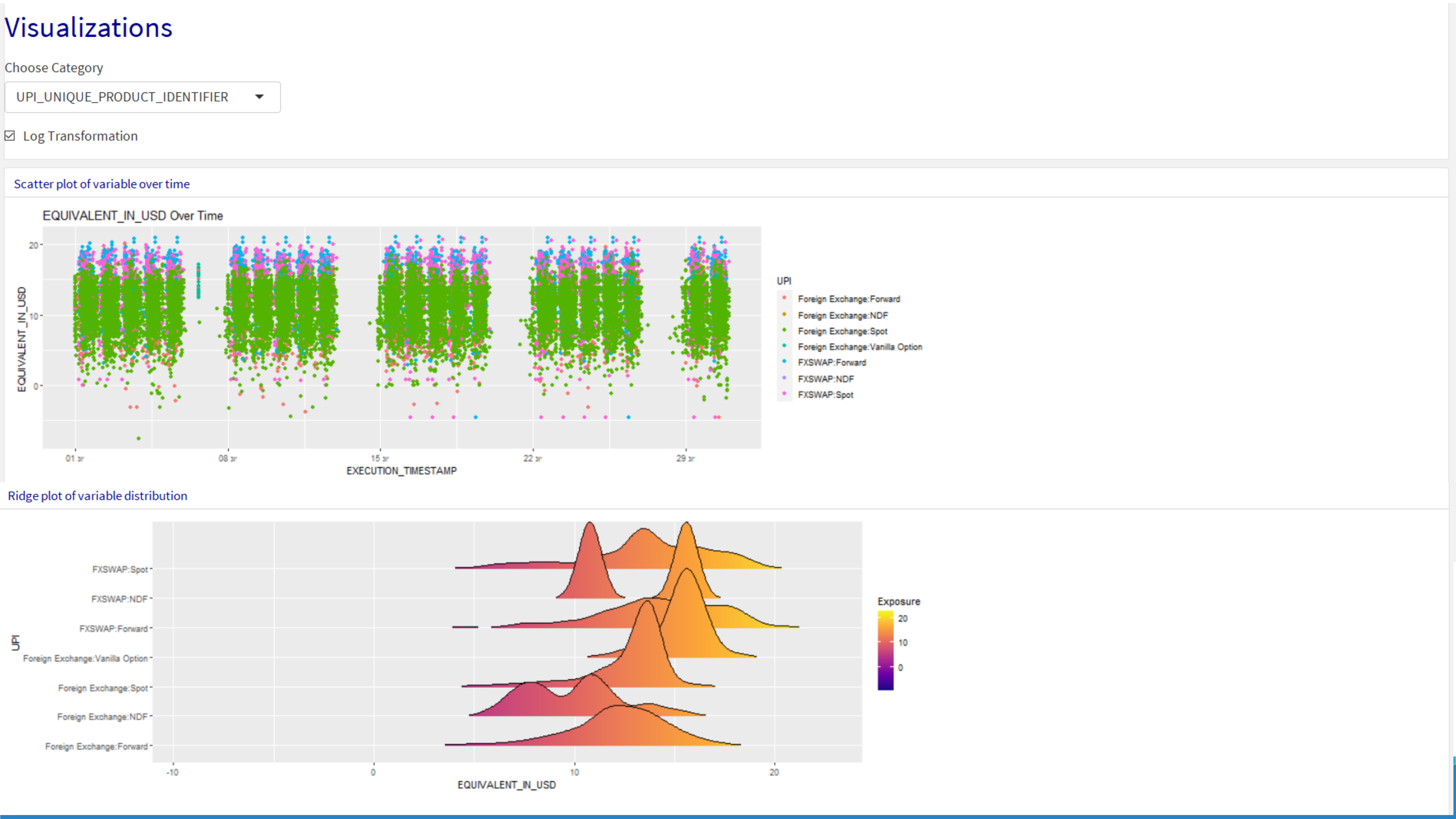
Showing 1 to 10 of 175,009 entries

Choose Variable to Examine



Choose Variable to Examine

EQUIVALENT_IN_USD ▾



Visualizations

Choose Category

UPI_UNIQUE_PRODUCT_IDENTIFIER ▼

☑ Log Transformation

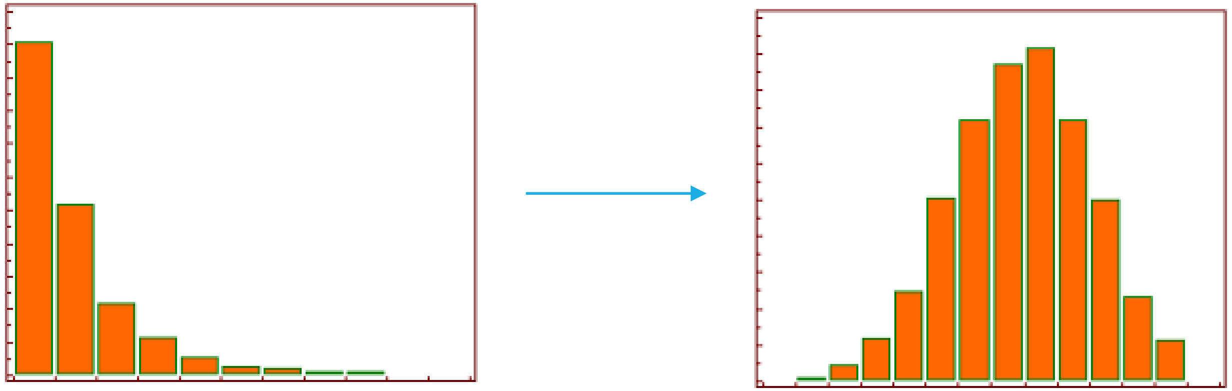
Scatter plot of variable over time

Ridge plot of variable distribution

Traditional Methods

Transformations

Financial data is characterized by **highly right tailed distributions**. Traditional parametric methods for anomaly detection often assume **symmetric** and sometimes **normal** distributions.



- Log transformation
- Box-Cox transformation: $y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & if\ \lambda \neq 0, \\ \ln y_i & if\ \lambda = 0 \end{cases}$

where optimal λ is one which results in best approximation of normal distribution curve

Measures of Variability

- SD – common practice, not robust to outliers, parametric approach - normal distribution
- IQR – common practice, non-parametric, semi-robust
- MAD – robust to outliers, non-parametric but does better with symmetric distributions.
- Double MAD – (*Rosenmai,2013*) like MAD robust to outliers and non-parametric, also takes skewness into consideration.

Measures of Variability

MAD - median absolute deviation

$$MAD(Y) = k * median(|Y_i - median(Y)|)$$

Where k is called the scale factor and is taken to be 1.4826 if data is normally distributed. We take k to be 1 since distribution is unknown.

Outlier threshold is taken to be *alpha* deviations from the median.

Measures of Variability

Double MAD - double median absolute deviation

Upper MAD:

$$\tilde{Y} = median(Y)$$

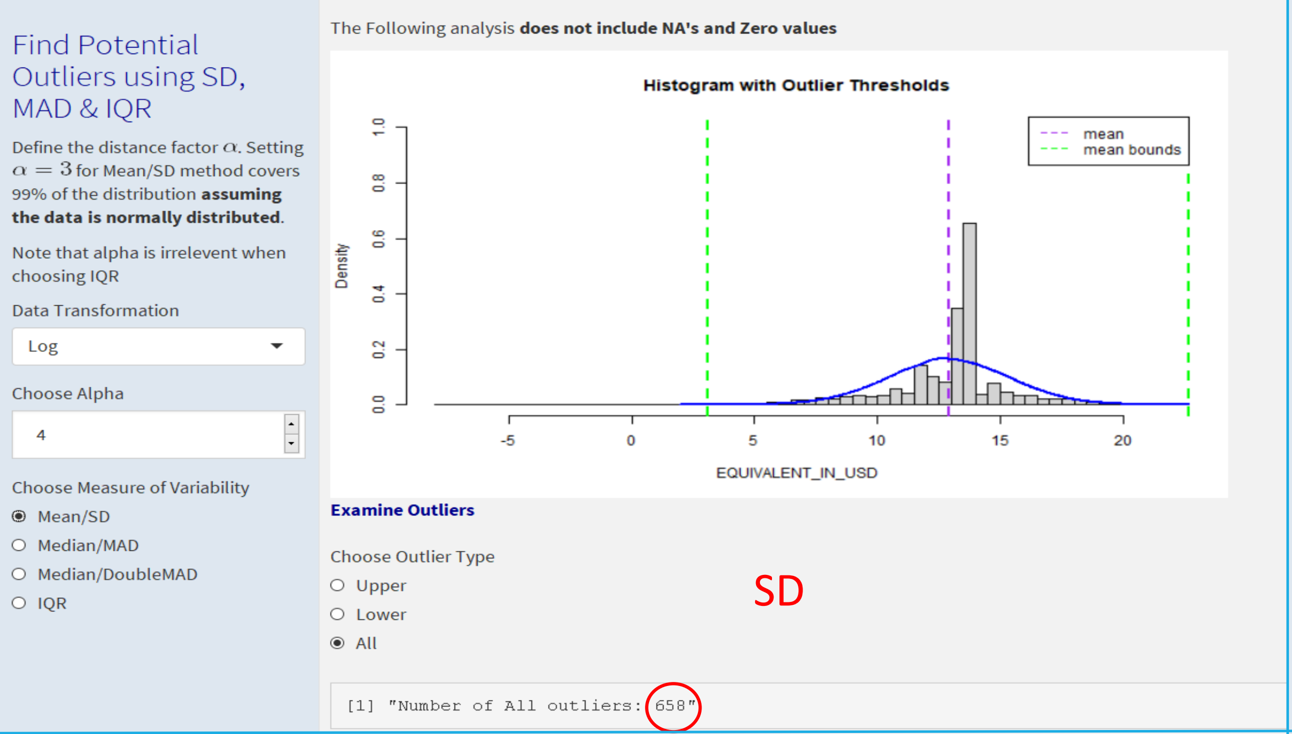
$$Y^{(u)} = \{y|y \in Y \cap y \geq \tilde{Y}\}$$

$$MAD^{(u)}(Y) = k * median\left(\left|Y_i^{(u)} - \tilde{Y}\right|\right)$$

Where k is called the scaling factor like in original MAD.

Similar for Lower MAD

Outlier threshold is taken to be *alpha* **upper** deviations, and *alpha* **lower** deviations from the median.



Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability

● Mean/SD
○ Median/MAD
○ Median/DoubleMAD
○ IQR

The Following analysis **does not include NA's and Zero values**

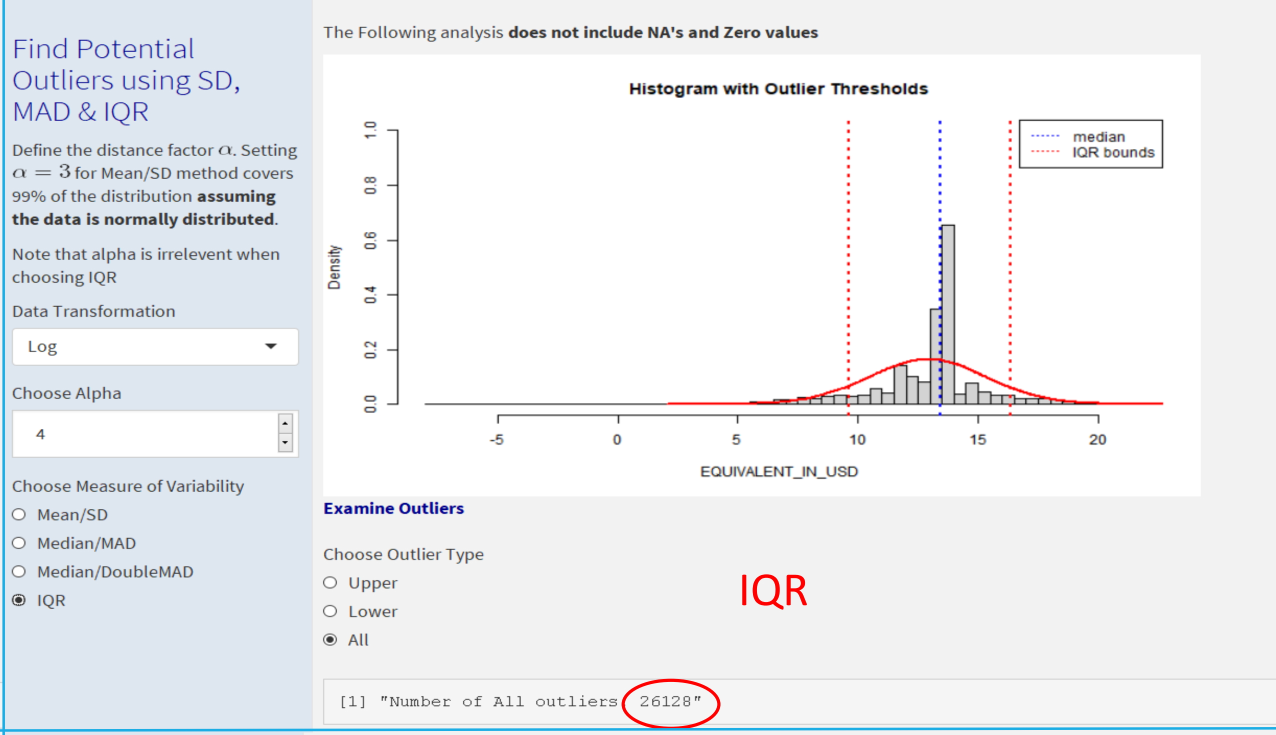
Examine Outliers

Choose Outlier Type

○ Upper
○ Lower
● All

```
[1] "Number of All outliers: 658"
```

SD



Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability

○ Mean/SD
○ Median/MAD
○ Median/DoubleMAD
● IQR

The Following analysis **does not include NA's and Zero values**

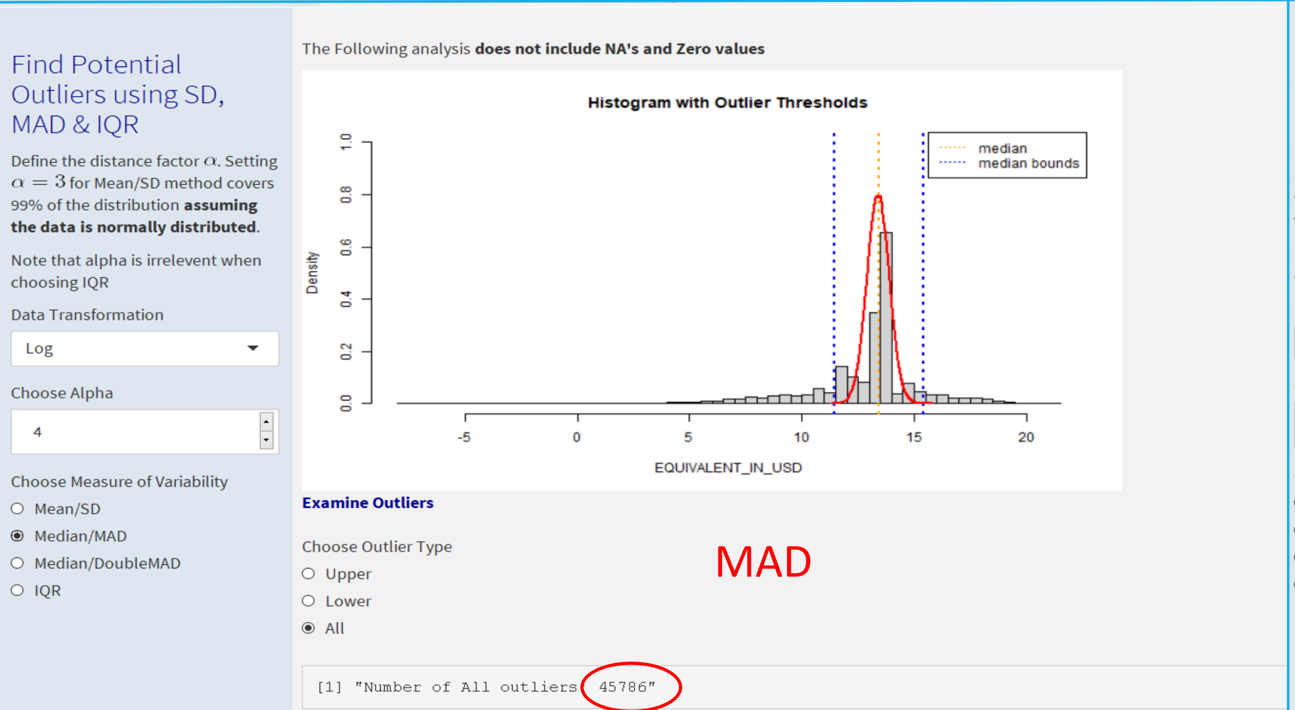
Examine Outliers

Choose Outlier Type

○ Upper
○ Lower
● All

```
[1] "Number of All outliers 26128"
```

IQR



Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability

○ Mean/SD
● Median/MAD
○ Median/DoubleMAD
○ IQR

The Following analysis **does not include NA's and Zero values**

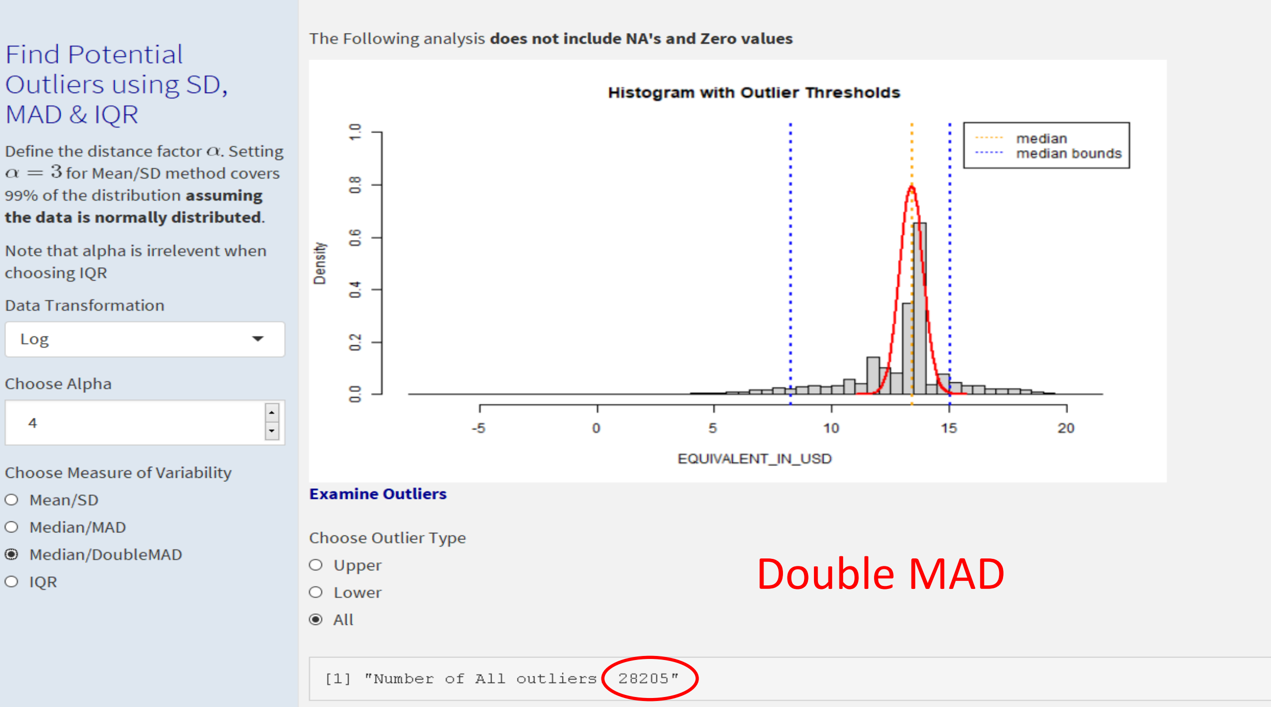
Examine Outliers

Choose Outlier Type

○ Upper
○ Lower
● All

```
[1] "Number of All outliers 45786"
```

MAD



Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability

○ Mean/SD
○ Median/MAD
● Median/DoubleMAD
○ IQR

The Following analysis **does not include NA's and Zero values**

Examine Outliers

Choose Outlier Type

○ Upper
○ Lower
● All

```
[1] "Number of All outliers 28205"
```

Double MAD

Non-Parametric Univariate and Multivariate Methods

Non-Parametric Methods

Applying **transformations** is not always enough to meet the **symmetry** or other **parametric assumptions** required for traditional anomaly detection methods.

A **distribution-free** test for outliers in data drawn from an **unknown data generating process** may give more reliable results.

Non-Parametric Methods

We offer two such methods:

1. Bootlier Plot and Bootlier Test:
   - "Bootlier-Plot – Bootstrap Based Outlier Detection Plot" *Kesar Singh and Minge Xie, 2003*
   - "A Distribution-free Test for Outliers, Discussion Paper Deutsche Bundesbank" *Bertrand Candelon and Norbert Metiu, 2013*
2. Isolation Forest:
   - "Isolation Forest" *Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou*

Bootlier Plot - *Singh and Xie, 2003*

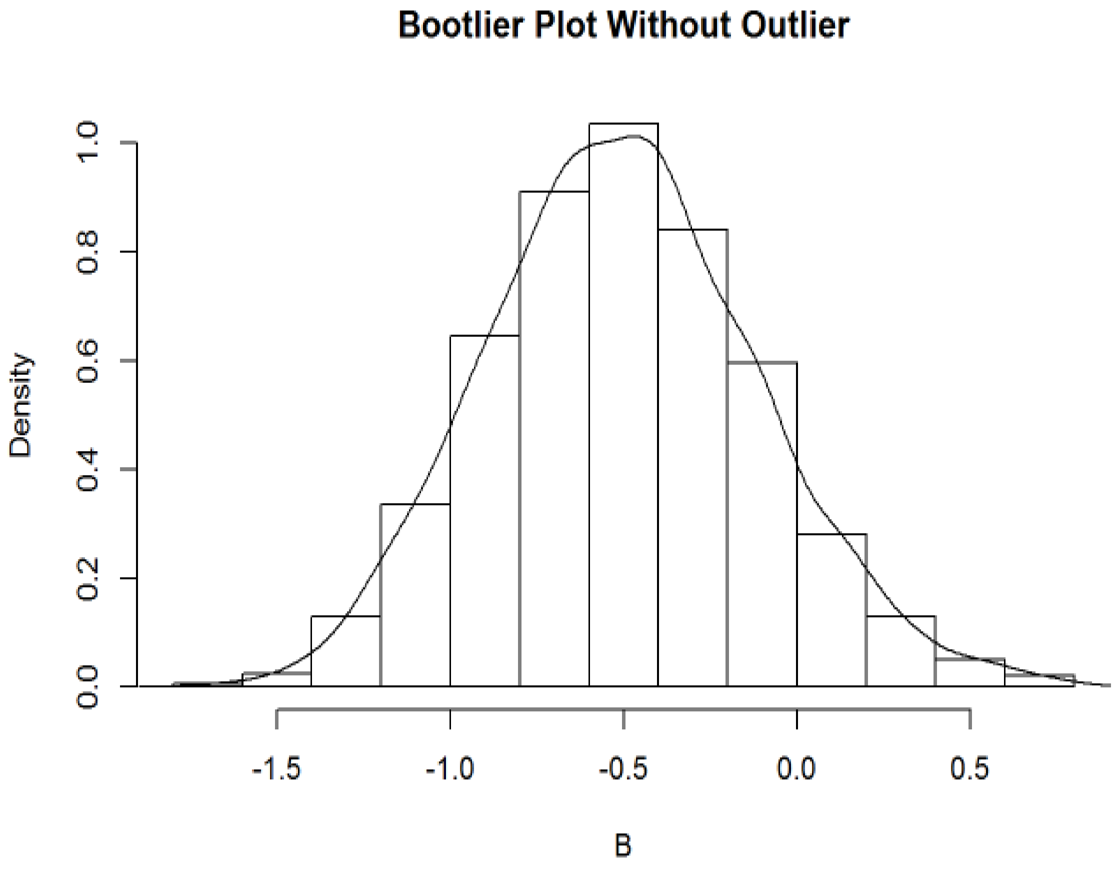
- Method is based on bootstrapping.
- When an outlier exists in a dataset, some bootstrap samples will contain the outlier while others will not.
- Presence of an outlier is expected to cause a significant increase or decrease in the bootstrap mean, and make the bootstrap distribution of the sample mean a mixture distribution.
- We expect the histogram of the sample mean to be multimodal.
- In order to make the bootstrap histogram more sensitive to a potential outlier, the chosen bootstrap statistic is the **"mean – trimmed mean".** Where the trimmed mean is the mean of the bootstrap sample after trimming *k* observations from each side of the sorted sample.

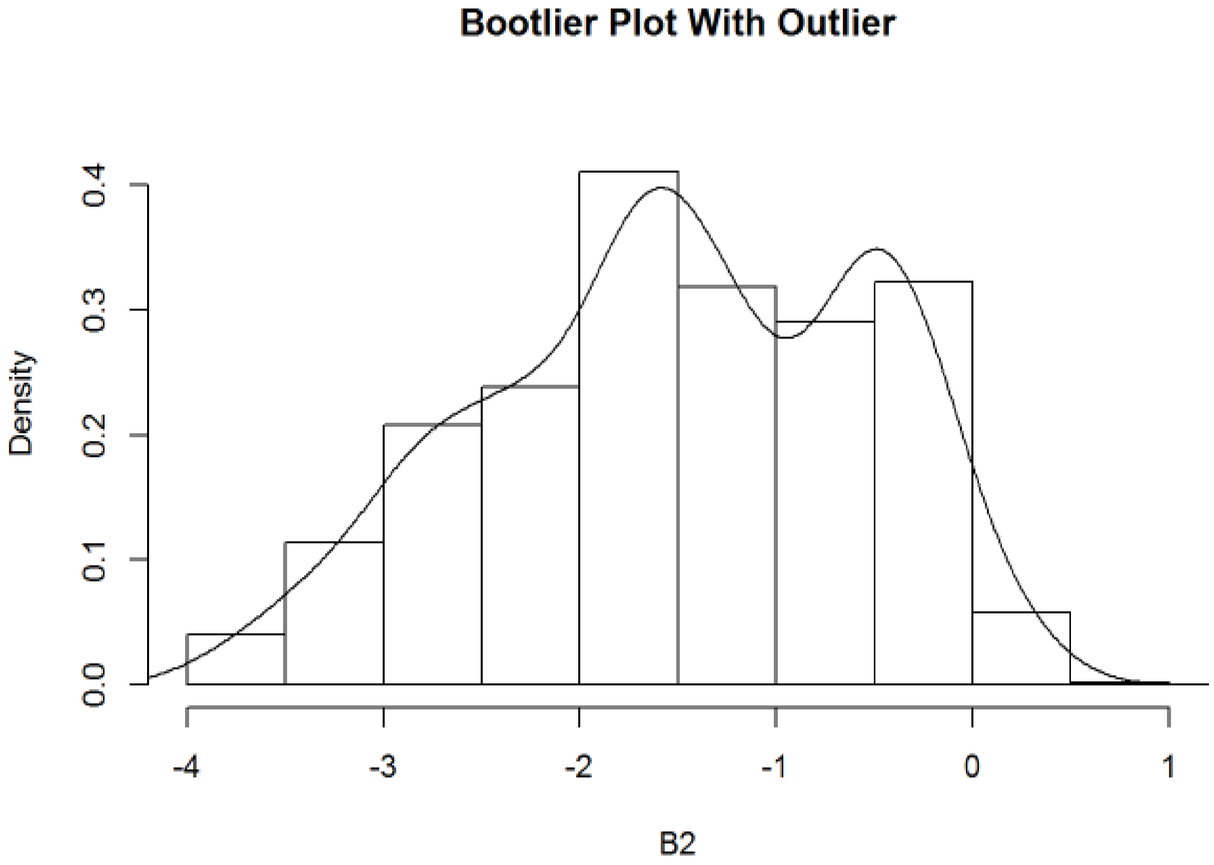
Mean-trimmed mean statistic:

$$T(Y^*) = \frac{1}{n}\sum_1^n Y_i^* - \frac{1}{n-2k}\sum_{k+1}^{n-k} Y_{(i)}^*$$

where $Y_1^*, Y_2^*, \ldots, Y_n^*$ denote bootstrap draws and $Y_{(1)}^*$'s be the corresponding order statistics

Bootlier Plot





Bootlier Plot - *Singh and Xie, 2003*

Bootlier Plot for large sample with numerous outlier candidates:

- Probability of having a bootstrap sample of size *n*, free from potential outliers is very small and we may not see a clear bump in Bootlier plot.
- Trick is to reduce bootstrap sample size to a fraction of *n*, i.e., *[αn]*, $\alpha \in (0,1]$
- **Practical recommendation:** look at several Bootlier plots at different bootstrap sample sizes – if any one is found bumpy that would be indicative of the presence of outliers.

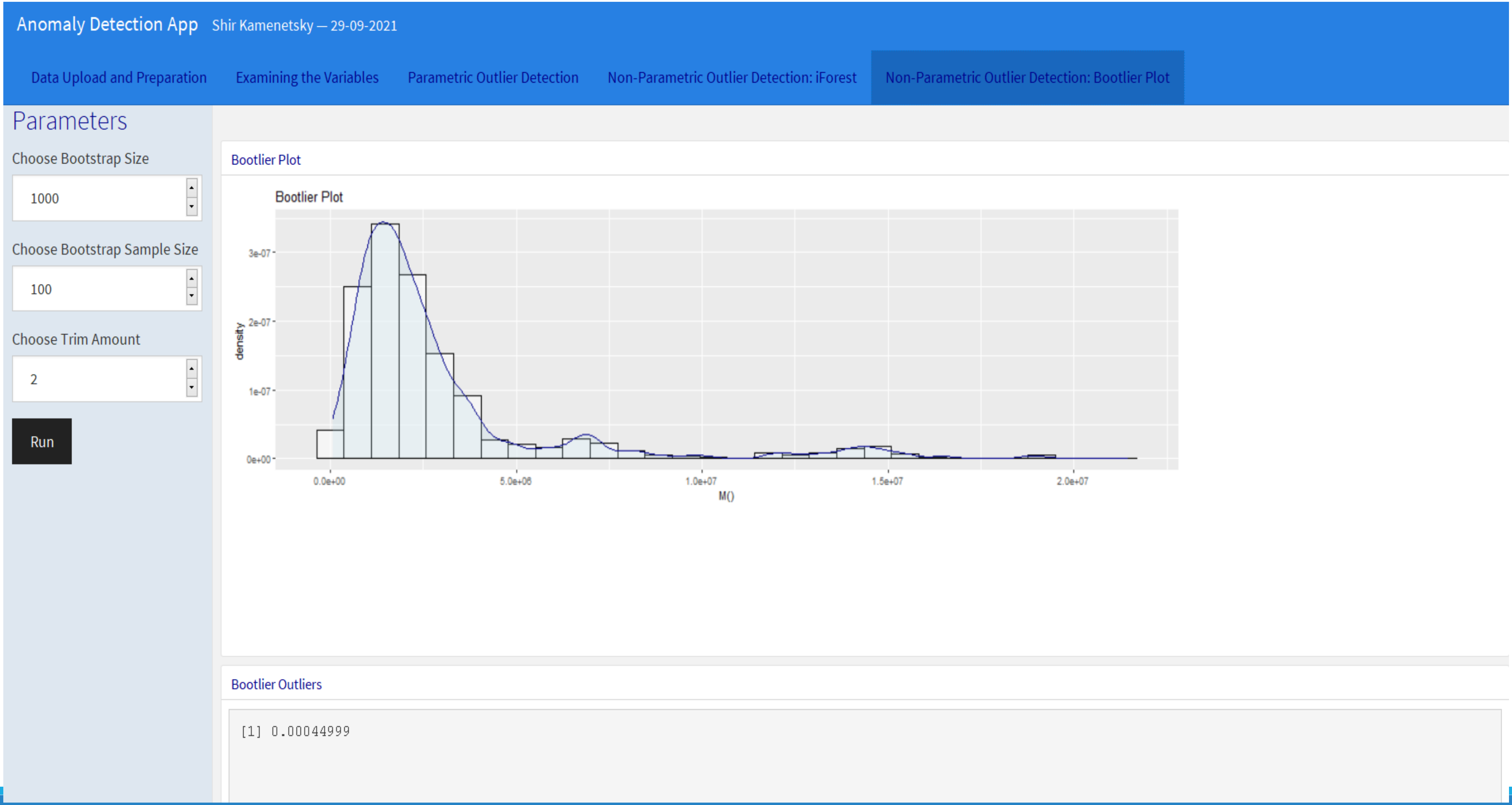
Bootlier Test - *Candelon and Metiu, 2013*

Identification of outliers based on Bootlier Plot

Bootlier Plot tells us whether there are outliers or not – how do we identify the outliers themselves?

Candelon and Metiu, 2013 address this in the Deutsche Bundesbank Discussion Paper – “A distribution-free test for outliers”.

- Two step process:
  1. **Test for multimodality: H0** – Bootlier plot has precisely one mode (and no local minimum), **H1** – Bootlier plot has more than one mode. Test hypothesis using:
     “Bootlier Test” - Bootlier plot coupled with distribution-free test for multimodality proposed by Silverman (1981)
  2. **Identify Outliers: 1)**Build subsamples by sequentially canceling observations from the tails of the original sample ordered in ascending order. **2)**Perform Bootlier test on each ordered subsample until the null hypothesis of unimodality cannot be rejected for a particular subset of observations. **3)**Data points not contained in this subset are the outliers.



Anomaly Detection App
Data Upload and Preparation
Examining the Variables
Parametric Outlier Detection
Non-Parametric Outlier Detection: iForest
Non-Parametric Outlier Detection: Bootlier Plot
Run
Bootlier Plot
density
3e-07
2e-07
1e-07
0e+00
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
M()

Shir Kamenetsky — 29-09-2021

Parameters

Choose Bootstrap Size: 1000

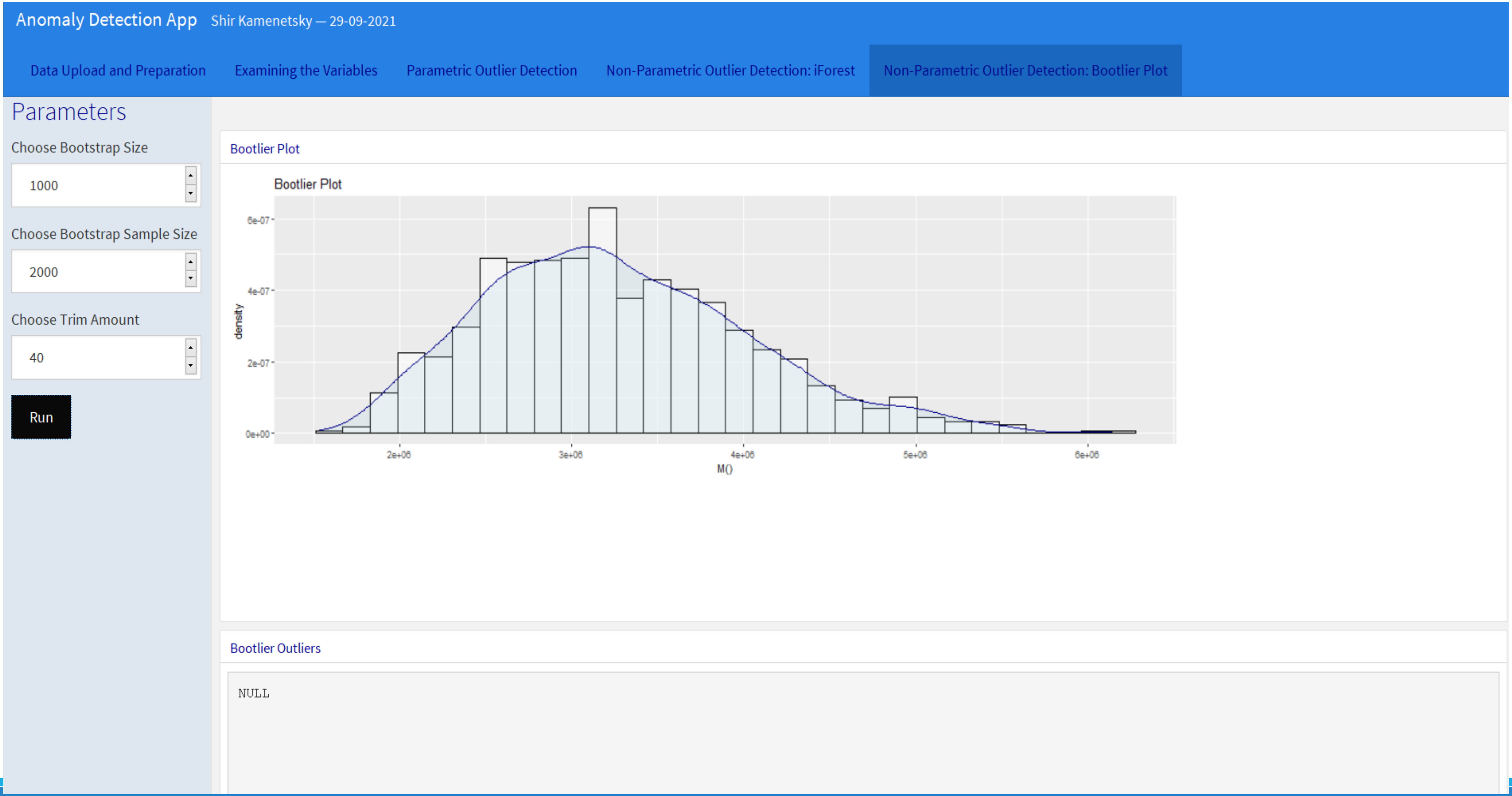
Choose Bootstrap Sample Size: 100

Choose Trim Amount: 2

Bootlier Plot

Bootlier Outliers

```
[1] 0.00044999
```



Anomaly Detection App Shir Kamenetsky — 29-09-2021

Data Upload and Preparation
Examining the Variables
Parametric Outlier Detection
Non-Parametric Outlier Detection: iForest
Non-Parametric Outlier Detection: Bootlier Plot

Parameters

Choose Bootstrap Size

1000

Choose Bootstrap Sample Size

2000

Choose Trim Amount

40

Bootlier Plot

Bootlier Plot
density
6e-07
4e-07
2e-07
0e+00
2e+06
3e+06
4e+06
5e+06
6e+06
M()

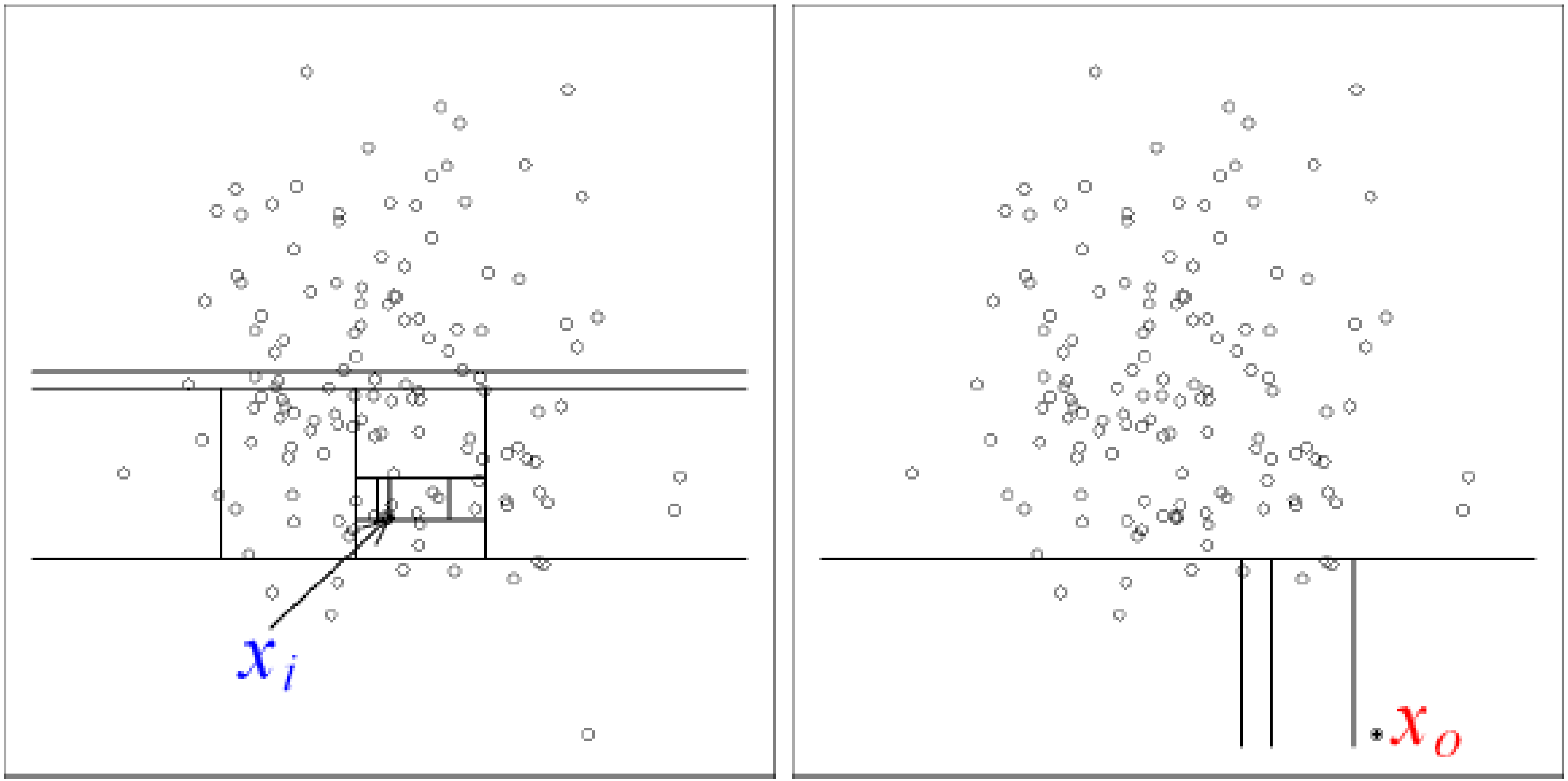
Bootlier Outliers

```
NULL
```

Isolation Forest - *Liu, Ting, and Zhou*

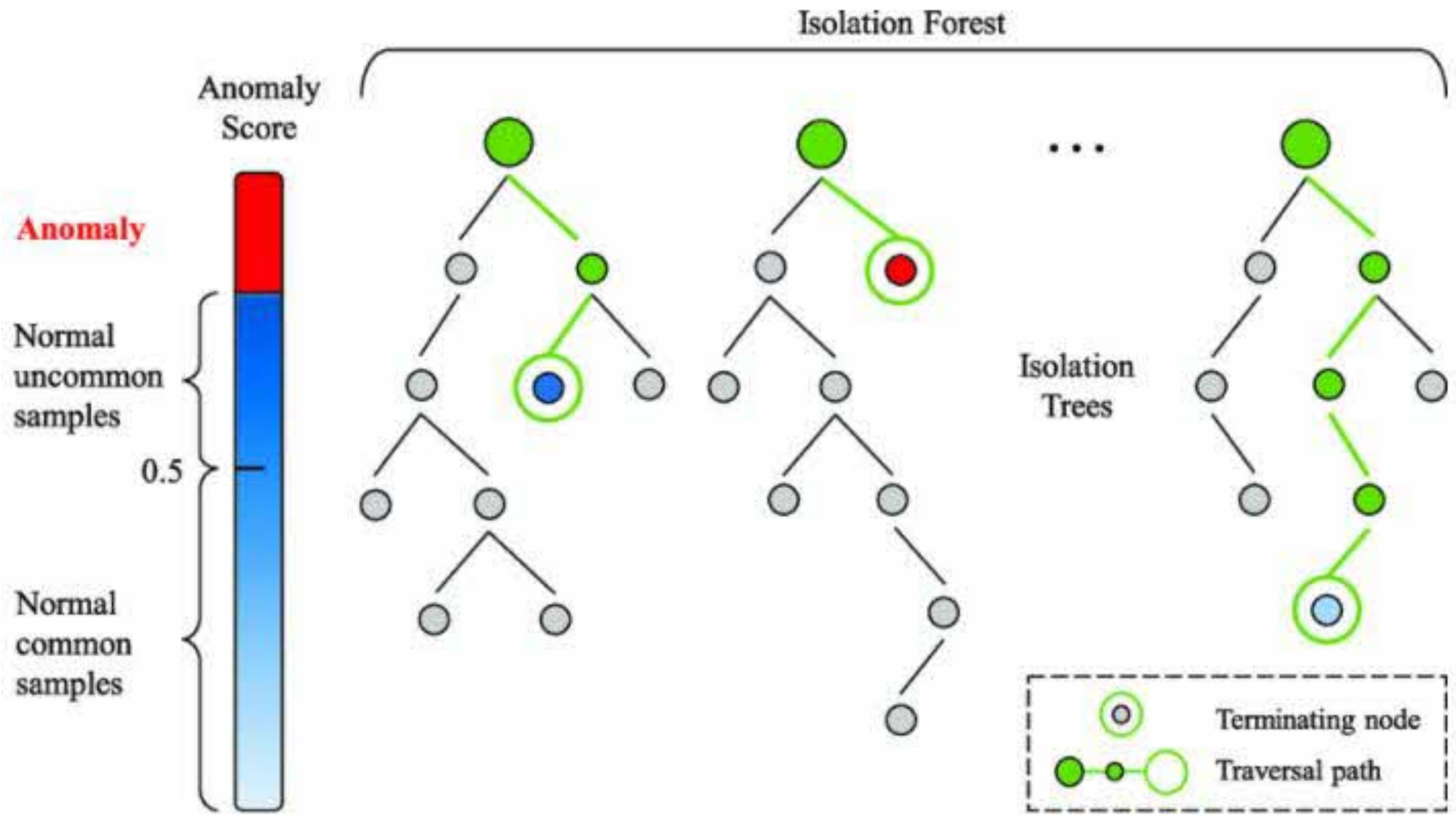
Basic idea is to **isolate anomalies** rather than profiling normal instances. Anomalies are **“few”** and **“different”** making them more susceptible to isolation than normal points. Isolation is conducted via construction of **tree structure**.

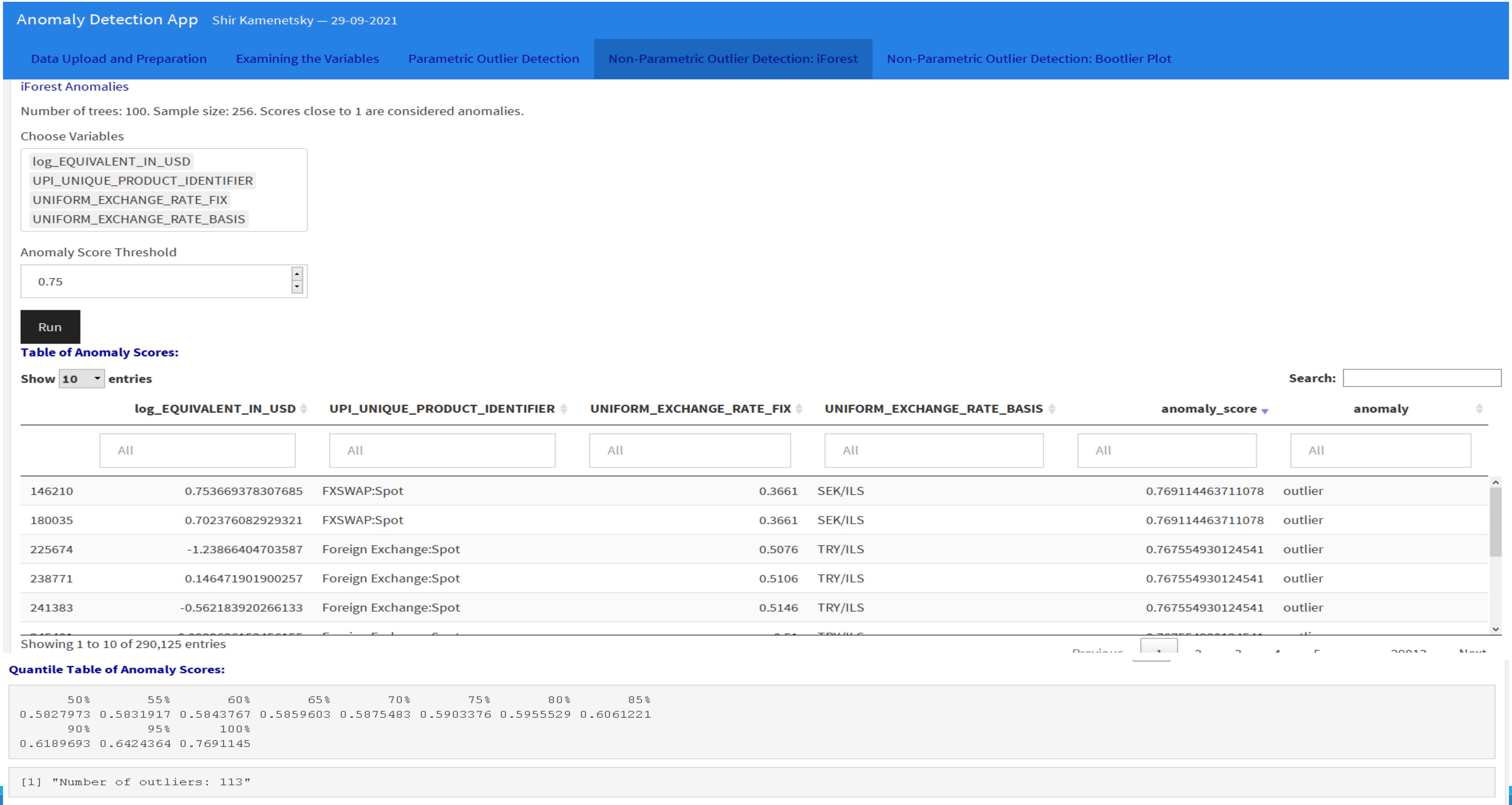
- **iTree (Isolation Tree)** – proper binary tree.
  - Partitions are generated by randomly selecting an attribute and then randomly selecting a split value between the maximum and minimum values of selected attribute.
  - This is done iteratively until each observation is isolated to its own node.
  - **Path length** is equivalent to number of partitions required to isolate a point, or in tree structure the number of edges from root node to external terminating node.
- **Iforest (Isolation Forest)** –
  - Builds ensemble of iTrees on sub-samples of data.
  - Calculate **Anomaly Score**: Aggregate iTree paths for each observation.
  - Anomalies are those instances which have short path lengths – or **anomaly scores close to 1**.



(a) Isolating x_i

(b) Isolating x_o


Isolation Forest
Anomaly Score
Anomaly
Normal uncommon samples
0.5
Normal common samples
Isolation Trees
Terminating node
Traversal path


Anomaly Detection App
Shir Kamenetsky — 29-09-2021
Data Upload and Preparation
Examining the Variables
Parametric Outlier Detection
Non-Parametric Outlier Detection: iForest
Non-Parametric Outlier Detection: Bootlier Plot
Run
Search:
Previous
1
2
3
4
5
29013
Next

iForest Anomalies

Number of trees: 100. Sample size: 256. Scores close to 1 are considered anomalies.

Choose Variables

log_EQUIVALENT_IN_USD
UPI_UNIQUE_PRODUCT_IDENTIFIER
UNIFORM_EXCHANGE_RATE_FIX
UNIFORM_EXCHANGE_RATE_BASIS

Anomaly Score Threshold

0.75

Table of Anomaly Scores:

Show 10 **entries**

	log_EQUIVALENT_IN_USD	UPI_UNIQUE_PRODUCT_IDENTIFIER	UNIFORM_EXCHANGE_RATE_FIX	UNIFORM_EXCHANGE_RATE_BASIS	anomaly_score	anomaly
	All	All	All	All	All	All
146210	0.753669378307685	FXSWAP:Spot	0.3661	SEK/ILS	0.769114463711078	outlier
180035	0.702376082929321	FXSWAP:Spot	0.3661	SEK/ILS	0.769114463711078	outlier
225674	-1.23866404703587	Foreign Exchange:Spot	0.5076	TRY/ILS	0.767554930124541	outlier
238771	0.146471901900257	Foreign Exchange:Spot	0.5106	TRY/ILS	0.767554930124541	outlier
241383	-0.562183920266133	Foreign Exchange:Spot	0.5146	TRY/ILS	0.767554930124541	outlier

Showing 1 to 10 of 290,125 entries

Quantile Table of Anomaly Scores:

```
      50%       55%       60%       65%       70%       75%       80%       85% 
0.5827973 0.5831917 0.5843767 0.5859603 0.5875483 0.5903376 0.5955529 0.6061221 
      90%       95%      100% 
0.6189693 0.6424364 0.7691145 
```

```
[1] "Number of outliers: 113"
```

Thank You!

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Unsupervised outlier detection in official statistics[1]

Nhan-Tam Nguyen, Deutsche Bundesbank,
and co-authors from the Deutsche Bundesbank
and the German Research Center for Artificial Intelligence

[1] This presentation 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 central banks and other institutions represented at the event.

Unsupervised Outlier Detection in Official Statistics

Tobias Cagala Jörn Hees Dayananda Herurkar Mario Meier
Nhan-Tam Nguyen Timur Sattarov Kent Troutman Patrick Weber[*]

November 1, 2021

Abstract

This paper presents a summary of a joint project conducted by the Deutsche Bundesbank and the German Research Center for Artificial Intelligence (DFKI). As a joint use case in the area of financial micro data, we evaluate the performance of all major classes of unsupervised learning algorithms for outlier detection and implement a complete machine learning workflow. Our workflow extends beyond pre-processing the data and flagging outliers by incorporating explainable AI methods and a possibility for the algorithm to exploit feedback by domain experts. We apply our approach to micro data sets that are typically collected by a central bank in the Euro area and that cover the structure and format of a wide range of financial data, namely the interest rates statistics (MIR), the money market statistics (MMSR), the (sectoral) securities holdings statistics (SHS-S), and the investment funds holdings statistics (IFS). With our work, we contribute to both the improvement of the data quality management work done by official statistical departments as well as to the literature on applied machine learning.

[*] Cagala: tobias.cagala@bundesbank.de; Hees: joern.hees@dfki.de; Herurkar: Dayananda.Herurkar@dfki.de; Meier: mario.meier@bundesbank.de; Nguyen: nhan-tam.nguyen@bundesbank.de; Sattarov: timur.sattarov@bundesbank.de; Troutman: kent.troutman@bundesbank.de; Weber: patrick.weber@bundesbank.de. The paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

Contents

1 Introduction

To meet the demand for timely provision of high-quality micro data in an environment of steadily rising data volumes, statistics departments of governmental organizations are increasingly turning to statistical learning methods from the fields of data science and machine learning (see for example Tissot et al. (2018)). The motivation is clear: these methods potentially promise higher process efficiency with the input of fewer (costly) human resources.

This paper presents a summary and lessons learned of a joint project conducted by the Deutsche Bundesbank and the German Research Center for Artificial Intelligence (DFKI). As a joint use case in the area of financial micro data, we evaluate the performance of all major classes of unsupervised learning algorithms for outlier detection and implement a complete machine learning workflow. Our workflow reflects the recursive nature of modern machine learning applications by extending beyond simple feature engineering and model estimation and into how to incorporate explainable AI methods and feedback by domain experts. We apply our approach to data sets that are collected by the Bundesbank and cover the structure and format of a wide range of financial data that include the interest rates statistics, the money market statistics, the sectoral securities holdings statistics, and the investment fund holdings statistics.

With our work, we contribute to the applied machine learning literature in two ways. First, we evaluate the performance of unsupervised learning algorithms in improving the data quality of micro data by flagging reporting errors that have characteristics of outliers. To this end, we collect reporting errors that domain experts (humans) detected in the past which gives us a unique labeled data set of errors and/or outliers that we can use to benchmark how well unsupervised methods recognize these errors. Second, we provide guidance on the implementation of all steps of an automated, unsupervised machine-learning pipeline that ranges from the pre-processing and selection of algorithms, to the application of explainable artificial intelligence (Explainable-AI) and active learning to enable and incorporate human feedback for official statistical data.

Our key findings are as follows: First, related to the performance of unsupervised algorithms, we show that most of the algorithms successfully isolate anomalous data points in micro data that were previously flagged in the data quality management (DQM) process by humans.[1] They achieve this without information on the labels by separating data points that deviate from the underlying structure of the data. In fact, we find that unsupervised algorithms can not only detect erroneous data points, but also hint at unusual data points and patterns that can further be analysed by data users. Second, we show that methods from the field of Explainable-AI provide domain experts with hints on how the models distinguish between anomalous and regular data points and can thereby inform business intelligence and allow statistics departments to issue more targeted DQM-related requests to reporting agents. Third, we address the challenge of incorporating the expertise of domain experts and reporting agents back into the production pipeline by implementing an active learning loop.

We conclude that unsupervised learning algorithms, applied to granular, financial data of the sort collected by a central bank, are not only suitable to detect incorrect reporting and thereby improve data quality, but that these methods can also detect unusual patterns in very heterogeneously structured data sets. However, our work also stresses that a production pipeline that is largely automated and that provides the possibility to actively incorporate (human) feedback is at least as important as a proper selection of algorithms.

The remainder of this paper is structured as follows: In Section 2, we summarize the different kinds of outlier detection methods, introduce measures for evaluating the performance of unsupervised outlier detection algorithms, describe our approach to counteract over-fitting, and describe the ecosystem in which we implemented the algorithms. In Section 3, we provide an overview of the data sets that we used for our study and descriptive statistics on outliers as well as the dimensions of each data set. Section 4 discusses pre-processing before running unsupervised algorithms, including how to deal with null values, how to handle categorical data, and how to scale data in the presence of dependencies. We further discuss the impact that feature

[1] Following Aggarwal (2015) an outlier or anomaly is a data point that is significantly different from the remaining data. In this paper, we use the terms outlier and anomaly interchangeably.

engineering can have on the performance of algorithms in this section. Next, we provide a broad conceptual overview of algorithms for the unsupervised detection of outliers in Section 5, including an outline of the strengths and weaknesses of each algorithm and their usefulness for detecting local versus global outliers. Section 6 moves away from single outlier detection methods to approaches that combine several machine learning algorithms into one single model, aiming at improving the performance of the final model. In Section 7 we provide an introduction to active learning. In Section 9, we open up the black box of unsupervised learning algorithms and introduce methods from the toolbox of Explainable AI. We close the section with an example on how explanations for outliers that were detected with autoencoders can provide novel insights to data producers and users alike. Section 8 shows cross-validated performance metrics for the detection of outliers in granular financial data sets with the approaches that we discussed in sections 5 and 6. Finally, Section 10 concludes this paper.

2 Background

2.1 What constitutes an outlier?

Detecting outliers or anomalies is a common data analysis task across various domains (e.g. health-care, quality assurance, financial data) with a variety of application scenarios (e.g. the identification of diseases, intrusions, mistakes, fraud, see e.g. Hodge and Austin (2004); Ahmed et al. (2016); Bhuyan et al. (2014)).

In general, outlier detection (OD) tries to solve a heavily imbalanced binary classification problem between a few points of interest (the true outliers) and the majority of other "normal" points (the true inliers). To solve the problem, outlier detection in general relies on the assumption that the true outliers can be distinguished from true inliers in the vector-/feature-space, e.g., by having a larger distance to their neighbors. Points showing such irregularities in feature-space are often called predicted outliers or simply outliers, while those similar to the majority are often called predicted inliers or simply inliers (Aggarwal, 2015; Aggarwal and Sathe, 2015; Chandola et al., 2009). It is worthwhile emphasizing the difference between "true outliers" and "predicted outliers": While the former are defined by experts, the latter are defined by distributions in feature-space. Whenever the distinction between the two is important, we will use the longer, more explicit names. In real world scenarios, the assumption that true outliers can be distinguished (easily or at all) from true inliers in feature-space is sometimes violated, leading to cases where true outliers can be predicted inliers and true inliers can be predicted outliers. What makes the detection of outliers an interesting use case is that an evaluation of these data points (the predicted outliers) might reveal certain patterns or problems (the true outliers) with a higher likelihood than when simply investigating a random sample of data points. This is particularly pertinent when the number of data points is prohibitively high and the fraction of true outliers is very low.

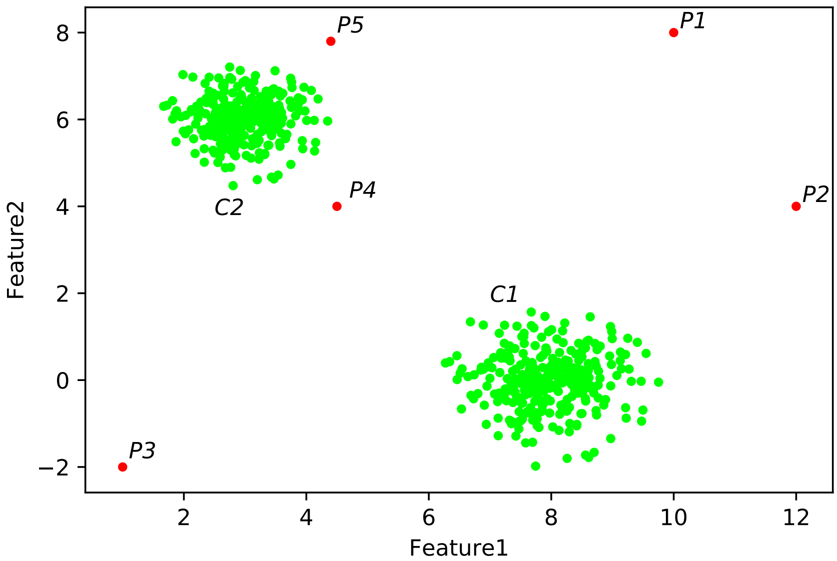
In the following, we will briefly describe the common sub-classes of outlier detection (also see Goldstein and Uchida (2016); Zhang et al. (2010)), based on the type of outliers of interest and the knowledge (if any) that is available about the true outliers and true inliers.

2.1.1 Global versus local outliers

Global outliers are data points which are classifed as anomalous to due to being (far) outside the overall distribution of the data set (Khoa and Chawla, 2010; Ernst and Haesbroeck, 2017; Dang et al., 2013; Goldstein and Uchida, 2016). An easy example for this class of outliers are points that are at least three standard deviations outside of an interval of an n-dimensional Gaussian that has been fitted to the whole data set. Outliers of this class are often data points that are orders of magnitude away from the others points and often caused by data entry mistakes.

In contrast, local outliers are points which are not anomalous on a global, but on a local scale (Khoa and Chawla, 2010; Ernst and Haesbroeck, 2017; Dang et al., 2013; Goldstein and Uchida, 2016). Such points deviate from the distribution/regularities of their local neighborhood. By this, local outliers are directly related to cluster analysis and account for the fact that many real world data sets can be better modelled as a composition of multiple distributions. Local outliers are then those data points which are close to such clusters, but which still behave different with respect to the local distribution of clustered points.

Figure 1: Example of local and global outliers in a 2D space



Notes: The figure shows an example for local and global outliers in a 2D space. C1 and C2 are two clusters of points. P1, P2, P3 represents global outliers, and P4, P5 represents local outliers with reference to cluster C2.

Local and global outliers are illustrated in Figure 1. While the detection of global outliers is often relatively trivial, it is typically much more difficult to detect local outliers, as such data points can reside well within the normal distribution limit of the data set. Unlike our illustration, real-world data sets typically have hundreds if not thousand of dimensions, making it challenging to find meaningful clusters and boundaries (Prieditis and Russell, 1995; Kriegel et al., 2005). Also, the algorithms that are optimized to find local outliers, often rely on a large variety of parameters to determine what is a "neighborhood" and what is "different". Still, the detection of not only global, but also local outliers is desired in many application areas. An overview of a variety of outlier detection algorithms can be found in Section 5.

2.1.2 Available labels

Depending on the availability of knowledge (also called ground truth labels or simply labels) about data points, outlier detection use cases can be divided into 3 main groups: unsupervised, supervised, and semi-supervised (Chandola et al., 2009; Chalapathy and Chawla, 2019; Goldstein and Uchida, 2016). In order to evaluate the outcomes of any algorithm and to compute the evaluation measures presented in Section 2.3, one needs at least some labelled data independent of the following groups.

Unsupervised outlier detection Unsupervised outlier detection is the process of detecting outliers without data labels, but solely by using density or distance measures of the data samples (Chalapathy and Chawla, 2019; Aytekin et al., 2018; Goldstein and Uchida, 2016). In this case the detection algorithm can only rely on the intrinsic properties of the data in feature-space to distinguish the abnormal samples (outliers) from the ordinary data (inliers). However, due to the advantage of not relying on often difficult or costly to acquire labels of the true outliers or true inliers, unsupervised OD is often the first choice in any OD application.

Supervised outlier detection For supervised outlier detection data labels are essential. The data points (or at least a subset of all points) have to be labelled as either true outlier or true inlier. A supervised outlier detection method is essentially a (strongly imbalanced) binary classifier with the task to classify a given data point into either an inlier or an outlier. The labelled data set is divided into at least training and test set so that the supervised OD model can be trained on the training set and evaluated on the test set, in order to evaluate its generalization to unseen data. In application oriented use-cases this approach is the least preferred one, because it is often difficult (or costly) to acquire a large enough amount of labelled data. The imbalanced nature of the data also complicates the acquisition of labels, as simple random sampling approaches often lead to situations in which the true outlier class suffers from too few samples to be well represented. In general it is debatable if the class of true outliers can be (or should be) well represented with examples, as focusing on such representations might hinder the detection of completely novel outliers in the future. Hence, while helpful for the detection of micro-clusters of outliers, it is advisable to combine fully supervised algorithms with those of the other classes.

Semi-supervised outlier detection Unlike the supervised context, semi-supervised algorithms are only trained on the true inlier labels. The underlying idea is that such an algorithm should model normality by learning the distribution of features from true inliers. Everything sufficiently deviating from this normality is then labelled as an outlier (Chalapathy and Chawla, 2019). In terms of classification, this is also called one-class classification. In practical use-cases, the acquisition of true inlier labels is often much simpler than that of true outlier labels. Especially based on previous unsupervised OD and a human review of the resulting predicted inliers, a large set of true inliers can often be generated with minimal human effort, making semi-supervised algorithms promising for practical applications.

2.2 On the importance of train-test splits

Splitting data between a train set for model estimation and a test set for validation is common practice in machine learning. In supervised learning, the goal is to avoid estimating a model that provides a tight fit to the relationship between features and predicted labels in the training data, but, due to modeling spurious relationships that do not generalize, does not provide accurate predictions of the labels in the test data. However, in unsupervised or semi-supervised learning applications, an algorithm cannot overfit on the prediction of a specific label, therefore the question whether to split the data into a train and test sample is more subtle.

The argument for a train-test-split in an unsupervised context is that it can prevent the estimation of overly-complex separation frontiers. If, for example, an unsupervised algorithm learns to distinguish clusters of observations, an over-complex separation frontier would be unstable.[2] Because overly-complex separation frontiers are partly driven by random, rather than structural relationships in the data, a model that returns different separation frontiers depending on random draws from the input data most likely suffers from over-fitting. A resulting measure of over-fitting in unsupervised settings is the cluster stability of a model.

There are different ways to split data into a train and test sample and in many settings, a random split of the data is sufficient to create proper train and test data sets. Many, if not most, data sets collected by central banks have a panel structure. In the context of financial (panel) data, three aspects should be considered: (a) the time-dimension of the data (b) the group structure of the data (i.e. holdings of bank A and holdings of bank B) and (c) the rarity of the outlier label. The time dimension is most subtle because financial data often contains features that reflect values from previous periods or changes across periods. If we split the data into a train and test set along the time dimension, using the later periods for testing, there could be data leakage from the train into the test set if features are serially correlated. Another source of data leakage are features that allow the algorithm to model serial correlation by including data from different time periods. In time series or panel data, we might include first differences as features. In this case, if we split the data randomly into train

[2] In our example, unstable separation frontiers result in different clusters if we train the unsupervised algorithm on different random samples from the same population. Stability is a desirable feature of a separation frontier because it implies that the algorithm learned structural and not spurious relationships in the data.

and test sets, information that we use in the train set can appear (e.g. in its lagged realization) in the test set. This can result in data leakage between the train- and the test set.

In our data sets, empirically either using a random split or attributing all months below a threshold date *t* to the train set and all months above the threshold to the test set does not have an effect on our findings because the algorithms cannot exploit this link in the data. The fact that there might be time persistence of outliers across time does not affect this logic. However, if there is a structural break in the data at some point in time, the random split might be more stable and better to extrapolate. If one splits the data according to some time threshold, it might happen that data before the structural break are the train set and the rest in the test set. This could considerably affect the performance in the test data. In contrast, time persistence should help to detect outliers more easily and should lead to a more stable algorithm if errors re-appear in *new data* that are fed into the algorithm.

Besides the aforementioned aspects, one needs to take into account that financial data often has a panel structure, with relationships within groups of time series. In our application, this group structure is relevant because reporting errors could be highly correlated within a certain (reporting or economic) group. In general, there are two ways to deal with the panel structure: First, it is possible to do a stratified split according to these groups so that the distributions across the train and test data sets are the same. The disadvantage, however, is that this could result in data leakage from the train to the test set. Second, it is possible to split the data set according to the groups themselves so that one group with all its observations is either always in the train or test data set. This, however, is only advisable if there are many small groups. However, since the data sets collected by central banks are relatively large, random splitting is usually sufficient to ensure that the distributions in the train and test data sets in terms of group belongings closely align. Still, it is advisable to do stratified sampling across relative membership categories such as banks, funds or sectors to ensure proper sample distribution by construction.

Another aspect that needs to be considered in the train-test split is that – by definition – the outlier label is heavily imbalanced. To ensure that the train and test data set contain the same fraction of outliers, we stratify the train-test-split according to the outlier label.[3]

Finally, the choice on the size of the test set should depend on the size of the data set and the fraction of outliers in the data set. For evaluation purposes it is necessary that an appropriate number of outliers is available in the test data set to avoid noise in the evaluation metrics due to the scarceness of the outlier label. However, this is sometimes not easy to achieve if the data set is too small. For our data sets, we have picked different test sizes.

To summarize, splitting and stratifying the data is important to properly evaluate the success of an outlier detection model and to avoid over-fitting and noisy model selection. Therefore, we always split our data, estimating the outlier detection model using train data and evaluating the model using test data.

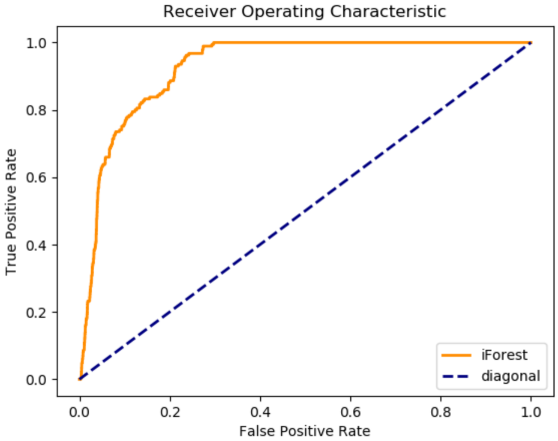
2.3 How to evaluate the success of unsupervised models?

For the evaluation of our models, we largely use two measures: the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve. These two measures principally measure the trade-off between different competing ideas of model performance. Each model we estimate produces a score, whether an instance is an outlier or not. An exception is One Class SVM that produces binary labels. However, here we can derive a score from the distance of the observation of the hyperplane that separates the classes. The threshold at which an instance is considered anomalous is often driven by the nature of the data and the problem to be solved. In our case, we are concerned with measuring a model's ability to detect rare and infrequent anomalies in our data and thus, we are faced in most cases with a severe imbalance between labelled outliers and inliers. For imbalanced data, the PR curve is often a fitting performance metric because it focuses more strongly on the minority class (Davis and Goadrich, 2006) . Thus, we mainly rely on the PR-curve to evaluate our models. Below is a brief description of the PR- and ROC curve.

[3] If no label is available in the data set, stratifying according to the label is, of course, not possible.

ROC curve The ROC curve plots the true positive rate (i.e. the recall) against the false positive rate. This curve shows for different threshold values of what constitutes an outlier and the number of true positives change versus the number of false positives. In Figure 2, we show a sample ROC-curve for an Isolation Forest model. For a model that perfectly separates outliers from inliers, the orange line would make a right angle to the upper left-hand corner, indicating that there is no trade-off because there are no false positives and all true positives have been isolated. The blue line indicates the curve a random model should produce. The area between the blue and orange lines (the area under the curve or AUC) measures the degree of the trade-off between true positives and false positives. For this model, the ROC-AUC score is 0.93, which is relatively high.

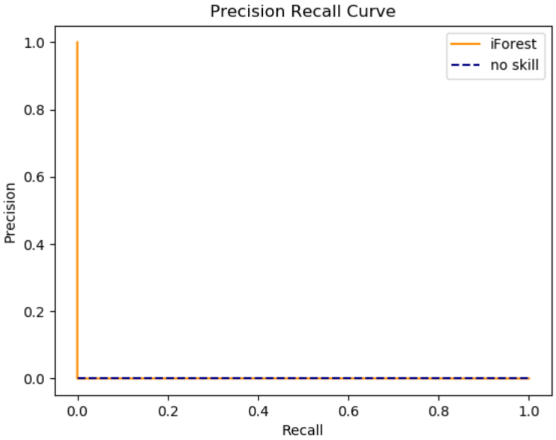
Figure 2: Sample ROC-curve for an isolation forest



Notes: The figure shows a sample ROC-Curve for an isolation forest model. The orange line is the ROC-curve for an isolation forest model, which shows the trade-off between the true positive and false positive rates for different thresholds of outlierness. A perfect model would form a right angle at the top left of this figure, whereas a skill-less model is represented by the diagonal blue line. Data source: WpInvest Aug-2017

PR curve However, we may not be interested in a model's ability to correctly predict inliers, but rather how well it predicts the much smaller outlier class. In this case, the PR-curve is much more useful, as it reflects the fraction of true positives among all positive predictions. Figure 3, which shows the PR-curve for the same model as above, makes clear that in the face of a class imbalance, PR-curves are a more appropriate measurement because potentially we may overestimate the ability of our model to predict the minority class on the basis of ROC AUC.

Figure 3: Sample PR-curve for an isolation forest



Notes: The figure shows a sample PR-Curve for an isolation forest model. The orange line is the PR-curve for an isolation forest model, which shows the trade-off between the precision and recall for different thresholds of outlierness. A perfect model would form a right angle at the top right of this figure, whereas a skill-less model is represented by the horizontal blue line. Data source: WpInvest Aug-2017

2.4 Computing environment

All experiments were conducted in *Python 3.6*. We used *Jupyter notebooks* during the evaluation phase as well as for the data exploration. The *scikit-learn* (Pedregosa et al., 2011) and *pyod* (Zhao et al., 2019) machine learning libraries were selected for training and evaluation of the models. The applicability of latter one was indeed good as it was specifically designed for anomaly detection tasks. The efficient training of the Autoencoder Neural Network was achieved by shifting expensive computations to GPUs (Nvidia V100). For such computations we used the *pytorch* (Paszke et al., 2019) deep learning library. In addition, the *statsmodels* (Seabold and Perktold, 2010) package was used for the estimation of the statistical models as well as for conducting statistical tests.

3 Data sets

The goal of our investigation is to apply the major classes of unsupervised algorithms to micro data sets which differ significantly in their dimensionality, frequency of collection, and their inherent properties. We will benchmark the performance of these unsupervised algorithm classes against information on all previously detected errors and outliers and evaluate their potential usefulness in a central bank's data quality management process. For our study, we use four different micro data sets that are typically collected by a central bank and for which we have the initially reported data set with all errors, next to a respective final data set where the errors were corrected and the outliers were flagged. A short summary of the data sets can be found in Table 1 and the following paragraphs.

The Investment Funds Statistics (IFS) collects all information about the individual holdings on a security-by-security basis for all investment funds issued by investment companies and public limited investment companies residing in Germany and subjected to the German Capital Investment Code. In addition to the granular holdings of each fund, a wide range of general information on the fund level is collected as well as the fund's key assets and liabilities[4]. Each line in the data set for the purpose of this paper corresponds to an asset or liability value submitted by the reporting entity at the end of each month.

The Securities Holdings Statistics (WpInvest) contains security-by-security information on all holdings of financial institutions registered in Germany for their domestic and foreign customers as well as the institution's own holdings. For each security – identified by the International Securities Identification Number (ISIN) – the nominal amount (or in some cases the number of units held) as well as the market value of the holding, the currency of the holding and the country of the holder are reported. In addition, flags are reported for securities repurchase and securities lending transactions. For the purpose of this paper, each line corresponds to a report by a single financial intuition for all of its customers located in Germany, broken down by the customers's sectoral classification (e.g. household, government, non-financial corporation etc.) and the customer's country of origin.

The MFI interest rate statistics covers all interest rates and the corresponding outstanding amounts (volumes) of existing and new business euro-denominated deposits and loans, broken down into the sectors households and non-financial corporations from the Euro area. The reporting is submitted by roughly 240 German banks on a monthly basis with month-end reporting values. Each line in the data set corresponds to a reported value by a bank for the interest rate or the corresponding outstanding amount, broken down by the aforementioned economic sectors for different (original) maturity buckets and loans and deposits respectively.[5]

Finally, the German part of the Money Market Statistical Reporting (MMSR) lists all transactions conducted in the money market of around 115 reporting agents from Germany. For the purpose of this paper, we focus only on the unsecured part of the MMSR which includes all unsecured transactions on a daily basis, covering

[4] Amongst others, this covers balance-sheet-like information on the total assets, the amount borrowed and loaned by the fund, the use of derivatives, and cash holdings in bank accounts. General information on the fund level cover information such as the number of fund shares outstanding, type of replication, assets under management and distributions.

[5] There a further breakdowns that are reported in the data set, for example the breakdown by the purpose of loans to households. For the purpose of this paper, however, we do not discuss those details of the data set.

borrowing and lending transactions for various instruments and for both fixed and variable rate contracts. Each row in our data set represents a single transaction of a bank with an eligible counterparty, including information on the counterparty itself, the agreed interest rate, the amount borrowed or lent as well as the maturity of the transaction and so forth.

Table 1: Overview of data sets

Name	Description
Investment Funds Statistics (IFS) 10k rows, 150 features, 5% outliers, Blaschke and Haupenthal (2020)	Monthly micro data on assets under management by German investment management and externally managed investment companies. Among other things, data consist of every security held by the respective investment fund on a security-by-security basis.
Security Holdings Statistics (WpInvest) 5M rows, 110 features, 0.001% outliers, Blaschke et al. (2020)	Securities reported by financial institutions domiciled in Germany which they hold for domestic or foreign customers. Furthermore, domestic banks provide information about their own holdings, irrespective of where the securities are held.
MFI Interest Rate Statistics (ZISTA) 40K rows, 12 features, 0.7% outliers, Bade and Krueger (2019)	The MFI interest rate statistics is composed as a representative sample of around 240 institutions. The MFI interest rate statistics measure the interest rates applied by domestic banks (MFIs) and the corresponding volumes for euro-denominated lending and deposit business with households and non-financial corporations domiciled in the euro area.
Money Market Statistics (MMSR) 25K rows, 33 features, 0.04% outliers, Bade et al. (2019)	The MMSR statistics provides the information on transactions carried out by monetary financial institutions on the euro money market. MMSR covers transactions in the secured, unsecured, foreign exchange swap and EONIA swap (euro overnight index swaps or OIS) market segments.

4 Data pre-processing and recommendations

4.1 General considerations

4.1.1 Encoded categorical values

Mixed data consists of numerical and categorical attributes. In order to work with categorical data, usually the non-ordinal categorical data is encoded with methods such as one-hot encoding or hashing (see Section 4.2 for a discussion of these approaches).[6] This change in representation, which in the case of hashing is not reversible, can lead to the generation of additional columns and different scaling. Because one-hot-encoding of categorical variables creates additional features, it can inflate the number of (encoded) categorical features relative to the number of (not encoded) numerical features. In financial data this can, for example, be the case for a feature that holds a large number of different currency codes. Creating a large number of features from categorical variables and keeping the number of numerical attributes unchanged can artificially increase the influence of categorical variables. If we use feature bagging in an isolation forest, for example, inflating the number of one-hot encoded categorical features increases the probability of drawing the categorical feature relative to drawing a numerical feature.[7] This problem is exacerbated by the fact that for each data point, there is at most one column in the expanded feature space for the categorical attribute that is set to one. Hence, the new feature space is quite sparse and data points are more likely to be equally far apart from one another (curse of dimensionality).

How to cope with this issue depends on the learning method. Trees in isolation forests need to grow deeper in order to capture the expansion of the feature space by a one-hot encoded column, more trees have to be generated, or more weight can be assigned to numerical columns by repeating them in the data set.

[6] For ordinal categorical data, order-preserving numeric encodings can be used.

[7] We use feature bagging in ensembles to reduce the correlation between estimators by training them on random samples of features instead of the entire feature set.

Alternatively, only a subset of the categories are encoded to keep the feature space small. For distance-based methods, measures such as Gower distance that take the mixed data structure into account or a reweighting of the individual columns are suitable.

4.1.2 Skewed and sparse distributions

This section will discuss how to deal with skewed and sparse data to detect outliers in the Bundesbank data. In particular, this section only covers numerical data, for a discussion of how to handle categorical data we refer the reader to Section 4.2.

Skewed data refer to data with long tails on either side of the distribution, which holds also true for multivariate distributions. One naïve approach to handle long tails of distributions is to truncate/clip them appropriately at empirically specified thresholds. In light of the fact that we are aiming to detect outliers, this needs to be handled with much caution in order to avoid truncating actual outliers. For the IFS data, we did some experiments truncating large values to a pre-specified threshold. This had two effects: (a) the focus slightly moved away from over-weighting large investment entities; however, (b) at the same time actual outliers were "truncated away" and could not be detected anymore. In the end, to avoid truncating away actual outliers and because most columns had no long tails, we did use truncation to deal with skewed data. The same applies to the WpInvest and ZISTA data sets.

Sparse data refer to the fact that in finite samples, high dimensional feature spaces are sparsely populated with data points. Because our data only provide us with a relatively small sample size, sparsity creates some difficulties in detecting outliers when a large number of features are included in the estimation. In section 4.1.4 we will discuss concrete approaches to reduce the dimensionality of the feature space to avoid estimating models in too sparsely populated feature spaces.

4.1.3 Missing values

The data sets used in this exercise are, relatively speaking, quite complete, as many of the values are required fields and do not pass basic validation checks if not filled. In cases where information in the training data is indeed missing, be it due to missing reference data or because the key is appearing for the first time (thus leading to missing values in the lagged values columns), the missing data is filled with zero or, in the case of categorical variables, assigned to a placeholder string. Additionally, for the ZISTA data, where continuous data is indeed missing, a separate categorical variable is created, which indicates the amount type, either positive, negative, or missing. The additional missing category had a negligible impact on the model's performance.

4.1.4 High dimensional data

The following table provides an overview of different approaches to reduce the dimensionality of the data. Our data sets have many categorical variables, and as such, are high dimensional following the aforementioned preprocessing steps. Aside from the bucketing/binning approach mentioned below, we explored using PCA and Autoencoders ways to reduce the dimensions prior to estimation, as well as bagging to reduce variance of the base estimators. Table 2 describes these methods and our evaluation thereof.

Table 2: High dimensional data reduction methods

Approach	Description	Evaluation
PCA	Linearly maps features into lower dimensional space.	PCA is efficient and easy to understand. However, it does not handle categorical variables or nonlinear relationships between features well.
Autoencoder (AE)	Uses a neural network architecture to compress or encode features into a lower dimensional space.	Is able to flexibly model nonlinearities between feature categories (unlike PCA). In particular, using it as a means of vectorising a large number of categorical variables is promising. AE suffers from a high degree of tuning parameters and long training times.
Feature Bagging	A meta estimator that combines (via mean or max) a number of base detectors on various sub-samples of the data set to improve the predictive accuracy and to control over-fitting. Features are randomly sampled from a subset of the features.	This is a useful technique we used during the intermediate and model validation stages of model development.
Feature Selection	Choosing or omitting features based on domain knowledge or empirical developments.	We did very little manual feature selection, except to omit non-reported features merged from reference data sets.

4.1.5 Large data sets

Large data sets may pose a problem for learning methods with high time and space complexity. Instance-based methods such as k-nearest neighbour (Ramaswamy et al. (2000); Angiulli and Pizzuti (2002)) and local outlier factor (Breunig et al. (2000)) have a time complexity of $O(nd)$ (or $O(log(n)d)$ when using an indexing structure) in the testing phase, where n is the training set size and d is the number of features. Thus, large training sets incur additional indexing structure construction cost in the training phase and query cost that depend on the training set size in the testing phase. In addition, instance-based methods need to keep the whole training set in memory during testing. Instead of working with the complete training data, a stratified subsample can used. However, the subsample may have an insufficient number of outliers. Therefore, for the WpInvest data, for example, we train on a subsample with a fixed normal-to-outlier class ratio of 10:1. In order to avoid inflated performance metrics, we have to ensure that we perform the evaluation on the complete (testing) set and not on the training data with over-sampled outliers.

4.1.6 Excluding extreme outliers

Exploratory analysis of the data has revealed a number of extreme outliers in the data. In particular, by simply plotting the development of a particular position over time in the ZISTA data set, one may notice the spike(s) in such series. It should be noted that such spikes are rare fluctuations of the business that do not represent errors but correctly reported data. As a result, training the model on such data may negatively influence the detection rate of the outliers on unseen data.

Considering the above-mentioned findings, we have added a pre-processing step where a subset of extreme outliers was removed from the data before the execution of the training cycle. The set of candidate samples for removal was selected according to the following criteria: the data point has to reside outside of three standard deviations from the mean of the distribution.[8]

Such pre-processing steps resulted in a higher detection rate and an overall positive outcome in terms of the defined performance metrics. We believe this step mainly affected the decision boundary of the model and

[8] There is a number of other techniques (like an Inter Quantile Range or Mean Absolute Deviation instead of the Standard Deviation) for removing the extreme outliers in the pre-processing step.

subsequently increased the generalization capabilities of the model. We found that this approach was a useful, computationally cheap method to improve the quality of the trained model.

4.2 Categorical variables

The variables in many financial data sets are of mixed types. Besides continuous variables, the data often include categorical variables. If the data set has a panel structure, for example, the cross-sectional and time dimension trivially correspond to categorical variables. We discuss four approaches to deal with categorical variables: One-hot Encoding, one-hot encoding a subset of categories, hierarchies, and hashing.

One-hot encoding One common approach to dealing with categorical variables is One-Hot Encoding. For each category, we create a binary variable that takes the value one if the categorical variable is equal to the category and zero otherwise. One advantage of this approach is its simplicity. A downside is that for categorical variables with many categories, the resulting number of binary variables is large. This can lead to a very sparsely populated feature space, especially if only small numbers of observations are part of each category. We can furthermore run into performance issues for algorithms whose computational cost increases with the number of features in the model. Another downside of one-hot encoding is that we lose the information that for a given categorical variable, the realizations of the binary (one-hot encoded) variables are not independent from each other. Because each observation belongs to one category, only one binary variable out of the group of one-hot encoded categories can take the value one. Although a model can learn this type of structure, we make the task of the model harder by not encoding information on the dependence between binary variables that stem from the same categorical variable.

One-hot encoding a subset of categories One way to counteract the large number of sparse features resulting from one-hot encoding is to one-hot encode only a subset of categories. Encoding only a subset can improve the performance metrics by counteracting overfitting to categories with a small number of observations. By resulting in a smaller number of encoded variables, we can also reduce the training time by selecting categories. A disadvantage is that the approaches require the selection of additional hyperparameters (e.g., a variance threshold of number of variables k). The approaches furthermore remove the distinction between non-encoded categories, which may pose problems for explainability. For the main part of our project, we used two ways to select a subset for encoding. The first approach is the *selection of top-k groups*. Instead of introducing a binary variable for each value of a categorical feature, we only create a binary variable for the top-k values, where the top-k values refer to the categories with the largest number of observations. Residual categories that are not in the top-k values are mapped to a separate binary variable. A special case of this approach is encoding the mode of the categorical variable. The second approach is using a *variance threshold*. Here, we calculate the variance of each one-hot vector and only include those with a variance above a certain threshold. Because the ranking of categories is identical if we use the variance or count the number of observations in a category, the variance threshold and the selection of top-k groups yield the same results as if we select an adequate threshold (value of k).

Hierarchies For many of the categorical variables in financial data, a grouping is possible. Because the number of groups is smaller than the number of categories, one-hot encoding groups results in a smaller number of binary variables than simple one-hot encoding. The approach preserves the information on the group level. For country codes, for example, we can group categories by larger geographic regions (Europe, Asia, ...) and one-hot encode these groups. This corresponds to mapping the categories to higher levels in hierarchical categorizations.

Hashing Another approach to encode categorical variables that results in a lower number of features than one-hot encoding is hashing. A hash function maps a categorical attribute with domain size k to a domain with size $k' \ll k$, thereby keeping the feature space small. However, relationships between categories are not preserved. Because the size of the target space of the categorical attributes can be set via the hyperparameter k', the feature space does not become uncontrollably large. Since hashing maps categorical values uniformly to the target space, collisions prohibit a one-to-one mapping from the feature space back to the original space.

Findings We find that one-hot encoding a subset of variables and using hierarchies improved the performance in terms of the success of the models in isolating outliers and running times of the training as compared to simple one-hot encoding. Out of the outlined approaches, hashing had the worst performance regarding the models' ability to isolate outliers.[9]

4.3 Numeric variables

4.3.1 Scaling

Independent scaling Here we scale features independently, i.e. without taking information from other features into account. For continuous variables, we apply min-max scaling to rescale variables between zero and one (one-hot-encoded categorical-type variables are naturally already scaled between zero and one). We also applied standard scaling, which has the effect of centring all inputs with a mean of zero and a variance of one. Scaling is essential for distance-based methods (such as LOF) to ensure equal weighting of features, and generally for efficient optimization of the cost function. For the particular case of the ZISTA data, we log-scaled features to normalized skewed distributions. In addition to global scaling, we also explored several stratified scaling approaches, described in the following section.

Scaling that captures dependencies The observations in our data sets are not independent. For example, own securities holdings that banks report in WpInvest data belong to reporting the banks' portfolio. If we feed the raw data to an algorithm, we do not exploit the domain knowledge on potential interdependencies between observations, i.e. positions in the same portfolio. To incorporate this information, we can scale numerical variables with aggregates by groups. For example, we can divide all own holdings of a bank in WpInvest by the aggregate size of own holdings of the bank. An alternative to scaling is to incorporate additional features that capture interdependencies. In the WpInvest example, we can include indicator variables for banks that allow the algorithm to model relationships between all own holdings of a bank, such as a larger average size of the holdings of the bank compared to the other banks.

For the IFS data we can normalize with the own-fund volume of investment funds. This gives an indication of the relative importance of funds positions and avoids a too large weight on large positions in absolute value. When we re-scale the numerical features in the IFS data, we do not find improvements in the overall performance of the outlier detection algorithms. However, the flagged outliers focused less on large funds and more on funds with relatively large positions in specific asset classes.

4.3.2 Binning

Binning is a method to discretize or smooth numerical data. Usually the continuous data is discretized in a fixed number of bins of equal width or by using quantiles to generate bins with an approximately equal number of observations. Then, for discretization, the result can be encoded in one-hot or in an ordinal format. For smoothing, values can be replaced, e.g., by their bin means.

For the WpInvest data, we observed that the performance of the quantile strategy was superior to equal-width binning. However, omitting the binning step altogether led to the best performance. One-hot encoding

[9] For the IFS data, some categorical classifications can be directly inferred from some of the numerical attributes. Hence, the information gain associated with these features might be small. Indeed, the results without categorical features are almost as good as with categorical features. In addition, because most outlier detection models are only properly specified for numerical data only this might be the cleanest approach for IFS data without losing much in terms of detecting outliers.

was inferior to ordinal encoding because of the increased size of the feature space and the loss of ordinal information.

4.4 Feature engineering

As for other types of data, feature engineering can have a large impact on the performance of algorithms that learn the structure of the data. We discuss three types to features that we can engineer in many financial data sets.

Past realizations In time series analysis, we commonly include lags to account for the influence of past realizations of a variable on future realizations. In outlier detection with unsupervised machine learning methods, past realizations provide context to the algorithm that can help to distinguish common from unusual data points. Large values of numerical variables, for example, can seem anomalous if we do not account for past realizations of the same variable in the previous period. A common method to account for past realizations is to calculate first differences. For a data generating process with

$$y_t = \mu + y_{t-1} + \epsilon_t, \tag{1}$$

first differencing leaves us with

$$y_t - y_{t-1} = \epsilon_t - \epsilon_{t-1}. \tag{2}$$

We eliminate the fixed component μ from the data, which our model then does not have to explain to model the structure of the data. We also make the implicit assumption that the previous realization's marginal effect on future realizations is one. If we want to leave the choice of the marginal effect size to the model or allow for different marginal effects for different groups of observations, we can include the previous realization as a feature, instead of calculating the first difference.

Aggregates Including aggregates also allows for a contextual evaluation by the model. For example, we can include the overall issued nominal value of a security as a yardstick for the model to compare to the size of the holdings. Because most algorithms can flexibly learn interactions between variables but cannot learn to aggregate values across rows, aggregation should be part of the feature engineering if we believe that it adds useful information to the data.

Context from other statistics Another source of contextual information can be other statistics. If we model banks' interest rates in the ZISTA data, for example, we can include interest rates, set by the Governing Council of the ECB.

Findings For all data sets, we find that first differencing and adding lagged features only slightly improved the performance of the algorithms. Adding aggregated interest rates per maturity and reporting period, led to a slight improvement for ZISTA data. In the IFS data and for the WpInvest data, the inclusion of aggregates also resulted in small improvements of the performance. Adding reference interest rates to the ZISTA data did not improve the performance of the model. One reason for the absence of a gain in performance is that the inclusion of one-hot encoded periods (time fixed effects), already allows the model to take into account contextual changes at time t. Therefore, additional information on changes in the interest rates, set by the ECB, do not add explanatory value. However, because they can allow for better explainability of the results by having a clear interpretation, the interest rates are superior to one-hot encoded time periods.

5 Approaches to detect outliers and recommendations

In this section, we strive to provide a broad overview of algorithms that allow for the unsupervised detection of anomalies.[10] Because of the abundance of resources on the methodology of the algorithms, we do not provide a detailed description of their inner workings in this paper. Instead, Table 3 refers the reader to the original paper that introduced the algorithm and further resources.

Table 3: Resources on the methodology of the anomaly detection algorithms

Algorithm	Original Paper	Further Reading
Isolation Forest	Liu et al. (2008)	
kNN	Ramaswamy et al. (2000); Angiulli and Pizzuti (2002)	
DBSCAN	Ester et al. (1996)	
LOF	Breunig et al. (2000)	
FINCH	Sarfraz et al. (2019)	
One Class SVM	Schölkopf et al. (1999)	
Autoencoder	Rumelhart et al. (1986)	Schreyer et al. (2017)
PCA & rPCA		
HBOS	Goldstein and Dengel (2012)	
ARIMA		Junttila (2001)

Table 4 shows a short intuition behind the algorithm and conceptual differences between the anomaly detection algorithms that we evaluated. Here, we distinguish between five groups of algorithms on the basis of their methodological approach. The first algorithm uses decision trees to isolate outliers. Algorithms in the second group use notions of distance or estimates density functions. Cluster based approaches use clustering algorithms, whereas SVM based algorithms rely on Support Vector Machines for classifying observations as outliers. Reconstruction based methods map the data to a lower dimensional space and then reconstruct the higher dimensional representation of the data. They then flag observations as outliers that have a high reconstruction error. Finally, we also discuss more classical statistical approaches.

[10] For a taxonomy and discussion of different anomaly detection algorithms, see, e.g., Goldstein and Uchida (2016) and Zhang et al. (2007).

Table 4: Overview of anomaly detection algorithms

		Intuition	Strengths	Weaknesses	Global vs. Local	Assessment
Tree Based	Isolation Forest	Anomalous instances in a data set are easier to separate from the rest of the sample (isolate), compared to normal data points. In order to isolate a data point, the algorithm recursively generates partitions on the sample by randomly selecting an attribute and then randomly selecting a split value for the attribute. When the iTree is fully grown, each data point is isolated at one of the external nodes. Intuitively, the anomalous points are those (easier to isolate, hence) with the smaller path length in the tree, i.e. points that are earlier separated at nodes of the tree.	• Fast to estimate • Easy to implement • Intuition of approach is easy to understand and certain degree of explainability (allows us to look at individual trees) • Few hyperparameters • Results are relatively robust to hyperparameter tuning	• Not tuned towards detecting local anomalies • Standadrd implementations of isolation forest cannot handle categorical data. One-hot vectors are treated equally to numerical data. This is problematic in the same way as for decision tree classifiers with random splits between categoricals	Global	The algorithm is well suited for our (mixed) data and a very good baseline model for comparison with other algorithms. IForests are also a very good starting point when implementing alternative models, feature spaces etc, because they are easy to implement, have few hyperparameters, fast to estimate, and are relatively robust.
Distance and Density Based	kNN	To determine the outlyingness of a data point, determine the (average) distance to its k(th)-nearest neighbour(s). Outliers are far away from their nearest neighbours, whereas inliers are similar(=close) to their nearest neighbors.	• Conceptually simple • Explainability • Distance metric takes information of the complete row into account	• Hyperparameters (number of neighbors and distance metric) make tuning more difficult • Not tuned towards detecting local anomalies • Standard implementations do not scale well for large or high-dimensional data sets • Suffers from curse of dimensionality	Global	The method is well suited for small to medium-sized data sets of low/medium dimension. For high dimensions both outlier-detection and computational performance suffer. Due to its conceptual simplicity, this algorithm serves as a good baseline model for benchmarking

... continued on next page

Table 4: Overview of Anomaly Detection Algorithms (continued)

		Intuition	Strengths	Weaknesses	Global vs. Local	Assessment
	DBSCAN	Density-Based Spatial Clustering of Applications with Noise determines core samples that are in a neighbourhood with high density. A neighbourhood is dense for a sample if there are at least a certain number of samples within a given distance. Data points that are close to a core sample form a cluster. Data points that are neither core samples nor close to them are considered outliers.	Can handle clusters of arbitrary shapes	• Does not provide anomaly scores out of the box (but binary labels) • Possibly slow training with high memory usage	Tendency towards detecting global outliers, but depending on the choice of hyperparameters DBSCAN can detect local outliers as well	The algorithm performs for outlier detection relatively well. However, the delicate interplay of hyperparameters and the feature space complicates the usage of this method. Incorrectly setting the hyperparameters leads to large training times and high memory usage. In addition, the calculation of anomaly scores has to be implemented separately.
	LOF	Anomalies are not located in densely populated neighbourhoods. The algorithm calculates the LOF score of an instance as the ratio of the average distance of the instance to its k-nearest neighbours over the average distances of the k-nearest neighbours to their respective neighbors. Anomalies will obtain large scores as they have low local density compared to normal observations.	Easy to implement	• Not very robust estimates • Slow to estimate with large number of neighbours	Depending on the number of neighbours (hyperparameter), the LOF can detect local as well as global outliers in the data	Like kNN the method works well with small and low-dimensional data. Large data sets and high-dimensionality pose a challenge to the algorithm that then becomes intractably slow.

... continued on next page

Table 4: Overview of Anomaly Detection Algorithms (continued)

		Intuition	Strengths	Weaknesses	Global vs. Local	Assessment
Cluster Based	FINCH	Forms chains by linking data points to their nearest neighbour. If data points have the same first neighbour, it links them to each other. The connected components of this graph form a cluster. To generate additional clusters,the algorithm performs the previous steps recursively on computed average data points.	• Conceptually simple • Few hyperparameters • Fast training and estimation, so it can be used for large and high dimensional data set	• Not designed as an anomaly detection method • Multiple solutions • No singleton clusters	–	The method was not considered due to related scalability issues in the reference implementation of the authors of this method.
Kernel Based	One Class SVM	One-Class SVM is a special case of the traditional SVM algorithm that is used for unsupervised scenarios. The main property of the traditional SVM is the ability to build a non-linear decision boundary by projecting the data to a high-dimensional (feature) space. The "Kernel trick" is used to perform the projection. In the feature space a "straight" hyperplane is built to separate the data to classes (positive / negative). The goal is to find the function that is positive for regions with high density and negative for low density.	• Ability to learn complex decision boundary. • Provides an "anomaly score" per sample (distance to the hyperplane)	• Compute and storage requirements increases rapidly with the number of training samples, due to the expensive kernel computation. • Might become sensitive to hyperparameters. Selection of the kernel, rejection rate, soft margin etc. have to be adjusted according to the data set structure. • Difficult interpretability of the model for high-dimensional data sets. • Cannot handle categorical data.	Captures global outliers almost always. For detection of local outliers tuning of model hyperparameters might be required.	OCSVM does not scale well to larger data sets (although more computionally efficient implementations are being developed). It can require a certain degree of hyperparameter tuning to achieve an acceptable performance.

... continued on next page

Table 4: Overview of Anomaly Detection Algorithms (continued)

		Intuition	Strengths	Weaknesses	Global vs. Local	Assessment
Reconstruction Based	Autoencoder	Performs non-linear data transformations by reducing the dimensionality to a lower level and then transforming it back to the original data space. The transformation may consist of multiple steps (hidden layers). Anomalies are those samples that performed worst in the reconstruction phase.	• Ability to capture non-linear relations in complex data structure • Multiple assessment of errors: reconstruction error, latent representation	• Computationally expensive • Lots of hyperparameters for tuning • Interpretation of the results is difficult • Sensitive to the attributes selected	Captures global outliers better than local outliers	The algorithm performs well detecting the global outliers.
	PCA and rPCA	Performs linear data transformation by reducing the dimensionality to a lower level and then transforming it back to the original data space. Anomalies are those samples that performed worst at the reconstruction phase.	• Not many hyper-parameters • Relatively fast • Level of explainability is relatively high • Multiple assessment of errors: reconstruction error, latent representation	• Poorly performs capturing nonlinear relationships • Sensitive to the attributes selected	Captures global outliers better than local	The algorithm showed a relatively good performance and could be well suited as a baseline model for comparison.
Statistical	HBOS	Uses the histogram approach for calculating the outlier score. The frequency (relative amount) of samples in a bin is used as density estimation. In multivariate anomaly detection, the scores obtained from each histogram are computed individually and combined afterwards.	• High level of explainability • Extremely fast • Small number of hyper-parameters • Good interpretability	Does not capture (unusual) relationships between the features and is sensitive to the feature selection	Captures global outliers better than local outliers	The algorithm showed good performance only for a particular set of feature combinations. Therefore, for successful model selection, the set of features must be done carefully.

... continued on next page

Table 4: Overview of Anomaly Detection Algorithms (continued)

		Intuition	Strengths	Weaknesses	Global vs. Local	Assessment
	ARIMA	Forecasting the time series data using the historical observations of the series.	• Good explainability • Few hyperparameters	• Each model has to be built separately for individual time series • Scaling of the anomaly scores across multiple series has to be done carefully	Captures global outliers better than local outliers	The algorithm showed a big potential for time series data. However, it needs to be calibrated carefully. We can imagine that some models need to be retrained/-calibrated from time to time because trends for some of individual time series change over time.

6 Combination of detectors and recommendations

In this section, we discuss how to detect outliers by combining the output of multiple outlier detection algorithms. These so called ensemble methods are meta-algorithms that combine several machine learning algorithms into one predictive model and thereby aim to improve/boost the performance of the final model (Opitz and Maclin, 1999; Rokach, 2010; Polikar, 2006). The outlier detection algorithms that are used to construct the ensemble are known as components. Outlier detection ensembles have many advantages over individual outlier detection algorithms. Often there are cases where a model that was trained on a data set will work well for the particular subset of the data and will fail when applied to other parts of the data. Also, in some instances, a trained model can perform well solving a task in one data set and fail performing the same task in other data sets. An ensemble model helps to leverage the different strengths of algorithms by not relying on a single model that could work well on only a particular data set. If one component under-performs in detecting outliers in a specific scenario, it is likely that this doesn't strongly impact the overall ensemble model's performance, as other components can work well for the same data points and thereby compensate. Hence, overall it can be observed that ensemble models often provide more stable / robust results when compared to individual models (Aggarwal, 2012; Aggarwal and Sathe, 2017). Generally, the design of an outlier detection ensemble model follows three steps (Aggarwal and Sathe, 2017):

1. *Model creation*: This step includes methodology or algorithms used to create the components.

2. *Normalization*: Ensemble models may consist of multiple heterogeneous components and the output from each component can be in different ranges. Therefore it is important to normalize the different scales of outlier scores from different components.

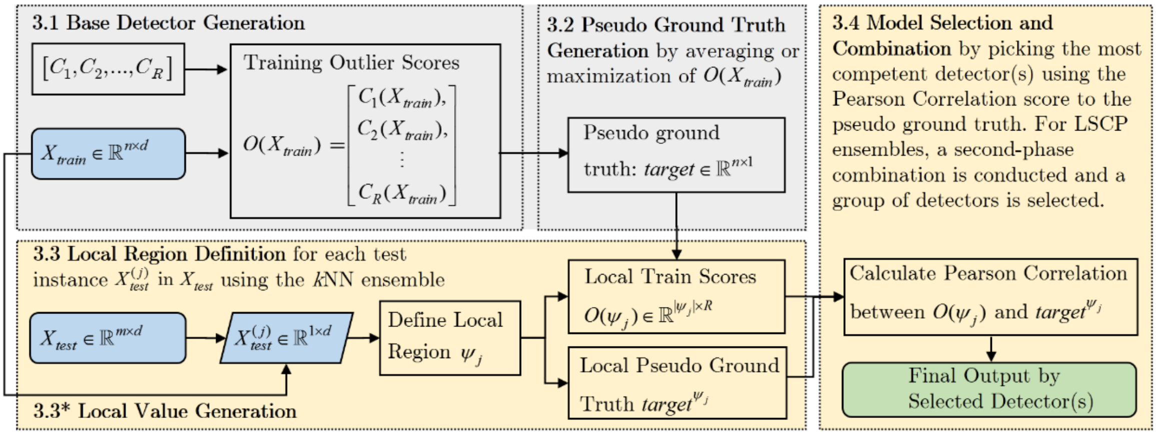
3. *Model combination*: We refer to the algorithm that combines individual components' outputs as fusion method. We have utilized and implemented different fusion methods (see below).

Outlier detection ensembles can be categorized into multiple groups, depending on either the type/class of components used or based on dependency within the components in the ensemble model (Aggarwal, 2012). We designed an ensemble model that is a hybrid (independent and model-centered) of different outlier ensemble groups. Drawing on Aggarwal (2012); Zhao and Hryniewicki (2018); Pasillas-Díaz and Ratté (2016); Zimek et al. (2014), we implemented different model combination functions. In the following, we distinguish three approaches.

6.1 Simple fusion methods

As the name suggests, simple fusion methods combine different components' outputs by using simple mathematical operations as combination functions. Among others, these functions are: maximum, average, damped averaging, pruned averaging, majority voting, normalized to one per component max, normalized to one per component average. Apart from their simplicity, advantages of simple fusion methods are that they are easy to implement, allow for easy interpretability, and are less computationally intensive. On the other hand, they exhibit limitations. The functions are not capable to learn the patterns in the component output, and the performance improvement depends on the diversity of the components' output. Further limitations are that *max* has a tendency to overestimate the outlierness and *average* tends to dilute the outlierness due to irrelevant components (Zimek et al., 2014; Aggarwal and Sathe, 2015). We recommend starting with these methods in the ensemble due to their simplicity, even though in our tests, they were not able to outperform individual components.

Figure 4: Flowchart of DCSO algorithm


3.1 Base Detector Generation
$[C_1, C_2, ..., C_R]$
Training Outlier Scores
$O(X_{train}) = \begin{bmatrix} C_1(X_{train}), \\ C_2(X_{train}), \\ \vdots \\ C_R(X_{train}) \end{bmatrix}$
$X_{train} \in \mathbb{R}^{n \times d}$
3.2 Pseudo Ground Truth Generation by averaging or maximization of $O(X_{train})$
Pseudo ground truth: $target \in \mathbb{R}^{n \times 1}$
3.4 Model Selection and Combination by picking the most competent detector(s) using the Pearson Correlation score to the pseudo ground truth. For LSCP ensembles, a second-phase combination is conducted and a group of detectors is selected.
3.3 Local Region Definition for each test instance $X_{test}^{(j)}$ in X_{test} using the kNN ensemble
$X_{test} \in \mathbb{R}^{m \times d}$
$X_{test}^{(j)} \in \mathbb{R}^{1 \times d}$
Define Local Region ψ_j
Local Train Scores $O(\psi_j) \in \mathbb{R}^{|\psi_j| \times R}$
Local Pseudo Ground Truth $target^{\psi_j}$
Calculate Pearson Correlation between $O(\psi_j)$ and $target^{\psi_j}$
Final Output by Selected Detector(s)
3.3* Local Value Generation

Notes: The figure shows a flowchart of DCSO algorithm with an explanation for each step. The figure was adapted from Zhao et al. (2018).

6.2 Unsupervised fusion methods

In an outlier ensemble, multiple outlier detection algorithms are used as the components which are applied on the input data for the outlier prediction. Later, these components output are used as input to the ensemble model for fusion. If labels are available (supervised learning) the optimization in the fusion-step can be based on the predictive performance of the labels. If labels are not available, wich usually is the case in outlier detection, we need to rely on an unsupervised fusion method. We implemented two unsupervised combination methods: Dynamic Combination of Detector Scores (Zhao and Hryniewicki, 2018) and Ensemble of detectors with correlation votes / Ensemble of detectors with variability votes (Pasillas-Díaz and Ratté, 2016).

Dynamic Combination of Detector Scores (DCSO) consists of two main steps: generation and combination. In the generation step, different and diverse base detector algorithms are selected. These base detectors can contain any outlier detection algorithm. In the combination step, a local region is defined for each observation by selecting the top-n most similar neighbors. Then, the base detector which delivered the best performance in the defined local neighborhood is selected as the competent detector for the observation. This competent detector is used to predict the outlier score for the selected test instance. DCSO focuses on local regions in the data for the computation of outlier scores, hence it can detect local outliers. All the steps and complete flowchart of DCSO are shown in Figure 4.

In the case of EDCV (Ensemble of detectors with correlation votes) and EDVV (Ensemble of detectors with variability votes), the outlier scores of all the algorithms for input samples are stored in a matrix F of size $m \times T$ where m is the number of samples and T is the number of algorithms (components). In the first step, vote matrix V of size $m \times T$ is calculated which contains the number of votes assigned by each algorithm for each data sample. A modified boxplot technique is used for the calculation of votes where a sample gets a vote if its score is greater than 150% of the Inter Quartile Range. In the next step, a weight matrix W is computed based on EDCV and EDVV approaches. In the EDCV method, a correlation coefficient matrix C between the output score F is calculated, and then by using the matrix C the corresponding weights of each component are calculated using the equation

$$W_n = \frac{(\sum_{m=1}^{T} C_m n) - 1}{T - 1}. \tag{3}$$

Similarly in EDVV, a matrix D of mean absolute deviations (MAD) between output scores F is calculated, and later, weights of each component are calculated using

$$W_n = \frac{\sum_{m=1}^{T} D_m n}{T-1}. \tag{4}$$

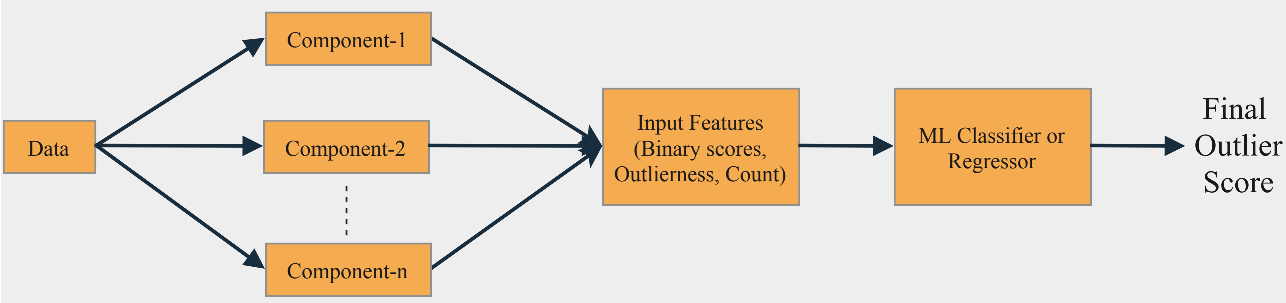
In the last step, the final score of each sample is calculated using the corresponding votes from V and weights from W. So ECVV and EDVV use the correlation and variability between individual components respectively to compute the final ensemble output.

Especially in the early stages of analyses, when labels are often missing, these methods can help to fuse the outputs of multiple components. However, due to large execution times, these methods sometimes have to be run on sub-sampled data sets. In our application, both methods were able to provide slight improvements in the results compared to the best output provided by any single outlier detection algorithm. However, because the performance improvements in our tests were not substantial, we opted to also investigate more complex fusion methods.

6.3 Complex fusion

In this approach, a supervised machine learning model is used as a fusion or combination function. In order to apply this approach, we need information on which observations are actual outliers as targets. We can then apply the fusion method to data, even if we have no information on the targets to isolate outliers. This method is similar to stacking or stacked generalization (Wolpert, 1992; Smyth and Wolpert, 1999; Breiman, 1996). The intuition behind this approach is that the outputs from several components (outlier detection algorithms) for an input sample are fed into another machine learning model to combine them into a single output. Here, the output of each component can be considered as a derived feature. So the derived feature can be a binary output (outlier/inlier), the outlierness score, or both. This fusion model can hence be seen as a meta-classifier/regressor that can use dependencies or identify patterns in prior components output. Figure 5 shows how this approach works. The figure illustrates that the input data is fed into different components of the ensemble, i.e. different outlier detection algorithms. The output prediction of each component takes the form of a binary output and an outlierness score. These outputs (features) are the input data for the ML classifier/regressor in the next stage. Here, the ML classifier/regressor is used for fusion by training the output from the previous step with given labels and predicts the final outlier score for the input sample. The advantage of this approach is that the ensemble methods are capable of learning from component outputs. Any supervised machine learning algorithm can be used as fusion method. Also, these algorithms do not rely on the diversity in previous components output but can identify patterns. Due to their learning capabilities, these methods were able to outperform all previous approaches substantially.

Figure 5: Complex fusion method

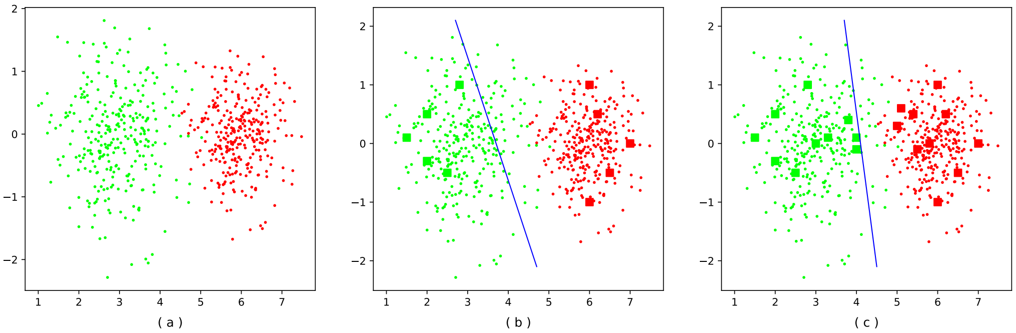

Data
Component-1
Component-2
Component-n
Input Features (Binary scores, Outlierness, Count)
ML Classifier or Regressor
Final Outlier Score

Notes: The figure illustrates the complex fusion method which takes prediction and outlier score output from each component as input. Then it fuses the input and learns the features to predict the final output from the ensemble model.

7 Active learning for outlier detection

Active learning is a unique type of machine learning where a learning model will frequently query the user/-expert for labels of selected samples for better performance (Settles, 2009; Rubens et al., 2011; Das et al., 2020). This method falls into the category of supervised learning in which only a small part of the data is labelled. In this method, human involvement in data labelling is treated as more valuable. This technique is used in cases where a large amount of unlabelled data is available and labelling is expensive (Settles, 2009). Figure 6 explains the advantages of this approach. Furthermore, with this approach, the outlier detection problem is started as unsupervised learning and then can gradually turn it into a supervised learning method.

Figure 6: Active learning



Notes: The figure shows an illustration of active learning. (a) Input unlabelled data consists of two clusters represented by colors green and red. (b) Classification result of active learning model on unlabelled data at the early stages. This approach is an iterative process where each iteration includes selecting few samples from unlabelled data based on the query strategies for the expert query, then labelling the selected samples by experts' feedback and later training the post-processing model (learner) on the labelled samples. These labelled samples are represented as squares. Here the decision boundary is represented as a blue line and is not optimal. (c) Result of active learning model on the unlabelled input data after few iterations. Here the decision boundary is more accurate in separating two clusters in the unlabelled data compared to the previous result due to the iterative learning process.

We designed active learning for outlier detection as an iterative process. In our case each iteration corresponds to a reporting period and is split into two steps (except in first iteration which includes only the first step). During the first step, an unsupervised outlier detection algorithm is applied on a new data set to detect potential outliers. We applied this approach on IFS data. Then, from the predicted output, the top 5% outliers are selected based on the outlierness score of each input sample. In the second step, a pre-trained supervised machine learning model is used which is also called as a post-processing model or a learner. The output from the first step is fed into the post-processing model that selects a number of (e.g.: top 30 or top 100 by output score of the post-processing model) samples to be reviewed by domain experts. The selection of samples for expert feedback depends on the scenario and on the implemented query strategies. We applied active learning in two scenarios and implemented two query strategies which we will discuss below.

- *Stream-Based Selective Sampling* (Lewis and Gale, 1994; Settles, 2009): Each unlabelled sample from a large corpus is drawn one at a time and fed to the learner. Then, the learner will decide whether to request the label of this sample from the expert or to discard it. We implemented this approach with a small modification that if the post-processing model can classify an unlabelled input sample with a score greater than some threshold then the learner itself can assign a label to such a sample, otherwise the sample will be dispatched for the expert query. We set the threshold to 90%. However, we couldn't find much progress with this approach due to the poor performance by the post-processing model in labelling the input samples.

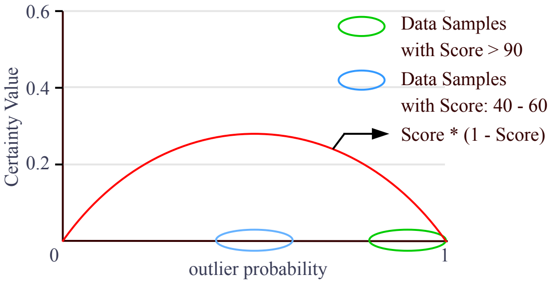
- *Pool-Based Sampling* (Settles, 2009): This is a scenario that is more commonly studied in active learning. In this type, a filter (in our case, an unsupervised model) is applied on a pool of unlabelled data and the samples are ranked based on the returned results. Then from the ranked list the top n samples are selected for the query, where n is a parameter. This parameter is application-specific and reflects the amount of resources that are available for labelling tasks. We tried n with values 30, 100, and 300. The filter is accompanied by a query strategy which we will discuss next.

The two query strategies used for this work are:

- *Certainty sampling* (Settles, 2009): In this query strategy, the samples for which the post-processing model is most certain in its outlier classification are selected for the expert query. Consequently, samples with the highest outlierness score are selected in each iteration (green zone in Figure 7).
- *Uncertainty Sampling* (Settles, 2009; Lewis and Gale, 1994; Pelleg and Moore, 2005): Here, the post-processing model selects such samples for the query for which it is least confident or least certain on how to label. The motivation behind this method is that having expert feedback / labels for the hardest samples will help the learner to improve its performance in the next iteration. Samples selected by this strategy are represented by the blue zone in Figure 7.

The selected samples by the post-processing model are queried for labels. Next, the post-processing model is trained using the new labelled samples and this two-step process is then repeated in each iteration.

Figure 7: Certainty sampling



Notes: The figure shows certainty sampling. Samples in the green zone are the ones for which the post-processing model scores above 90% and the blue zone represents the samples for which the model is least certain about its class.

8 Results and evaluation

The success of the algorithms and their relative performance is, of course, highly domain specific. Depending on the working definition of an outlier in financial data, the performance of the algorithms will differ. We still report detailed results for all algorithms. What is more, we show the performance with and without feature engineering and parameter tuning. The reason is that conditional on a data set and our definition of outliers, the results provide suggestive evidence on the necessity and benefits of feature engineering and parameter tuning as well as an indication of the heterogeneity in performance between different algorithms.

Tables 5 and 6 and Tables 7 and 8 present the evaluation results based on ROC-AUC and PR-AUC for the securities holdings statistics (WpInvest), the investment funds statistics (IFS), the interest rate statistics (ZISTA) and the money market statistics (MMSR) respectively. For each data set, we depcit the baseline results in the first column, the results with feature engineering in the second column, the results with parameter tuning in the third column and - where applicable – the results with feature bagging in the forth column. The table provides three main insights. First, we find that there is profound heterogeneity regarding the success of algorithms in isolating outliers. Second, an algorithm that, in our application, showed a good performance across different data sets is the Isolation Forest. Even without time consuming feature engineering and parameter tuning, the isolation forest provided a good performance relative to other approaches. Without a prior intuition which algorithm successfully isolates outliers in financial data, our findings suggest that Isolation Forests are a good starting point. Third, the importance of feature engineering, parameter tuning and (if applicable) feature bagging for the performance depends on the algorithm and data set. Even for Autoencoders that, due to their complex structure, can discover features, we find that feature engineering can provide sizeable improvements.

Table 5: Results (ROC-AUC) for unsupervised outlier detection algorithms and combination of detectors (WpInvest & IFS)

		Evaluation Results (ROC AUC)							
		WpInvest Total 3.9M samples and Outliers 0.001%				IFS Total 120000 samples and Outliers 4.6%			
		Baseline	+Feature Engineering	+Parameter Tuning	+Feature Bagging	Baseline	+Feature Engineering	+Parameter Tuning	+Feature Bagging
Ensemble (Homogeneous Learners)	Isolation Forest	0.965	0.963	0.982	–	0.597	0.641	0.652	0.638
Distance/Density Based	kNN	0.805	0.882	0.882	–	0.599	0.628	0.628	0.623
	DBSCAN	0.880	0.945	0.945	–	0.554	0.609	0.617	–
	LOF	0.626	0.740	0.740	–	0.570	0.569	0.570	0.652
	OCSVM	–	–	–	–	–	–	–	–
Cluster Based	Autoencoder	0.516	0.552	0.595	–	0.607	0.608	0.665	0.652
	PCA and rPCA	0.780	0.632	0.632	0.591	0.607	0.665	0.669	0.651
	HBOS	–	–	–	–	–	–	–	–
	ARIMA								
Combination of Detectors		(Simple Fusion) 0.9832	(Unsupervised Fusion) 0.954	**(Complex Fusion)** **0.9877**	–	(Simple Fusion) 0.6429	(Unsupervised Fusion) 0.6473	**(Complex Fusion)** **0.7648**	–

Notes: The table shows results (ROC-AUC) of all the unsupervised outlier detection algorithms and Combination of Detectors method for WpInvest and IFS data sets.

Table 6: Results (ROC-AUC) for unsupervised outlier detection algorithms and combination of detectors (ZISTA & MMSR)

		Evaluation Results (ROC AUC)							
		ZISTA Total 4.3M samples and Outliers 0.7%				MMSR Total 2.2M samples and Outliers 0.04%			
		Baseline	+Feature Engineering	+Parameter Tuning	+Feature Bagging	Baseline	+Feature Engineering	+Parameter Tuning	+Feature Bagging
Ensemble (Homogeneous Learners)	Isolation Forest	0.499	0.637	0.639	–	0.925	0.935	**0.935**	–
Distance/Density Based	kNN	0.702	0.733	0.738	–	0.894	0.905	0.905	–
	DBSCAN	–	–	–	–	–	–	–	–
	LOF	0.576	0.733	0.736	–	0.901	0.908	0.908	–
	OCSVM	0.506	0.625	0.626	–	0.921	0.928	0.928	–
Cluster Based	Autoencoder	0.701	0.704	0.708	–	–	–	–	–
	PCA and rPCA	0.714	0.744	**0.745**	–	0.844	0.865	0.865	–
	HBOS	0.695	0.735	0.735	–	0.915	0.928	0.929	–
	ARIMA	0.665	0.665	0.691					
Combination of Detectors		(Simple Fusion) –	(Unsupervised Fusion) –	(Complex Fusion) –	–	(Simple Fusion) –	(Unsupervised Fusion) –	(Complex Fusion) –	–

Notes: Result (ROC-AUC) of all the unsupervised outlier detection algorithms and Combination of Detectors method for ZISTA and MMSR data sets.

Table 7: Results (PR-AUC) for unsupervised outlier detection algorithms and combination of detectors (WpInvest & IFS)

		Evaluation Results (PR AUC)							
		WpInvest Total 3.9M samples and Outliers 0.001%				IFS Total 120000 samples and Outliers 4.6%			
		Baseline	+Feature Engineering	+Parameter Tuning	+Feature Bagging	Baseline	+Feature Engineering	+Parameter Tuning	+Feature Bagging
Ensemble (Homogeneous Learners)	Isolation Forest	0.0008	0.0009	0.0013	–	0.048	0.055	0.056	0.054
Distance/Density Based	kNN	0.0002	0.001	0.001	–	0.064	0.097	0.097	0.082
	DBSCAN	0.0001	0.0007	0.0007	–	0.052	0.092	0.092	–
	LOF	0.00004	0.00006	0.00006	–	0.055	0.057	0.057	0.073
	OCSVM	–	–	–	–	–	–	–	–
Cluster Based	Autoencoder	0.0001	0.0001	0.0002	–	0.063	0.064	0.085	0.082
	PCA and rPCA	0.023	0.0004	0.0004	0.0002	0.063	0.085	0.086	0.082
	HBOS	–	–	–	–	–	–	–	–
	ARIMA	–	–	–					
Combination of Detectors		(Simple Fusion) 0.001	(Unsupervised Fusion) 0.0009	**(Complex Fusion)** **0.003**	–	(Simple Fusion) 0.065	(Unsupervised Fusion) 0.056	**(Complex Fusion)** **0.3441**	

Notes: The table shows results (PR-AUC) of all the unsupervised outlier detection algorithms and Combination of Detectors method for WpInvest and IFS data sets.

Table 8: Results (PR-AUC) for unsupervised outlier detection algorithms and combination of detectors (ZISTA & MMSR)

		Evaluation Results (PR AUC)							
		ZISTA Total 4.3M samples and Outliers 0.7%				MMSR Total 2.2M samples and Outliers 0.04%			
		Baseline	+Feature Engineering	+Parameter Tuning	+Feature Bagging	Baseline	+Feature Engineering	+Parameter Tuning	+Feature Bagging
Ensemble (Homogeneous Learners)	Isolation Forest	0.013	0.018	0.019	–	0.012	0.015	0.015	–
Distance/Density Based	kNN	0.012	0.023	0.024	–	0.012	0.015	0.015	–
	DBSCAN	–	–	–	–	–	–	–	–
	LOF	0.014	0.026	0.027	–	0.015	0.017	0.017	–
	OCSVM	–	–	–	–	–	–	–	–
Cluster Based	Autoencoder	0.071	0.078	**0.079**	–	–	–	–	–
	PCA and rPCA	0.013	0.070	0.072	–	0.013	0.016	0.016	–
	HBOS	0.013	0.022	0.028	–	0.012	0.018	**0.019**	–
	ARIMA	0.015	0.016	0.019	–	–	–	–	–
Combination of Detectors		(Simple Fusion) –	(Unsupervised Fusion) –	(Complex Fusion) –	–	(Simple Fusion) –	(Unsupervised Fusion) –	(Complex Fusion) –	

Notes: The table shows results (PR-AUC) of all the unsupervised outlier detection algorithms and Combination of Detectors method for ZISTA and MMSR data sets.

As mentioned in Section 6, we also applied different combination functions in outlier detection ensembles. We evaluated a total of 16 different combination functions, mainly categorized into three categories. The performance across the three categories and all the methods on the IFS data set are summarized in Table 9. There were some glimpses of performance improvement by simple and unsupervised methods but the difference was not large. Simple and unsupervised fusion methods couldn't outperform the best individual component result at a significant level which might be due to the lack of diversity in output between the individual components used as input to the combination functions. However, the performance of complex fusion methods were promising as they clearly outperformed all other methods by a large margin. The complex fusion methods were able to learn the patterns and relation between each component output for the prediction of output. So there were sizeable improvements in the ROC-AUC and PR-AUC values by complex fusion methods.

Table 9: Comparison of different fusion methods

	Combination functions	**F1-Score**	**PR-AUC**	**ROC-AUC**
	Best individual component	0.1021	0.0857	0.6643
Simple fusion method	Min	0.1195	0.056	0.5804
	Max	0.0834	0.0828	0.6539
	Avg	0.0834	0.0673	0.6601
	Normalize-one-per-Component-Max	0.0991	0.0696	0.6539
	Normalize-one-per-Component-Avg	0.0834	0.0655	0.6598
	Majority_Voting	0.0834	NA	NA
Unsupervised fusion method	DCSO-AVG	0.1092	0.063	0.588
	DCSO-MAX	0.115	0.0799	0.6273
	DCSO-AOM	0.0913	0.0629	0.6389
	DCSO-MOA	0.1109	0.0889	0.6336
	EDCV	0.0834	0.0650	0.6591
	EDVV	0.0834	0.0676	0.6595
Complex fusion method	SVM-Model	0.1123	0.074	0.6412
	LR-Model	0.1068	0.0923	0.6494
	KNN-Model	0.2912	0.3042	0.7045
	RFC-Model	**0.3515**	**0.3148**	**0.7734**

Notes: The table shows a comparison of results if we apply different fusion methods to a combination of detectors using the IFS data set. Complex fusion method outperformed the other fusion methods as well as best individual component used for outlier ensemble.

Regarding active learning, for the IFS data set, we observed slight improvements using either query strategies (see Table 10). Uncertainty sampling outperformed certainty sampling in all considered performance metrics.

Table 10: Comparison of query strategies

Query strategy	**Avg PR-AUC**	**AVG ROC-AUC**	**Avg Precision**	**Avg Recall**
Certainty sampling	0.3131	0.6077	0.1372	0.4772
Uncertainty sampling	**0.379**	**0.7221**	**0.3278**	**0.5925**

Notes: The table shows a comparison of query strategies using IFS data set.

Taken together, we find that there is no "one size fits all" solution to outlier detection in the financial data sets that we evaluated. Complex fusion can help to overcome the necessity of selecting a single approach by rendering the selection of an algorithm an empirical exercise. Therefore, although fusion did not dramatically improve the performance relative to the best single algorithm, combination of detectors with a fusion method can be helpful in some contexts. To go beyond a strictly data-driven isolation of outliers, active learning provides possibilities to take domain knowledge into account that goes beyond the detection capabilities of the algorithms.

9 Explainable AI

If our sole interest lies in successfully flagging outliers and we are confident that optimization with cross validation (see Section 2.2) leads to internally and externally valid results, we can treat the outlier detection algorithms as a black box.[11] However, in many business applications, we need a better explanation of our decision. If we apply outlier detection in DQM, we often need an explanation for data users why an observation was excluded that goes beyond referring to the decision of an algorithm. The same holds if we make inquiries at reporting agents regarding data points that were flagged as outliers by an algorithm. Likewise, to use outlier detection algorithms to uncover – economically meaningful – unusual structures in the data, we need better insight into what distinguishes an outlier from an inlier.

To provide such explanations, we can either use a transparent model that we can interpret directly to detect outliers, such as a histogram, or we can use a surrogate model that provides an explanation for a more complex algorithm's classification[12]. The idea of a surrogate model is to treat the prediction of a more complex algorithm as an outcome and use an algorithm to model the relationship between the outcome and the input features. We discuss methods that are applied globally in Section 9.1 and techniques that are applied locally in Section 9.2. Then. we further develop ways to explain the results of outlier detection with Autoencoders in Section 9.3.

To motivate the need for local explanations, the following briefly explains the difference between local and global explanations in the context of our use case.

9.1 Global methods

Global methods are applied to the full data set and include linear regressions and decision trees on which we elaborate below.

9.1.1 Linear regression

One method for approximating the decision function of a 'black box' model is to approximate it linearly; that is, to estimate the output of the non-linear model as a linear function of the inputs. To do this, we can estimate a linear regression, with the original model input as the independent variables and the model's output (either a probability score or class based on the probability score) as the dependent variable. We can then make inferences about the model's decision function based on the coefficients of the linear surrogate model. The r-squared, or explained variance, of the surrogate model tells us how well the surrogate model approximates the original model.

While this approach has the advantage that it is highly flexible and easily interpretable via the estimated coefficients, the disadvantages are manifold. First, if the surrogate model does a very good job of explaining the original model – for example with an R-squared of 98 percent – then it would behove us simply to use the linear model in the first place. If it does not explain the surrogate model well, then we cannot rely on the explanations it provides us with. Especially for our use case, where we want to explain rare and anomalous instances, they must necessarily not be able to be well-explained by a linear model.

It is important to reiterate that the estimates from a linear regression surrogate model, and any other surrogate model for that matter, are making inferences about the model, not about the data itself. Thus, unlike the original black box model itself which is an abstraction of the real data-generating process, the surrogate model is in turn an abstraction of the black box model.

[11] Following Patino and Ferreira (2018), internal validity is defined as the extent to which the observed results represent the truth in the population we are studying and, thus, are not due to methodological errors. Once the internal validity of the study is established, the researcher can proceed to make a judgment regarding its external validity by asking whether the study results apply to similar patients in a different setting or not.

[12] For a discussion of common approaches to explainable AI, see e.g., Molnar (2019)

9.1.2 Decision trees

One simple and intuitive method to understand the outlier scores of outlier detection algorithms is to estimate and visualize a decision tree.

An advantage of the decision tree over linear regression is that it is more flexible, i.e., it can capture non-linear relationships in the data and can model interactions between input variables that are not additive separable. To implement this method, the outlierness score and the data features are used to estimate a decision tree regressor where the outlierness score serves as the label.

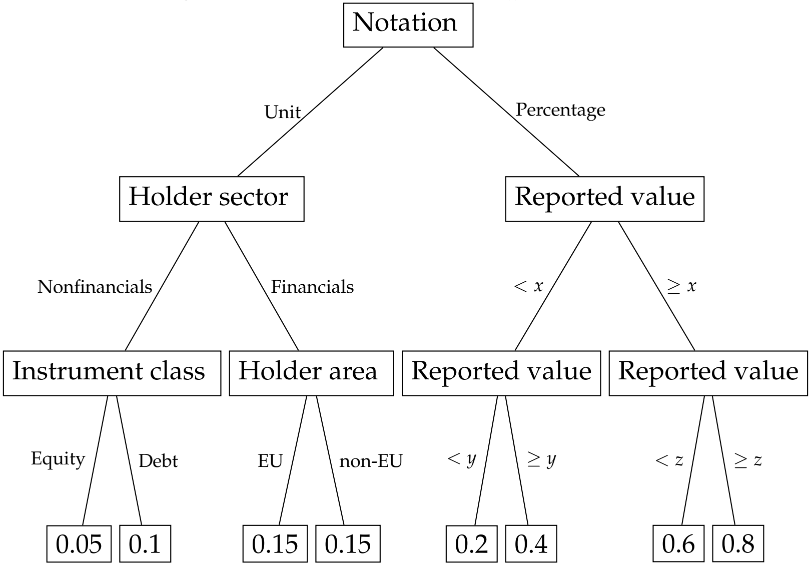
A decision tree highlights the most important features (in terms of entropy) for the outlier detection algorithm. In other words, a decision tree highlights those features that are most relevant for the attribution of a high or low outlierness score of a certain data point. It further delivers interpretable decision rules for these features that can be understood by applying some domain knowledge.

Alternatively, one can estimate a decision tree classifier using the predicted outliers as labels. The tree should be specified relatively simple, with only a small depth, so that it can be interpreted and visualized more easily.

To illustrate this method we apply it to the WpInvest data. Each observation in this data set consists of the aggregate amount that a bank holds for a client in custody for each security, holder sector, and holder area. Let A be a classifier that has been trained to detect reporting errors. Given a set of prediction scores for a reporting period by classifier A, we estimate a decision tree (see Figure 8 for a simplified representation). With the help of the decision tree, one could say that the trained classifier A assigns higher outlierness scores to observations where `Holder sector = Financials` and the reported values are large both in the previous and in the current period.

The illustration in Figure 8 can therefore be interpreted as follows: Since we normalized the outlierness scores to the interval $[0, 1]$ the numbers at the end of the tree can be interpreted as probabilities. Hence, if the notation is percentage notation and the reported value is larger than the respective cutoff values, then in 80% of cases the data point is classified as an outlier by the outlier detection algorithm (right most path in Figure 8). If the notation is unit notation, the holder sector is financials and the holder area is EU, then the data point is only in 15% of cases an outlier according to the outlier detection algorithm. This logic applies equally to all leafs of the tree.

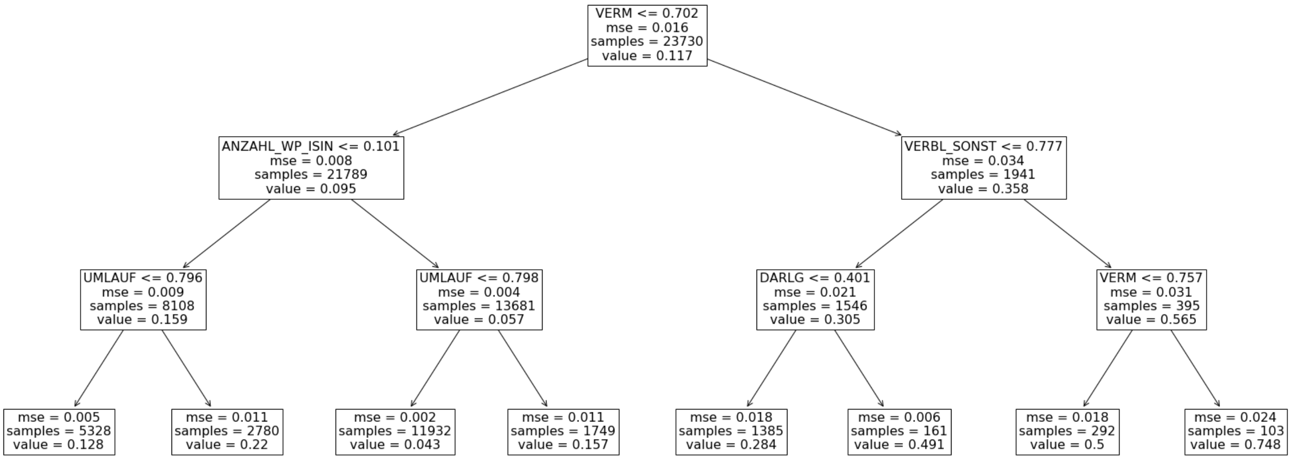
Figure 8: Decision trees for Explainable AI



Notes: The figure illustrates the use of decision trees for explainable AI. A stylized representation of a decision tree that was estimated using outlierness score and the data features. A leaf node gives the outlierness score that the classifier likely assigns to an instance. x, y and z represent cutoff values in the decision tree in case of continuous features.

Next, we apply decision trees to the IFS data to better understand the results of the outlier detection algo-

Figure 9: Example of a decision tree using IFS data



Notes: The figure shows an example of a decision tree using IFS data[13]. Underlying outlier detection model is baseline isolation forest.

rithm. In Figure 9, we plot the decision tree for our baseline Isolation Forest algorithm in the IFS data. The interpretation of the tree is the same as discussed earlier in this section. We find that in the IFS data funds with unusually large non-standard balance sheet positions like other equity and (other) liabilities (VERM and DARLG, VERBL_SONST) are more likely to be classified as outliers by the algorithm. This can be seen on the right side of the figure where funds with other equity positions greater than 0.757 and other liabilities greater than 0.777 get assigned a very high probability of being classified as outliers by the algorithm. This is an interesting economic relation that the algorithm detects because funds with large liabilities or non-standard equity positions are relatively uncommon. The result also suggests that outlier detection in the IFS is capable of detecting economically anomalous data points and that this particular model is less focused on data quality issues concerning funds. As one can see from this example, the decision tree is a very helpful tool to understand the estimated model and on which aspects of the data the algorithm is focused. With different input features or alternative models, the result could be very different leading to alternative insights about the data.

9.2 Local methods

If we have estimated a well-performing model which indicates that a particular instance is anomalous, the model has likely detected that this instance is meaningfully different from similar instances. For example, if a security has a market value in euros of €10 billion and is nominally denominated in Egyptian pounds, whilst all other securities of €10 billion are denominated in euros, the model might predict that this instance is an outlier; given the neighbourhood (market value in euros), the reported nominal currency is anomalous. This simplified example illustrates the need to estimate a surrogate model with input instances that are similar to the instance that is to be explained.

This need to have an explanatory method which is locally valid is one component of a more general framework for instance-specific explanatory methods. Although there is no single agreed-upon definition of what constitutes a good explanation, Table 11 summarises commonly used standards for valid local explanatory models. [14]

[13] `VERM` is other equity, `ANZAHL_WP_ISIN` denotes the number of ISIN-securities in the portfolio, `VERBL_SONST` is the amount of other liabilities, `UMLAUF` gives the units outstanding, and `DARLG` are loans to property companies.

[14] For a broader discussion, see Alvarez-Melis and Jaakkola (2018), Antwarg et al. (2019), and Lundberg and Lee (2017)

Table 11: Summary of common characteristics of a valid explanation

Local accuracy / faithfulness	An explanation should be accurate within the local proximity of the instance in question
Missingness	If a feature value is missing, it should receive a weight of zero in the explanation model
Explicitness / intelligibility	If a model changes, and the contribution of a feature in the new model increase relative to the old model, the surrogate model's attribution to that feature should not decrease. In other words, the explanation is consistent with human intuition.

There are only two (related) methods that satisfy at least two of these axioms. These methods two are described below.

9.2.1 Local interpretable model-agnostic explanations (LIME)

Following up on the idea that a linear regression provides a good linear approximation, even for nonlinear relationships in the data, LIME aims to estimate a regression in the local area around an instance to approximate the contribution of input features to the output value for that instance. The idea is very similar to that of a kernel regression.

For a given instance, LIME takes random samples of instances from the input space and perturbs them. These perturbed instances are plugged into the model and a linear regression is estimated on the resulting output. Importantly, the weight that each perturbed instance received in the regression is based on exponential smoothing kernel.

The choice of kernel is a major drawback for LIME, as with tabular data with possibly many binary variables, different distance kernels can lead to very different explanations (see Molnar (2019)). This drawback is addressed with the second method.

9.2.2 Shapley additive explanations (SHAP)

Directly picking up on drawbacks inherent in LIME's weighting function, SHAP uses a concept from game theory — Shapley values — to attribute to each feature their "fair" weight in local surrogate regression model.

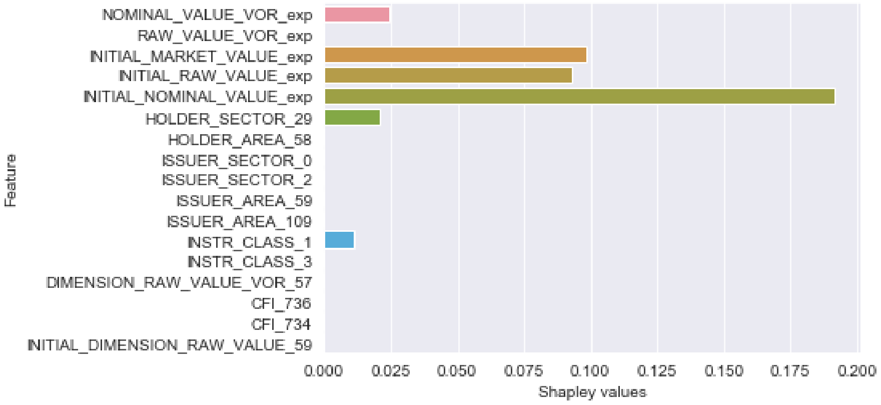
Originally used to fairly attribute payoffs to players in a multiplayer game, Shapley values are a permutation-based method repurposed to provide explanations that satisfy the conditions stated above. Briefly summarised, Shapley values are the average marginal contribution of each player to the output in a multiplayer game as they are present or absent in all possible coalitions with other players, therefore rendering its computation expensive. Yet a sufficient estimation of Shapley values can be obtained by focusing on small and large coalitions. SHAP uses this insight to create the SHAP Kernel, which uses the weights that each coalition would receive in the Shapley value calculation to weight the perturbed instances for the local surrogate regression model. Thus, SHAP provides us with an explanation method with strong theoretical grounding and all the desirable properties outlined in Table 11.

9.2.3 Applications of local explanations

Having local explanations of individual instances is useful in many parts of the machine learning pipeline. Below we identify two use cases in which we found local explanations to be helpful.

Use case 1: Individual explanations for business experts and end users This is one of the most common uses for individual explanations, and one we encountered often during the model validation phase. For example, after we had a working model that was showing good results on the partially labelled data we had available, we obtained predicted outliers on unlabelled validation data and wanted to have domain experts evaluate our predictions. Given the large number of features in the data set, there are many ways in which a particular data point may be anomalous. We wanted to be able to provide guidance on what – in particular – our model found anomalous with a particular instance. To this end, we applied SHAP to obtain instance-level feature attribution for a sample of highly anomalous instances. Figure 10 shows the Shapley values for one particularly anomalous instance.

Figure 10: SHAP explanations for a single instance



Notes: The figure shows SHAP explanations for a single instance. On the x-axis are the SHAP values, and on the y-axis are the input features into the original model. The figure shows how the features contribute (across all alternative "coalitions" of feature values) to the outlier score of the instance. Thereby, SHAP allows us to interpret a single feature's importance not only in the relation to the realization of other features' values that characterize the instance but also for alternative values of the remainder of features.

The Shapley values clearly point to three features as contributing most to the outlieredness of the instance: INITIAL_MARKET_VALUE, INITIAL_RAW_VALUE, and INITIAL_NOMINAL_VALUE. After consulting with business experts, this instance was indeed anomalous due to an incorrect reporting of the INITIAL_RAW_VALUE feature. This exercise also pointed out a limitation of the model: in so far as the input features are correlated[15], as is the case with all three of these features, the SHAP and LIME will not be able to distinguish between them.

Providing this additional information to domain experts is an improvement over a simple outlier-inlier indicator, and additionally help researchers understand the behaviour and edges of their model.

Use case 2: Feature influence During the development of the model, it is often illuminating to understand how the model's output is influenced by certain input features. This in turns can help developers with further tuning and feature engineering. Feature dependence plots on feature values and their corresponding Shapley values from the model help to do this.

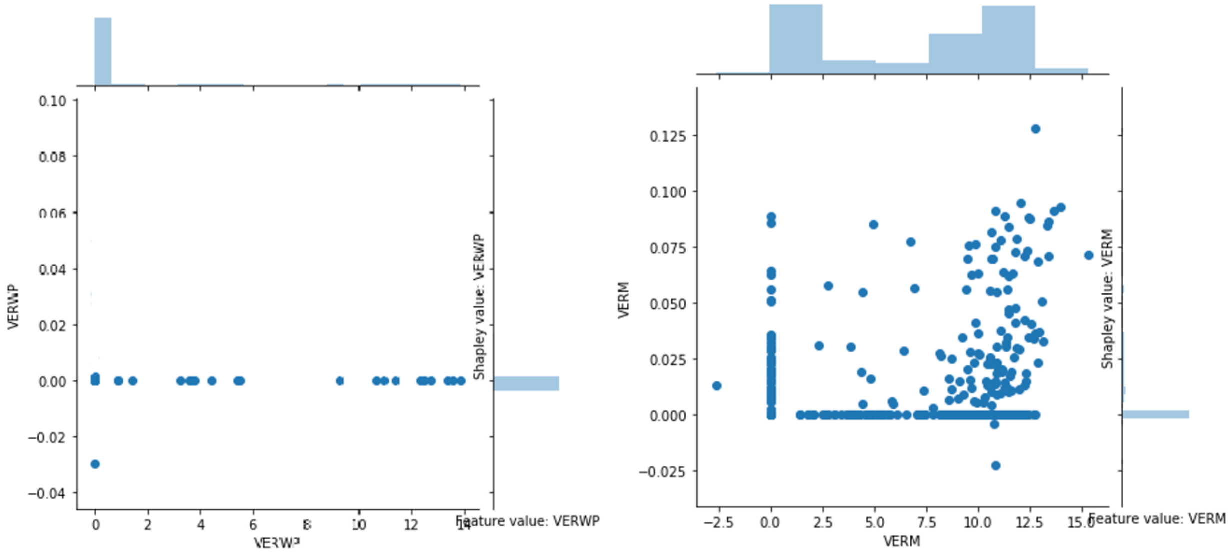
Figure 11 shows the SHAP explanations for two features in a model, and along with their respective feature values. The left panel shows that the feature, VERWP, is unimportant for this model's predictions. The right panel on the other hand indicates that this feature, VERM, is moderately important for the model, where values of zero are also meaningful for some instances. Without these explanation methods, researchers are often left to deduce which features are important by observing the models post-hoc performance with and with a particular set of features. In the unsupervised setting, this is not possible and thus such explanations are very useful for model comparison and diagnostics.

9.2.4 Evaluation

Although these methods are very useful additions to the unsupervised toolbox, they are both costly in terms of computation time, and in the case of LIME as discussed above, considerable practical downsides when working with tabular data.

[15] MARKET_VALUE is the market value of the holdings; NOMINAL_VALUE is nominal or book value of the holdings; RAW_VALUE is the originally reported book value of holdings (i.e. original currency, etc). The INITIAL means that this was the value of the first reporting of the bank, which may have been subsequently changed. See Blaschke et al. (2020) for more details on the data.

Figure 11: Individual model explanations and corresponding feature values



Notes: The figure shows individual model explanations and corresponding feature values. Both figures show the SHAP values on the y-axis and their corresponding feature values on the x-axis. On the left, the SHAP values are not meaningfully different than zero, whereas for the subfigure on the right, larger values of the feature are associated with larger SHAP values. Data source: IFS Jan-2019

In terms of runtime, Table 12 summarises an experiment computing local explanations for the top one hundred instances of a data set by predicted outlier score using LIME and SHAP.

Table 12: Test runtimes for 100 individual local explanations

Method	Runtime
LIME	3m31s +/- 13s
KernelSHAP	11m22 +/- 44s

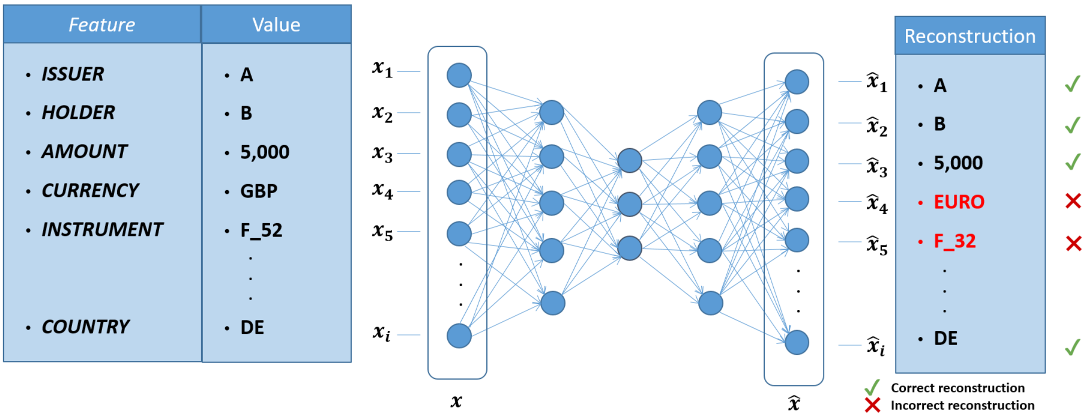
That KernelSHAP is computationally expensive is certainly an important consideration when deciding at what point in the model development process such explanations warrant the time to compute them. For early stages of model development, global surrogate models may suffice to give researchers rough intuition as to feature importance. For later stages, such as interfacing with business experts or regulatory stakeholders, SHAP and to a lesser extent LIME are a worthwhile tool to reach for.

9.3 Autoencoder neural network

In the last decades, neural-network-type models have shown a remarkable progress in various domains. However, understanding the decision making process of such complex models remains a challenging task for domain experts. In the following, we outline our approach to providing explanations for outliers that were detected with an autoencoder.

The goal of the Autoencoder Neural Network is to perform a lossy compression of the data into a lower dimensional space (encoder) and to reconstruct the data in its original dimensionality as accurately as possible (decoder). Because reconstruction is imperfect, the deviation of the reconstructed data from the original data – the reconstruction error – reflects the success of the model in reconstructing a sample. Because the reconstruction error is tightly linked to how well a sample fits into the structure of the data that is preserved in the lower-dimensional space, it provides us with a measure of the outlyingness of a sample. Consequently, we can use the reconstruction error to separate *inliers* that follow a common pattern (low reconstruction error) from *outliers* that deviate from the common structure of the data (high reconstruction error). One challenge of this approach is that the reconstruction error – a single scalar – is not sufficient to answer the question why a sample is flagged as an outlier. To provide an explanation why an observation was flagged as an outlier we study the reconstruction error and its properties in more detail. Instead of collecting the reconstruction errors on the instance level we collect them on the attribute level. In particular, we unfold the reconstruction error of an instance and study corresponding reconstruction errors per individual attribute. This way we are able to identify whether a particular field was reconstructed or not. The subset of fields that were not reconstructed trigger high reconstruction error of an instance. As a result this subset of entries are most likely contain structurally unusual pattern. Figure 12 depicts a schematic overview of the reconstruction on the attribute level. In this example two attributes (*CURRENCY* and *INSTRUMENT*) were reconstructed incorrectly. Moreover, given the values of the other attributes the model predicts that the *CURRENCY* should be ’EURO’ instead of ’GBP’ and *INSTRUMENT* should be ’F_32’ instead of ’F_52’. Therefore, there is a high chance that these two fields contain an error and as a result have to be screened. Technically this is achieved by applying the softmax function on the final/output layer of the decoder network per categorical attribute. Since we use the one-hot encoded representation of a categorical attribute the result of the softmax provides the normalized scores for all categories of an attribute which can be used as a probability estimates. Finally the element with the highest probability score is selected as the prediction category. If such category differs from the corresponding category of the input instance then the field is flagged as incorrectly reconstructed. Such methodology also provides an opportunity to the domain expert to order potential reporting errors correspondingly and start the auditing process from such field(s) that most-likely contain the most *severe* error(s).

Figure 12: Correct and incorrect reconstructions



Feature	Value
ISSUER	A
HOLDER	B
AMOUNT	5,000
CURRENCY	GBP
INSTRUMENT	F_52
⋮	⋮
COUNTRY	DE

	Reconstruction	
$\hat{x}_1$	A	✓
$\hat{x}_2$	B	✓
$\hat{x}_3$	5,000	✓
$\hat{x}_4$	EURO	✗
$\hat{x}_5$	F_32	✗
⋮	⋮	
$\hat{x}_i$	DE	✓

Notes: The figure provides a schematic overview of the correct and incorrect reconstructions on attribute level of the Autoencoder Neural Network. Each field is flagged correspondingly based on the reconstruction errors collected per attribute field.

Such technique allows us to flag a set of attribute fields that affect the reconstruction quality of an instance at most. In other words, the combination of attributes that were reconstructed incorrectly appears to become an anomaly pattern. Most likely this set of attribute values is what makes the sample anomalous and as a result some of these fields might contain an error. We believe that such features provide more detailed explanation of a particular decision(s) made by the Autoencoder Neural Network and could serve as an important supplement to the domain experts’ toolbox.

10 Conclusion

Steadily rising data volumes and an increasing complexity of statistical reporting of micro data led to a surge in the interest of statistics departments to employ statistical learning methods from the fields of data science and machine learning to provide data user with the highest possible data quality. Reducing the burden on the reporting agents in the data quality management process and achieving an overall higher operational efficiency of statistics departments with minimal (costly) human input are equally important goals when moving to data science and machine learning methods.

In this paper, we outlined the steps that we took to implement a prototype that is capable (i) to detect outliers on an unsupervised basis (ii) to provide explanations of data points that seems suspicious, and (iii) to incorporate the feedback of domain experts in the process. We apply our pipeline to data sets that are collected by the Bundesbank and cover the structure and format of a wide range of financial data, including interest rates, money market statistics, sectoral securities holdings, and investment fund holdings. In addition, since we had information on previous reporting errors of all the aforementioned statistics, we were able to evaluate the performance of the various unsupervised algorithms in detecting unusual data points.

We conclude that unsupervised learning algorithms, applied to granular financial data that was collected by a central bank, are not only suitable to detect incorrect reporting and thereby improve the data quality, but, in conjunction with explainable AI, can also provide explanations for what distinguishes outliers from inliers. However, our work also stresses that a production pipeline that is largely automated and that provides the possibility to actively incorporate (human) feedback is at least as important as a proper selection of algorithms.

References

Aggarwal, C. C. (2012). Outlier ensembles: position paper. *SIGKDD Explorations*, 14(2):49–58.

Aggarwal, C. C. (2015). Outlier analysis. In *Data mining*, pages 237–263. Springer.

Aggarwal, C. C. and Sathe, S. (2015). Theoretical foundations and algorithms for outlier ensembles. *SIGKDD Explor. Newsl.*, 17(1):24–47.

Aggarwal, C. C. and Sathe, S. (2017). *Outlier Ensembles: An Introduction*. Springer Publishing Company, Incorporated, 1st edition.

Ahmed, M., Naser Mahmood, A., and Hu, J. (2016). A survey of network anomaly detection techniques. *J. Netw. Comput. Appl.*, 60(C):19–31.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. *CoRR*, abs/1806.07538.

Angiulli, F. and Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. In *European conference on principles of data mining and knowledge discovery*, pages 15–27. Springer.

Antwarg, L., Shapira, B., and Rokach, L. (2019). Explaining anomalies detected by autoencoders using SHAP. *CoRR*, abs/1903.02407.

Aytekin, C., Ni, X., Cricri, F., and Aksu, E. (2018). Clustering and unsupervised anomaly detection with l_2 normalized deep auto-encoder representations. In *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6.

Bade, M., Doll, H. C., Hirsch, C., Hubrich, A., and Schulz, F. (2019). Money Market Statistical Reporting, Data Report 2019-08 – Metadata Version MMSR-Data-Doc-v1-0. *Deutsche Bundesbank, Research Data and Service Centre*.

Bade, M. and Krueger, M. (2019). MFI interest rate statistics, Data Report 2019-05 – Metadata Version 3. *Deutsche Bundesbank, Research Data and Service Centre*.

Bhuyan, M. H., Bhattacharyya, D. K., and Kalita, J. K. (2014). Network anomaly detection: Methods, systems and tools. *IEEE Communications Surveys Tutorials*, 16(1):303–336.

Blaschke, J. and Haupenthal, H. (2020). Investment Funds Statistics Base, Data Report 2020-05 – Metadata Version 3-1. *Deutsche Bundesbank, Research Data and Service Centre*.

Blaschke, J., Sachs, K., and Yalcin, E. (2020). Securities Holdings Statistics Base plus, Data Report 2020-14 – Metadata Version 3-1. *Deutsche Bundesbank, Research Data and Service Centre*.

Breiman, L. (1996). Stacked regressions. *Mach. Learn.*, 24(1):49–64.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104.

Chalapathy, R. and Chawla, S. (2019). Deep Learning for Anomaly Detection: A Survey. *CoRR*, abs/1901.03407.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15.

Dang, X. H., Micenková, B., Assent, I., and Ng, R. T. (2013). Local outlier detection with interpretation. In Blockeel, H., Kersting, K., Nijssen, S., and Železný, F., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 304–320, Berlin, Heidelberg. Springer Berlin Heidelberg.

Das, S., Wong, W.-K., Dietterich, T., Fern, A., and Emmott, A. (2020). Discovering anomalies by incorporating feedback from an expert. *ACM Trans. Knowl. Discov. Data*, 14(4).

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Ernst, M. and Haesbroeck, G. (2017). Comparison of local outlier detection techniques in spatial multivariate data. *Data Min. Knowl. Discov.*, 31(2):371–399.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231.

Goldstein, M. and Dengel, A. (2012). Histogram-based Outlier Score (HBOS): A Fast Unsupervised Anomaly Detection Algorithm. *Poster and Demo Track of the 35th German Conference on Artificial Intelligence (KI-2012)*, pages 59–63.

Goldstein, M. and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126.

Junttila, J. (2001). Structural breaks, arima model and finnish inflation forecasts. *International Journal of Forecasting*, 17(2):203–230.

Khoa, N. L. D. and Chawla, S. (2010). Robust outlier detection using commute time and eigenspace embedding. In Zaki, M. J., Yu, J. X., Ravindran, B., and Pudi, V., editors, *Advances in Knowledge Discovery and Data Mining*, pages 422–434, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kriegel, H. ., Kroger, P., Renz, M., and Wurst, S. (2005). A generic framework for efficient subspace clustering of high-dimensional data. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8 pp.–.

Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. *CoRR*, abs/cmp-lg/9407020.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.

Lundberg, S. and Lee, S. (2017). A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Molnar, C. (2019). *Interpretable Machine Learning*. lulu.com. https://christophm.github.io/interpretable-ml-book/.

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198.

Pasillas-Díaz, J. and Ratté, S. (2016). An unsupervised approach for combining scores of outlier detection techniques, based on similarity measures. *Electronic Notes in Theoretical Computer Science*, 329:61–77.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Patino, C. M. and Ferreira, J. C. (2018). Internal and external validity: can you apply research study results to your patients? *Jornal brasileiro de pneumologia*, 44:183–183.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pelleg, D. and Moore, A. W. (2005). Active learning for anomaly and rare-category detection. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 1073–1080. MIT Press.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45.

Prieditis, A. and Russell, S. J., editors (1995). *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*. Morgan Kaufmann.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438.

Rokach, L. (2010). Ensemble-based classifiers. *Artif. Intell. Rev.*, 33(1–2):1–39.

Rubens, N., Kaplan, D., and Sugiyama, M. (2011). *Active Learning in Recommender Systems*, pages 735–767. Springer US, Boston, MA.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). *Learning Internal Representations by Error Propagation*, page 318–362. MIT Press, Cambridge, MA, USA.

Sarfraz, M. S., Sharma, V., and Stiefelhagen, R. (2019). Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943.

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., and Platt, J. (1999). Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 582–588, Cambridge, MA, USA. MIT Press.

Schreyer, M., Sattarov, T., Borth, D., Dengel, A., and Reimer, B. (2017). Detection of anomalies in large scale accounting data using deep autoencoder networks. *CoRR*, abs/1709.05254.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Smyth, P. and Wolpert, D. (1999). Linearly combining density estimators via stacking. *Machine Learning - ML*, 36:59–83.

Tissot, B., Widjanarti, A., Zulen, A. A., Ari, H. D., and Wibisono, O. (2018). The use of big data analytics and artificial intelligence in central banking. *IFC Bulletin*, 50.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241 – 259.

Zhang, Y., Meratnia, N., and Havinga, P. (2007). A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets. *Rap. tech., Centre for Telematics and Information Technology University of Twente*.

Zhang, Y., Meratnia, N., and Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *Communications Surveys and Tutorials, IEEE*, 12:159 – 170.

Zhao, Y. and Hryniewicki, M. K. (2018). DCSO: Dynamic Combination of Detector Scores for Outlier Ensembles. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection De-constructed (ODD v5.0)*.

Zhao, Y., Hryniewicki, M. K., Nasrullah, Z., and Li, Z. (2018). LSCP: locally selective combination in parallel outlier ensembles. *CoRR*, abs/1812.01528.

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7.

Zimek, A., Campello, R. J., and Sander, J. (2014). Ensembles for unsupervised outlier detection: Challenges and research questions a position paper. *SIGKDD Explor. Newsl.*, 15(1):11–22.

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Restoration of omissions in the quarterly indicators of financial statements for the other financial institutions in the Bank of Russia[1]

Anna Borisenko, Denis Koshelev, Petr Milyutin and Alieva Piruza,
Central Bank of the Russian Federation

[1] This presentation 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 central banks and other institutions represented at the event.

Restoration of omissions in the quarterly indicators of financial statements for the Other Financial Institutions in the Bank of Russia

Piruza Alieva, Anna Borisenko, Petr Milyutin, Denis Koshelev

Abstract

Quarterly financial accounts and sectoral balance sheets' statistics in the context of financial instruments and sectors of the economy, formed on the basis of microdata, is a reliable information basis for a comprehensive and deep macroeconomic analysis. Most organizations in the Financial Corporations sector (S12) report on an annual and quarterly basis, but for some of them, including organizations of the Other Financial Institutions subsector (S125), which perform non-licensed activities, data is only available on an annual basis. On a quarterly basis, only a small part of these organizations' reporting is available. Therefore, to ensure the completeness of the range of companies in the formation of statistics of financial accounts and sector balance sheets, it is necessary to restore gaps in the quarterly indicators of financial statements of organizations. In this article the results of restoring omissions in the quarterly indicators of financial statements for the Other Financial Institutions subsector (S125) in the Russian Federation, which perform non-licensed activities, are presented. In particular, the results of the traditional methods (regression analysis, individual growth rates, cluster analysis) and Machine learning-based methods, that can be applicable to recover data, such as random forest and generative neural network.

Keywords: financial accounts and sectoral balance sheets' statistics, Other Financial Institutions subsector in the Russian Federation, restorations of omissions in the quarterly indicators, Machine learning-based methods

JEL classification: C8

Contents

Introduction

In accordance with Clause 16.1 of Article 4 of Federal Law No. 86-FZ, dated 10 July 2002, 'On the Central Bank of the Russian Federation (Bank of Russia)', the Bank of Russia develops the methodology for compiling the Russian Federation financial accounts in the System of National Accounts (SNA) and organises the compilation of the Russian Federation financial accounts (hereinafter, SNA financial accounts). The Bank of Russia's obligations to form the measures of the financial accounts and financial balance sheets of the SNA on a quarterly and annual basis are also stipulated in Recommendation No. 8 'Sectoral Accounts' within the G20 Data Gaps Initiative (DGI-II). The Bank of Russia has been publishing the financial accounts and financial balance sheets, being part of the System of National Accounts of the Russian Federation, since 2015 on a quarterly and annual bases. The Bank of Russia relies on the System of National Accounts 2008 (2008 SNA)[1] manual as a conceptual and methodological framework for compiling financial accounts and financial balance sheets.

Statistics provided in quarterly financial accounts and sectoral balance sheets, broken down by financial instrument and economic sector, which are compiled based on microdata, expand the opportunities for enhancing the efficiency and the depth of macroeconomic research. This in turn improve the understanding of interconnections between the real sector and the financial industry of the country's economy.

A significant advantage of SNA financial accounts is the fact that they are an important source of data used to analyse activity in the economic sectors failing to provide detailed information, e.g. in the subsector 'Other financial corporations' (of the sector 'Financial corporations') comprising a large number of organisations not reporting to the Bank of Russia.

The key challenge in compiling statistics on organisations in the subsector 'Other financial corporations' is that, in contrast to the majority of organisations in the sector 'Financial corporations' (S12) regularly reporting to the Bank of Russia on a monthly, quarterly and annual basis, information on all organisations in the subsector 'Other financial corporations' (S125) can be obtained only on an annual basis. The main source of data for compiling statistics on the subsector 'Other financial corporations' is annual accounting (financial) statements producing the main portion of processed statistics (whereas quarterly statements are only submitted by a small number of these organisations).

In this regard, when quarterly SNA financial accounts are compiled, it is essential to have comprehensive information as of quarterly dates when only a part of organisations submit their reporting, as well as to ensure that annual and quarterly statistics are comparable. For this purpose, it is needed to close data gaps in organisations' accounting statements as of quarterly dates.

[1] System of National Accounts 2008 (European Commission, United Nations, Organisation for Economic Cooperation and Development, International Monetary Fund, World Bank).

This paper outlines statistical and machine learning methods which will be used to close data gaps in the measures of accounting statements as of quarterly dates for the subsector 'Other financial corporations' of the sector 'Financial corporations' (hereinafter, OFCs). In particular, the paper considers the results of using the individual growth method and cluster analysis, and describes the currently applied method for restoring missing data (the '1/4' and dynamics extension methods). Among machine learning methods, the authors consider the random forest method for regression, gradient boosting model and the generative adversarial network (GAN).

The authors explore the following balance sheet measures (Form No. 0710001) as the data to be restored on the quarterly bases, because these measures are the key ones for compiling SNA financial accounts:

- Loans (short- and long-term ones),
- Accounts receivable,
- Accounts payable,
- Equity and investment fund shares.

The methods employed were compared based on the following criteria.

1) Discrepancies between restored and actual values in an artificial sample as of 1 January 2017 and 1 January 2020.
2) Interpretable dynamics (identification of organisations accounting for a rise or a decline in a particular measure).

Overview of methods for closing data gaps in quarterly measures of accounting statements

There is a vast number of statistical methods generating quarterly data based on annual values. Specifically, the quarterly financial accounts for the 1950s published by the Federal Reserve System were obtained based on the interpolation method.[2] This method estimates quarterly values of various financial measures in the form of a fixed weighted linear combination of annual values for the periods t-1, t and t+1. Thus, interpolation determines unknown intermediate values in a time series as a linear combination of bordering annual values.

To estimate unknown quarterly values of financial measures, the Federal Reserve System also applies the ratio method.[3] Under this method, the first step is to calculate the ratio R_h for each quarter using known data, according to the formula:

$$R_{t,h} = \frac{x_{t,h}}{\sum_{h=1}^{4} x_{t,h}}$$

[2] Board of Governors of the Federal Reserve System, 2000.

[3] Financial Accounts: History, Methods, the Case of Italy and International Comparisons, 2008, p. 118–122.

where $x_{t,h}$ is the base time series with known quarterly values, $h = 1, \dots, 4$ is the order number of a quarter, and $t = 1, \dots, N$ is the order number of annual values.

The resulting ratio is then used to derive the unknown value of a particular measure for the relevant quarter, according to the formula:

$$y_{t,h} = R_{t,h} y_t$$

where $y_t,\ t = 1, \dots, N$ is a series of annual values.

The above methods are easy to use and are in line with time aggregation limits. However, the quarterly series generated by these methods omit possible dynamics of an unknown quarterly series.

Most central banks use the Chow–Lin method[4] to derive quarterly values from annual figures. This method assumes that a time series of annual figures and the dynamics of quarterly values are strongly correlated with each other and have the same order of integration equal one, which means that the above time series are cointegrated. According to the Chow–Lin method, the coefficients of a linear regression model are estimated with the autocorrelation of order 1 errors using the generalised least squares method. Then, based on the estimated coefficients of the model, the quarterly values of variables are derived.

If the assumption of the correlation between the annual values of a particular measure and its disaggregated values (quarterly values) is rejected, Fernández's method[5] or Litterman's method[6] is then employed. Fernández's method estimates the coefficients of a linear regression model with random errors. Contrastingly, Litterman estimates a linear regression model with the autocorrelation of order 2 errors. It should be noted that Litterman extended the method proposed by Fernández with the autocorrelation coefficient equalling zero.

To restore missing values in the measures of accounting statements as of quarterly dates, it is also possible to apply machine learning methods. Specifically, the generative adversarial network (GAN) is one of the most widespread methods used to close data gaps. The GAN is an algorithm of unsupervised machine learning built on a combination of two neural networks. One of them (the generative network) generates candidates (the generative model), whereas the other evaluates them (the discriminative network) trying to distinguish generated candidates from true data (the discriminative model). Hence, the idea of this method is to produce objects that would resemble true ones. This method is often applied to restore missing parts in images. However, the technique of producing realistic objects can also be applied to generate the vectors whose elements are balance sheet measures in accounting statements corresponding to companies' actual behaviour in the market. The Wasserstein GAN (WGAN), which is an extension to the conventional GAN offering an alternative way to train the model, improves the approximation of the distribution of data

[4] Chow, G. and Lin, A. Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series. The Review of Economics and Statistics 53, 1971, p. 372–375.

[5] Fernández, R. A Methodological Note on the Estimation of Time Series. The Review of Economics and Statistics 63, 1981, p. 471–478.

[6] Litterman, R. A Random Walk, Markov Model for the Distribution of Time Series. Journal of Business and Economic Statistics 1, 1983, p. 169–173.

observed in the training sample. The benefit of the WGAN is that the training process is more stable and less sensitive to the model architecture and the choice of hyperparameter configurations.[7]

It is also possible to find a direct functional dependence between balance sheet measures by using ensemble trees algorithms to solve regression problem. Two of the most popular algorithms used: random forest[8] and gradient boosting[9]. Both models are ensemble decision tree models. In random forest model trees are built independently and final result is mean output of all decision trees. In gradient boosting every new tree helps to correct errors made by previously trained tree, thus, final model is sequentially connected trees. Gradient boosting often is more accurate but prone to overfitting. So, both models were used.

Description of data

Other financial corporations are financial institutions providing financial services, except credit institutions, insurers, pension funds, and financial auxiliaries. OFCs comprise leasing companies, financial holdings, factoring companies, investment companies, mortgage companies, mortgage agents, and other organisations.

A part of OFCs perform activities supervised by the Bank of Russia. These OFCs include pawnshops, microfinance organisations, consumer credit cooperatives and agricultural consumer credit cooperatives, professional securities market participants, and housing savings cooperatives. However, the largest portion of OFCs' financial operations are performed by organisations that are not subject to the Bank of Russia's supervision. This is the group of financial institutions that this paper deals with.

OFCs that are not subject to the Bank of Russia's supervision account for a rather significant portion of information impacting the dynamics of released official statistics on SNA financial accounts. OFCs carrying out activities not supervised by the Bank of Russia account for over 90% of the total balance of OFCs (line 1600 in Form No. 0710001 'Balance sheet' and make more than 70% of their overall number over the entire period under review (Table 1).

[7] Martin Arjovsky, Soumith Chintal, Léon Bottou. Wasserstein GAN, 2017.

[8] Leo Breiman. Random Forests, 2001

[9] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001, p. 1189–1232

Table 1. Portion of the total balance and number of OFCs, whose activities are beyond the scope of the Bank of Russia's supervision, in the total balance and number of all organisations classified as OFCs

	Portion of unsupervised OFCs' total balance in all OFCs' total balance	Ratio of the number of unsupervised OFCs to the overall number of OFCs
01.01.2015	97.34%	72.75%
01.01.2016	97.48%	76.97%
01.01.2017	96.07%	77.31%
01.01.2018	95.29%	83.45%
01.01.2019	99.28%	84.47%
01.01.2020	99.54%	85.52%

The main descriptive statistics for the measures of accounting statements as of annual dates (1 January 2016–1 January 2020) for OFCs not supervised by the Bank of Russia are presented in Table 2.

Table 2. Descriptive statistics for the main measures of accounting statements of the subsector of OFCs not supervised by the Bank of Russia, as of annual dates, mln of rubles

		Maximum	Mean value	Standard deviation
01.01.2016	Accounts receivable	360 629.53	60.44	2 081.14
	Accounts payable	345 204.79	54.96	1 948.86
	Loans	420 355.01	156.40	3 898.93
	Equity and investment fund shares	675 510.8	153.52	4 348.28
01.01.2017	Accounts receivable	713 452.34	69.66	3 479.26
	Accounts payable	724 855.60	66.87	3 454.53
	Loans	764 417.40	166.15	5 484.51
	Equity and investment fund shares	522 092.58	153.92	3 760.17
01.01.2018	Accounts receivable	672 230.23	72.46	3 618.33
	Accounts payable	607 838.35	64.71	3 413.29
	Loans	784 792.92	200.77	7 029.77
	Equity and investment fund shares	570 229.35	175.67	4 211.96
01.01.2019	Accounts receivable	644 381.76	97.04	4 056.20
	Accounts payable	645 739.15	90.86	3 934.06
	Loans	1 079 909.35	274.70	8 979.45
	Equity and investment fund shares	1 003 973.45	249.19	6 820.11

01.01.2020	Accounts receivable	698 177.71	110.31	4 002.05
	Accounts payable	684 351.64	96.69	3 660.11
	Loans	1 022 043.26	325.50	9 538.12
	Equity and investment fund shares	1 163 698.12	297.53	8 302.93

According to Table 2, organisations of the subsector of OFCs whose activities are beyond the scope of the Bank of Russia's supervision are rather heterogeneous over the entire period under review. This is evident from the high values of variation indicators (variance, standard deviation, variation coefficient).

The main source of information for compiling quarterly SNA financial accounts of OFCs whose activities are not supervised by the Bank of Russia is primary statistics submitted according to federal statistical forms No. P-3 'Data on organisations' financial standing' (hereinafter, form No. P-3) and No. P-6 'Data on financial investment and liabilities' (hereinafter, form No. P-6). However, the coverage in the above forms had been low during several years (see Table 3). Loans have the highest coverage ratio, while Equity and Accounts payable – the lowest coverage ratios.

Table 3. Dynamics of the coverage according to forms No. P-3 and No. P-6 of the main financial instruments for OFCs whose activities are not supervised by the Bank of Russia, as of annual dates, %

	01.01.2016	01.01.2017	01.01.2018	01.01.2019	01.01.2020
Accounts payable	4.77	18.60	11.46	10.74	13.38
Accounts receivable	16.67	17.90	14.91	22.67	29.02
Loans	30.76	36.36	31.90	56.60	58.79
Equity and investment fund shares	3.37	15.60	16.16	12.98	24.31

As regards OFCs whose activities are beyond the scope of the Bank of Russia's supervision, data on all these organisations are only available on an annual basis. As to quarterly reporting of these organisations, only a small part of it is available. Chart 1 shows changes in the main financial measures of accounting statements of OFCs whose activities are not supervised by the Bank of Russia.

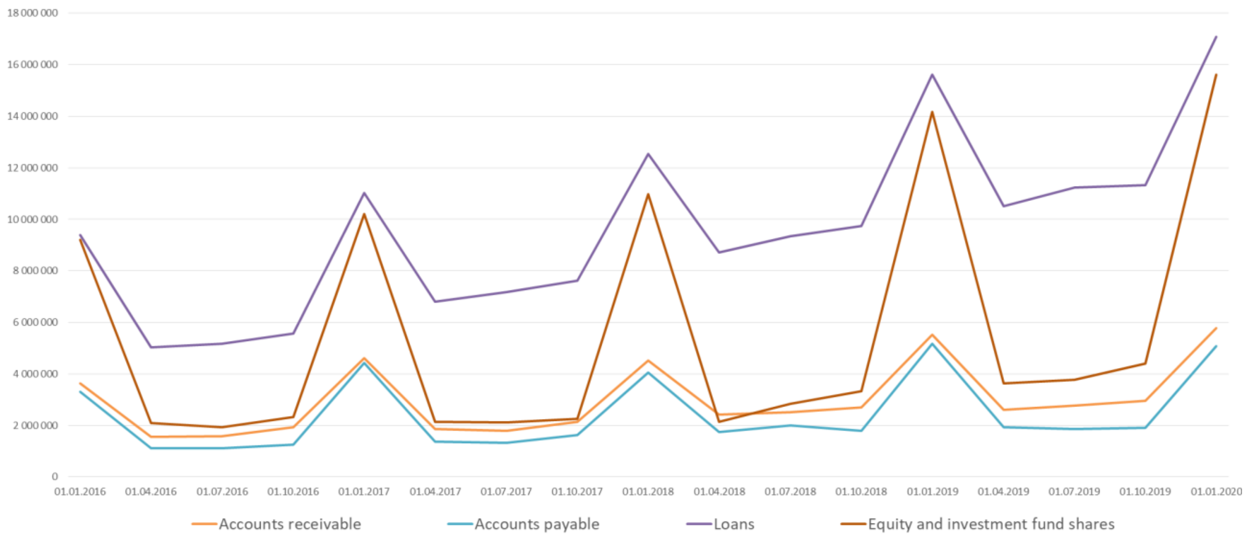


Chart 1. Changes in the main financial measures of accounting statements of OFCs not supervised by the Bank of Russia, 01.01.2016–01.01.2021, mln of rubles

Moreover, the number of OFCs submitting statements as of quarterly dates, including federal statistical forms (forms No. P-3 and No. P-6), is considerably smaller than the number of organisations submitting reporting on an annual basis (see Chart 2). These forms are submitted predominantly by large organisations of the OFC subsector.

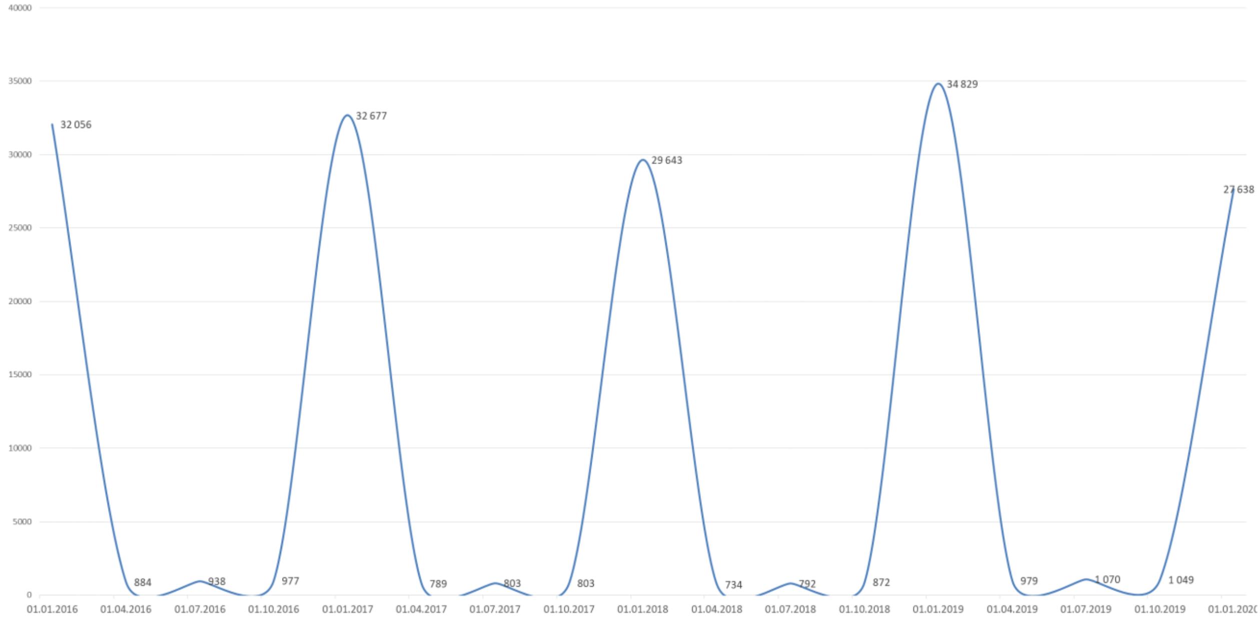


Chart 2. Changes in the number of statements submitted by OFCs not supervised by the Bank of Russia, 01.01.2016–01.01.2020

Chart 2 evidences that OFCs mostly submit statements as of annual dates. According to the analysis of changes in the main measures of accounting statements, the most widespread pattern of gaps over the period from 1 January 2016 to 1 January 2020 is missing data only as of quarterly dates. Specifically, over the period under review, 22.08% of OFCs provided data on Accounts receivable only as of annual dates, 22.04% – on Accounts payable, 20.58% – on Loans, and 27.66% – on Equity.

Chart 3 shows the behaviour of stable organisations of the OFC subsector not supervised by the Bank of Russia, that is, of the organisations that submitted data on the main measures of accounting statements as of all quarterly dates over the period from 1 January 2016 to 1 January 2020. Stable organisations primarily demonstrate bucket dynamics of the main financial measures over the considered period, with declines as of quarterly dates.

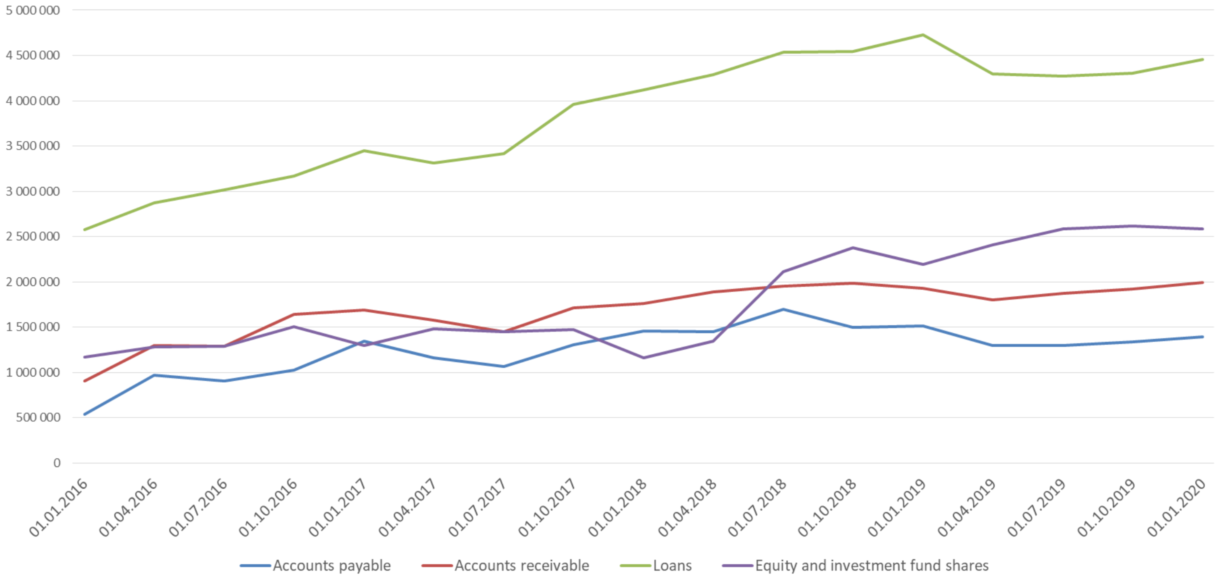


Chart 3. Changes in OFCs' main financial measures, 01.01.2016–01.01.2020, mln of rubles

Results of different methods used to close data gaps

The previous stage of the research made it clear that it is impossible to compile financial balance sheets on a quarterly basis for the subsector of OFCs whose activities are not supervised by the Bank of Russia relying solely on the data submitted to the Bank of Russia. Only a small part of reporting is available on a quarterly basis. Therefore, to obtain comprehensive information on all companies when compiling statistics on financial accounts and sectoral balance sheets, it is necessary to restore missing values in the measures of organisations' accounting statements as of quarterly dates. Below are the main results of various statistical and machine learning methods employed to restore missing data in the measures of accounting statements as of quarterly dates. It should be noted that the analysis encompassed all OFCs, including those subject to the Bank of Russia's supervision, in order to improve the quality of data restoration.

Cluster analysis

As evident from the analysis of descriptive statistics for the main measures of accounting statements, the subsector of OFCs whose activities are not supervised by the Bank of Russia is highly heterogeneous. For this reason, k-means cluster analysis was carried out at the first stage. The scree test formed 11 clusters (see Chart 4).

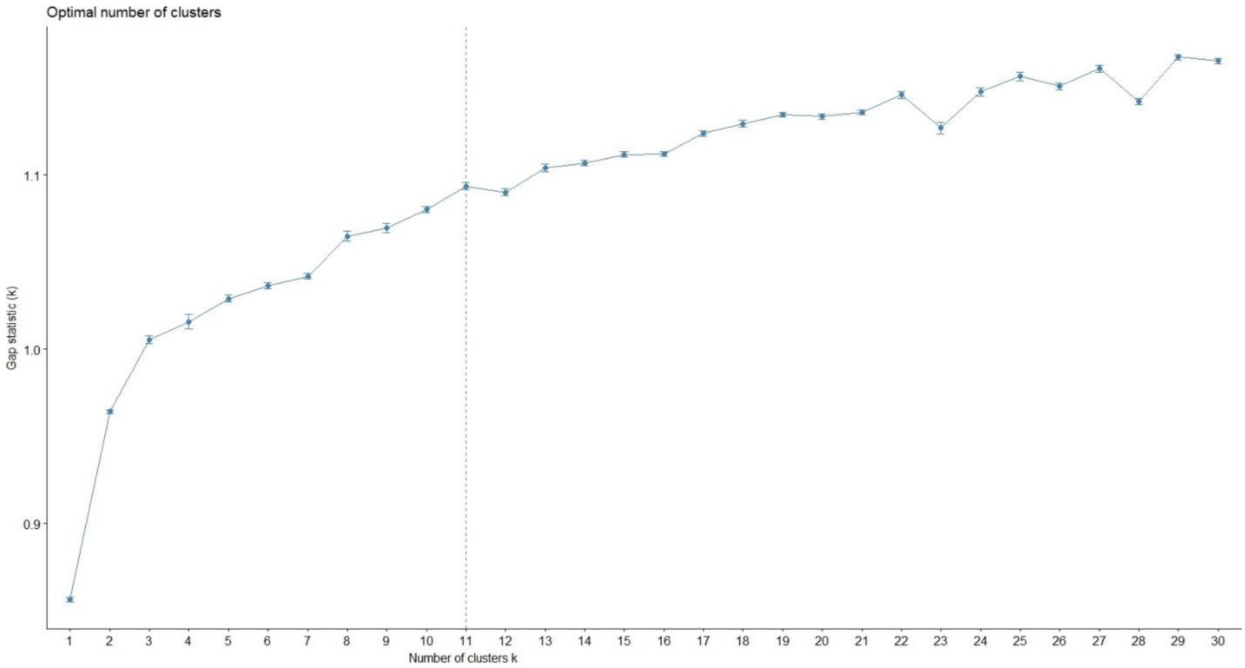


Chart 4. Determining the optimal number of clusters

The variation coefficient for the measures of accounting statements for OFCs not supervised by the Bank of Russia by cluster is presented in Table 4.

Table 4. Variation coefficient for clusters

#	Accounts payable	Accounts receivable	Loans	Equity and investment fund shares
1	6.09	7.39	19.97	7.68
2	17.45	6.65	29.76	7.78
3	15.96	7.13	11.39	5.44
4	19.83	4.95	12.81	12.32
5	17.03	11.97	6.73	17.84
6	5.52	6.54	4.89	4.34
7	10.08	11.54	13.13	4.57
8	6.24	4.83	5.05	9.18
9	6.94	4.04	7.79	6.54
10	21.47	17.25	16.53	8.74
11	8.73	7.46	5.62	6.32

According to Table 4 cluster analysis doesn’t solve the problem of heterogeneity, as the value of variation coefficient for each financial instrument in each cluster remain high.

Individual growth rates

As demonstrated by the analysis of missing values in the dynamics of the main measures of accounting statements, it is possible to apply interpolation. This method estimates quarterly values of various financial measures in the form of a fixed weighted linear combination of bounding annual figures.

Let us assume that the value as of the end of the third quarter (y_3) is missing, whereas all other values are known. In this case, the value of the unknown quarterly measure is calculated according to the formula:

$$y_3 = y_2 + \frac{Y_3 - y_2}{2},$$

where y_3 is the value of the measure at the end of the third quarter, Y_3 is the value of the measure as of the end of the year, and y_2 is the value of the measure as of the second quarter. Other cases are considered in a similar way.

The results of interpolation are presented in Chart 5. Changes in the resulting main financial measures resemble the dynamics of measures in accounting statements submitted by stable companies of the OFC subsector shown in Chart 3. The reason for this is that the coverage ratio as of quarterly dates is very low, due to which the result of data restoration is very similar to stable companies' behaviour. However, there is no reason to believe that the behaviour of organisations not submitting statements on a quarterly basis repeats the behaviour of large market participants that submit such reporting.

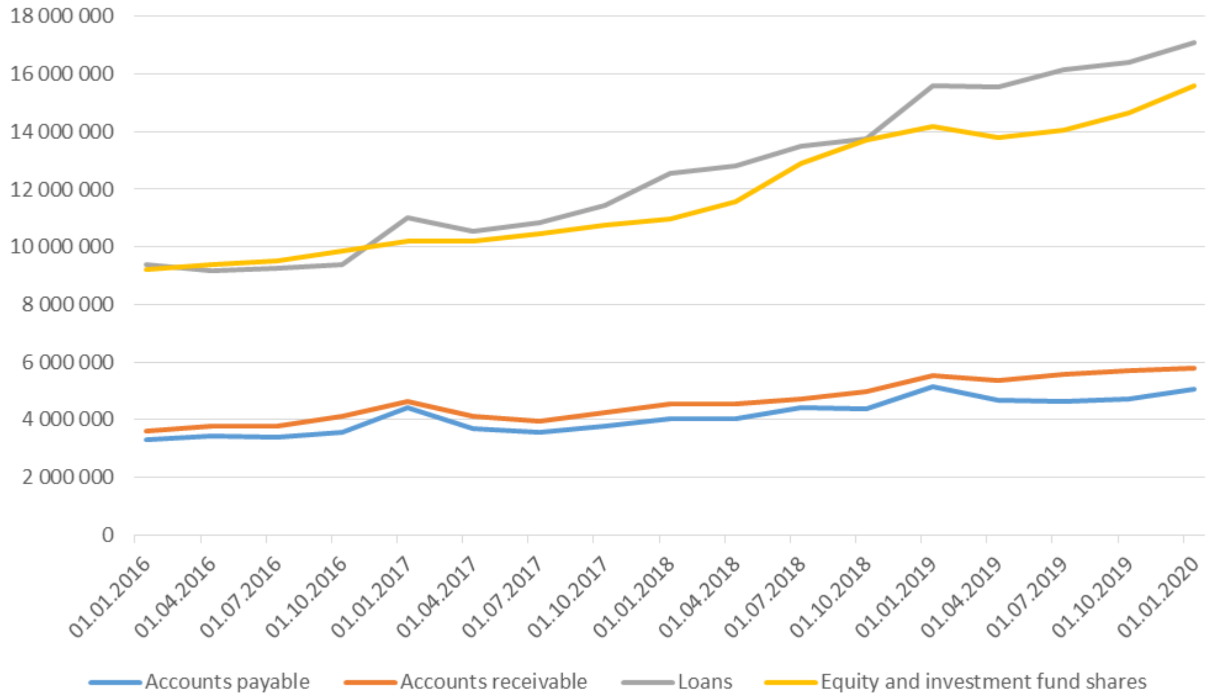


Chart 5. Results of restoring the main financial measures using the individual growth method, mln of rubles

'1/4' and dynamics extension methods (current method)

Currently, the main method to restore missing values on a quarterly basis is the so-called '1/4' method. It assumes that all quarterly dynamics of balance sheet measures in the OFC subsector are proportionately equal to the annual dynamics of the same financial measures. In particular, the values of unknown quarterly measures are calculated according to the formulas: $y_{1,i} = 0{,}25(Y_{2,i} - Y_{1,i})$, $y_{2,i} = 0{,}5(Y_{2,i} - Y_{1,i})$, $y_{3,i} = 0{,}75(Y_{2,i} - Y_{1,i})$, where $y_{1,i}, y_{2,i}, y_{3,i}$ are unknown values of the measure as of the first, second, and third quarters, respectively, for the i organisation and $Y_{1,i}, Y_{2,i}$ are known annual values as of the beginning and the end of the year, respectively, for the i organisation. When the value as of the end of the relevant year is unknown, the latest known value of a given measure over the year is taken as the unknown quarterly value.

By attributing a proportionate change over the year to each quarter, it is possible to uniformly distribute the annual growth or decline of a particular financial measure, thus smoothing the dynamics. Furthermore, when the latest known values are attributed as of quarterly dates where there is no closing annual date, the balances of the additionally calculated measure change only based on known data. We thus avoid significant errors in the dynamics when data as of the next annual date are received. The results of current method are presented in Chart 6.

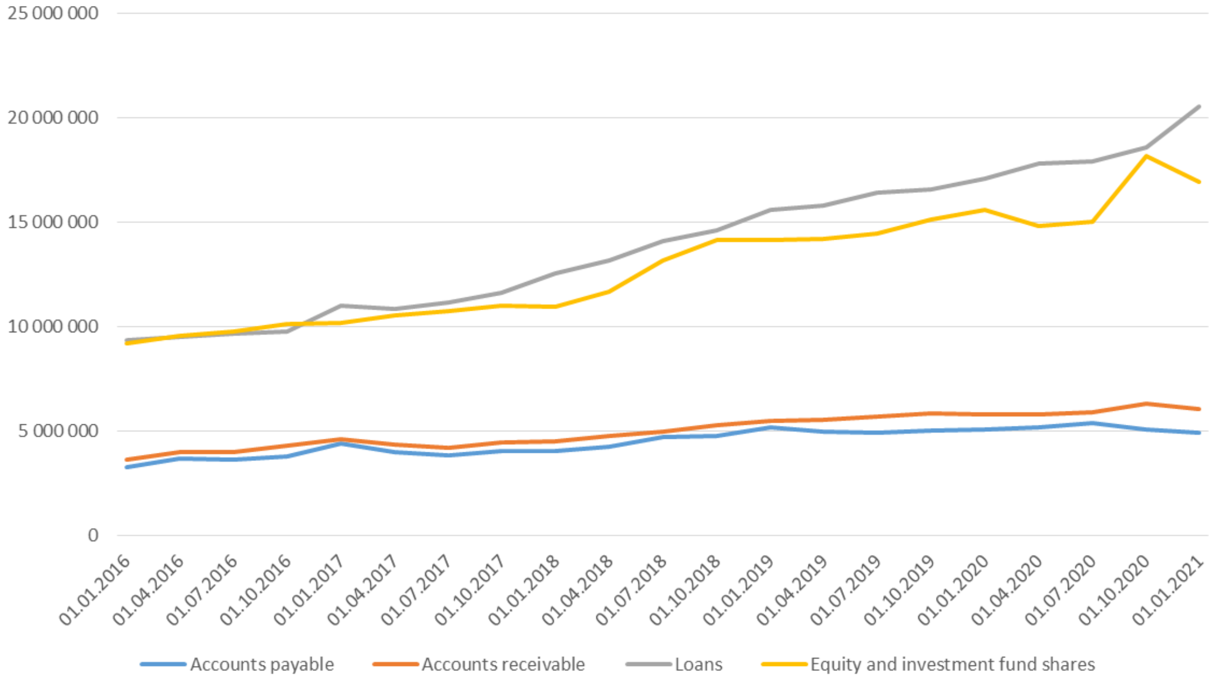


Chart 6. Results of restoring the main financial measures using ‘1/4’ and dynamics extension methods, mln of rubles

Random forest and gradient boosting

As an alternative to classic models we use machine learning algorithms for data recovery such as random forest and gradient boosting. We assume that there is functional relationship between current value of balance indicator and previous values of company’s balance indicators. So the relationship for indicator i in period t (y_t^i) looks like this:

$$y_t^i = f_t^i(y_{t-1}^1, .., y_{t-1}^N, y_{t-2}^1, .., y_1^N)$$

Small training sample and large variance in sizes and structures of balances lead to the fact, that fitting this model doesn’t give good results. Instead we use following relationship:

$$\frac{y_t^i - y_{t-1}^i}{S_{t-1}} = f_t^i\left(\frac{y_{t-1}^1}{S_{t-1}}, .., \frac{y_{t-1}^N}{S_{t-1}}, \frac{y_{t-2}^1}{S_{t-1}}, .., \frac{y_{t-2}^N}{S_{t-1}}, S_{t-1}\right)$$

Where S_t – aggregate balance of company. The dependence of the share of the increase in the balance sheet indicator is estimated depending on the distribution of the shares of indicators in the two previous periods. For each indicator, training sample contains companies, with this indicator filled in in the current period, and all indicators filled in past and before last periods. Thus, part of the data recovered by the algorithm before is not included in the learning process.

Random forest and gradient boosting models are used for estimation f_t^i. To tune model hyperparameters (number of trees, maximum tree depth, etc.), the assumption is used that the dependence of the value of any indicator on the values of indicators in past dates has a similar structure to the dependence of the annual value on past annual values. Thus, to select the hyperparameters for each indicator, tests were made on the annual data which is fully completed.

The results of these methods are presented in Chart 7 and Chart 8 respectively.

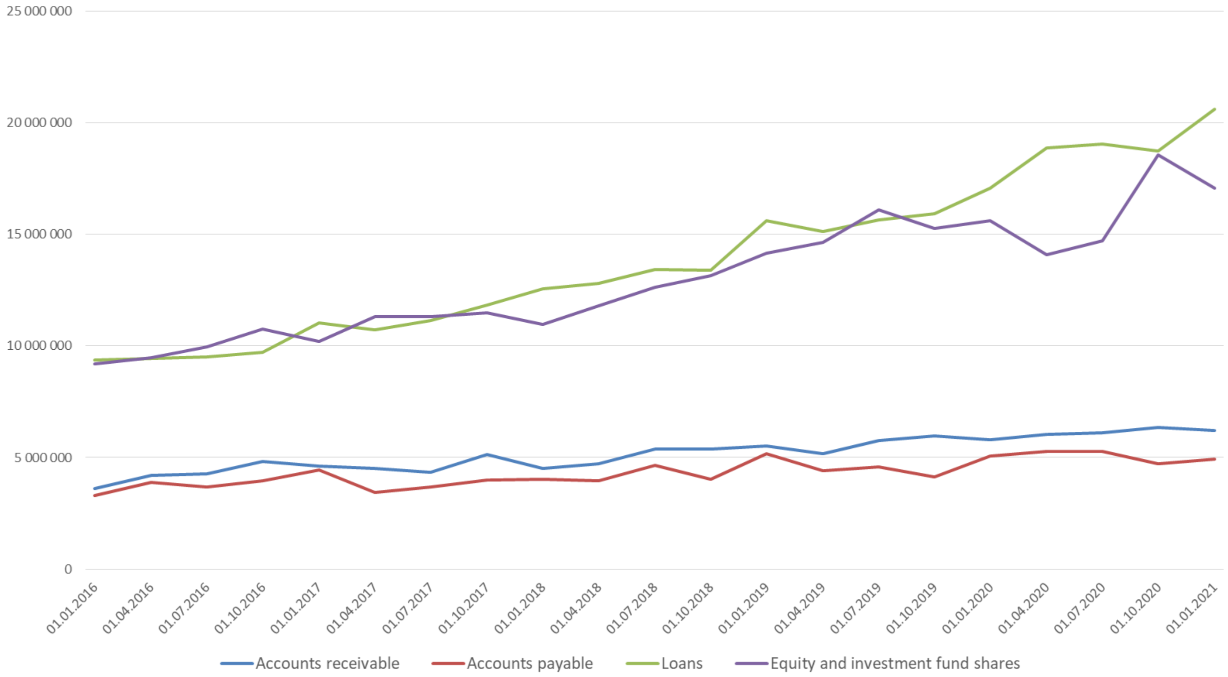


Chart 7. Results of data restoration using random forest, mln of rubles

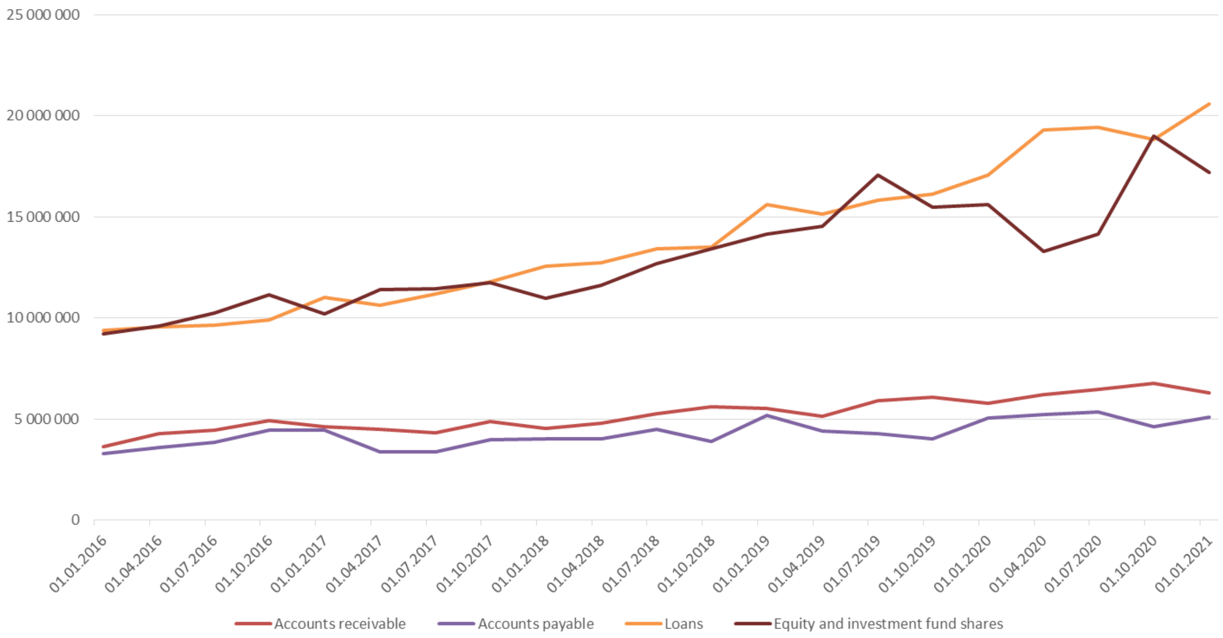


Chart 8. Results of data restoration using Gradient boosting model, mln of rubles

Generative adversarial network

Each measure of financial statements is considered separately. The vector of values for each company is a time series of 17 elements, each of which is a quarterly value of the measure.

The generative network should be trained using the maximum possible number of real companies will all data filled in. However, the array of source data is very sparse as a large portion of companies' quarterly information is unavailable. To increase the number of companies in the training sample, it is usual to also include companies with 'almost all' data filled in. In this case, these are companies having three missing values at most. These missing values are filled in based on mean (quarterly) growth rates calculated for all companies (except the outliers where a specific company's quarterly growth rates are beyond the range [0.7, 1.5]). This generates a training sample that is significantly larger than the original sample of all filled-in data, which uses 'almost all' filled-in data with the minimal impact of growth rates.

After the GAN is trained, the noise generated by it as inputs transforms into the vector of financial measures of a simulated company. To achieve a high accuracy, the input noise is changed iteratively using the Adam optimiser so as to make inputs closer to the filled-in values of the vector. The generative network thus simulates a company which is most similar to the one having data gaps. The generative network outputs are used to close data gaps.

Missing data for all companies are restored sequentially for each financial measure. The results of this method are presented in Chart 9.

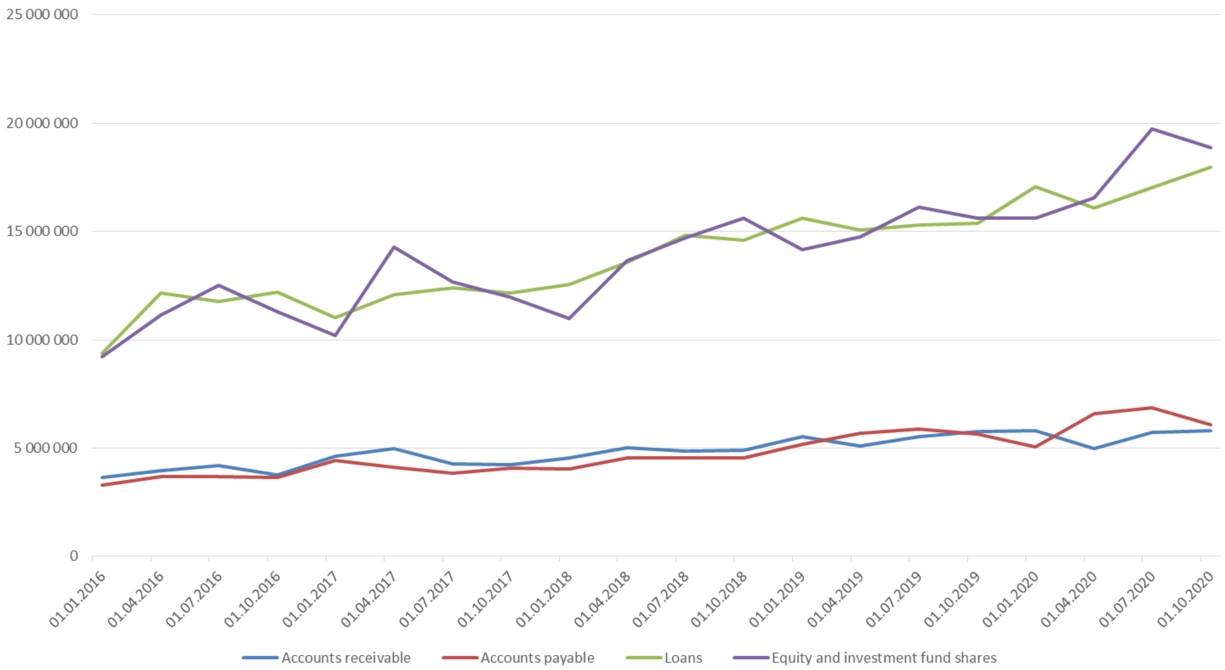


Chart 9. Results of data restoration by the GAN, mln of rubles

Comparative analysis of the methods used to restore quarterly values of the measures of accounting statements and conclusions

At the previous stage of the research, we presented the main results of the data restoration methods employed. Below are the results of the comparison between the currently applied '1/4' and the dynamics extension method, as well as machine learning methods, namely the Random Forest algorithm for regression, gradient boosting model and the generative adversarial network (GAN).

To compare the results of the methods used to close data gaps for each of the reviewed measures as of 1 January 2017 and 1 January 2020, we formed a sample of organisations from the ensemble of other financial corporations carrying out unlicensed activities and made an additional calculation for this sample. The sample for each financial measure included organisations that had not submitted their statements according to federal statistical forms No. P-3 and No. P-6 as of the dates under review.

To compare the methods used, we calculated the ratio of the deviation of the restored values from the actual ones to the total value of each financial measure. The results of closing data gaps in the main financial measures are given in Table 5.

Table 5. Comparison of the results of different methods used to close data gaps, as of 1 January 2017 and 1 January 2020, %

		01.01.2017	01.01.2018	01.01.2019	01.01.2020
Current method	Accounts receivable	-7.13	6.80	-8.16	-2.90
	Accounts payable	-11.01	8.48	-13.13	-9.84
	Loans	-3.44	2.07	-10.50	-4.43
	Equity and investment fund shares	-9.93	-10.33	-16.05	-3.88
Random forest	Accounts receivable	-2.73	11.23	1.96	0.08
	Accounts payable	-4.88	13.26	-1.81	5.25
	Loans	-3.95	7.27	-8.14	-0.83
	Equity and investment fund shares	-0.06	-6.45	-13.33	-8.11
Gradient boosting model	Accounts receivable	3.46	10.30	7.77	2.74
	Accounts payable	-6.63	6.65	2.88	23.44
	Loans	-4.23	7.98	-8.11	2.44
	Equity and investment fund shares	0.19	-5.14	-12.13	-8.25
GAN	Accounts receivable	-18.32	-1.18	-5.19	11.03
	Accounts payable	-14.27	-1.51	2.24	3.76
	Loans	12.90	12.25	-8.97	-4.59
	Equity and investment fund shares	8.94	39.72	-25.13	-8.00

Source: the authors' calculations.

As demonstrated by the analysis of the results presented in Table 5, we can't distinguish the best approach to close data gaps. According to the Table 5 random forest was the best algorithm in many cases. Contrastingly, generative adversarial network method showed the highest ratio of the deviation. This is so, because, first of all, there are not enough points to fit on the GAN model. Secondly, the learning set is more likely to have missed many maxima and minima of the loss function (that finds the weights of neural network), i.e. points where the gradient is zero. This is occured when the domain of inputs is not dense, i.e. the input values are not closely clustered (typical case for sparse data).

The main disadvantage of machine learning methods is their bad interpretive properties, whereas the current method identify organisations accounting for a rise or a decline in a particular measure. Unfortunately, it is hard to tell which method will be preferable for a particular indicator at a certain point in time. Furthermore, there is a significant difference in the results of the calculations of each measure: on average, the best results in data restoration can be achieved in Accounts receivable, whereas the worst results – in Equity. This may be associated with both the low coverage as of quarterly dates, as compared to annual dates, and high volatility of this indicator.

Literature

1. Board of Governors of the Federal Reserve System, 2000.

2. Leo Breiman. Random Forests, 2001

3. Chow, G. and Lin, A. Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series. The Review of Economics and Statistics 53, 1971, p. 372–375.

4. Fernández, R. A Methodological Note on the Estimation of Time Series. The Review of Economics and Statistics 63, 1981, p. 471–478.

5. Financial Accounts: History, Methods, the Case of Italy and International Comparisons, 2008, p. 118–122.

6. Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001, p. 1189–1232

7. Litterman, R. A Random Walk, Markov Model for the Distribution of Time Series. Journal of Business and Economic Statistics 1, 1983, p. 169–173.

8. Martin Arjovsky, Soumith Chintal, Léon Bottou. Wasserstein GAN, 2017.

9. System of National Accounts 2008 (European Commission, United Nations, Organisation for Economic Cooperation and Development, International Monetary Fund, World Bank).

Bank of Russia

RESTORATION OF OMISSIONS IN THE QUARTERLY INDICATORS OF FINANCIAL STATEMENTS FOR THE OTHER FINANCIAL INSTITUTIONS IN THE BANK OF RUSSIA

PIRUZA ALIEVA AND ANNA BORISENKO,
STATISTICS DEPARTMENT
PETR MILYUTIN AND DENIS KOSHELEV,
RESEARCH & FORECASTING DEPARTMENT

Workshop on Data Science in Central Banking
19-22 October 2021

Agenda

1. Overview of the Other Financial Intermediaries in the Russian Federation
2. The results of cluster analysis for the Other Financial Intermediaries
3. Current approach of the restoration of omissions in the quarterly indicators of financial statements for the Other Financial Intermediaries
4. Results of methods
    1. Individual growth rates method
    2. Random forest
    3. Gradient boosting model
    4. Generative adversarial networks
5. Comparison of methods
6. Conclusions

Overview of the Other Financial Intermediaries in the Russian Federation

Table 1. Unsupervised OFIs’ statistics

	Portion of unsupervised OFIs’ total balance in all OFIs’ total balance	Ratio of the number of unsupervised OFIs number to the overall number of OFIs
01.01.2015	97,34%	72,75%
01.01.2016	97,48%	76,97%
01.01.2017	96,07%	77,31%
01.01.2018	95,29%	83,45%
01.01.2019	99,28%	84,47%
01.01.2020	99,54%	85,52%

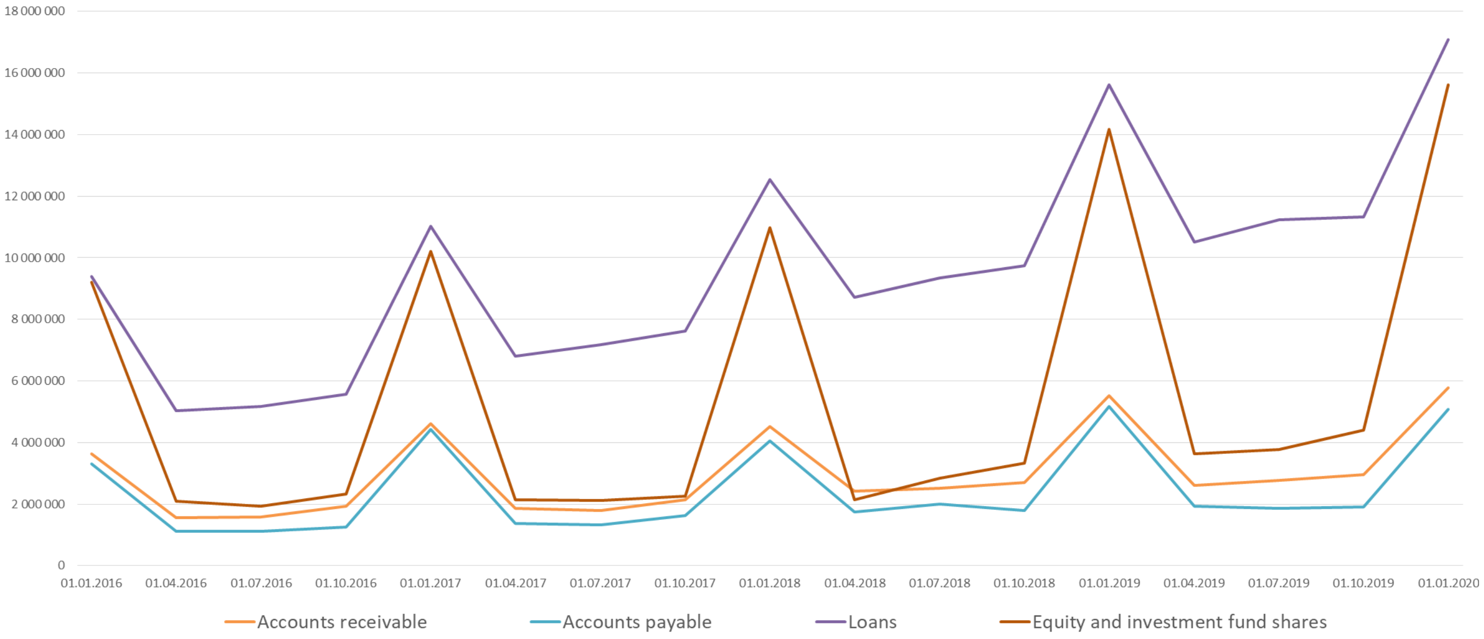


Figure 1. Dynamics of the main financial measures of accounting statements of unsupervised OFIs’, 2016 – 2020, mln of rubles

Cluster analysis

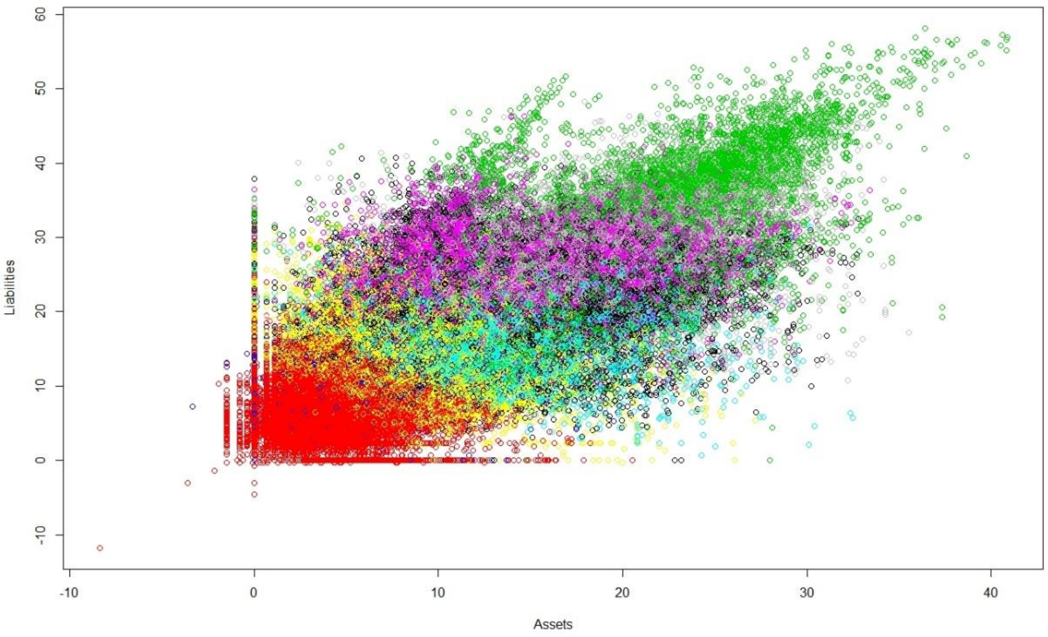


Figure 2. Results of cluster analysis

Table 2. Variation coefficient for clusters

#	Accounts payable	Accounts receivable	Loans	Equity and investment fund shares
1	6,09	7,39	19,97	7,68
2	17,45	6,65	29,76	7,78
3	15,96	7,13	11,39	5,44
4	19,83	4,95	12,81	12,32
5	17,03	11,97	6,73	17,84
6	5,52	6,54	4,89	4,34
7	10,08	11,54	13,13	4,57
8	6,24	4,83	5,05	9,18
9	6,94	4,04	7,79	6,54
10	21,47	17,25	16,53	8,74
11	8,73	7,46	5,62	6,32

Current approach of the restoration of omissions in the quarterly indicators

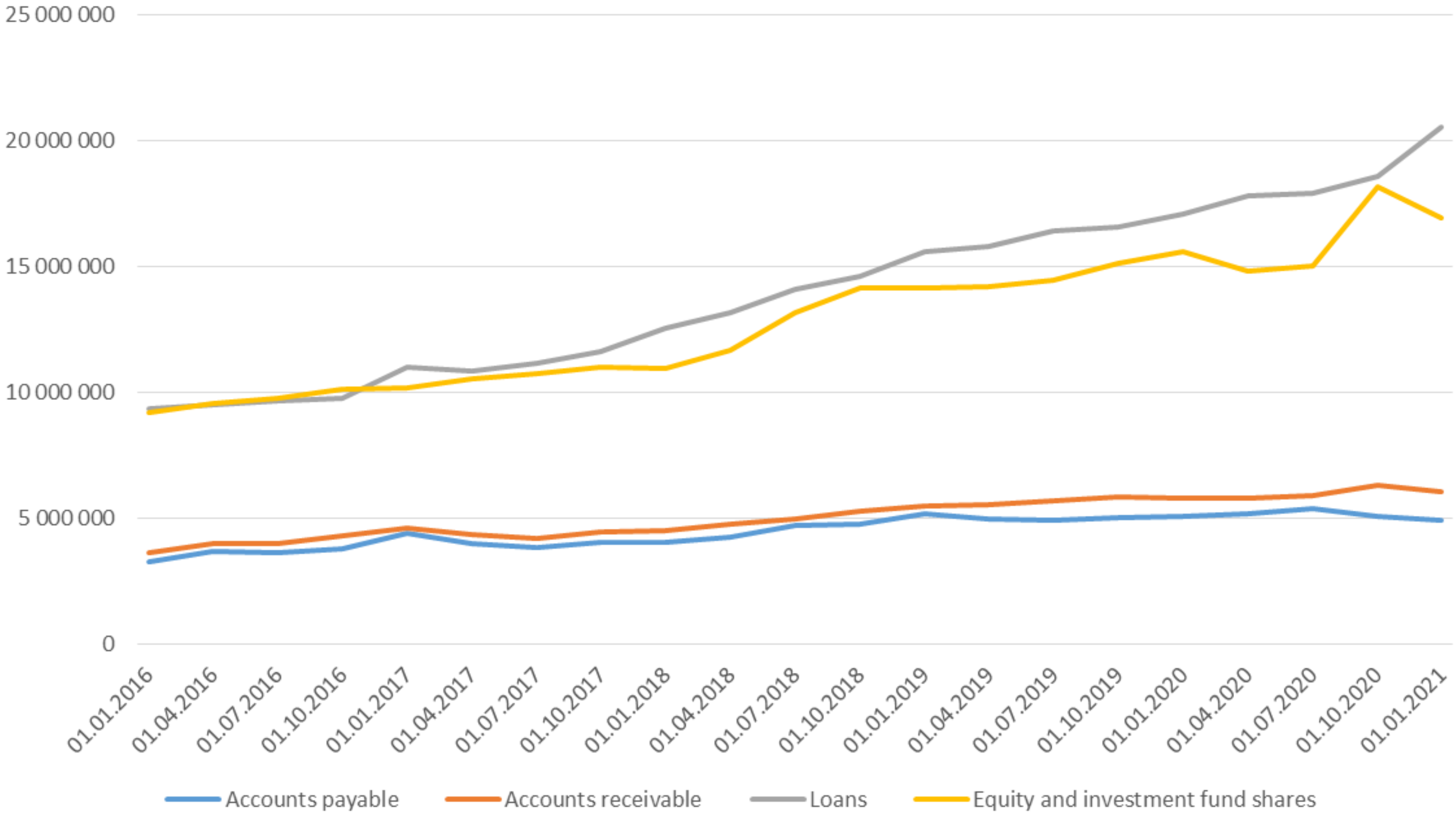


Figure 3. Results of the restoration of omissions in the quarterly indicators (mln of rubles)

Individual growth rate method

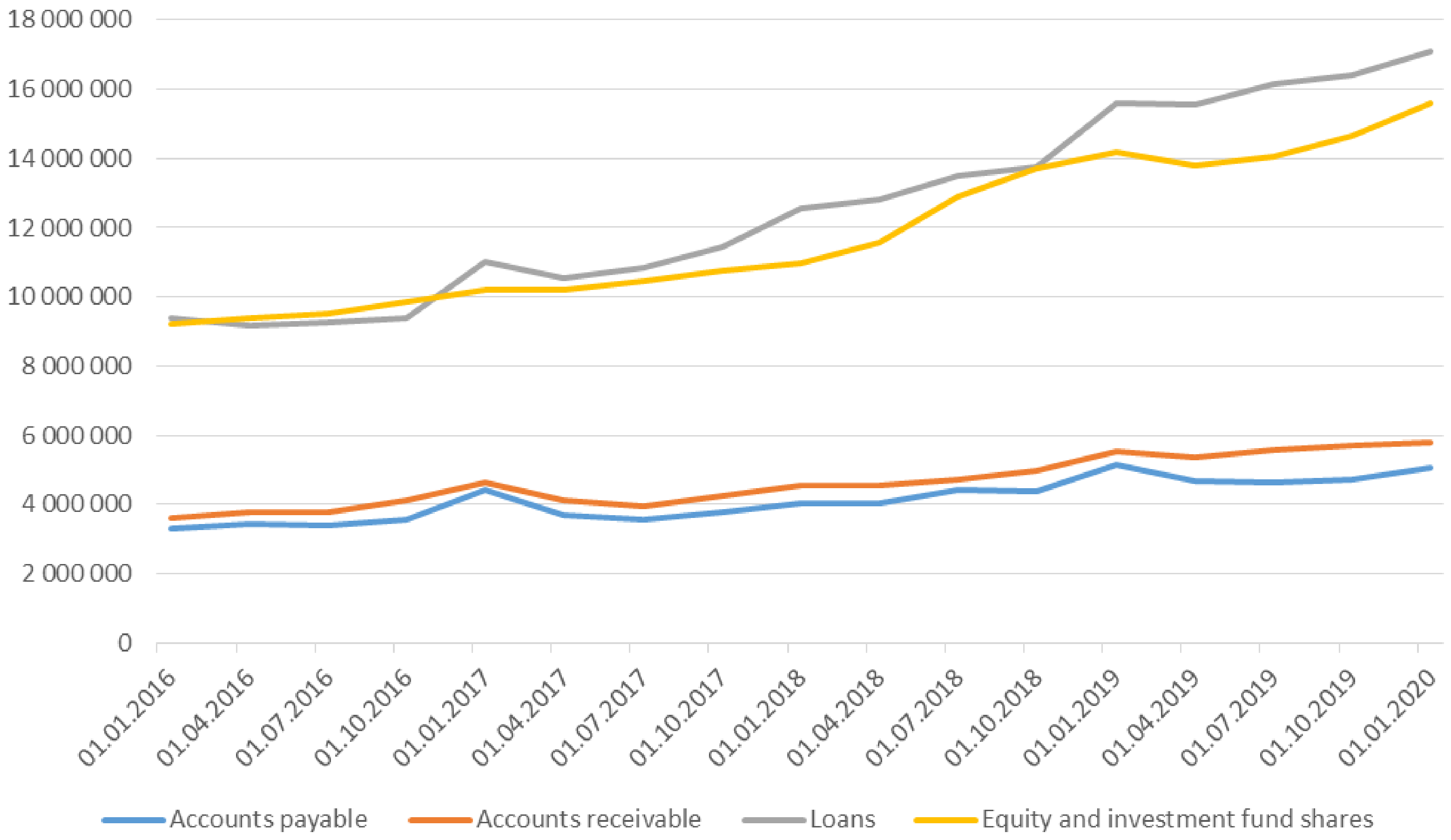


Figure 4. Results of the restoration of omissions in the quarterly indicators (mln of rubles)

Stable OFIs’ main financial measures

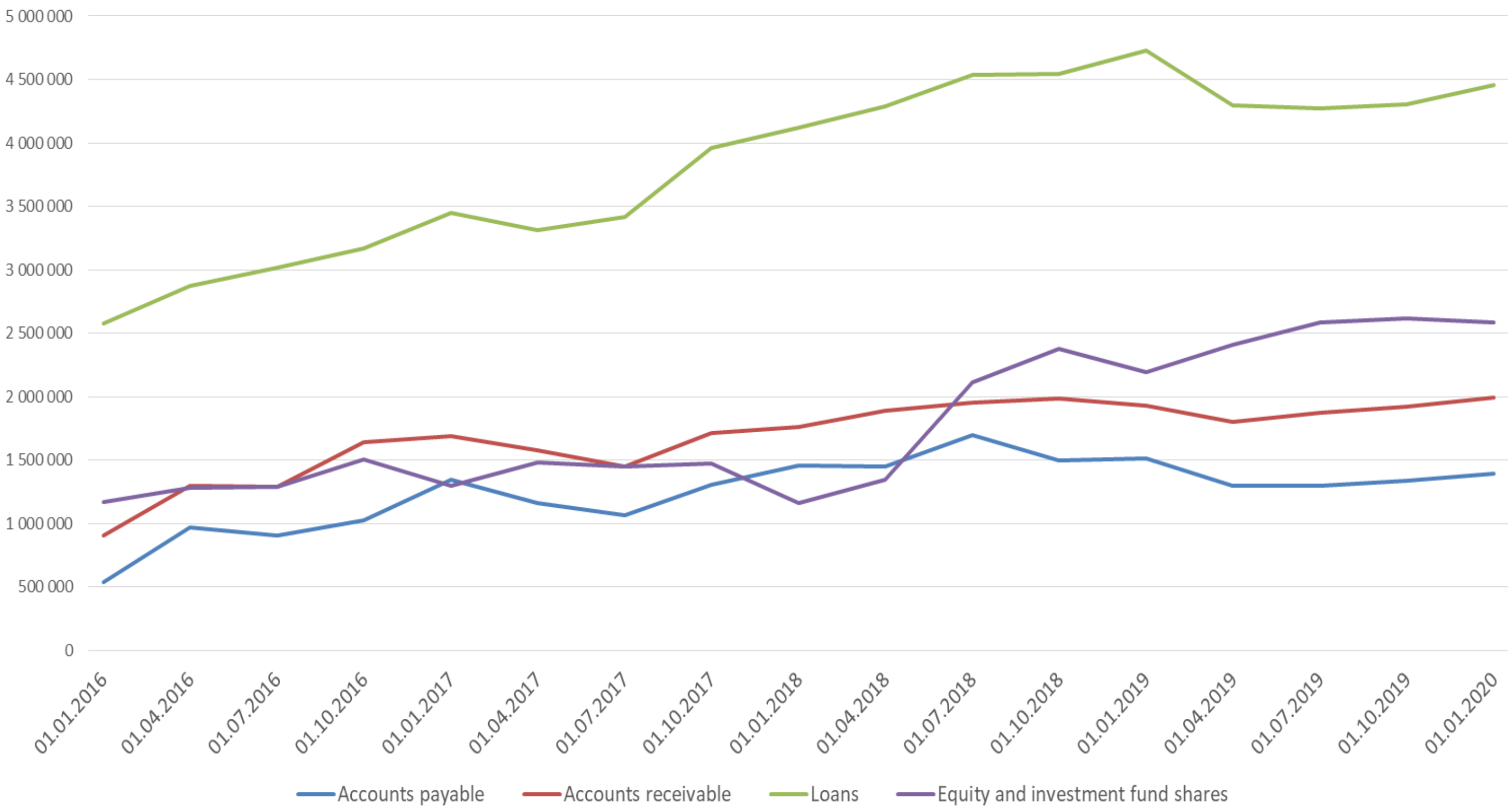


Figure 5. Changes stable OFCs’ main financial measures, 01.01.2016–01.01.2020, mln of rubles

Random forest

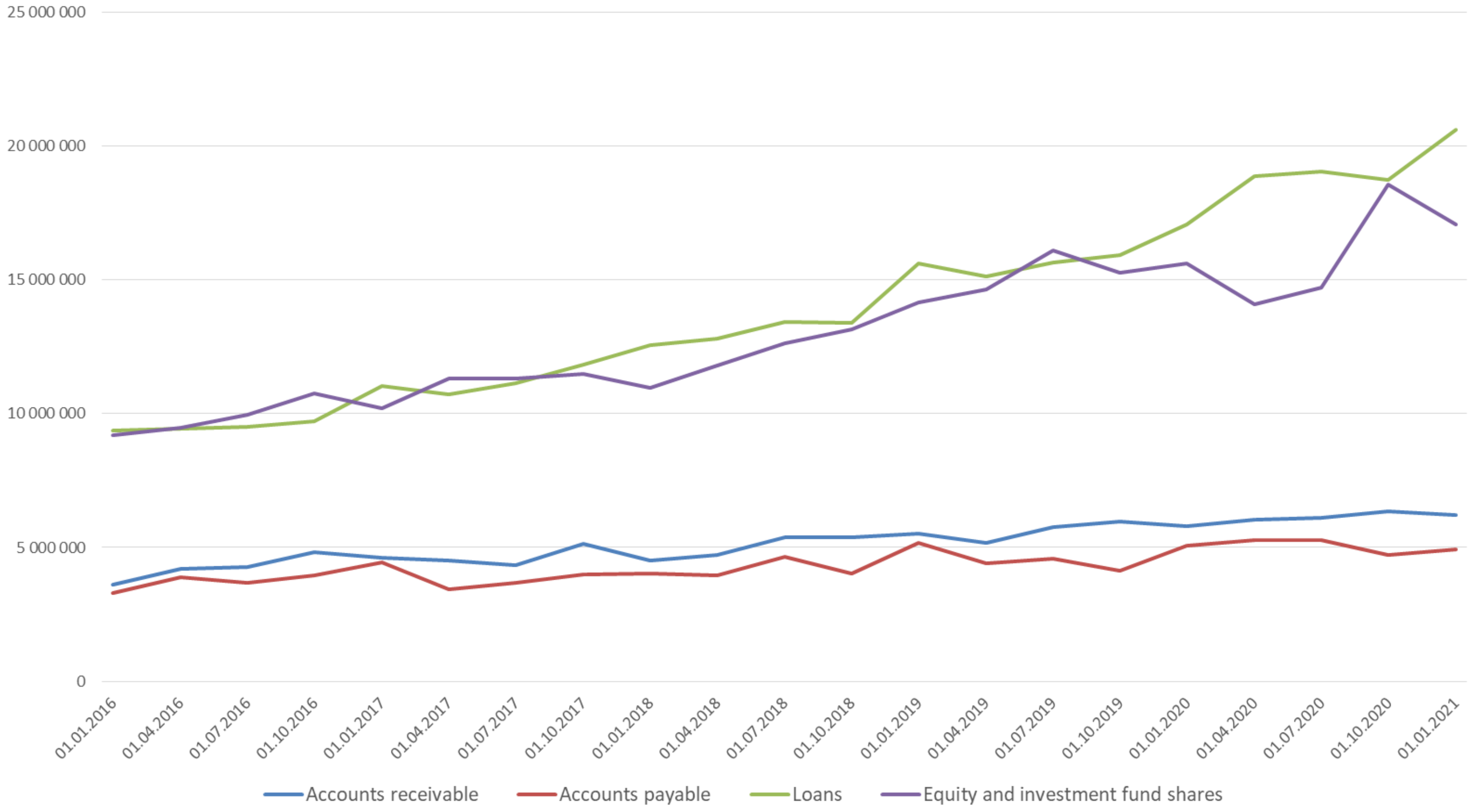


Figure 6. Results of the restoration of omissions in the quarterly indicators (mln of rubles)

Gradient boosting model

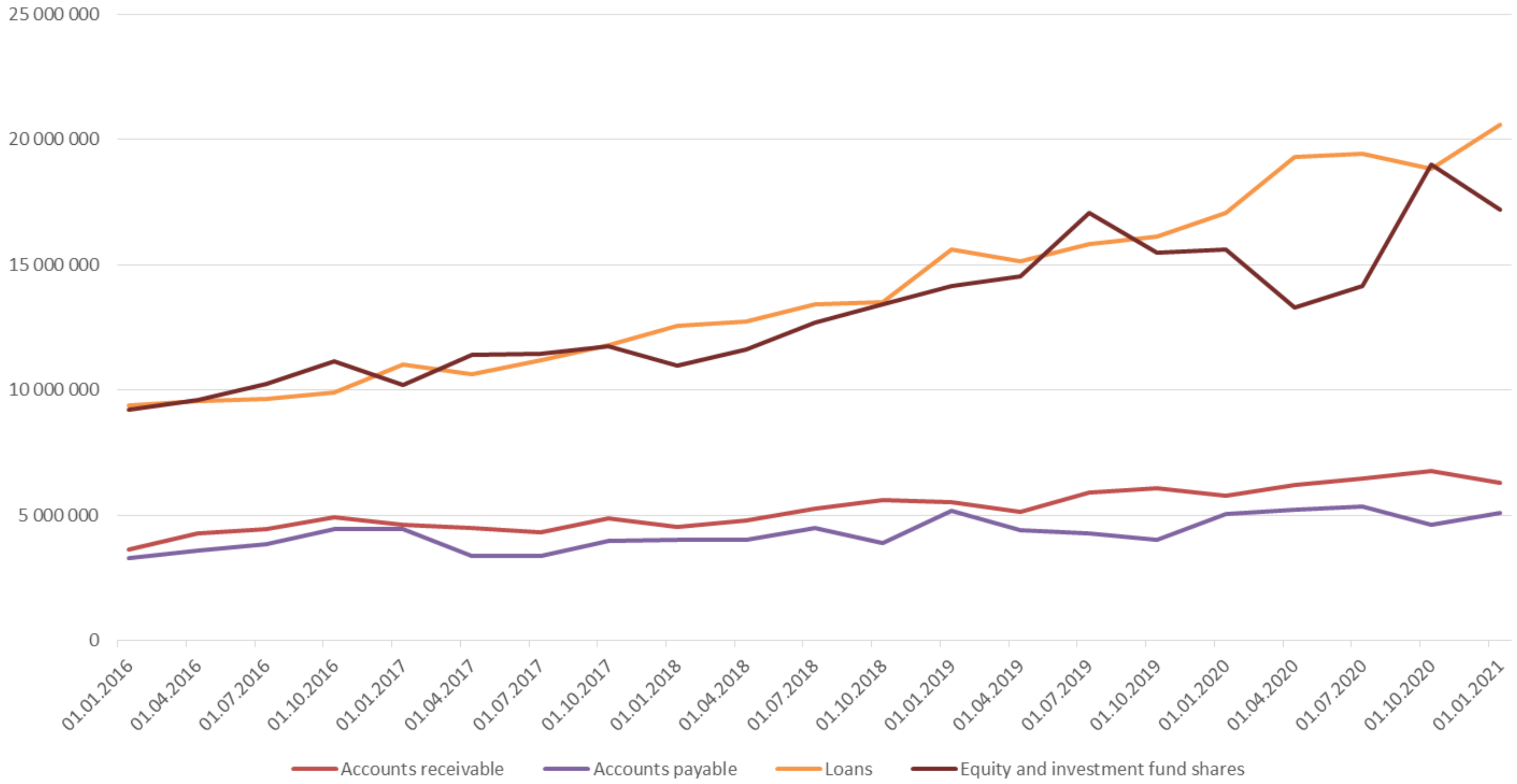


Figure 7. Results of the restoration of omissions in the quarterly indicators (mln of rubles)

Generative adversarial networks (GAN)

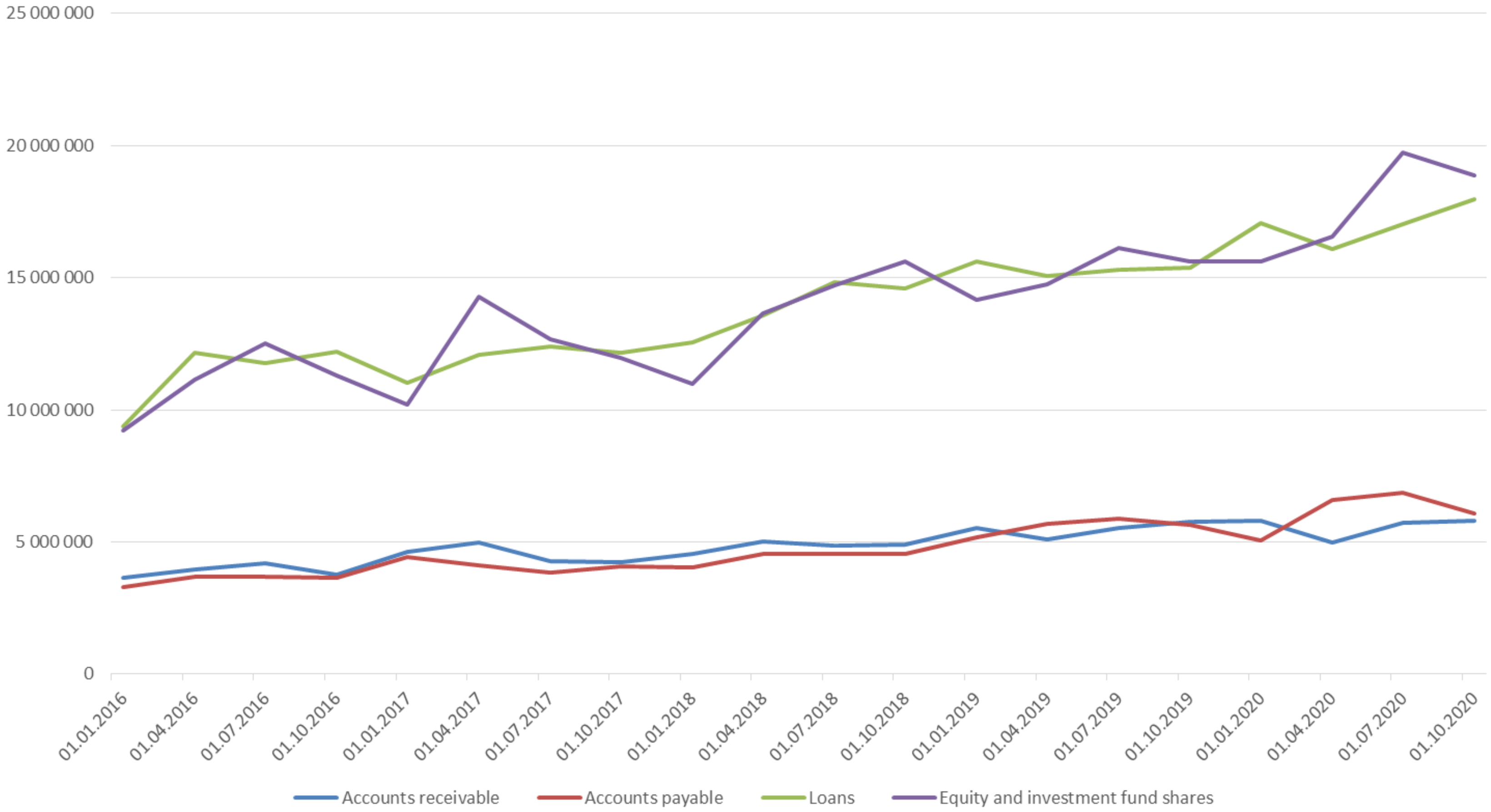


Figure 8. Results of the restoration of omissions in the quarterly indicators (mln of rubles)

Comparison of methods and conclusions

Table 3. Results of comparison of methods (deviation of estimated value from real number)

		01.01.2017	01.01.2018	01.01.2019	01.01.2020
Current method	Accounts receivable	-7,13%	6,80%	-8,16%	-2,90%
	Accounts payable	-11,01%	8,48%	-13,13%	-9,84%
	Loans	**-3,44%**	**2,07%**	-10,50%	-4,43%
	Equity and investment fund shares	-9,93%	-10,33%	-16,05%	**-3,88%**
Random forest	Accounts receivable	**-2,73%**	11,23%	**1,96%**	**0,08%**
	Accounts payable	**-4,88%**	13,26%	**-1,81%**	**5,25%**
	Loans	-3,95%	7,27%	-8,14%	**-0,83%**
	Equity and investment fund shares	**-0,06%**	-6,45%	-13,33%	-8,11%
Gradient boosting model	Accounts receivable	3,46%	10,30%	7,77%	2,74%
	Accounts payable	-6,63%	6,65%	2,88%	23,44%
	Loans	-4,23%	7,98%	**-8,11%**	2,44%
	Equity and investment fund shares	0,19%	**-5,14%**	**-12,13%**	-8,25%
Generative adversarial networks	Accounts receivable	-18,32%	**-1,18%**	-5,19%	11,03%
	Accounts payable	-14,27%	**-1,51%**	2,24%	3,76%
	Loans	12,90%	12,25%	-8,97%	-4,59%
	Equity and investment fund shares	8,94%	39,72%	-25,13%	-8,00%

Bank of Russia

THANK YOU FOR YOUR ATTENTION

RESTORATION OF OMISSIONS IN THE QUARTERLY INDICATORS OF FINANCIAL STATEMENTS FOR THE OTHER FINANCIAL INSTITUTIONS IN THE BANK OF RUSSIA

PIRUZA ALIEVA AND ANNA BORISENKO, STATISTICS DEPARTMENT
PETR MILYUTIN AND DENIS KOSHELEV, RESEARCH & FORECASTING DEPARTMENT

2021

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Supervised machine learning for estimating the institutional sectors of legal entities on a large scale[1]

Francesca Benevolo, Thomas Gottron, Ilaria Febbo and Nicolò Pegoraro,
European Central Bank

[1] This presentation 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 central banks and other institutions represented at the event.

Supervised machine learning for estimating the institutional sectors of legal entities on a large scale

Francesca Benevolo, Thomas Gottron, Ilaria Febbo, Nicolò Pegoraro[1]

Abstract

The Register of Institutions and Affiliates Data (RIAD) is the European System of Central Banks' (ESCB) shared register providing master data for more than 10 million legal entities. One of the key RIAD features is the provision of institutional sector classification according to the ESA 2010 methodology. The distinction between different types of financial institutions, non-financial corporations and private versus public sector is of high importance for several ESCB tasks. In fact, information on the institutional sectors is mandatory for all entities in RIAD and is maintained on an on-going basis by experts at National Central Banks and at the European Central Bank. Though, the process of classifying entities by institutional sector is currently manual and time consuming – as necessary to ensure the requested accuracy – and therefore hardly applicable on a large scale, e.g. when a high number of entities need to be imported from external registers.

To address this use case, we present an automated, high-quality approach for the bulk classification of entities according to their (ESA) institutional sector. The estimates produced serve as good preliminary information supporting the expert assessment and the final entity classification. The approach is based on supervised machine learning with a two-level-approach. It makes use of publicly available information on legal entities, e.g. their name, residence, registration authority or legal form. We use a hierarchical setup of ensemble methods tailored to suit the business needs for the hierarchy of the ESA institutional sectors. Furthermore, we use deep neural networks to create semantic embeddings for company names which have shown to improve classification performance. The approach has been tested and evaluated on a dataset of approximately 550,000 known entities and it was applied to estimate the institutional sector for nearly 1 million entities not yet in RIAD.

Keywords: ESA 2010, institutional sector, Legal Entity Identifier, RIAD, machine learning, estimation

JEL classification codes: C130

[1] Disclaimer: 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.

Contents

Introduction

The European System of National and Regional Accounts (ESA 2010) is an internationally compatible EU accounting framework for a systematic and detailed description of an economy (Eurostat, 2013). At the European Central Bank (ECB), the ESA sector classification is used in several business processes and databases.

The Register of Institutions and Affiliates Data (RIAD) is the ESCB's shared register providing master data for more than 10 million legal entities such as banks, private enterprises, and public institutions. One of the key RIAD features is the provision of institutional sector classification according to the ESA 2010. The information about ESA sectors in RIAD is mandatory for all entities and it is well maintained by experts at National Central Banks (NCBs) and the ECB.

ESA sectors are used to distinguish between types of legal entities, e.g. non-financial, deposit-taking corporations, insurance corporations or governmental bodies. The ESA sector classification in RIAD comprises a code with prefix "S" followed by a number with maximum three digits (cf. Table 1). The nine ESA sectors falling under the financial sector start with "S12". The other seven ESA sectors not belonging to the financial sector start with "S11", "S13", "S14" and "S15".

The process of classifying RIAD entities (especially those residing outside EU) by institutional sector is currently performed manually. The process is time consuming because it needs to ensure a high level of accuracy. This approach is hardly applicable on a large scale, e.g. when a high number of entities need to be imported from external registers such as the Global Legal Entity Identifier Foundation (GLEIF). For these use cases, automated approaches are necessary for providing a preliminary ESA sector estimate. The motivation to work with ESA sector preliminary estimates is to streamline experts' work. In this way, ECB and NCBs experts can prioritise their work and focus on entities of higher relevance for the central banking tasks, i.e. financial entities.

In this paper we present an automated, machine learning based approach to estimate the ESA sector of GLEIF entities based on publicly available reference data (e.g. name, address, legal form, registration authority).

The approach leverages on the overlap between GLEIF and RIAD populations and uses this overlap as labelled training data. The labelled data provides the basis for training a two-level supervised machine learning model based on a Random Forest classifier.

We make the following contributions in this paper:

- We investigate the potential of using supervised machine learning for estimating the legal entity ESA sectors solely based on entity reference data. To the best of our knowledge there is little prior work dealing with ESA sector estimation and no prior work performing this task only via reference data.
- We investigate methods for feature engineering of reference data (e.g. feature selection, semantic embeddings, one-hot encoding).
- We systematically test and evaluate different supervised machine learning methods and identify the best solution for the task.
- We demonstrate that the final solution is of high quality and can safely be integrated in the RIAD production environment.

Table 1: ESA sector classification

ESA sector	Description
S11	Non-financial corporations
S121	Central banks
S122	Deposit-taking corporations except the central bank
S123	Money Market Funds (MMFs)
S124	Non-MMF investment funds
S125	Financial corporations other than MFIs, non-MMF investment funds, financial auxiliaries, captive financial institutions and money lenders, insurance corporations and pension funds
S126	Financial auxiliaries
S127	Captive financial institutions and money lenders
S128	Insurance corporations
S129	Pension funds
S1311	Central government (excluding social security funds)
S1312	State government (excluding social security funds)
S1313	Local government (excluding social security funds)
S1314	Social security funds
S14	Households
S15	Non-profit institutions serving households

Related work

There is a wide range of research addressing the estimation of economic activity codes or institutional sectors of business units. A recent investigation by ONS (Noyvirt, 2021) looking into machine learning approaches to solve this task concluded that achieving a high accuracy remains a challenge. Approaches for estimating the economic activity or sector of an entity mainly differ in terms of input data, methods and target codification schemes.

In the context of classifying counterparties in the EMIR dataset (Lenoci & Letizia, 2021) external sources were used to provide context. This context information helped in identifying the type of activity of an entity, e.g. because its information was obtained from the ECB's list of monetary financial institutions. The overall solution then involved a knowledge-based classification system.

Many approaches leverage the availability of national codifications for economic types of activity for assigning NACE codes (Eurostat, 2008) to entities. Different machine learning techniques are used in settings where no one-to-one translation between different codification systems is available. The techniques range from the use of multi-level classification systems (Giudice, Massaro, & Vannini, 2020), matching pre-processed textual descriptions (Colasanti, Macchia, & Vicari, 2009) tokenising web texts and generating descriptive features (Kühnemann, van Delden, & Windmeijer, 2020) or supervised solutions like Naïve Bayes, Random Forest, Support Vector Machines, k-Nearest Neighbours or voting ensemble methods (Roelands, van Delden, & Windmeijer, 2018).

A general survey of how to model a probabilistic approach for capturing overlaps of text tokens for coding the occupation sector of survey respondents is discussed in (Gweon, Schonlau, Kaczmirek, Blohm, & Steiner, 2017).

The paper at hand presents an approach that is new in respect to the work available in the literature. To the best of our knowledge there is little prior work addressing the estimation of ESA sector classification. Our work offers a new method of estimating the ESA sector classification, based on machine learning techniques such as text analytics, neural networks, and random forests.

Methodology

A thorough analysis of the classification task and users' needs led to specific methodological choices, i.e. the use of a two-level supervised machine learning model and semantic embeddings.

The reason for selecting a two-level supervised machine learning model derives from the primary need to distinguish financial versus non-financial entities. The first step identifies financial entities, i.e. S12 (independently from the ESA sector detail). The second step estimates the full three-digit ESA sector code (S122, S123, etc.).

The aim of exploring semantic embeddings is to make entity legal name information more manageable and valuable. The entity legal name constitutes the most valuable resource for identifying an entity's nature as well as the largest challenge in processing it. To verify the statistical relevance of the legal name we analysed the words frequency (to find the most explicative for ESA sectors) and generated semantic embeddings using neural networks (with the scope of retrieving hidden meaning from the legal names).

In this section we provide further insights into these two choices.

A supervised learning approach

We modelled the task as a supervised learning approach because we could leverage on the existing overlap between GLEIF and RIAD, i.e. 548,464 legal entities belong to both databases. The common entities were used to align GLEIF features with the target RIAD variables. This aligned data was needed to train and test our models.

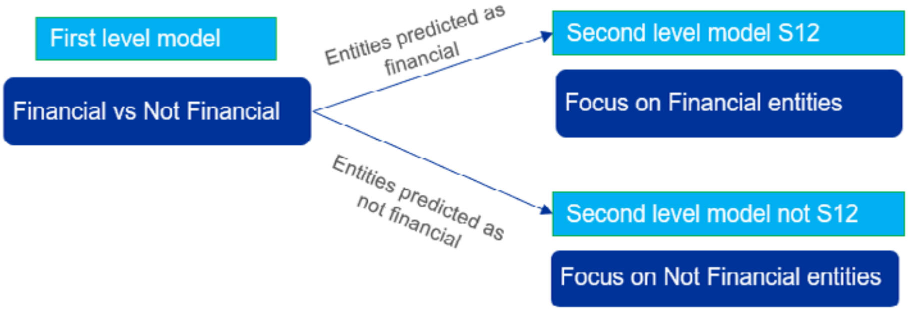
The main steps undertaken to build the model are illustrated in Figure 1 and can be summarised as follows:

1. **Building a first level model to classify observations into financial and non-financial entities.** This task is a binary classification problem aiming to predict if an entity belongs to the financial class. Accordingly, the target variable was re-shaped into a binary variable with a value of 1 if the ESA sector started with "S12" (financial) and a value of 0 otherwise. The decision of this high-level distinction was driven by the business need to primarily distinguish between financial and non-financial entities. The use case of prioritising a further manual assessment implied a conservative approach which is rather biased towards assigning uncertain entities to the financial class. We addressed this requirement by using weighted classes in the classification task. The weights reflected priorities in the

outcome and the corresponding preference for different types of errors in the classification.

2. **Building the two second level models which subsequently predict the detailed ESA sector class.** The second level consists of two sub-models. The first sub-model aims to detail the financial entity class (i.e. distinguish among credit institutions, money market funds, pension funds, central banks, etc.). The second model aims to detail the different types of non-financial entities (e.g. non-financial corporations, governments, households). The financial and non-financial domains are quite heterogeneous. The advantage of having two separate sub-models is to better fine-tune our algorithms and deal with such heterogeneity.

Figure 1: Process design


First level model
Financial vs Not Financial
Entities predicted as financial
Entities predicted as not financial
Second level model S12
Focus on Financial entities
Second level model not S12
Focus on Not Financial entities

Legal name analysis

In this paragraph we discuss the impact of the entity legal names on our work, i.e. the analysis of individual words used in legal names as well as the benefit of using semantic embeddings.

We analysed the words frequency in the legal names. The most frequent words were slightly different when considering only entities belonging to the S12 classes (financial). Our assumption was that some words in the legal names were more likely to be used by financial entities since the company name may include indicative hints on the main business of an entity. Legal names containing terms like “bank”, “fund” or “insurance” provide good evidence for a financial entity. Instead, if the name contains terms like “manufacturing”, “travel” or “transport”, the entity probably belongs to the non-financial sector. We decided to include the most frequent words as features for our models, using one-hot encoding. We also included some additional words that were defined as very relevant by the RIAD experts. The business expertise on certain words being indicative for S12 or not S12 was a very good motivation to consider individual words as features.

In the semantic embeddings work, the challenge lies in the text complexity of company names. Information in text is encoded on the semantic level and involves a deeper understanding of concepts and their relations expressed by words. While to a certain extend this knowledge can be encoded in a rule-based static knowledge base it is tedious to set up and maintain such a knowledge base. As an example, the multilingual context was challenging in the sense that we had to deal with words such “bank”, “banca” or “banque” that refer to the same concept. Semantic embeddings are a technique to represent texts as high dimensional, numeric vectors, where similar vectors represent similar concepts or even meaning. It is a well established approach for multiple use cases of text processing.

We computed semantic embeddings using a deep recurrent neural network. For our use case we made good experiences using a relatively simple topology of a bi-directional recurrent neural network composed of LSTM neurons neurons (Hochreiter & Schmidhuber, 1997), following by a multi-layer dense network of ReLU nodes. The network was trained on a large number of legal names taken from GLEIF. The values observed in the last layer of the dense network served as the semantically rich representation of the legal names. These embeddings were included as features in the actual ESA sector classification task.

Data

Our analysis leverages two data sources: GLEIF and RIAD. GLEIF contains a rich feature set for legal entities covering basic reference data. Among others, this feature set contains information on the name, address, legal identifiers, registration authorities or relations to parent companies and subsidiaries. Some GLEIF information items, like the address information, are structured further, e.g. to distinguish between legal address, headquarter address, other addresses, or transliterated addresses. All these features are potential raw input variables for an ESA sector prediction model.

When we started our exercise, it was possible to establish a link between records in RIAD and GLEIF via LEI for approximately 550,000 entities. The first step was to partition the dataset into training, testing and validation data. We decided to take 20% out as blind holdout data before creating the models. The blind holdout data provides the basis for a final evaluation of the machine learning model's performance after it has been trained and tested. It was not used to make decisions about which model to use or for improving or tuning algorithms. The remaining 80% of the labelled data was used to build, train, and test the models. We used a simple random split for partitioning the data.

We started to analyse the data in an explorative phase to get a better understanding of the task. As extensively discussed, this paper aims to estimate the ESA sector for entities in GLEIF. The distribution of the ESA sector classes is unbalanced in our dataset (cf. Table 2). The financial entities (starting with S12) represent the 21.7% of the total.

The non-financial entities (S11, S13, S14, S15) represent the 78.3% of the overall volume. Among those, the non-financial corporations (S11) play the main role: with 415,426 units, they cover 75% of the entities. We took these percentages as benchmark of our model, at least for the binary problem of predicting if an entity was financial or not. Our goal was to build a model that could estimate the ESA sector significantly better that an approach based on flipping a coin with probability 0.783.

Table 2: ESA sector frequency in the data

ESA sector	Count	Frequency
S11	415,426	75.74%
S124	51,226	9.34%
S127	28,321	5.16%
S125	14,065	2.56%
S126	11,990	2.19%
S15	8,840	1.61%
S122	6,471	1.18%
S128	3,404	0.62%
S129	2,754	0.50%
S1313	2,463	0.45%
S14	1,381	0.25%
S1311	804	0.15%
S1314	481	0.09%
S123	444	0.08%
S1312	311	0.06%
S121	83	0.02%

Experiments

First level model

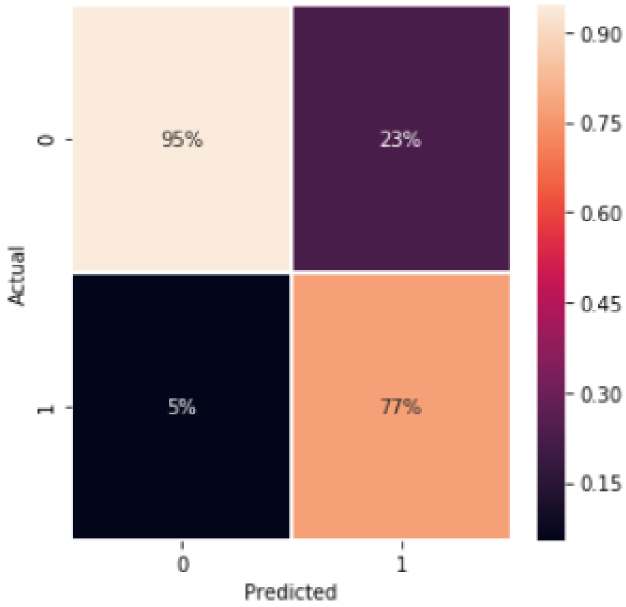
The first level model was a Random Forest Classifier with 100 trees, optimised parameters (from parameter search techniques) and weighted classes. The model was trained and tested on a dataset of 438,771 records with cross validation methods. It was finally validated on our blind holdout dataset of size 109,693. Being a binary classifier, the first level model must deal with two types of errors: false positives and false negatives. The RIAD experts wanted to reduce the error of classifying entities as not financial when they were financial. We translated the business constraint by reducing the false negatives.

Methodologically, we accepted higher false positive, while keeping as low as possible the false negative, i.e. the error of predicting as "Not Financial" the entities that were "Financial" in reality. Therefore, we included a higher weight for the class "Financial". After several experiments, we decided that a triple weight was a good compromise among keeping the percentage of false negatives under 5% and having an acceptable percentage of false positives (<30%).

The confusion matrix in Figure 2 shows the percentage of the real entity type versus the prediction of being financial or not in the validation set. Among the entities predicted as "Not Financial" (Predicted = 0), only 5% (5 480) were "Financial" (Actual=1). The number of false negatives resulted to be quite low, as requested by the RIAD experts. As consequence, we accepted a higher error for the false positives.

In our data, 23% of the entities 79,863) predicted as “Financial” (Predicted = 1) were “Not Financial” (Actual = 0) in reality.

Figure 2: Results on blind hold out data Weighted confusion matrix for first level model



We used the accuracy metric to evaluate the models. The accuracy is the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined. Accuracy is used as a statistical measure indicating how well a binary classification test correctly identifies a condition.

We selected the Random Forest Classifier with triple weights on financial classes, all variables, and semantic embeddings. The selected model was validated on the blind holdout dataset with 109,693 labelled records, which was not used in the test and training phase to guarantee the test integrity. The overall accuracy score of the first level model was 90% on the blind holdout dataset. In other words, in 90% of the cases the first level model could predict correctly if an entity was financial or not. Please note that this value of accuracy must be taken with a grain of salt, as it also reflects the bias of the model to declare entities as financial entities. This tendency of reducing false negative results has a slightly negative impact on the overall accuracy of the model.

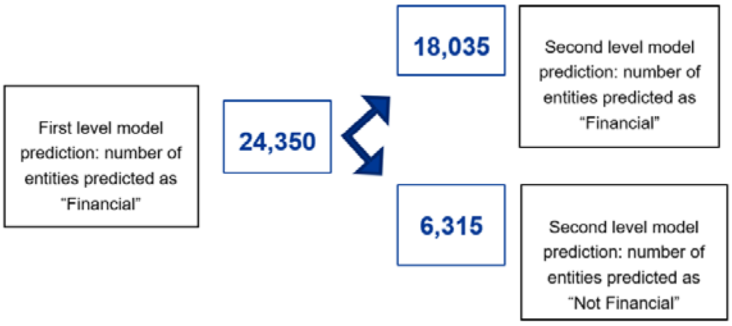
Second level models

The second level model for financial entities aimed to assign the ESA sector to entities that were estimated as “Financial” by the first level model. We trained this second level model on entities that were predicted as S12 (financial) in the first level model. The selected model for financial entities was a Random Forest Classifier for multi-class target with 100 trees and selected parameters. We run some experiments considering XGBoost. Like the first level model, we excluded the XGBoost classifier due to an observed accuracy lower than 90%.

The selected model was validated on 24,350 records from the blind holdout dataset that were predicted as "Financial" by the first level model.

In Figure 3, the number of well predicted "Financial" entities by the second level model is 18,035 on the blind holdout dataset, corresponding to the 73% of the data. This means that the three digits ESA classification was correctly predicted as financial in 73% of the cases. The accuracy dropped significantly from the 90% of the first level model. The reason is related to the business choice of reducing false negatives. Among the 24,350 entities that were predicted as "Financial", we have a 23% of wrongly classified entities from first level model. This percentage is amplified in the second level model, as the errors made earlier cannot be corrected by the second level models.

Figure 3: Second-level model for financial entities - Results on blind hold out data


First level model prediction: number of entities predicted as "Financial"

24,350

18,035

Second level model prediction: number of entities predicted as "Financial"

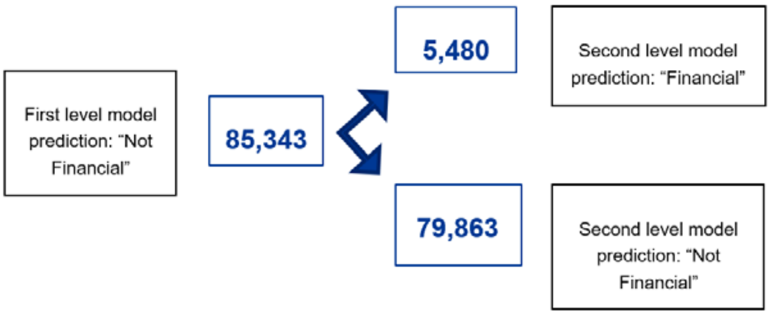
6,315

Second level model prediction: number of entities predicted as "Not Financial"

The second level model for not financial entities aimed to assign the ESA sector classification for entities that were predicted as "Not-Financial" by the 1st level model. The input data includes 85,343 records. The experiments motivated us to select a Random Forest Classifier with 100 trees and ad-hoc parameters. As in the previous models, alternatives to the Random Forest Classifiers demonstrated inferior performance.

The model was validated on records from the blind holdout dataset that were predicted as "Not Financial" by the first level model. In Figure 4 the number of well predicted "Not-Financial" entities by the second level model is 79,863 on the blind holdout dataset, corresponding to the 93% of the data. This means that the three digits ESA sector was correctly predicted as not financial in 93% of the cases.

Figure 4: Second-level model for not financial entities - Results on blind hold out data


First level model prediction: "Not Financial"

85,343

5,480

Second level model prediction: "Financial"

79,863

Second level model prediction: "Not Financial"

The model performance was measured by the accuracy scores on the blind holdout dataset, that resulted to be in line with the accuracy scores of the test phase. This means that the model's performance on new data was accurate as much as on the test data.

Overall, we compared the accuracy scores calculated on the validation set with those obtained in the testing phase. The results were very promising and made us confident in the model generalising well. The accuracy scores on the test data were very similar to those calculated on the blind holdout dataset, as shown in Table 3. Our models were predicting the ESA sectors well, compared to the performance during the test phase. The models did not show overfitting or underfitting problems.

Table 3: Accuracy score on test and blind data

	Accuracy score	
Model	Test data	Blind holdout dataset
First level: S12 vs not S12	0.9018	0.9018
Second level: S12	0.7326	0.7301
Second level: not S12	0.9874	0.9367

Conclusions

In this paper we estimated the institutional sector classification for legal entities according to the ESA 2010, based on basic entity reference information publicly available on the GLEIF website. The aim was to assign the correct institutional sector to GLEIF entities, in view of their potential registration in RIAD.

The proposed solution was leveraging data on approx. 550,000 entities from GLEIF which are already recorded in RIAD with their respective ESA sector classification. Such data was used to build a gold standard for training, testing and validating a supervised machine learning approach. The approach is composed of a two-level design, with a first level model to distinguish between financial and non-financial entities and two second level model to perform the fine-grained classification into the final ESA sectors. We employed semantic embeddings methods, feature engineering and selection, and random forest to address the task. The solution was tested and evaluated on a blind holdout dataset. The evaluation showed that the solution achieves high accuracy: 90% for the first-level model, 73% for the second-level model on financial entities and 93% for the second-level model on not financial entities. The adopted strategy also considered business-specific needs to bias the first level classifier to ensure a high recall for identifying financial entities.

The resulting automated, high-quality process for the bulk classification of entities according to their (ESA) institutional sector can be directly reused in the future. The produced estimates will serve as high-quality preliminary information supporting the expert assessment and the final entity classification. As a result, the institutional sector of nearly one million entities from GLEIF have been made available for the assessment of the RIAD experts before the potential recording in RIAD.

References

Colasanti, C., Macchia, S., & Vicari, P. (2009). The automatic coding of Economic Activities descriptions for Web users. *New Techniques and Technologies for Statistics*.

Eurostat. (2008). *NACE Rev. 2 - Statistical classification of economic activities.* Eurostat.

Eurostat. (2013). *The European System of Accounts — ESA 2010. Official Journal as Annex A of Regulation (EU), 549. doi:10.2785/16644*.

Giudice, O., Massaro, P., & Vannini, I. (2020, March). Institutional sector classifier, a machine learning approach. *Occasional Papers (Questioni di Economia e Finanza)*(548). Retrieved from https://www.bancaditalia.it/pubblicazioni/qef/2020-0548/QEF_548_20.pdf

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three Methods for Occupation Coding Based on. *Journal of Official Statistics, 33*(1), 101-122.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780.

ISTAT. (2009). *Classificazione delle attività economiche Ateco 2007.* Roma: Istituto nazionale di statistica. Retrieved from https://www.istat.it/it/files//2011/03/metenorme09_40classificazione_attivita_economiche_2007.pdf

Kühnemann, H., van Delden, A., & Windmeijer, D. (2020). Exploring a knowledge-based approach to predicting NACE codes of enterprises based on web page texts. *Statistical Journal of the IAOS, 36*(3), 807-821.

Lenoci, F., & Letizia, E. (2021). Classifying counterparty sector in EMIR data. In *Data Science for Economics and Finance.* Springer.

Noyvirt, A. (2021). *FinBins – granular classification of the UK's financial sector*. (Office for National Statistics - Data Science Campus) Retrieved April 28, 2021, from https://datasciencecampus.ons.gov.uk/project/finbins-granular-classification-of-the-uks-financial-sector/

Roelands, M., van Delden, A., & Windmeijer, D. (2018). *Classifying businesses by economic activity using web-based text minin.* Statistics Netherlands.

EUROPEAN CENTRAL BANK

EUROSYSTEM

Estimating the institutional sectors of legal entities on a large scale

A supervised learning approach

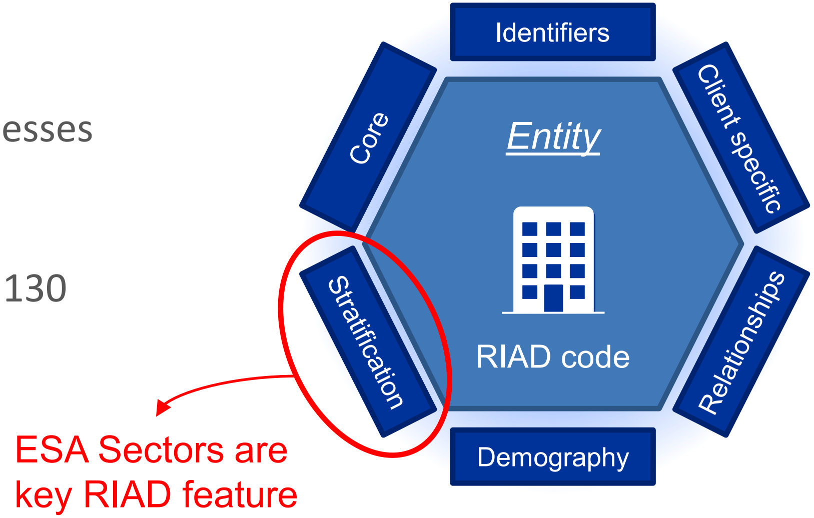
20th October 2021

Francesca Benevolo, Thomas Gottron,
Ilaria Febbo, Nicolò Pegoraro
European Central Bank

RIAD: shared master dataset on legal entities

- The Register of Institutions and Affiliates Data (RIAD):
  - is a shared master-dataset
  - supports several clients and business processes across the ESCB[1], SSM[2] and EBA[3]
  - Has more than 12 Mn entities, more than 130 attributes

RIAD



ESA Sectors are key RIAD feature

[1] European System of Central Banks
[2] Single Supervisory Mechanism
[3] European Banking Authority

ESA 2010 sector classification*

European System of Accounts (ESA) is internationally compatible accounting framework for a systematic and detailed description of a total economy

	ESA sector	Description
	S11	Non financial corporations
Financial sector	S121	Central banks
	S122	Deposit-taking corporations except the central bank
	S123	Money Market Funds (MMFs)
	S124	Non-MMF investment funds
	S125	Financial corporations other than MFIs, non-MMF investment funds, financial auxiliaries, captive financial institutions and money lenders, insurance corporations and pension funds
	S126	Financial auxiliaries
	S127	Captive financial institutions and money lenders
	S128	Insurance corporations
	S129	Pension funds
	S1311	Central government (excluding social security funds)
	S1312	State government (excluding social security funds)
	S1313	Local government (excluding social security funds)
	S1314	Social security funds
	S14	Households
	S15	Non profit institutions serving households

* As represented in RIAD

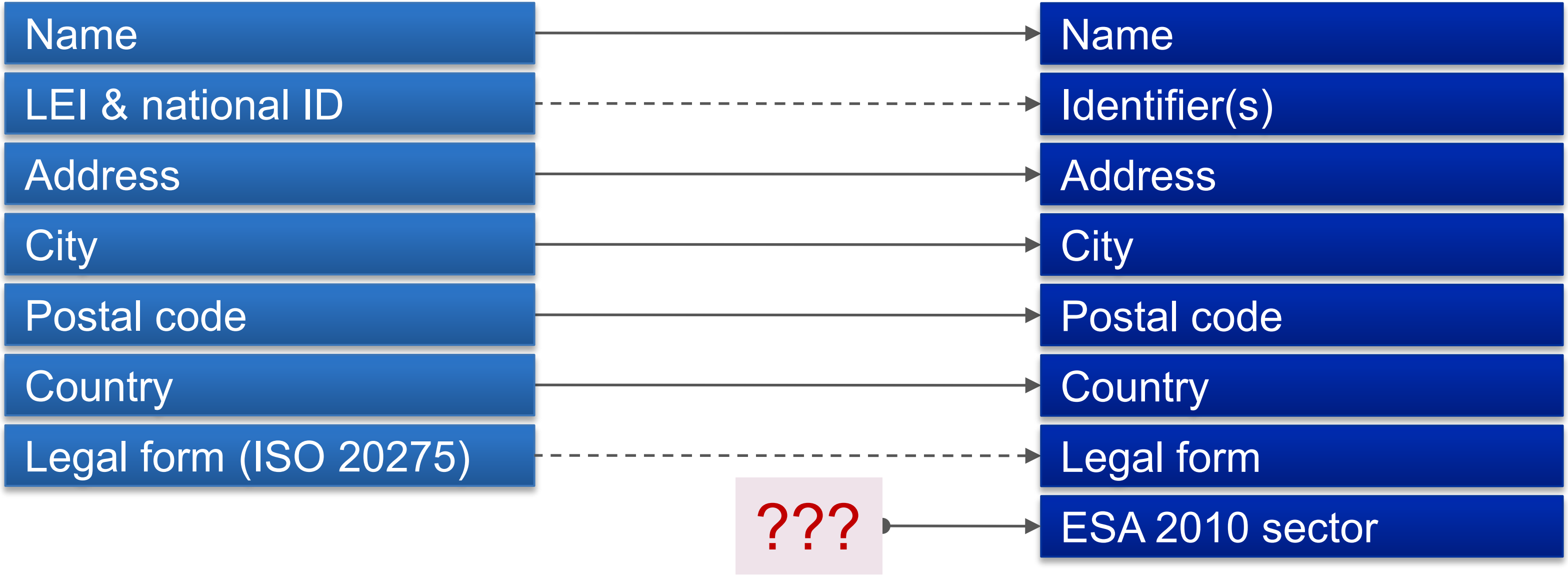
GLEIF: Legal Entity Identifiers and reference data

- GLEIF is a non-profit organisation
- GLEIF provides legal entity identifiers (LEI) for corporations and other organisations
- Contains information on approx. 1,6 Mn entities
- Entities involved in financial transactions need to have an LEI

Attributes in GLEIF & RIAD

LEI GLOBAL LEGAL ENTITY IDENTIFIER FOUNDATION

RIAD



Name → Name

LEI & national ID → Identifier(s)

Address → Address

City → City

Postal code → Postal code

Country → Country

Legal form (ISO 20275) → Legal form

??? → ESA 2010 sector

Business case

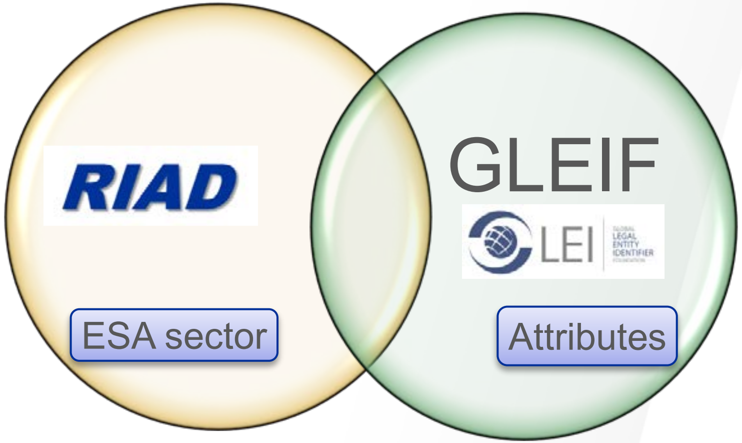
Problem: Assigning ESA sectors to newly recorded RIAD entities is a manual and time consuming process (esp. for non-EU).

Scope: Creating a model based on public GLEIF information able to automatically estimate ESA sectors so to streamline RIAD experts' work.

Question: How to estimate the ESA sector classification based only on GLEIF data starting from a simple LEI code?

Solution: A supervised learning approach trained on RIAD data applicable for present and future needs.

Supervised Learning Approach



545,541 entities in both GLEIF and RIAD

➡ These entities have the ESA sector available

TRAIN, TEST, VALIDATE

963,652 entities only in GLEIF

➡ ESA sector to be predicted for these entities

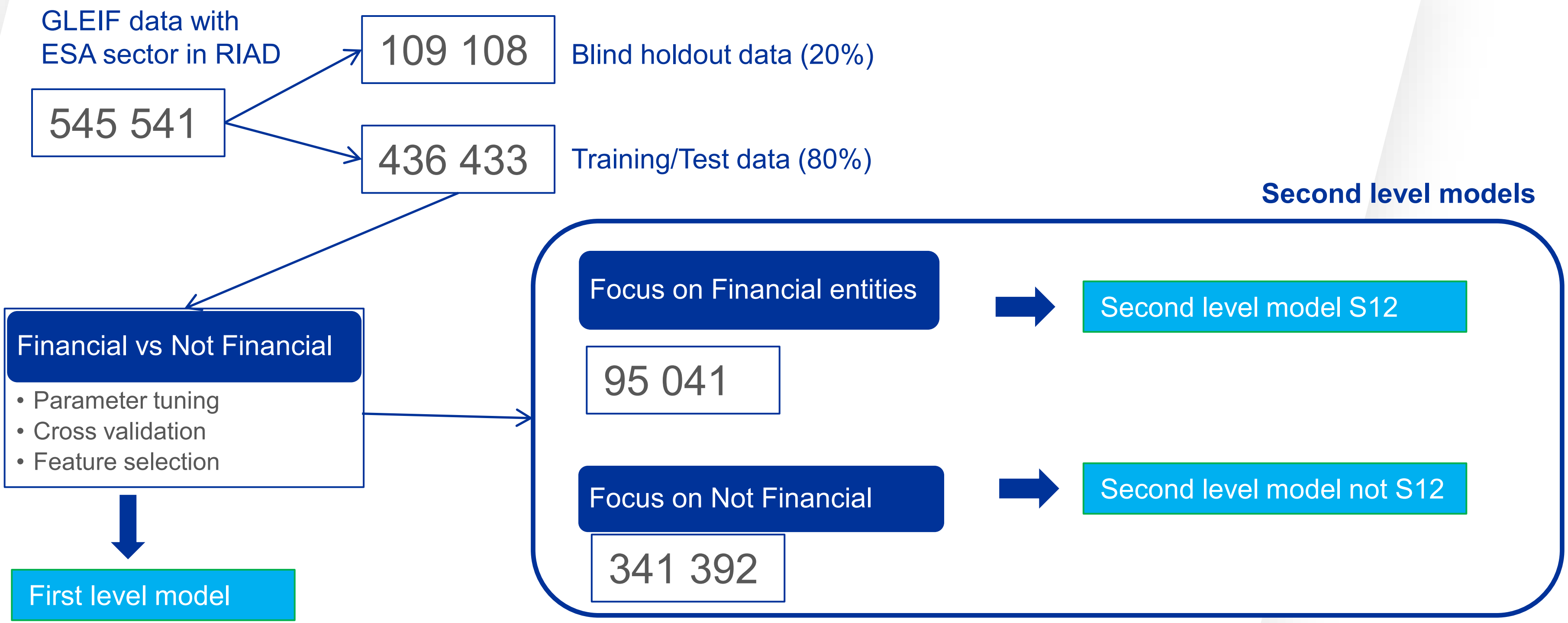
REAL DATA

Training/Test data: entities in both databases.

Target variable: ESA sector.

Predictors: GLEIF attributes.

Process design



GLEIF data with
ESA sector in RIAD

545 541

109 108

Blind holdout data (20%)

436 433

Training/Test data (80%)

Second level models

Financial vs Not Financial

- Parameter tuning
- Cross validation
- Feature selection

First level model

Focus on Financial entities

95 041

Second level model S12

Focus on Not Financial

341 392

Second level model not S12

Methodology

FEATURE ENGINEERING

Legal Name was encoded with semantic embedding to improve the predictions

PARAMETERS TUNING

Comparison of Random Forest input parameters to find the best combination

CROSS VALIDATION

The best model was selected among 72 options based on the accuracy.

BLIND HOLDOUT DATA

Additional 100 000 entities used to confirm the quality of the models in the end

Two levels model

Priority: to find all Financial entities (S12)
→ reduce false negative

First level model:
Predict if an entity is financial (S12) or not.

Second level models:
Predict ESA sector 3-digits code.

Distribution of ESA sector in the data

	Frequency	Percentage
Financial S12	118,554	**22%**
Not financial S12	426,987	**78%**

First level model accuracy score: 90%

Improvement from baseline (78%): the first level model distinguishes financial and not-financial entities with 90% probability

Second level model accuracy score: 73%

Improvement from baseline (43%): the second level model predicts the 3-digits ESA sector for financial entities with 73% probability

Conclusions

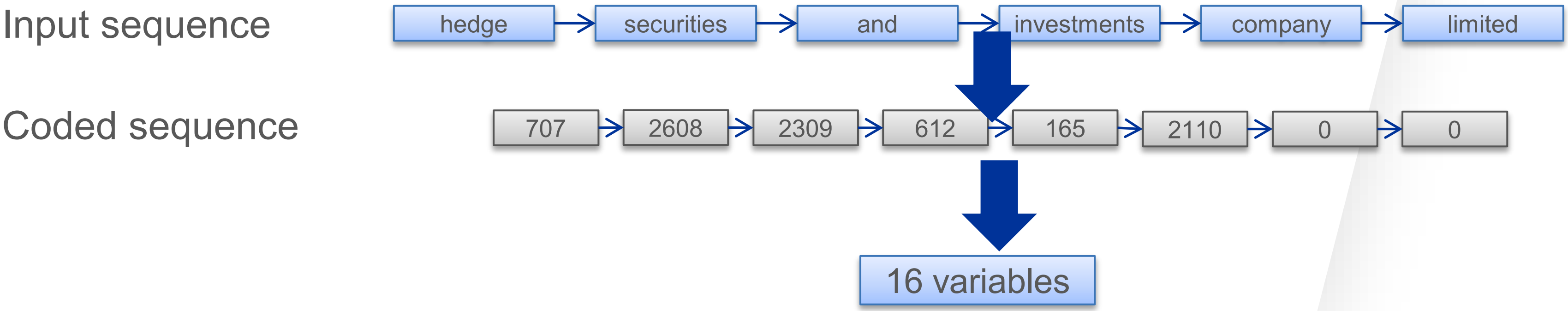
- The application estimated the ESA sector for 963,652 GLEIF entities.
- The semantic analysis on legal names added value to the models.
- The parameters fine tuning and cross validation search helped to find the best model.
- Benefits for the business areas (efficiency gain, prioritisation, data availability)
- The innovative aspect of our work was to estimate missing data using reference data only.

Appendix

Appendix: Embedded variables

The Legal Name was encoded with semantic embedding.

Results: 16 embedded variables were generated and used as models predictors, improving the overall accuracy.



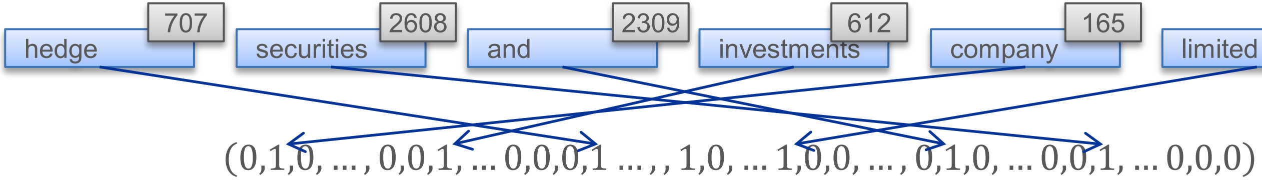
Appendix: Embedded variables

Embedded variables: incorporate name information in the classification task.

HEDGE SECURITIES AND INVESTMENTS COMPANY LIMITED

Traditional approach: Bag-of-words

- ➢ Each word corresponds to an index number (using a dictionary)
- ➢ Vector setting the index entry to 1 if the word is present.



hedge 707

securities 2608

and 2309

investments 612

company 165

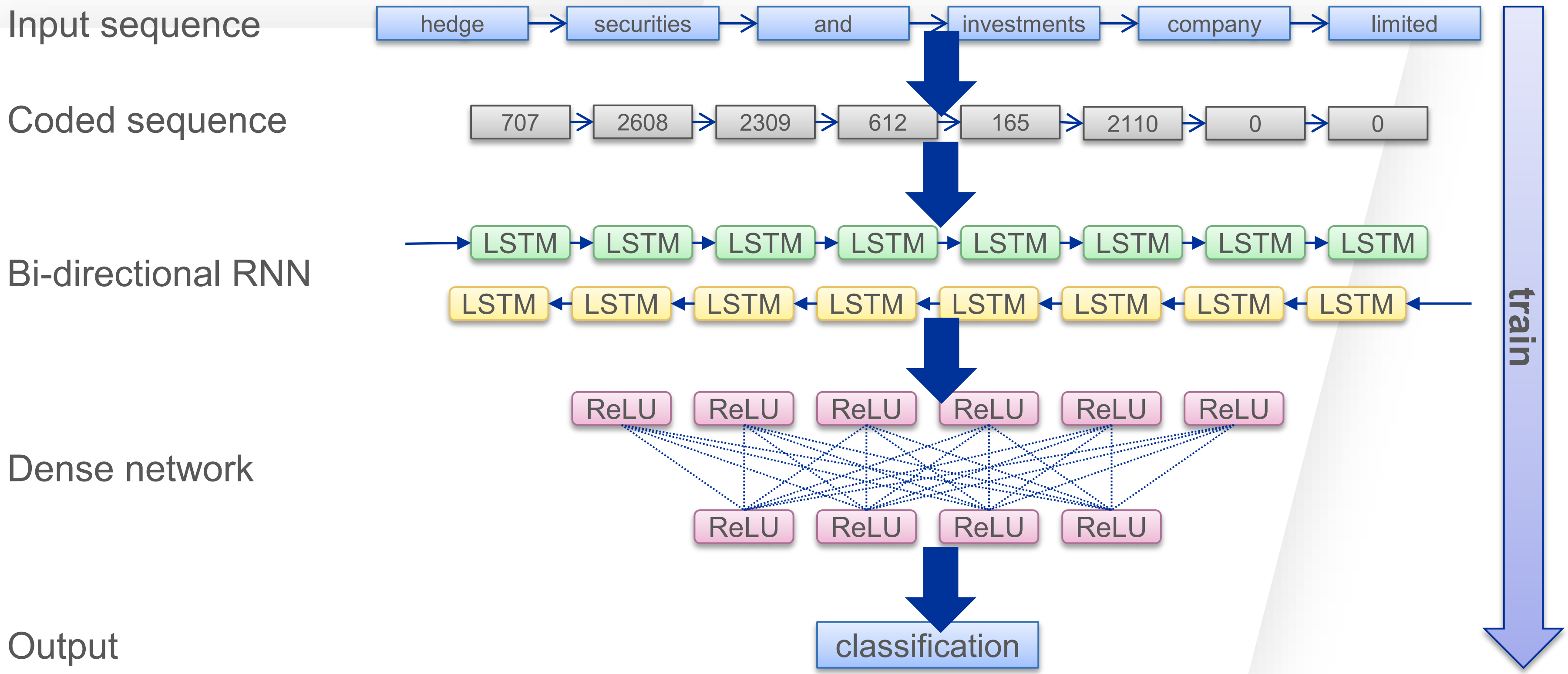
limited 2110

(0,1,0, ... , 0,0,1, ... 0,0,0,1 ... , , 1,0, ... 1,0,0, ... , 0,1,0, ... 0,0,1, ... 0,0,0)

Drawback:

- ➢ Space of words is of very high dimension and sparsely populated
- ➢ Word order is lost in this representation

Appendix: Embedded variables



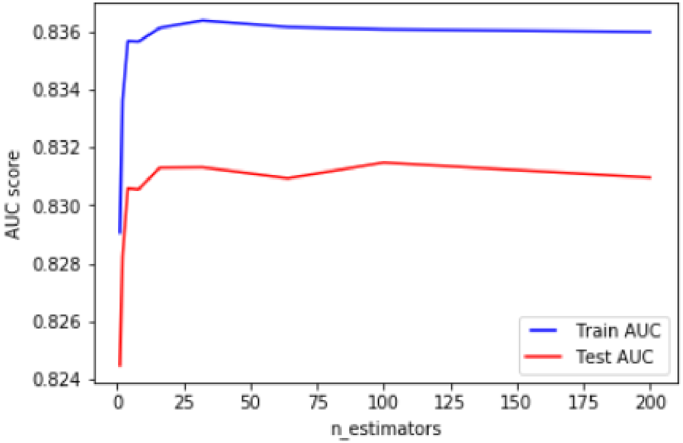
Appendix: Methodology

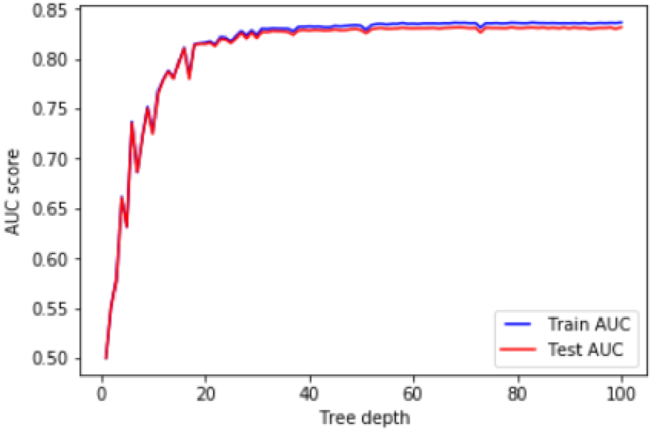
Top features used to predict the ESA sector:

- Category FUND
- Embedded variables from the semantic analysis of legal name
- Luxemburg as legal basis
- Legal form
- Presence of words HOLDING, INVEST, BANK, FUND in the legal name
- Registration authority

Methodology: Parameters fine tuning

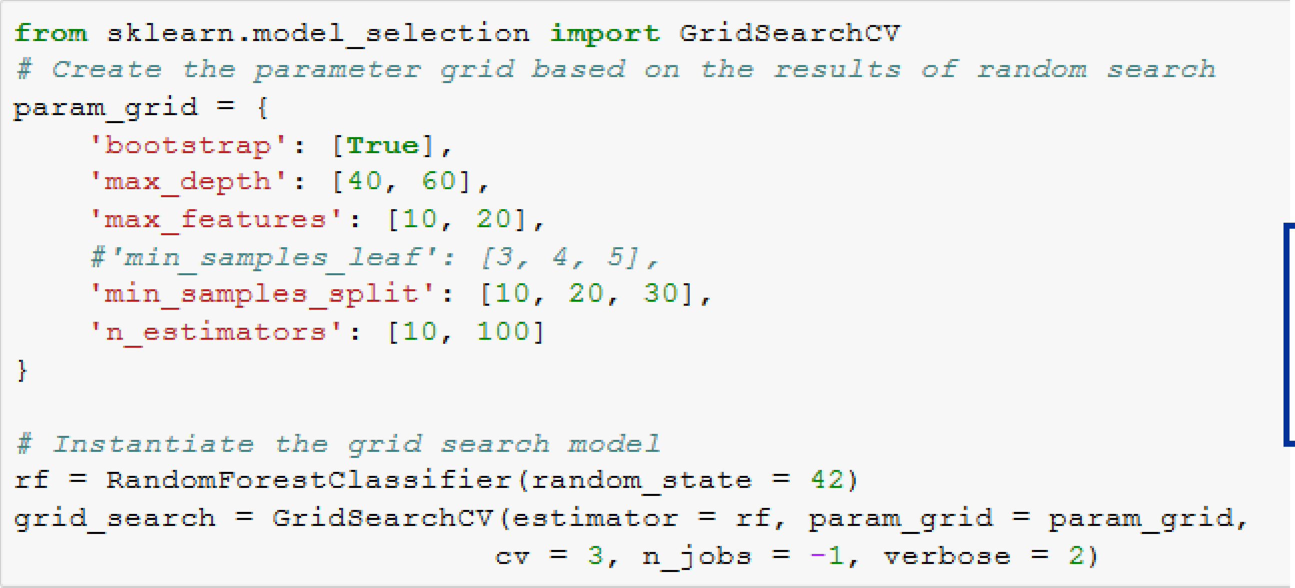
Random Forest parameters: N estimators, tree depth, minimum sample leaf, min sample split, max features.





Methodology: Cross validation

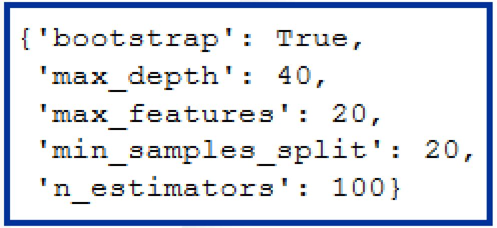
Random Forest parameters grid with 3 folds for 24 combinations, totalling 72 fits



```
from sklearn.model_selection import GridSearchCV
# Create the parameter grid based on the results of random search
param_grid = {
    'bootstrap': [True],
    'max_depth': [40, 60],
    'max_features': [10, 20],
    #'min_samples_leaf': [3, 4, 5],
    'min_samples_split': [10, 20, 30],
    'n_estimators': [10, 100]
}

# Instantiate the grid search model
rf = RandomForestClassifier(random_state = 42)
grid_search = GridSearchCV(estimator = rf, param_grid = param_grid,
                           cv = 3, n_jobs = -1, verbose = 2)
```

Best model from Grid Search CV



```
{'bootstrap': True,
 'max_depth': 40,
 'max_features': 20,
 'min_samples_split': 20,
 'n_estimators': 100}
```

Methodology: Blind holdout dataset

The models were evaluated on a blind holdout dataset to verify the accuracy.

Result: The accuracy scores on the blind holdout dataset were very close to the accuracy scores on test data.

→ The models are stable

Irving Fisher Committee on
Central Bank Statistics

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Data science opportunities with non-cash transactional payments[1]

Per Nymand-Andersen,
European Central Bank

[1] This presentation 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 central banks and other institutions represented at the event.

EUROPEAN CENTRAL BANK

EUROSYSTEM

Data science opportunities with non-cash transactional payments

IFC & Bank of Italy
Workshop on
“Data Science in Central Banking”

22/10/2021

Per Nymand-Andersen
European Central Bank
Per.nymand@ecb.int; percannes@gmail.com

Overview

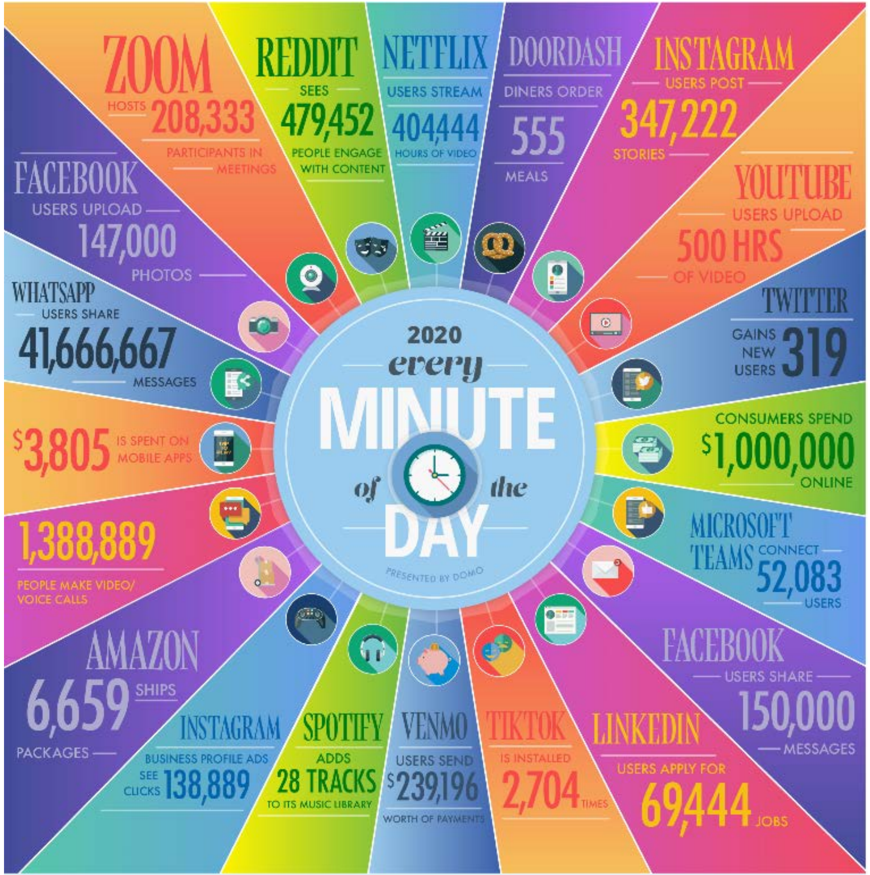
1 Fintech & data science - Data never sleeps

2 Data science pilots - Data mania versus phobia

3 Discovery - Partnership - insights for tool-kits

Data never sleeps – Alternative sources

J. Powell (July 29, 2020):"Non-standard, high frequency data has become a very important thing"



Source: 2020: What Happens In An Internet Minute, Lori Lewis, domo.com

Pandemic triggering (daily, weekly)

- Increase in use of digital platforms

(online shopping, social media, music and film streaming, booking & trading platforms)

- Digital data economy – tracking
  - Consumer's spending & savings
  - Credit and debit cards
  - Booking platforms (restaurants, air, housing)

A borderless market for digital data

European Data Strategy

- Data service act – Governance
- Cloud services
- Artificial Intelligence
- Machine readable digital formats
- Identifiers (entities, instruments, transactions)
- General regulation on data protection
- Public, academia, Private data exploration for knowledge

Fintech – A paradigm of borderless records

Digital transformation in finance and economics

E- trading

Clearing & Payment systems

Credit cards

Mobile pay

Price scans

S-media

Digital data

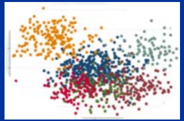
- DLT, Block chain
- Crypto-currency
- Digital assets
- S-contracts
- Token trading
- Supply chain management
- Ecosystem management

- Kickstarter & Indiegogo
- Lending & financing
- Peer to peer lending
- Robo advisors

- Alibaba,
- Ant Finance
- Ant Fortune
- Zhima scores
- Alipay
- Credit pay

Data Science lab

Systematic acquire, Structure, Process,

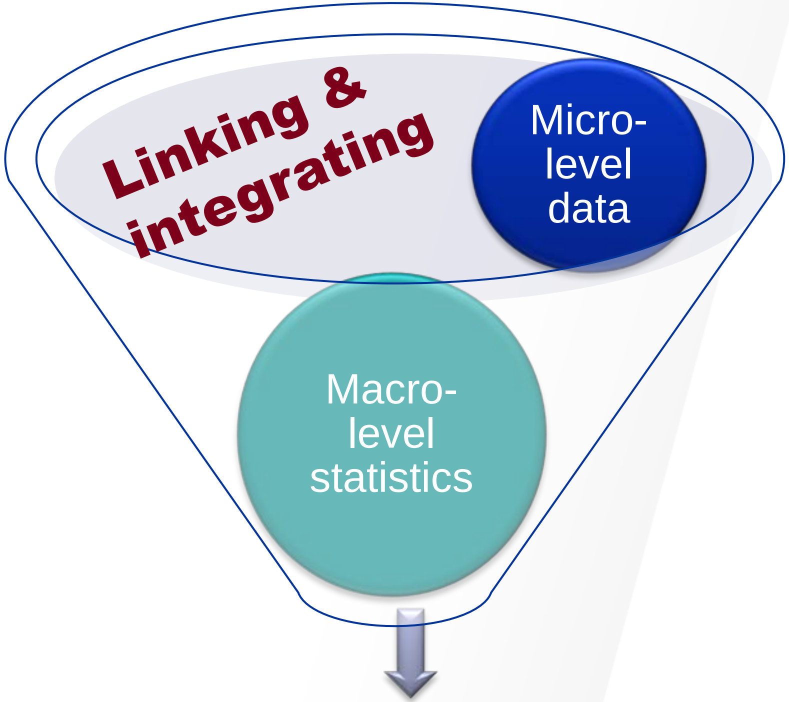


Statistical algorithm, pattern detections, Machine learning, AI

Linking insights for new services & competitive advantages

Source: **Data Science in Economics and Finance for Decision Makers,** Riskbooks.com 2021, Per Nymand-Andersen

The paradigm shift following the financial crisis



Data science analytics

Micro-level statistics

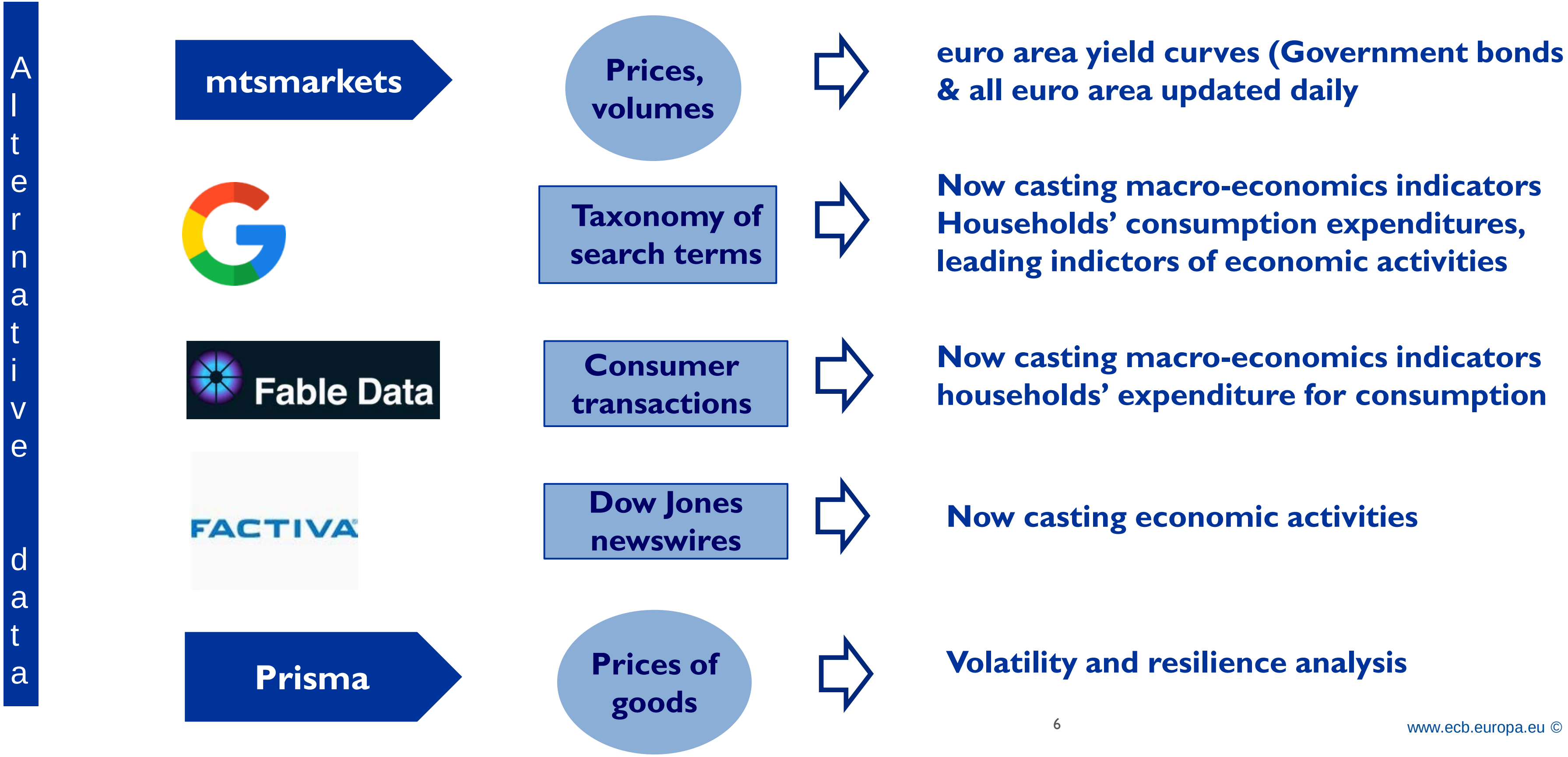
- Security-by-security issuance/price
- Holdings of individual securities
- Interbank lending and holdings
- Banks individual loans to corporates
- Identifiers of Financial Institutions
- Individual bank supervisory data

Macro-level statistics

- Linking to sector analysis
- Securities issues & Banks interest rates
- Government finance & financial accounts

Data Science Analytics – Engage with your data

Advanced analytics – 30 experimental pilots



Alternative data

mtsmarkets → Prices, volumes → euro area yield curves (Government bonds & all euro area updated daily

Taxonomy of search terms → Now casting macro-economics indicators Households' consumption expenditures, leading indictors of economic activities

Fable Data → Consumer transactions → Now casting macro-economics indicators households' expenditure for consumption

FACTIVA → Dow Jones newswires → Now casting economic activities

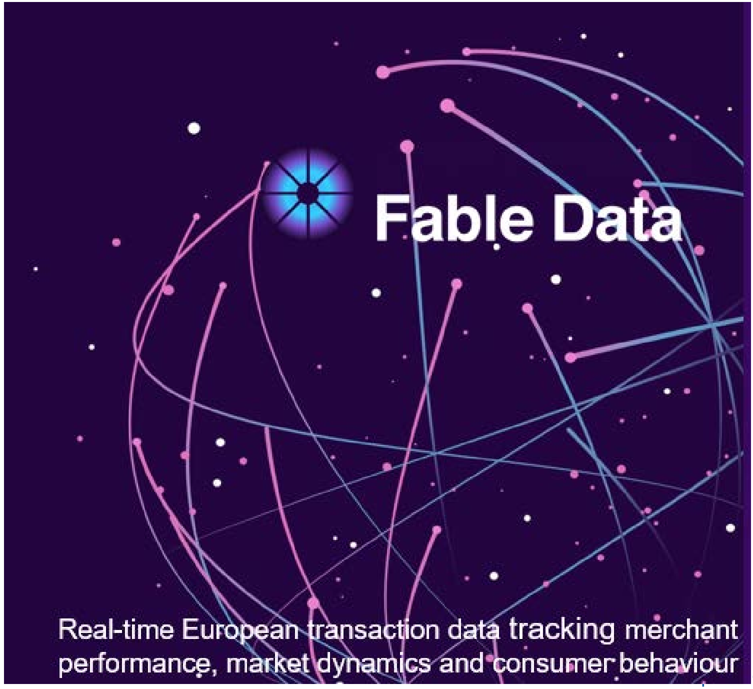
Prisma → Prices of goods → Volatility and resilience analysis

FinTech and partnerships

Case studies

None-cash consumer spending using credit card transactions partnership with alternative data sources

Picture: conns.com

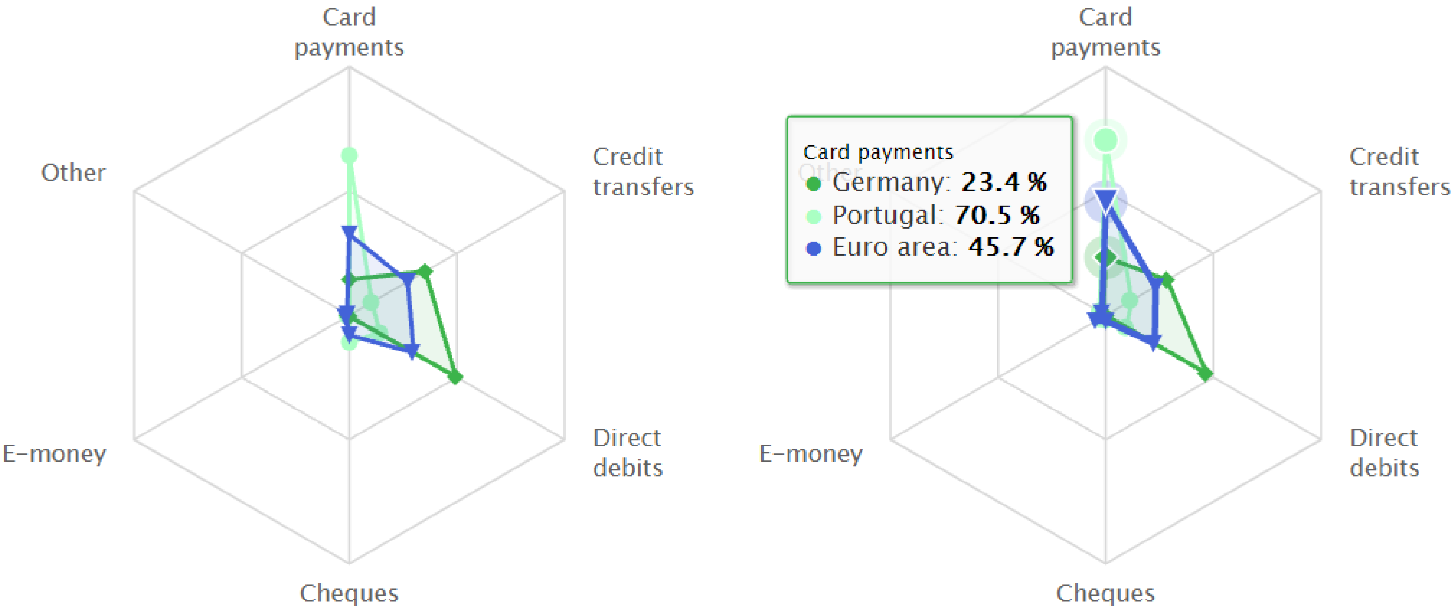

Fable Data
Real-time European transaction data tracking merchant performance, market dynamics and consumer behaviour

©

Non-cash transactional payments

Euro area card payments double in a decade

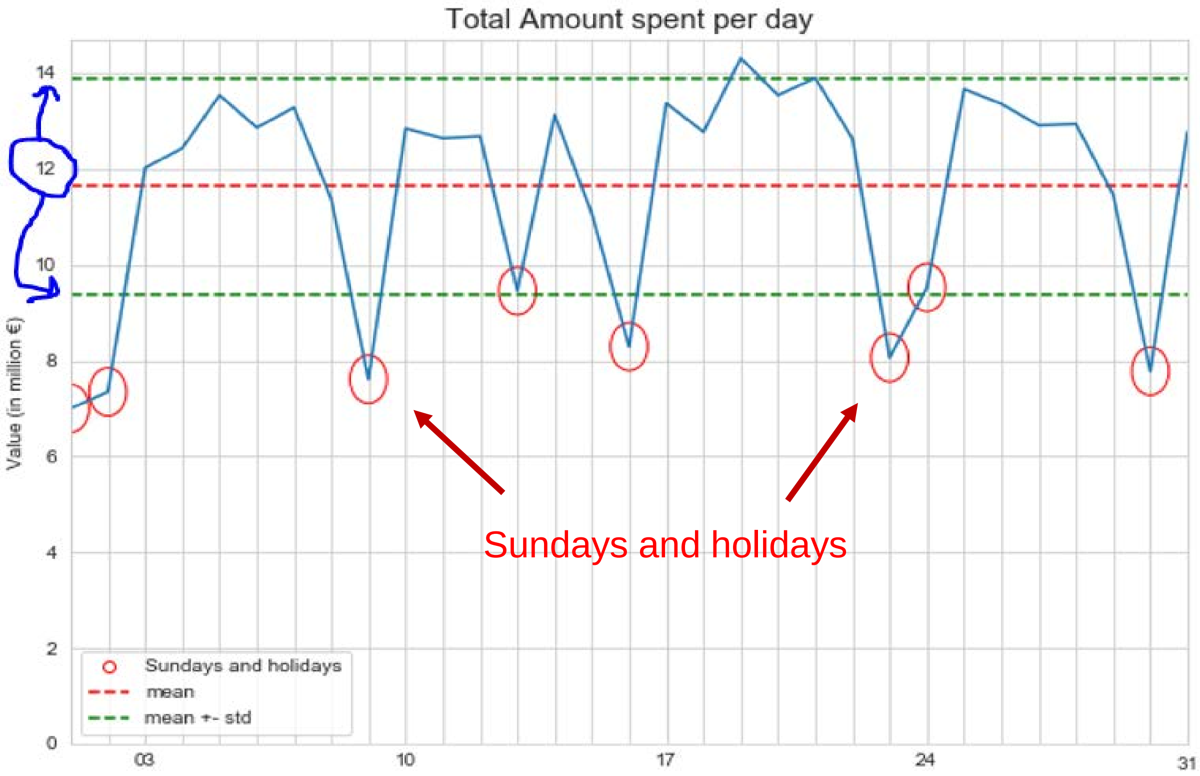
Share of total number of non-cash payments per payment method



Today, Europeans are most likely to reach for a card when they want to make a payment without using cash. While the number of card transactions in the euro area has more than doubled in the last decade, the average value of each transaction has fallen

Non-cash transactional payments

Fable data sample: May 2021 - credit card transactions (1 million observations)



Standard deviation

Average

Standard deviation

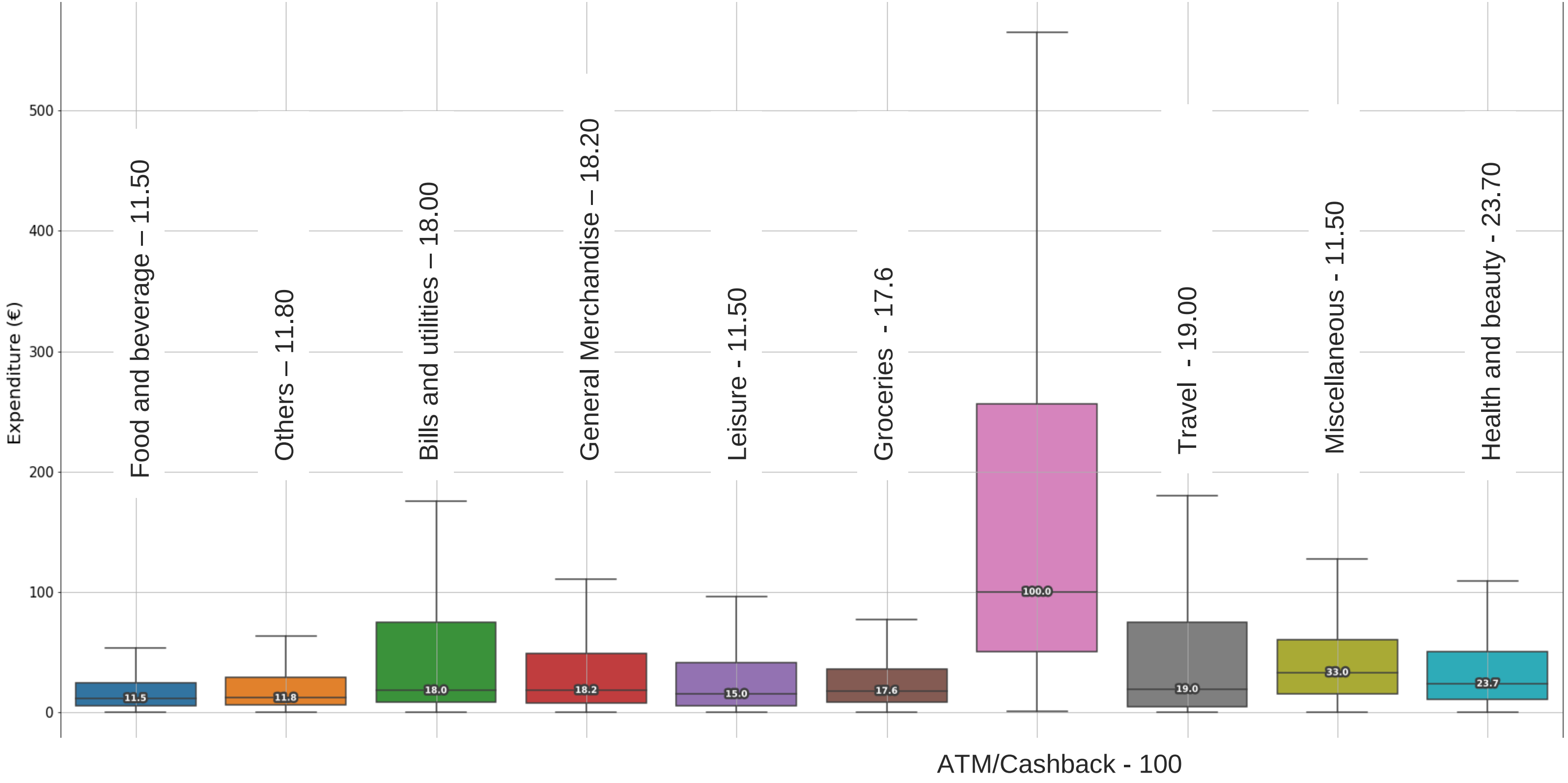
Sundays and holidays

Days within month

This is work in progress and suggestions are welcome !

Non-cash transactional payments

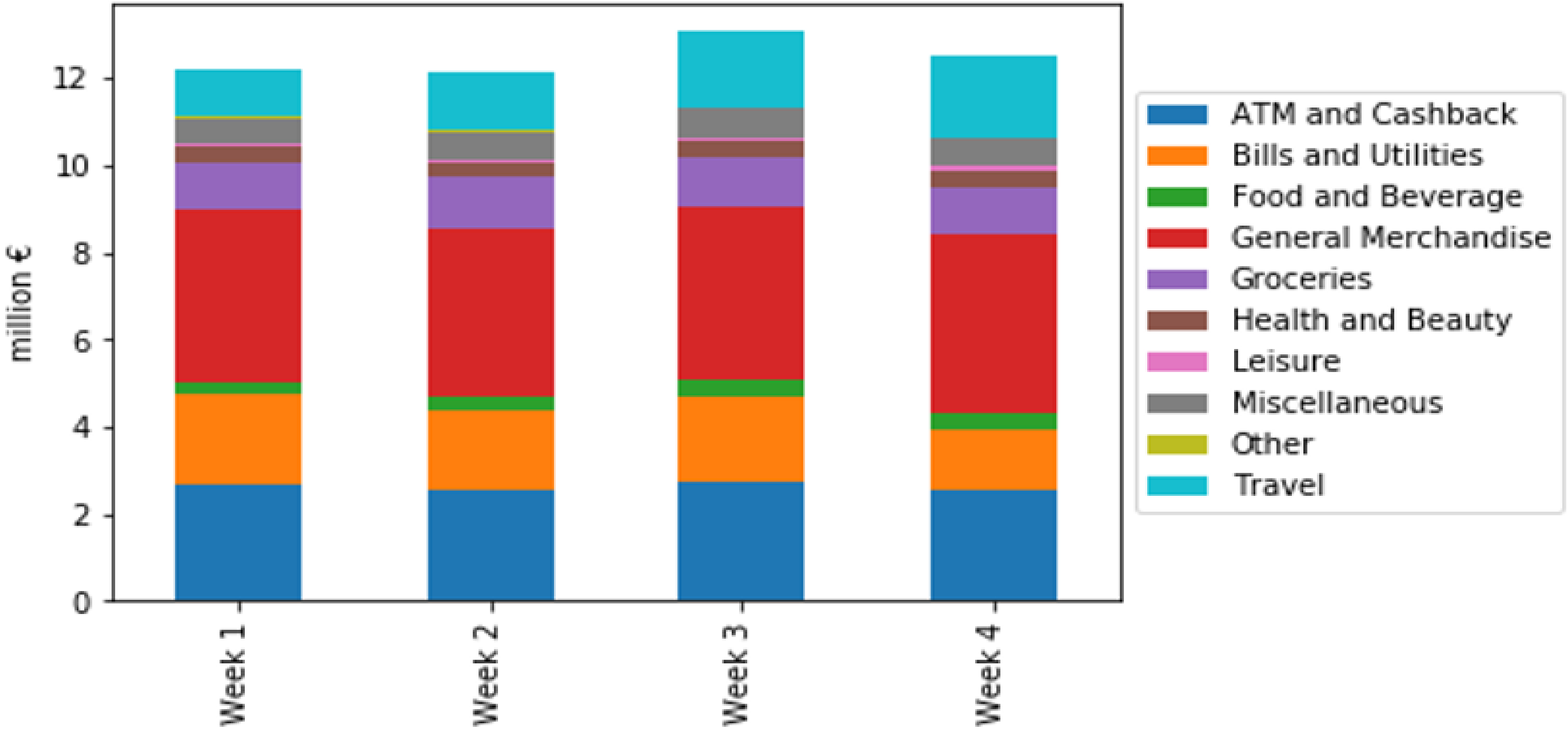
Average/high/low expenditures/transaction in sector classification (Fable categories)



b.europa.eu ©

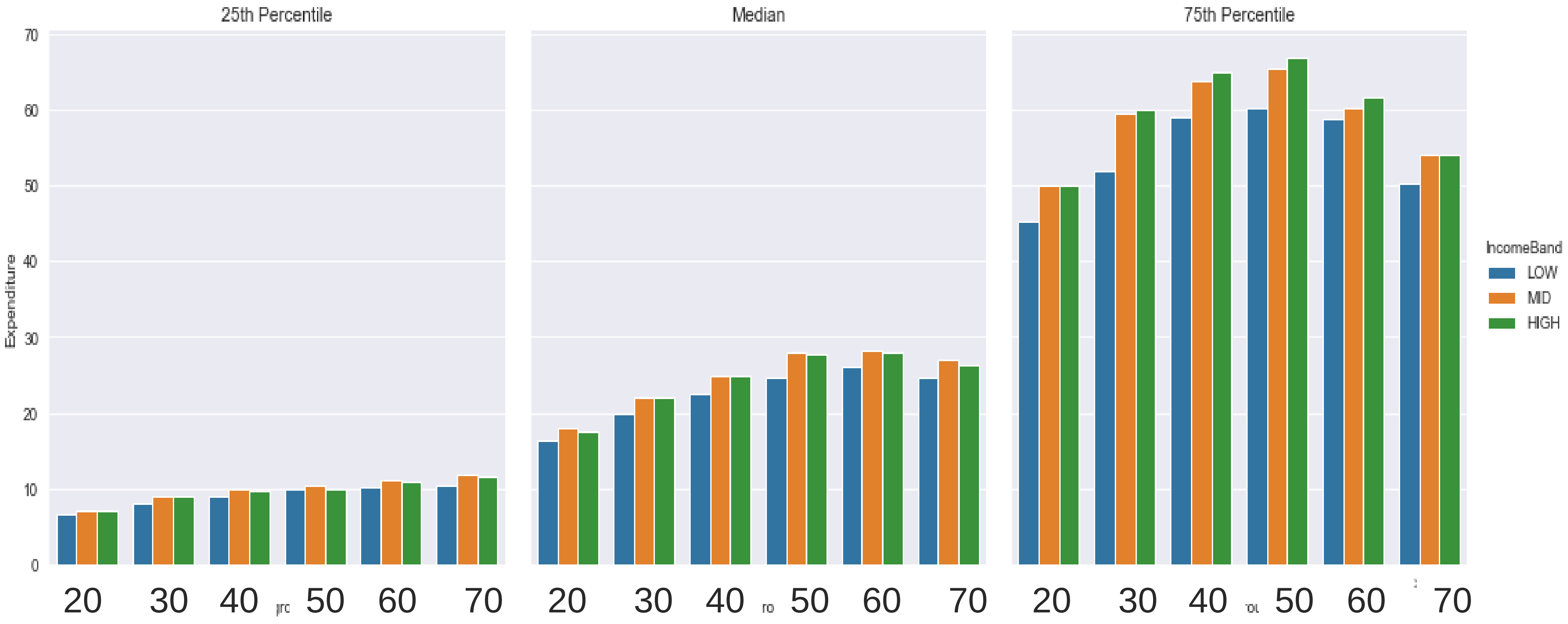
Non-cash transactional payments

Average daily spending per week per sector in May 2021



Non-cash transactional payments

Expenditure by age and income group (high, medium, low)
25th – median – 75th percentile

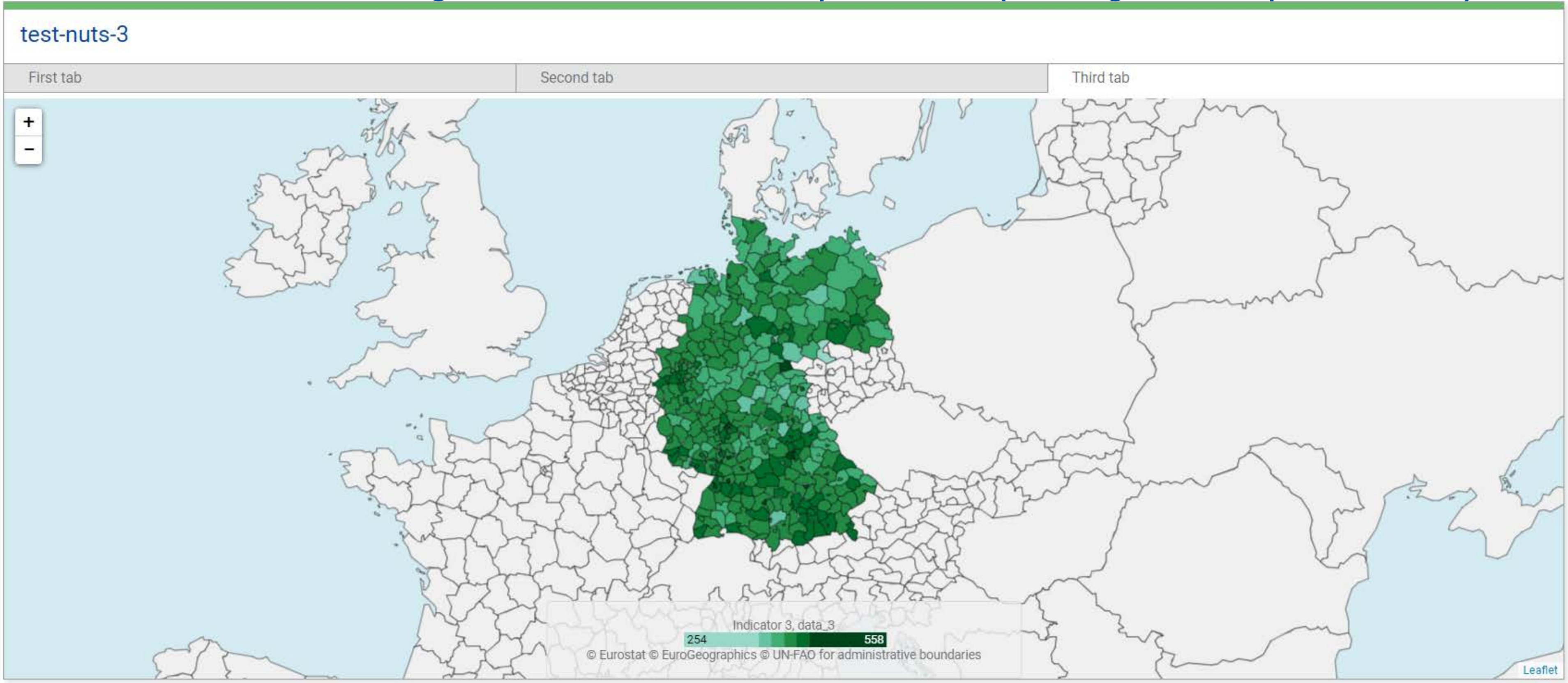


Non-cash transactional payments

EURO AREA STATISTICS

ENGLISH

Demo zoom in on regional distributional expenditure (average, total, perhabitant)

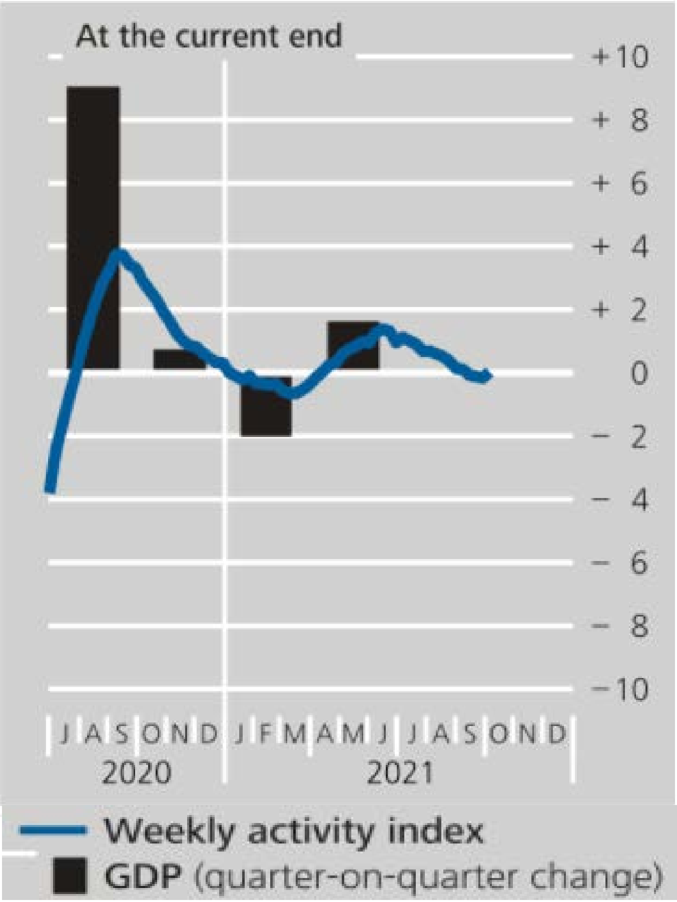


test-nuts-3

© Eurostat © EuroGeographics © UN-FAO for administrative boundaries

Data science tools to answer research questions!

Timely Weekly Activities Index (WAI) calculated by Deutsche Bundesbank



The WAI is a weekly index designed to measure real economic activity in Germany in a timely manner

Similar to the index published by the Federal Reserve Bank of New York (Lewis et al., 2020).

Based on a 10 high-frequency alternative data sources available covering various economic sectors

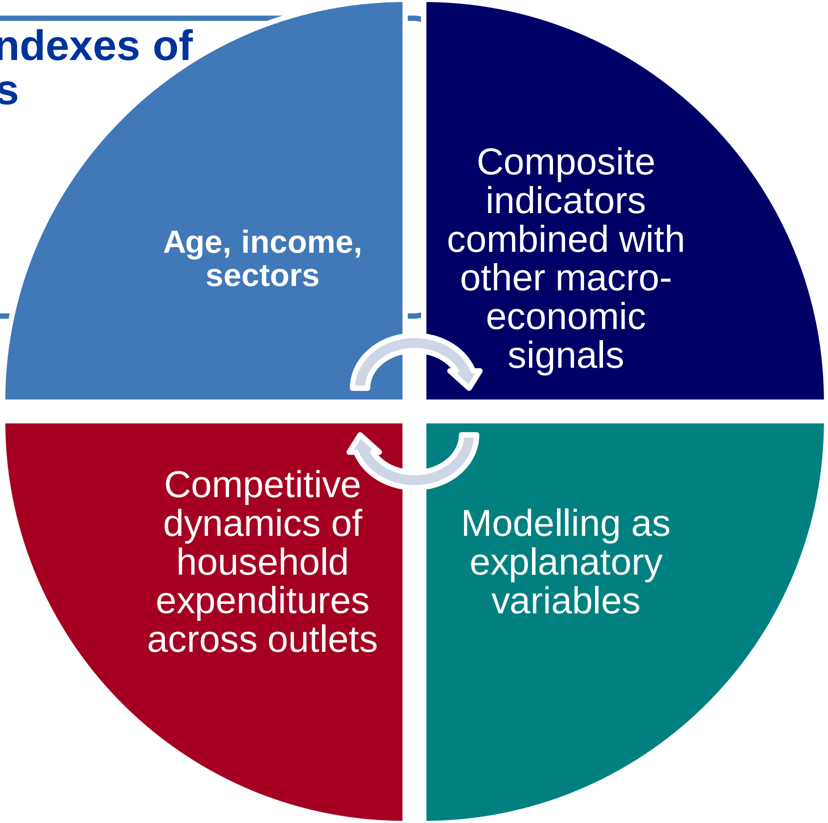
- Credit card data & Google search data, consumer confidence, electricity consumptions, toll index (trucks)….

Source: Eraslan, S. and T. Götz (2020), An unconventional weekly economic activity index for Germany, Deutsche Bundesbank Technical Paper, 02/2020.

Data science tools to answer research questions!

Time series analysis of household expenditures for monetary policy, financial stability, system risks analysis, banking supervision,

- **Creating indexes of economics activities**



Age, income, sectors

Composite indicators combined with other macro-economic signals

Modelling as explanatory variables

Competitive dynamics of household expenditures across outlets

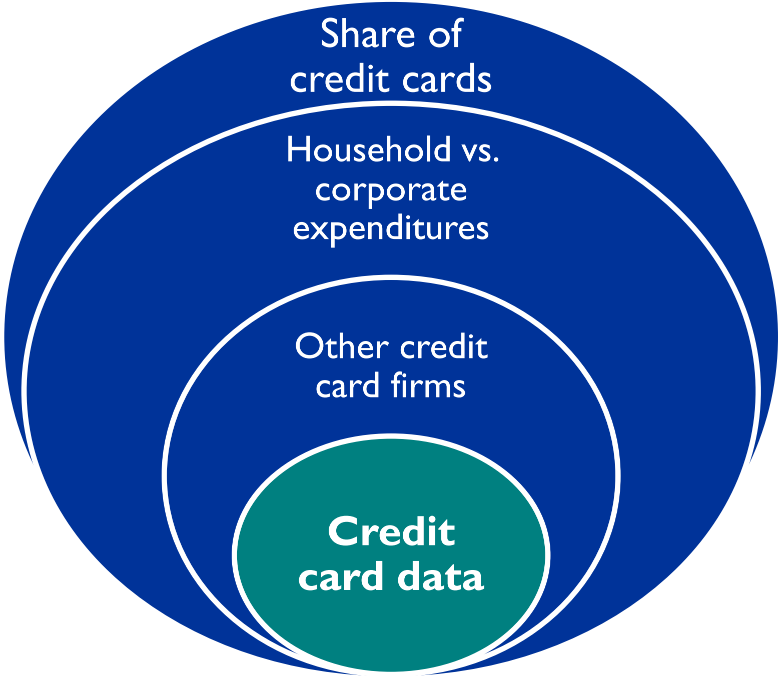
Data Science analytics – Engage with your data !

One **misperception** of big data is that we **do not need** to worry about **sample bias and representativeness**, as large volumes of information supersede standard sampling theory, since big data provide census-type information

1

Digital recording of operations

Census?



2

Unit measurement

Not people –
Number of transactions

3

Event driven

Volume changes may not necessary refer to changes in demand (more credit card payments during lock-down/covid-19?)

www.ecb.europa.eu ©

Data Science and partnerships

Three takeaways

Progress lies in experimenting

Amara’s law

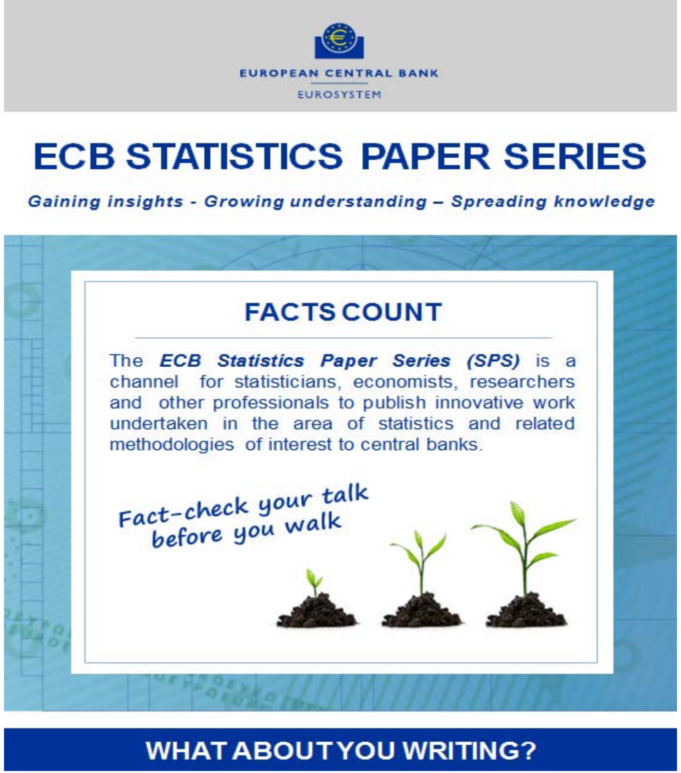
Valuable source for economic activities

Data Science labs

Moving from experimenting to tool-kits

Leveraging on Fintech partnerships for excellence

Questions and (hopefully) answers



EUROPEAN CENTRAL BANK

EUROSYSTEM

ECB STATISTICS PAPER SERIES

Gaining insights - Growing understanding – Spreading knowledge

FACTS COUNT

The ***ECB Statistics Paper Series (SPS)*** is a channel for statisticians, economists, researchers and other professionals to publish innovative work undertaken in the area of statistics and related methodologies of interest to central banks.

Fact-check your talk before you walk

WHAT ABOUT YOU WRITING?

IFC-Bank of Italy Workshop on “Machine learning in central banking”

19-22 October 2021, Rome / virtual event

Using twitter data to measure inflation perception[1]

Julien Denes, Ariane Lestrade and Lou Richardet,
Bank of France

[1] This presentation 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 central banks and other institutions represented at the event.

Using Twitter Data to Measure Inflation Perception

A working paper

Julien Denes **Ariane Lestrade** **Lou Richardet**

Banque de France

Abstract

Anchoring inflation expectations and measuring the current and future impact of prices evolution is a crucial issue for central banks. With the current rise of social networks, a new source of information has appeared to measure inflation perception. In this study, we propose an indicator of inflation perception based on Twitter data, focusing on the specific community of users who retweeted posts of the Banque de France account. Despite the bias induced, this strategy allows to efficiently extract information from an expert community on economic and financial subjects, capturing a potentially more relevant signal. We combine supervised machine learning and natural language processing methods with dictionary-based filters to classify all tweets posted by these retweeters, by first detecting whether they relate to prices issues, and then assessing which of the selected tweets mention inflation, deflation or is off-topic. Finally, we create a Twitter indicator of inflation perception, which is the difference between the number of tweets about inflation and the number of tweets about deflation. The resulting Twitter indicator is consistent with monthly household surveys on inflation expectations and perception, and is highly correlated with the inflation rate. We also show that it has a strong anticipatory power of the future inflation. These results suggest that it is both possible and relevant to use Twitter data to construct a daily measure of inflation perceptions.

JEL Classification: E31, C53, C55, D84, E58.

Keywords: inflation, inflation perceptions, inflation expectations, Twitter data, text mining, machine learning, big data, real-time data, high-frequency data.

This working paper describes preliminary results of ongoing research. It is made available to the public solely to elicit discussion and comments. Views expressed in the paper are those of the authors and do not reflect the position of the Banque de France.

Authors' email address: julien.denes@banque-france.fr.

1. Introduction

Maintaining price stability over the medium term is one of the most essential task of central banks as it highly influences the trust economic agents have on the future. This raises the key question of how to measure inflation, but most importantly how to measure the inflation expectations of the population in order to anticipate their behavior. Formally, inflation is defined by the French National Institute for Statistics (INSEE) as the loss of the purchasing power of money, which translates into a general and sustainable increase in prices.[1] To measure this phenomenon on the long run, the Harmonized Index of Consumer Prices (HICP) is the most used tool, in particular because it was designed to allow for international comparison. Calculated monthly, it makes it possible to estimate, between two given periods, the average change in the prices of products consumed by households. The methodology is harmonized within the European Union to allow for comparisons between each member state national Consumer Prices Index (CPI).

The most crucial question however is rather how economic actors perceive, anticipate, and react to inflation. To measure this perception of inflation, several sources can be used. First, some indicators are constructed using surveys. For instance, the European Central Bank (ECB) produces the Survey of Professional Forecasters, who are asked to forecast inflation rate among many other macroeconomic values. Likewise, the Consensus Economics survey and the Eurozone Barometer are both published every month. Some surveys focus specifically on inflation, such as the Atlanta Fed Business Inflation Expectations, in which participants provide their estimations of future inflation as values, or assign probabilities that inflation is within predefined ranges. Some other indicators rather survey the perception of non-professional individuals within the general population. For instance, INSEE's Monthly Household Survey includes a few questions about perceived past inflation and expected future inflation rate.

Second, indicators can also be extracted from financial markets, for instance using "break-even" inflation rates and inflation swap markets. Break-even inflation rate is the difference between the yield of a nominal bond and an inflation-linked bond of the same maturity. A swap is a product that converts an inflation-indexed loan (or borrowing) into a fixed-rate loan (or borrowing). Unlike surveys, these indicators are available at a high frequency and react more quickly about changes in the economy.

However, both those sets of indicators have some limits. For instance, if respondents to the ECB Survey of Professional Forecasters are convinced of the Bank's credibility, their answers will reflect this opinion rather than their true inflation expectations. Market indicators also have their limits, since observed prices incorporate risk and liquidity premia and may carry a seasonality bias, which can be problematic for extracting information on inflation expectations. The best way to obtain an unbiased measure of inflation is therefore to have as much indicators as possible, taking into account the limitations of each of them.

It is it this perspective that this study is inscribed. Our goal is to provide an alternative measure of inflation perception, by exploiting information made available by social networks. The current boom in social networks indeed provides a new source of information to find out how individuals feel about price trends. Twitter in particular allows relatively open access to its massive data, with millions of tweets being published each month, just in French. In order to keep the amount of data collected and analyzed within a reasonable range, this study focuses on a specific subset of users, namely the retweeters of the Banque de France Tweeter account. This community has been the subject of a previous internal study of Banque

[1] https://www.insee.fr/fr/metadonnees/definition/c1473

de France (Kintzler, 2018), which highlights among others the following characteristics: most retweeters are French, located in Paris and work in the banking, finance or economic sector. They seem therefore to be well informed about price evolution than the general population, and make them a population of interest to survey. Some bias may of course arise in comparison with data obtained from the general population, but we hypothesize that the opinion of informed experts on the topic are of higher value to estimate the general perception of inflation.

Recent studies have explored the contribution of indicators that measure sentiment perception about macroeconomic issues based on social media or newspaper data. For instance, Bertoli, Combes and Renault (2017) have calculated a media sentiment indicator for short-term employment forecasting. Thorsrud (2016) uses the content published by some Norwegian media to obtain a leading indicator of activity in Norway. Finally, Baker, Bloom and Davis (2016) create an economic policy uncertainty (EPU) index based on newspaper coverage frequency and demonstrate it proxies well for movements in policy-related economic uncertainty. Altig *et al.* (2020) then generalize it by successfully applying their methodology to Twitter, creating a high frequency uncertainty index. Despite this quite rich literature, very few studies focus on analyzing social media data regarding inflation problematics.

Angelico *et al.* (2021) from Banca d'Italia published one of the pioneering studies in the use of Twitter data for measuring inflation perception. The authors first collect all tweets in Italian related to prices, and then filtered appropriate ones using topic models to reduce noise. They then identify tweets related to increase of inflation and decrease of inflation using keywords, and finally combine create an indicator of perceived inflation.

Inspired by these results, we propose a novel approach than also combines the two approaches, namely keywords and natural language processing methods. Our final objective is to improve the performance of the final indicator, in particular by using modern state-of-the art supervised machine learning techniques rather than unsupervised topic models. Our methodology is conducted in three steps, focusing on the case of France. First, we collect all tweets published by all accounts that have retweeted a tweet from the Banque de France account. Second, we keep only those related to prices by classifying them using word2vec embeddings and random forests. Third, we classify each tweet in one of three categories: inflation, deflation, or other. Finally, we construct our indicator as the difference in the number of tweets related to inflation minus those related to deflation. We finally show that our indicator is highly correlated to more traditional survey-based metrics for the perception of inflation.

The remainder of the paper is structured as follows. We first describe the Twitter data used to create the indicator of price perception. Then, we will explain our methodology and its consecutive steps. Finally, we present the results and show their predictive power of the general perception of inflation.

2. Twitter data

In this study, we use Twitter data, which are increasingly used in economic news reporting. Twitter is a social network where users can publish short posts called tweets. Each day, millions of tweets are posted in French. From preliminary experiments, we even measured that thousands of tweets are posted each day that mention the simple keyword "prix" (which translates to both "price" and "prices" in French).

For the purpose of this study, in order to keep the amount of tweets reasonable, this study focuses on a specific subset of users, namely those who retweeted at least one post of the official Banque de France Twitter account. A previous internal research paper from Banque de France (Kintzler, 2018) collected this list of users and analyzed its population. It appears to be mostly French, located in Paris, and working in the banking, finance or economic sector, as almost half of them have keywords related to finance or economy in their account description. Appendix 7.4.3 presents the methodology used for this analysis. Even though these users are not representative of the general population, extracting signals from their tweets could be insightful because of their interest in economic matters. In future works, we will endeavor to generalize our methodology to a much larger and more representative subset of Twitter users, and if allowed to the whole population.

In order to not be limited by the restricted number of tweets imposed by the official Twitter API, we used the open-source tool Twint[2] to construct our database. Using our list of Banque de France retweeters, we collected for each of them the whole history of their public post, referred to as "timelines", with the only condition that the detected language of the tweet must be French. In total, the list of retweeters is composed of 3,548 accounts. We collected tweets from January 2008 to June 2021, although further filtering in applied on the following steps. In total, the number of all posts in all 3,548 timelines amounts to 10,382,847 tweets.

The information available at the tweet level is very rich: the text of the tweet, the date of creation, the geolocation, the language (in our case only French), the number of times it was shared or bookmarked, etc. A twitter account is also characterized by several variables: the date of creation of the profile, the short biography provided by the user, the number of accounts followed, the number of followers, and so on.

3. Methodology

The biggest challenge of this study relies on succeeding to detect the tweets related to prices among all possible topics discussed. Preliminary experiments showed us that using only is insufficient, because the targeted theme is not precise enough to avoid the problematic of linguistic polysemy of the French language. A very concrete example is the word "prix", which means both "price" (both singular and plural), and "award" or "prize". Therefore, using a list of keywords that is too large will lead to selecting too many tweets not related to inflation ("false positives"). On the other hand, a very restrictive list of keywords may lead to filter out many relevant posts ("false negatives"). Using machine learning, which is a much more flexible and precise tool, arises as a promising alternative. However, filtering the whole database of Tweets could be much too expensive in terms of computing power and time, since machine learning models rely on complex calculation, whereas keywords identification are extremely simple and fast.

To leverage both tools as efficiently and as precisely as possible, we use multi-level filtering, in which we combine keywords filters and machine learning filters to detect tweets relevant for the analysis. The methodology consists of four steps filtering and classifying tweets. First, a dictionary-based filter is applied to keep only tweets containing specific keywords related to the topic of prices. Second, a supervised Machine Learning model is trained on a random sample of 800 manually labelled tweets, and is used to the rest of the tweets in order to clean the residual noise of the first filter and retaining only tweets related to prices matter. Third, a keywords-based method is used to classify the direction of prices evolution mentioned in the tweets between "inflation", "deflation", or "other". Fourth, tweets mentioning foreign prices are excluded from the analysis since the study focuses on French prices.

The following sections detail each of these four steps, and present evaluations of the performance of each of the filter used when applicable.

[2] https://github.com/twintproject/twint

3.1. First step: retrieve tweets related to the lexical field of inflation

This selection process relies on a dictionary-based filter build on six types of lexical fields related to prices problematics. If a tweet contains one of the keywords belonging to one of these lexical fields, it is selected. Otherwise, it is removed from the database.

As collected tweets are in French, most of the keywords are in French. However, since the language specified by Tweeter is automatically detected, some tweets might be in English. Moreover, some users may also use English expressions within tweets mostly written in French. For both reasons, a small set of keywords in English has also been defined. Table 1 displays the six lexical fields defined. It can be noted that the filter has been intentionally set broad to avoid missing any relevant tweet.

Lexical Field	Keywords originally used (French)	Keywords translated in English
Lexical field of inflation with economical terms	Inflation, déflation, stagflation, désinflation, inflationniste, déflationniste, antiinflationniste, antidéflationniste, IPC, IPCH	Inflation, deflation, stagflation, disinflation, inflationary, deflationary, anti-inflationary, anti-deflationary, IPC, IPCH
Lexical Field of being expensive	Onéreux, cher, prohibitif, couteux, élevé, exorbitant, inabordable, conséquent, inaccessible, excessif, anormal, dispendieux, arnaque, arnaquer, ruineux, faramineux, hors de portée, rondelette, inconcevable, rédhibitoire	Expensive, expensive, prohibitive, costly, high, exorbitant, unaffordable, consequential, inaccessible, excessive, abnormal, expensive, rip-off, rip-off, ruinous, outrageous, out of reach, roundabout, inconceivable, prohibitive
Lexical field of being cheap	Faible, modique, avantageux, brader, imbattable, dérisoire, alléchant, réduit, occase, occasion, défiant toute concurrence, aubaine, modeste, clopinettes, bon prix, attrayant, clopinette, abordable, raisonnable, compétitif, accessible, acceptable, normaux, moyen, équitable, intéressant, convenable, négligeable.	Low, modest, advantageous, discounted, unbeatable, derisory, attractive, bargain, bargain price, attractive, bargain, affordable, reasonable, competitive, accessible, acceptable, normal, fair, interesting, suitable, negligible
Lexical field of prices and costs	Prix, tarif, montant, coût, loyer, vente, achat, location, frais, abonnement, facture, coûter, facturer, payer, tarifer, vendre, devis, paiement, rabais, tarifaire, croissance, promotion, remise, ristourne	Price, tariff, amount, cost, rent, sale, purchase, lease, fee, subscription, bill, cost, charge, pay, rate, sell, quote, payment, discount, tariff, growth, promotion, rebate, rebate
Lexical field of statistical institutions related to the inflation's measure	BCE, banque centrale, banque central, Banque de France, INSEE, FED, taux directeur, taux intérêt	ECB, central bank, central bank, Banque de France, INSEE, FED, key rate, interest rate

Additional keywords in English	Price, prices, cost, costs, rent, rents, bill, bills	

Table 1: List and content of lexical fields used to detect tweets related to prices matters

After this filtering procedure is applied, only 5% of the tweets remains, that is precisely 504,664 tweets. This lazy procedure needs not to be completed by a more demanding filtering, which will allow to remove the residual noise. Given the relatively small amount of remaining tweet, it is now possible to train a use a tailored supervised machine learning algorithm.

3.2. Second step: detect tweets related to prices

A supervised machine learning model is a model that provided numerical or categorical features of a tweet, will output a predicted label. It is called supervised because in the first place, it needs to be trained with a set of correct examples, *i.e.* of features associated with the correct label. In our case, this label is a binary value, which is set to 1 if the tweet is related to prices and 0 otherwise. In what follows, we explain the various components of our model: how the features of each tweet are constructed, what type of model is used and how it works, its performance, as well as how we manually labeled the training sample to train the model.

3.2.1. Explanatory variables: 200 word2vec variables

The goal of word embedding models, also called language models, is to create numerical vectors from textual data. Multiple word embedding models are available, and this study make use of the word2vec model (Mikolov *et al.*, 2013b), a probabilistic representation of words that take advantage of neural networks. In this model, the word is represented as a vector in the Euclidean space. Two words that appear frequently next to each other in the text will be close to each other in the Euclidean space. This proximity between the vectors can be measured using cosine similarity. In Appendix 7.1.3, we illustrate for instance what are the closest words from "price", "inflation" and "deflation" according to a word2vec model. Word2vec is a very popular and easy to use language model thanks to numerous implementations and an easy training that required no human labeling.

Many works have trained and used word2vec models, and consequently many pre-trained models are available online. It is relevant to use them when the data at hand is not sufficient to train correctly one's own word2vec algorithm. However, it can be interesting to train a word2vec model one's specific data when the database is large enough as in our case. In fact, most of the time, pre-trained models are trained on a huge amount of generic text data, such as Wikipedia articles or crawled websites. If the data at hand is specific to a topic or to a writing style, using such pre-trained models as is will result in poor performances. In this situation, training a new model better tailored to the data is often much better as it enables to capture the semantic specificities of the topic and style. In our study, the data at hand is specific because Twitter data is composed of short sentences about economics, hence with a specific style and vocabulary.

Therefore, we trained our own word2vec model on tweets. To make sure the model is trained on enough data, we use it on the full collected dataset of more than 10 million tweets.

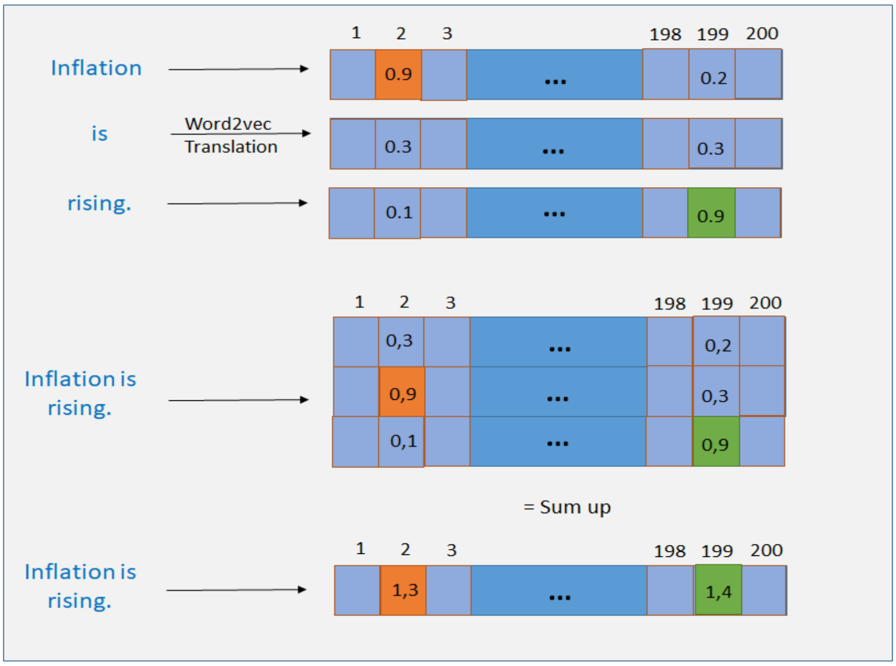


Figure 1 - Reading note: the diagram illustrates how the tweet "Inflation is rising" can be transformed into a vector with 200 coordinates. The three words "inflation", "is" and "rising" are transformed into a 200-coordinate vector. These coordinates characterize the word in question. For example, high values on coordinate 2 may indicate the presence of a price lexical field (orange color), and high values on coordinate 199 may indicate the presence of the increase lexical field (green color). The tweet can be represented as a matrix with 3 rows and 200 columns. To simplify this representation, we average over the rows: the tweet is simply characterized by 200 coordinates. Repeating this process for all tweets, each of them is characterized by 200 word2vec variables, representing as many lexical field signals more or less relevant for our analysis. (Values are given for explanatory purposes only).

In a word2vec representation, each word in the tweets is represented as a vector with K coordinates, as illustrated in Figure 1. The number of coordinates is a parameter of the model that needs to be chosen. This parameter is fixed by choosing the number of neurons that compose the hidden layer of the model. In our study, this number K was set to 200, as it is the case in most word2vec applications. Since the explanatory variable is to be computed at the tweet level, which are composed on several words, an aggregation method of each word's embedding must be applied. We chose to compute the average of the word2vec representations of each words in the tweet, along each dimension, resulting in a new 200 dimensions tweet embedding.

3.2.2. Explanatory variables: 38 additional dictionary-based variables

It is unclear, however, which of those 200 dimensions will (or will not) be related to the lexical field of price evolution. In order to add interpretability in the features used to characterize a tweet, we add 38 additional variables to each feature vector, which indicate the presence of 38 relevant lexical fields. For each tweet t, each indicator variable $X_{t,j}$ is set to 1 is the tweet contains a keyword belonging to the lexical field j, and 0 otherwise.

These lexical fields were built to detect a more precise notion among the topic of prices present in a tweet. They indicate the presence of characteristics in the tweet that can be grouped into five categories: the presence of words related to price topics or statistical institutions; the presence of words related to directional evolutions of price; the presence of words related to the lexical field of "cheap" or "expensive"; the presence of degree adverbs and adjective or negation terms; and the presence of words to be excluded (false friends of keywords that belong to the lexical field of prices).

The majority of the variables rely on French keywords whereas seven variables rely on English ones. In further developments, we could create more variables based on the English vocabulary. The variables also distinguish between verb and noun in order to integrate grammatical features into the model. In Appendix 7.5, we detail further the keywords used for each variable.

3.2.3. The labelling process

To train the model, we manually labelled a random sample of 800 tweets. As a recall, label 1 was assigned if the tweet is related to prices and 0 otherwise. This quite straightforward task required no specific knowledge, but if was however realized by trained economists to ensure consistency. Table 3 shows a typical example of the obtained labeled dataset. In a second step, which will be explained further, we tagged each tweet identified as "about price" according to one of the five following categories depending on its precise content: "deflation", "inflation", "disinflation", "price stability" or "unspecified". The "unspecified" label is important because most of the tweets mentioning prices actually do not mention any price evolution or perception of such evolution. For more information about the labelling process, please refer to Appendix 7.3.

Original text (French)	Translated text (for paper purpose)	Label
L'OPEP veut voir les prix du pétrole revenir à un niveau « raisonnable »	OPEC wants to see oil prices return to a "reasonable" level	1
Prix d'excellence Alassane Ouattara 2018	Alassane Ouattara Excellence Award 2018	0

Table 2: Typical result of the labelling process (without categories)

3.2.4. The model: random forest

Using the resulting 238 obtained features computed for each tweet, we train a random forest model to detect tweets whether a tweet is related to prices matter. This type of model architecture is well tailored for classification and is particularly relevant with high dimensional data. Its specificity lies in combining a multitude of decision trees, and outputs the class obtained from a majority vote of all decision trees. This methodology has also the advantage to produce little overfitting. Appendix 7.2 offers additional information about random forests and its functioning.

To optimize a random forest, two key parameters need to be tuned: the number of decision trees generated to include in the random forest, also called *ntree*, and the number of explanatory variables that will be considered at each tree split, also called *mtry* (see Kern, 2019). In general, the number of trees is set to 500 and the number of variable to consider at each split at the square root of the number of explanatory variables used in the random forest. However, these values are only starting points and they need to be calibrated in function of the data at hand and the problem we need to model. A large set of values for these parameters – between 50 and 2500 for *ntree* and between 0 and 30 for *mtry* – have been tested using cross-validation.

Figure 2 reports the result of this hyper-parameters optimization. After a certain point, increasing *ntree* or *mtry* does not improve significantly the performance, according to both accuracy and kappa metrics. In the case of *ntree*, a value greater of equal to 500 seems to stabilize the performance, but increasing this value also increase the computation time of the algorithm. After analyzing these results, we decide to set *mtry* to 30 and *ntree* to 500. Appendix 7.6.2 offers more information about these evaluation metrics.

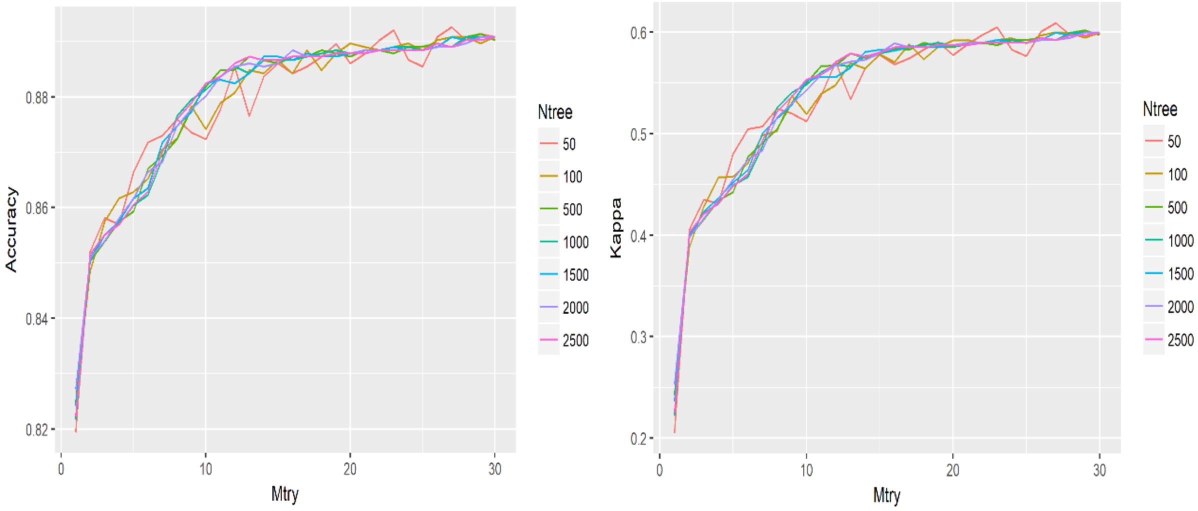


Figure 2: Optimizing the random forest with ntree and mtry

3.2.5. Setting the probability threshold

The raw output of such random forest is a predicted probability that a tweet is related to prices. To classify a tweet based on this probability, it is necessary to set a threshold that will act as a frontier: tweets with a probability above this threshold will be considered as talking about prices, otherwise they are considered as off-topic. To determine the optimal threshold, we compute the false positive rate and the false negative rate for different threshold value. A higher threshold value has two opposite effects on these rates (Figure 3). First, the false positive rate decreases, since the probability to belong to the "on topic" class predicted by the model has to be higher and higher so that only the "most certain" ones remain; and second, the false negative rate increases, since the model is more selective and misses more tweets related to prices matters.

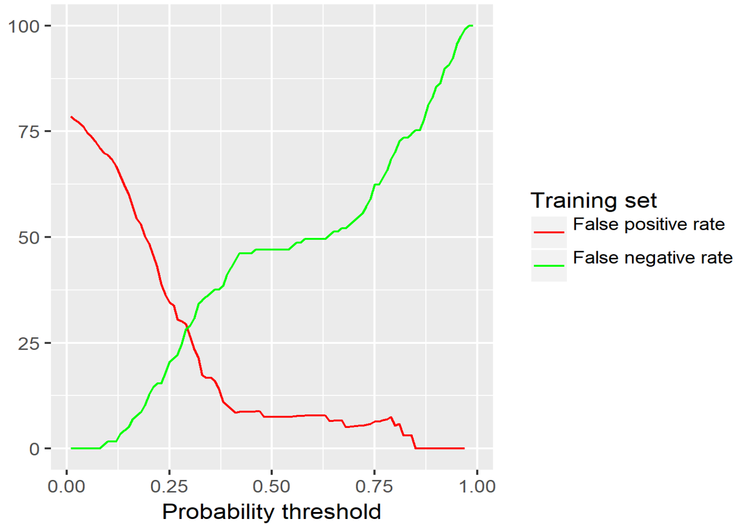


Figure 3: Setting the probability threshold

We chose to select the threshold in such a way as to obtain as many false positives as false negatives. For our study, these two types of error have a priori the same importance and are both to be minimized. We do not want to decrease one at the risk of increasing the other. The probability threshold is therefore set to 0.3, since is exactly the point where the false positive rate equals to the false negative rate in our training data as seen in Figure 3. In other words, it means that the model classifies all the tweets with a probability greater than 30% in the category "related to price matters", and to the "off-topic" category otherwise.

3.2.6. Variable importance

One key advantage of random forests is that is makes it possible to identify which features contribute the most to the predictions by computing the variable importance as the average decrease in Gini metric. Each node that composes a decision tree in the random forest is a condition based on a single predictor that split the data in two datasets (for instance, split observations into those having feature $k \geq 0.5$, and those having $k < 0.5$). The "Gini impurity" is the most often used metric to get this optimal (local) condition. It is possible to calculate afterwards by how much each variable decreases the average "impurity" of the tree in total. In the case of a random forest, the importance of each predictor is obtained by averaging the impurity in each tree of the forest.

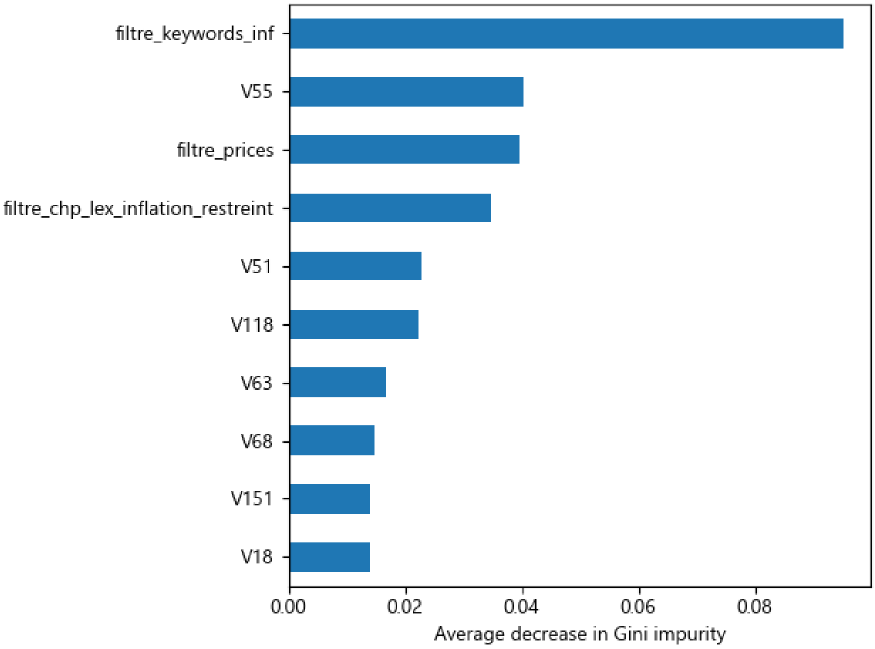


Figure 4: Gini feature importance

Figure 4 displays the obtained feature importance analysis four our model. Both word2vec and dictionary-based features seem to contribute significantly to predictions. Using these two kind of variables was justified as it provides relevant information. However, these results should be interpreted cautiously: other variables may also be important and still not appear on the figure because of multicolinearity between variables.

3.2.7. Performance evaluation

The performance of the random forest can be assessed using the area under the ROC curve (AUC) metric on train (560 tweets, 70% of the labeled data) and test set (240 tweets, 30% of the data). An AUC metric much higher on the test set than on the train test could be a sign of overfitting. As displayed on Figure 5, our model shows little overfitting as the ROC curve for train and test sets are similar. Overall, the model performs very well: the values for the AUC metrics are very high for the train and test set.

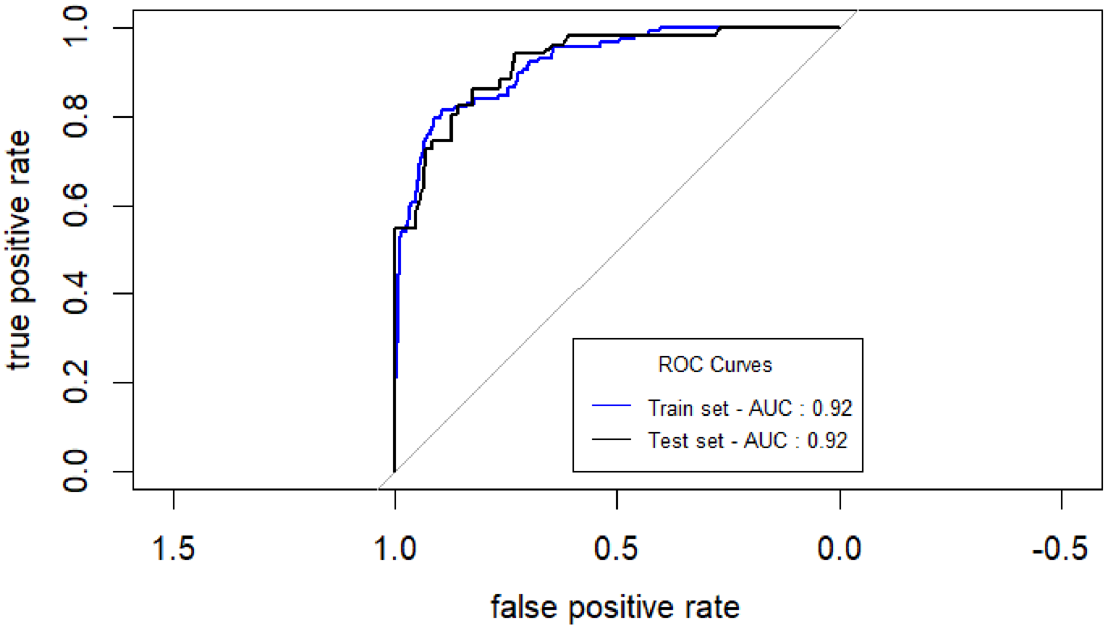


Figure 5: Comparing ROC curves between train and test sets

Other evaluation metrics, displayed in Table 5, also demonstrate performance of the model. We computed the metric both on the training set and on the test set, using standard evaluation metrics. Please refer to Appendix 7.6.2 for more information about their definitions. Obtained values are close to each other, which indicates little overfitting. To assess the quality of the model, it is necessary to consider and compare multiple metrics because the response variable is significantly imbalanced. This imbalanced data increases artificially the AUC or the accuracy. We rather use precision, recall and F-score, which are better tailored to handle such situation. These two indicators are indeed slightly lower but still high enough to warrant a good performance of the model.

Metrics	**Training sample**	**Test sample**
Accuracy	0.912	0.879
F1-score	0.819	0.739
Precision	0.817	0.732
Recall	0.822	0.746

Table 3: Evaluation metrics of the model

3.3. Third step: identify the direction of price mentioned

In previous steps, we identified tweets related to prices in general using keywords and machine learning methods. In this last step, our goal is to go one-step further and categorize what opinion about price evolution each tweet conveys. We define four categories (or topics) of interest in which each tweet could fall: inflation, deflation, disinflation and prices stability. However, a preliminary analysis shows that even among tweets related to prices, almost half of them do not match any type. We therefore add an additional "Other" category, in which fall all tweets not mentioning any of the four topics of interest. Table 3 displays an example of tweet (translated in English) for each type of content.

Tweet text (translated for paper purpose)	Category
Fukushima: electricity prices increase lead to 10 times more deaths than the accident itself.	Inflation
Another one bites the dust… #retailapocalypse #diesel #realestate #lifestyle #disruption #deflation	Deflation
Euro Area inflation expectations keep dropping	Disinflation
So one of the euro's concrete objectives, to control inflation, has been achieved.	Stability
Are you broke? Your banker is rubbing his hands. Cost per client: 58.91 euros.	Other (out of topic)

Table 4: Illustration of the classification of the tweets regarding prices evolution

As mentioned earlier, we annotated a sample of 800 tweets, both with a binary label to identify whether they are related to prices, but also according to one of our five topics if they were. However, our initial ambition to train another supervised machine learning model to automatically recognize the topic mentioned cannot be achieved with our annotated database. In fact, with only 168 tweets being tagged as related to prices, and then about a few dozen of them falling into the "Other" category, there remains less than a dozen example for each category of interest. Labelling more tweets to get significantly more training data would be necessary to train such a classifier. Further development will undertake this task.

For this reason, we chose to use a keywords-based method to identify the direction of prices mentioned in the tweets. Appendix 7.4.2 presents the combination of lexical fields used, which focused on the hard task to disentangle the thin differences of some of our four topics of interest. We make the hypothesis that, at this point, using keywords related to the increase, decrease, and stability's lexical fields could be enough to identify whether a tweet is talking about inflation, deflation, disinflation or prices stability, as most of the noise has already been ruled out. Further development will focus on improving the reliability of this last task.

3.4. Last step: excluding tweets mentioning foreign prices

Despite our focus on French perception of inflation and thus tweets about French prices, a few tweets explicitly concerning prices of foreign countries remain in our database. In the labelled dataset, only 3% (24 tweets) of the tweets were concerned, yet we decided to exclude such tweets from our general database using simple rules. First, we remove tweets that mention a country name, in French or English, other than France. Second, we remove tweets that mention a nationality name, also in French or English, other than "French". Note that we excluded tweets that mention global or European prices, as we attempt to keep a certain purity in the index despite the probable correlation between European and French prices evolution.

3.5. Review of the quality of this four-step methodology

It is possible to estimate the overall quality and the performance of our methodology. The aim is to estimate the quality of the detection of the tweets falling in our four categories of interest, obtained by a two-step classification with machine learning followed by a keywords-based method. We compute the metrics using our database of 800 labeled tweets despite the little counts in each category. Table 5 displays those metrics.

Category	Precision	Recall	F-score
Inflation	0.64	0.56	0.60
Deflation	0.78	0.97	0.86
Disinflation	0.00	0.00	0.00
Stabilization	0.56	0.71	0.63
Other	0.78	0.86	0.82

Table 5: Overall model evaluation

Results can be read as followed, in the case of the inflation topic. Among tweets returned as "about inflation", 64% are indeed about inflation (recall) and 36% are false positive. On the other hand, when considering all tweets truly about inflation, 56% are identified as such by our methodology, and hence 44% mentioned are missed and not classified as "about inflation" (recall). All results read the same for each category, with the notable case of disinflation where all scores as set to zero. This can be explained by the fact that only 16 examples are labeled as such in our training dataset, and therefore it is hard to obtain statistically significant results.

Overall, metrics show that our methodology has satisfying success when it comes to identify tweets according to their topical content; yet the score greatly depends on the category. The methodology for instance succeeds in detecting tweets related to deflation and those falling in the out-of-topic category, but encounters more difficulties to detect tweets about inflation, disinflation or prices stability. This highlights the limit of using just keywords in the last step of our methodology. Once again, this highlights the necessity to improve this last step in future works.

4. Results

4.1. Twitter indicator of perceived inflation

The methodology detailed in the previous section enables us to identify tweets mentioning the problematic of price, and then to detect tweets related to inflation, deflation, and other minor topics. The indicator we propose builds on this methodology with a simple computation. We simply count, for each given period (day, week, or month) the number of tweets mentioning inflation, and the number of tweets mentioning deflation. Figure 6 displays both those counts at a monthly scale. Inspired by the work of Angelico *et al.* (2021), the indicator if the number of tweets about inflation minus the number of tweets about deflation. Note that we do not take into account tweets mentioning disinflation or price stability, mainly because the quality of the detection for those categories lack precision. We also compute a smoothed version of the indicator, using a backward-looking exponential weighted moving average with parameter α set to 0.35.

It can be noticed that from the huge volume of collected data, with more than 10 million tweets, only a minority pass all filters and end up will contributing to the construction of the Twitter indicator. Each month, about 13,000 tweets are published on average by Banque de France retweeters, but only 58 relate to inflation and 27 to deflation.

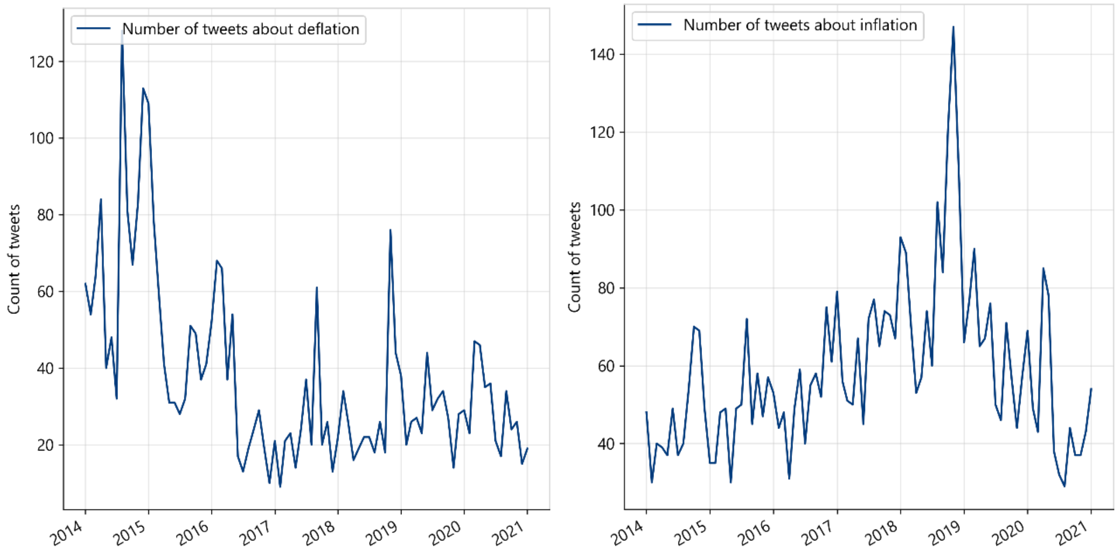


Figure 6: Comparing the number of tweets mentioning deflation or inflation over time

The Twitter indicator and its smoothed transformation appear on Figure 7. The indicator is negative just at the beginning of the period, due to the spike of tweets talking about deflation at the beginning of the year of 2015 as seen in Figure 6(a). To assess further the quality of the index, it is also necessary to compare it with other measures of perceived inflation, as well as the true inflation rate.

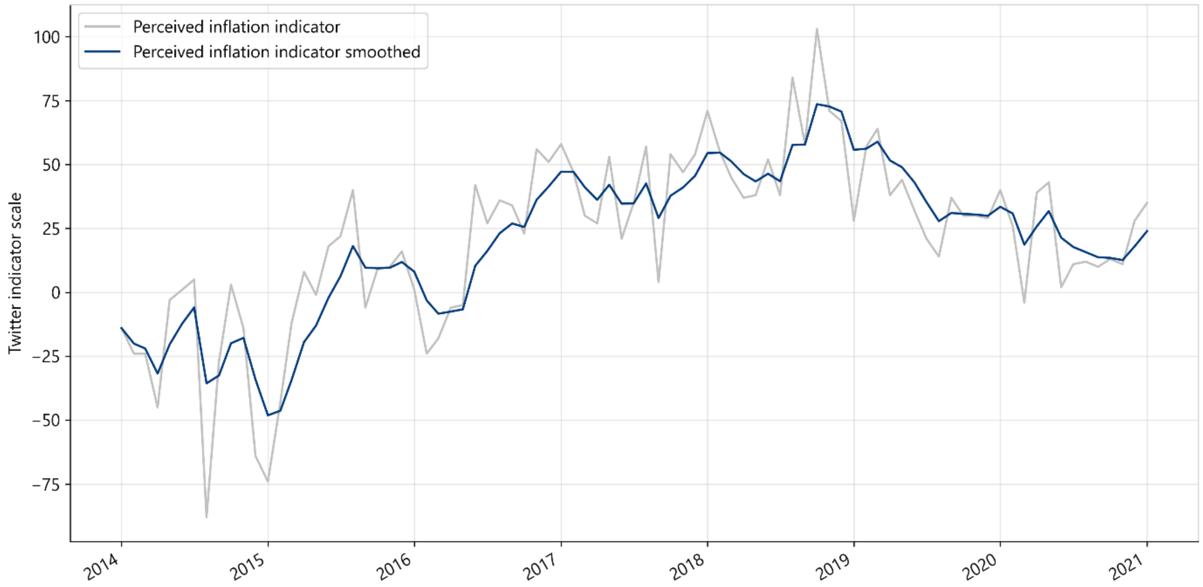


Figure 7: Twitter indicator

4.2. Twitter indicator consistency with households surveys

The French National Institute for Statistics (INSEE) produces its household survey each month, whereby it questions a sample of French households about their opinion on their economic environment.[3] Among other subjects, respondents are asked about their perceptions of inflation and their anticipations on its evolution. Respondents are asked to choose between "increase" and "decrease" on the two following questions: "how do you think prices have evolved over the past 12 months?" and:

[3] Data are available at: https://www.insee.fr/fr/statistiques/series/102414547?INDICATEUR=2874666%2B2874667

“how do you think prices will evolve over the next 12 months?”. INSEE constructs two indicators from those answers: a past evolution indicator, which is the percentage of households that think prices have increased over the past 12 months, and a future evolution indicator, as the percentage of households that think prices will increase in the next 12 months. Both can be used as reference indicators as the overall French population’s perception of inflation.

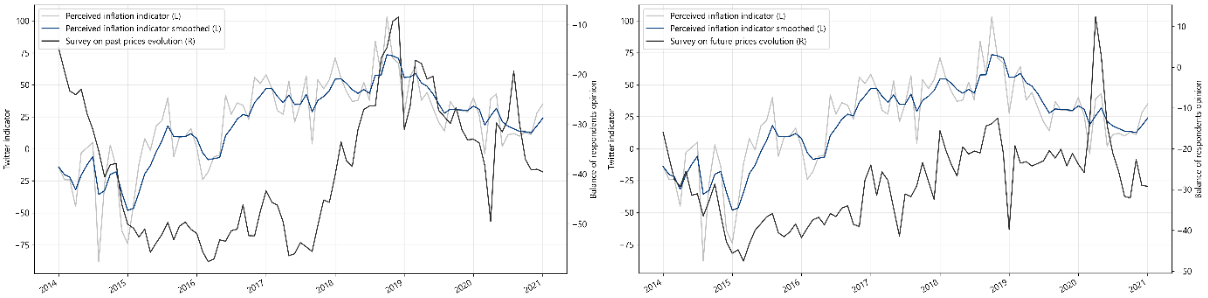


Figure 8: Households’ perceived evolution of past prices (left) and future prices (right) compared to Twitter indicator

Our findings show that the Twitter indicator is correlated with both INSEE indicators (past and future), using both the Pearson and Spearman correlation coefficients, as shown in Table 4 (see Appendix 7.6.1 for more details about both metrics). The correlation is notably stronger when the Twitter indicator has been smoothed. In Figure 8, we display both INSEE indicators along with our Twitter index. We see that in both cases, curves seem to follow the same trends, but also peak at the same time. Our index therefore seems consistent with those reference indicators. An interrogation arises on whether the Twitter indicator measures anticipation of future inflation, or the perception of past inflation. Since correlation coefficients are always higher with the past evolution INSEE indicator than with the future evolution one, we can hypothesize that it is rather related to anticipation of future inflation.

Correlation between…	Pearson	Spearman
Past evolution perception and Twitter indicator	0.213	0.219
Past evolution perception and smoothed Twitter indicator	0.305	0.329
Future evolution perception and Twitter indicator	0.460	0.487
Future evolution perception and smoothed Twitter indicator	0.505	0.558

Table 7: Correlation summary Table of the INSEE statistics and the Twitter indicator

4.3. Twitter indicator consistency with the inflation rate

INSEE also publishes the Consumer Price Index (CPI), which is the base for the calculation of the inflation rate as the evolution of the variation over month of the CPI.[4] Table 8 shows that the Twitter indicator is strongly correlated to the inflation rate, and Figure 9 displays the two series together. Pearson and Spearman correlation metrics are indeed high and always between 0.6 and 0.8. The correlation is once again stronger when the Twitter indicator is smoothed.

Correlation between…	Pearson	Spearman
Inflation rate and Twitter indicator	0.642	0.664

[4] Data are available at: https://www.insee.fr/fr/statistiques/serie/001763852

Inflation rate and smoothed Twitter indicator	0.735	0.792

Table 8: Correlation summary Table of the inflation rate and the Twitter indicator

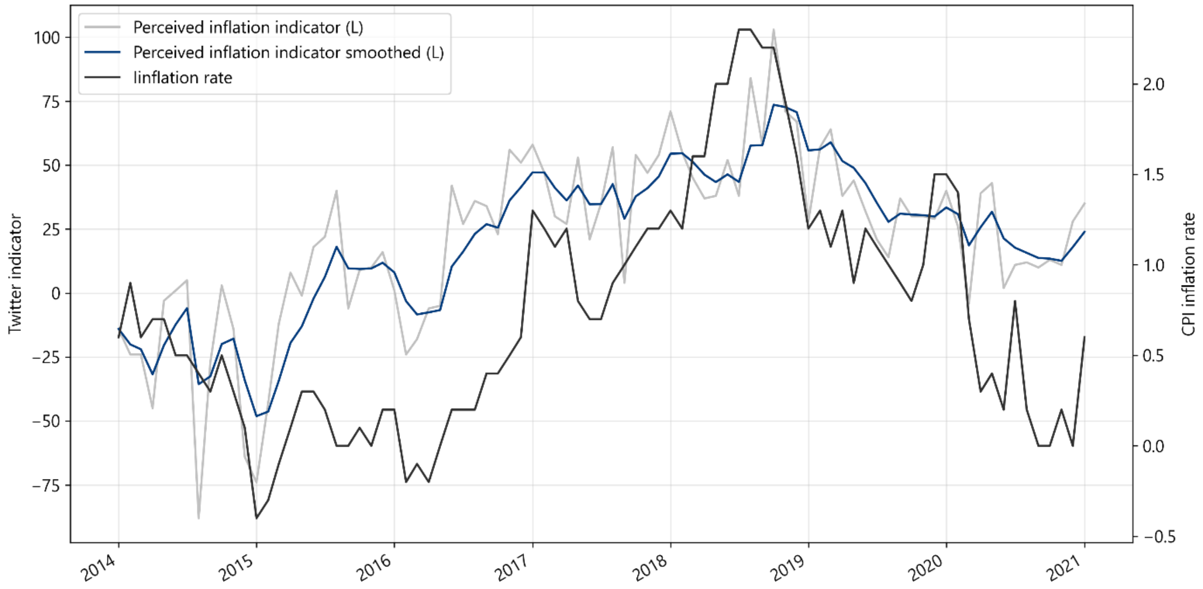


Figure 9: Twitter indicator and true inflation measured with CPI change

One interesting question if also to investigate whether our indicator can be anticipatory of the future inflation rate, or rather that it reflects past inflation rate. To do so, we analyze the time-lagged correlations between the inflation rate and the Twitter indicator. Figure 10 displays the Pearson and Spearman correlations' variations when the Twitter indicator is compared to the inflation rate in a range from 12 months earlier to 12 months later. Here is how to read the leftmost point: the Pearson correlation between the Twitter indicator at month *m* and the inflation rate at month *m*-12 (a year before) is about 0.195. The rightmost point reads similarly: the Spearman correlation between the Twitter indicator at month m and the inflation rate at month *m*+12 (a year later) is about 0.580.

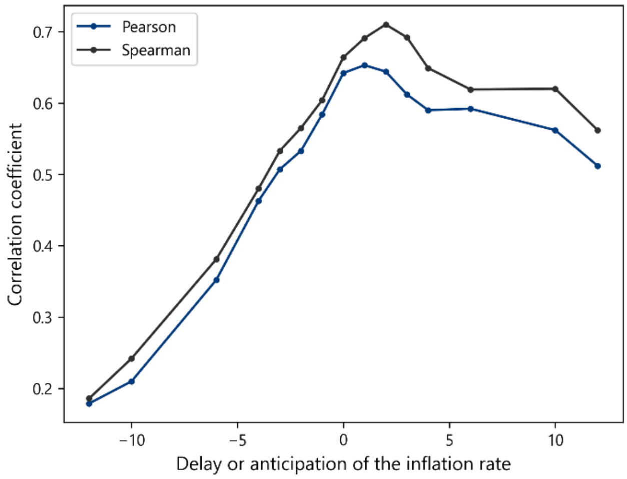


Figure 10: Time-lagged correlations between the smoothed Twitter indicator and the inflation rate

A clear conclusion appears from this plot: the Twitter indicator is much more correlated to the future inflation rate than to the past inflation rate. Interestingly, inflation rate 12 months in the past has a low correlation coefficient of 0.195, but inflation rate 12 months in the future keeps a high correlation coefficient, between 0.5 and 0.6. More interestingly, the indicator is more highly correlated to the inflation rates of month $m+1$ and month $m+2$ than it is with the one of month m. This clearly indicates that our Twitter indicator is a forward-looking measure, which provides indications about *perceptions of future of inflation* (i.e. anticipations) rather than *perceptions of past inflation*.

4.4. Twitter indicator and Covid-19

A robustness check of our indicator can be conducted by studying its behavior during the beginning of the Covid-19 economic crisis. Interestingly, the indicator is quite stable during this period. In fact, 30% more tweets have been posted during the French lockdown period (March to May 2020) than before March 2020, and a significant amount of them concerns the pandemic: 16% of the tweets posted between March and May 2020 are related to the Covid-19 disease[5]. It also had an effect on the number of tweets concerning inflation or deflation matters, which increased by 28% during the first French lockdown compared to the previous period.

Period	Number of tweets (monthly)	Number of tweets about Covid-19	Number of tweets about deflation or inflation
Before the lockdown of March 2020	13 450	80	92
During the lockdown (March to May 2020)	17 296	2 781	125

Table 6: Comparing tweets volume before and during the French lockdown of March 2020

Interestingly, it seems from Figure 6 that tweets related to inflation and those related to deflation counterbalance each other, resulting in a stability of the Twitter indicator during that period (see Figure 6). This behavior is coherent with the behavior of the true inflation rate, which has increased in some sectors (food prices for instance), and decreased in other sectors (e.g. energy), resulting in a global inflation rate remained globally stable during this period (see INSEE, 2020).

4.5. How do people talk about prices?

Word clouds enable us to visualize words that appear most often in the tweets and thus contribute the most to the Twitter indicator. Then, particular topics can be highlighted. Figure 11 displays the word cloud on the subset of the tweets related to either deflation or inflation topics.

Naturally, words like "prices", "price", "inflation", "costs", "bill", or "prix" (French word for "price" and "prices") are the most represented. They belong to the set of keywords used in the first place to filter tweets. What has more value is to analyze words that do not belong to the lexical field of price evolution. Some sectors come up more often than others do. In particular, the oil ("oil" or the French word "pétrole"), the housing ("housing" or the French word "logement", "rents" or the French word "loyer", "immobilier" which means "real estate" in French), and the energy ("energy") sectors. It could be interesting to isolate these tweets to build an indicator measuring prices on these particular sectors. The area of Paris is also more mentioned that others. Some major events that are likely to impact prices such as "Brexit" or "Covid-19" pandemic are also present in the word cloud.

[5] Keywords used are to detect them are "covid", "coronavirus", "pandemic", and "epidemic", both in English and French.

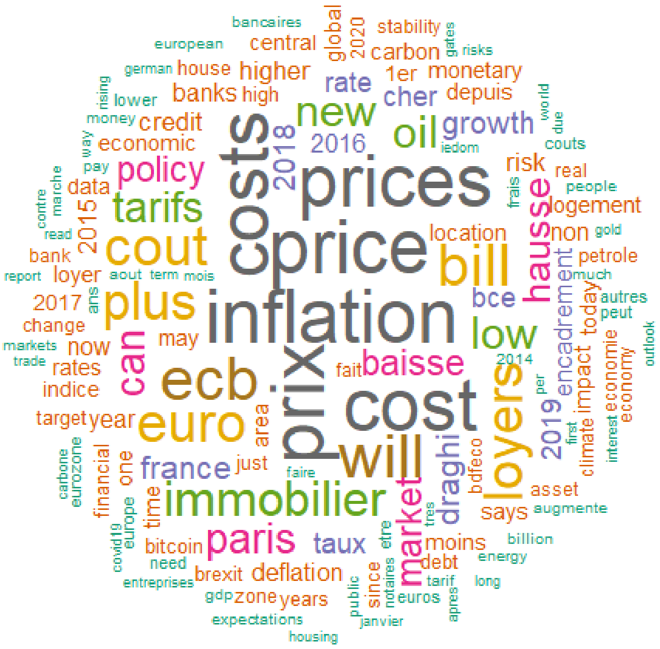


Figure 11: Word cloud of the tweets behind the Twitter indicator

5. Conclusion and future works

Finding the tools to classify and evaluate the methodology was one of the main challenges of the study. Combining machine learning methods and keywords filtering was a good way to fulfill those tasks and provided good results. The performance metrics of the classifier are high and the resulting Twitter indicator seems to provide useful information. The indicator is indeed consistent with the monthly household surveys on inflation expectations and is highly correlated with the inflation rate. Overall, our study demonstrates that it is both possible and relevant to use Twitter data to measure inflation perceptions. The results provide reassurance about the bias in the data, restricted to the Banque de France retweeters. Measuring the perception of specific users proved to be an appropriate approach. In addition to enable profile characterization, restricting the study to expert profiles gives relevant result because their perception seems to be more accurate of the reality by being based on scientific grounds.

Of course, there are still open questions that remain and motivate further developments. First and above all, the data is restricted to a small panel of users who post rarely about prices matters despite their expert profile. It would be most useful to expand the initial data at hand by extending the number of expert profiles analyzed. Three main approaches for instance be explored: following a social community approach by integrating the contacts of users used in this version of the indicator, including the retweeters of the European Central Bank, or even use the list of predetermined economists who have a Twitter account. Of course, the most complete approach would be to include all tweets mentioning prices in French, irrespective of the users, as to obtain a fuller picture. This raises of course the question of the volume of the data that would be collected, and would require computing tools accordingly.

The quality of the various classification and filtering steps could be improved by extending the labelled database, which would exempt us from using keywords to categorize the content of the tweets regarding prices evolution. This would make it possible to train a supervised machine learning model to classify the content directly and not just detecting tweets related to prices matters as in this study.

Finally, no normalization is actually involved in the Twitter indicator. This was a choice driven by the idea that an increase of tweets about inflation was still an interesting signal to measure even though the total number of posts also increased. However, testing several types of normalizations to capture additional information would be relevant. For instance, one could think of normalizing by the sum of number of tweets talking about inflation, deflation or off-topic, as this would have the merit of taking into account "off-topic" (or rather, "neutral") tweets. Another idea could be to normalize by the total number of number of tweets that mention economic topics, not just inflation. This normalization would make it possible to measure the relative importance of inflation within economic topics, but would require properly defining what an economic topic is.

Finally, a crucial flaw in this work is that we do not yet make the distinction between tweets expressing a personal opinion about inflation and those that merely reflect official announcements of statistical institutions about inflation. This "sounding board" effect is interesting to study, particularly with regard to the impact of institutional communication, but it may also introduce biases to the measurement. It should therefore be more clearly investigated and the two effects should be separated.

6. References

Altig D., Baker S., Barrero J. M. , Bloom N., Bunn P., Chen S., Davis S. J., Leather J., Meyer B., Mihaylov E., Mizen P., Parker N., Renault T., Smietanka P. and Thwaites G. (2020). "Economic Uncertainty Before and During the COVID-19 Pandemic". *Journal of Public Economics* 191.

Angelico C., Marcucci J., Miccoli M., and Quarta F. (2021). "Can we Measure Inflation Expectations Using Twitter?". *Banca d'Italia Temi di discussione* n°1318.

Baker S. R., Bloom N. and Davis S. J. (2016). "Measuring economic policy uncertainty". *The Quarterly Journal of Economics*, 131(4).

Bec F. and Mogliani M. (2013). "Nowcasting French GDP in Real-Time from Survey Opinions: Information or Forecast Combinations?". *Banque de France Working Paper Series* n°436.

Bertoli C., Combes S., Renault T. (2017). "Comment prévoir l'emploi en lisant le journal". *Note de conjoncture de l'INSEE* of 03/2017.

Kern C. (2019). "Tree-based Machine Learning Methods for Survey Research". *Survey Research Methods* 13(1).

Kintzler E. (2018). "La Banque de France sur Twitter : impact des publications et réseaux d'influence". *Banque de France Internal Research Document* n°18-025.

Mikolov T., Chen K., Corrado G., and Dean J. (2013a). "Efficient Estimation of Word Representations in Vector Space". *Proceedings of Workshop at ICLR 2013*.

Mikolov T., Sutskever I., Chen K., Corrado G., and Dean J. (2013b). "Distributed Representations of Words and Phrases and their Compositionality". *Proceedings of NIPS 2013*.

Thorsrud L. A. (2016). "Nowcasting using news topics, Big Data versus big bank", *Norges Bank Working Paper* n° 20/2016.

7. Appendix

7.1. Word2vec model

7.1.1. How it works

The Word2vec model is a word embedding method based on a probabilistic representation of words and on neural networks. It has made it possible to rethink the concept of word embeddings by representing words in a vector space where words used in similar contexts are represented close to each other. The Word2vec representation of a word depends indeed on its "context", it means it depends on the words surrounding the term considered in the sentences of interest. The similarity between two word-vectors can be measured with the cosine similarity metric, a little further.

The word2vec model relies on a two-layer neural network. Two types of neural architectures can be used. In the Continuous Bag of Words (CBOW) architecture, the neural network tries to predict a word according to its context. In the Skip-Gram architecture, the neural network tries to predict the context according to the given word. In both cases, the neural network takes unstructured text as input, and modifies its neural weights using unsupervised learning to reduce the prediction error of the algorithm. It is possible to fix the number of word-vector coordinates obtained by the model by choosing the number of neurons number of the hidden layer.

The word2vec model has multiple advantages. For its training, word2vec only needs raw text data that do not require to be labelled. Therefore, a large corpus of unstructured set is enough to estimate a model with good performance. Finally, the algorithm is efficient and can be run on a huge volume of data in a minimum of time. This is mainly due to its simple neural network structure.

7.1.2. Optimization of hyper parameters

Many hyper-parameters can be tuned to improve the model performance. We highlight three of them. The *dimension of the vector space* it is the number of numerical predictors used to describe the words (between 100 and 1000 in general), in other words the number of coordinates characterizing the vector representation of a word. The *architecture of the neural network* is a second parameter. It must be chosen between Continuous Bag of Words (CBOW) and Skip-Gram. Finally, the *size of the context* is a last crucial parameter. It refers to the number of terms surrounding the word in the sentence. According to the creators of word2vec, it is recommended to use contexts of size 10 with the Skip-Gram architecture and 5 with the CBOW architecture (see Mikolov *et al.*, 2013a).

7.1.3. Application in this study

As explained on the paper, many pre-trained word2vec models are freely available online. They is even more useful when the volume of data at hand is not sufficient to train correctly the Word2vec model. Most of the time, pre-trained models rely on a huge volume of generic data. Therefore, the resulting word-vector representations are also generic. In the case of specific data, training the model on this data can be judicious as it allows capturing semantic relations specific to the field of study. With its short sentences, Twitter data is atypical and fits in this kind of use case.

Training a Word2vec model on one's own data can be done under Python with the *Gensim* package or under R with the *wordVectors* package. In our study, we trained a word2vec model on tweets using Python and thus Gensim. To train it on enough data, all tweets collected were used for the training, before any filtering was applied. In the end, the Word2vec model has been trained on more than 200 million French and English words. The hyper-parameters chosen are listed in Table 7.

Parameter	Value
Dimension of the vector space	200
Context size window	5
Number of times to process the entire corpus for training	100
Type of neural architecture	CBOW
Minimum times a word must appear in all tweets to be included in the training process	5

Table 7: Hyper-parameters for the word2vec model of the study

After training, to check if the Word2vec representation is coherent, it is possible to look at the 10 words that are the closest to terms specific for our analysis. In Table 8, we do so for words “prix” (“prices” and “price” in English), “inflation”, and “deflation”. We provide words in French along with their English translation when needed.

Rank	“Prix”	“Inflation”	“Deflation”
1	croissance (*growth*)	fed	deflationniste (*deflationist*)
2	achat (*purchase*)	recession	inflation
3	tarif (*rate*)	infl	recession
4	cout (*cost*)	insee	fed
5	moyen (*means*, probably for *means of payment*)	hicp	easing (probably for *quantitative easing*)
6	loyer (*rent*)	eurozone	spiral
7	cher (*expensive*)	contraction	bulle (*bubble*)
8	frais (*fees*)	ipc *(cpi)*	qe (for *quantitative easing*)
9	occasion (*second-hand*)	evaporation	deflationnaire (*deflationary*)
10	vendre (*sell*)	draghi	eclatement (*bursting*)

Table 8: Closest words to relevant terms for the study

This example demonstrates that the word2vec representation has indeed captured the context of terms that usually go along with a word: the closest words belong indeed to the lexical field of inflation, prices and deflation context of words.

7.2. Random forest

The random forest model is appropriate for categorical (so-called "classification" problem) or numerical outcome variables (so-called "regression" problem). Input predictors (or features) can be both categorical and numerical variables. This methodological appendix concerns the case where the response variable is categorical (and even binary, as in this study).

Random forests are the result of combining a multitude of decision trees from the CART algorithm. A decision tree creates multiple partitions of data using a set of rules to predict the class of each observation. Decision trees are intuitive and interpretable algorithms but they also tend to overfit data. Random forests solve this issue. The bagging approach is used to generate the trees. Sub-samples are generated using a random sampling without replacement. Then a CART-type algorithm is applied on each sub-sample to build a decision tree. Not all variables are input of the algorithm but a random sample without replacement. These samplings enable to build independent decision trees being trained with different observations and different variables. That is why random forests tend to present less overfitting and are better generalized on new data.

7.2.1. Hyper-parameters optimization

To optimize random forest, two key parameters have to be considered: the number of decision trees (*ntree*) and the number of explanatory variables that each decision node will take as input (*mtry*). In general, *ntree* is set at 500 and *mtry* is set at the square root of the number of predictors used in the random forest. However, these values are only starting points and they need to be calibrated in function of the data at hand and the problem we need to model. In the case of *ntree*, increasing this value also increase the computation time of the algorithm. That is why we rather keep this value not too high.

7.2.2. Random forests output

A random forest produces for each observation a prediction. For a classification problem, each tree computed by the random forest gives a prediction - in our study, it would be "the tweet is related to prices matters" or "the tweet is not related to prices matters". The final prediction is computed as the most frequent prediction among all predictions produced by all trees.

A random forest can also produce the probability for an observation to belong to a class. This is what is used in our study. The computation of probabilities actually depends on the implementation of the random forest algorithm. In our study, we used the *scikit-learn* package on Python, which provides the proportion of decision trees classifying a tweet as related to prices matters.

7.2.3. Variables importance

Inspecting the importance of the variables is crucial to determine which predictors have the most contributed to the predictions. In this perspective, two measures are possible in the classification case. First, Mean Decrease Accuracy, which is constructed by swapping the values of a given predictor and look at the impact produced by calculating the decrease of accuracy. The more important predictors are, the more significantly the accuracy of the model decreases.

Second, Mean Decrease Gini. In a decision tree, each node that composes a decision tree is a condition based on a single predictor. To get an optimal (local) condition, the metric often used is the "Gini impurity". When training a tree, it is possible to calculate the impact of each variable on the average "impurity" of the tree. At the scale of a random forest, it is possible to calculate the importance of each predictor by averaging the "impurity" obtained for each tree of the forest.

Inspecting the importance of the variables is crucial to understand the relationships between the explanatory variables and the response variable. By being an interpretable model, the random forest provides a descriptive analysis of the data. However, a prerequisite to a good analysis is the absence of collinearity between variables. This issue is not a real concern in a predictive approach but it prevents from analyzing causal links. Indeed, when two (or more) variables are strongly collinear, only one of the variables is more likely to capture all the importance. The importance of the other variables is somehow "hidden" by the one capturing all the importance. For this reason, we have not done a deep analysis of the variables importance in our study. The variables importance mostly gives information about the forest construction and does not really enable us to deduce any causal links. To do such causal analysis, a more in-depth study on collinearity links within the predictors would have been necessary. This could be a line of research for further developments

7.3. Description of the human labelling process

To train a machine learning model able to detect when tweets relate to prices matters, we randomly selected a sample of 800 tweets among the tweets containing at least one of the keywords related to the lexical field of prices. Then, we labelled this sample to train a supervised machine learning model to detect tweets related to prices. The manual annotation process consisted in creating four variables by reading the text of the tweet. Not all of them are used in the project, but may be useful in future developments.

Variable *"is_inf"*: asked the question: "is the tweet related to prices matters?". It is a binary variable, which is assigned to 1 if the answer is yes, 0 otherwise. This variable enables us to detect tweets related to our problematic. Despite having used a preliminary dictionary-based filter, only 21% of the tweets are found to concern prices in our training dataset.

Variable "*what_info*" answered to the following question, if a tweet was first labeled as being related to prices: "what is its content?". Possible answers were one of our five categories of interest described in the paper: "inflation", "disinflation", "deflation", "prices stability", or "other" (out of topic). The variable helps to describe the content of the tweet regarding the prices problematic.

Variable "*what_prices*" asked, "if the tweet is related to prices matters, what kind of prices are mentioned?". Two answers were anticipated either "global prices" or "prices concerning particular sectors". This variable allows determining if a significant part of the tweets concerns particular sectors. Almost half of the tweets classified as talking about prices matters concern indeed particular sectors.

Variable "*what_loc*" was finally concerned with the following question: "if the tweet is related to prices matters, what geographical localization is it referring to?". The goal was to filter tweets that explicitly mentioned something else than France, rather than having a precise location. Its possible values are "Global/French prices" or "prices concerning explicitly another country". Around 15% of the tweets related to prices matters concern explicitly prices from another country. These later tweets were removed from our database.

Finally, Figures 12 to 15 describe the counts of each modality for our four variables in our labelled database of 800 tweets.

7.4. Dictionary-based filters

In this study, we used several kind of filters relying on keywords. The tables below explain the lexical fields used and their corresponding keywords. The lexical fields and dictionary-based filters have been set with an expert perspective and by confronting multiple examples of tweets.

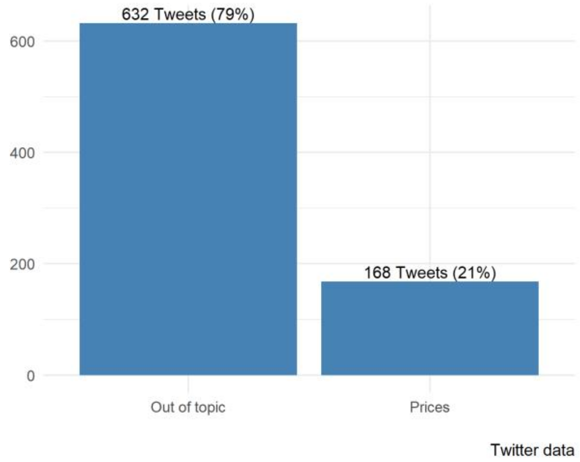


Figure 12: Description of variable is_inf

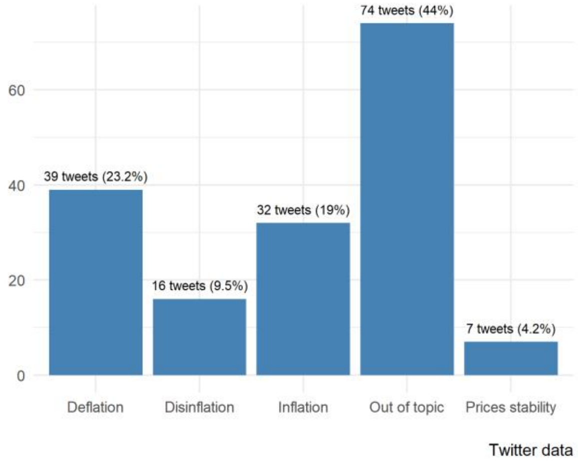


Figure 13: Description variable what_info

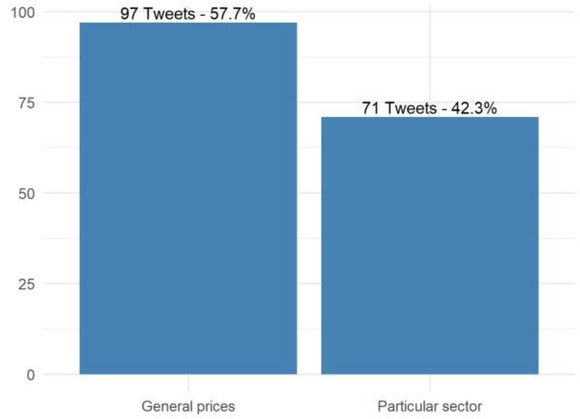


Figure 14: Description of variable what_prices

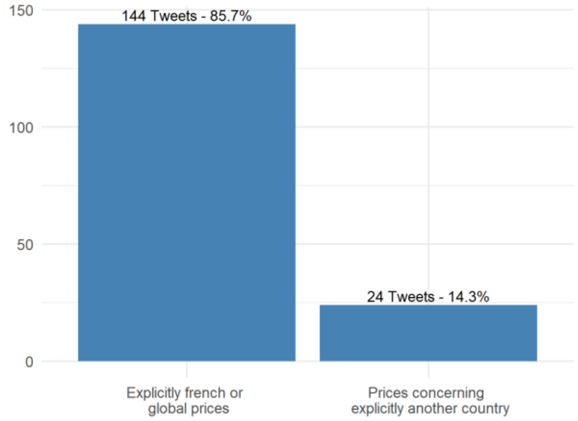


Figure 15: Description of variable what_loc

7.4.1. Detecting tweets related to prices problematics

To detect tweets related to prices matters, the study relies on a dictionary-based filter build on six types of lexical fields related to prices problematics. If a tweet contains one of the keywords belonging to one of these lexical fields, it is selected (first step of the methodology detailed in the main of this paper). The keywords are mainly in French, but some English words have also been included. The filter has been intentionally set broad to avoid missing any relevant tweet. However, it also captures a lot of noise that will be removed in the next steps of the methodology.

Lexical Field	Keywords
Lexical field of inflation with economical terms	inflation, déflation, stagflation, désinflation, inflationniste, déflationniste, antiinflationniste, antidéflationniste, ipc, ipch
Lexical Field of being expensive	onéreux, cher, prohibitif, couteux, élevé, exorbitant, inabordable, conséquent, inaccessible, excessif, anormal, dispendieux, arnaque, arnaquer, ruineux, faramineux, hors de portée, rondelette, inconcevable, rédhibitoire

Lexical field of being cheap	faible, modique, avantageux, brader, imbattable, dérisoire, alléchant, réduit, occase, occasion, défiant toute concurrence, aubaine, modeste, clopinettes, bon prix, attrayant, clopinette, abordable, raisonnable, compétitif, accessible, acceptable, normaux, moyen, équitable, intéressant, convenable, négligeable.
Lexical field of prices and costs	prix, tarif, montant, coût, loyer, vente, achat, location, frais, abonnement, facture, coûter, facturer, payer, tarifer, vendre, devis, paiement, rabais, tarifaire, croissance, promotion, remise, ristourne.
Lexical field of statistical institutions	bce, banque centrale, banque central, banque de France, insee, fed, taux directeur, taux intérêt
Additional keywords in English	price, prices, cost, costs, rent, rents, bill, bills.

Table 9: List and content of lexical fields used to detect tweets related to prices matters

7.4.2. Detecting tweets content towards prices evolutions

In this section, we aim to specify the content of the tweet towards prices evolutions – inflation, deflation, disinflation, and prices stability – using keywords in French and in English from Table 9. For instance, if a tweet contains one of the words of the lexical field of inflation, it is classified as related to inflation. If a tweet does not contain any of the keywords included in the listed lexical fields, it is classified as "other" (out of topic). In Table 10, we summarize the heuristic rules applied, which combined lexical fields of Table 9 in this fashion.

Category	**Heuristic rule**
Inflation	[Lexical field of acceleration *OR* Lexical field of increase] *EXCLUDING* Lexical field of deflation.
Deflation	Lexical field of decrease *OR* Lexical field of deflation
Disinflation	[Lexical field of decrease *OR* Lexical field of slowdown OR "disinflation"] *EXCLUDING* Lexical field of deflation
Prices stability	Lexical field of prices *AND* lexical field of stabilization

Table 10: List of the keywords to specify the tweet's content. Reading Note: a tweet is classified as talking about "inflation", if it contains one of the keywords of the lexical field "acceleration" or one of the keywords of the lexical field "increase", but does not contain any words belonging to the lexical field of "deflation".

7.4.3. Keywords list to target economists and finance experts among retweeters

Each user has completed the sidebar "Description" where they describe their Twitter account content and profile. This information is available in our data and enables us to identify the users' profile. We lists the keywords used to target the community of economists and finance experts among the Banque de France retweeters with the information available in this variable "description", in French: *banque, actuaire, bancaire, banque, bank, assurance, finance, marche, market, investissement, bourse, business, economi, economy, credit, monnaie, entreprise, pme, tpe, eti, monetary*. If any of those words appears in their biography, then a user is considered as an economist or finance professional.

7.5. List of additional explanatory variables

This appendix describes the additional explanatory dictionary-based variables used in the Random Forest that predicts whether tweets are related to prices matters. For each tweet, each variable checks whether at least one word of a precise lexical field is found. For instance, the variable "acceleration" is set to 1 if one of the keywords related to the lexical field of acceleration is found. The variables have been built in an expert way to better characterize tweets regarding the prices problematic, according to five dimensions:

- Variables to check the presence of words related to prices matters or statistical institutions involved with the inflation problematic;
- Variables to check the presence of words related to directional evolutions to specify the prices evolution;
- Variables to check the presence of words related to the "cheap" or "expensive" lexical field to specify the perception of prices levels;
- Variables to check the presence of degree adverbs/adjective, negation terms;
- Variables to check the presence of words to be excluded (fake friends of keywords that belong to the lexical field of prices).

Most of the variables are in French, but seven variables are based on English keywords. In further developments, more variables based on English words could be added. The variables also distinguish between verb and noun in order to integrate grammatical logics into the model. Table 11 lists the additional dictionary-based variables, by describing what they attempt to capture, as well as the language and the grammatical type of their keywords. Those keywords themselves are not displayed in this paper for concision purposes.

Variable	Variable lexical field	Language	Grammatical type
1	acceleration	French	Noun
2	to accelerate	French	Verb
3	increase	French	Noun
4	to increase	French	Verb
5	decrease	French	Noun
6	to decrease	French	Verb
7	slowdown	French	Noun
8	to slow down	French	Verb
9	stabilization	French	Noun
10	to stabilize	French	Verb
11	stagnation	French	Noun
12	to stagnate	French	Verb
13	change	French	Noun
14	to change	French	Verb
15	stative verbs	French	Verb
16	expensive	French	Noun/Adjective

17	cheap	French	Noun/Adjective
18	affordable	French	Noun/Adjective
19	prices	French	Noun
20	discount	French	Noun
21	inflation (economical words)	French	Noun
22	negation terms	French	Adverb
23	little	French	Degree Adverb
24	much	French	Degree Adverb
25	little	French	Degree adjective
26	much	French	Degree adjective
27	terms to exclude	French	(no difference)
28	prices (very restrictive list)	French	(no difference)
29	prices (restrictive list)	French	(no difference)
30	prices (large list)	French	(no difference)
31	statistical institutions	French	Names
32	increase	English	Noun
33	to increase	English	Verb
34	decrease	English	Noun
35	to decrease	English	Verb
36	stabilization	English	Noun
37	to stabilize	English	Verb
38	prices	English	Noun

Table 11: List of additional dictionary-based explanatory variables

7.6. Definition of the metrics

7.6.1. Pearson and Spearman correlations coefficients

Pearson and Spearman correlation coefficients are indicators used to assess how well two variables are correlated.

The Pearson coefficient is used to measure the linear correlation between two variables. It is defined as the covariance of the two variables divided by the product of their standard deviations. It has a value between +1 and −1. A value of +1 means a positive collinearity, 0 means no linear correlation, and −1 means a negative collinearity.

Spearman's rank correlation coefficient is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It is defined as the Pearson correlation between the rank values of those two variables. While Pearson correlation assesses linear relationships, Spearman correlation assesses monotonic relationships not necessary linear. With no repeated data values, a Spearman correlation of +1 or −1 occurs when one variable is exactly a monotonic function of the other. The Spearman correlation between two variables will be high when the observations of two variables have a similar rank, and low otherwise. Spearman coefficient is appropriate for both continuous and discrete ordinal variables.

7.6.2. Metrics for model evaluation

Our study aims to predict the value of a binary variable Y as a function of a number of explanatory variables X. More precisely, if a tweet concerns prices matters (Y=1) or not (Y=0). Supervised machine learning methods rather produce a probability than a classification. The idea is then to set a probability threshold to classify a tweet in a category. This is a classification rule and model evaluation consists in comparing predicted and true outcome values by varying this threshold. Different metrics can be used to evaluate the quality of a classification.

Confusion matrix

A classification can be qualified by a confusion matrix. It provides more insight into the performance of a predictive model by describing which classes are predicted correctly, and what types of errors are being made. In the case of a two-class classification problem, the confusion matrix is:

	Positive prediction	**Negative prediction**
Positive class	Count of True Positive (TP)	Count of False Negative (FN)
Negative class	Count of False Positive (FP)	Count of True Negative (TN)

Table 13: Confusion matrix

True positive is when the actual value is 1, and the predicted value is 1. For instance, the tweet concerns prices matters, and the model predicted it would. True negative is when the actual value is 0 and the predicted value is 0. For instance, the tweet does not concern prices matters, and the model predicted it would not. False positive is when the actual value is 0 and the predicted value is 1. For instance, the tweet does not concern prices matters, and the model predicted it would. False negative is when the actual value is 1 and the predicted value is 0. For instance, the tweet concerns prices matters, and the model predicted it would not.

Evaluation metrics given a probability threshold

From the confusion matrix, it is possible to calculate the following metrics to evaluate the quality of predictions by combining each of the four classes. Table 14 summarizes them, with a short explanation of their meaning.

Name	**Computation**	**Meaning**
Accuracy	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$	What proportion of the observations are correctly classified?
Precision	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$	Among actually relevant observations, what proportion is correctly identified?
Recall	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$	Among what is identified as relevant, what proportion actually is?
F-score	$2 \times \frac{\mathrm{precision}\ \times \mathrm{recall}}{\mathrm{precision}+\mathrm{recall}}$	A harmonic combination of precision and recall

Table 14: Most standard model evaluation metrics

The most common metric used is the accuracy, because it is the most straightforward to understand. However, it can overestimate the performance of a model in the case of imbalanced data. Then, the recall and precision metrics become handy and need to be closely analyzed.

Receiver Operating Characteristic curve and Area Under the Curve

The ROC curve (or Receiver Operating Characteristic curve) is a plot that summarizes the performance of a binary classification model. Each point indicates the False Positive Rate and the True Positive Rate, for a given threshold. At (0, 0), the classifier assigns to all the observations the prediction of Y=0: there are no false positives, but also no true positives. At (1, 1), the classifier assigns to all the observations the prediction of Y=1: there are no true negatives, but also no false negatives. At (0, 1), the classifier predicts no false positives and no false negatives, and is therefore perfectly accurate. At (1, 0) the classifier has no true negatives nor true positives, and is therefore always being wrong. Simply reversing its predictions allow getting a perfectly accurate classifier. A random classifier draws a line from (0, 0) to (1, 1). The ROC curve makes it easy to compare the prediction quality of several models by simply comparing their respective ROC curves. The closer a ROC curve is from the top left corner, better is the model.

The Area Under the Curve (AUC) is an indicator of the quality of a model. The AUC values range between 0 and 1 (1 for a perfect model, 0.5 for a random model, 0 for a model always wrong). It is a great tool for model evaluation but it is not well tailored for imbalanced data because it tends to overestimate the quality of the model.

BANQUE DE FRANCE

EUROSYSTÈME

Using Twitter Data to Gauge Inflation Perception

Data Science in Central Banking: Machine Learning Applications

Julien Denes

Banque de France

19-22 October 2021

Motivations

Why measuring inflation?

- ensuring price stability is a key role of central banks
- observing past inflation is easy, but much less interesting than future inflation
- what truly matters is how people anticipate future inflation, as they will act accordingly

Why using Twitter?

- granular data as a complement to household surveys thanks to the rich available information about the user and the tweet
- real time, high frequency data
- accessible in nearly open data and free

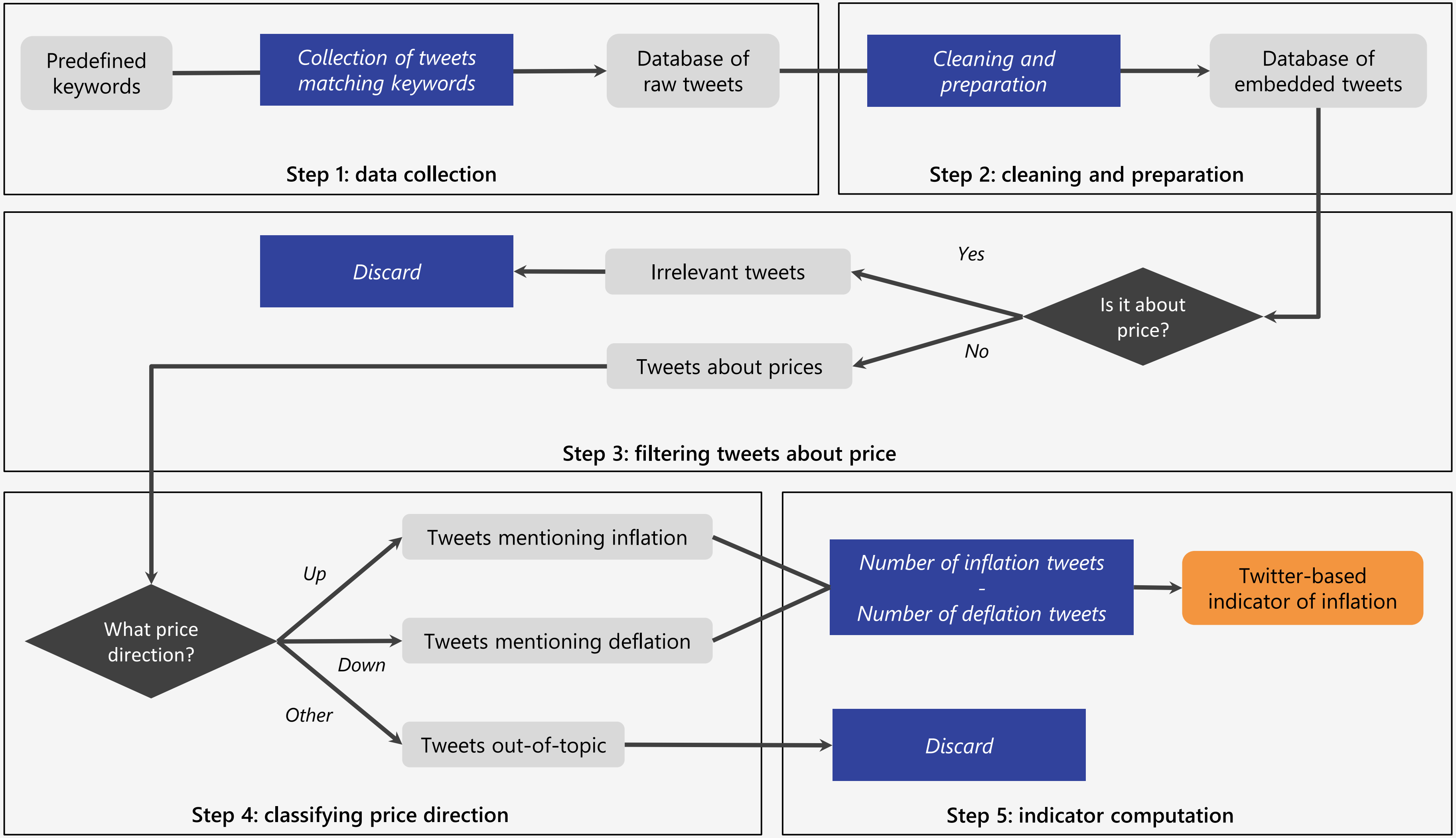
Introduction

Summary

1. Presentation of the pipeline
2. Detailed methodology of each step
3. Resulting indicators
4. Improvements and future works

Presentation of the pipeline

What do people think about the evolution of prices?



Predefined keywords

Collection of tweets matching keywords

Database of raw tweets

Step 1: data collection

Cleaning and preparation

Database of embedded tweets

Step 2: cleaning and preparation

Is it about price?

Yes

Irrelevant tweets

Discard

No

Tweets about prices

Step 3: filtering tweets about price

What price direction?

Up

Tweets mentioning inflation

Down

Tweets mentioning deflation

Other

Tweets out-of-topic

Step 4: classifying price direction

Number of inflation tweets

-

Number of deflation tweets

Twitter-based indicator of inflation

Discard

Step 5: indicator computation

Detailed methodology

Step 1: Data collection

Keywords matching

- collecting any tweet matching a broad set of keywords, from economical terms to price-related terms and expert inflation vocabulary (≈100 words)

Additional filtering

- from January 2008 to June 2021, and only in French
- only a subset of users: retweeters of Banque de France’s tweets (≈3,500 users)

Resulting amount of data

- more than 500,000 tweets

Detailed methodology

Step 2: Data cleaning and preparation

A standard cleaning

- removing stop words
- applying lemmatization (using only words roots) and stemming (splitting)

Text transformation using embeddings

- embedding is a representation of a text (here a tweet) as a numerical vector
- we combine word2vec embeddings and keywords-based indicators

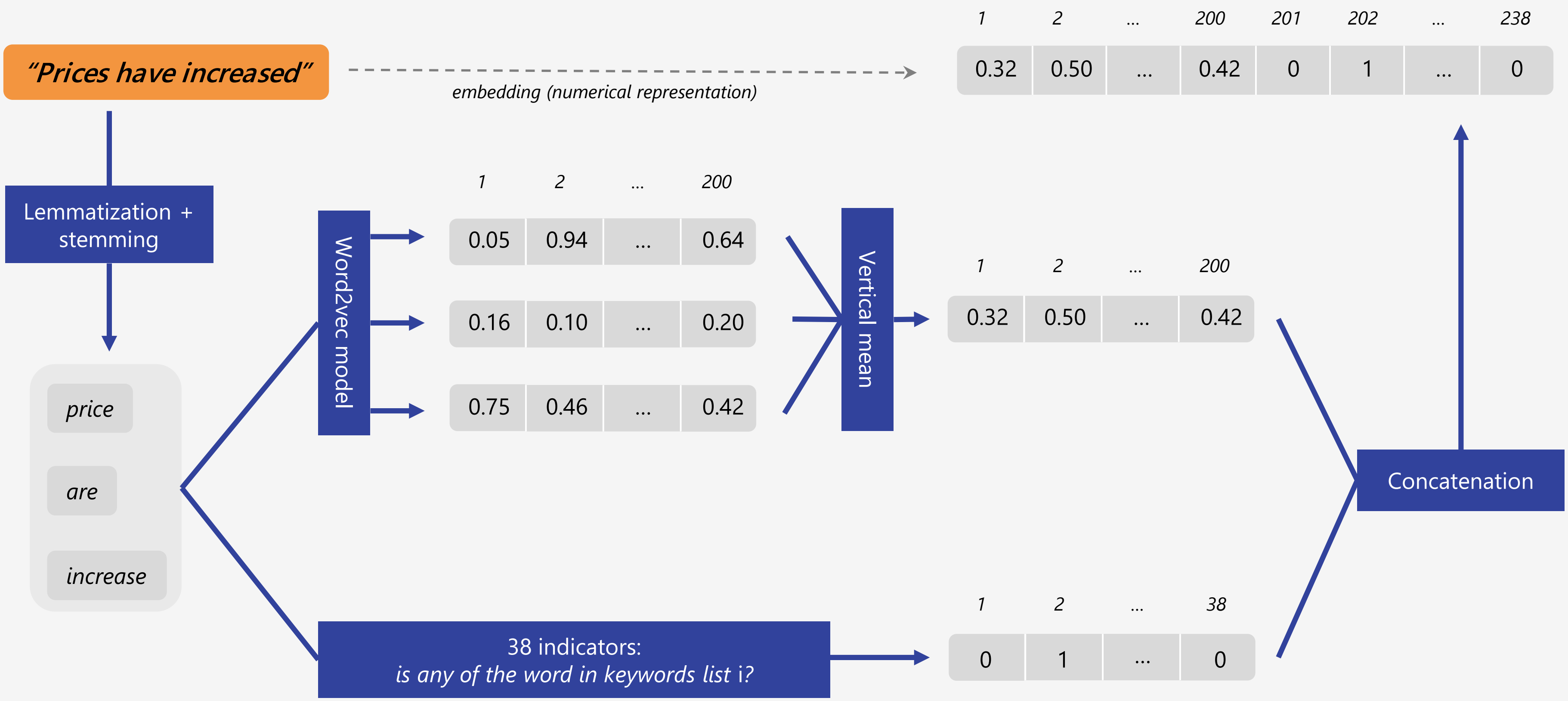
Why two types of embeddings

- word2vec provides general contextual information about the words used
- keywords indicators target the presence of specific lexical fields

Detailed methodology

Step 2: Data cleaning and preparation

An example of tweet full preparation



"Prices have increased"

embedding (numerical representation)

1	2	...	200	201	202	...	238
0.32	0.50	...	0.42	0	1	...	0

Lemmatization + stemming

price

are

increase

Word2vec model

1	2	...	200
0.05	0.94	...	0.64
0.16	0.10	...	0.20
0.75	0.46	...	0.42

Vertical mean

1	2	...	200
0.32	0.50	...	0.42

38 indicators:
is any of the word in keywords list i?

1	2	...	38
0	1	...	0

Concatenation

Detailed methodology

Step 3: Filtering tweets about price

Aim of the step

- input: a database of tweets, with their embeddings
- output: the subset of tweets that relate to price (i.e. relevant in a broad way)

Chosen model

- a random forest made of 500 trees
- major pros: fast to train and infer, light, and interpretable

Dataset used for training

- 800 tweets labelled binary as either “about price” (1) or not (0)

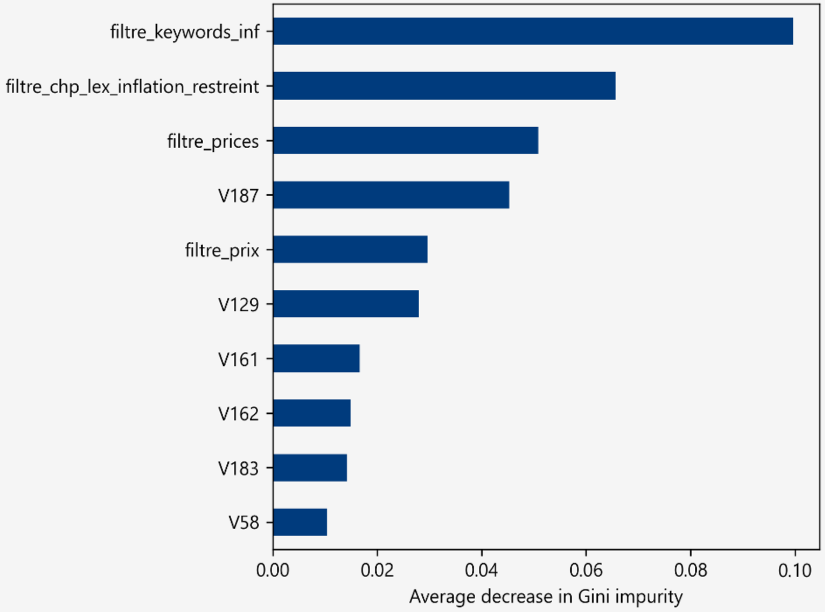
Detailed methodology

Step 3: Filtering tweets about price

Performance of the model

Metrics	Value on testing sample
Accuracy	90.00
F1-score	88.69
Precision	91.27
Recall	87.22

Feature importance



Detailed methodology

Step 4: Classifying according to price direction

Aim of the step

- input: a database of tweets about price, with their embeddings
- output: each tweet is tagged as mentioning prices going up, down, or anything else

Chosen model

- a random forest made of 500 trees (again)

Dataset used for training

- 1,100 tweets labelled as either “up”, “down”, or “other” (multi-label task)

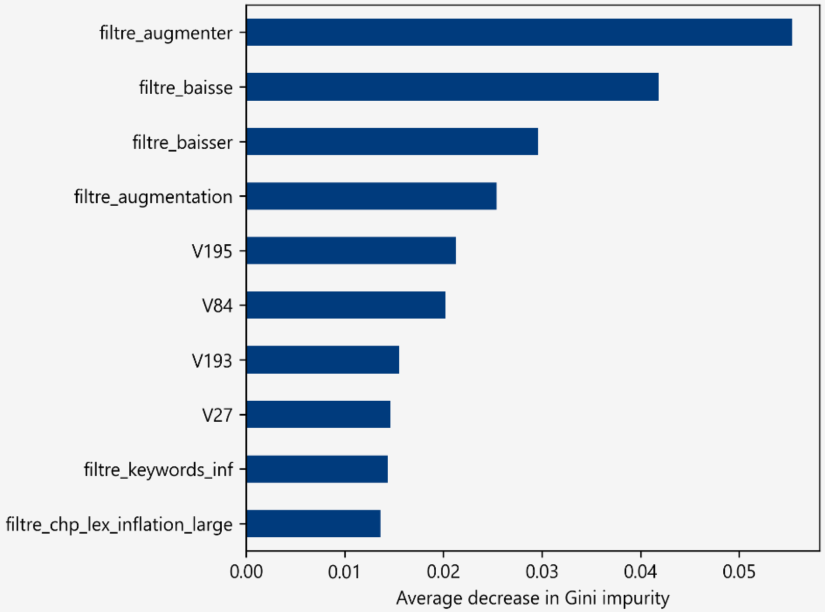
Detailed methodology

Step 4: Classifying according to price direction

Performance of the model

Metrics	Value on testing sample
Accuracy	85.58
F1-score	84.53
Precision	84.64
Recall	84.43

Feature importance



Detailed methodology

Step 5: Computing the indicator

Aim of the step

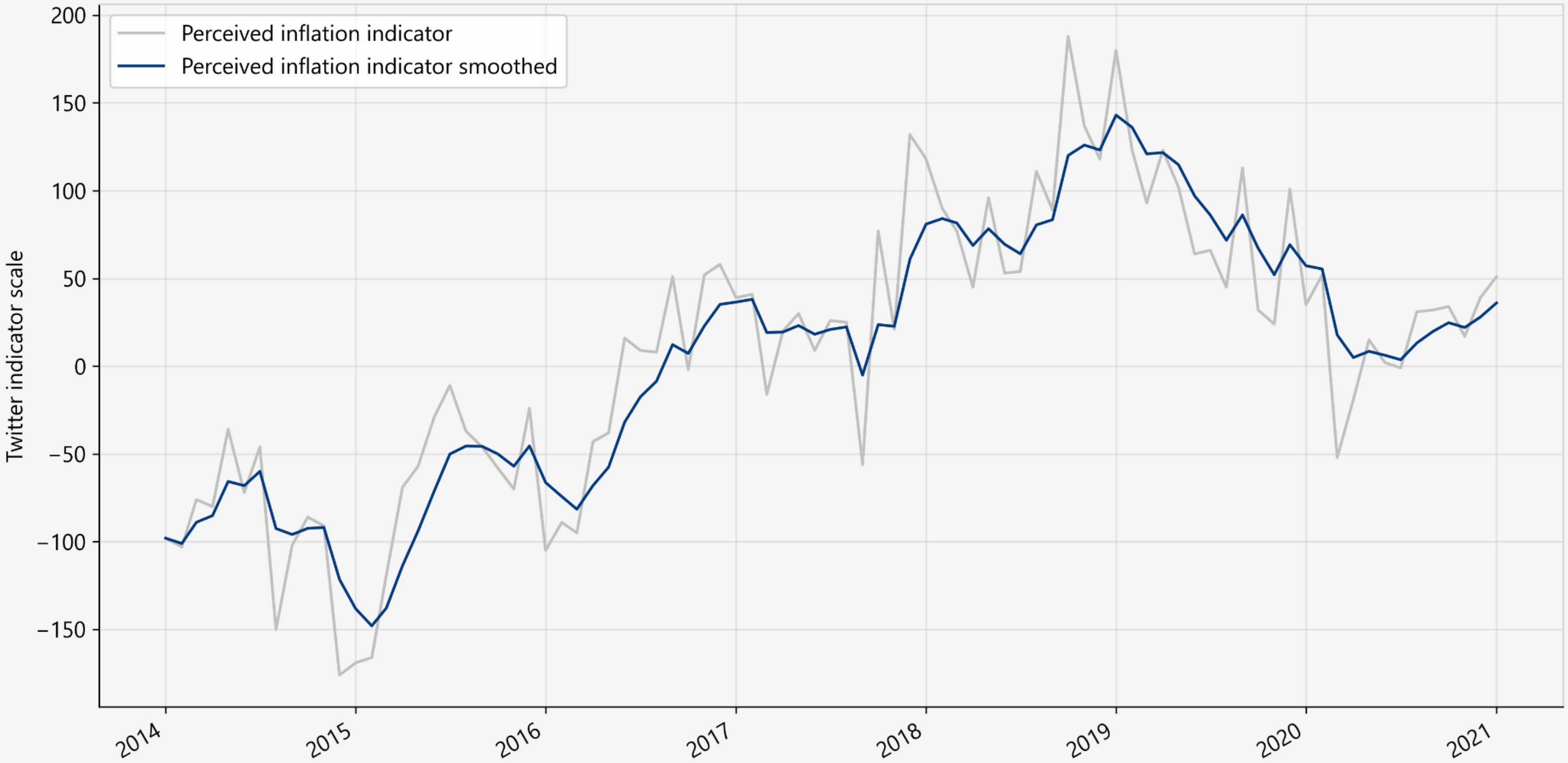
- input: a database of tweets tagged as “prices going up”, “prices going down”, or anything else
- output: an indicator of inflation as perceived by Twitter

Chosen method

- for each period of time (e.g. day, week, month), the value of the indicator is the difference between the number of tweets mentioning prices going up and those mentioning prices going down
- voluntarily simple and naïve as to mimic a balance of respondents opinion in household surveys (where answers are binary)

Results

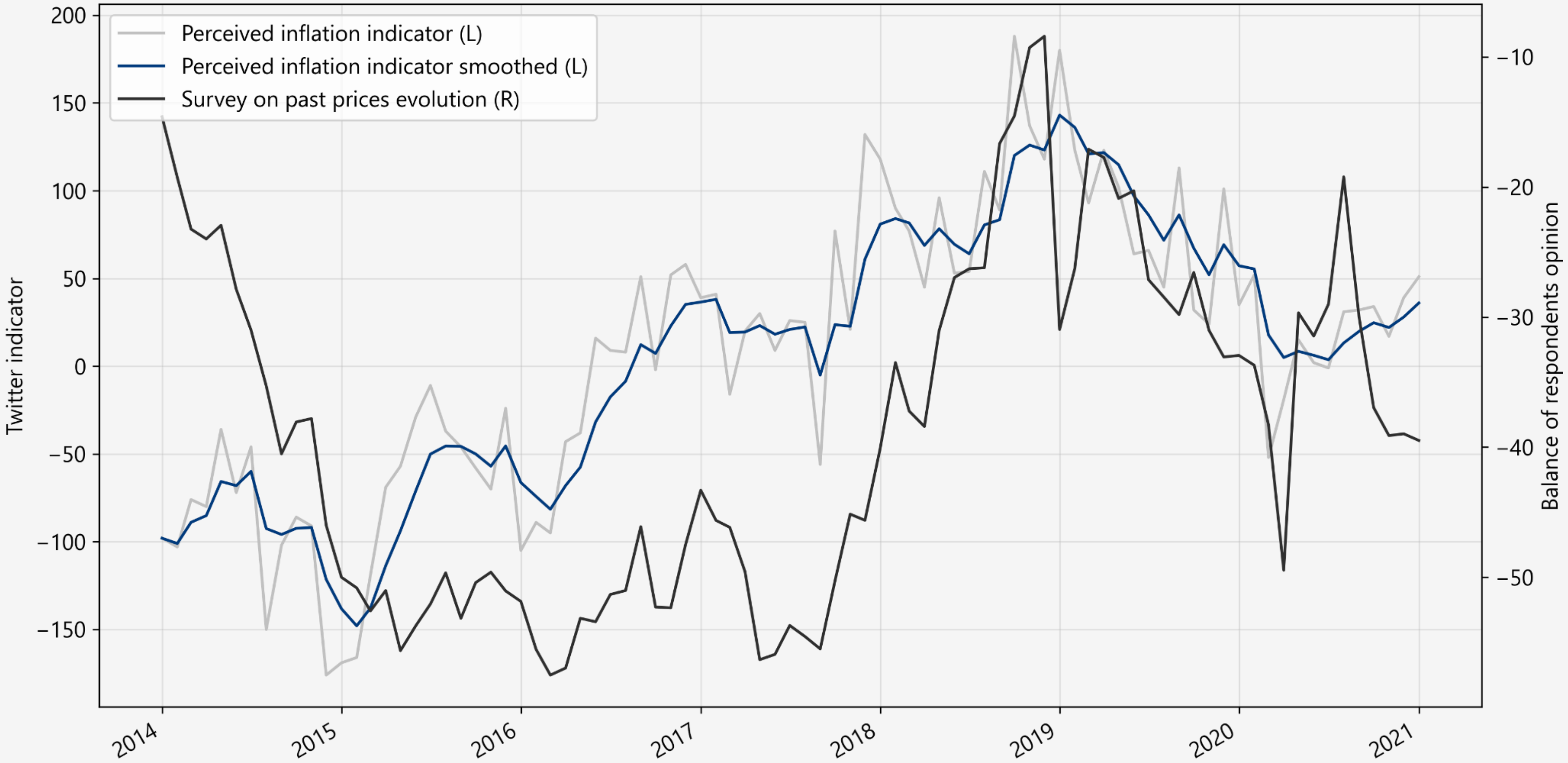
Twitter indicator of perceived inflation



BANQUE DE FRANCE
EUROSYSTÈME

Results

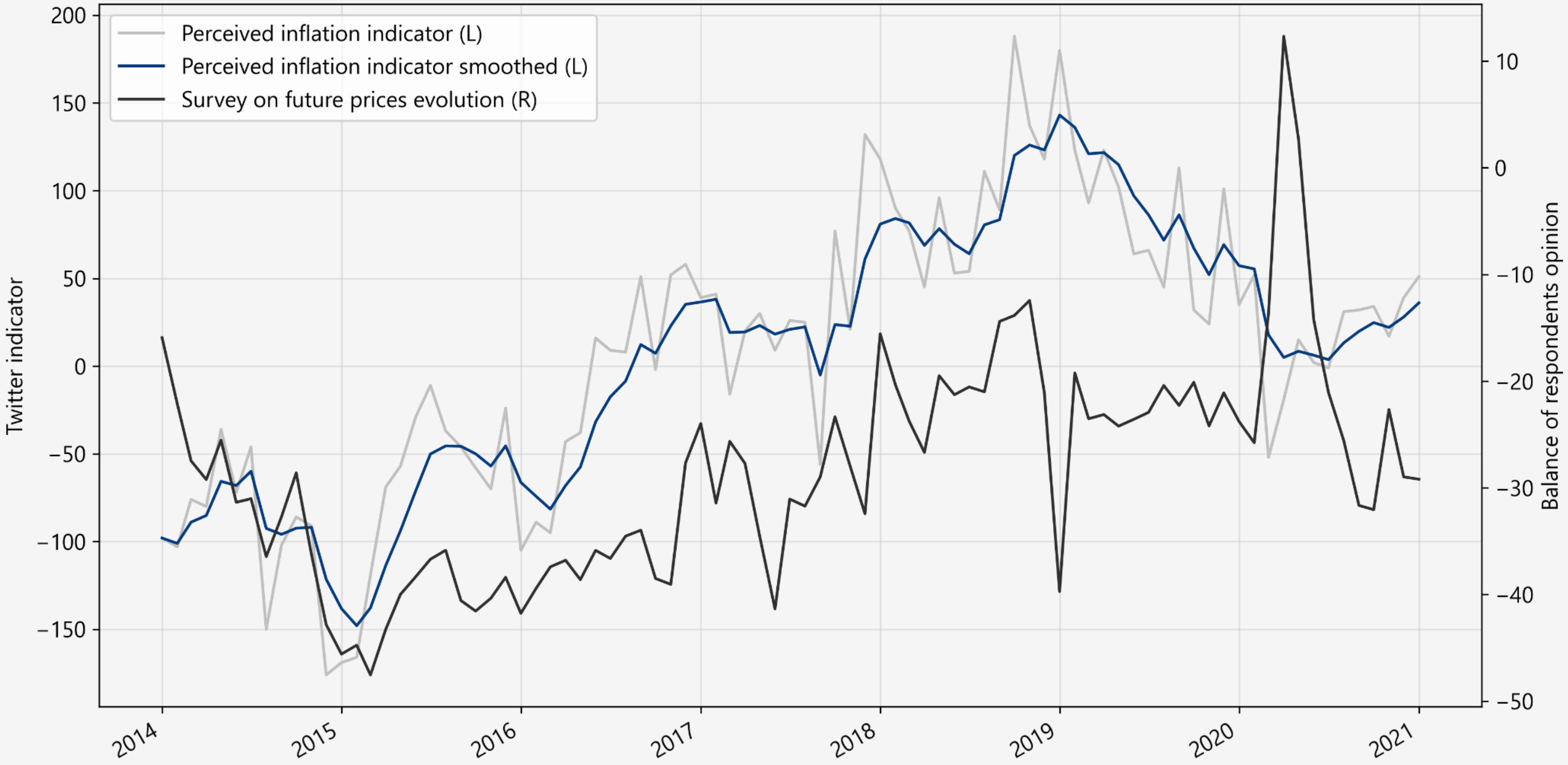
Consistency with households surveys: past evolution



BANQUE DE FRANCE
EUROSYSTÈME

Results

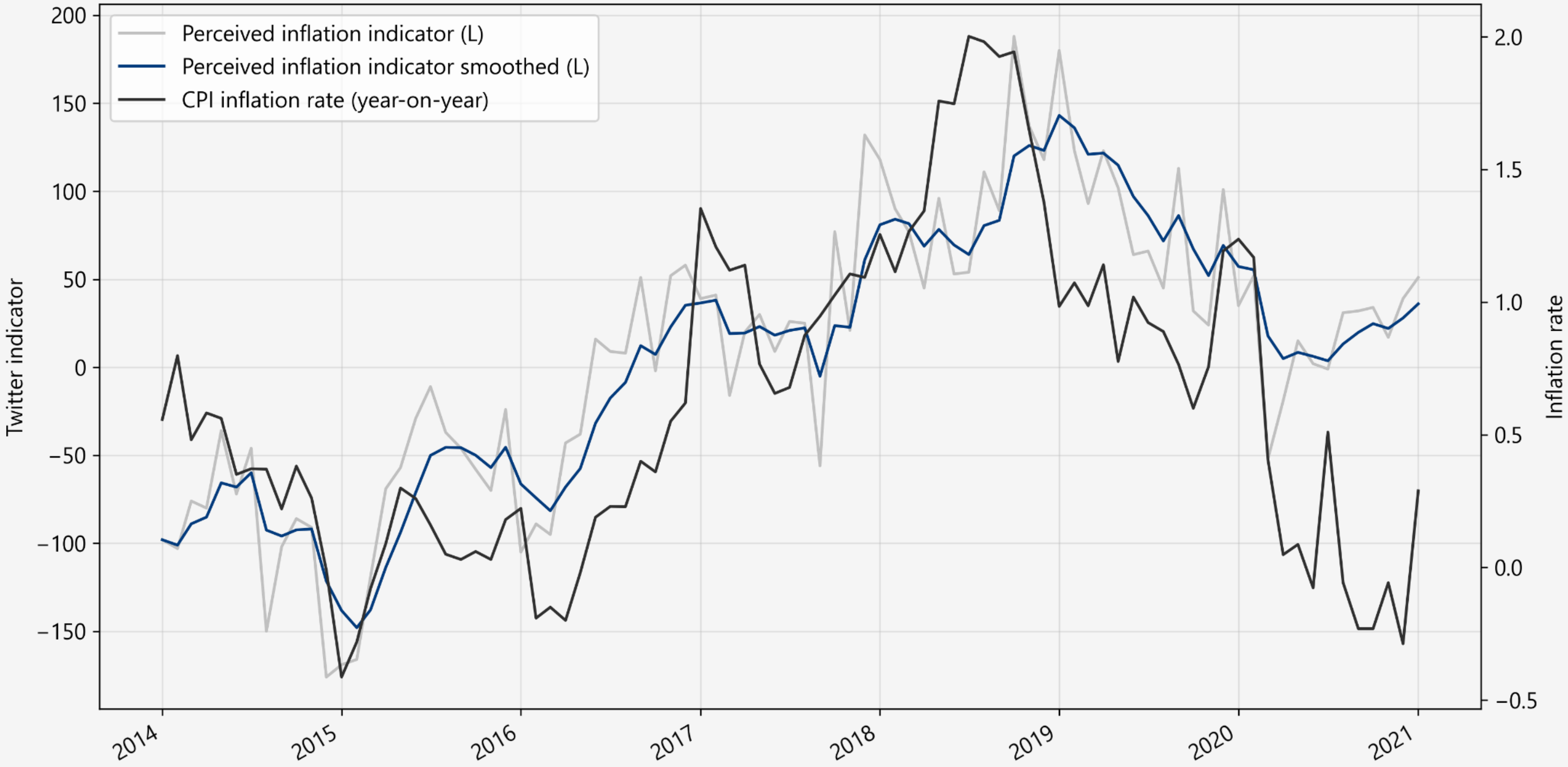
Consistency with households surveys: future evolution



BANQUE DE FRANCE
EUROSYSTÈME

Results

Consistency with true inflation rate



BANQUE DE FRANCE
EUROSYSTÈME

Results

Consistency: correlation coefficients

Correlation between...	Pearson	Spearman
Past evolution perception and Twitter indicator	0.213	0.219
Past evolution perception and smoothed Twitter indicator	0.305	0.329
Future evolution perception and Twitter indicator	0.460	0.487
Future evolution perception and smoothed Twitter indicator	0.505	0.558
Inflation rate and Twitter indicator	0.642	0.664
Inflation rate and smoothed Twitter indicator	0.735	0.792

Improvements and future works

Refining data collection

- lifting the restriction on Twitter users whose tweets are collected
- potential issue: huge increase of the volume of data

Distinguishing perception and anticipation

- the true goal is to capture how people anticipate future inflation
- requires an additional step to distinguish between past/present and future

Declining topics mentioned

- eliciting what topics are mentioned, to construct sectorial indicators: consumption prices, transportation, housing, raw material, etc.

BANQUE DE FRANCE
EUROSYSTÈME

Using Twitter Data to Gauge Inflation Perception

Julien Denes
julien.denes@banque-france.fr

Direction Générale du Système d'Information (DGSI)
Direction des Données et des Services Analytiques (DDSA)
Service Intelligence Artificielle, Données, Expérimentations (ILIADE)

Appendix

Predefined keywords for tweets collection

Lexical field	Keyword
Inflation (economic vocabulary)	inflation, déflation, stagflation, désinflation, inflationniste, déflationniste, antiinflationniste, antidéflationniste, ipc, ipch
Expensive	onéreux, cher, prohibitif, couteux, élevé, exorbitant, inabordable, conséquent, inaccessible, excessif, anormal, dispendieux, arnaque, arnaquer, ruineux, faramineux, hors de portée, rondelette, inconcevable, rédhibitoire
Cheap	faible, modique, avantageux, brader, imbattable, dérisoire, alléchant, réduit, occase, occasion, défiant toute concurrence, aubaine, modeste, clopinettes, bon prix, attrayant, clopinette, abordable, raisonnable, compétitif, accessible, acceptable, normaux, moyen, équitable, intéressant, convenable, négligeable
Prices and costs	prix, tarif, montant, coût, loyer, vente, achat, location, frais, abonnement, facture, coûter, facturer, payer, tarifer, vendre, devis, paiement, rabais, tarifaire, croissance, promotion, remise, ristourne
Statistical institutions	bce, banque centrale, banque central, banque de france, insee, fed, taux directeur, taux intérêt

BANQUE DE FRANCE
EUROSYSTÈME

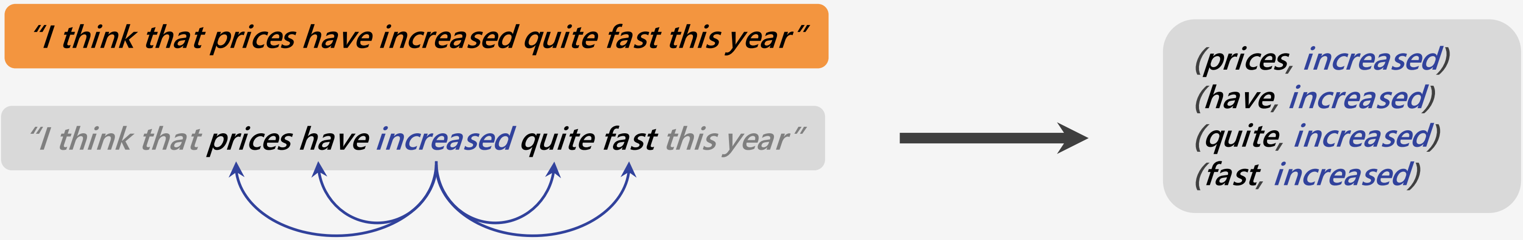
Appendix

Lexical fields used for embeddings indicators

Variable	Variable lexical field	Language	Grammatical type
1	acceleration	French	Noun
2	to accelerate	French	Verb
3	increase	French	Noun
4	to increase	French	Verb
5	decrease	French	Noun
6	to decrease	French	Verb
7	slowdown	French	Noun
8	to slow down	French	Verb
9	stabilization	French	Noun
10	to stabilize	French	Verb
11	stagnation	French	Noun
12	to stagnate	French	Verb
13	change	French	Noun
14	to change	French	Verb
15	stative verbs	French	Verb
16	expensive	French	Noun/Adjective
17	cheap	French	Noun/Adjective
18	affordable	French	Noun/Adjective
19	prices	French	Noun
20	discount	French	Noun
21	inflation (economical words)	French	Noun
22	negation terms	French	Adverb
23	little	French	Degree Adverb
24	much	French	Degree Adverb
25	little	French	Degree adjective
26	much	French	Degree adjective
27	terms to exclude	French	(no difference)
28	prices (very restrictive list)	French	(no difference)
29	prices (restrictive list)	French	(no difference)
30	prices (large list)	French	(no difference)
31	statistical institutions	French	Names
32	increase	English	Noun
33	to increase	English	Verb
34	decrease	English	Noun
35	to decrease	English	Verb
36	stabilization	English	Noun
37	to stabilize	English	Verb
38	prices	English	Noun

Appendix

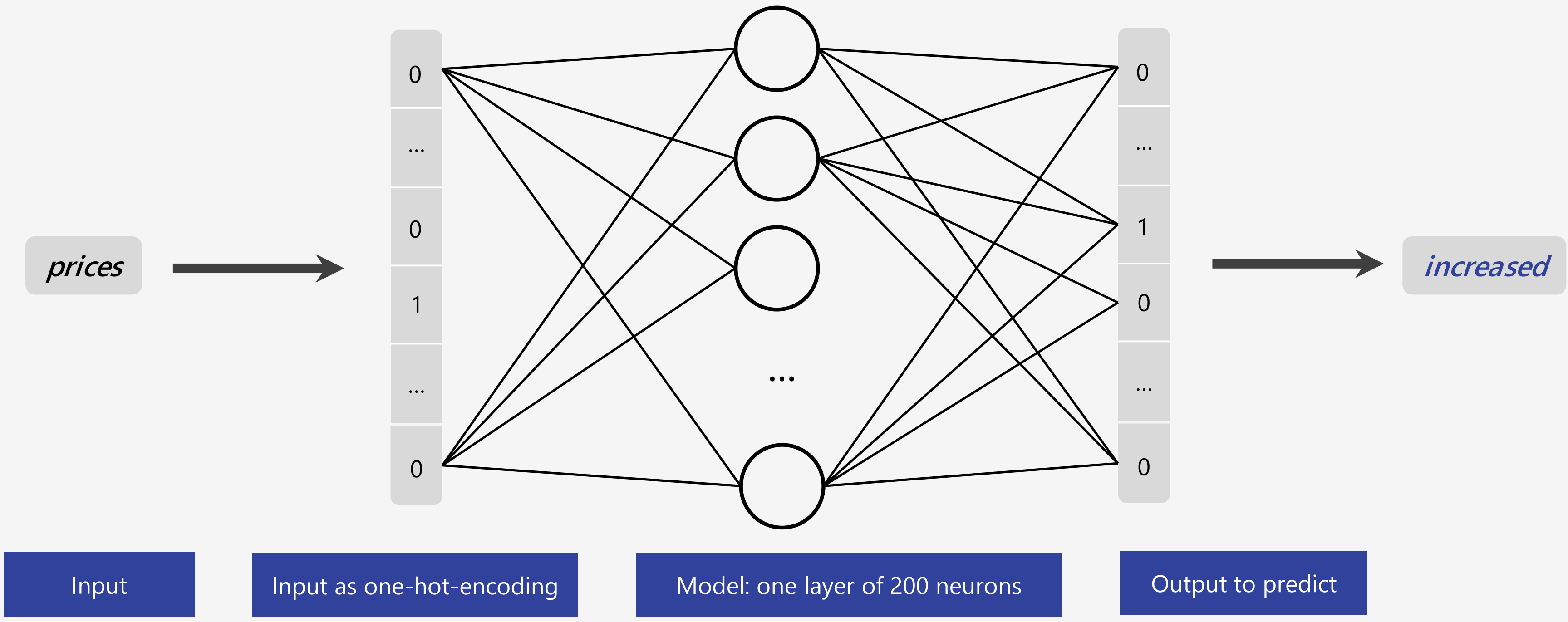
Focus on word2vec: how it works



"I think that prices have increased quite fast this year"

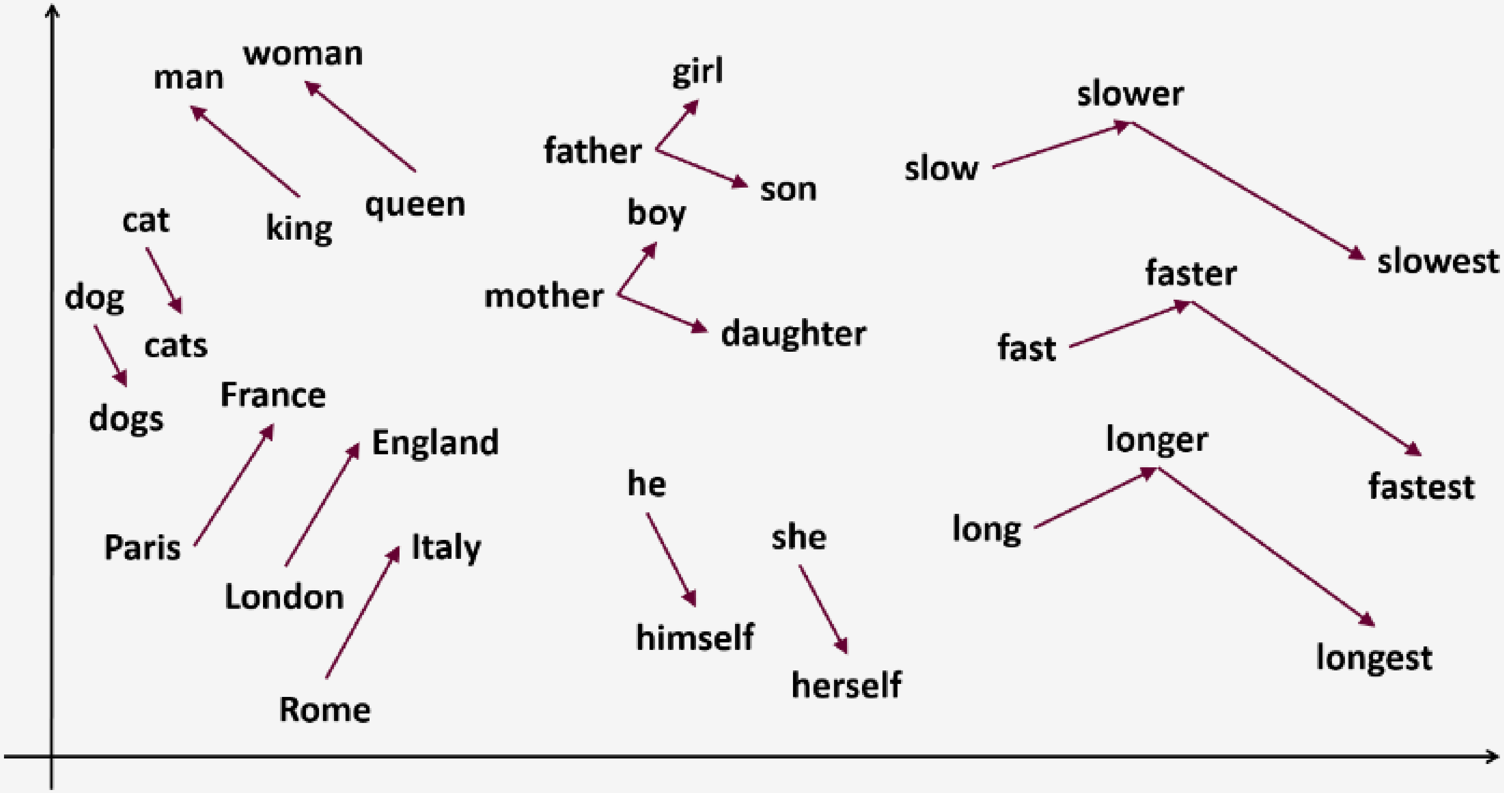
*"I think that **prices have** increased **quite fast** this year"*

(prices, increased)
(have, increased)
(quite, increased)
(fast, increased)



Appendix

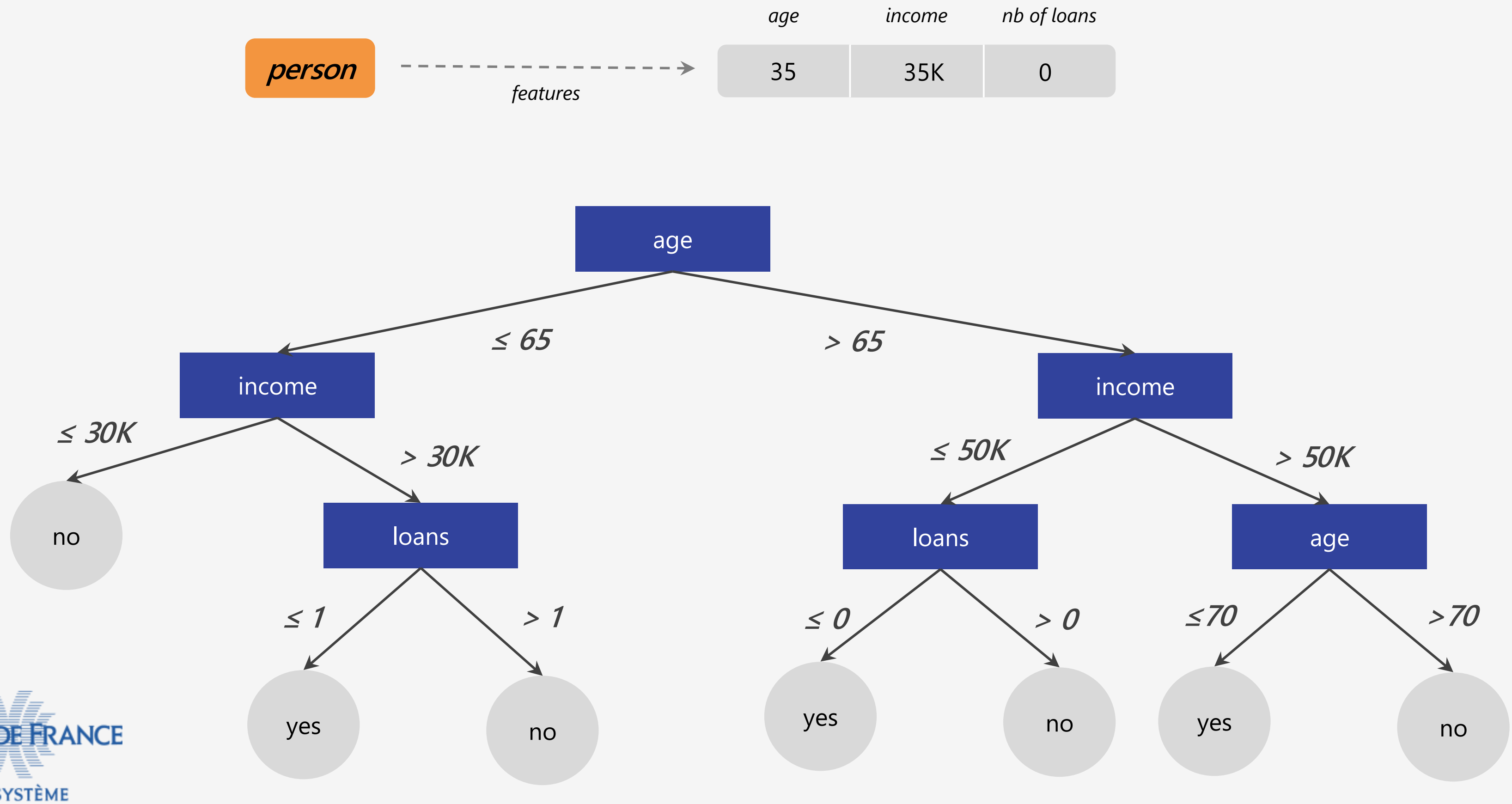
Focus on word2vec: properties



Appendix

Focus on random forests: decision tree

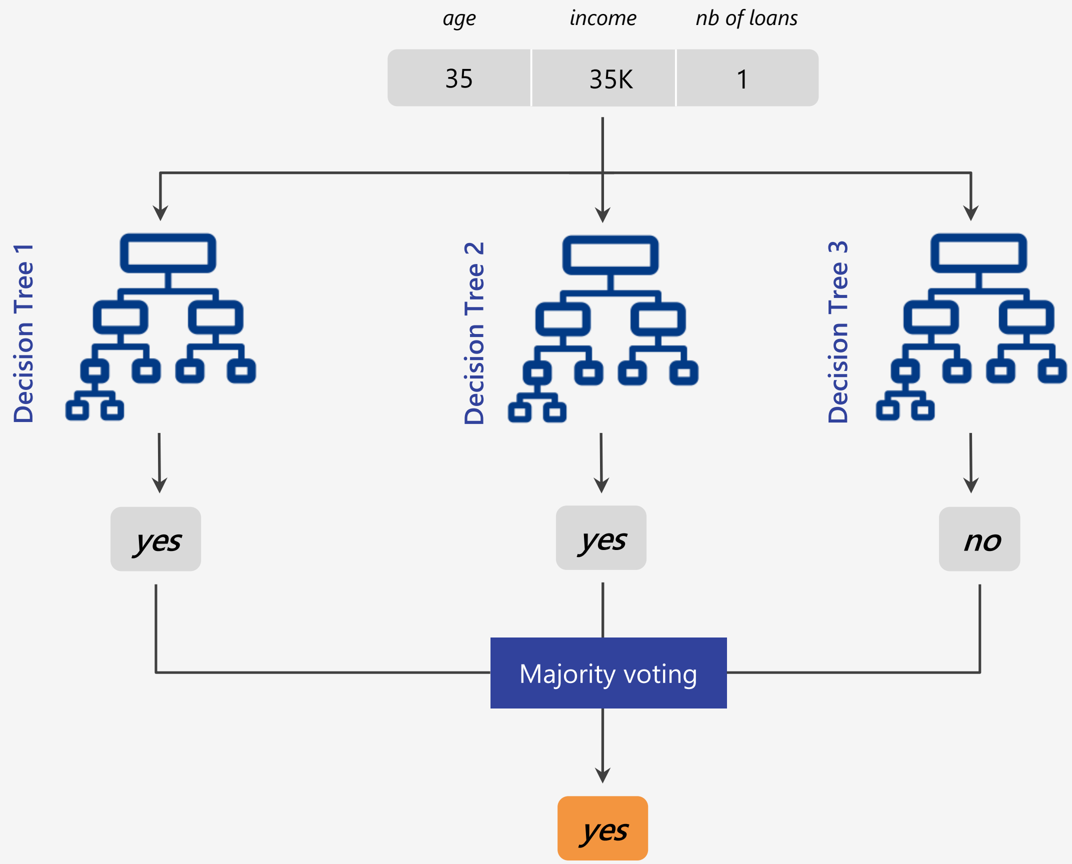
Will a person get a loan from their bank?



Appendix

Focus on random forests: forests

Will a person get a loan from their bank?



age	income	nb of loans
35	35K	1

BANQUE DE FRANCE
EUROSYSTÈME

Appendix

Sources and references

Seminal work of Banca d'Italia

- Angelico C., Marcucci J., Miccoli M. and Quarta F. (2021). Can we measure inflation expectations using Twitter? *Temi di discussione* n° 1318.

INSEE monthly household surveys

- Series on opinions about prices evolution.

INSEE consumption price index, base 2015

- Series and documentation. Inflation rate is computed by us as the 12-month relative difference.

BANQUE DE FRANCE

EUROSYSTÈME

This presentation describes preliminary results of ongoing research. It is made available to the public solely to elicit discussion and comments. Views expressed in the presentation are those of the authors and do not reflect the position of the Banque de France.

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Fostering European SMEs' internationalization using big data: the BIZMAP application[1]

Jean-Noel Kien, Etienne Kintzler and Theo Nicolas,
Bank of France

[1] This presentation 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 central banks and other institutions represented at the event.

Fostering European SMEs' internationalization Using Big Data: The BIZMAP Application[1]

Jean-Noël Kien, Etienne Kintzler, Théo Nicolas

Banque de France. E-mail: Jean-Noel.KIEN@banque-france.fr; Etienne.KINTZLER@banque-france.fr; Theo.NICOLAS@banque-france.fr. Address: 37 Rue du Louvre 75002 Paris.

Abstract

This paper proposes a decision-making tool (BIZMAP) that enables European small and medium-sized enterprises (SMEs) to visualize the most economically attractive European regions for the internationalization of their business activities. Building on more than 80 variables coming from seven different open access databases, we take advantage of big data and machine learning methods to include the most relevant ones in a standard gravity model of trade. In the end, we implement an interactive data visualization tool inside our BIZMAP application. Depending on the sector and the home country, we provide SMEs with a ranking of most promising European countries. Importantly, BIZMAP not only enables SMEs to understand what are the main drivers of this score but also offers the possibility to compare the 281 European regions with each other. Hence, by reducing information uncertainty abroad, BIZMAP is likely to improve the SMEs' analysis of new markets through the visualization of harmonized territorial attractiveness indicators.

Keywords: SMEs, Trade, FDI, Big Data.

JEL codes: F14, F17, F15, F31, F21.

[1] The views expressed in this paper are those of the authors and do not necessarily coincide with those of the Banque de France or the Eurosystem. The authors are grateful to Jérôme Coffinet, Jean-Brieux Delbos, Louis-Marie Harpedanne, Guillaume Gaulier and Antoine Berthou.

Non-technical summary

The internationalisation of economic activities opens up news opportunities for SMEs. However, some obstacles to their exploitation remain. Among them, the information deficit turns out to be one of the most salient. To tackle this issue, the BIZMAP application offers a decision-making tool that enables SMEs to identify the most economically attractive EU countries or regions for their internationalisation (exports or foreign direct investments – FDI).

The principle of the application is straightforward: after filling out all the necessary fields (sector, home country and type of internationalisation), BIZMAP first provides the SME manager with an interactive visualisation of the most promising national markets ranked by scores. The latter are based on a wide range of criteria aggregated into a five-dimension indicator: economic perspectives, standard of living, infrastructure, financial conditions and institutional environment. Then the application goes even further in the analysis by zooming on the 281 EU regions. In this regard, BIZMAP does not provide one unique solution but encourages the SME manager to explore and compare the different areas and criteria used to compute the scores. This user-friendly visualisation is particularly addressed to practitioners and can be understood without technical background.

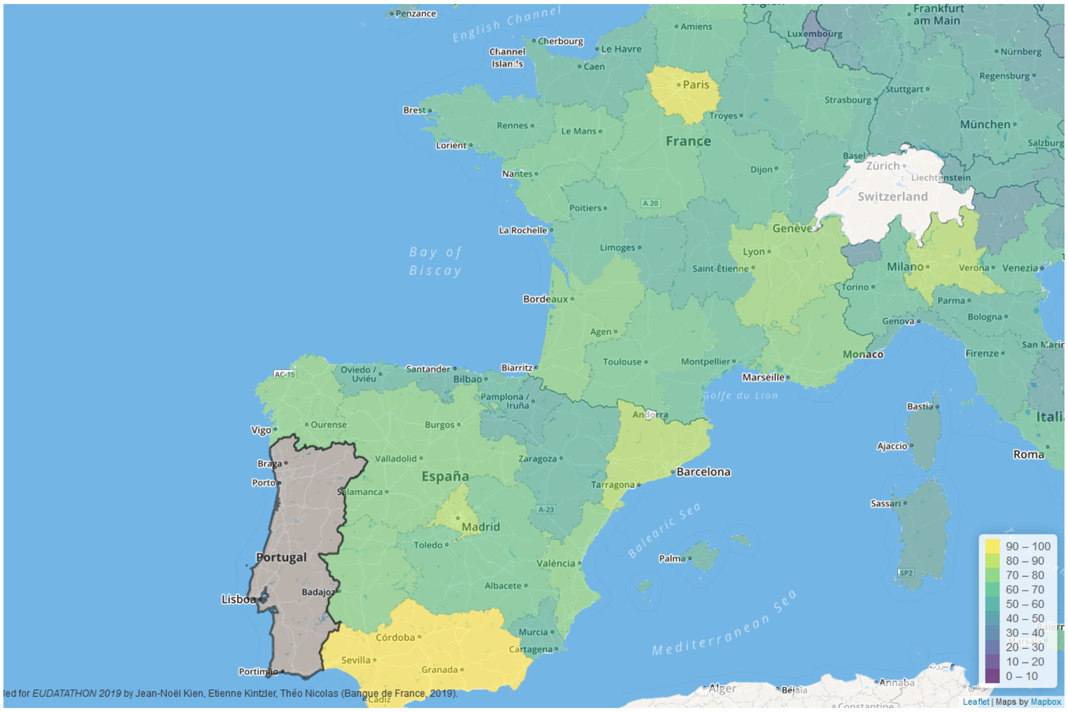
The application builds on 7 different open access databases: Eurostat, OECD, World Bank, European Central Bank, European Investment Bank, European Commission, CEPII. After harmonization, we take advantage of machine learning methods to select the best predictors of bilateral flows of imports and FDI. In this model, the distance between the two countries as well as their gross domestic products are crucial for both types of flows. In addition, the legal framework regarding insolvency and the cost associated with border compliance and domestic transport play a significant role in bilateral flows of imports. Concerning FDI flows, taxes on goods and services as well as air freight are the most important factors.

By reducing informational uncertainty abroad, BIZMAP enhances traditional evaluation of commercial opportunities and enables SMEs to target some EU markets before launching

more accurate research. Hence, we enjoin the entrepreneur to use BIZMAP in complement with information coming from governmental agencies or sectoral market studies that could provide him more qualitative data.

This working paper aims at presenting in detail the methodology used in the application. It allows comments, suggestions and reactions to be collected from practitioners and researchers. In particular, one way to improve significantly the model would consist in using products classification, combined with countries, to model bilateral trade flows. Thus, the SME manager would be able to choose in the application not only the sector but also the product. Ultimately, BIZMAP is intended to be shared and used among SMEs which are seeking for new opportunities abroad.

Figure 1: Zoom on regional scores for a Portuguese SME in the construction sector



led for *EUDATATHON 2019* by Jean-Noël Kien, Etienne Kintzler, Théo Nicolas (Banque de France, 2019).

1. Introduction

In the post-1950 period, the global increase in the flows of trade, capital and information has helped push the world economy into a state of globalization, in which most of economies are highly interconnected (Masson, 2001). In this context, the firm-level internationalization refers to the expansion of international business operations such as exports, international partnerships or foreign direct investment (FDI). By fostering innovation and facilitating spill-overs of technology, this participation in global markets may create opportunities to enhance productivity and can therefore be an important driver of employment growth (Wagner, 2012).

However, engaging in such activities can be expensive and usually only the most productive firms can afford to do so (Melitz, 2003; Helpman et al., 2004; Bernard et al., 2007). For instance, the entry into foreign markets implies transaction costs or fixed costs that can generate significant barriers (Eden & Miller, 2004). Given their small size, small and medium-sized enterprises (SMEs) suffer from typical obstacles which affect their ability to increase their activity abroad (Hollenstein, 2005; Paul et al., 2017). The latter can be classified as either internal, such as lack of internal resources, or external, such as uncertain institutional environments.

Hence, despite their importance in terms of activity and employment, SMEs only account for a small share of exports (OECD, 2015). In most OECD countries, for instance, SMEs represent more than 95% of all enterprises, about two-thirds of total employment and more than half of the value added of the business sector. Yet, their contribution to overall exports stands between 20% and 40% for most OECD economies (see figure 2).

Thus, although the fragmentation and specialization of global economic activity opens up a number of opportunities for SMEs, some obstacles to their exploitation remain. Among them, the lack of information is one of the most salient for example when it comes to selling goods and services on foreign markets (Lloyd-Reason et al., 2009). This patchy knowledge limits their ability to choose the geographical areas most suited to their business.

To overcome these difficulties, this paper combines many economic and financial data in open access to determine a multidimensional indicator of the attractiveness of European territories. The latter makes it possible to evaluate, according to SMEs' activity, which are the most promising markets based on a wide range of criteria. By reducing information uncertainty abroad, the BIZMAP application enables SMEs to improve their analysis of new markets through the use of an harmonized territorial attractiveness indicator.

To capture the protean nature of attractiveness at local level, the application builds on 7 different open access databases coming from Eurostat, the European Central Bank (ECB), the Organisation for Economic Cooperation and Development (OECD) , the European Investment Bank (EIB), the European commission (AMECO), the World Bank and the Research and Expertise on the world economy (CEPII) which is a French institution specialized in international trade. Based on our expert judgment, we end up with an unified database encompassing more than 80 preselected variables for the 28 members of the European Union over the period 2015-2021.[2]

Our approach relies on big data methods. First, we aggregate the time series and impute missing values with either random forest techniques or Kalman filters. Second, given the high dimensionality of our dataset, the most relevant variables are selected according to Lasso (Least Absolute Shrinkage Selection Operator) regressions applied to a gravity model of trade using either imports or FDI as dependent variable.

In the end, we obtain the contribution of each variable to exports or FDI in order to weight the variables we use to compute the indicators of geographical attractiveness and we propose a data visualization of our results inside our BIZMAP application. The principle is straightforward: after filling out all the necessary fields on the application (sector and home country), BIZMAP provides the SME with an European ranking based on an interactive visualization

[2] Note that, for some variables, our dataset both incorporates the 3-year economic forecast of the European Commission and our own forecasts based on Kalman filter or random forest methods. See section 3 for more details.

which indicates what are the most attractive European countries for its specific activity.[3] Importantly, BIZMAP also enables SMEs to understand what are the main drivers of this score by presenting the contributions of the most important variables. Finally, BIZMAP offers the possibility to compare the 281 European regions with each other using the Eurostat NUTS 2 classification[4]. Looking at countries heterogeneity, SMEs are therefore able to have a clearer picture of the most attractive European areas.

Our paper relates to the literature focusing on the firm decision to engage in international activity. While the traditional trade theories discuss the importance of differences in technology (David, 1817) and factor endowments (Heckscher & Ohlin, 1933) across economies to highlight comparative advantages, the New Trade Theory developed a model of monopolistic competition in which only the most productive firms internationalize their business (Melitz, 2003; Helpman et al., 2004). In contrast, we focus on the determinants of internationalization based on the economic potential of foreign markets. In particular, BIZMAP aims to reinforce the European economic integration which is likely to increase the growth potential of its members through higher regional trade (Vamvakidis, 1998).

The challenge of selecting the main drivers of the external performance among a wide range of possible variables was discussed in the economic growth literature under the so-called issue of "openendedness of theories" (Brock & Durlauf, 2001). In this case, one faces both the traditional problem of estimation uncertainty and the additional one of model uncertainty related to the choice of covariates. We tackle this issue by implementing Lasso methods, which provide a formal treatment of model uncertainty by considering all possible sets of variables.

[3] The determinants of attractiveness are studied according to the Statistical classification of economic activities in the European Community (NACE Rev.2).

[4] The current NUTS 2016 classification is valid from 1 January 2018 onwards and lists 104 regions at NUTS 1, 281 regions at NUTS 2 and 1348 regions at NUTS 3 level. The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU for the purpose of socio-economic analyses of the regions. More information are available in the following page `https://ec.europa.eu/eurostat/en/web/nuts/background`

The remainder of the paper is structured as follows. Section 2 presents the preselected potential drivers of internationalization. Section 3 deals with the empirical strategy. Section 4 discusses the results. Section 5 exhibits the BIZMAP application. Section 6 concludes.

2. Potential drivers of internationalization

The potential drivers of internationalization are numerous and incorporate among other aspects growth prospects, demography, education, quality of institutions, access to finance. Consequently, the latter can be searched in very wide areas of economics.

The first conceptual step consists in choosing the dependent variables which captures the economic potential of a foreign market. Here we focus on the balance of payments defining either flows of imports or FDI as measures of a country's commercial dynamism. More precisely, we look at bilateral flows in order to evaluate trade potential of each European country with respect to a given country. Coming from the Eurostat database, bilateral flows of imports and FDI are timely and harmonized across European countries.

While constructing the set of potential drivers of internationalization, we considered the following themes: (i) economic prospects; (ii) infrastructure; (iii) institutional environment; (iv) financial conditions ; and (v) demography and standard of living. The final dataset comprises 82 explanatory variables coming from 7 different open access databases over the period 2015-2021[5]. The main categories of potential drivers of internationalization are discussed below (see Table 1 and Table 2 for names of variables, sources and the granularity of data available).

2.1. Trade related variables

This block of variables refers to various statistics that are informative for bilateral trade outcomes based on the detailed trade data of the French center of Research and Expertise on

[5] Note that, for some variables, our dataset both incorporates the 3-year economic forecast of the European Commission and our own forecasts based on Kalman filter or random forest methods. See section **??**for more details.

the world economy (CEPII). Regarding the trade distance, we use the distance measures made available by the CEPII which hinge on city-level data to assess the geographic distribution of population (in 2004) inside each nation. The basic idea is to calculate distance between two countries based on bilateral distances between the biggest cities of those two countries, those inter-city distances being weighted by the share of the city in the overall country's population. [6] We also add two dummies: while the dummy *trade contiguity* takes the value 1 whether two countries are adjacent or 0 otherwise, the *Common official language* one takes the value 1 whether two countries share a common official language. In our case, we assume that a lower distance, a common language and a neighbouring country increase the probability of trading.

2.2. Economic prospects variables

To assess the economic potential of European countries, we rely on macroeconomic variables which describe the structure of the economy. Stemming from 4 different providers (the European commission, Eurostat, ECB, EIB), theses variables are available with different levels of granularity: country, macro-sector, NUTS 2 region, and NACE Rev. 2 activity.

As regards the database of the macro-economic database of the European Commission (AMECO), we select the gross value added at 2010 price, the harmonised consumer price index, the unemployment rate, the private final consumption expenditure and the gross fixed capital formation to account for productive capacity and growth prospects at the country-level. We also include the ECU-EUR exchange rates to take into account the effect of exchange rate volatility on countries attractiveness within the European Union.

Drawing on Eurostat, we complement these measures by the GDP and the unemployment growth rate at the regional level. We also use sector-specific variable such as labour costs or sentiment indicators. The latter are made up of five sectoral confidence indicators : industrial confidence indicator, services confidence indicator, consumer confidence indicator, construc-

[6] See (Head & Mayer, 2002) for more details on distance measures)

tion confidence indicator and retail trade confidence indicator. At the NACE Rev 2 level, we include the amount of firm turnover, the wage adjusted labour productivity, the average personnel costs, the growth rate of employment, the gross operating surplus and the investment rate. In addition, we add the house price index as well as the amount of R&D expenditures at the national level.

Turning to financial variables we consider that foreign credit cycles may drive external demand. Thus, using the ECB database, the outstanding amounts of household and corporate credit are incorporated into the dataset. Finally, to capture the business cycle, we make use of the annual EIB Group Survey on Investment and Investment Finance (EIBIS). Encompassing all EU countries, this survey gathers qualitative and quantitative information on investment activities by small and medium-sized businesses and larger corporations, their financing requirements and the difficulties they face. It thus provides a wealth of unique firm-level information about investment decisions and investment finance choices. Restricting the survey to SMEs, we retain questions that focus on the expected investment, the share of companies that invest, the demand for products or services and the uncertainty about the future. Importantly, the EIBIS allows to gather the answers according to macro-sectors.[7]

2.3. Institutional environment

The institutional environment of a given country plays a crucial role in attracting foreign firms. To proxy the quality of institutions and the rule of law which encourages international trade we make use of three different databases coming from the World Bank and the OECD. First we rely on the Worldwide Governance Indicators (WGI) which aggregate governance indicators for over 200 countries over the period 1996–2018, for six dimensions of governance: voice and Accountability, political Stability and absence of violence, government effectiveness, regulatory quality, rule of law, control of corruption. These aggregate indicators combine the views of a large number of enterprises, citizens and expert survey re-

[7] Note that in this paper macro-sectors refer to four different macro-sectors: industry, services, construction and retail

spondents in industrial and developing countries. They are based on over 30 individual data sources produced by a variety of survey institutes, think tanks, non-governmental organizations, international organizations, and private sector firms.

Second, we choose the *Doing Business* indicators of the World Bank which provide objective measures of business regulations and their enforcement across 190 economies on the following topic: trading across borders, time to import, starting a business, resolving insolvency, regulatory quality, registering property, protecting minority investors and paying taxes.

Finally, we gather information about the tax environment of all European countries using the OECD tax database which provides comparative information on a range of tax statistics - tax revenues, personal income taxes, non-tax compulsory payments, corporate and capital income taxes and taxes on consumption - that are levied in the 35 OECD member countries.

2.4. Infrastructure

The quality of infrastructure is also of major importance in order to facilitate delivering freight from a given country to every country of the European Union. To proxy the beneficial effect of infrastructure, we add 8 more variables stemming from the World Bank, Eurostat and the EIB. While we select the variable *Getting electricity* of the *Doing Business* projet we also include transport network information available at the NUTS2 level in the Eurostat database such as the motorway network, the railway network, the air freight and the ocean freight. Then, exploiting the EIB's survey on investment, we select questions dealing with energy costs, access to digital infrastructure and availability of adequate transport infrastructure.

2.5. Financial conditions

Since the onset of the crisis, financial variables have gained prominence in explaining the performance of both firms and countries. Based on Eurostat, the ECB database and the EIB's survey on investment, we consider measures characterizing financing conditions including the debt of households, non-financial corporations and governments. In addition, we look at the effect of financial stability measures such as non-performing loans, the country-level core tier one ratio of European banks or the financial stress indicator of the ECB [8]. Besides we add measures of SMEs access to finance using answers of the EIB survey about the amount of credit obtained, the cost of the external finance obtained and even the collateral required.

2.6. Demography and standard of living

The trade attractiveness of a country is tightly connected with the demography and the standard of living. In particular, we include the Eurostat share of labour force with secondary and tertiary education, as well as answers about availability of staff with the right skills provided by the EIB survey to capture the skill endowment of the labour force. Besides, the total population or the level of inequality or poverty may also play an important role in determining the volume and the nature of the external demand. Finally, we complete the database with Eurostat information on environmental policies of the EU countries such as the share of renewable energy or the level of gas emissions.

3. Methodology

Building on this large amount of information, we take advantage of big data techniques to assess the relative importance of potential drivers of internationalization. To obtain an unified database, we first deal with the imputation of missing data for the 28 EU members. Depending on the nature of these data (partially or completely missing), two different algorithms are implemented. On the one hand, series where a year observation is missing are imputed using time series technique such as Kalman filtering (section 3.1.1). On the other hand, if the data are unavailable for a given geographical area (country or region) then multivariate imputation methods such as missForest are implemented to make use of the observed link between the

[8] The Country Level Index of Financial Stress (CLIFS) includes six, mainly market-based, financial stress measures that capture three financial market segments: equity markets, bond markets and foreign exchange markets. In addition, when aggregating the sub-indices, the CLIFS takes the co-movement across market segments into account. See Duprey et al. (2017)for more details.

missing variable and the others in the areas where all data are available (section 3.1.2). From there, we model bilateral flows of imports and FDI through a gravity model of trade using a post-lasso OLS which consists in running an OLS on variables selected using a Lasso model.

3.1. Missing values imputation

As explained previously, our data are available at 4 different levels of aggregation: country, macro-sector, sector and region (see Figure 3). On each of the four levels, some data are missing. The first need is thus to impute these missing data. The Figure 4 show the average percentage of missing data for each year and each geographical area. For a given variable and a given geographical area (national or regional), the data can be either partially or totally missing. We address the first case using time serie techniques, while we rely on multivariate imputations for the second one. As a result, for each level, the missing data are imputed according to a specific algorithm (see Algorithm 1). The Figure 5 gives the count of the available observations and the imputed values according to the different techniques.

Algorithm 1 Impute missing values for this level of granularity

```
for serie in this level do
  for area (nuts0 or nuts2) where some data is available do
    if enough observations ($n \geq 3$) and non null variance then
      Impute using Kalman over the period 2015-2021
    else
      Impute using the mean value
    end if
  end for
  for area without available data do
    Impute using missForest with observations of the region with non missing values
  end for
end for
```

3.1.1. Kalman filtering

In this step, we complete missing values of time series taken individually. To do so, we model time series as state-space models based on a decomposition of the series into a number of components (Harvey, 1990). The estimation of such state-space models is then done by a Kalman filtering algorithm. The Figure 6 shows the output of such this algorithm for various level of missingness.

More specifically, in our study, time series are modelled by a state-space model named *local linear model*. This model is defined as follow:

$$x_t = \mu_t + \epsilon_t \,, \qquad \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{1}$$

$$\mu_{t+1} = \mu_t + \eta_t + \xi_t \,, \qquad \xi_t \sim N(0, \sigma_\xi^2) \tag{2}$$

$$\eta_{t+1} = \eta_t + \zeta_t \,, \qquad \zeta_t \sim N(0, \sigma_\zeta^2) \tag{3}$$

where x_t are the observations defined as the sum of a time-varying slope μ_t (unobserved) and a noise ϵ_t of variance σ_ϵ^2. The time-varying slope is made of a random walk of variance σ_ξ^2 which models the trend with and an additional random walk η_t of variance σ_ζ^2 which models the fact that the trend can vary over time. ϵ_t is called the noise of the observations whereas ξ_t and ζ_t are called the noise of the model or the noise of the system. The Kalman algorithm is used to estimate the variance of the three parameters ϵ_t, ξ_t and ζ_t. Once these variances have been estimated, the equation system of the state-space model enables us to complete missing data in the time series.

3.1.2. missForest algorithm

Afterwards, we focus on the imputation of missing data where no observation is available for a given geographical area (or a macro-sector/sector). For this purpose, we implement the algorithm `missForest` (Stekhoven & Bühlmann, 2011), which is based on a random forest predictor (Breiman, 2001). The benefits of using this classifier are numerous such as allowing for interactive and non-linear effect, not relying on strong statistical hypothesis on the data, providing prediction even though data are missing in its inputs (contrary to linear models for instance), gracefully handling mixed data (such as categorical). Importantly, the absence of a-priori statistical hypotheses on the data is a desirable feature of `missForest`. Furthermore, one should note that`missForest` outperforms others standard approach with a decrease of imputation error of 50% in some cases(Stekhoven & Bühlmann, 2011).

This multivariate method consists in predicting the missing values using a random forest trained on the observed parts of the dataset. In other words, it makes use of all the other variables to predict the variable with missing values. At the first iteration of the imputation process, the missing data are imputed to their mean. The process stops as soon as the difference between the newly imputed data matrix and the previous one increases for the first time.

3.2. Trade flow modelling

3.2.1. A machine learning selection using Lasso

For now, we have preselected 82 variables based on our expert judgment. However, we still have an issue for the estimation of imports and FDI. Indeed, in situations where the dimensionality of the data may exceed the length of the sample size, overfitting concerns arise (Hawkins, 2004). In our database, variables are typically available at an annual frequency and available only for few years. In this case, least squares estimation cannot yield unique coefficient estimates and it is necessary to reduce the number of covariates included in the model. Consequently, before plugging all these variables in any model, we decide to go for a variable selection procedure.

Among the different existing methods, we implement Lasso (Least Absolute Shrinkage Selection Operator) regressions to set to zero covariates for which the absolute value of their estimates is lower than a level λ (Tibshirani, 1996). The difference between a Least Squares regression and Lasso regression lies in the optimization problem solved. In fact, the Lasso regression adds a penalty term to the least squares term as follows:

$$\min_{\theta} \sum_{i=1}^{n} \Big(y_i - x_i\theta \Big)^2 + \lambda \, ||\theta||_1 \tag{4}$$

where y_i is the i^{th} observation of the independent variable, x_i denotes the covariates of the i^{th} observation, θ corresponds to the estimates, $||\,.\,||_1$ is the L_1 norm, and λ is the penalty parameter. The penalty parameter helps reducing the number of the covariates included in the model. The optimal λ is determined by cross-validation. The latter refers to a resampling technique which helps to find a parameter value that ensures a proper balance between bias and variance (or flexibility and interpretability).

The cross-validation used is the so-called K-fold cross-validation method that divides the dataset randomly into K different subsets. One subset is kept for validation while the model is estimated over the remaining K-1 subsets. This procedure is repeated for each subset and each λ. The best penalty parameter value is the one yielding the lowest K-fold estimate. In our study, we chose a K-fold cross-validation with K=10 which is the default value for K-fold cross-validation.

Since the Lasso biases the coefficients towards zero, the estimates might not be consistent. This is even more true in presence of highly correlated covariates. Besides, Belloni et al. (2013) have shown that the post-Lasso OLS performs at least as well as the Lasso under mild additional assumptions. We therefore decide to use a two-step estimation procedure in which we regress our variables of interest on the subset of covariates chosen by the Lasso.

3.2.2. Gravity model equation

Once we have selected the most relevant variables according to the Lasso criteria, we then are able to incorporate those variables into a gravity model of trade. The gravity equation in international trade is one of the most robust empirical finding in economics (Chaney, 2018): bilateral trade between two countries is proportional to their respective sizes, measured by their GDP, and inversely proportional to the geographic distance between them. The traditional gravity model applied to bilateral trade flows is the following:

$$X_{ij} = G \cdot \frac{Y_i^{\beta_1} Y_j^{\beta_2}}{D_{ij}^{\beta_3}} \tag{5}$$

Where the trade flow $X_{i,j}$ is explained by Y_i and Y_j that are the masses of the exporting and importing country (e.g. the GDP) and D_{ij} that is the distance between the countries. A logarithmic operator can be applied to form a log-linear model, which yields the following equation[9]:

$$\log X_{ij} = \beta_0 + \beta_1 \log Y_i + \beta_3 \log Y_j + \beta_4 \log D_{ij} + \epsilon_{ij} \tag{6}$$

Additional bilateral variables such as contiguity (the fact that two countries share the same border), common language or regional trade agreement[10] are often include in the equation. In the case of our dataset which includes more information, we estimate two different equations which are the following:

$$\log M_{ij} = \beta_0 + \beta_1 \log Y_i + \beta_3 \log Y_j + \beta_4 \log D_{ij} + \boldsymbol{\beta_5 V_{ij}} + \boldsymbol{\beta_6 Z_i} + \epsilon_{ij} \tag{7}$$

$$\log FDI_{ij} = \beta_0 + \beta_1 \log Y_i + \beta_3 \log Y_j + \beta_4 \log D_{ij} + \boldsymbol{\beta_5 V_{ij}} + \boldsymbol{\beta_6 Z_i} + \epsilon_{ij} \tag{8}$$

where the dependent variable is either M_{ij} and refers to the bilateral flow of imports of country i coming from country j or FDI_{ij} and represents the bilateral flow of FDI of country i coming from country j. Besides, in both equations, $\boldsymbol{Z_i}$ is the matrix of potential drivers of attractiveness related to country i and $\boldsymbol{V_{ij}}$ is the matrix of bilateral variables between country i and j such as contiguity and common language.

[9] Note that constant G becomes part of the β_0. Also for easier interpretation we decide not to invert the log of the distance, contrary to what would be implied by taking logarithm of Equation 5.

[10] Since the countries under interest are in the Eurozone, this variable is *de facto* excluded.

3.2.3. Using different level of granularity

Finally, one last step of data processing is necessary before being able to run regressions. Indeed, since the explained variables are available at the country level only, the explanatory variable available at a more disaggregated level (i.e. macro-sector, sector or NUTS 2) must be aggregated to a country level. To this end, two ways of aggregating values are used: summation and product. The summation is used for the value that are absolute (i.e. not in percentage) whereas the product is used when the value is in percentage.[11] In addition, for a prediction made at a more disaggregated level than the national level, only a subset of covariates are available for this level. Hence, for the variables not available at this level, we take the values of these variables at a more aggregated level. For instance, if the prediction is made at the sector level, the values we use for the variables not available at the sector level are the ones of the corresponding macro-sector. Similarly, if some variables are not available at the macro-sector level, then the values at the national level are retained.

4. Empirical results

4.1. Selected variable by Lasso

The gravity equations 7 and 8 are first estimated using Lasso regression. The selection process for the import equation retains 10 different variables, including the core variables of gravity models such as distance, the contiguity, the common language, the GDPs of the exporting and importing countries and their total population. With regards to the FDI equation, the Lasso regression selects 9 variables, and includes as well the core variables of the gravity model. In those two sets of selected variables, there are 5 common variables. Thus, even though some variables can explain both phenomena (imports and FDI flows), we still have some variables that do not overlap which means that some factors explaining each process are specific.

[11] Since the product is not weighted, note that we assume that each of the sector/macro-sector/region have the same weight.

4.2. Estimates from gravity models

The estimates of the variables selected in the Lasso are shown in Table 3 and Figure 7. In the latter, note that the explanatory variables are all standardized to allow direct comparison of the magnitudes of the effects. Regarding the bilateral FDI estimation in column (1) of Table 3, the three variables with the biggest effects are the GDPs of both countries, the logarithm of the distance and the importance of the air freight of the country attracting FDI. These effects are consistent and significant at the 1% level. An increase in the GDP of the investing country, as well as a decrease in the distance between the two countries lead to higher FDI. Also, better air transports in the receiving country are associated with higher FDI. Unsurprisingly, the coefficients of the GDP and the corporate credit of the country attracting FDI are both positive and significant. Indeed, these are overall demand factors that directly influence the investors decision to invest in a foreign country . In addition, from an investor perspective, higher level of taxation directly reduce financial profitability and thus has negative effect on investment volume. Hence, higher receiving country's taxes on good and services have a negative effect on FDI. Conversely, sharing a common language has a positive impact on those inflows.

Regarding the estimations of imports flows, column (2) shows the same prominence of the core variables of the gravity model (GDPs of both countries and distance) that are all significant at the 1% level. Similarly, the coefficient related to air transports is still positive and significant. Yet, sharing a common language is no more significant while the contiguity variable turns out to increase the global volume of imports. Indeed, importing from a country might require less cultural proximity than investing in the long run through FDI. Instead, trading with a neighbouring country is of major importance even after controlling for the effect of the distance. Other variables related to institutions quality such as solvency rules or reduced time and costs associated with the logistical process of importing goods have also a positive and significant effect on these inflows. Furthermore, the share of corporate non-performing loans, which captures financial fragility, has a negative and significant effect on imports but its magnitude is lower than the previous variables. Finally, turning to the demographic factors, the higher the population of the importing country, the higher the imports.

5. Implementation in BIZMAP

5.1. Software and hardware used

BIZMAP is a web application built within the shiny framework in R. Regarding the user interface, the core package `shiny` combines `shinydashboard` and `shinydashboardPlus` to enhance the user experience. In order to guide the user, a tutorial has been created with the package `rintrojs`. The latter is available on the menu `Help` of the application. Furthermore, some custom CSS and JS scripts have been developed to enhanced style and dynamics of the application. Turning to the web infrastructure, the application has been deployed on an Amazone EC2 instance with the following configuration: variable ECU, 2 vCPU, 2.3 GHz, Intel Broadwell E5-2686v4, 8 Go memory, EBS only. A shiny server has also been installed and configured on the AWS instance to receive the application locally developed.

5.2. Operating instructions

When an user opens the app, he has access to a left menu where he is asked to fill in some information. First, the user has to determine whether he wants to export or make a Foreign Direct Investment for his business as described in Figure 8. Depending on whether the user is interested in current indicators or predictions, he has to choose the period he is interested in within the 2015-2021 period (see Figure 9). Then, the user has to select the country where his company is located as shown in Figure 10. In fact, our indicators rely on geographical distances between this localization and all other EU countries. Finally, the user has to fill in the macro-sector and the sector of his company to obtain results tailored to his business (see Figure 11). All in all, it is possible to choose among a total of 21 macro-sectors and 88 sectors.

Once the left menu is completed, models are running and indicators for each countries and each regions are computed. The application is buffering layers based on the value of the indicators. Values are scaled between 0 and 100 (see Figure 12) and the higher the value, the better the user has interest to export or make a Foreign Direct Investment. Graphically, the most attractive European territories are represented with the warmest colors.

For example, consider the case of a Portuguese SME specialized in the retail of Portuguese wine that intend to export its production in Europe. The firm wants to know where are the best opportunities in Europe for its products so it fills in the information needed in the left menu presented in Figure 13. From there, the application provides the user with a ranking of the Top 10 best countries to export (see Figure 14). The application also enables to have a global view of the scores of all the countries of the European Union (see Figure 15). In our example, the SME should export to France (100), Spain (97.9) and United Kingdom (89.6). Importantly, it is possible to display the score of any country by hovering the mouse over it. In addition, by clicking on a country, the user has access to an in-depth analysis that explains the scores obtained through the visualization of the contribution of each theme to the score (see Figure 16). In the case of our Portuguese entrepreneur, the SME has interest in exporting to France mainly because of a better demography, a higher standard of living, a well developed infrastructure and a strong institutional environment.

The user has also access to even deeper analysis with the *Analytics* menu on top of the app. First, the number of best countries can be selected from 1 to 10 in order to display the ranking of the top countries (see Figure 17). Second, each country score can be broken down into our five different themes (see Figure 18) or the top 8 most impacting variables (see Figure 19). Figure 17 displays the same information available in the table exhibiting the ranking. However, a dashed line is added to represent the average score of all EU countries in order to enable a cross-country comparison between the different scores. Figure 18 displays the same information available on the map by clicking on a country but here, the information is displayed for the top countries all together making any comparison easier. A point for each theme is added to represent the mean contribution of a theme across all countries in order to have a better idea of the significance of the difference between the values. Besides, Figure 19 presents some new analytics. For each country, the contribution of the 8 most impacting variables is presented with the mean contribution of each variables across all countries. Again, this reference point enables to have a better evaluation of any value.

Finally, the application allows to obtain all the previous result at the regional level. The user only has to go back to the map and scroll up to zoom. Then, BIZMAP updates all the predictions to compute the score of territorial attractiveness at the regional level. Returning to our Portuguese example, Figure 20 presents the new ranking at the regional level, while Figure 21 exhibits the scores of all the EU regions. Once again, the user can go to the *Analytics* part to explore the results at the regional level.

6. Conclusion

In this paper, we have built a web application (BIZMAP) that enables SMEs to improve their analysis of foreign markets through the use of an harmonized territorial attractiveness indicator. Many challenges arise from the construction of such application. As a matter of fact, starting from the collection and manipulation of big data provided by 7 different open access databases, we deal with missing values and aggregation issues. From there we start using machine learning methods such as random forest and Kalman filtering. Then, we used a two-steps estimation procedure by combining Lasso regression to select the most relevant variables and a gravity model of trade to determine the most attractive region of Europe. Last but not least, we end up with a synthetic indicator easily readable by SMEs to help them in their decision-making process.

Our datascience pipeline enables us to build a flexible application that covers all the 28 members of the European Union at both a national and regional level, thus providing an indicator about the territorial attractiveness concerning 21 macro-sectors and 88 sectors from 2015 to 2021. Hence, by reducing information uncertainty abroad, BIZMAP is likely to improve the SMEs' analysis of new markets through the visualization of harmonized territorial attractiveness indicators.

This project can still be improved in many ways. More data available at a NUTS 2 level concerning sectors will reduce the number of artificial completions that we made, thus making our models more reliable. Moreover, data more related to the core business of the SMEs would be more useful for them. So, it might be interesting to make an analysis at a deeper sectoral level. Another way to improve the application is to explore the residuals of the gravity models in order to understand which factors are missing. Finally, we plan to extend the application to a wider range of countries, for instance the members of the OECD.

References

Belloni, A., Chernozhukov, V. et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*, 521–547.

Bernard, A. B., Jensen, J. B., Redding, S. J., & Schott, P. K. (2007). Firms in international trade. *Journal of Economic Perspectives*, *21*, 105–130.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brock, W. A., & Durlauf, S. N. (2001). What have we learned from a decade of empirical research on growth? growth empirics and reality. *The World Bank Economic Review*, *15*, 229–272.

Chaney, T. (2018). The gravity equation in international trade: An explanation. *Journal of Political Economy*, *126*, 150–177.

David, R. (1817). On the principles of political economy and taxation. *publicado en*, .

Duprey, T., Klaus, B., & Peltonen, T. (2017). Dating systemic financial stress episodes in the eu countries. *Journal of Financial Stability*, *32*, 30–56.

Eden, L., & Miller, S. R. (2004). Distance matters: Liability of foreignness, institutional distance and ownership strategy. In *" Theories of the Multinational Enterprise: Diversity, Complexity and Relevance"* (pp. 187–221). Emerald Group Publishing Limited.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge university press.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, *44*, 1–12.

Head, K., & Mayer, T. (2002). *Illusory border effects: Distance mismeasurement inflates estimates of home bias in trade* volume 1. Citeseer.

Heckscher, E., & Ohlin, B. (1933). Factor-endowment and factor proportion theory.

Helpman, E., Melitz, M. J., & Yeaple, S. R. (2004). Export versus fdi with heterogeneous firms. *American Economic Review*, *94*, 300–316.

Hollenstein, H. (2005). Determinants of international activities: are smes different? *Small Business Economics*, *24*, 431–450.

Lloyd-Reason, L., Ibeh, K., & Deprey, B. (2009). Top barriers and drivers to sme internationalisation, .

Masson, M. P. R. (2001). *Globalization facts and figures*. 1-4. International Monetary Fund.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, *71*, 1695–1725.

OECD, W. (2015). Inclusive global value chains. policy options in trade and complementary area for gvc integration by small and medium enterprises and lowincome developing countries.

Paul, J., Parthasarathy, S., & Gupta, P. (2017). Exporting challenges of smes: A review and future research agenda. *Journal of World Business*, *52*, 327–342.

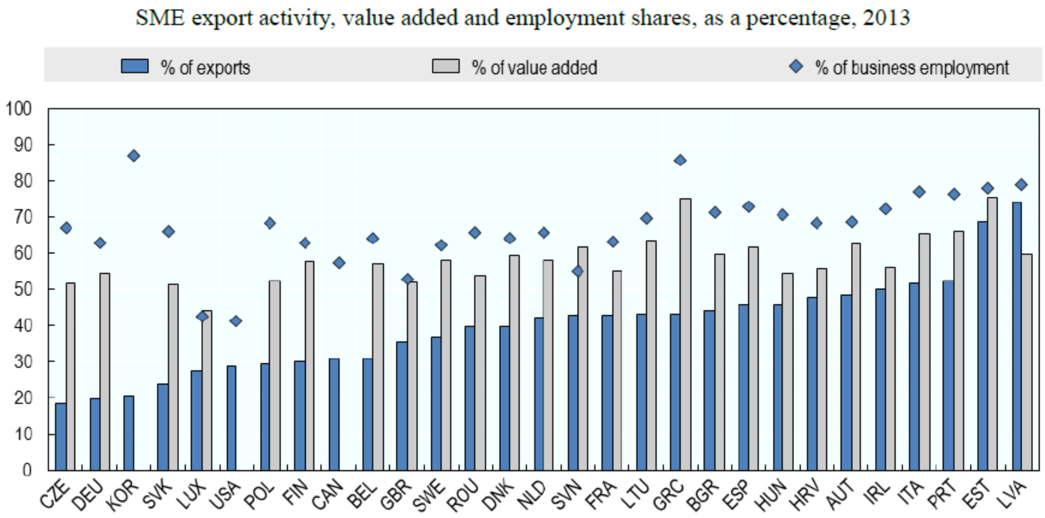
Stekhoven, D. J., & Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*, 112–118.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, *58*, 267–288.

Vamvakidis, A. (1998). Regional integration and economic growth. *The World Bank Economic Review*, *12*, 251–270.

Wagner, J. (2012). International trade and firm performance: a survey of empirical studies since 2006. *Review of World Economics*, *148*, 235–267.

Figure 2: Economic importance of SMEs as compared to their contribution in global trade



Source: OECD Structural and Demographic Business Statistics and Trade by Enterprise Characteristics databases.

Figure 3: The various levels of granularity

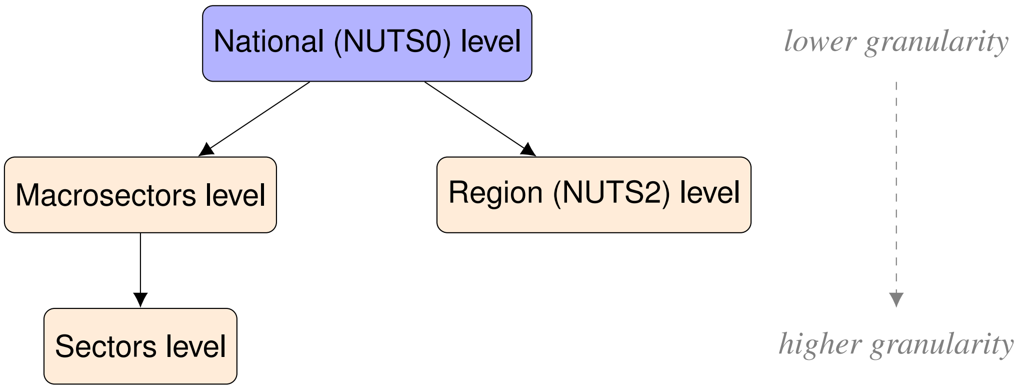


Table 1: List of variables

Variable	Units	Database	Granularity	Theme
Bilateral flows of imports	Million €	Eurostat	Country	Trade
Bilateral flows of FDI	Million €	Eurostat	Country	Trade
Bilateral trade distance (weighted)	Km	CEPII	Country	Trade
Trade contiguity	-	CEPII	Country	Trade
Common official language	-	CEPII	Country	Trade
Total population	Number	Eurostat	NUTS 2 region	Demography and standard of living
Young population	Number	Eurostat	NUTS 2 region	Demography and standard of living
Household income	Million €	Eurostat	NUTS 2 region	Demography and standard of living
Poverty rate	%	Eurostat	NUTS 2 region	Demography and standard of living
Share of renewable energy	%	Eurostat	Country	Demography and standard of living
Income share of the bottom 40%	%	Eurostat	Country	Demography and standard of living
Greenhouse gas emission	Tonnes per capita	Eurostat	Country	Demography and standard of living
Women in senior mangement position	%	Eurostat	Country	Demography and standard of living
High educational level	% of positions	Eurostat	Country	Demography and standard of living
Availability of staff with the right skills : major obstacle	%	European Investment Bank	Macro-sector	Demography and standard of living
Gross Value Added at 2010 prices	Billion €	AMECO	Country	Economic prospects
Gross Value Added at 2010 prices	Billion €	AMECO	Macro-sector	Economic prospects
Harmonised consumer price index	Index	AMECO	Country	Economic prospects
Unemployment rate	%	AMECO	Country	Economic prospects
Private final consumption expenditure	Billion €	AMECO	Country	Economic prospects
Gross fixed capital formation	Billion €	AMECO	Country	Economic prospects
ECU-EUR exchange rates	Number	AMECO	Country	Economic prospects
GDP	Billion €	Eurostat	NUTS 2 region	Economic prospects
Sentiment indicators	Index	Eurostat	Macro-sector	Economic prospects
Consumer Sentiment indicators	Index	Eurostat	Country	Economic prospects
Unemployment rate	%	Eurostat	NUTS 2 region	Economic prospects
Household credit	Billion €	European Central Bank	Country	Economic prospects
NFC credit	Billion €	European Central Bank	Country	Economic prospects
Labor costs index	Index	Eurostat	NACE Rev. 2 activity (2 digit)	Economic prospects
House price index	Index	Eurostat	Country	Economic prospects
R&D expenditures	Billion €	Eurostat	Country	Economic prospects
Expected investment : increase	%	European Investment Bank	Macro-sector	Economic prospects
Share of companies that invest : increase	%	European Investment Bank	Macro-sector	Economic prospects
Demand for product or service : major obstacle	%	European Investment Bank	Macro-sector	Economic prospects
Uncertainty about the future: major obstacle	%	European Investment Bank	Macro-sector	Economic prospects
Turnover	Billion €	Eurostat	NACE Rev. 2 activity (2 digit)	Economic prospects
Wage adjusted labour productivity	%	Eurostat	NACE Rev. 2 activity (2 digit)	Economic prospects
Average personnel costs (personnel costs per employee)	Thousand €	Eurostat	NACE Rev. 2 activity (2 digit)	Economic prospects
Growth rate of employment	%	Eurostat	NACE Rev. 2 activity (2 digit)	Economic prospects
Gross operating rate (gross operating surplus/turnover)	%	Eurostat	NACE Rev. 2 activity (2 digit)	Economic prospects
Investment rate (investment/value added at factors cost)	%	Eurostat	NACE Rev. 2 activity (2 digit)	Economic prospects

Notes : Continues on the next page.

Table 2: List of variables (continued)

Variable	Units	Database	Granularity	Theme
Household debt (% of GDP)	%	Eurostat	Country	Financial conditions
NFC debt (% of GDP)	%	Eurostat	Country	Financial conditions
Public debt (% of GDP)	%	Eurostat	Country	Financial conditions
Core tier one ratio	%	European Central Bank	Country	Financial conditions
Financial stress indicator	Index	European Central Bank	Country	Financial conditions
Household non-performing loans	%	European Central Bank	Country	Financial conditions
Corporate non-performing loans	%	European Central Bank	Country	Financial conditions
Availability of finance : major obstacle	%	European Investment Bank	Macro-sector	Financial conditions
The amount of credit obtained : dissatisfied	%	European Investment Bank	Macro-sector	Financial conditions
The cost of the external finance you obtained : dissatisfied	%	European Investment Bank	Macro-sector	Financial conditions
The collateral required : dissatisfied	%	European Investment Bank	Macro-sector	Financial conditions
Getting electricity	Index	World Bank	Country	Infrastructure
Motorway network	Km	Eurostat	Country	Infrastructure
Railway network	Km	Eurostat	Country	Infrastructure
Air freight	Thousand tonnes	Eurostat	Country	Infrastructure
Ocean freight	Thousand tonnes	Eurostat	Country	Infrastructure
Energy costs	Index	European Investment Bank	Macro-sector	Infrastructure
Access to digital infrastructure : major obstacle	Index	European Investment Bank	Macro-sector	Infrastructure
Availability of adequate transport infrastructure : major obstacle	%	European Investment Bank	Macro-sector	Infrastructure
Voice and Accountability	Index	World Bank	Country	Institutional environment
Trading across borders	Index	World Bank	Country	Institutional environment
Time to import	Index	World Bank	Country	Institutional environment
Starting a business	Index	World Bank	Country	Institutional environment
Rule of Law	Index	World Bank	Country	Institutional environment
Resolving insolvency	Index	World Bank	Country	Institutional environment
Regulatory Quality	Index	World Bank	Country	Institutional environment
Registering property	Index	World Bank	Country	Institutional environment
Protecting minority investors	Index	World Bank	Country	Institutional environment
Political Stability	Index	World Bank	Country	Institutional environment
Paying taxes	Index	World Bank	Country	Institutional environment
Government Effectiveness	Index	World Bank	Country	Institutional environment
Getting credit	Index	World Bank	Country	Institutional environment
Enforcing contracts	Index	World Bank	Country	Institutional environment
Control of Corruption	Index	World Bank	Country	Institutional environment
Social security contributions (% of GDP)	%	OECD tax database	Country	Institutional environment
Tax on corporate profit (% of GDP)	%	OECD tax database	Country	Institutional environment
Tax on payroll (% of GDP)	%	OECD tax database	Country	Institutional environment
Tax on goods and services (% of GDP)	%	OECD tax database	Country	Institutional environment
Labour market regulations : major obstacle	%	European Investment Bank	Macro-sector	Institutional environment
Business regulations : major obstacle	%	European Investment Bank	Macro-sector	Institutional environment

Figure 4: Percentage of missing value as function of year and country

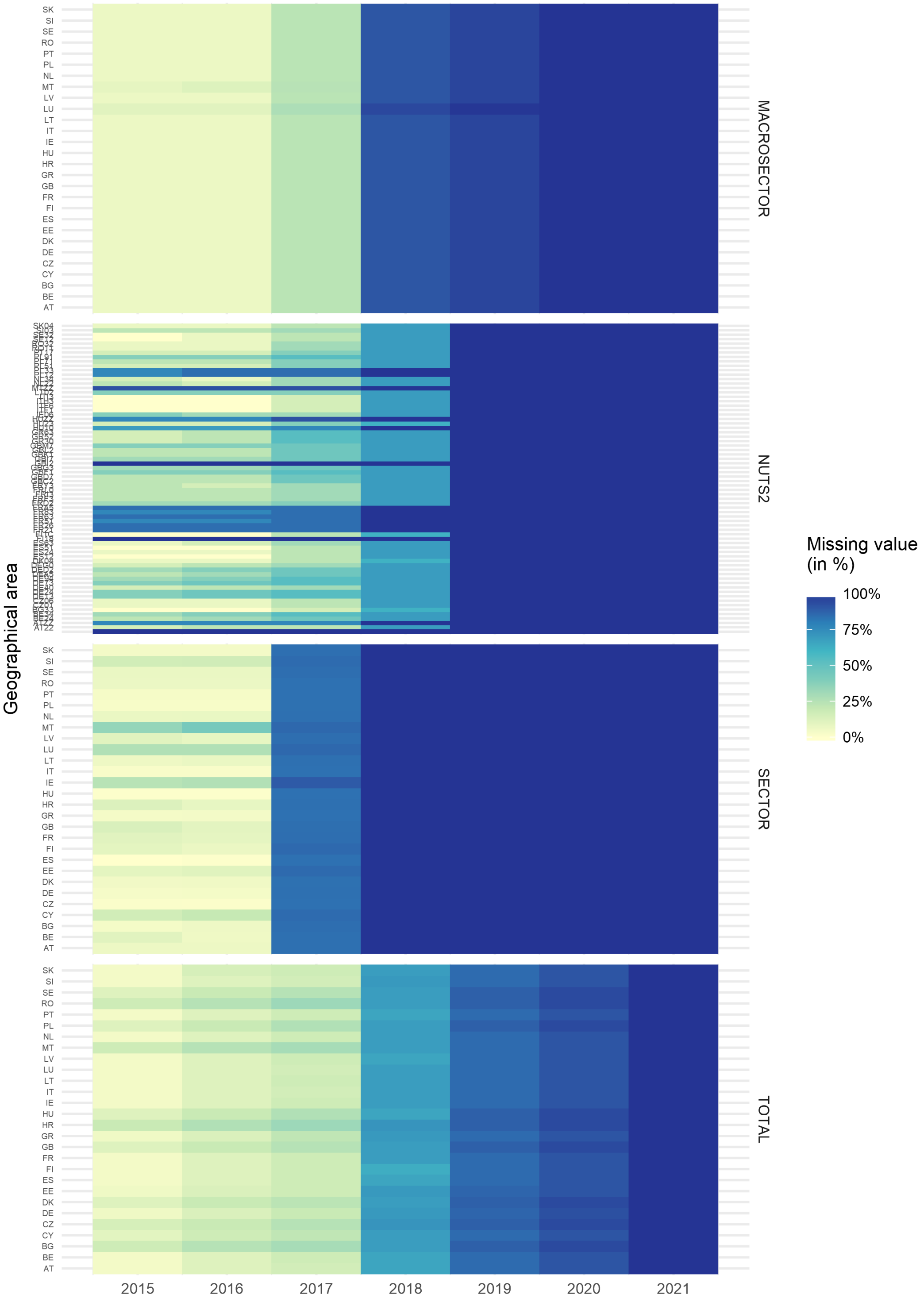


Figure 5: Available and imputed data

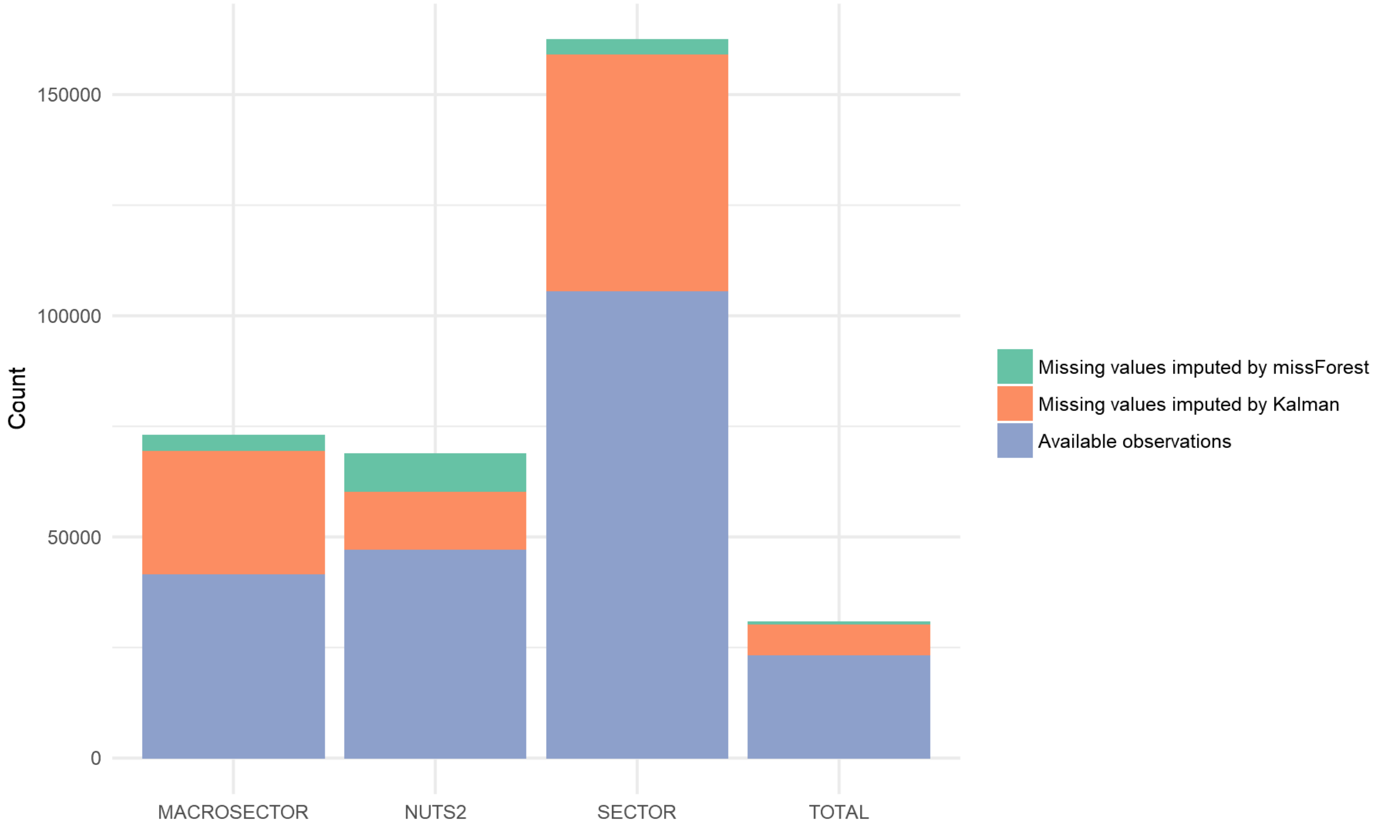
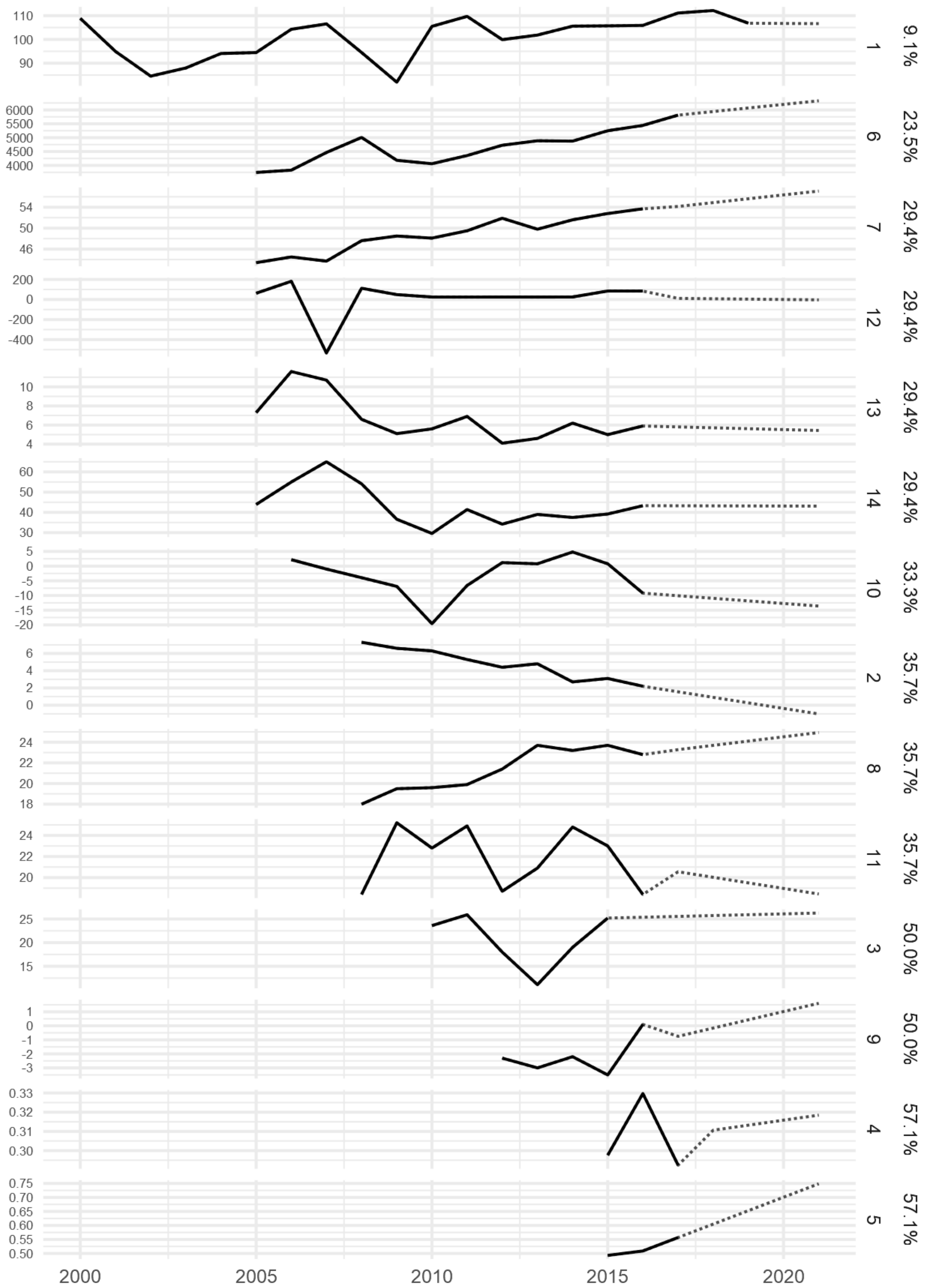


Figure 6: Kalman filtering examples



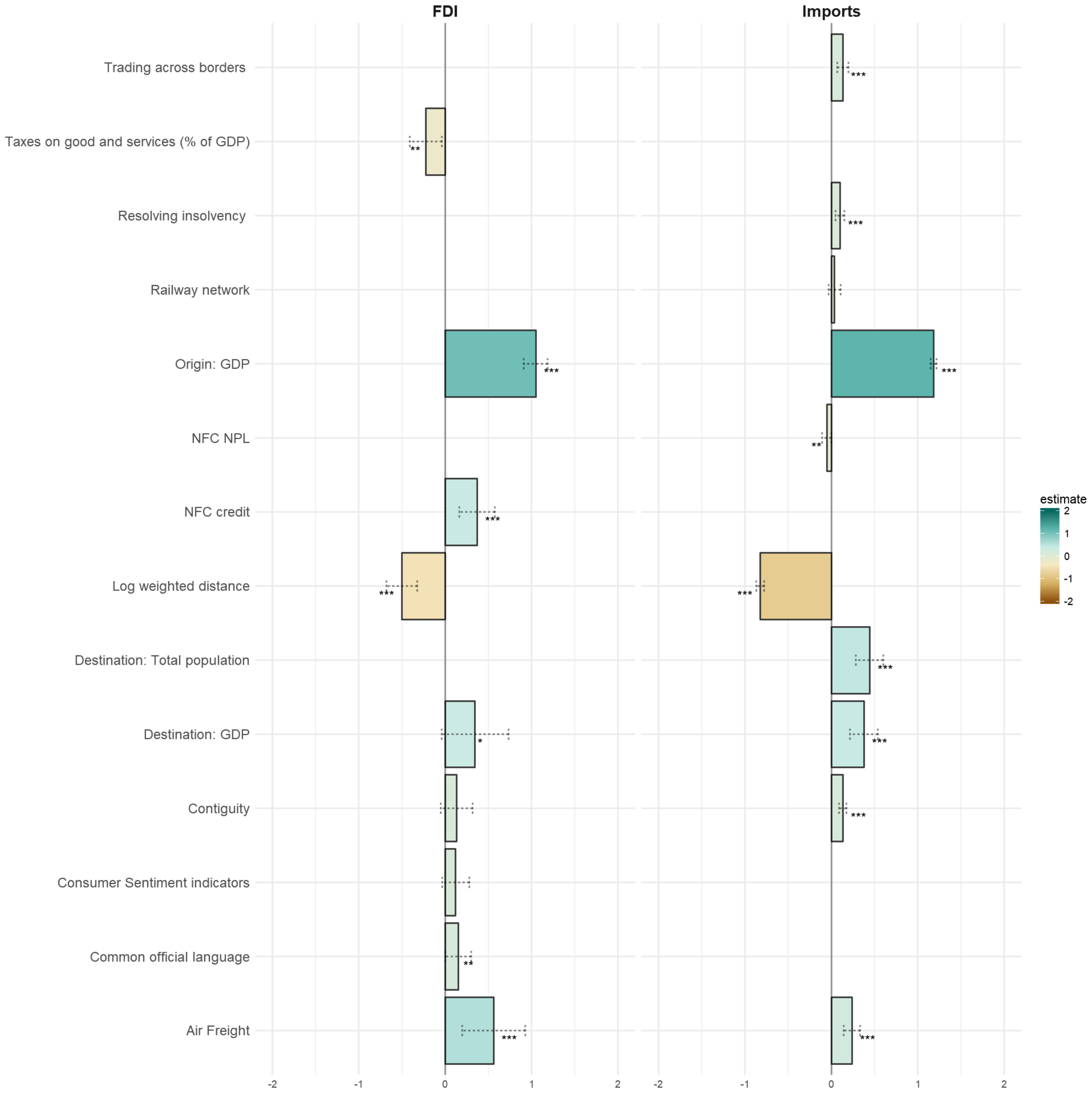
Explanation: The graph show the output of the Kalman filtering for different levels of missing values (from 9.1% to 57.1%). The solid black lines show the original series and the grey dotted line the Kalman filtering extrapolations

Table 3: Estimations of the gravity model of imports and FDI

Model	FDI (1)	Imports (2)
Trade		
Contiguity	0.134	0.132***
	(0.092)	(0.021)
Log weighted distance	-0.501***	-0.824***
	(0.089)	(0.023)
Common official language	0.151**	
	(0.075)	
Economic prospects		
Origin: GDP	1.048***	1.183***
	(0.069)	(0.017)
Destination: GDP	0.347*	0.377***
	(0.194)	(0.081)
Consumer sentiment indicators	0.123	
	(0.078)	
Institutional environment		
Resolving insolvency		0.099***
		(0.025)
Taxes on good and services	-0.224**	
	(0.093)	
Trading across borders		0.133***
		(0.032)
Infrastrucrure		
Air Freight	0.562***	0.238***
	(0.183)	(0.047)
Railway network		0.038
		(0.035)
Financial conditions		
Corporate NPL		-0.054**
		(0.027)
Corporate credit	0.370***	
	(0.103)	
Demography and standard of living		
Destination: total population		0.443***
		(0.079)
Observations	896	1,486
Countries	28	28
R^2	0.441	0.900
Adjusted R^2	0.435	0.899

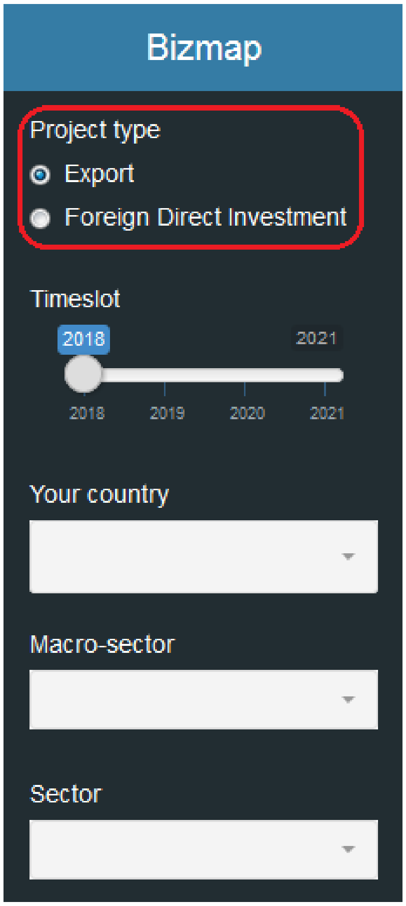
Notes: The table shows the results of equation 7 and 8. All variable definitions are presented in Table 1 and 2.*, ** and *** indicate significance levels at 10%, 5% and 1% respectively.

Figure 7: Normalized gravity model estimations



Notes: The figure shows the results of equation 7 and 8.*, ** and *** indicate significance levels at 10%, 5% and 1% respectively. All explanatory variables are normalized using standardization (i.e variables are centered and reduced) to allow direct comparison of the effects.

Figure 8: Choice between exportation and Foreign Direct Investment



Bizmap

Project type

● Export

○ Foreign Direct Investment

Timeslot

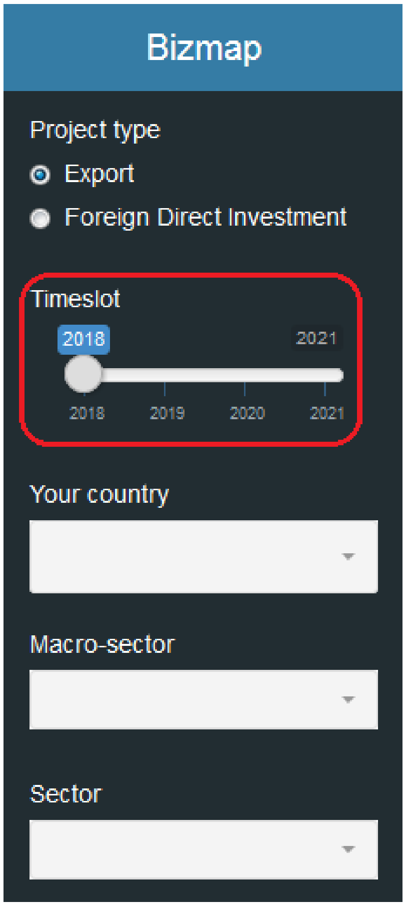
2018 2021

Your country

Macro-sector

Sector

Figure 9: Choice of the period



Bizmap

Project type

● Export

○ Foreign Direct Investment

Timeslot

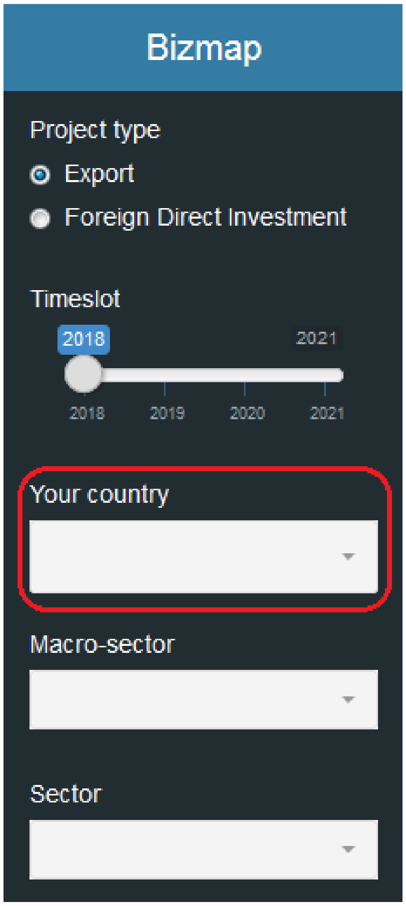
2018 2021

Your country

Macro-sector

Sector

Figure 10: Choice of the country



Bizmap

Project type

◉ Export

○ Foreign Direct Investment

Timeslot

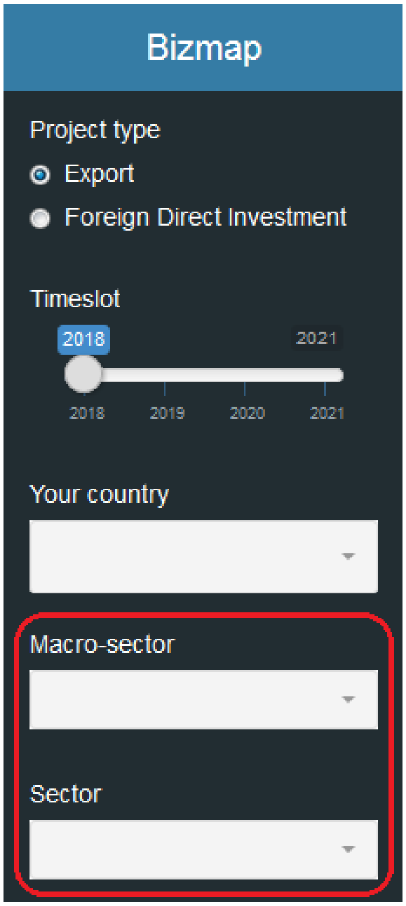
2018 2021

Your country

Macro-sector

Sector

Figure 11: Choice of the macro-sector and sector



Bizmap

Project type

● Export

○ Foreign Direct Investment

Timeslot

2018 2021

Your country

Macro-sector

Sector

Figure 12: Scale of the indicator displayed on the application

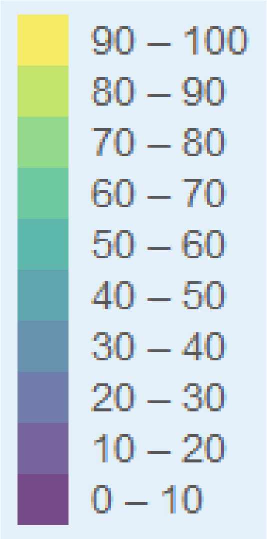
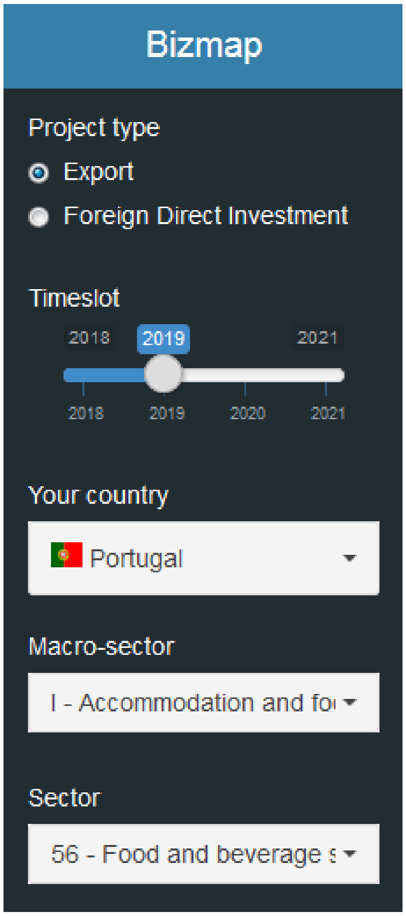


Figure 13: Informations filled by a Portuguese specialized in the retail of Portuguese wine willing to export



Bizmap

Project type

● Export

○ Foreign Direct Investment

Timeslot

2018 2019 2021

Your country

Portugal

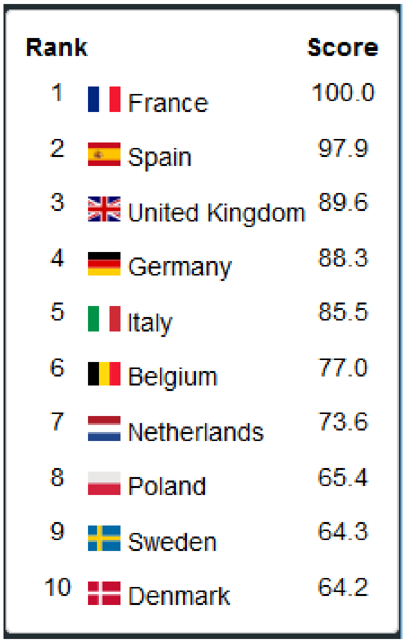
Macro-sector

I - Accommodation and foo

Sector

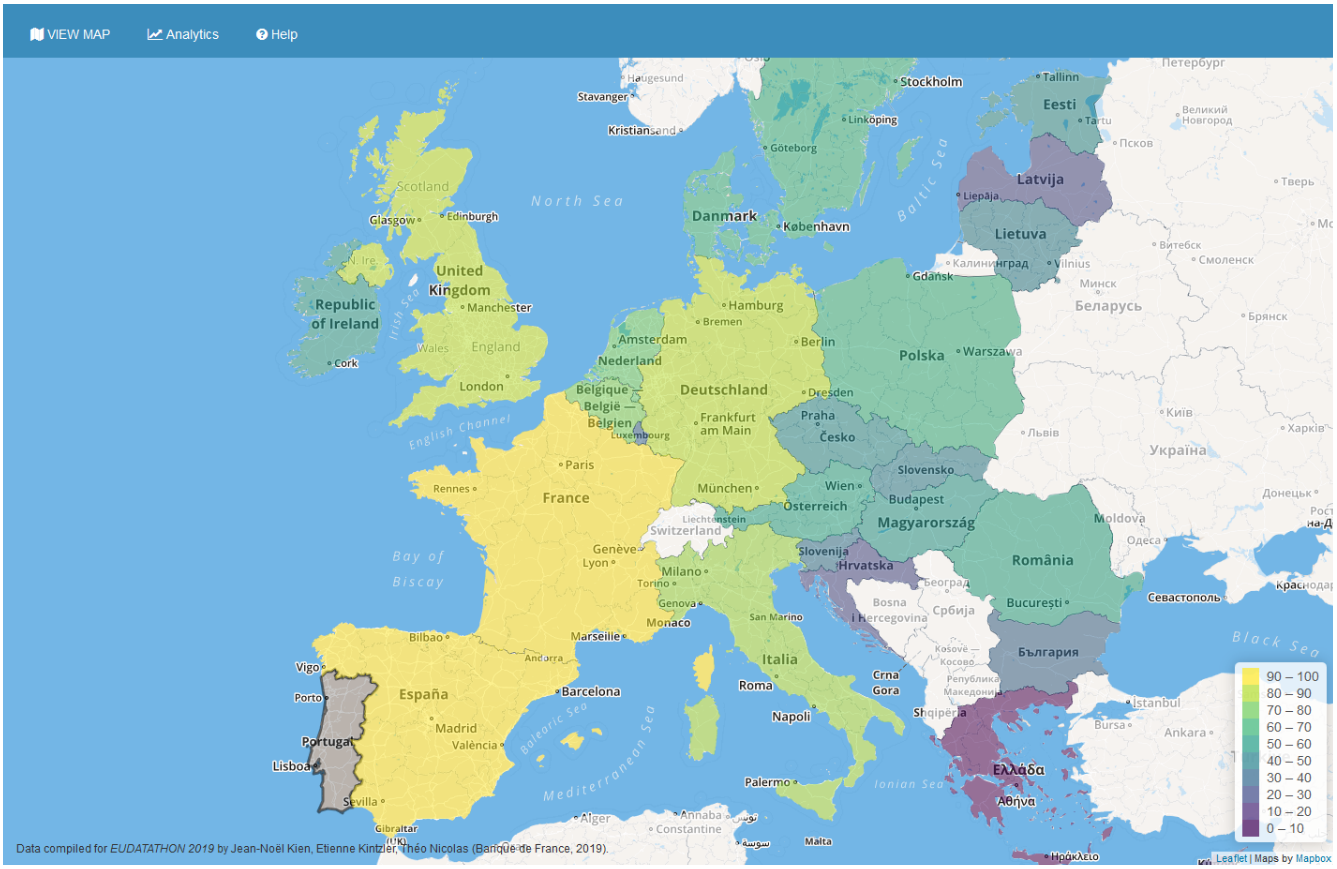
56 - Food and beverage s

Figure 14: Top 10 best countries to export for the Portuguese company



Rank		Score
1	France	100.0
2	Spain	97.9
3	United Kingdom	89.6
4	Germany	88.3
5	Italy	85.5
6	Belgium	77.0
7	Netherlands	73.6
8	Poland	65.4
9	Sweden	64.3
10	Denmark	64.2

Figure 15: Scores of all the countries of the European Union



Data compiled for *EUDATATHON 2019* by Jean-Noël Kien, Etienne Kintzler, Théo Nicolas (Banque de France, 2019).

Figure 16: Contribution of each theme to the score

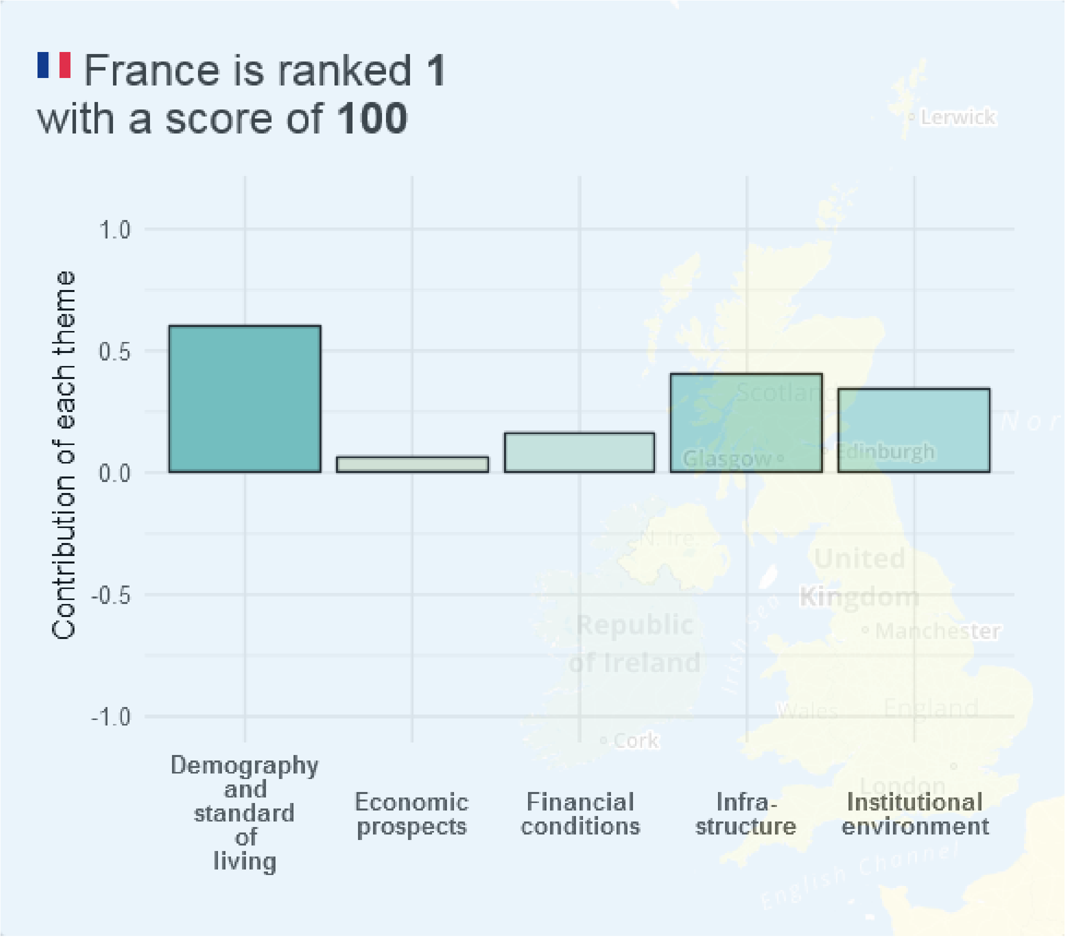


Figure 17: Ranking of the top countries

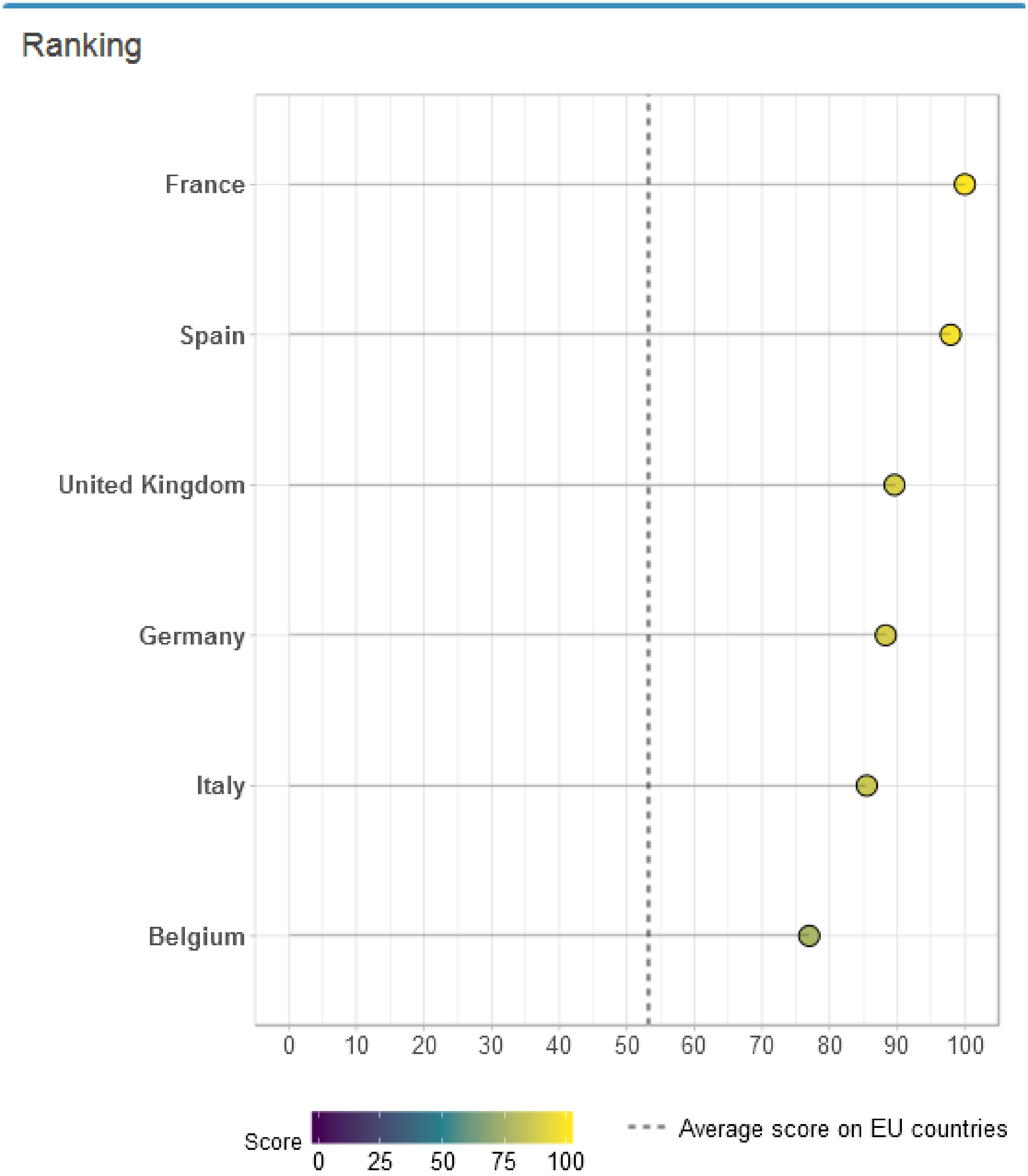


Figure 18: Contribution of each theme to the score

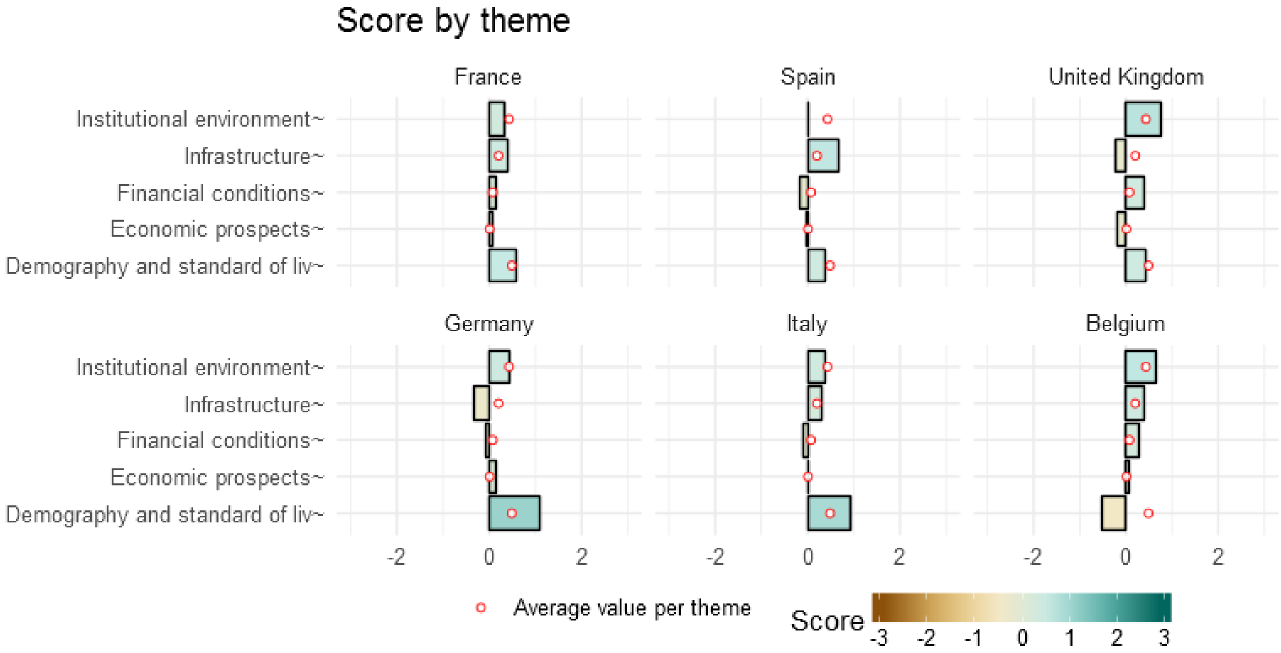


Figure 19: Contribution of each 8 most impacting variables

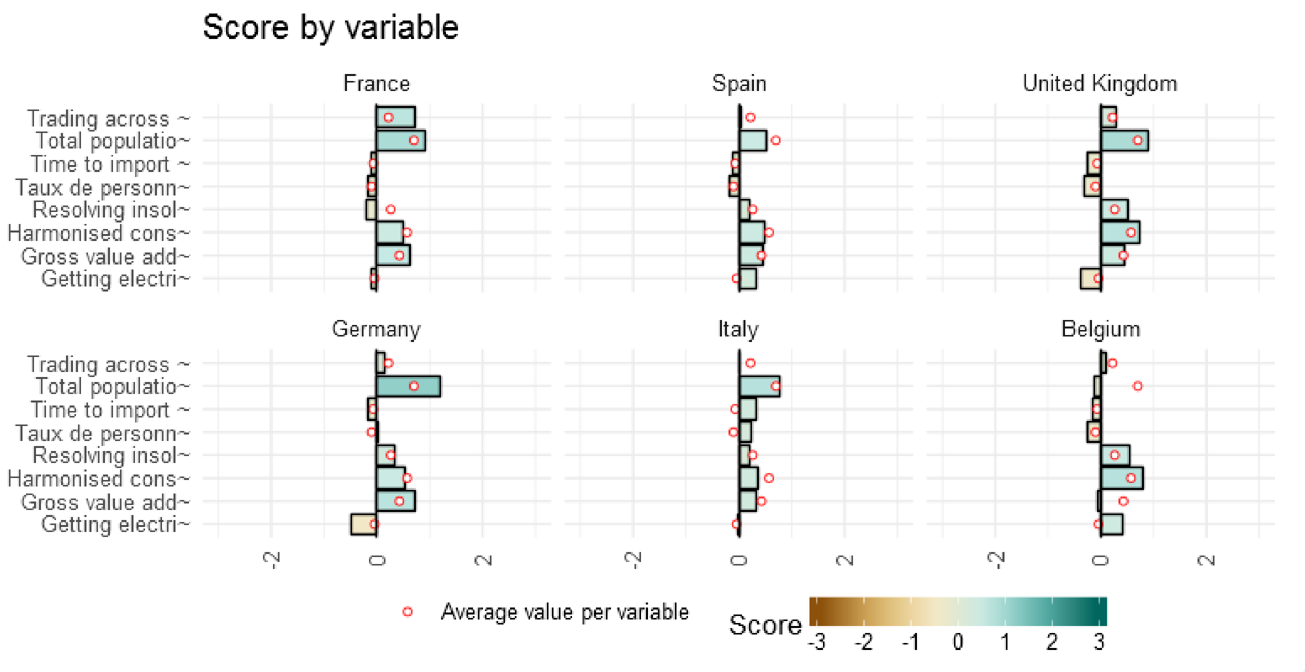
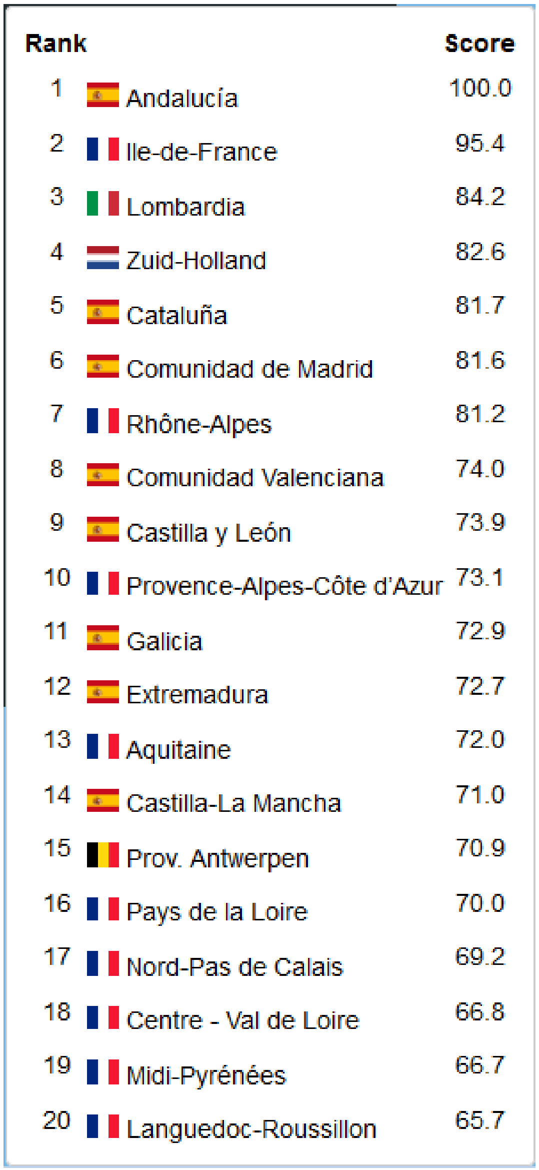
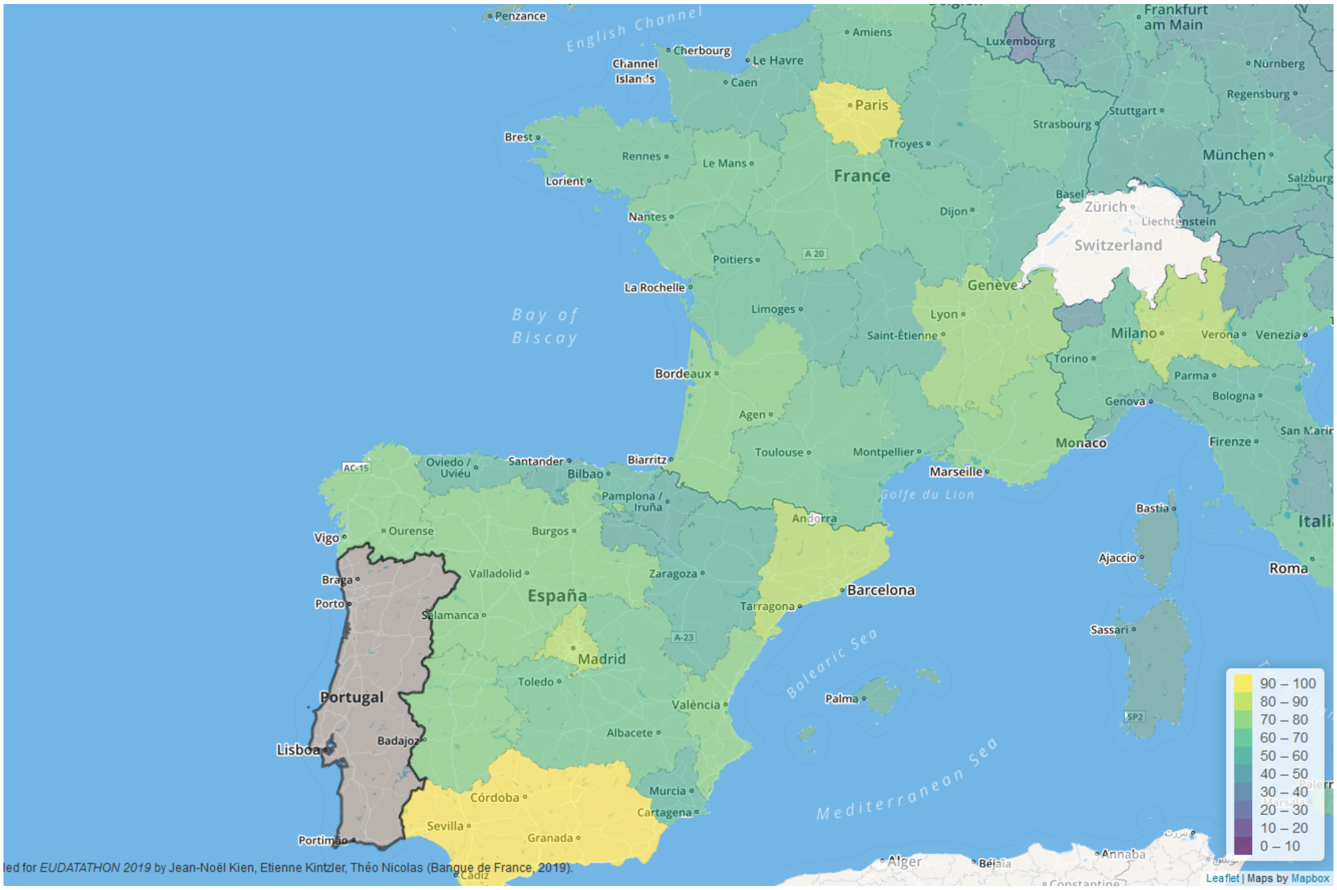


Figure 20: Top 20 best regions to export for the Portuguese company



Rank		Score
1	🇪🇸 Andalucía	100.0
2	🇫🇷 Ile-de-France	95.4
3	🇮🇹 Lombardia	84.2
4	🇳🇱 Zuid-Holland	82.6
5	🇪🇸 Cataluña	81.7
6	🇪🇸 Comunidad de Madrid	81.6
7	🇫🇷 Rhône-Alpes	81.2
8	🇪🇸 Comunidad Valenciana	74.0
9	🇪🇸 Castilla y León	73.9
10	🇫🇷 Provence-Alpes-Côte d'Azur	73.1
11	🇪🇸 Galicia	72.9
12	🇪🇸 Extremadura	72.7
13	🇫🇷 Aquitaine	72.0
14	🇪🇸 Castilla-La Mancha	71.0
15	🇧🇪 Prov. Antwerpen	70.9
16	🇫🇷 Pays de la Loire	70.0
17	🇫🇷 Nord-Pas de Calais	69.2
18	🇫🇷 Centre - Val de Loire	66.8
19	🇫🇷 Midi-Pyrénées	66.7
20	🇫🇷 Languedoc-Roussillon	65.7

Figure 21: Scores of all the regions of the European Union



led for *EUDATATHON 2019* by Jean-Noël Kien, Etienne Kintzler, Théo Nicolas (Banque de France, 2019).

Leaflet | Maps by Mapbox

Fostering European SMEs' internationalization using Big Data : the BIZMAP application

IFC workshop on *Data Science in central banking*

The views expressed in the presentation are the sole responsibility of the authors and do not necessarily represent those of the Banque de France or the Eurosystem. All remaining errors are our own responsibility.

Chloe Brochet Lostie de Kerhor
Yasmine Houri
Jean-Noël Kien
Etienne Kintzler
Lou Richardet

ILIADE-DDSA-DGSI
Banque de France

20 OCTOBER 2021

1. Introduction

2. Overview

3. Methodology

4. Results

5. Perspectives

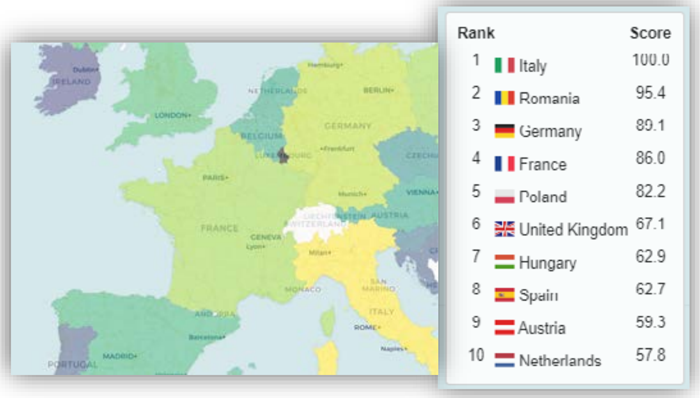
1. INTRODUCTION

Birth: 2019 EU Datathon of the European Commission - 3rd prize

Ambition: to help companies to export (in particular SMEs)

- Economic weight of SMEs (95% of companies, 50% of employment)
- Obstacle to export: lack of resources, especially information (30% of exports)

Solution: identify the attractiveness for exports of European territories



Rank		Score
1	Italy	100.0
2	Romania	95.4
3	Germany	89.1
4	France	86.0
5	Poland	82.2
6	United Kingdom	67.1
7	Hungary	62.9
8	Spain	62.7
9	Austria	59.3
10	Netherlands	57.8

Use cases :

- Visualize a model
- Help a French company to define an export strategy
- Demonstrate France's attractiveness abroad
- Provide quantitative data to experts
- Advise public authorities

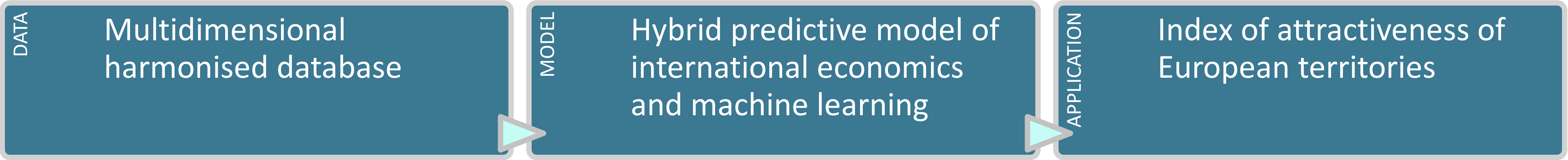
1. Introduction
2. Overview
3. Methodology
4. Results
5. Perspectives

2. OVERVIEW

9 open access data providers

European Commission · eurostat · EUROPEAN CENTRAL BANK · European Investment Bank · THE WORLD BANK IBRD • IDA | WORLD BANK GROUP · OECD · CEPII RECHERCHE ET EXPERTISE SUR L'ÉCONOMIE MONDIALE

TDM Trade Data Monitor · BLAVATNIK SCHOOL OF GOVERNMENT UNIVERSITY OF OXFORD



DATA — Multidimensional harmonised database

MODEL — Hybrid predictive model of international economics and machine learning

APPLICATION — Index of attractiveness of European territories

80 variables divided into 6 themes

- Institutional environment
- Economic perspectives
- Infrastructure
- Standard of living
- Financial conditions
- Geographical and cultural distance

3 dimensions

- Time (years)
- Spatial (countries, regions)
- Sectoral

Imputation of missing values

missForest + Kallman filters

Model

- Gravity model
- Augmented with ML: lasso
- Predictive
- Calibrated on export flows
- Provides attractiveness scores

Index of attractiveness

- Multidimensional, sectoral thanks to the database
- Possibility to look at future attractiveness through the model

Visualisations of the indicator

- Dynamics
- Ranking of territories
- Map with zoom on regions
- Contribution of the variables to the indicator

Need for interpretability and pedagogy

Allows a company to determine a custom export strategy

1. Introduction

2. Overview

3. Methodology

4. Results

5. Perspectives

3. METHODOLOGY IMPUTATION OF MISSING VALUES

🟩	available
🟥	missing
🟦	imputed

	2015	2016	2017	2018	2019	2020
AT	🟩	🟩	🟩	🟩	🟥	🟥
BE	🟩	🟩	🟩	🟩	🟥	🟥
DE	🟩	🟩	🟩	🟥	🟥	🟥
ES	🟩	🟥	🟩	🟩	🟥	🟥
FR	🟩	🟩	🟩	🟩	🟥	🟥
IT	🟩	🟩	🟩	🟥	🟥	🟥
NL	🟩	🟩	🟩	🟩	🟥	🟥

1. Completion **between countries** with missForest*

	2015	2016	2017	2018	2019	2020
AT	🟩	🟩	🟩	🟩	🟥	🟥
BE	🟩	🟩	🟩	🟩	🟥	🟥
DE	🟩	🟩	🟩	missForest	🟥	🟥
ES	🟩	missForest	🟩	🟩	🟥	🟥
FR	🟩	🟩	🟩	🟩	🟥	🟥
IT	🟩	🟩	🟩	missForest	🟥	🟥
NL	🟩	🟩	🟩	🟩	🟥	🟥

2. (For each country) completion **through time** with Kalman filters

	2015	2016	2017	2018	2019	2020
AT	🟩	🟩	🟩	🟩	Kalman filtering	Kalman filtering
BE	🟩	🟩	🟩	🟩	Kalman filtering	Kalman filtering
DE	🟩	🟩	🟩	🟦	Kalman filtering	Kalman filtering
ES	🟩	🟦	🟩	🟩	Kalman filtering	Kalman filtering
FR	🟩	🟩	🟩	🟩	Kalman filtering	Kalman filtering
IT	🟩	🟩	🟩	🟦	Kalman filtering	Kalman filtering
NL	🟩	🟩	🟩	🟩	Kalman filtering	Kalman filtering

MissForest = nonparametric missing value imputation using random forest

3. METHODOLOGY
ESTIMATED EQUATION

- Standard gravity model augmented with economic variables

$$Y_{ij}^{k} = \beta_0 + \sum G_{ij} + GDP_i + GDP_j + \sum X_j^k$$

Y_{ij}^{k} : exports from country ***i*** to country ***j*** for sector ***k***

G_{ijp} & $GDP_{i\ or\ j}$: variables of the standard gravity model

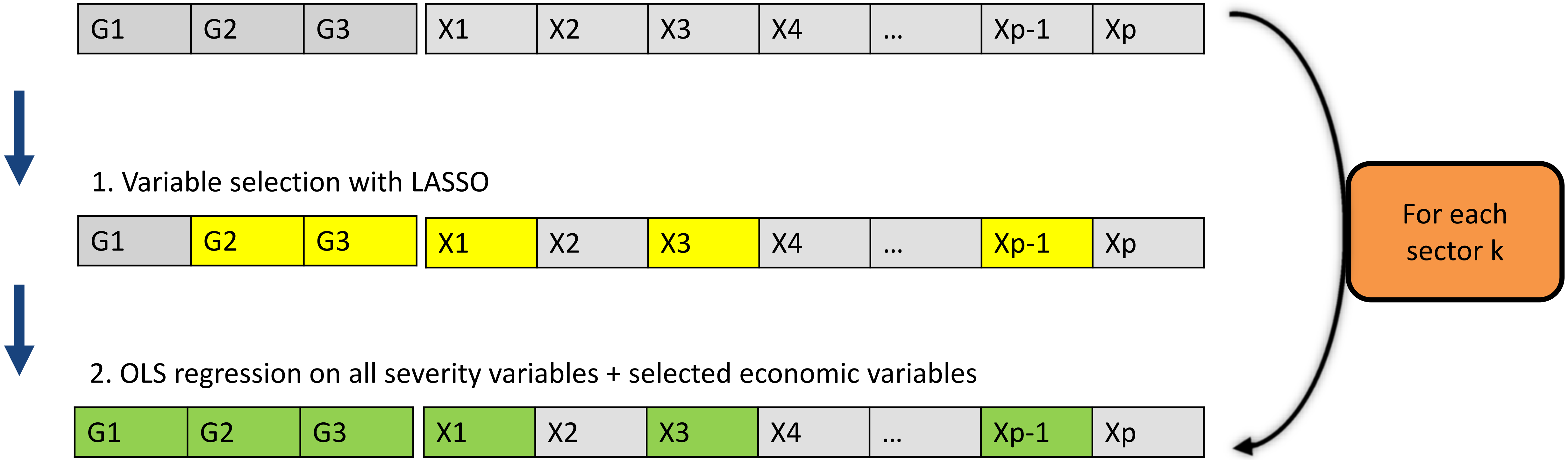
- geographical, bilateral and cultural variables
- economic mass of countries

X_j^k : Economic variables for country ***j*** and sector ***k*** (if available)

- Institutional environment
- Economic outlook
- Infrastructure
- Life standards
- Financial conditions

3. METHODOLOGY
ESTIMATION OLS POST-LASSO BY SECTORS

Gravity model variables (G) + 80 economic variables (X) selected by experts



1. Variable selection with LASSO

2. OLS regression on all severity variables + selected economic variables

⇒ Gravity variables are always selected

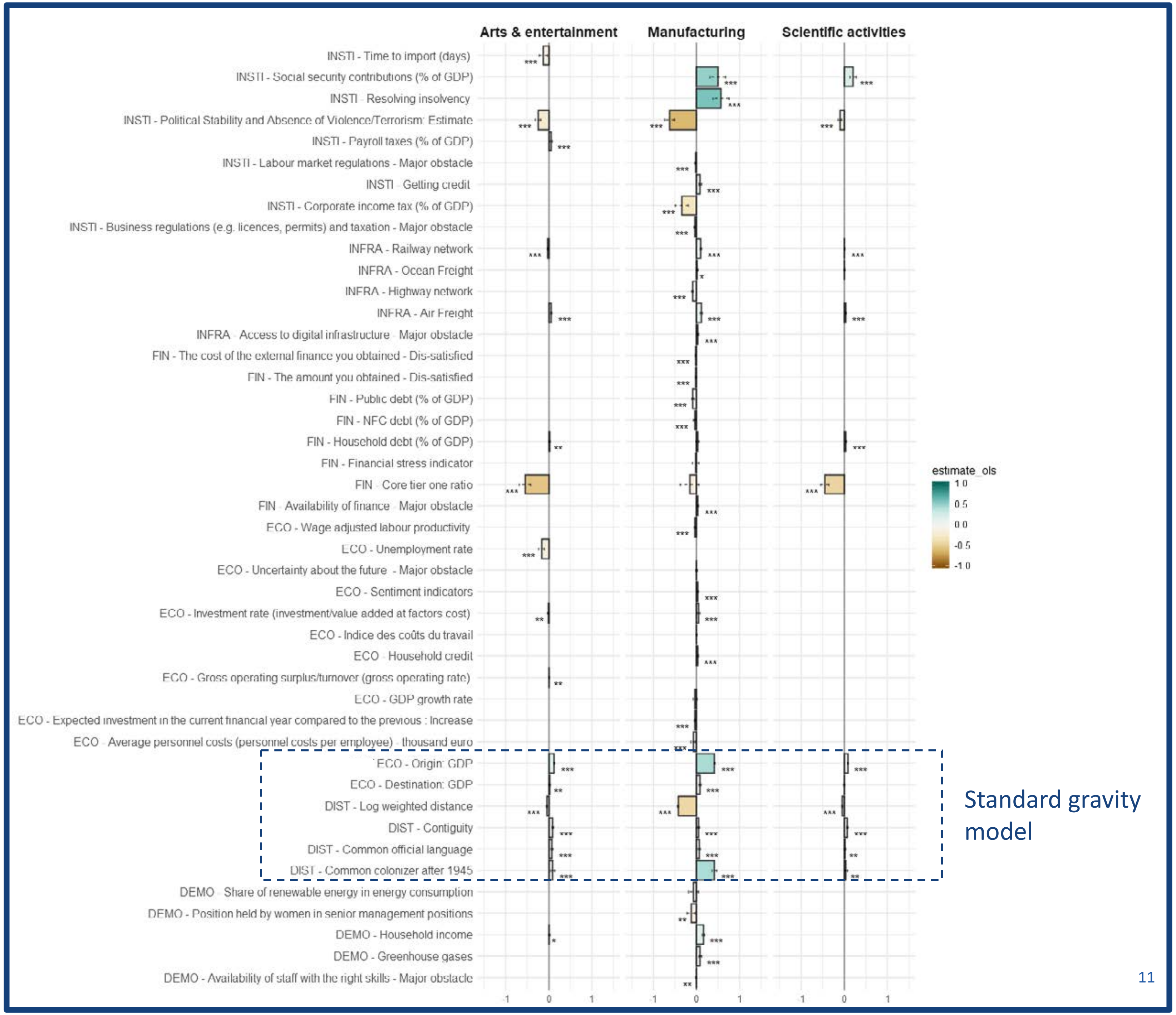
⇒ Different economics variables are selected for different sectors

Reference: *Least squares after model selection in high-dimensional sparse model* (BELLONI and CHERNOZHUKOV, 2013) https://www.jstor.org/stable/23525734?seq=1

1. Introduction
2. Overview
3. Methodology
4. Results
5. Perspectives

4. RESULTS
COEFFICIENTS PER SECTOR

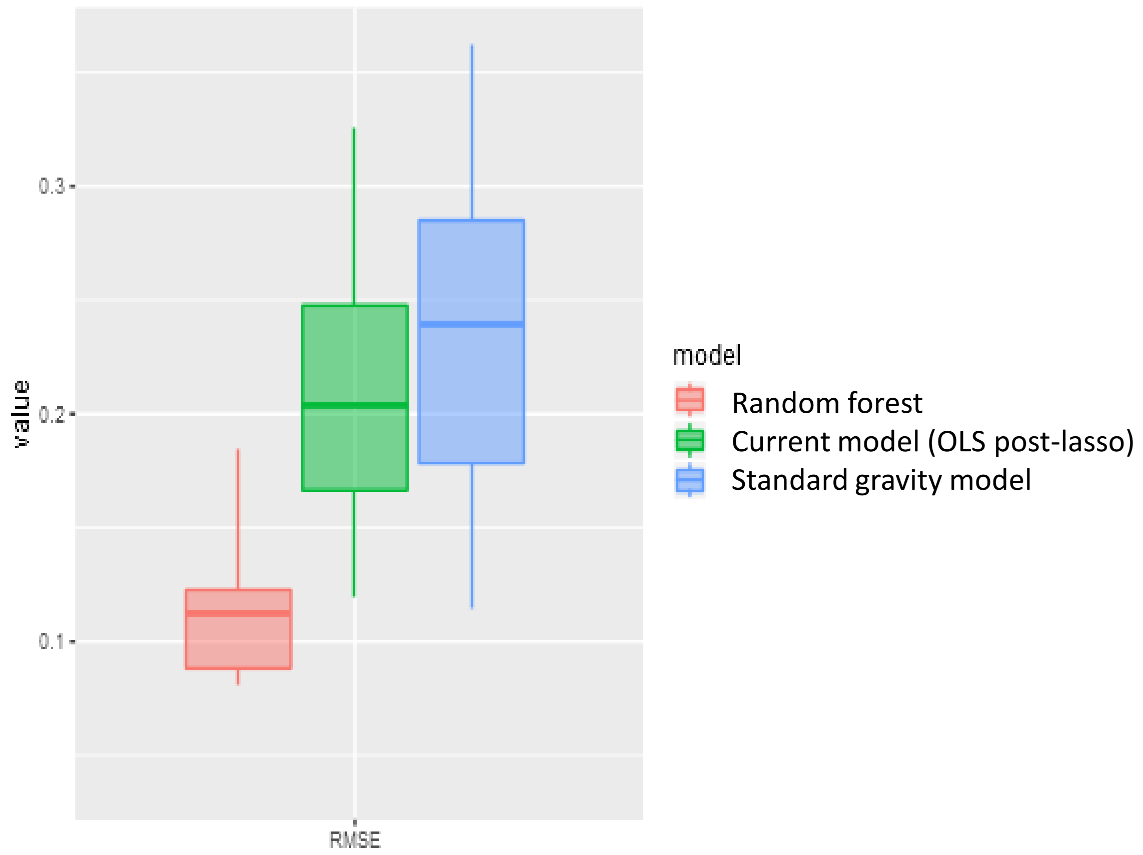
- The manufacturing sector is the biggest macrosector, it includes many subsectors (food, beverage, wood, iron...): many variables are selected
- The arts and sciences are smaller and more specific sectors: fewer variables are selected
- The severity variables are robust



Standard gravity model

4. RESULTS
MACHINE LEARNING IMPROVES PERFORMANCE

Distribution of performances among sectors



- Performance: random forest (RF) >> hybrid model > standard gravity model
- Machine learning (RF) is more accurate but more difficult to interpret:
  - Trade-off between performance and interpretability
  - Specific methods can be used: feature importance, partial dependence plot, shapley value
- Other models can be tested: Gradient boosting (XGBoost, Catboost, etc.)

BANQUE DE FRANCE
EUROSYSTÈME

BANQUE DE FRANCE
EUROSYSTÈME

1. Introduction
2. Overview
3. Methodology
4. Results
5. Perspectives

PROSPECTS

Recently completed project

- Automate the update of data

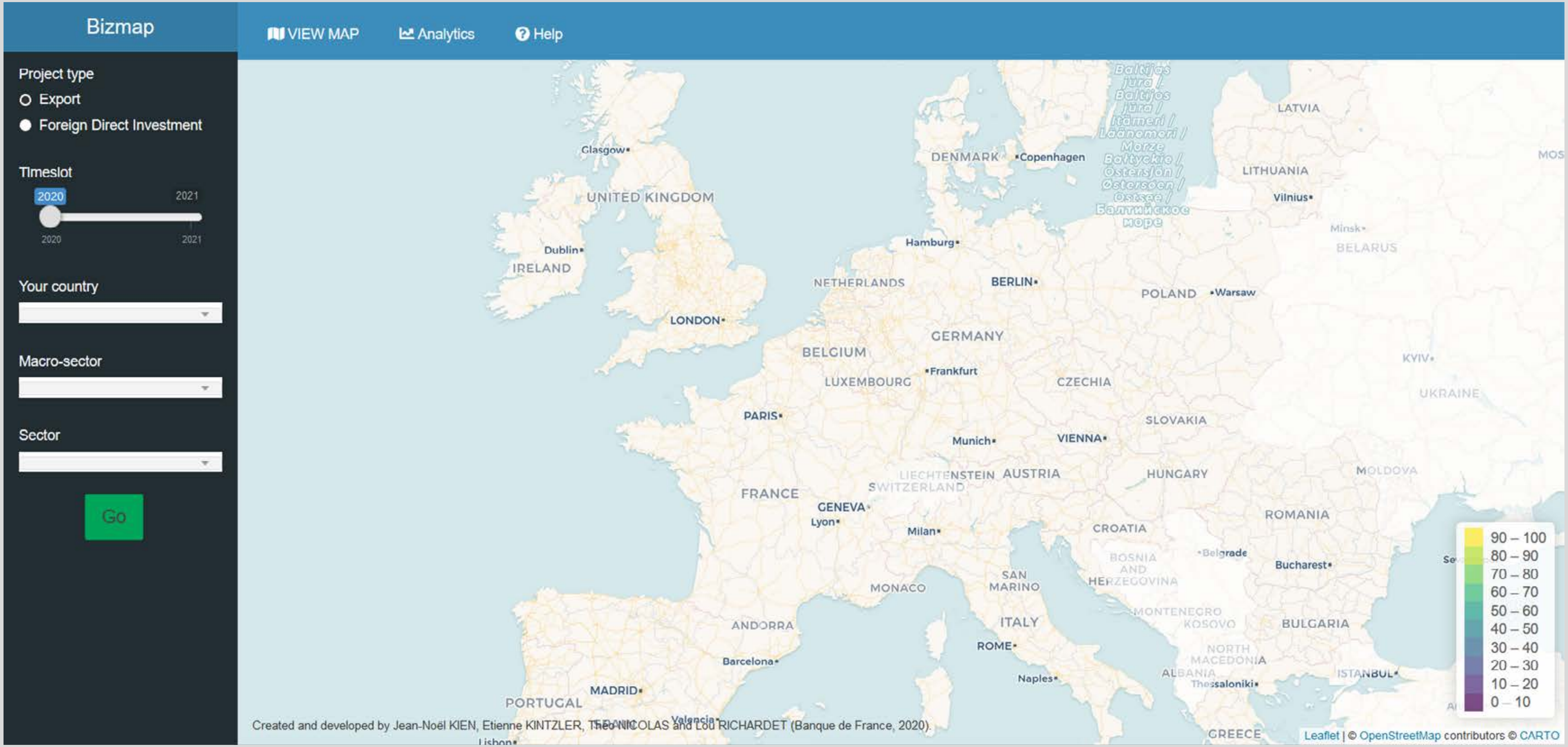
Current work

- Take into account the effects of the health crisis (adapt the model)
- Testing other ML models
- Facilitating the interpretation of attractiveness scores
- Adding information and visualisations to the application
- Exchange with companies and international trade experts to improve the application

Possible developments

- Looking at the attractiveness of French regions to each other
- Create indicators of market potential / business survival

ANNEX : APP DEMO



Bizmap

Project type

○ Export

● Foreign Direct Investment

Timeslot

2020 2021

Your country

Macro-sector

Sector

Go

Created and developed by Jean-Noël KIEN, Etienne KINTZLER, Théo NICOLAS and Lou RICHARDET (Banque de France, 2020).

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on “Machine learning in central banking”

19-22 October 2021, Rome / virtual event

Applications of variational inference in the Bank of Russia[1]

Sergei Seleznev and Ramis Khabibullin,
Central Bank of the Russian Federation

[1] This presentation 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 central banks and other institutions represented at the event.

Bank of Russia

APPLICATIONS OF VARIATIONAL INFERENCE IN THE BANK OF RUSSIA

Ramis Khabibullin
Bank of Russia,* HSE

Sergei Seleznev
Bank of Russia*

19 October 2021
IFC and Bank of Italy Workshop ‘Data Science in Central Banking’
Part 1: Machine Learning Techniques

*The views expressed in the presentation are solely those of the authors and do not necessarily represent the official position of the Bank of Russia. The Bank of Russia is not responsible for the contents of the presentation.

Bayesian estimation of macro models

Vector Autoregressions

- Litterman (1980), Doan, Litterman and Sims (1984), Sims (1993), Villani (2009), Banbura, Giannone and Reichlin (2010), Koop and Korobilis (2010), Giannone, Lenza and Primiceri (2015).

Dynamic Factor Models

- Otrok and Whiteman (1998), Kim and Nelson (1998), Aguilar and West (2000), Blake and Mumtaz (2012).

Dynamic Stochastic General Equilibrium Models

- Smets and Wouters (2003, 2007), Fernandez-Villaverde and Rubio-Ramirez (2007), Justiniano and Primiceri (2008), Herbst and Schorfheide (2015).

Agent Based Models

- Grazzini, Richardi and Tsionas (2017), Gatti and Grazzini (2018), Lux (2018).

Bayesian techniques for intractable posterior

Sampling (asymptotically sample from exact posterior)

- Gibbs Sampling (Casella and Goerge (1992));
- Importance Sampling (Owen (2013));
- Metropolis-Hastings (Chib and Greenberg (1995));
- Hamiltonian Monte Carlo (Neal (2011));
- No-U-Turn Sampling (Hoffman and Gelman (2014));
- Sequential Monte Carlo (Doucet, De Freitas and Gordon (2001));
- PDMP Continuous-Time Monte Carlo (Fearnhead et al. (2016)).

Optimization (allows to estimate larger models)

- MAP estimation;
- Expectation Propagation algorithm (Minka (2001));
- Variational Bayes estimation (Wainwright and Jordan (2008));
- α-divergense (Li and Turner (2016)).

What does VB offer?

Approximate posterior

- Chooses the posterior approximation from a family of distributions that allows independent sampling.

Solving via optimization

- Finds an approximate posterior minimizing KL divergence between posterior and approximation family;
- Allows achieving the same accuracy faster than MCMC methods;
- Allows estimating posterior simultaneously maximizing the marginal likelihood with respect to hyperparameters.

Large-scale applications

- Estimates models with thousands/dozens of thousands of parameters on Desktop PC.

Variational Bayes

A VB algorithm maximizes the lower bound of the logarithm of the marginal likelihood with respect to an approximate density and hyperparameters:

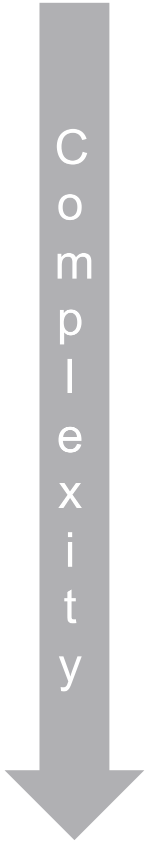
$$\log p(y|x,\varphi) = \log\int p(y,\theta|x,\varphi)d\theta = \log\int \frac{p(y|\theta,x,\varphi)p(\theta|\varphi)}{q(\theta)}q(\theta)d\theta \geq$$

$$\int (\log p(y,\theta|x,\varphi) - \log q(\theta))q(\theta)d\theta =$$

$$\log p(y|x,\varphi) - \int (\log q(\theta) - \log p(\theta|y,x,\varphi))q(\theta)d\theta =$$

$$\log p(y|x,\varphi) - KL\big(q(\theta)||p(\theta|y,x,\varphi)\big) \ = ELBO(q,\varphi)$$

Different approximate densities might be useful in different situations


Complexity

Mean-field approximations (Independent Gaussian approximation)

- All components are independent distributions (Wainwright and Jordan (2008)).

Gaussian approximation

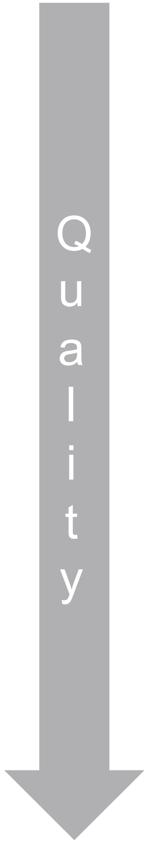
- Gaussian distribution (Tan, Bhaskaran and Nott (2019)).

Neural network (Normalizing flows)

- Simple distribution is passed through a neural network (Rezende and Mohamed (2015)).

Non-parametric approximation (Stein variational inference)

- Approximate functional space optimization (Liu and Wang (2019)).


Quality

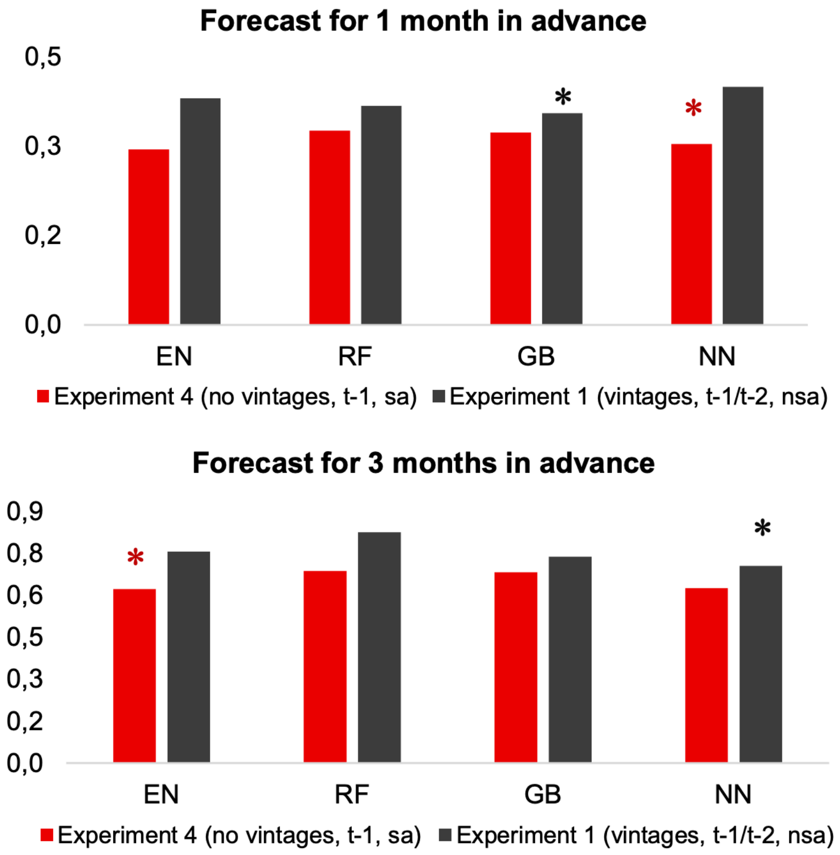
Sparse Bayesian neural network for inflation forecasting

Neural network

- Highly non-linear and flexible model;
- Sparse Bayesian regularization (Tipping (2001)) to avoid overfitting and minimize cross-validation dimensionality.

Model performance

- Bayesian neural network is usually comparable with or better than other ML models for inflation prediction;
- Due to a sparse structure, it can be easy to find the most important features (Khabibullin and Seleznev (2020)).



Forecast for 1 month in advance

Forecast for 3 months in advance

RMSFEs of inflation forecasting for different ML models developed by Mamedli and Shibitov (2021)

**Mean-field variational approximation is used.*

Seasonal adjustment model for financial flows

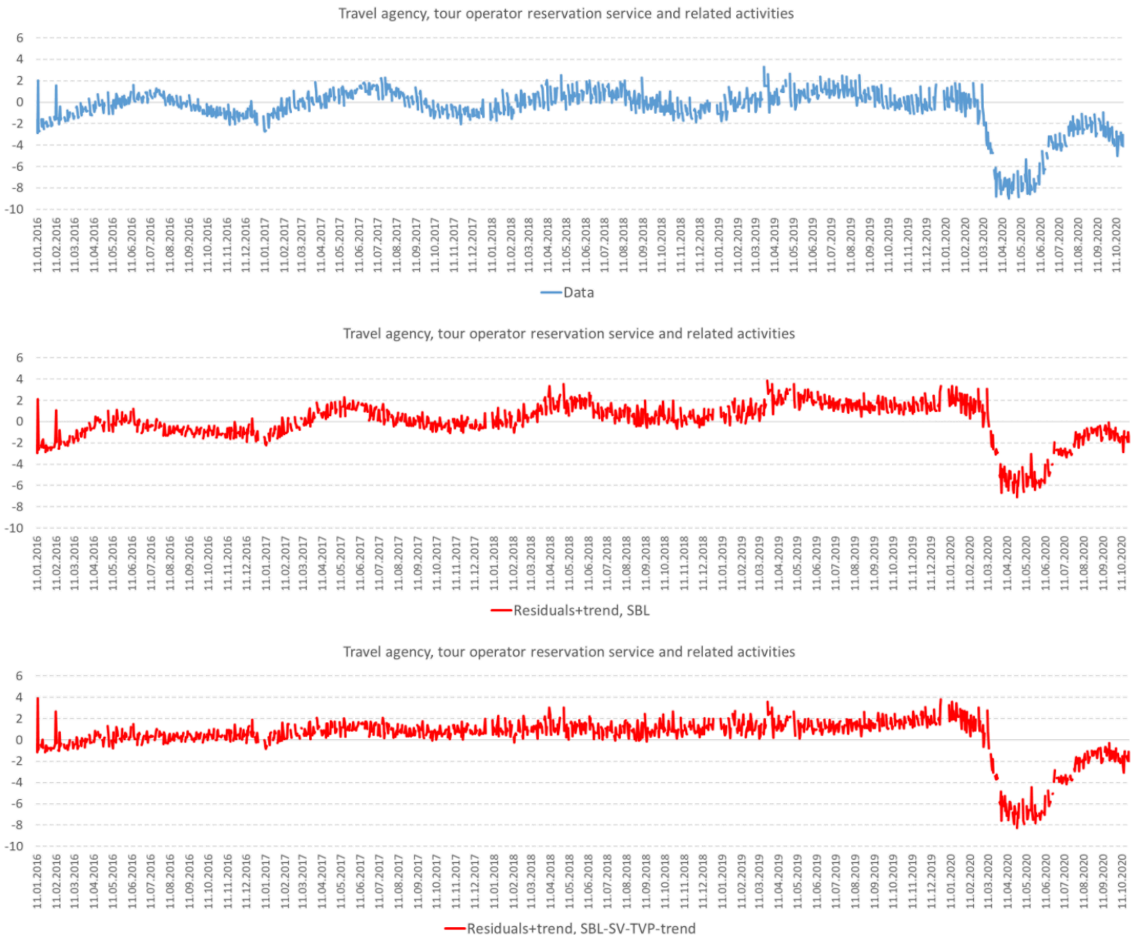
Financial flow data

- A financial flow report which helps monitor real-time industry inflows and outflows was created at the start of the COVID-19 crisis;
- Existence of daily, weakly and monthly seasonality greatly complicates the understanding of reasons behind changes in indicator dynamics.

Seasonal adjustment

- Hundreds of series are smoothed every day;
- Models without flexible trends (RW trend with SV) show poor performance for some series.

**Mean-field and normalizing flows variational approximation is used. For uncertainty estimation, Stein VI produces significantly more accurate results.*



Daily seasonal adjustment during the COVID-19 crisis using models with and without a flexible trend proposed by Khabibullin et al. (2020)

Incorporating economic judgment into nonlinear models

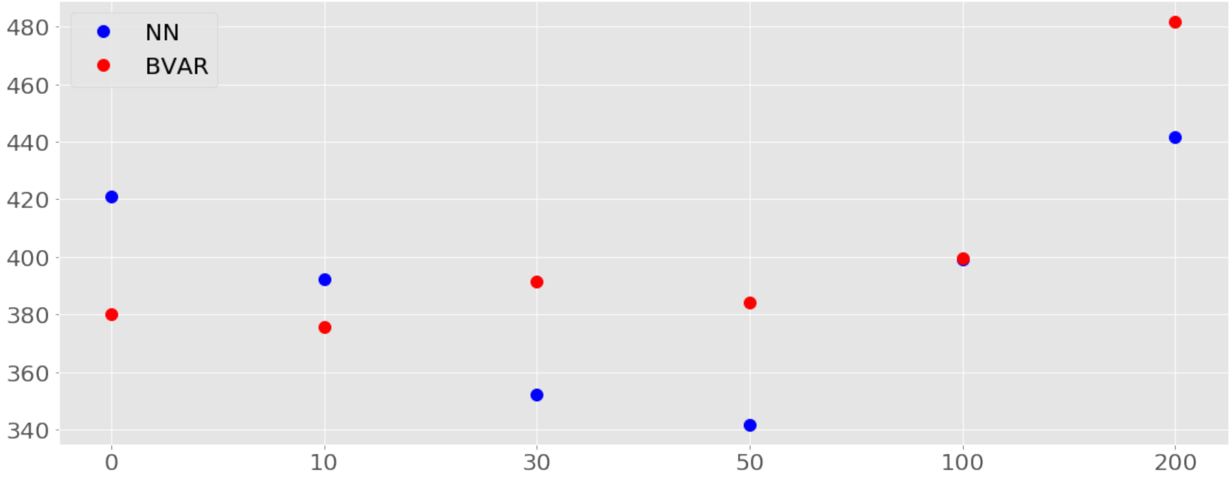
Guided non-structural model

- An extension of the idea suggested by Del Negro and Schorfheide (2004);
- Real data and artificial data from a structural model are used for the estimation of a non-structural model;
- The final algorithm for the joint structural and non-structural (hyper)parameters estimation is a double variational inference or ADVIL (Li et al. (2019)).

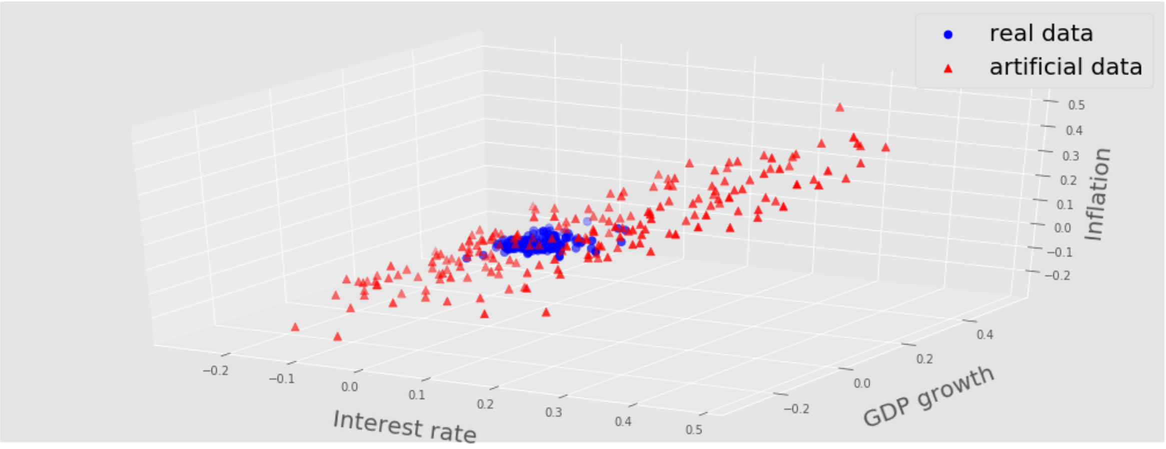
Results

- Economic judgment from the DSGE model helps to improve BNN model results.

**Mean-field and normalizing flows variational approximations are used.*



NLL for a different choice of the number of artificial points



Real and artificial points for BNN

Conclusion

- Variational inference helps in the estimation of models that cannot be estimated via classical Bayesian techniques;
- Our experience shows that even poor mean-field approximation is usually sufficient for practical applications.

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Deep learning solutions for dynamic stochastic general equilibrium models[1]

Mo Ashtari and Vladimir Skavysh,
Bank of Canada

[1] This presentation 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 central banks and other institutions represented at the event.

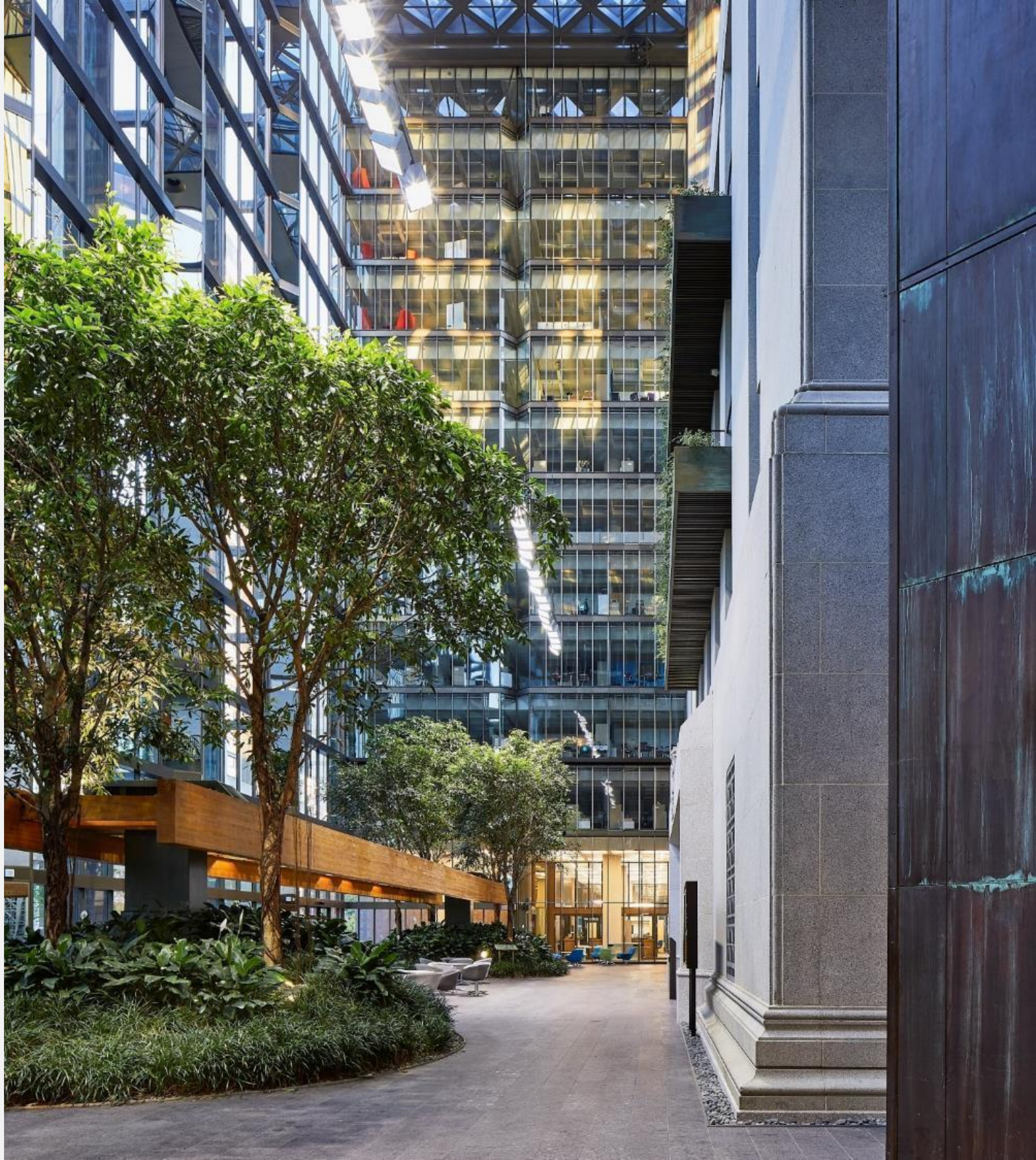
BANK OF CANADA
BANQUE DU CANADA

OCTOBER 2021

Deep Learning Solutions for Dynamic Stochastic General Equilibrium Models

Mo Ashtari, Vladimir Skavysh

The views expressed in this presentation are solely those of the authors and may differ from official Bank of Canada views. No responsibility for them should be attributed to the Bank.



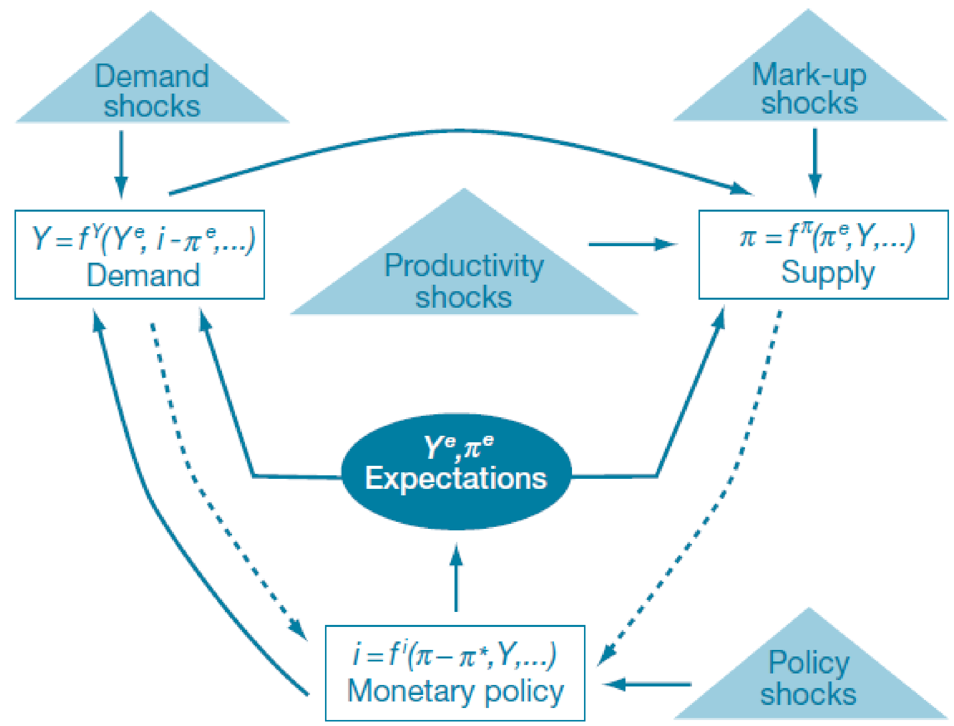
DSGE Models

❑ Dynamic: there are intertemporal problems and agents rationally form expectations

❑ Stochastic: exogenous stochastic processes may shift aggregates

❑ General Equilibrium: all markets are in equilibrium, although unpredictable shocks disturb this equilibrium for a while

The Basic Structure of DSGE Models



Deep Learning

❑Machine Learning has shown to be a promising solutions in many fields as well as economics.

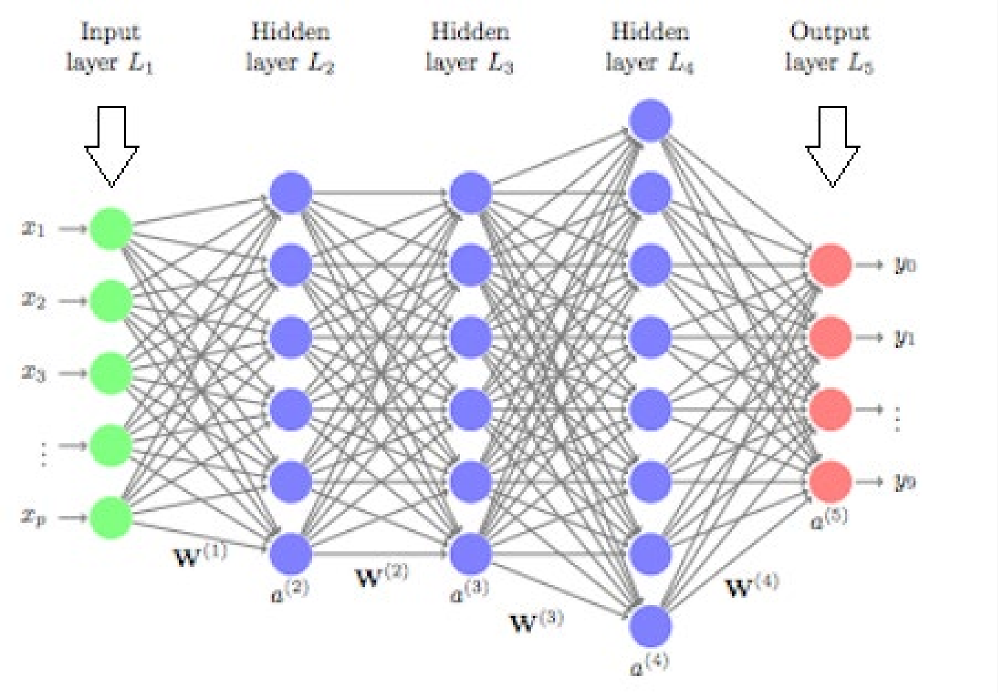
❑ Universal Approximation Theorem: A neural network with at least one hidden layer can approximate any Borel measurable function mapping finite-dimensional spaces to any desired degree of accuracy. (Hornik, Stinchcombe, and White, 1989)

❑ Breaking the curse of dimensionality: A one-layer neural network achieves integrated square errors of order $O(1/M)$, where M is the number of nodes. In comparison, for series approximations, the integrated square error is of order $O(1/(M^{2/N}))$ where N is the dimensions of the function to be approximated. (Barron 1993)

❑ Here, we used deep learning methods to solve DSGE models, starting with the simple Neoclassical growth model

❑ When applied to DSGE models, neural networks (NN) offer the following advantages:

- ❑ Ability to solve high dimensional problems without the curse of dimensionality
- ❑ High approximation power outside of the steady state



The Neo-Classical Growth Model

❑ In an economy with a representative household

$$\max_{c_t,k_{t+1}} E\left(\sum_{t=0}^{\infty} \beta^t u(c_t)\right)$$

where c_t is the consumption and $\mathrm{k_t}$ is the capital at time t, $E$is the expectation operator, β is the discount factor, and u is the utility function. The value function is defined by the Bellman operator:

$$V(k_t, z_t) - \max_{k_{t+1}}[u(e^{z_t}k_t^{\alpha} + (1-\delta)k_t \; - k_{t+1}) + \; \beta V(k_{t+1}, z_{t+1})] = 0$$

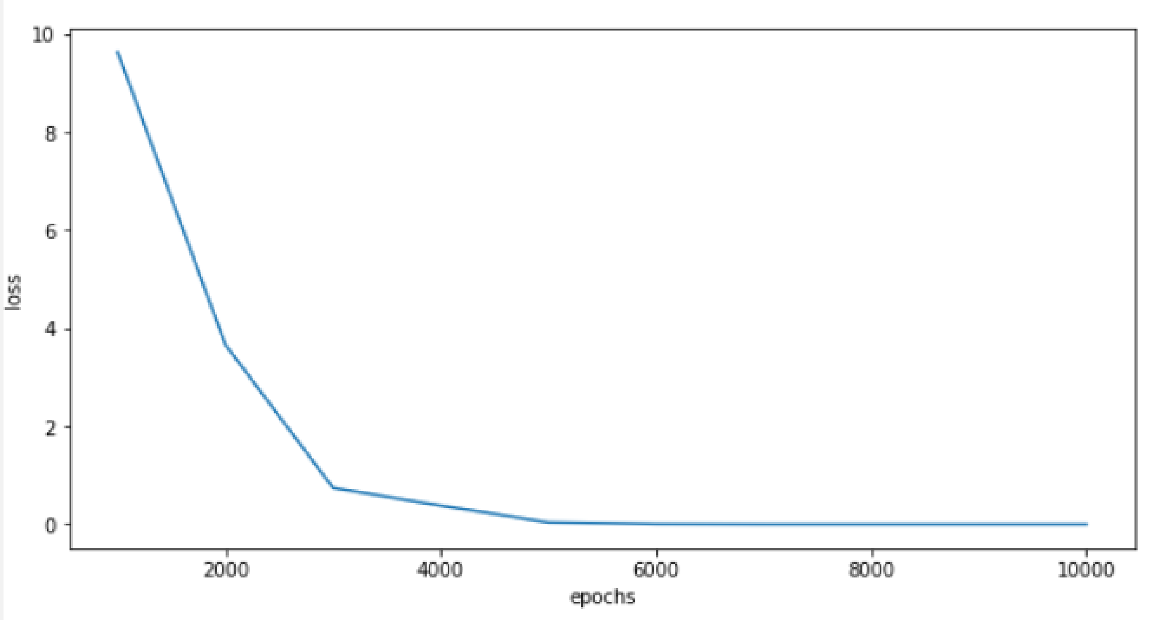
where δ is the depreciation rate, and k and z are two state variables of the economy, capital and productivity

❑ We use deep learning to solve for the value function using the following steps:

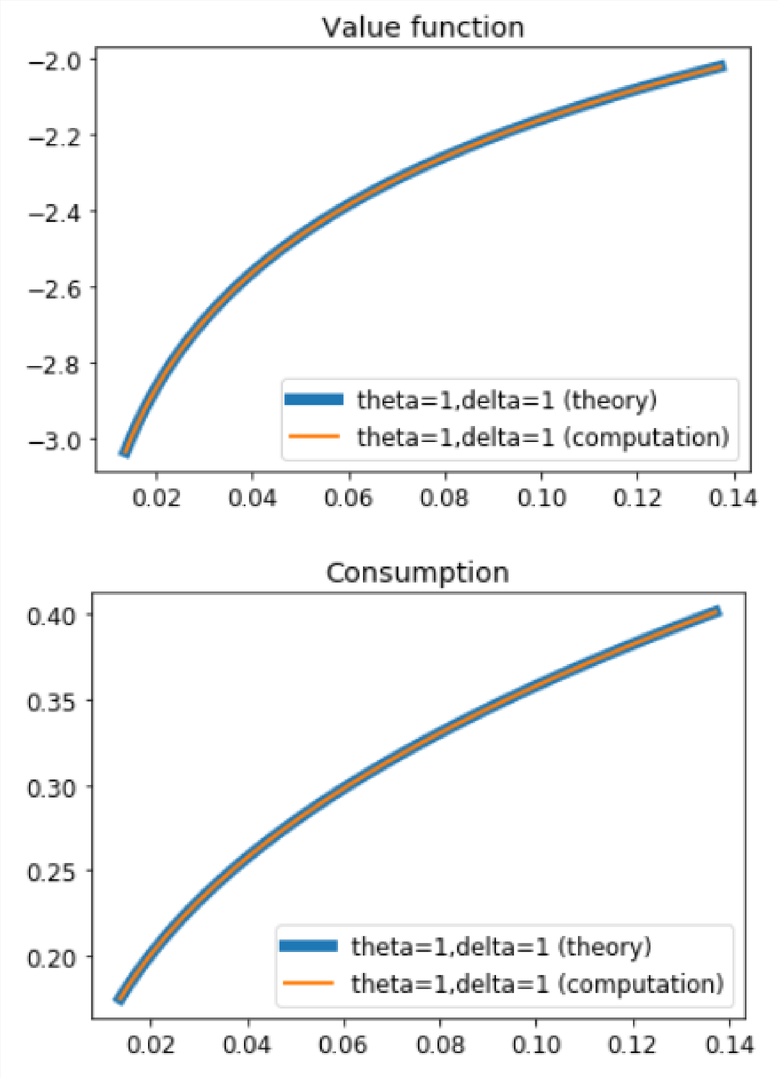
- ❑ Initialize V with a neural network parametrized by $\{\theta\}$
- ❑ Make a random draw of k and z, as well as of future shocks $z_{t+1,1}$, and $z_{t+1,2}$
- ❑ Calculate $\frac{\partial V(k,z)}{\partial k}$ and then solve for c
- ❑ Solve for k_{t+1}
- ❑ Compute the maximized Bellman equation
- ❑ Compute the Bellman error
- ❑ Update $\{\theta\}$ to minimize the error term. If it is smaller than a certain threshold stop. Otherwise loop back to make new random draws.

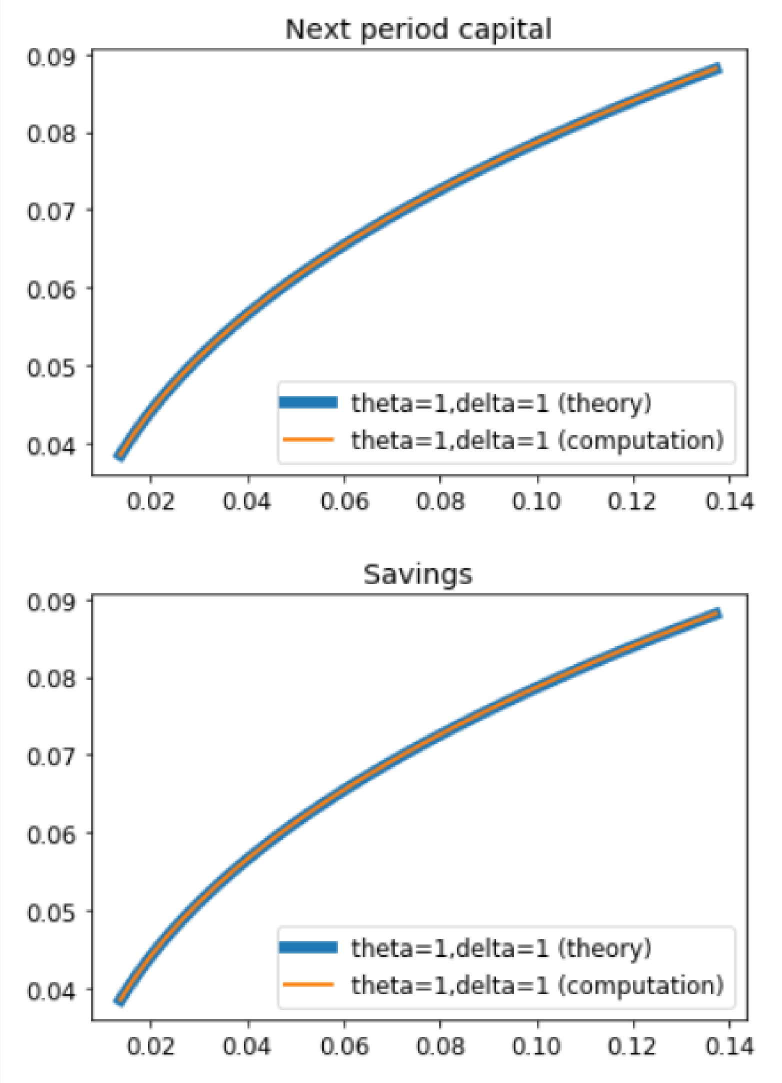
The Deep Learning Model

- ❑ We used both PyTorch and Tensorflow to tackle this problem.
- ❑ The model was tested with and without random market shocks
- ❑ k and z values were randomly chosen from a uniform and log-normal distributions, respectively
- ❑ In each epoch, the derivative of the network with respect to k is calculated and used to update the value of c
- ❑ A custom loss is defined and calculated using $||V(k,z) - T[V(k,z)]||$ and the network is minimized for loss using all trainable-weights
- ❑ The learning rate was set to 0.001 and is reduced step-wise after certain number of epochs



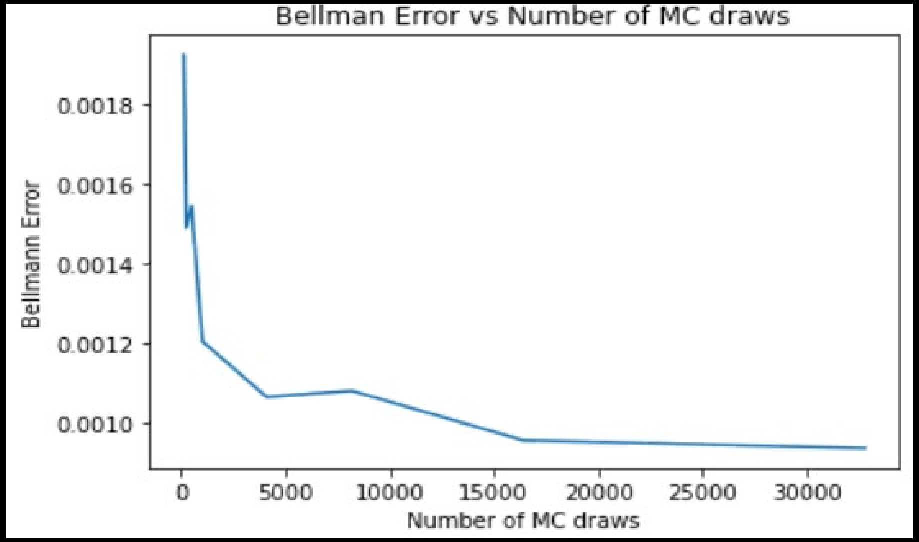
The Deep Learning Model





Limitation of the Deep Learning Framework for DSGE

- Monte Carlo is essential
  - provides unbiased estimator of the stochastic gradient with respect to random variables
  - possible to simultaneously estimate the decision function and to integrate with respect to future economic shocks
- Downside:
  - Low square-root rate of convergence



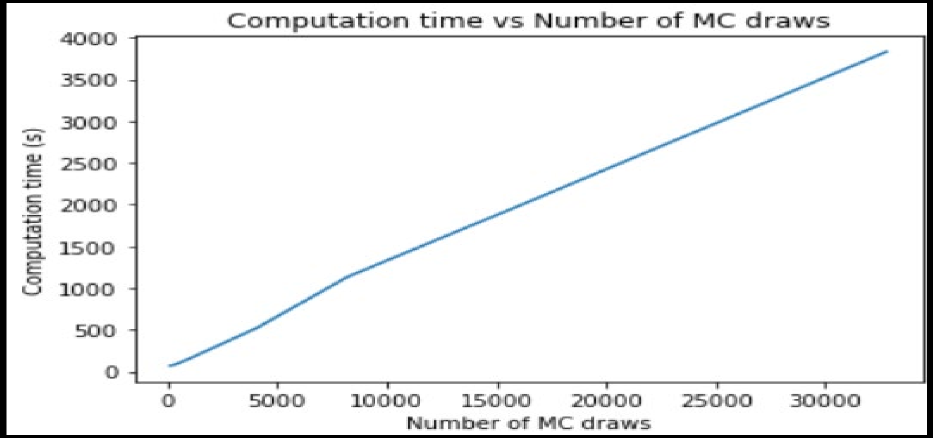


Figure: Code from Maliar, Lilia & Maliar, Serguei & Winant, Pablo, 2019. "Will Artificial Intelligence Replace Computational Economists Any Time Soon?," CEPR Discussion Papers 14024, C.E.P.R. Discussion Papers, Different numbers of MC draws were submitted to obtain Bellman Error

Classical vs Quantum Monte Carlo

Classical (samples): $N \propto \frac{1}{\varepsilon^2}$

X samples

↓

0.1 error

100X samples

↓

0.01 error

Quantum (gates): $N = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$

X gate applications

↓

0.1 error

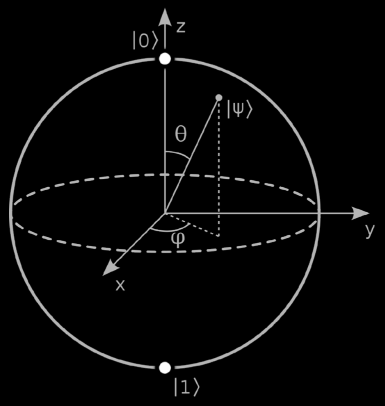
10X gate applications

↓

0.01 error

How Quantum Works

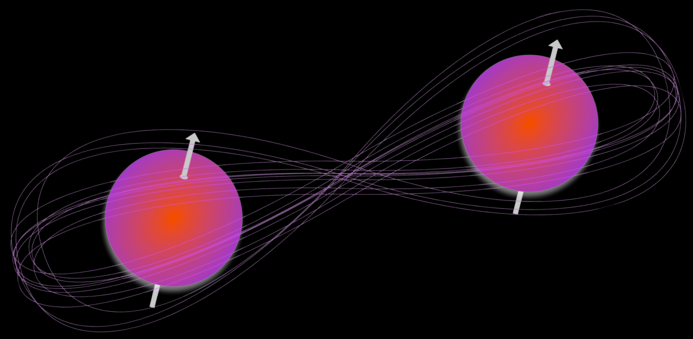
Superposition



Qubits can be a combination of 0 and 1 due to superposition, whereas binary bits can only represent 0 or 1

+

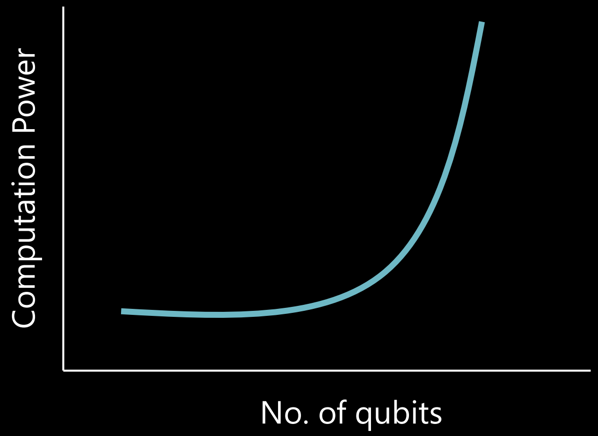
Entanglement



Quantum entanglement enables qubits to have special correlations between states, enabling greater information density

=

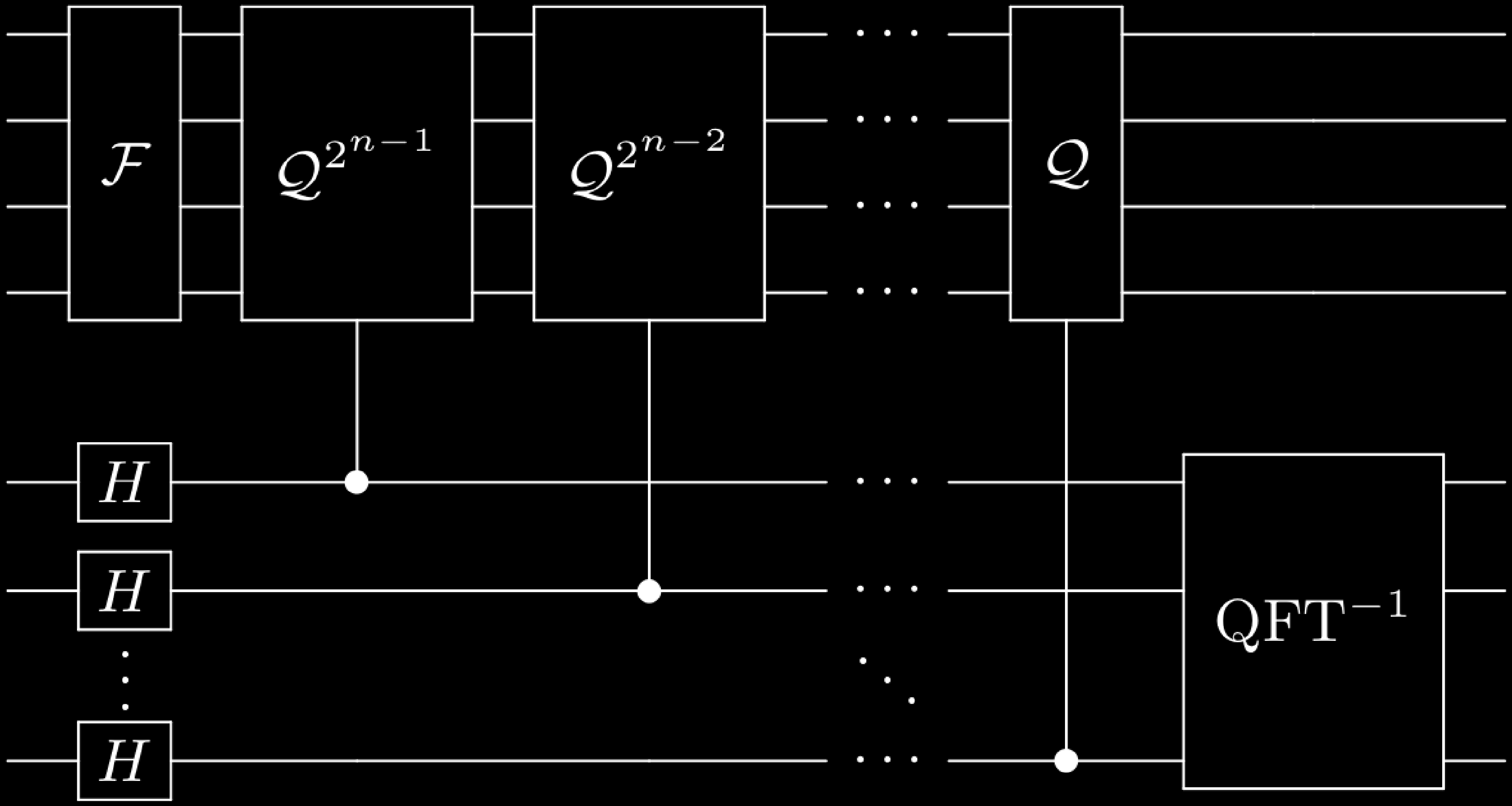
Faster computing



The power of quantum computers grows exponentially as individual qubits are added to the system

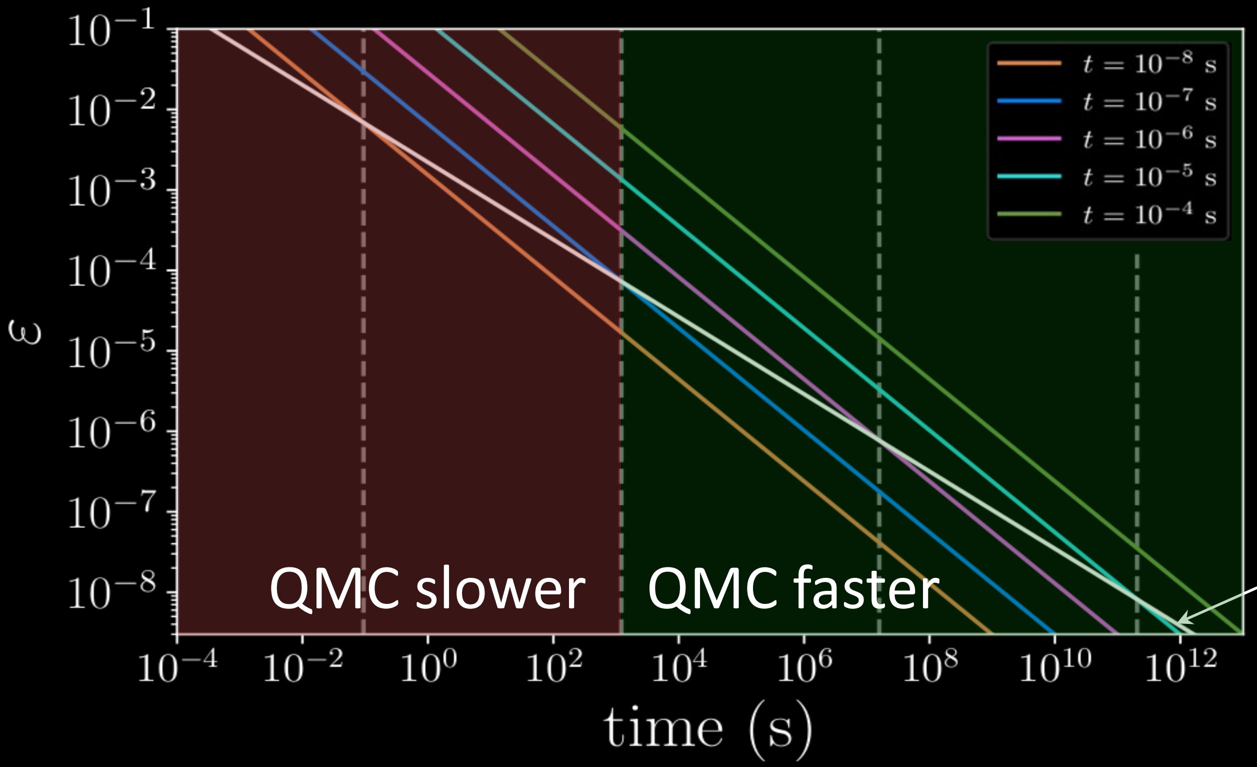
Figures curtesy of Xanadu Quantum Technologies, Inc.

Quantum Algorithm for Monte Carlo Estimation



Results for the Neoclassical Model

The good news: Assuming 10^{-7} s gate times, QMC is faster for simulations above 20 minutes



Different quantum gate times

Existing Monte Carlo solution

The bad news: requires circuit depth of 10,000+ (perhaps 5-15 years into the future)

Figure from Guala D., Skavysh V., Priazhkina S., Bromley T., 2021. “Economic Applications of Monte Carlo on a Quantum Computer"

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Using news sentiment for economic forecasting: a Malaysian case study[1]

Eilyn Chong, Chiung Ching Ho, Zhong Fei Ong and Hong H Ong,
Central Bank of Malaysia

[1] This presentation 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 central banks and other institutions represented at the event.

Using News Sentiment for Economic Forecasting

A Malaysian Case Study

Chiung Ching Ho, Eilyn Chong, Zhong Fei Ong and Hong H. Ong[1]

Abstract

Newspaper text sentiment can be informative about the prevailing macroeconomic conditions at a high frequency level and can be used to improve forecasts of macroeconomic indicators. In this paper, we extract the sentiment from the business and financial section of local newspaper articles in Malaysia using a simple dictionary method, and then evaluate the relationship with existing survey-based sentiment measures and macroeconomic growth outcomes. Specifically, this paper investigates the forecasting power of newspaper sentiment for GDP growth and its demand-side components using linear models, non-linear machine learning models and long-short term memory (LSTM) neural network. Our findings show that the news sentiment could nowcast the survey-based business sentiment measure. Using linear regression and non-linear machine learning models, we also show that the news sentiment has a reliable predictive ability for private investment growth within the two to three-quarter forecast horizon. Nevertheless, we find no significant improvement in using news sentiment to forecast other demand-side components of GDP growth across forecast periods, suggesting that the extracted news sentiment provides limited information content for the broader economy.

Keywords: Newspaper sentiment analysis, text analysis, dictionary, macroeconomic indicator, time-series, machine learning

JEL classification: C53, C82, C45

[1] This paper was prepared for the IFC and Bank of Italy Workshop on "Data Science in Central Banking" held virtually on 19-22 October 2021. All authors are from the Central Bank of Malaysia. Emails: hocc@bnm.gov.my; eilyn@bnm.gov.my; zhongfei@bnm.gov.my; onghh@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 or of anyone else associated with the Central Bank of Malaysia. The authors would like to thank Ong Li Ming, Mohd Zhafry Zainol, Chang Wen Huei and Gulraj Singh Sidhu for their invaluable comments on the paper.

1. Introduction

Central banks care about the sentiment of economic agents given their prospective influence on changes in real economic activity. For consumers, their expectations and sentiments about the economic conditions can affect their consumption and saving decisions. Likewise for businesses, their expectations and sentiments about the economic conditions would influence how much they invest and how they set prices and wages. One popular way of measuring sentiment is to directly survey these economic agents about their views on the current and future economic conditions. Prominent examples in Malaysia include the Malaysian Institute of Economic Research (MIER)'s Business Confidence Index and Consumer Sentiment Index. However, surveys can be challenging to conduct frequently, especially if they need to cover a representative sample of the population. Moreover, during periods of macroeconomic stress such as that caused by the COVID-19 pandemic, businesses that temporarily ceased trading may not respond to surveys, thereby affecting the quality of the survey responses.

Two recent advances have offered an alternative approach to measuring sentiment. First is the wide availability of newspapers and annual reports in digital format. These media especially digital newspapers may capture the high-frequency information that consumers and businesses refer to for decision making. The second advancement relates to developments in computational linguistic methods and the rapid growth in computing power. Put together, these enable us to process massive volume of text from newspapers and reports, and extract the sentiment contained therein.

Such techniques have been explored to provide quantitative measures of economic policy uncertainty (Alexopoulos & Cohen, 2015; Baker et al., 2016), daily economic sentiment (Buckman et al., 2020; Thorsrud, 2020), measure of political leanings of media outlets (Gentzkow & Shapiro, 2010) and central bank's objective function (Shapiro & Wilson, 2019).

The derived text-based measures are used to correlate with a variety of economic and financial outcomes. Nyman et al. (2021), for example, suggest text-based measures of excitement could pre-empt an impending financial system distress. Similarly, Manela & Moreira (2017) used The Wall Street Journal articles to construct a news implied volatility (NVIX), which peaks during the financial crises, stock market crashes, times of policy-related uncertainty and world wars. Prominent examples that link sentiment from text with financial market reaction include Calomiris & Mamaysky (2019) and García (2013) who link sentiment from news with stock returns as well as Jegadeesh & Wu (2013) and Loughran & McDonald (2011) who correlated sentiment of text in firms' annual reports with financial market reaction. Others have explored the links between text-based measures of uncertainty and the business cycle (Baker et al., 2016; Bloom, 2014; Moore, 2017). Our paper is closest to the literature that use news sentiment to track and predict a range of macroeconomic variables of interest. Papers such as Aguilar et al. (2021); Fraiberger (2016); Kalamara et al. (2020); Larsen & Thorsrud (2019); Nguyen & Cava (2020); Rambaccussing & Kwiatkowski (2020) show that newspaper text contains information on the future path of certain macroeconomic variables, including gross domestic product (GDP).

To investigate whether their findings would similarly apply to Malaysia, we extract sentiment measures for the Malaysian economy by applying text analytics on local news articles. A study by the Reuters Institute reports that around 86% of

Malaysian users rely on online media as a dominant source of news in recent years (Newman et al., 2020). We explore the feasibility of using online news articles to obtain timely and forward-looking cues about economic conditions that could inform policymaking, especially given the evolving COVID-19 situation, whose impact cannot be immediately observed from the lagging official statistics. Specifically, our news corpus comprises over 720,000 business and financial news articles from 16 major news portals, some of which have digital news archive since year 2001. We develop the sentiment measures based on the net balance of positive and negative words used in news articles based on the pre-defined word lists in dictionaries by Loughran & McDonald (2011) and Correa et al. (2017). We also leverage on the sentiment-scoring model developed by Shapiro et al. (2020) that cater specifically to economic news articles. We then aggregate the word counts or sentiment scores from individual article into monthly time-series indexes. The monthly index is found to comove with the business cycle and key economic events.

Specifically, we find that the news sentiment measures can nowcast movements in the survey-based sentiment indicators that are released on a quarterly basis by MIER. Of the growth variables we attempt to forecast using the news sentiment, we find that the news sentiment measures perform well especially in forecasting private investment growth especially within 2 – 3 quarters ahead. This is true even during periods of macroeconomic stress, highlighting the advantages of using higher frequency news sentiment to inform movements in private investment growth. Similar results are obtained when using non-linear machine learning algorithms. Nevertheless, we find no significant improvement in using news sentiment to forecast other components of economic activity, such as private consumption. This suggests the extracted news sentiment provides limited information content for the broader economy.

The rest of the paper is organised as follows: we first describe our newspaper text data in Section 2. Section 3 then discusses the methods of transforming text into time series and the nowcasting exercises that we perform. In Section 4, we look at the forecast performance of several economic variables with the text-based sentiment using simple linear regression, non-linear machine learning models as well as long short-term (LSTM) neural networks. Section 5 discusses the overall results and concludes.

2. Data

The raw data used in constructing the news sentiment consist of daily newspaper articles taken from our internal subscriptions of 16 Malaysian online newspapers portal either from the official websites or via third party services. Based on a study commissioned by the Reuters Institute in 2020, online news websites remained as one of the predominant sources of news for Malaysians in recent years. Over 720,000 online newspaper articles in the English language were used for our analytical dataset. The selection of newspaper articles was motivated by the availability of digital archives, allowing articles to be extracted from as early as 2001, all the way up to June 2021.

We were particularly interested in investigating the sentiment of newspaper articles in the year 2020, which is representative of the developments of the COVID-19 pandemic and its impact to the economy. As we intend to correlate news sentiment with economic indicators, we consider only news articles from the business- or financial-related sections to increase the signal-to-noise ratio. The 16 news portal used in our dataset, sorted by the average number of articles published per month, is shown in Table 1.

Based on the study by Reuters Institute, The Star and Astro Awani are among the most popular news portals with readership[2] of 30% and 35% respectively. It is important to note, however, that these represent the general readership pattern of Malaysians for all types of news, which is different from the goal in this study which focuses on business- or financial-related news articles. Hence, beyond the popular news portals for general news, we also include the more business-centric news portals such as The Edge Markets and i3investor that are not featured in the Reuters Institute's study.

On top of English news, Malay, Tamil and Chinese are also common languages used in news portals in Malaysia. We decided to analyse only English news articles for two main reasons. The first is the limited availability of well-established algorithms to analyse the sentiment of text in vernacular languages. The second is to avoid double counting news articles that are written in multiple languages in the same news portal.

[2] Based on the Reuters Institute Digital News Report 2020 (Newman et al. 2020), readership is defined as the share of respondents who consumed the media at least once a week.

Descriptive statistics of articles from selected local news portals

Table 1

	Average number of articles per month	Date of first online article	Readership[3]
i3investor	7898	03/03/2020	-
The Star	949	01/01/2003	30%
The Edge Markets	912	16/01/2009	-
Malay Mail	827	18/06/2013	8%
Bernama	654	04/03/2020	-
The Malaysian Reserve	487	10/01/2017	-
Free Malaysia Today	384	31/12/2015	15%
The Borneo Post	284	23/12/2009	-
The Sun Daily	280	15/11/2017	-
SoyaCincau	273	29/01/2020	-
New Straits Times	238	20/05/2014	10%
MPOB Palm News	165	01/03/2020	-
paultan.org	78	25/03/2007	-
Astro Awani	59	01/01/2013	35%
Daily Express	54	15/01/2001	-
MARC	18	20/05/2020	-

Sources: Authors' calculation

3. Methodology

Text pre-processing

Text pre-processing is a common practice of cleaning and preparing text data for subsequent natural language processing tasks. We took the common steps of cleaning the raw newspaper text, including:

- Removal of punctuations, hyperlinks, hypertext markup language (HTML) tags, special characters and extra white spaces;
- Dropping of common stop words using the word list by Nothman et al. (2019) - words that are not by themselves informative and differentiative of sentiment, such as *and*, *is* and *the*;
- Setting all words to lowercases.

Given that we extract sentiment using dictionaries that include the stem words and their inflections (for example, *decline, declining* and *declined*), we do not use stemming or lemmatisation.

[3]

Sentiment scoring

Sentiment from text is not directly observable and would require text analytical approaches to extract and quantify the sentiment contained in the text. There are two such general approaches. The first is the lexical- or dictionary-based approach that associates predefined lists of words with specific scores indicating how positive or negative it is, without any element of learning. Generally, these dictionaries have ternary classifications of 1, 0, and -1 for positive, neutral, and negative sentiment respectively, but certain lexicons such as Vader have a range of scores. While such word-matching method measures the sentiment of a given corpus of text based on the prevalence of negative vs positive words, it ignores the word's context and compositionality.

To capture the specific contextual characteristics and nuances in human language beyond heuristic rules, machine learning (ML) techniques can be employed to probabilistically predict the sentiment of any given set of text. An ML model is typically trained on a large set of text containing a mapping between textual utterances and sentiment ratings assigned by humans. These have been applied for example on social media data, such as tweets on Twitter combined with user feedback to identify the sentiment of the tweets. While this approach can better capture the nuances in sentiment expression, constructing a large, labelled training dataset is time-consuming and expensive.

In this paper, we adopt the simpler dictionary-based method to measure the sentiment of the news corpuses and construct the news sentiment index. We use the financial stability dictionary by Correa et al. (2017) (hereafter Correa) and the finance-oriented dictionary by Loughran & McDonald (2011) (hereafter LM). Our news sentiment is constructed by counting the number of times that negative and positive words appear in the cleaned text of articles and measuring the net balance of words. When the news contains more positive words and/or fewer negative words, it indicates better sentiment in the economy.

In addition to the lexicons above, we also leveraged on the lexicon created by Shapiro et al. (2020) (hereafter SSW), who scored a corpus of U.S. economic news articles with Vader[4]. This consists of 20,000 words labelled from -4 to +4, corresponding to most negative to most positive. While specific to the context of the U.S., we adopt this alternative lexicon as an attempt to incorporate words that are more specific to the economics rather than financial domain, and that have a wider scoring scale that differentiates between weaker and heightened sentiments. For example, the word *declined* is assigned a score of -0.12, *dropped* -0.16, *downturns* -1.22 and *sluggish* -1.97.

We also took an additional step to swap the sentiment for the words *positiv*e and *negative* that are in the dictionaries but may be associated with COVID-19 in the more recent newspaper text as it will contribute falsely to the economic sentiment that we intend to construct.

To construct an index of the news sentiment, the articles are sorted by the date of publication. In the case of Correa and LM, we compute the sentiment score for each news article i by subtracting the count of negative words from the count of positive words and then dividing by total word count. In the case of SSW, we take a

[4] An open-source Vader python package developed by Hutto & Gilbert (2014)

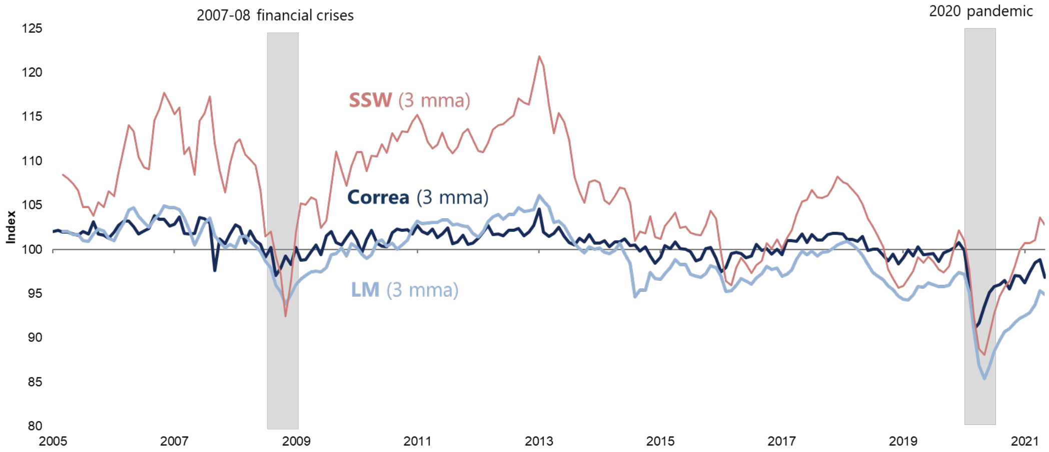
sum of the sentiment scores associated with the words w in each news article i and then dividing by total word count. We then express these sentiment scores as net count per 1000 words and convert them into an index. An index above 100 indicates better sentiment, and index below 100 indicates otherwise.

$$\text{Correa and LM:} \quad sentiment\ index_i = 100 + \frac{\sum Positive_i - \sum Negative_i}{Word\ count_i} \times 1000$$

$$\text{SSW:} \quad sentiment\ index_i = 100 + \sum_w \frac{Score_{w,i}}{Word\ count_i} \times 1000$$

Given that the overall volume of articles varies across newspapers and time, we scale the index by the total number of articles in the same newspaper and time period, which yields a news sentiment time series for each news. We then take average across the 16 news portals by a chosen frequency, whether it be daily, monthly or quarterly. Figure 1 shows the monthly news sentiment indices. The sentiment measures exhibit a pronounced drop during the economic downturns that happened during the 2007-08 financial crisis and the onset of COVID-19 pandemic in Malaysia, as well as the subsequent increase following the economic recovery from these crises.

Figure 1: Monthly news sentiment for Malaysia



Nowcasting survey-based measures of sentiment

To investigate the information content of the news sentiment, we first explore whether news sentiment can help nowcast the Business Condition Index (BCI) and Consumer Sentiment Index (CSI) published by the Malaysian Institute of Research (MIER). While these survey-based measures are closely followed by economic analysts, they are released only every quarter and with a lag of 2 months after the end of the reporting quarter. In contrast, the news sentiment can be constructed at a higher frequency, thereby providing information about economic sentiment between releases of these survey-based measures.

In this nowcasting exercise, we test whether the monthly news sentiment within the quarter can help predict the current quarter's BCI_t and CSI_t. We estimate the following for $m = 1, 2, 3$:

$$BCI_t = \alpha + \beta\, BCI_{t-1} + \eta\, x_{t,m} + \varepsilon_t \tag{1}$$

$$CSI_t = \alpha + \beta\, CSI_{t-1} + \eta\, x_{t,m} + \varepsilon_t \tag{2}$$

where $x_{t,1}$ is the value of the sentiment in the first month of the current quarter t, $x_{t,2}$ is the average value for the first two months of the quarter t and $x_{t,3}$ is the average value for the full quarter.

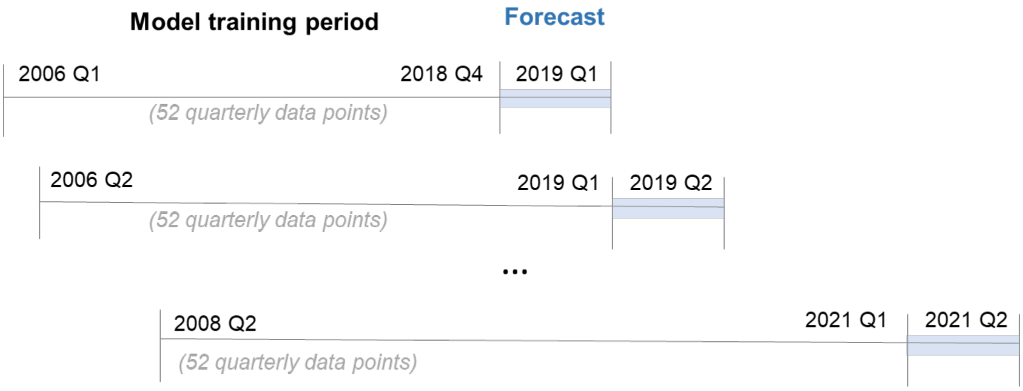
Forecasting GDP components using linear models with news sentiment

Next, we assess the forecasting ability of the news sentiment using linear models. Our forecasting targets variables are aggregated real GDP year-on-year growth and some of its demand-side components, namely year-on-year growth in private consumption, private investment, exports and imports. All target variables are at quarterly frequency and are standardised to z-scores for the forecasting exercises.

Our forecast exercises involve estimating a model over a specified training period using each of the three sentiment measures in turn. We train the models recursively on a rolling window, followed by producing out-of-sample predictions of target variables at horizon *h* = 1; 2; 3 quarters ahead. In other words, our forecast exercise seeks to mimic a scenario in which policymakers at time t have historical data of the target variable (i.e., $y_{\mathrm{t-1}}, y_{\mathrm{t-2}}$,...) and are anticipating/forecasting the official statistics $y_{\mathrm{t-1+h}}$ while having access to the news sentiment x_t.

For *h* =1, this replicates an actual forecasting situation starting from 2019 Q1 and moving forward a quarter at a time through to 2021 Q2 (or 2018 Q3 onwards for *h* = 3). For example, for the first vintage of the data, the models are estimated over the period 2006 Q1 to 2018 Q4 using data for both the chosen target variable, its lag and the news sentiment. The fitted models are then used to nowcast the response variable in 2019 Q1. As an example, Figure 2 illustrates the selected horizons for model training and 1-quarter ahead forecast period. Overall, we generate 10 real-time nowcasts of the response variables. The chosen period will also put to test the informational content of the proxy indicators during the recent macroeconomic stress caused by the pandemic.

Figure 2: Training period and 1-quarter ahead forecast horizon



We use the ordinary least square (OLS) linear regression method, and also include two more modified versions of linear regression:

- The Ridge regression which performs L2 regularisation where the model is penalised for the sum of squared value of the magnitude of the coefficients. λ is the parameter that determine the relative impact of the penalty terms. When λ = 0, we have the standard OLS approach.

$$\beta^{Ridge} = argmin \left[\sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} \mathrm{x}_{ij}\beta_j \right) + \lambda \sum_{j=1}^{p} \beta_j^2 \right]$$

- The Huber regression which is robust to outlier by reducing the weight of large residuals. In Huber weighting, observations with small residuals get a weight of 1 and the larger the residual, the smaller the weight. This is defined by the weight function:

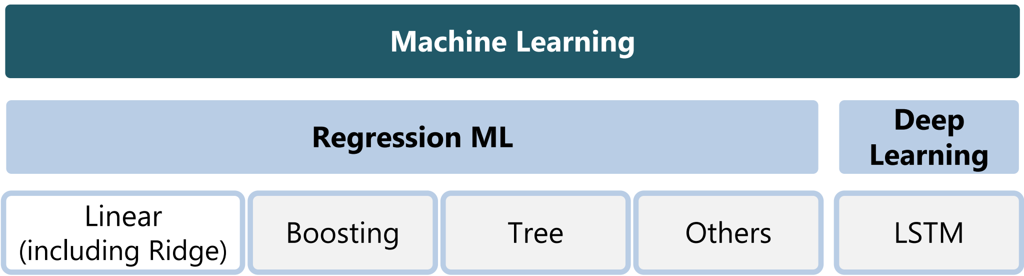
$$w(e) = \begin{cases} 1 & for \quad |e| \leq \mathrm{k} \\ \dfrac{k}{|e|} & for \quad |e| > \mathrm{k} \end{cases}$$

Forecasting GDP components using non-linear machine learning models with news sentiment

Beyond linear regression models, we also run a set of non-linear machine learning models. More generally, machine learning is a subset of artificial intelligence which aims to learn representation of knowledge of data in order to generate meaningful insights concerning both the data on hand and also on data that is unknown from the future. Supervised machine learning algorithms take an input (e.g., the sentiment of newspaper articles) to predict a future outcome, e.g., an economic indicator such as a country's GDP. Supervised regression depends on labelled data, which are pairs of inputs and outputs that has been sampled from historical data. The parameters of a supervised machine learning algorithm are tuned in order to minimise in the prediction error, for example the root mean squared error (RMSE).

In our paper, we wanted to investigate the viability of using supervised non-linear regression machine learning techniques in order to predict the growth variables for h = 1; 2; 3 quarters ahead. As with the linear regression models, we train the models recursively with news sentiment used as a feature on a rolling window, followed by producing out-of-sample predictions of the target variables, but with two distinct approaches: a suite of non-linear machine learning regression models and long short-term memory (LSTM) (only for h = 1).

Figure 3: Machine learning groups



The regression machine learning algorithms used are Light Gradient Boosting Machine, Random Forest Regressor, Extra Tree Regressor, Orthogonal Matching Pursuit, Gradient Boosting Regressor, Decision Tree Regressor, Adaboost Regressor and Passive Aggressive Regressor, which are grouped into distinct groups as shown in Figure 3.

One group of non-linear machine learning algorithms is called Boosting, and it groups weak machine learning algorithms in an ensemble in order to reduce bias and variance which reduces predictive power. Algorithms in this group include Light Gradient Boosting Machine, Gradient Boosting Regressor and the Adaboost Regressor. The following group, Tree, are highly interpretable algorithms which divides the prediction boundaries into simpler regions. In this group, we have algorithms such as the Random Forest Regressor, Extra Tree Regressor and the Decision Tree Regressor. The final group Others, contains algorithms such as Orthogonal Matching Pursuit.

Recurrent neural network (RNN) is a type of neural network with an internal memory. Because of this memory, RNN can remember information about the input that they have received, and this helps RNN to predict what future values precisely. This attribute is very useful for predicting the values of data that is organised in sequences, including time-series data. Long term short memory (LSTM) is an extension to RNN, in that it extends the memory. It is therefore very well suited to learn from data that has very long time-gaps in between. LSTM has been used to successfully predict macroeconomic indicators such as global merchandise exports value and volume (Hopp, 2021). LSTM is considered to be a type of deep learning algorithm. The learning process in deep learning algorithm is 'deep' because the underlying neural network has many layers that captures data (input layer), process data (hidden layer) and contains the predicted state of the data (output layer)

Regression machine learning algorithms and deep learning algorithms were chosen as we wanted to explore the fundamental differences in how prediction is made. Regression machine learning algorithms needs to be told how to represent data (the independent variables) while deep learning algorithms are able to create new data representation through data processing.

In the LSTM experiments, we wanted to ascertain the suitability of this technique given the frequency of data available in our context. Different combinations of independent variables were used to predict macroeconomic indicators, using different length of data.

We use the settings in Table 2 for the LSTM machine learning experiments. The evaluation measure used in the experiments is the average root mean squared error for the sliding window that we are using. The sliding window will start from 2006Q1 and move 1 quarter at each step. From this table, we can conceptually understand that the LSTM deep learning network will consider the past four quarter of observations in order to predict the following one-quarter ahead. We further considere different combinations of macroeconomic economic indicators and sentiment scores in order to predict a target macroeconomic indicator.

LSTM Experimental Setting

Table 2

Time period	2006Q1 to 2021Q2
Sliding window size	4
Number of quarters to predict ahead	1
Number of layers	5
Optimizer	Adam
Loss function	Huber

Forecast evaluation

For all models mentioned above, the out-of-sample root mean squared error (RMSE) is used as a metric to compare the nowcasting properties of the sentiment measures. RMSE is calculated on the out-of-sample forecast period with the standard formula:

$$RMSE = \sqrt{\sum_{t=1}^{N} \frac{(\hat{y}_t - y_t)^2}{N}}$$

where $\hat{y}_t$ is the predicted value for the time period t, y_t is the actual value and N is the total number of predicted observations.

We compare the performance of the model with news sentiment to a pure OLS-AR(1) model, which is the baseline model without the sentiment measure x_t. For example, in the case of the linear models, we compare

AR(1) with sentiment: $y_{t+h} = \alpha + \beta\, y_{t-1} + \eta\, x_t + \epsilon_t$

OLS- AR(1) baseline: $y_{t+h} = \alpha + \beta\, y_{t-1} + \epsilon_t$

where $x_t \in \{Correa_t, LM_t, SSW_t\}$

For each forecast, we calculate the ratio RMSE which is the model's RMSE relative to the OLS-AR(1) model.

$$Ratio\ RMSE = \frac{RMSE}{RMSE_{AR1}}$$

A ratio RMSE of less than 1 indicates that the model performed better compared to the benchmark model. Conversely, a ratio RMSE of greater than 1 indicates that the model has performed worse compared to the benchmark model. For a given target variable, we calculate the average ratio RMSE across the out-of-sample periods.

4. Results and Discussion

Pearson correlation coefficient

Table 3

	Full sample 2006 Q1 – 2021 Q2			Non-crisis 2011 Q1 – 2020 Q4		
	Correa	LM	SSW	Correa	LM	SSW
Macroeconomic variables:						
Aggregate GDP	0.57	0.52	0.41	0.31	0.27	0.32
Private investment	0.49	0.51	0.48	0.56	0.61	0.65
Private consumption	0.60	0.53	0.38	0.23	0.16	0.18
Exports	0.25	0.22	0.22	-0.03	-0.11	-0.13
Imports	0.28	0.28	0.28	0.23	0.18	0.17

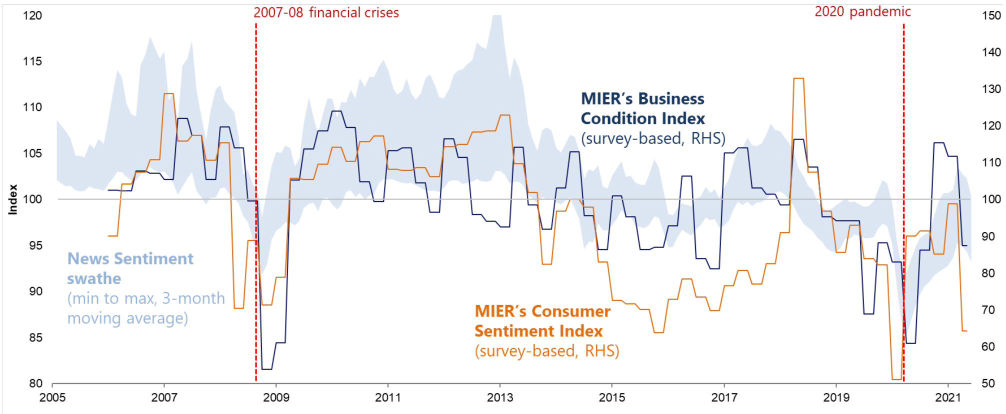
Sources: Authors' calculation

As a starting point, we compare each news sentiment to the time series of aggregate GDP growth and its components by looking at their contemporaneous correlations as shown in Table 3. On average, the news sentiments' correlation with aggregate GDP growth as well as private sector economic activities are relatively higher and of the expected sign. Trade activities (exports and imports) appear to be weakly correlated with the news sentiment measures, possibly reflecting the coverage of topics in our news corpuses that may lean towards domestic-oriented economic activities. Another possibility is a stronger lead-lag relationship between news sentiment and trade activities. After excluding the 2007-08 financial crisis and the recent pandemic-induced crisis, there is a noticeable decline in the correlations between the macroeconomic growth variables and news sentiments, with the exception of private investment. This suggests that even during non-crisis period, the news sentiment may contain information regarding investment activities in the private sector.

Nowcasting survey-based measures of sentiment

We now assess the information content of the news sentiment vis-à-vis the survey-based measures of economic sentiment. Both news and survey-based sentiment appear to move in tandem and exhibit sharp declines during the 2007-08 financial crisis and onset of COVID-19.

Figure 4: Monthly news sentiment and survey-based sentiment measures



Tables 4a documents the results of using news sentiment to nowcast the survey-based MIER's BCI. Before adding the news sentiment measures, we find that the prior quarter's release of the survey-based sentiment is statistically significant in explaining the current quarter's release. Beyond column (1), when we add the news sentiment measure, reflecting the information set available with each passing month in the quarter, we find that this news sentiment measure is statistically significant throughout. Perhaps not surprisingly, the coefficient of the sentiment measure and adjusted R-squared also generally increase as more news within the quarter are incorporated into the news sentiment measure, suggesting increasing information content in nowcasting MIER's BCI in the current quarter. Similar results are also obtained when we repeat the exercise with MIER's CSI in Table 4b, but the statistical significance is weaker compared to that of BCI.

Nowcasting Quarterly Survey-Based Sentiment

Dependent variable: MIER's Business Conditions Index (BCI)
Sample: 2006 Q1 – 2021 Q2

Table 4a

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	AR1	Correa	Correa	Correa	LM	LM	LM	SSW	SSW	SSW
MIER's Sentiment	0.49***	0.36***	0.33***	0.33***	0.35***	0.34***	0.35***	0.35***	0.31***	0.33***
(1 quarter prior)	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)
News sentiment		0.30**			0.32**			0.41***		
(first month of quarter)		(0.14)			(0.14)			(0.14)		
News sentiment			0.41***			0.39***			0.52***	
(average of first 2 months)			(0.15)			(0.14)			(0.15)	
News sentiment				0.45***			0.40***			0.48***
(average for the quarter)				(0.14)			(0.13)			(0.14)
Constant	-0.01	-0.01	-0.01	0.02	-0.03	-0.02	0.00	-0.02	-0.03	-0.02
	(0.11)	(0.11)	(0.11)	(0.10)	(0.11)	(0.11)	(0.11)	(0.10)	(0.10)	(0.10)
Adjusted R^2	0.22	0.29	0.34	0.35	0.30	0.32	0.32	0.35	0.39	0.35
Observations	61	61	61	61	61	61	61	61	61	61

Heteroscedastic and autocorrelation robust (HAC) standard errors in parentheses. * p < 0:10 ** p < 0:05 *** p < 0:01

Source: Authors' calculation

Nowcasting Quarterly Survey-Based Sentiment

Dependent variable: MIER's Consumer Sentiment Index (CSI)
Sample: 2006 Q1 – 2021 Q2

Table 4b

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	AR1	Correa	Correa	Correa	LM	LM	LM	SSW	SSW	SSW
MIER's Sentiment	0.71***	0.65***	0.59***	0.55***	0.54***	0.54***	0.52***	0.50***	0.44***	0.42***
(1 quarter prior)	(0.09)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.13)
News sentiment		0.10			0.27*			0.32***		
(first month of quarter)		(0.14)			(0.14)			(0.08)		
News sentiment			0.22			0.30**			0.44***	
(first 2 months of quarter)			(0.16)			(0.14)			(0.10)	
News sentiment				0.30*			0.35***			0.48***
(full data for the quarter)				(0.15)			(0.13)			(0.11)
Constant	-0.01	-0.01	-0.01	0.00	-0.03	-0.02	-0.00	-0.02	-0.03	-0.02
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.08)	(0.08)	(0.08)
Adjusted R^2	0.47	0.47	0.49	0.51	0.51	0.51	0.52	0.51	0.55	0.55
Observations	61	61	61	61	61	61	61	61	61	61

Heteroscedastic and autocorrelation robust (HAC) standard errors in parentheses. * p < 0:10 ** p < 0:05 *** p < 0:01

Source: Authors' calculation

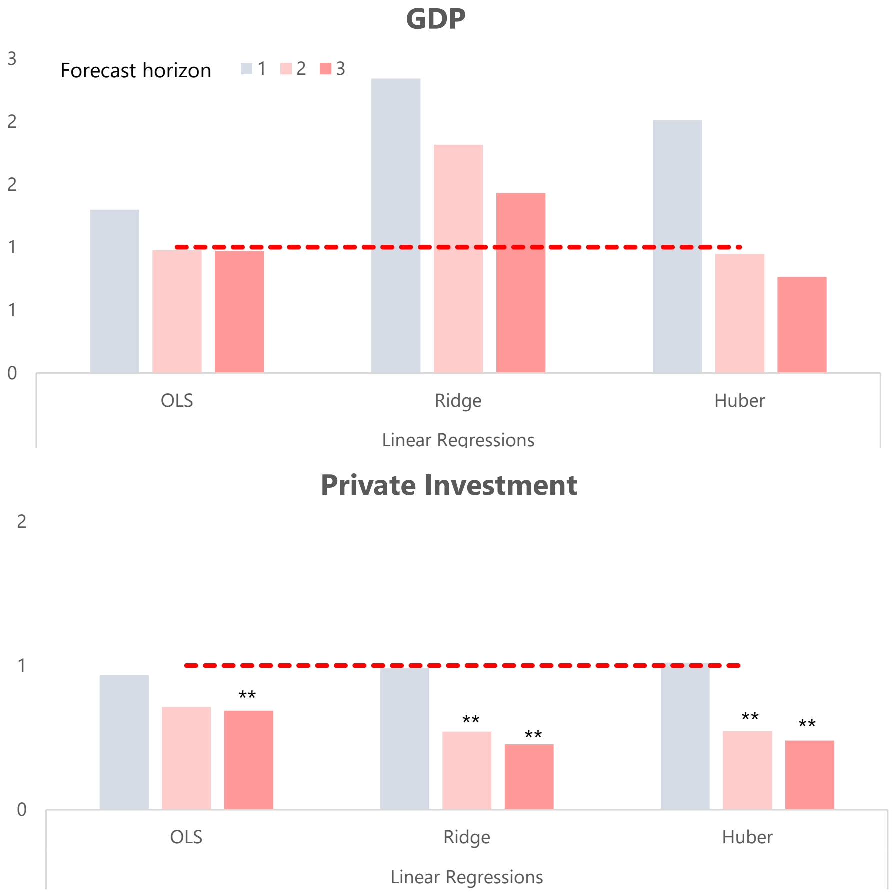
Forecasting GDP components using linear models

We now look at the predictive ability of the news sentiment for GDP components in a linear model setting. Figure 5 shows the average ratio RMSEs (across forecast period and news sentiment measures) for the out-of-sample forecast relative to the OLS-AR(1) baseline model for each target variable. Bars below the red line indicate ratio RMSE values of less than 1, which may be interpreted as an improvement of the forecast over the OLS-AR(1) model

Across the OLS, Ridge, and Huber regressions, we observe that the forecasts of private investment over the 2- and 3-year forecast horizon have ratio RMSEs less than 1, which indicates improvement in performance with the addition of news sentiment compared to the baseline model. Bars with ** further indicate ratio RMSEs that are below 1 for at least 90% of the chosen forecast periods. Furthermore, Figure 6 shows that despite worsening performance across all models following the onset of the pandemic in Q2 2020, news sentiment still offers a significant improvement in the forecast performance relative to the benchmark OLS-AR(1) model.

For aggregate GDP and other GDP components, however, the results are relatively weaker across all forecast horizons compared to the AR(1) model, except for trade activities at the 3-quarter horizon. These suggest that news sentiment has a relatively stronger influence on investment decision but not so much on the broader economic activity.

Figure 5: Ratio RMSE using linear regression models


GDP

Private Investment

Figure 5 (continued): Ratio RMSE using linear regression models

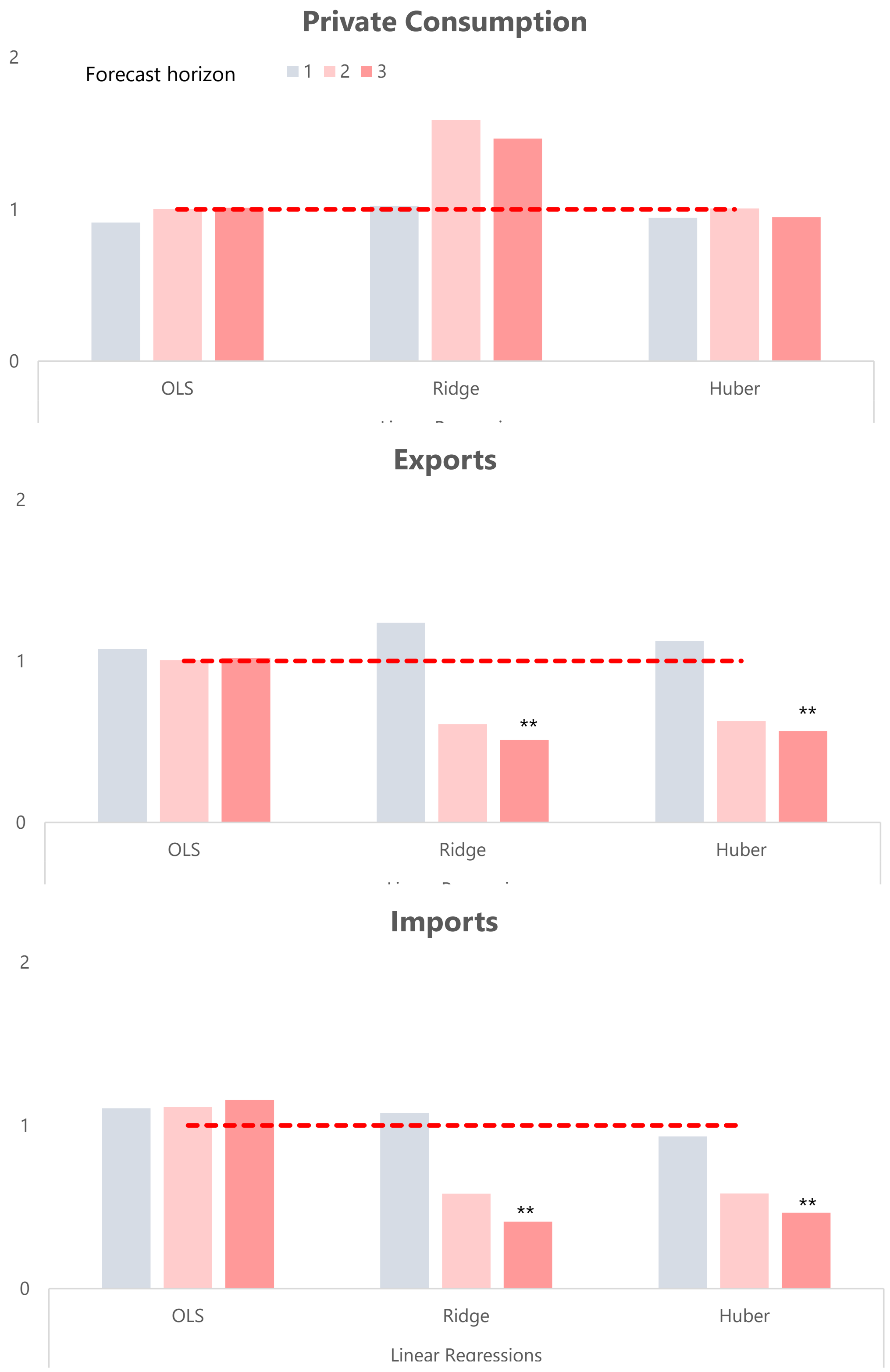
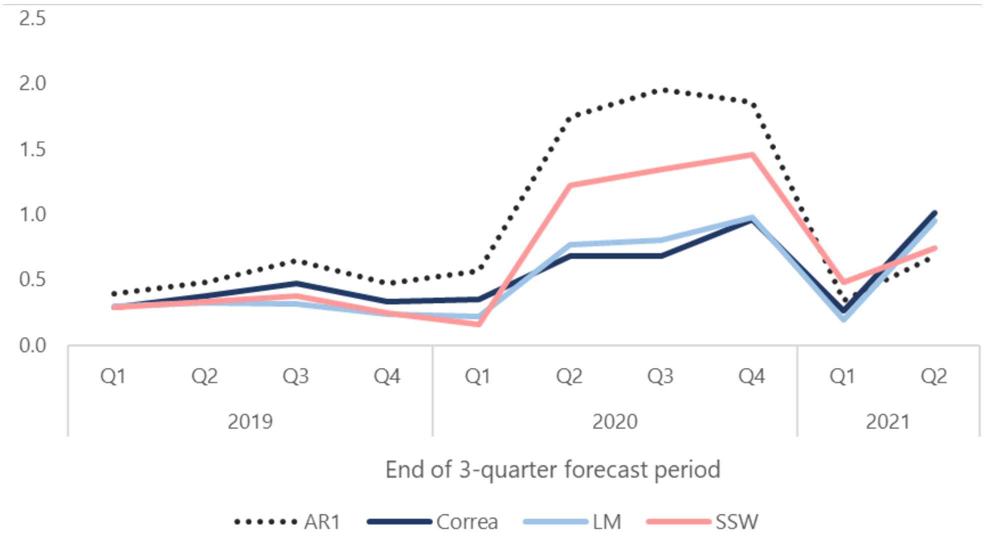


Figure 6: Comparison of out-of-sample RMSEs (3-quarter ahead) between the AR(1) model and the OLS model with news sentiment.



Forecasting GDP components using non-linear machine learning models

Figure 7 expands the earlier results of predicting GDP and its various components using a range of non-linear machine learning algorithms from the *Boosting*, *Tree* and *Others* groups with news sentiment. Again, bars below the red line can be interpreted as an improvement of the forecast over the OLS-AR(1) model, and bars marked with ** further indicate ratio RMSEs that are below 1 for at least 90% of the chosen forecast periods.

Similar to the results from linear regressions, we observe notable improvements in the performance of forecasting private investment growth 2- and 3-quarter ahead. In this regard, the AdaBoost Regressor, Random Forest Regressors and Extra Trees Regressors show reliable improvement in forecasting private investment growth across the forecast periods. As with the linear regressions, the forecast improvements shown by the machine learning models with news sentiment persist even during the crisis in 2020 (results not shown).

The machine learning algorithms also exhibit improvements for the 2- and 3-quarter predictions of exports and imports growth, although with less consistency across time period compared to private investment growth. The Light Gradient Boosting Machine nevertheless performs well for predicting imports growth. Overall, non-linear machine learning algorithms performed poorly when it comes to predicting GDP and private consumption. Further investigation on the usage of machine learning to estimate macroeconomic indicators is however warranted due to good results achieved recently (Cicceri et al., 2020; Richardson et al., 2018).

Figure 7: Ratio RMSE using non-linear machine learning algorithms

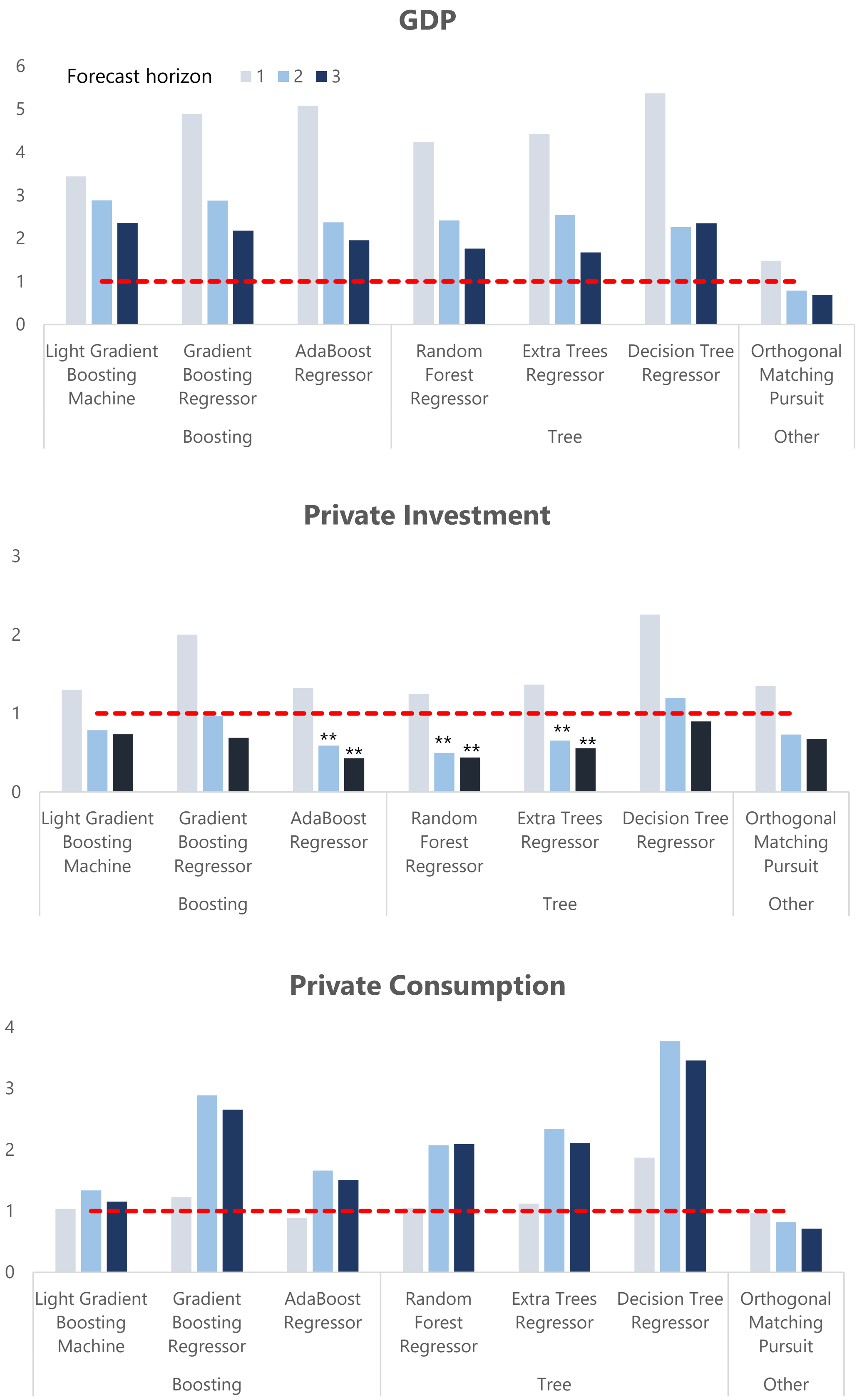
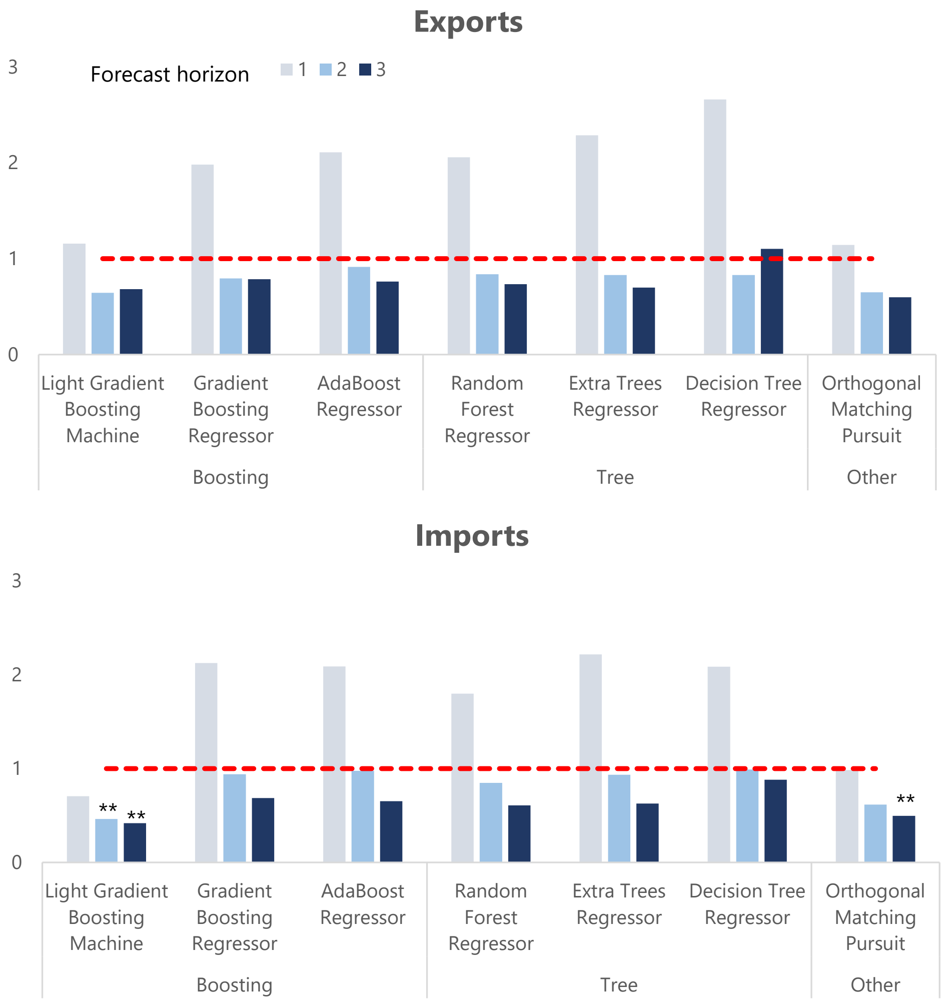


Figure 7 (continued): Ratio RMSE using non-linear machine learning algorithms



Experimental Results using LSTM

Average RMSE across sliding windows for LSTM

Table 5

Input variable(s) in LSTM (with past four quarter of observations)		Target variables				
		GDP	Private investment	Private consumption	Exports	Imports
(1)	GDP	0.58				
(2)	Private investment		0.32			
(3)	Private consumption			0.29		
(4)	Exports				0.95	
(5)	Imports					0.80
(6)	GDP and Correa	0.55				
(7)	GDP and LM	0.53				
(8)	GDP and SSW	0.56				
(9)	Private investment and Correa		0.30			
(10)	Private investment and LM		0.27			
(11)	Private investment and SSW		0.18			
(12)	Private consumption and Correa			0.19		
(13)	Private consumption and LM			0.18		
(14)	Private consumption and SSW			0.18		
(15)	Exports and Correa				1.20	
(16)	Exports and LM				1.16	
(17)	Exports and SSW				1.08	
(18)	Imports and Correa					1.06
(19)	Imports and LM					0.97
(20)	Imports and SSW					0.88
(21)	GDP, Private investment, Private consumption, Exports, Imports, Correa, LM, SSW	0.55	0.19	0.14	1.02	0.90

In Table 5, we show our experimental results using LSTM. Our LSTM experiments were focused on investigating the effect of newspaper sentiment on predicting macroeconomic indicators. The shaded cells indicate the lowest value for RMSE for each target variable.

Rows 6 till 21 show the effect of newspaper sentiment on predicting macroeconomic indicators and here, the results are mixed but broadly similar to the findings from the linear and non-linear models above. Compared to the RMSEs of models where the lagged values of the target variable are the only input in the LSTM network (rows 1 to 5), there is an improvement for GDP, private consumption and especially for private investment when newspaper sentiment is included as an additional input to the LSTM. For exports and imports, however, the prediction results worsened after incorporating newspaper sentiment. What happens if we were to incorporate more variables in the LSTM model? In Row 21, eight features were used as inputs and the prediction results improved for private investment, private consumption and GDP but worsened again for exports and imports.

Nevertheless, this study deals with only a relatively short time series for LSTM. The architecture of the LSTM model, the number of times the entire time series is passed through the network (epoch), the choice of the optimizer and loss function, batch size and network constructs (two autoencoders in this case) would all need to be investigated thoroughly for a conclusive answer.

5. Conclusion

Our observations suggest that the sentiment embodied in the business and financial news from online newspaper portals corresponds to the survey-based business sentiment measure and can provide forward-looking indication of investment activities in Malaysia. Specifically, we find that the monthly news sentiment – even in the early part of the quarter – can explain movements in the quarterly survey-based business sentiment indicators that are often published with a lag.

We investigated the extent to which news sentiment can help predict economic growth outcomes, finding that the news sentiment can consistently forecast private investment growth better than the benchmark OLS-AR(1) model, especially within 2 – 3 quarters ahead, but not other components of economic activity. The forecast gains for private investment holds true even during the recent period of macroeconomic stress following the pandemic. This suggests the extracted news sentiment can provide a timelier read of investment activities in Malaysia even during economic turning points but has limited information content for the broader economy.

There are several avenues for future research. First, we only explore and compare the sentiment in English local news in Malaysia, but to be truly representative of the Malaysia's newspaper readership in terms of ideological predisposition, one needs to also draw text from vernacular-based newspapers (e.g. Malay, Chinese and Tamil). To our knowledge, well-established algorithms to analyse the sentiment of such text remain limited. Yet another avenue for future research is to consider the ideology of a news article as a feature for machine learning models. In addition, one may also consider identifying the sentiment in news articles from other sections. For example, while we find that business and financial news are associated with investment activity, the language used in non-business articles is plausibly associated with changes in consumer sentiments as well as household spending. Finally, using non-linear machine learning models such as LSTM to improve forecasting performance seems to be hindered by the limited data points as we collapse article text into a single time series. To get the most out of text, it may be worthwhile to have more granular data points, similar to work by Kalamara et al. (2020) who retain thousands of terms retained from text as a larger set of time series that can be used with machine learning models.

References

Aguilar, P., Ghirelli, C., Pacce, M., & Urtasun, A. (2021). Can news help measure economic sentiment? An application in COVID-19 times. *Economics Letters*, *199*, 109730. https://doi.org/10.1016/j.econlet.2021.109730

Alexopoulos, M., & Cohen, J. (2015). The power of print: Uncertainty shocks, markets, and the economy. *International Review of Economics and Finance*, *40*. https://doi.org/10.1016/j.iref.2015.02.002

Ardia, D., Bluteau, K., & Boudt, K. (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting*, *35*(4), 1370–1386. https://doi.org/10.1016/j.ijforecast.2018.10.010

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Bloom, N. (2014). Fluctuations in Uncertainty. *Journal of Economic Perspectives*, *28*(2). https://doi.org/10.1257/jep.28.2.153

Buckman, S. R., Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). News Sentiment in the Time of COVID-19. *FRBSF Economic Letter*, *08*.

Calomiris, C. W., & Mamaysky, H. (2019). How news and its context drive risk and returns around the world. *Journal of Financial Economics*, *133*(2). https://doi.org/10.1016/j.jfineco.2018.11.009

Cicceri, G., Inserra, G., & Limosani, M. (2020). A Machine Learning Approach to Forecast Economic Recessions—An Italian Case Study. *Mathematics*, *8*(2), 241. https://doi.org/10.3390/math8020241

Correa, R., Garud, K., Londono-Yarce, J.-M., & Mislang, N. (2017). Constructing a Dictionary for Financial Stability. *IFDP Notes*, *2017*(33), 1–7. https://doi.org/10.17016/2573-2129.33

Fraiberger, S. P. (2016). *News Sentiment and Cross-Country Fluctuations*.

García, D. (2013). Sentiment during Recessions. *Journal of Finance*, *68*(3). https://doi.org/10.1111/jofi.12027

Gentzkow, M., & Shapiro, J. M. (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, *78*(1). https://doi.org/10.3982/ECTA7195

Hopp, D. (2021). Economic Nowcasting with Long Short-term Memory Artificial Neural Networks (LSTM). *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3855402

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International AAAI Conference on Weblogs and Social Media*.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2020). Making Text Count: Economic Forecasting Using Newspaper Text. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3610770

Larsen, V. H., & Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, *210*(1), 203–218. https://doi.org/10.1016/j.jeconom.2018.11.013

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Moore, A. (2017). Measuring Economic Uncertainty and Its Effects. *Economic Record*, *93*(303). https://doi.org/10.1111/1475-4932.12356

Newman, N., Fletcher, R., Schulz, A., Andi, S., & Kleis Nielsen, R. (2020). *Reuters Institute Digital News Report 2020* (p. 99). Oxford: Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Nguyen, K., & Cava, G. L. (2020). *Start Spreading the News: News Sentiment and Economic Activity in Australia*. https://www.rba.gov.au

Nothman, J., Qin, H., & Yurchak, R. (2019). *Stop Word Lists in Free Open-source Software Packages*. https://doi.org/10.18653/v1/w18-2502

Nyman, R., Kapadia, S., & Tuckett, D. (2021). News and narratives in financial systems: Exploiting big data for systemic risk assessment. *Journal of Economic Dynamics and Control*, *127*. https://doi.org/10.1016/j.jedc.2021.104119

Rambaccussing, D., & Kwiatkowski, A. (2020). Forecasting with news sentiment: Evidence with UK newspapers. *International Journal of Forecasting*, *36*(4), 1501–1516. https://doi.org/10.1016/j.ijforecast.2020.04.002

Richardson, A., Mulder, T., & I Vehbi, T. (2018). Nowcasting New Zealand GDP using machine learning algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3256578

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2020.07.053

Shapiro, A. H., & Wilson, D. J. (2019). Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives Using Text Analysis. *Federal Reserve Bank of San Francisco, Working Paper Series*. https://doi.org/10.24148/wp2019-02

Thorsrud, L. A. (2020). Words are the New Numbers: A Newsy Coincident Index of the Business Cycle. *Journal of Business & Economic Statistics*, *38*(2), 393–409. https://doi.org/10.1080/07350015.2018.1506344

Using News Sentiment for Economic Forecasting

A Malaysian Case Study

Eilyn Chong, Chiung Ching Ho, Zhong Fei Ong and Hong H. Ong

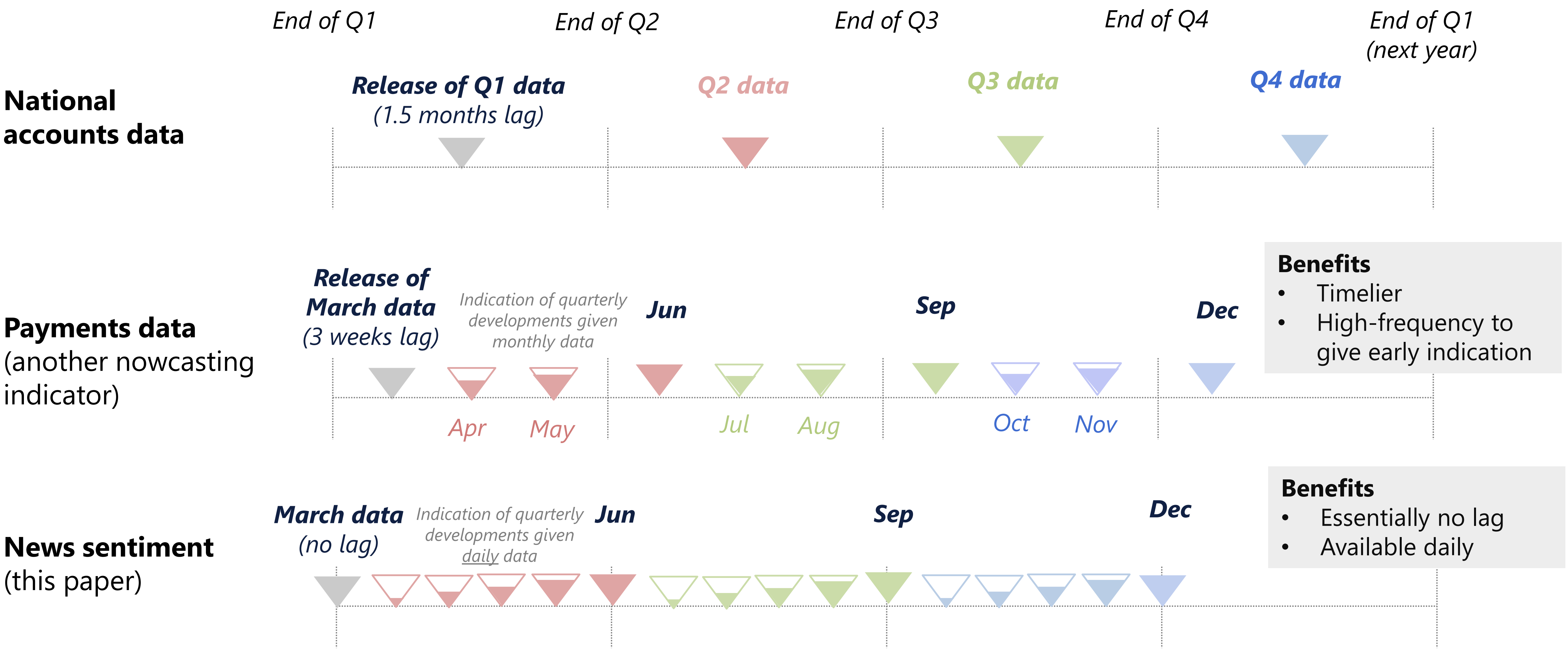
Bank Negara Malaysia

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 or of anyone else associated with the Central Bank of Malaysia.

Outline

- Motivation
- Literature Review
- Data
- Methodology
- Selection of results and conclusions

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



End of Q1
End of Q2
End of Q3
End of Q4
End of Q1 (next year)
Release of Q1 data (1.5 months lag)
Q2 data
Q3 data
Q4 data
Release of March data (3 weeks lag)
Indication of quarterly developments given monthly data
Jun
Sep
Dec
Apr
May
Jul
Aug
Oct
Nov
March data (no lag)
Indication of quarterly developments given daily data
Jun
Sep
Dec

National accounts data

Payments data (another nowcasting indicator)

Benefits

- Timelier
- High-frequency to give early indication

News sentiment (this paper)

Benefits

- Essentially no lag
- Available daily

A growing literature has made use of the wide availability of text in digital format, developments in computational linguistic methods and better computing power

- **Computational linguistic methods have been applied to text in digital format to provide quantitative measures**
  - Economic policy uncertainty (Alexopoulos and Cohen, 2015; Baker, Bloom and Davis, 2016)
  - Daily economic sentiment (Buckman et al., 2020; Thorsrud, 2020)
  - Others: Media slant (Gentzkow & Shapiro, 2010) and central bank's objective function (Shapiro & Wilson, 2019)

- **The papers also relate the text-based measures to a variety of economic and financial outcomes.**
  - Text from news with imminent financial distress (Manela & Moreira, 2017; Nyman et al. , 2021)
  - Text from news/firms' annual reports with stock returns (Calomiris & Mamaysky, 2019; García, 2013; Jegadeesh & Wu, 2013,' Loughran & McDonald, 2011)
  - Text-based measures of uncertainty with business cycle (Bloom, 2014; Moore, 2017)

- **This paper is closest to literature that use news sentiment to track and predict a range of macroeconomic variables.**
  - Aguilar et al. (2021); Fraiberger (2016); Kalamara et al. (2020); Larsen & Thorsrud (2019); Nguyen & Cava (2020); Rambaccussing & Kwiatkowski (2020) show that newspaper text contains information on the future path of certain macroeconomic variables, e.g. GDP.

Data: Our news corpus consists of over 720 thousands economic and financial news articles from major news portals since early 2000s

Descriptive statistics of articles from selected local news portals

Table 1

	Average number of articles per month	Date of first online article	Readership (selected news media)[1]
klse.i3investor	7898	03/03/2020	
The Star	949	01/01/2003	30%
The Edge Markets	912	16/01/2009	
Malay Mail	827	18/06/2013	8%
Bernama	654	04/03/2020	
The Malaysian Reserve	487	10/01/2017	
Free Malaysia Today	384	31/12/2015	15%
The Borneo Post	284	23/12/2009	
The Sun Daily	280	15/11/2017	
SoyaCincau	273	29/01/2020	
New Straits Times	238	20/05/2014	10%

Note: 1. Based on the Reuters Institute Digital News Report 2020 (Newman et al. 2020). Readership is defined as the share of respondents who consumed the media at least once a week.

Sources: Authors' calculation

Methodology – Sentiment scoring: We apply the dictionary-based method on the news corpus to construct the news sentiment index

- Leverage on the lexicon approach using:
  - Financial stability dictionary by Correa et al. (2017) (hereafter Correa)
  - Finance-oriented dictionary by Loughran & Mcdonald (2011) (hereafter LM)
  - Lexicon created by Shapiro, Sudhof and Wilson (2020) (hereafter SSW), who scored a corpus of U.S. economic news articles with Vader (scores ranging from -4 to 4)

- Construction of news sentiment index using words w in each news article i :
  - Correa and LM: $sentiment\ index_i = 100 + \frac{\sum Positive_i - \sum Negative_i}{Word\ count_i} \times 1000$
  - SSW: $sentiment\ index_i = 100 + \sum_w \frac{Score_{w,i}}{Word\ count_i} \times 1000$

- $sentiment\ index_i$ is then averaged for a chosen frequency, whether it be daily, monthly or quarterly.

Methodology: We assess whether news sentiment can nowcast survey-based sentiment & compare the forecasting performance of models with news sentiment to AR (1) model

Various daily news sources

Text pre-processing

- Remove punctuations, hyperlinks, HTML tags, special characters, extra white spaces
- Drop stop words
- Set words in lowercases

Sentiment indices using lexicons, x_t
(2006 – 2021 Q2)

- Correa
- LM
- SSW

1 **Nowcasting quarterly survey-based sentiment**
(2006 – 2021 Q2)

- Business conditions index
- Consumer sentiment index

2 **Forecasting target variables,**
y_t (2006 – 2021 Q2)

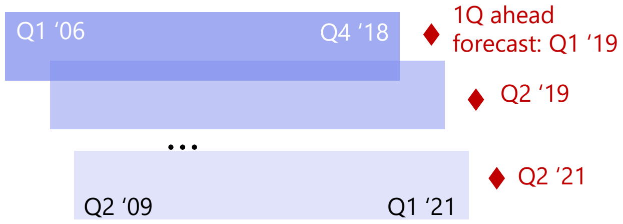
- Aggregate GDP
- Private consumption
- Private investment
- Imports
- Exports

Split into train-test period

Train **Test**

Rolling-window estimations comparing forecasts from…

- **OLS-AR1**
  $y_{t+h} = \alpha + \beta\, y_{t-1} + \epsilon_t$
- **AR1 with sentiment**
  $y_{t+h} = f\,(y_{t-1}, \boldsymbol{\eta}\, \boldsymbol{x_t})$
  where $x_t \in \{Correa_t, LM_t, SSW_t\}$

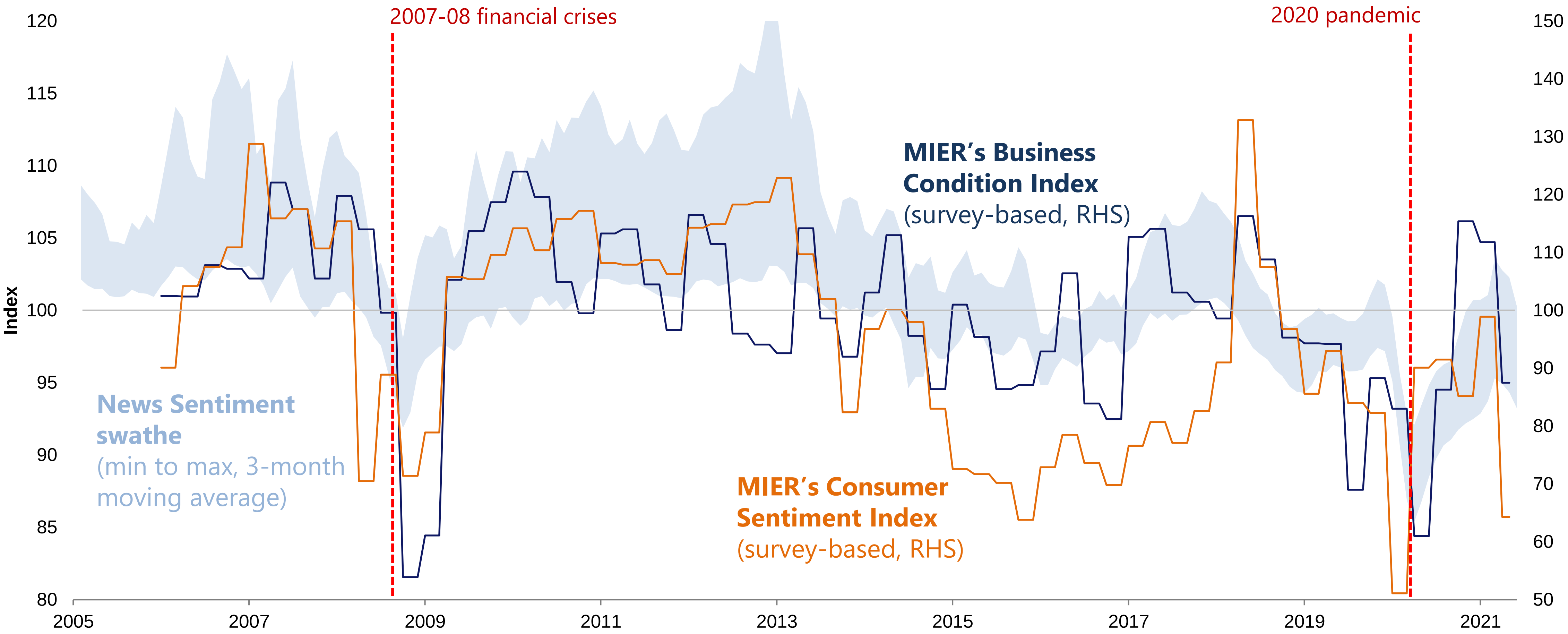


Evaluation of h-quarter ahead forecast

$$Ratio\ RMSE = \frac{RMSE_{model_i}}{RMSE_{AR1}}$$

$$for\ h = 1,2,3$$

The monthly news sentiment measures move with fluctuations in economic conditions and exhibit the sharp declines during the 2007-08 financial crisis and onset of COVID19



Monthly news sentiment provides information in nowcasting movements in the quarterly survey-based measures of business sentiment

Nowcasting Quarterly Survey-Based Sentiment

Dependent variable: MIER’s Business Conditions Index (BCI) in the same quarter
Sample: 2006 Q1 – 2021 Q2

Table 3

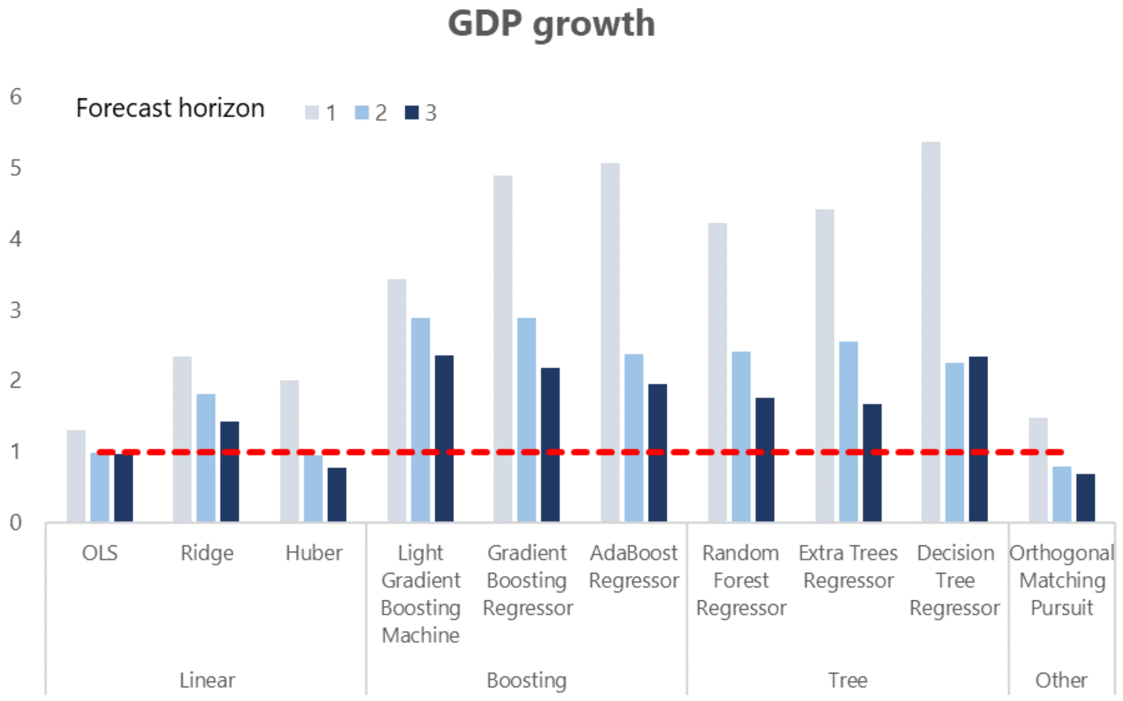
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	AR1	Correa	Correa	Correa	LM	LM	LM	SSW	SSW	SSW
MIER’s Sentiment (1 quarter prior)	0.49***	0.36***	0.33***	0.33***	0.35***	0.34***	0.35***	0.35***	0.31***	0.33***
	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)
News sentiment (first month of the quarter)		0.30**			0.32**			0.41***		
		(0.14)			(0.14)			(0.14)		
News sentiment (average first 2 months of the quarter)			0.41***			0.39***			0.52***	
			(0.15)			(0.14)			(0.15)	
News sentiment (full data for the quarter)				0.45***			0.40***			0.48***
				(0.14)			(0.13)			(0.14)
Constant	-0.01	-0.01	-0.01	0.02	-0.03	-0.02	0.00	-0.02	-0.03	-0.02
	(0.11)	(0.11)	(0.11)	(0.10)	(0.11)	(0.11)	(0.11)	(0.10)	(0.10)	(0.10)
Adjusted R^2	0.22	0.29	0.34	0.35	0.30	0.32	0.32	0.35	0.39	0.35
Observations	61	61	61	61	61	61	61	61	61	61

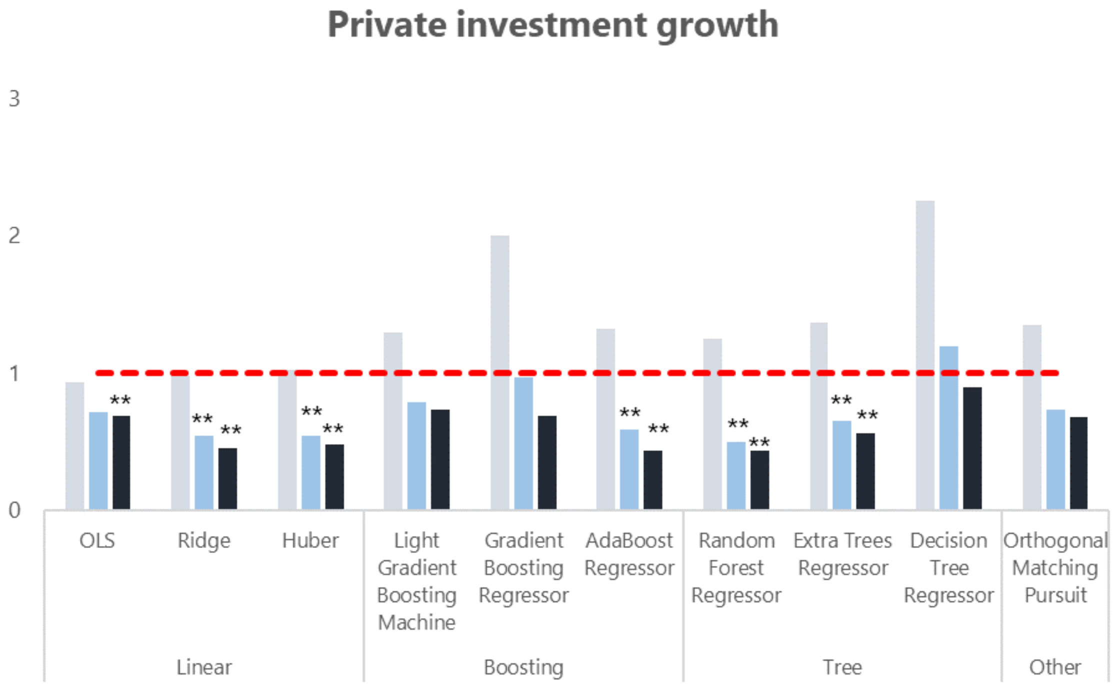
Heteroscedastic and autocorrelation robust (HAC) standard errors in parentheses. * p < 0:10 ** p < 0:05 *** p < 0:01

Sources: Authors’ calculations

Among the GDP components, the news sentiment is predictive of private investment growth 2-3 quarters ahead

Ratio of RMSE of models with news sentiment vs. RMSE of OLS-AR(1)
(average across time period and sentiment measures)





Bars below the dashed red line indicate an improvement in forecast performance, and those with ** exhibit a forecast improvement for at least 90% of the forecast periods when compared to the OLS with just the AR(1) term only

Conclusion

1. 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.

2. Among the GDP components, the news sentiment is able to reliably forecast private investment growth 2 – 3 quarters ahead.

Future work

Explore and compare the sentiment in news of vernacular languages (e.g. Malay, Chinese and Tamil).

Generate new models to show better predictive accuracy than existing models by taking into the account of articles' ideological predisposition.

IFC-Bank of Italy Workshop on “Machine learning in central banking”

19-22 October 2021, Rome / virtual event

Machine learning real-time CPI forecasting[1]

Mariam Mamedli,
National Research University, Higher School of Economics, Moscow

[1] This presentation 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 central banks and other institutions represented at the event.

Forecasting Russian CPI with data Vintages and Machine learning techniques[1]

Mariam Mamedli
National Research University Higher School of Economics
Email: mmamedli@hse.ru

Denis Shibitov
Bank of Russia
Email: ShibitovDS@mail.cbr.ru

Abstract

We show, how the forecasting performance of models varies, when certain inaccuracies in the pseudo real-time experiment take place. We consider the case of Russian CPI forecasting and estimate several models on not seasonally adjusted data vintages. Particular attention is paid to the availability of the variables at the moment of forecast: we take into account the release timing of the series and the corresponding release delays, in order to reconstruct the forecasting in real-time. In the series of experiments, we quantify how each of these issues affect the out-of-sample error. We illustrate, that the neglect of the release timing generally lowers the errors. The same is true for the use of seasonally adjusted data. The impact of the data vintages depends on the model and forecasting period. The overall effect of all three inaccuracies varies from 8% to 17% depending on the forecasting horizon. This means, that the actual forecasting error can be significantly underestimated, when inaccurate pseudo real-time experiment is run. We underline the need to take these aspects into account, when the real-time forecasting is considered.

Kew words: inflation, pseudo real-time forecasting, data vintages, machine learning, neural networks.

JEL-classification: C14, C45, C51, C53.

[1] The views expressed in this paper are solely those of the authors and do not necessarily reflect the official position of the Bank of Russia. The Bank of Russia assumes no responsibility for the contents of the paper. Any reproduction of these materials is permitted only with the express consent of the authors.
The authors are grateful to Sergei Seleznev (Bank of Russia), Dmitry Gornostaev (Bank of Russia), participants of 27th Annual Symposium of the Society for Nonlinear Dynamics and Econometrics at FRB Dallas (2019) and the participants of seminars at the Bank of Russia for their helpful comments.

Table of Contents

Introduction

Although it is common in empirical macroeconomics to work with the revised data, a growing body of literature suggests, that analysis using real-time data often leads to the substantially different conclusions, than the work which ignores data revisions (Croushore, Stark, 2001; Orphanides, 2001; Koenig, 2003; Molodtsova et al., 2008). More and more evidence indicate the importance of using real-time data, when constructing forecasting models and performing monetary policy analysis (Fernandez et al., 2011).

In this research, on basis of the unique for Russia real-time dataset we aim to define, how different inaccuracies during pseudo-real time experiments affect the estimates of the model performance. Namely, we focus on the use of not-seasonally adjusted vintages of timeseries (data which include initial values and revisions), consider it as the data for an "ideal" experiment and consequently add one of the inaccuracies to define the impact, each of them has on the estimation of out-of-sample model performance (linear regression, random forest, gradient boosting, neural networks). Our focus is not on the choice of the optimal modelling techniques, but on constructing real-time forecasting exercises.

The ability to forecast accurately inflation is a crucial for the development of monetary policy by the central bank, that is why we consider the CPI forecasting and investigate, how these aspects can affect the CPI forecasting errors. Within this task, we construct the forecasts for four sets of data in order to define an impact of several inaccuracies of pseudo-real-time forecasting. In the *first estimation experiment* we run our models on the vintages of original, not seasonally adjusted, timeseries, taking into account the timing of the data releases. By the *second experiment* we determine, how the results would change, if we include all data without any regard to the data availability, when the forecasts are made. This aspect is important, because macroeconomic and financial data usually have different release lags. However, often macroeconomic forecasting is based on the data, that include all available up to that point timeseries. If the forecasts were constructed in real time, some of these series may not have been yet released. By comparing the first and the second experiment we can quantify the effect of data release timing. By the *third experiment* we define the role of data vintages: how models' forecasting performance would change, if we neglect data revisions and take the final values, which in reality are not available, when the forecast is made. In the *last experiment* the forecasts are based on seasonally adjusted data without vintages and not taking into account the release dates. Seasonal adjustment is a subjective procedure, which leads to the analysis of unobservable series, highly dependent on the applied method of seasonal adjustment. Moreover, the use of seasonally adjusted data complicates the comparison between research results, while not seasonally adjusted data allow to compare the results in terms of observable variables, making the forecasting with not seasonally adjusted data more correct. Nevertheless, the combination of these data characteristics (seasonally adjusted revised data and the neglect of the data availability) often occurs in the forecasting literature, underlining the need to define their contribution to the forecasting error.

Main results of this research come from the comparison of the results between the experiments. *First,* we show that the estimates based on the data, which incorporate the series with different lags (when we take into account the release timing of series) generally have higher out-of-sample forecasting error, in comparison to the case, when all the series are treated as available at the time of forecast (for 8 out of 12 cases). This is an expected result, since in this case for some timeseries we use less recent data: when the forecasts are made, vintages are available with two months lag. This difference amounts to 10,5% on average for all four models for one month in advance (8,0% and 1,7% for three and six lags) *Second,* we compare estimation results based on the vintages and the data after revisions. In this case, the results are inconclusive, and do not point in favor of one experiment or the other. Some models and forecasting horizons are more sensitive to the use of data vintages and have lower forecasting error, some do not (in 7 out of 12 cases forecasting on revised data has an advantage). *Finally,* the last comparison shows, that the use of seasonal adjustment leads to significantly lower forecasting errors for most of the models and forecasting horizons. Overall, the joint effect of all three inaccuracies varies from 7,6% to 16,8% of average monthly CPI depending on the forecasting horizon.[2] This may result in the considerable underestimation of the error in the case of real-time forecasting: when part of timeseries is not published yet (different timing of releases), for some of them only preliminary numbers are available (data vintages) and forecasts are based on the seasonally adjusted data, which lower the error further.

In addition to the error decomposition depending on certain inaccuracies in prediction exercises, we contribute to the literature on Russian CPI forecasting in the following way:

This is the first research, which uses Russian data vintages in economic forecasting.[3]

We propose first reproducible real-time benchmark for Russian inflation forecasting.

In our research we consider a variety of popular machine learning models (including ensemble methods and Bayesian neural networks).

Although, the methodology of the research is relatively new for the forecasting on the Russian data, the application of machine learning (ML) techniques is known to have a great potential for macroeconomic forecasting. Short overview of the related literature on the application of ML models in economic forecasting is presented in the following section.

Alternative, traditional for macroeconomic research approaches to the forecasting in data-rich environment include pooling or averaging of bi-variate forecasts (Stock, Watson (2003); Rossi, Sekhposyan, 2010) and direct pooling of information using a high-dimensional models (see Forni et al. (2000, 2005); Stock, Watson (2002a, b) for DFM, Banbura et al. (2010) for Bayesian VAR).[4] However, an analysis of a wide range of models is out of the scope of this paper, so we do not cover these methods in details here.

The paper is organized as follows. Next section gives an overview of the relevant research on analysis on real-time data and examples of ML applications in economic forecasting. Third section provides an overview of the considered experiments and details on the data. Next cross-validation, optimization procedure and models, applied to the CPI forecasting, are described. Then we provide the estimation results and show the impact of each data characteristic to the out-of-sample forecasting error. The last section concludes.

2. Related research

2.1. Real-time data

Croushore (2011) provides an overview of the existing research on real-time data analysis, dividing it into six areas: data revisions, structural macroeconomic modelling, forecasting, monetary policy, current analysis and revisions to conceptual variables. The research shows that real-time data matter in a variety of contexts. Overall, the results show, that the forecasting ability in real time is much worse, than the forecasting ability,

[2] Average monthly CPI was calculated on the basis of data included in the test set.

[3] The details on the data are provided in Ponomarenko et al. (2021).

[4] Among the recent research on the forecasting of Russian CPI one may name Styrin (2019), where the CPI dynamics is predicted using dynamic model averaging.

when the revised data are used. Moreover, forecasts of levels are very sensitive to data revisions, whereas forecasts of growth rates are much less sensitive (Howrey, 1996). Filardo (1999) shows, how unreliable in real time are models, which attempt to predict recessions, as these models are usually based on revised data. As for the inflation, Koenig (2003) shows, that while markup is a useful predictor of inflation with revised data it fails to predict inflation in real-time. Orphanides, Norden (2005) illustrate, that in real-time the estimation of output gaps is so much affected by uncertainty, that they cannot be reliably used in inflation forecasting. Forecasts of exchange rates are known to be even more sensitive to real-time data issues (Faust et al., 2003; Molodtsova, 2008; Molodtsova et al., 2008).

The question of whether the use of real-time data leads to the different forecasts, than when the latest-available data are used goes back to the first paper by Denton, Kuiper (1965), who on the case of Canada find, that the use of real-time data or latest-available data leads to the significant differences in the forecasts. Cole (1969) also shows, that data errors can reduce forecast efficiency and lead to biased forecasts, as the result, there can be significant differences between forecasts made with different data sets. Similar results were obtained by Trivellato, Rettore (1986), who showed on Italian data, that data errors in a simultaneous-equations model have large effects.

Series of papers afterwards advocated in favor of using real-time data and bring attention to the consequences of using real-time data as opposed to the latest available. Stark, Croushore (2002) discuss, how the forecasts are affected by the use of real-time data rather than latest-available data. They bring attention to the fact, that in the forecasting literature the results are usually obtained using the data set available to the model's developer, but the data would not have been available to him in real-time. They investigate, how the vintage data affect such forecasts and advocate in favour of real-time data rather than latest-available data. Forecasts made for a particular date can be quite different, depending on the vintage of data used: the RMSEs and MAEs of forecasts can differ between real-time data and latest-available data, when only short spans of observations are used, and been misleadingly low when latest-available data are used. They also find, that inflation forecasts are more sensitive to the choice between real-time and latest-available data, than real output forecasts. Stark, Croushore (2002) also show that the choice of lag length depends on whether latest-available data or real-time data are used.

Croushore, Stark (2003) examine the nature of data revisions and investigate the robustness of the results of several key papers in macroeconomics (Kydland, Prescott, 1990; Hall, 1978; Blanchard, Quah, 1989) to different vintages. They show, that only the results of the former paper are robust to the use of different data vintages, underlining an importance of real-time data.

Koenig et al. (2003) argue, that analysts should generally use data of as many different vintages as there are dates in the samples. More specifically, at every date within a sample, right-hand-side variables ought to be measured as they would have been at that time (so called "real-time-vintage data"). They consider three different ways of using real-time data (depending on whether the real-time-vintage data on the left-hand side are used) and show on the example of GDP forecasting that out-of-sample forecasting performance of the model estimated using real-time-vintage data is superior to the one, obtained using conventional estimation. They advocate, that the most popular approach of using end-of-sample-vintage data (estimation strategy 3 in our case) should generally be avoided. Kishor, Koenig (2012) present a method of adopting VAR analysis to account for data revisions. They apply the technique to employment and unemployment rate, real GDP and the GDP/consumption ratio and show, that in each case the proposed procedure outperforms the conventional VAR analysis.

Clements, Galvao (2009) provide evidence in favor of real-time vintage data within MIDAS model. In the following paper Clements, Galvao (2011) show, that a certain class of models can be used to forecast 'fully revised' or 'post-revision' values of past and future observations, estimating the value of those forecasts in terms of their contribution to improving real-time estimates of the output gap, trend inflation and inflation gap. Their research was followed by Clements, Galvao (2013), who conduct an evaluation of vintage-based VAR model forecasts for US inflation and output of a range of maturities of data using a variety of different target variables (forecasting future observations or revisions to past data). They show, that VAR models on a single variable estimated on data vintages can be successful in forecasting the data revisions process of inflation, but are less useful for US output growth. They also find only some evidence, that vintage-based VAR models provide more accurate forecasts of output growth, but clear evidence, that revisions to past inflation data are predictable. In the case of predicting Russian CPI, revisions can play a role only via explanatory variables since CPI series are not subject to any revisions.

Being a brand-new practice for Russia, real-time datasets are more common in other countries. Croushore, Stark (2001) compiled and analyzed a large real-time dataset for macroeconomists on the US economy starting from 1965, bringing the attention to this subject. Later McCraken, Ng (2016) formed a big database for macroeconomic research, updated in real-time through Federal Reserve Economic Data (FRED) database. The data include historical vintages from august 1999 and is widely used in the literature.

There exist several datasets with vintages on European economies as well. Giannone et al. (2012) presented a real-time database for the euro area. Fernandez et al. (2011) introduce a new international real-time dataset on OECD countries and illustrate the importance of using real-time data in macroeconomic analysis by considering several economic applications conducted in real-time perspective on the data on G7 economies. Being one of the first multicounty real-time datasets Fernandez et al. (2011) contributed to the papers with datasets on individual countries. Some examples of the data vintages on other countries include Egginton et al. (2002), who presented a real-time macro dataset for the UK, Clausen, Meier (2005); Sauer, Sturm (2007) and Gerberding et al. (2005), who collected the real-time data for Germany, and Nikolsko-Rzhevskyy (2011), who proposed a methodology of estimation forward-looking Taylor rules in real-time and illustrated it on the example of UK, Germany and Canada.

2.2. Applications of machine learning in economic forecasting

Although been relatively new in application to the Russian data, ML models are widely used in international economic research. Tiffin (2016) along with Chakraborty, Joseph (2017) and Kapetanios, Papailias (2018) outline the potential of ML for central banking and policy analysis and provide a broad overview of the key ML techniques, their advantages and limitations. The comparison of forecasts of the standard and ML models provides evidence in favour of ML models (e.g. Cook, Hallyz, 2017). There is an evidence in favour of ML techniques, especially neural networks (NNs), in forecasting the CPI in many other countries (Moshiri, Cameron, 2000; Chen et al., 2001; Szafranek, 2019; Hanif et al., 2018). The performance of NNs in forecasting inflation was also evaluated within a cross-country comparison with the results in favor of ML models (McAdam, McNelis, 2005; Choudhary, Haider, 2012). At the same time, there is opposite evidence, that NNs in fact does not outperform the standard models (Kock, Tersvirta, 2013; Catik, Karauka, 2012; lvarez-Daz, Gupta, 2015; Sermpinis et al., 2014; Zhang, Li, 2012).

There is less evidence of applying boosting techniques in the CPI forecasting. However, Buchen, Wohlrabe (2011) apply it to the forecasting US industrial production, showing that boosting can be a serious competitor to other methods in the short run and that it performs best in the long run. Dpke et al. (2017) show that boosting has a better out-of-sample performance, than probit models for the recessions prediction in Germany.

To the best of our knowledge, this is the first research on forecasting abilities of neural networks, gradient boosting and random forest with optimal architectures in the case of Russian CPI. The closest to this research is Baybuza (2018), who shows, that random forest and gradient boosting can outperform standard methods on the horizon of two and more months in forecasting Russian CPI. However, he does not choose optimal hyperparameters, considering the predetermined number of estimators. We consider data vintages and illustrate the impact of the data used on ML models with optimal architectures.

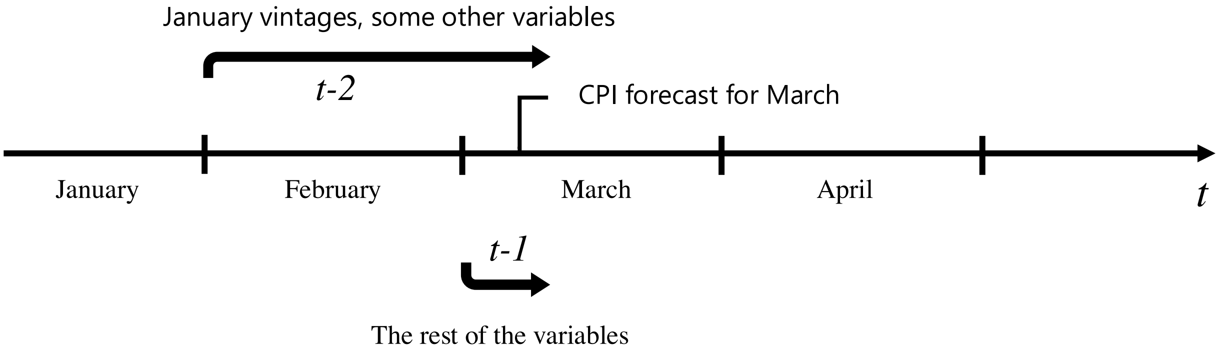
Taking into account, that the results from the other countries are inconclusive with respect to the forecasting abilities of these models, our research, along with its main goal, fills this gap by analysing the performance of ML models in application to the Russian CPI.

3. Experiments and data

In our key estimation experiment we use vintages of official data releases, described in detail in Ponomarenko et al. (2021). The data cover the period from January 2001 to July 2019 and consist of timeseries of macroeconomic variables for each release date, including all revisions. At each point in time, the forecast is made on the basis of the latest (not revised) data. The data include 27 time series on economic activity, interest rates, price indexes, international trade and others. For most of the variables the vintages of data are used: timeseries at each date, as they were initially published by the Federal statistical service. Table A1 in Appendix provides an overview of series included in the dataset.[5] For experiments with not seasonally adjusted data we include months dummies.

We consider the forecasts for four datasets in order to define the impact of main data characteristics. In the *first experiment* we estimate models on the vintages of not seasonally adjusted timeseries, taking into account the timing of the data releases (they are presented in Table A1). In each experiment we assume, that we construct our forecasts at the beginning of each month, when the data on vintages for two months ago are released. By this moment the CPI data and most of the financial variables for the previous month are already released (Figure 1). Therefore, at the moment of forecast we have part of the data for the previous month ($t - 1$ lag) and the rest of the data with two months delay ($t - 2$ lag). In order to reconstruct properly the real-time forecasting procedure this difference in the data for the latest available dates should be taken into account.

Figure 1. Example of the timeline of data releases


January vintages, some other variables
t-2
CPI forecast for March
January
February
March
April
t
t-1
The rest of the variables

[5] Data sources are Federal State Statistical Service, Central Bank of Russia, EIA, Roskazna.

In the *second experiment* we include timeseries without the regard to the data availability at the moment of forecast: all variables are taken with one month lag. In the *third experiment* we estimate models on the revised data and define, how the use of data vintages affects the forecasting performance. In the *last experiment* the forecasts are based on seasonally adjusted data without vintages and not taking into account the release dates.[6] We advocate for the forecasting of not seasonally adjusted CPI, first, because it is the commonly recognized indicator, as opposed to the seasonally adjusted data, which is the result of the adjustment procedure, which may differ from one research to the other. Second, the original CPI values are not revised, while in the case of seasonal adjustment historical values are revised each time, which may alter the estimations in the unknown manner. The last experiment illustrates the joint effect of all three inaccuracies. Table 1 provides an overview of the experiments with the references to corresponding subsections.

Table 1. Different characteristics of the experiments

Experiment	Starting lags	Vintages	Seasonality
1	$t-1/t-2$	vintages	NSA
2	$t-1$	vintages	NSA
3	$t-1$	regular	NSA
4	$t-1$	regular	SA

Note: yellow – subsection 5.2, blue – subsection 5.3, green – subsection 5.4.

4. Models and estimation procedure

4.1 Cross-validation and estimation procedure

In order to choose the optimal model architecture, we apply the cross-validation procedure. We split the dataset in train and test subsets: train set includes observations up to December of 2012, following observations are included in the test set. On the train set we apply cross-validation with an expanding window.[7] Starting from a minimum size window of 48 observations we train each model and test it on the next observation from the train set.[8] We consequently add next observation to the training data and re-estimate the model. We evaluate the out-of-sample model performance on the test set.

In order to define the optimal architectures of the models, we use Bayesian optimization instead of the grid search. The parameters are chosen so to minimize RMSE on the training set. This procedure allows us to decrease significantly the estimation time (roughly six times faster). Moreover, preliminary results, based on the cross-validation with grid-search, show that Bayesian optimization not only is a faster joint optimization tool, but allows us to define more optimal in terms of out-of-sample RMSE parameter combinations due to the search on continuous intervals instead of fixed combinations. For details on Bayesian optimization see Shahriari et al. (2015).

[6] The seasonal adjustment was conducted in Demetra program with tramoseats method and rsa3 specification.

[7] For the first experiments we have compared the results for both, expanding and rolling window. RMSE levels were lower in the case of
expanding window for all models and both experiments.

[8] The size of minimum window was chosen experimentally.

Each of four models (linear regression, random forest, gradient boosting, and Bayesian neural network) are applied to forecast Russian CPI for one, three and six months in advance. For the last two horizons we predict the accumulated inflation (CPI for three months and for six months in advance).

We estimate each model for each forecasting horizon on the dataset with one, three or six lags. This brings the total number of estimations for each experiment to 36. Overall, 144 estimations were made. Next subsections briefly describe the estimated models.

4.2. Regression with regularization

As a linear model we consider a regression with regularization (elastic net), which is advised, when features are correlated with each other (Friedman et al., 2010). This model combines ridge regularization penalty and Lasso penalty. Via cross-validation we choose the optimal type of regularization as well as other optimization parameters.

Formally it is defined as follows:

$$\min_{w} \frac{1}{2N}\left(\|Xw - y\|_2^2 + \alpha\rho\|w\|_{l_1} + \frac{\rho(1-\alpha)}{2}\|w\|_{l_2}^2\right), \tag{1}$$

where y is a target variable, X is a vector of predictors, β is vector of coefficients and N is the number of observations.

4.3. Random forest

Random forest is an ensemble model proposed by Breiman (2001).[9] The algorithm of random forest is based on decision trees. Each tree is a graph model, which consists of a set of rules on explanatory variables to obtain the target variable. This model has a tree structure with nodes as decision points. The split occurs according to a certain criterion on one of the explanatory variables, while terminal nodes (leaves) contain the value of the target variable. The decision tree is built in a stepwise manner: first, the sample is split into two subsamples according to the specified criterion, then each of subsamples is consequently split further, until a certain stop criterion is not reached.[10]

The random forest model can be expressed as follows. Suppose we have N observations (x_i, y_i) for $i = 1,2,\dots,N$ with $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and M regions $R_1, R_2, \dots, R_M$. The response is modeled as a constant c_m in each region:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m). \tag{2}$$

We consider a splitting variable j and split point s and define the pair of half-planes:

$R_1(j,s) = \{X | X_j \leq s\}$ and $R_2(j,s) = \{X | X_j > s\}$.

[9] For the estimation we use scikit-learn Python package.

[10] Algorithm is presented in detail in Friedman et al. (2009), p. 308.

Then the splitting variable *j* and the split point *s* are defined via minimization problem with the chosen minimization criterion. After the best split is defined, the data are divided accordingly and the splitting process is repeated for these two regions. The procedure is repeated on all of the resulting regions.

Via cross-validation we choose the architecture, which ensures the best forecasting performance of the models. Considered parameters are presented in Table A2 in Appendix.

4.4. Gradient boosting

Gradient boosting is another ML algorithm based on the combination of predictive models, decision trees in our case, so to minimize the loss function.[11] In this form it was proposed by Friedman (2001). Gradient boosting can be used both for classification and regression, in our case the problem belongs to the second type. Boosting allows to identify outliers and to exclude them from the training set. However, it is known to have a tendency to overfit, while the stepwise approach of this algorithm can lead to a non-optimal set of weak learners. That is why it is highly important to choose optimally the combination of the number of estimators and learning rate, which can help to avoid overfitting.

Analytically gradient boosting is expressed as an additive sum of simpler models:[12]

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x), \tag{3}$$

where $h_m(x)$ are decision trees, and $\gamma_m(x)$ is a step length.

Gradient boosting is built in a stepwise manner in the following way:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x). \tag{4}$$

On each step the decision tree $h_m(x)$ is chosen optimally from the minimization of the loss function L with a given F_{m-1} and $F_{m-1}(x_i)$:

$$h_m(x) = \underset{h,\beta}{\operatorname{argmin}} \sum_{i=1} L\big(y_i, F_{m-1}(x_i) + \beta h(x_i)\big). \tag{5}$$

The minimization problem is approximately solved via fitting steepest descent directions (negative gradient of the loss function evaluated at the current model F_{m-1}) by new weak learner. The step length $\gamma_m(x)$ is chosen according to the equation (6):

$$\gamma_m(x) = \underset{h,\beta}{\operatorname{argmin}} \sum_{i=1} L\big(y_i, F_{m-1}(x_i) - \gamma h_m(x)\big). \tag{6}$$

When defining the optimal architecture, we consider different model characteristics, including the specification of loss function, maximum depth of a tree and the number of trees. The complete set of hyperparameters is presented in Table A3 in Appendix.

[11] In the baseline case it is a least squares regression.

[12] For the estimation we use scikit-learn Python package.

4.5. Bayesian neural networks

We consider a sparse Bayesian neural network model described in Khabibullin, Seleznev (2020).[13] The model consists of two hidden layers with 30 and 10 nodes. Analytically the model can be presented as follows:

$$h_{1,t} = f(W_1 x_t + b_1), \tag{7}$$

$$h_{2,t} = f\left(W_2 h_{1,t} + b_2\right), \tag{8}$$

$$y_t = W_3 h_{2,t} + b_{y,t} + \varepsilon_t, \tag{9}$$

where x_t are input data, W_i are weights on the layers $i = 1, 2$ and output layer, b_i and b_y are bias, $f(\cdot)$ is an activation function, h_i is the output of the hidden layer i, y_t is the output of the neural network at time t and ε_t is an error term. We consider two model specifications: Normal distribution of the error term and Tanh activation function and tStudent distribution with ReLu activation function.[14]

5. Results

We estimate elastic net, random forest, gradient boosting and Bayesian neural network on four different datasets. For each dataset we construct forecasts for one, three and six months in advance. Along with the optimal model architectures we consider the different numbers of lags for each horizon. In the next subsection we provide the results for the benchmark experiment (experiment 1), conducted on the vintages of not seasonally adjusted time series, where the release dates are taken into account. In the following subsections it is compared to the results of the experiments. In the second subsection the role of release timing is studied (experiment 1 and 2, 'starting lags' in Table 1). Next, we study, how the estimation with the vintages can affect model performance (experiment 2 and 3, 'vintages' category). In the last subsection we show, how the estimates are affected by the use of revised and seasonally adjusted data (experiment 3 and 4, 'seasonality').

5.1. Benchmark experiment

We start by providing the results for a set of considered models, estimated on the data vintages of not seasonally adjusted data. We also take into account the release dates of timeseries: the official data are released with a delay and different lags. In our case, the publications of vintages have a one-month delay comparing to the financial variables and the CPI, which are not revised. In order to replicate a real-time forecasting experiment, we take these publication lags into account and include lagged variables starting in different point in time: exchange rates, oil price and other financial variables are included in the dataset starting from $t-1$ (previous month), for vintages data first observations correspond to two months lag ($t-2$). The details on the data and the delay of first observations are provided in Table A1 in Appendix. In the

[13] The mean-field approximation was used.

[14] Initially for the first experiment all four the combinations of model parameters were considered. These two combinations were chosen as they allow to achieve the lower RMSE level on cross-validation.

cases of three and six months lags we include the equal number of lags starting from $t-1$ or $t-2$ for corresponding series.

Table 2 provides RMSEs, estimated on vintages of not seasonally adjusted timeseries, when the release dates are taken into account (experiment 1 in Table 1). RMSE levels on cross-validation and test for all experiments are provided in Appendix (Tables A7-A10). The results are provided for combination of hyperparameters, which allows to achieve the lowest RMSEs for each forecasting horizon and number of lags (Tables A4-A6, Appendix).

Table 2. The out-of-sample RMSE levels for different models

lags	Elastic Net	Random Forest	Gradient Boosting	Neural network	AR
1 month					
1	0,480	0,368	0,356	0,409	0,404
3	0,403	0,373	0,382	0,431	0,427
6	0,380	0,389	0,396	0,400	0,392
3 months					
1	0,756	0,826	0,820	0,705	1,414
3	0,760	0,883	0,759	0,728	1,393
6	0,789	0,895	0,737	0,860	1,390
6 months					
1	0,758	0,963	0,789	0,687	2,693
3	0,643	0,932	0,716	0,752	2,681
6	0,758	0,986	0,842	0,876	2,664

Note: AR model was estimated with seasonal dummies for proper model comparison.
The lowest RMSE levels for each forecasting horizon are marked with grey.

The results suggest that forecasts on the basis of the gradient boosting model have lower out-of-sample RMSE in the forecasting for one month in advance, random forest forecast is the second best, with both models performing better than AR.[15] For three months neural network forecast has the lowest RMSE, followed by elastic net model. For six month in advance elastic net has the lowest error. In forecasting for three and six months all ML models outperform the AR benchmark, which predicts poorly on these horizons.

Higher AR errors on these horizons are explained by its univariate nature: other models include different macroeconomic variables, informative in predicting future economic developments. Let us consider an elastic net, as more interpretable one among ML models, and a forecasting period of three months with one lag. Additional estimations suggest, that elastic net trained on the whole train set, with optimal hyperparameters “picks” non zero coefficients for variables such as export and import, exchange rate, oil price, production of eggs and meat as well as retail of nonfood goods and others while seasonal dummies have zero or close to zero coefficients. Additional variables capture some seasonal fluctuations and mitigate the autoregressive spikes in CPI if needed. In the case of AR with seasonal dummies this cherry picking is impossible, which makes AR forecasts excessively volatile, when longer horizons are considered due to the change in the seasonal fluctuations of the CPI on the test set with comparison to the train set (see Figure 1 in Appendix).

[15] The models outperform RW benchmark as well, the results of which are omitted in the table with the aim of comparability. For 1, 3 and
6 months in advance the RMSE for RW is 0.41, 1.76 and 3.81, correspondingly.

5.2. The role of release timing

Next, we consider, how taking into account the release timing affects the results. We examine two experiments with not seasonally adjusted data vintages. In the first one we form dataset with the regard to the time, when the latest value of an indicator is released ($t-1/t-2$). In the second experiment we treat all timeseries as available and include them in the dataset with $t-1$ lag. The results for these experiments are provided in Table 3.

Unsurprisingly, for most of the cases an assumption, that the most recent data are available, leads to the lower forecasting errors (for 10,5%, 8% and 1,7% on average for one, three and six months in advance).[16] The exceptions are gradient boosting and random forest forecasts for one month in advance and elastic net forecast for six months in advance. Since usually linear models are used for macroeconomic forecasting the results for elastic net can be more representative: for this model an inclusion of all "available" data can lead to an underestimation of the real forecasting RMSE from 5,5% to 10,6% for one and three months in advance, correspondingly.

Table 3. Comparison of RMSEs of models estimated on data with the same and different lags

	Elastic Net		Random Forest		Gradient Boosting		Neural network	
lags	$t-1$	$t-1/t-2$	$t-1$	$t-1/t-2$	$t-1$	$t-1/t-2$	$t-1$	$t-1/t-2$
				1 month				
1	**0,359**	0,480	0,393	**0,368**	0,378	**0,356**	**0,338**	0,409
3	0,362	0,403	0,446	0,373	0,404	0,382	0,349	0,431
6	0,447	0,380	0,428	0,389	0,375	0,396	0,411	0,400
				3 months				
1	**0,676**	0,756	**0,772**	0,826	**0,710**	0,820	**0,626**	0,705
3	0,697	0,760	0,816	0,883	0,789	0,759	0,787	0,728
6	0,720	0,789	0,814	0,895	0,790	0,737	0,765	0,860
				6 months				
1	0,722	0,758	0,914	0,963	0,735	0,789	**0,685**	0,687
3	0,701	**0,643**	**0,903**	0,932	**0,716**	0,716	0,698	0,752
6	0,743	0,758	0,927	0,986	0,809	0,842	0,916	0,876

Note: The lowest RMSE levels for each forecasting horizon and number of lags are marked with grey.
The lowest RMSE among two experiments for each model is marked with bold.

5.3. The role of data vintages

We can define the role of data vintages in macroeconomic forecasting by comparing the estimates based on not seasonally adjusted data vintages and model results estimated on not seasonally adjusted data with final releases (experiments 2 and 3 in Table 1). The RMSE levels on test sets for four models are presented in Table 4.

For one and three months in advance gradient boosting and random forest provide lower errors, when vintages are not used. For elastic net and neural network, on the contrary, for these forecasting horizons the use of data vintages allows to achieve lower errors. For longer horizon of six months the neural network

[16] Estimations were made for each model with the optimal number of lags. The change is calculated as the difference between 2nd and 1st experiment's RMSEs relative to the latter one.

estimated on the data without vintages provides the lowest RMSE with the significant difference comparing to its results on vintages data. Elastic net and random forest also provide lower errors, when the vintages data are not used, yet the difference between errors is lower and insignificant in the latter case. For gradient boosting the use of vintage data leads to lower RMSE.

Table 4. Comparison of RMSEs of models estimated with and without data vintages

	Elastic Net		Random Forest		Gradient Boosting		Neural network	
lags	vintages	no vintages	vintages	no vintages	vintages	no vintages	vintages	no vintages
1 month								
1	**0,359**	0,369	0,393	**0,373**	0,378	**0,334**	**0,338**	0,348
3	0,362	0,399	0,446	0,386	0,404	0,366	0,349	0,402
6	0,447	0,392	0,428	0,399	0,375	0,392	0,411	0,414
3 months								
1	**0,676**	0,697	0,772	**0,732**	0,710	**0,677**	0,626	**0,615**
3	0,697	0,711	0,816	0,824	0,789	0,803	0,787	0,766
6	0,720	0,751	0,814	0,881	0,790	0,848	0,765	0,822
6 months								
1	0,722	0,767	0,914	**0,901**	0,735	0,767	0,685	0,669
3	0,701	**0,690**	0,903	0,940	**0,716**	0,731	0,698	**0,657**
6	0,743	0,690	0,927	0,972	0,809	0,781	0,916	0,889

Note: The lowest RMSE levels for each forecasting horizon and number of lags are marked with grey.
The lowest RMSE among two experiments for each model is marked with bold.

Overall, the results suggest, that the sensitivity to the use of data vintages varies among the models and the considered forecasting horizons. In these two experiments, if the choice of a model with the lowest errors has to be made, neural network has the lowest or comparable out-of-sample errors in comparison to the other models. For the forecasts for one month in advance the use of vintages data leads to the lower error. For three and six months, on the contrary, the use of data vintages only increases the error.

Overall, in 19 cases out of 36 (for three forecasting horizons, three lags for each and four models) the use of data vintages leads to the lower RMSE. However, when the model with the optimal number of lags is considered in 8 out of 12 cases (four models and three forecasting horizons) the forecasts on the revised data provide lower RMSEs. Anyway, usually forecaster is not faced with the choice what data to use. In real time only preliminary releases and data published with delay are available. The comparison of these two estimation experiments shows, that in real-time forecasting RMSE levels can differ significantly from the estimates, obtained on the revised data. Depending on the model, number of lags used and forecasting horizon this impact may vary. However, the results suggest, that model performance on all available data is not representative in real-time estimations and should be considered with caution.

5.4. The role of seasonal adjustment

Next, we consider two experiments on data, which incorporate all revisions (no vintages). In the one experiment all variables with seasonal fluctuations are seasonally adjusted (experiment 4 in Table 1), while in the second one the variables are not altered and seasonal dummies are added in the dataset (experiment 3 in Table 1).[17] This comparison is aimed to show, how forecasting results depend on the use of unobservable data transformations and whether it can alter the forecasting error and the choice of the model. Table 5 summarizes the results.

In almost all cases the use of seasonally adjusted timeseries leads to the lower forecasting errors. The difference in the value errors varies from 12,1% for one month in advance to 8,5% and 9,3% for three and six months, correspondingly. The results suggest, that the use of seasonal adjustment in the CPI forecasting can lead to a significant underestimation of forecast errors.

Table 5. RMSEs estimated on seasonally adjusted and not seasonally adjusted data

	Elastic Net		Random Forest		Gradient Boosting		Neural network	
lags	sa	nsa	sa	nsa	sa	nsa	sa	nsa
1 month								
1	0,348	0,369	**0,326**	0,373	**0,323**	0,334	**0,304**	0,348
3	**0,295**	0,399	0,346	0,386	0,344	0,366	0,347	0,402
6	0,377	0,392	0,353	0,399	0,344	0,392	0,391	0,414
3 months								
1	**0,622**	0,697	**0,686**	0,732	0,700	**0,677**	0,625	**0,615**
3	0,675	0,711	0,715	0,824	0,682	0,803	0,652	0,766
6	0,664	0,751	0,738	0,881	0,732	0,848	0,716	0,822
6 months								
1	0,666	0,767	**0,843**	0,901	**0,671**	0,767	0,837	0,669
3	**0,599**	0,690	0,846	0,940	0,791	0,731	0,804	**0,657**
6	0,658	0,690	0,880	0,972	0,827	0,781	0,921	0,889

Note: The lowest RMSE levels for each forecasting horizon and number of lags are marked with grey.
The lowest RMSE among two experiments for each model is marked with bold.

5.5. Joint effect of three discrepancies

Here we compare two 'corner' experiments: the benchmark one, when the estimates are based on vintages of not seasonally adjusted data and release timing is taken into account (experiment 1), and the last one, estimated on series with final revisions and seasonally adjusted data, treating all variables as available. Table 6 summarizes the results.

Table 6. Comparison of RMSEs of models the first and fourth experiment

	Elastic Net		Random Forest		Gradient Boosting		Neural network	
lags	exp. 4	exp.1	exp. 4	exp.1	exp. 4	exp.1	exp. 4	exp.1
1 month								
1	0,348	0,480	**0,326**	0,368	**0,323**	0,356	**0,304**	0,409
3	**0,295**	0,403	0,346	0,373	0,344	0,382	0,347	0,431
6	0,377	0,380	0,353	0,389	0,344	0,396	0,391	0,400
3 months								

[17] The seasonal adjustment is conducted in Demetra program with tramoseats method and rsa3 specification.

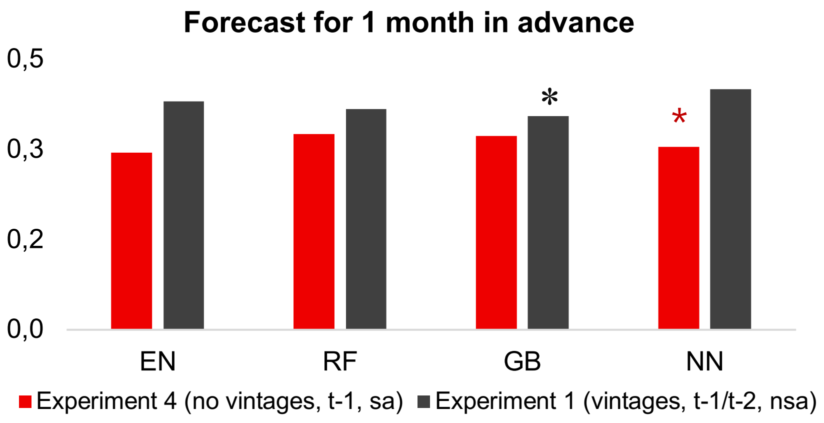
1	**0,622**	0,756	**0,686**	0,826	0,700	0,820	**0,625**	0,705
3	0,675	0,760	0,715	0,883	**0,682**	0,759	0,652	0,728
6	0,664	0,789	0,738	0,895	0,732	0,737	0,716	0,860
6 months								
1	0,666	0,758	**0,843**	0,963	**0,671**	0,789	0,837	**0,687**
3	**0,599**	0,643	0,846	0,932	0,791	0,716	0,804	0,752
6	0,658	0,758	0,880	0,986	0,827	0,842	0,921	0,876

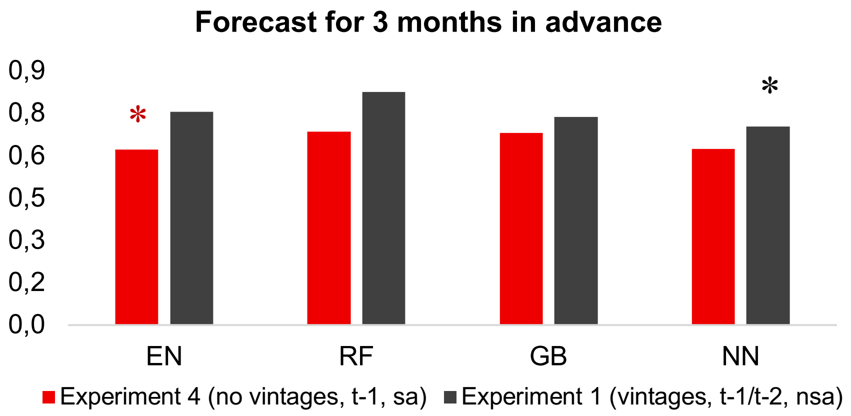
Note: Exp. 1 does not include vintages or take into account of release timing, with seasonal adjustment;
Exp.4 is based on vintages, in it we take into account release timing and use not seasonally adjusted data.
The lowest RMSE levels for each forecasting horizon and number of lags are marked with grey. The lowest RMSE among two experiments for each model is marked with bold.

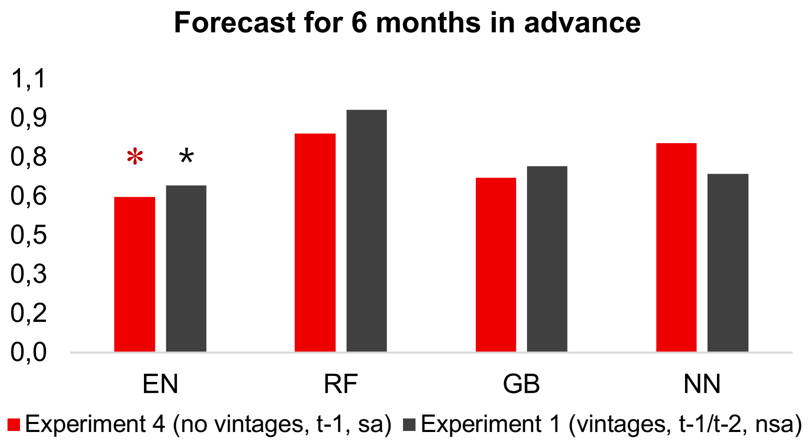
We see, that for all four models experiment 4, which incorporates all inaccuracies, has a lower forecasting error than the 1 experiment (32 out of 36 cases), which reconstructs the real time forecasting procedure, how it would have been at each point in the past.

If we consider the forecasts with the optimal for each horizon number of lags, the evidence is even more conclusive (Figure 2). The RMSE gap of the models within two experiments varies depending on the model and the forecasting horizon. On average the difference between the experiments is 16,8% for one month in advance, 13,4% for the three months and 7,6% for forecasts for six months. Moreover, the choice of the optimal model for two out of three horizons changes depending on the experiment, so these results may also affect the comparison of models, based on their performance. Therefore, neglecting these three inaccuracies in the actual real-time forecasting can lead to a significant underestimation of the forecasting error and the incorrect choice of the forecasting model.

Figure 2. Comparison of the RMSEs of the models with optimal number of lags for first and fourth experiment for different forecasting horizons







Conclusion

We construct four forecasting experiments in order to define, how different inaccuracies in pseudo real-time forecasting affect the performance of several machine learning models. We analyze these effects in the case of forecasting Russian CPI. As a benchmark for the comparison, we use the case, where the forecasts are based on the data with three main characteristics. First, we use data vintages, which is the first case of their use in the forecasting excises on Russian data. Second, in our real-time experiment we take into account, when each variable is released, fix at what part of the month we make our forecasts and include in the estimation only available data with the corresponding release delay. Finally, we do not apply any seasonal adjustment procedures, and include seasonal dummies in the dataset along with timeseries.

Main results come from the comparison of the benchmark experiment with others. First, we show, that the neglect of the release timing of series lead to the significant underestimation of the forecasting error. Second experiment provides inconclusive evidence concerning an impact of the use of vintages. For most of the models the estimation on the revised data leads to the lower errors. We show, that for some models the use of ordinary data can lead to an artificially low forecasting error. In reality, all we have at each point in time are preliminary data before any revisions occur. With the help of the last experiment we show, that the use of seasonally adjusted data lowers artificially the forecasting error. This means, that when a particular model is considered, lower forecasting errors can be misleading and partly be the result of the use of seasonally adjusted data. Finally, we compare the results of the benchmark experiment with the results of the estimations, conducted on revised time series (not vintages) of seasonally adjusted data and with no account for the difference in release dates. Overall effect of these three discrepancies is 16,8% for the best number of lags for each model for one month in advance (13,4% and 7,6% for three and six months, correspondingly). Therefore, these aspects could lead to a significant underestimation of forecasting error in actual real-time forecasting and should be taken into account in the macroeconomic forecasting.

Within the benchmark experiment we also compare the performance of ML models and show, that these models have a great potential in forecasting macroeconomic timeseries. For forecasting for one month in advance gradient boosting and neural network have the comparable RMSE, for three months in advance neural network has the best performance, for six months elastic net provides lower error. In all three cases AR forecasts are outperformed by these ML models.

References

1. Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector auto regressions. Journal of applied Econometrics, 25(1), 71-92.
2. Baybuza I. (2018). Inflation Forecasting Using Machine Learning Methods. Russian Journal of Money and Finance 77(4) pp. 4259. doi: 10.31477/rjmf.201804.42.
3. Blanchard, O. J., & Quah, D. (1989). The dynamic effects of aggregate demand and supply disturbances: Reply. The American Economic Review, 79, 655-673.
4. Buchen T. & Wohlrabe K. (2011). Forecasting with many predictors: Is boosting a viable alternative?. Economics Letters 113(1) 16-18.
5. Catik A. N. & Karauka M. (2012). A comparative analysis of alternative univariate time series models in forecasting Turkish inflation. Journal of Business Economics and Management 13(2) 275-293.
6. Chakraborty C. & Joseph A. (2017). Machine learning at central banks. Staff Working Paper No. 674 Bank of England
7. Chen X. Racine J. & Swanson N. R. (2001). Semiparametric ARX neural-network models with an application to forecasting inflation. IEEE Transactions on neural networks 12(4) 674-683.
8. Choudhary M. A. & Haider A. (2012). Neural network models for inflation forecasting: an appraisal. Applied Economics 44(20) 26312635.
9. Clausen, J. R., & Meier, C. P. (2005). Did the Bundesbank Follow a Taylor Rule? An Analysis Based on Real-Time Data. Swiss Journal of Economics and Statistics (SJES), 141(II), 213-246.
10. Clements, M. P., & Galvão, A. B. (2009). Forecasting US output growth using leading indicators: An appraisal using MIDAS models. Journal of Applied Econometrics, 24(7), 1187-1206.
11. Clements, M. P., & Galvão, A. B. (2011). Improving Real-time Estimates of Output Gaps and Inflation Trends. VAR models. Discussion paper, Department of Economics, University University of Warwick.
12. Clements, M. P., & Galvão, A. B. (2013). Forecasting with vector autoregressive models of data vintages: US output growth and inflation. International Journal of Forecasting, 29(4), 698-714.
13. Cole, R. (1969). Data errors and forecasting accuracy. In Economic forecasts and expectations: analysis of forecasting behavior and performance (pp. 47-82). NBER.

14. Cook T. R. & Smalter Hall A. (2017). Macroeconomic Indicator Forecasting with Deep Neural Networks (No. RWP 17-11).

15. Croushore, D. (2011). Frontiers of real-time data analysis. Journal of economic literature, 49(1), 72-100.

16. Croushore, D. and T. Stark (2001) A real-time data set for macroeconomists. Journal of Econometrics, 105 pp. 111-130.

17. Croushore, D., & Stark, T. (2003). A real-time data set for macroeconomists: Does the data vintage matter?. Review of Economics and Statistics, 85(3), 605-617.

18. Denton, Frank T., and John Kuiper. The Effect of Measurement Errors on Parameter Estimates and Forecasts: A Case Study Based on the Canadian Preliminary National Accounts, Review of Economics and Statistics 47 (May 1965), pp. 198-206.

19. Dpke J., Fritsche U. & Pierdzioch C. (2017). Predicting recessions with boosted regression trees. International Journal of Forecasting 33(4) 745-759.

20. Egginton, D. M., Pick, A., & Vahey, S. P. (2002). 'Keep it real!': A real-time UK macro data set. Economics Letters, 77(1), 15-20.

21. Faust, J., Rogers, J. H., & Wright, J. H. (2003). Exchange rate forecasting: the errors we've really made. Journal of International Economics, 60(1), 35-59.

22. Fernandez, A. Z., Branch, H., Koenig, E. F., & Nikolsko-Rzhevskyy, A. (2011). A real-time historical database for the OECD. Globalization and Monetary Policy Institute Working Paper, 96.

23. Filardo, A. J. (1999). How reliable are recession prediction models?. Economic Review-Federal Reserve Bank of Kansas City, 84, 35-56.

24. Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. Review of Economics and statistics, 82(4), 540-554.

25. Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. Journal of the American Statistical Association, 100(471), 830-840.

26. Friedman J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics 1189-1232.

27. Friedman J. Hastie T. & Tibshirani R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33(1) 1.

28. Gerberding, C., Seitz, F., & Worms, A. (2005). How the Bundesbank really conducted monetary policy. The North American journal of economics and finance, 16(3), 277-292.

29. Giannone, D., Henry, J., Lalik, M., & Modugno, M. (2012). An area-wide real-time database for the euro area. Review of Economics and Statistics, 94(4), 1000-1013.

30. Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. Journal of political economy, 86(6), 971-987.

31. Hanif M. N., Mughal K. S. & Iqbal J. (2018). A Thick ANN Model for Forecasting Inflation (No. 99). State Bank of Pakistan Research Department.

32. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

33. Howrey, E. P. (1996). Forecasting GNP with noisy data: a case study. Journal of Economic and Social Measurement, 22(3), 181-200.

34. Kapetanios G. & Papailias F. (2018). Big data & macroeconomic nowcasting: Methodological review. ESCoE Discussion Paper (12).

35. Khabibullin R. & Seleznev S. (2020) Stochastic Gradient Variational Bayes and Normalizing Flows for Estimating Macroeconomic Models. Bank of Russia working paper series No. 61.

36. Kishor, N. K., & Koenig, E. F. (2012). VAR estimation and forecasting when data are subject to revision. Journal of Business & Economic Statistics, 30(2), 181-190.

37. Kock A. B. & Tersvirta T. (2013). Forecasting the Finnish Consumer Price Inflation Using Artificial Neural Network Models and Three Automated Model Selection Techniques. Finnish Economic Papers 26(1) 13-24.

38. Koenig, E. (2003) Is the makeup a useful real-time predictor of inflation? Economic Letters, 80 pp. 261-267.

39. Koenig, E. F., Dolmas, S., & Piger, J. (2003). The use and abuse of real-time data in economic forecasting. Review of Economics and Statistics, 85(3), 618-628.

40. Kydland, F. E., & Prescott, E. C. (1990). Business cycles: Real facts and a monetary myth. Real business cycles: a reader, 383.

41. lvarez-Daz M. & Gupta R. (2015). Forecasting the US CPI: Does Nonlinearity Matter?. Department of Economics University of Pretoria Working Paper No 12.

42. McAdam P. & McNelis P. (2005). Forecasting inflation with thick models and neural networks. Economic Modelling 22(5) 848-867.

43. McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics, 34(4), 574-589.

44. Molodtsova, T., & Ince, O. (2008). Real-Time Exchange Rate Predictability with Taylor Rule Fundamentals. Manuscript, Emory University.

45. Molodtsova, T., A. Nikolsko-Rzhevskyy, and D. Papell (2008) Taylor rules and real-time data: A tale of two countries and one exchange rate. Journal of Monetary Economics, 55 pp. S63-S79.

46. Moshiri S. & Cameron N. (2000). Neural network versus econometric models in forecasting inflation. Journal of forecasting 19(3) 201-217.

47. Nikolsko-Rzhevskyy, A. (2011). Monetary policy estimation in real time: Forward-looking Taylor rules without forward-looking data. Journal of Money, Credit and Banking, 43(5), 871-897.

48. Orphanides, A. (2001) Monetary policy rules based on real-time data. American Economic Review, 91(4) pp. 964-985, September.

49. Orphanides, A., & Van Norden, S. (2005). The reliability of inflation forecasts based on output gap estimates in real time. Journal of Money, Credit and Banking, 583-601.

50. Ponomarenko A., Gornostaev D., Seleznev S., Sterhova A. (2021) A Real-Time Historical Database of Macroeconomic Indicators for Russia. Bank of Russia working paper series.

51. Rossi, B., & Sekhposyan, T. (2010). Have economic models' forecasting performance for US output growth and inflation changed over time, and when?. International Journal of Forecasting, 26(4), 808-835.

52. Sauer, S., & Sturm, J. E. (2007). Using Taylor rules to understand European Central Bank monetary policy. German Economic Review, 8(3), 375-398.

53. Sermpinis G. Stasinakis C. Theofilatos K. & Karathanasopoulos A. (2014). Inflation and unemployment forecasting with genetic support vector regression. Journal of Forecasting 33(6) 471-487.

54. Shahriari B. Swersky K. Wang Z. Adams R. P. & De Freitas N. (2015). Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE 104(1) 148-175.

55. Stark, T., & Croushore, D. (2002). Forecasting with a real-time data set for macroeconomists. Journal of Macroeconomics, 24(4), 507-531.

56. Stock, J. H., & W Watson, M. (2003). Forecasting output and inflation: The role of asset prices. Journal of Economic Literature, 41(3), 788-829.

57. Stock, J. H., & Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. Journal of the American statistical association, 97(460), 1167-1179.

58. Stock, J. H., & Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics, 20(2), 147-162.

59. Styrin K. (2019). Forecasting Inflation in Russia by Dynamic Model Averaging. Russian Journal of Money and Finance 78(1) 3-18.

60. Szafranek, K. (2019). Bagged neural networks for forecasting Polish (low) inflation. International Journal of Forecasting, 35(3), 1042-1059.

61. Tiffin A. (2016). Seeing in the dark: A machine-learning approach to nowcasting in Lebanon. WP/16/56 IMF Working paper.

62. Trivellato, U., & Rettore, E. (1986). Preliminary data errors and their impact on the forecast error of simultaneous-equations models. Journal of Business & Economic Statistics, 4(4), 445-453.

63. Zhang L. & Li J. (2012). Inflation Forecasting Using Support Vector Regression. In Information Science and Engineering (ISISE) (pp. 136140). IEEE.

Appendix

Tables

Table A1. Variables and data sources[18]

Name of series	Type	Source	Period
Consumer price index	regular	FSSS	t-1
Industrial production	vintages	FSSS	t-2
Unemployment rate	vintages	FSSS	t-2
Real wage	vintages	FSSS	t-2
Real agricultural production	vintages	FSSS	t-2
Eggs production	vintages	FSSS	t-2
Meat production	vintages	FSSS	t-2
Milk production	vintages	FSSS	t-2
Freight	vintages	FSSS	t-2
Railway freight	vintages	FSSS	t-2
Commercial freight	vintages	FSSS	t-2
Real retail output	vintages	FSSS	t-2
Food retail	vintages	FSSS	t-2
Nonfood retail	vintages	FSSS	t-2
Services	vintages	FSSS	t-2
Public catering	vintages	FSSS	t-2
Construction	vintages	FSSS	t-2
Export of goods	vintages	RSSS	t-2
Import of goods	vintages	RSSS	t-2
Nominal exchange rate	regular	CBR	t-1
Real effective exchange rate	regular	BIS	t-1
Interbank interest rate	regular	CBR	t-1
Deposit interest rate	regular	CBR	t-2
International reserves	regular	CBR	t-1
Monetary aggregate M2 (real)	regular	CBR	t-1
Total government deficit	regular	Roskazna	t-2
Crude oil (Brent) price	regular	EIA	t-1

[18] FSSS – Federal State Statistical Service, CBR – Central Bank of Russia, BIS – Bank for international settlements, EIA – Energy Information Administration.

Table A2. Values of the parameters used in the estimation of random forest

Hyperparameters	Values
criterion	Mse, mae
maximum number of features	Auto,log2, sqrt
number of estimators	(5, 300)
maximum depth of a tree	None, (1, 14)
min. number of samples required to split a node	(2, 10)

Note: default values are criterion – mse, number of estimators = 100, maximum depth – None, maximum number of features – auto, maximum depth of a tree – None, min. samples split = 2.

Table A3. Values of the parameters used in the estimation of gradient boosting

Hyperparameters	Values
loss function	least squares least absolute deviation Huber function
criterion	Friedman mse mse mae
learning rate	(0.001 0.2)
number of estimators	(10 400)
maximum depth	(2 10)
subsample	(0.2 1.0)

Note: default values are loss function - least squares, criterion - Friedman mse, learning rate = 0.1, number of estimators = 100, maximum depth = 3, subsample = 1.0.

Table A4. Optimal parameters of random forest chosen on cross-validation

hyperparameters	one month	three months	six months
first experiment			
criterion	mae	mae	mae
maximum number of features	auto	auto	auto
number of estimators	21	30	33
maximum depth of a tree	8	10	14
min. samples split	8	2	2
max. leaf nodes	47	50	50
number of lags	1	1	3
second experiment			
criterion	mse	mse	mse
maximum number of features	auto	auto	auto
number of estimators	10	39	34
maximum depth of a tree	9	14	14
min. samples split	4	2	4
max. leaf nodes	30	None	None
number of lags	1	1	3
third experiment			
criterion	mae	mae	mae
maximum number of features	auto	auto	auto
number of estimators	39	244	204
maximum depth of a tree	10	13	14
min. samples split	2	5	6
max. leaf nodes	34	45	25
number of lags	1	1	1
fourth experiment			
criterion	mae	mse	mse
maximum number of features	auto	auto	auto
number of estimators	300	194	92
maximum depth of a tree	12	7	14
min. samples split	7	3	2
max. leaf nodes	47	41	43
number of lags	1	1	1

Note: default values are criterion – mse, number of estimators = 100, maximum depth – None, maximum number of features – auto, maximum depth of a tree – None, min. samples split = 2.

Table A5. Optimal parameters of gradient boosting chosen on cross-validation

hyperparameters	one month	three months	six months
first experiment			
loss function	ls	ls	ls
criterion	mse	friedman_mse	mse
learning rate	0,041	0,223	0,145
number of estimators	323	348	314
maximum depth	2	2	2
subsample	0,386	0,768	0,995
number of lags	1	6	3
second experiment			
loss function	ls	ls	huber
criterion	mse	mse	mae
learning rate	0,068	0,181	0,103
number of estimators	399	163	344
maximum depth	2	2	2
subsample	0,552	0,889	0,792
number of lags	6	1	3
third experiment			
loss function	ls	ls	huber
criterion	mse	friedman_mse	mse
learning rate	0,086	0,171	0,205
number of estimators	128	334	211
maximum depth	3	2	2
subsample	0,620	0,619	0,982
number of lags	1	1	3
fourth experiment			
loss function	huber	ls	ls
criterion	friedman_mse	friedman_mse	mse
learning rate	0,050	0,4	0,4
number of estimators	238	10	28
maximum depth	9	2	2
subsample	0,733	1	1
number of lags	1	3	1

Note: default values are loss function - least squares, criterion - Friedman mse, learning rate = 0.1, number of estimators = 100, maximum depth = 3, subsample = 1.0.

Table A6. Optimal parameters of elastic net chosen on cross-validation

hyperparameters	one month	three months	six months
first experiment			
alpha	0,005	0,108	0,044
l1_ratio	0,547	0,907	0,999
fit_intercept	TRUE	TRUE	TRUE
normalize	TRUE	FALSE	FALSE
max_iter	268	237	1821
selection	random	cyclic	cyclic
number of lags	6	1	3
second experiment			
alpha	0,003	0,061	0,067
l1_ratio	0,138	0,987	1,000
fit_intercept	TRUE	TRUE	TRUE
normalize	TRUE	FALSE	FALSE
max_iter	1280	1669	463
selection	random	cyclic	cyclic
number of lags	1	1	3
third experiment			
alpha	0,017	0,201	0,007
l1_ratio	0,653	0,965	1,000
fit_intercept	TRUE	TRUE	TRUE
normalize	FALSE	FALSE	TRUE
max_iter	568	1447	1261
selection	random	random	random
number of lags	1	1	6
fourth experiment			
alpha	0,022	0,057	0,069
l1_ratio	0,836	0,923	0,922
fit_intercept	TRUE	TRUE	TRUE
normalize	FALSE	FALSE	FALSE
max_iter	488	1009	1239
selection	cyclic	random	random
number of lags	3	1	3

Note: default values are alpha = 1 , l1_ration = 0,5, fit_itercept = True, normalize = False, max_iter = 1000, selection = ‘cyclic’.

Table A7. RMSE levels on cross-validation and test in the first experiment

Lags	Gradient Boosting		Random Forest		Elastic Net		Neural network	
	CV	test	CV	test	CV	test	CV	test
1 month								
1	0,382	0,356	0,371	0,368	0,377	0,480	0,421	0,409
3	0,400	0,382	0,378	0,373	0,387	0,403	0,439	0,431
6	0,394	0,396	0,405	0,389	0,387	0,380	0,447	0,400
3 months								
1	0,598	0,820	0,644	0,826	0,631	0,756	0,694	0,705
3	0,648	0,759	0,715	0,883	0,673	0,760	0,739	0,728
6	0,646	0,737	0,717	0,895	0,598	0,789	0,891	0,860
6 months								
1	0,837	0,789	0,922	0,963	0,649	0,758	0,769	0,687
3	0,853	0,716	0,939	0,932	0,632	0,643	0,756	0,752
6	0,805	0,842	0,949	0,986	0,586	0,758	0,724	0,876

Table A8. RMSE levels on cross-validation and test in the second experiment

Lags	Gradient Boosting		Random Forest		Elastic Net		Neural network	
	CV	test	CV	test	CV	test	CV	test
1 month								
1	0,360	0,378	0,379	0,393	0,363	0,359	0,363	0,338
3	0,393	0,404	0,387	0,446	0,397	0,362	0,425	0,349
6	0,385	0,375	0,401	0,428	0,370	0,447	0,464	0,411
3 months								
1	0,566	0,710	0,641	0,772	0,587	0,676	0,672	0,626
3	0,650	0,789	0,689	0,816	0,632	0,697	0,847	0,787
6	0,644	0,790	0,700	0,814	0,561	0,720	0,757	0,765
6 months								
1	0,830	0,735	0,961	0,914	0,646	0,722	0,689	0,685
3	0,890	0,716	1,031	0,903	0,684	0,701	0,729	0,698
6	0,839	0,809	1,003	0,927	0,604	0,743	0,712	0,916

Table A9. RMSE levels on cross-validation and test in the third experiment

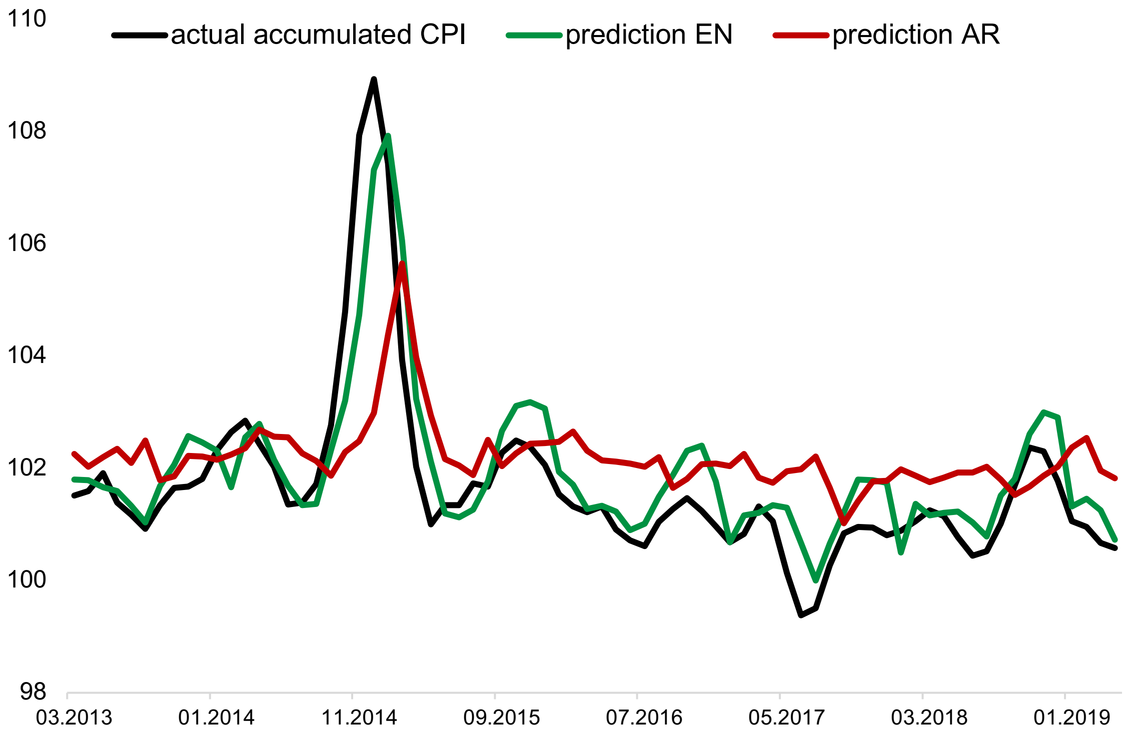
Lags	Gradient Boosting		Random Forest		Elastic Net		Neural network	
	CV	test	CV	test	CV	test	CV	test
1 month								
1	0,348	0,334	0,378	0,373	0,395	0,369	0,378	0,348
3	0,387	0,366	0,393	0,386	0,387	0,399	0,407	0,402
6	0,390	0,392	0,396	0,399	0,353	0,392	0,471	0,414
3 months								
1	0,576	0,677	0,666	0,732	0,592	0,697	0,712	0,615
3	0,655	0,803	0,697	0,824	0,632	0,711	0,729	0,766
6	0,629	0,848	0,720	0,881	0,610	0,751	0,731	0,822
6 months								
1	0,829	0,767	0,929	0,901	0,610	0,767	0,731	0,669
3	0,877	0,731	0,972	0,940	0,639	0,690	0,743	0,657
6	0,842	0,781	0,963	0,972	0,590	0,690	0,723	0,889

Table A10. RMSE levels on cross-validation and test in the fourth experiment

Lags	Gradient Boosting		Random Forest		Elastic Net		Neural network	
	CV	test	CV	test	CV	test	CV	test
1 month								
1	0,293	0,323	0,290	0,326	0,314	0,348	0,313	0,304
3	0,304	0,344	0,302	0,346	0,374	0,295	0,310	0,347
6	0,300	0,344	0,307	0,353	0,313	0,377	0,349	0,391
3 months								
1	0,460	0,700	0,458	0,686	0,444	0,622	0,541	0,625
3	0,480	0,682	0,488	0,715	0,494	0,675	0,561	0,652
6	0,478	0,732	0,494	0,738	0,505	0,664	0,581	0,716
6 months								
1	0,596	0,671	0,627	0,843	0,502	0,666	0,513	0,837
3	0,656	0,791	0,663	0,846	0,512	0,599	0,640	0,804
6	0,636	0,827	0,692	0,880	0,553	0,658	0,788	0,921

Figures

Figure A1. Comparison of AR and elastic net forecasts for three months in advance with one lag



Bank of Russia

Machine learning real-time CPI forecasting

Mariam Mamedli
NRU Higher School of Economics
Denis Shibitov
Bank of Russia

IFC workshop ”Data science in central banking”
Machine learning applications
October, 2021

This views expressed herein are those of the author and do not necessarily represent the position of the Bank of Russia

Aim:

To replicate the construction of CPI forecasts in real time and to define how different aspects of forecasting affect the model accuracy:

- The use of vintages, data availability and the use of SA procedure.

Tasks:

- to build ML models with the optimal architecture (hyperparameters, lags, CV).
- to define how the model performance changes depending on:
  - the type of ML model applied;
  - data availability;
  - the use of vintage data;
  - the use of seasonal adjustment.

Relevant research

- Importance of vintage data (Koenig et al., 2003; Clements, Galvao, 2009)
- The role of data revisions (Stark, Croushore, 2002)

Application of ML models in CPI forecasting:

- Potential of ML models in economic forecasting (Chakraborty, Joseph, 2017)
- The use of neural networks:
  - one country cases (Moshiri, Cameron, 2000; Chen et al., 2001; Szafranek, 2017, Hanif et al., 2018);
  - panel data (Choudhary, Haider, 2012, McAdam, McNelis, 2005);
  - results not in favor of neural networks (Kock, Terasvirts, 2013; Catik, Karacuka, 2012).
- Other models:
  - random forest (Chakraborty, Joseph, 2017; Butavyan, 2019, Baybuza, 2018);
  - gradient boosting (Baybuza, 2018);
  - SVR (Zhang, Li, 2012; Sermipinis et al., 2014; Plakandaras et al., 2017).
- Only one paper on the application to the Russian data (Baybuza, 2018).

Value added

Data:

- Unique for Russia vintage data.

Methodology:

- Application of ML models (one of the first for Russia) + Bayesian optimization;
- The choice of the optimal architecture via CV and different lags.

Concept:

Model comparison in a series of experiments so define the role of:

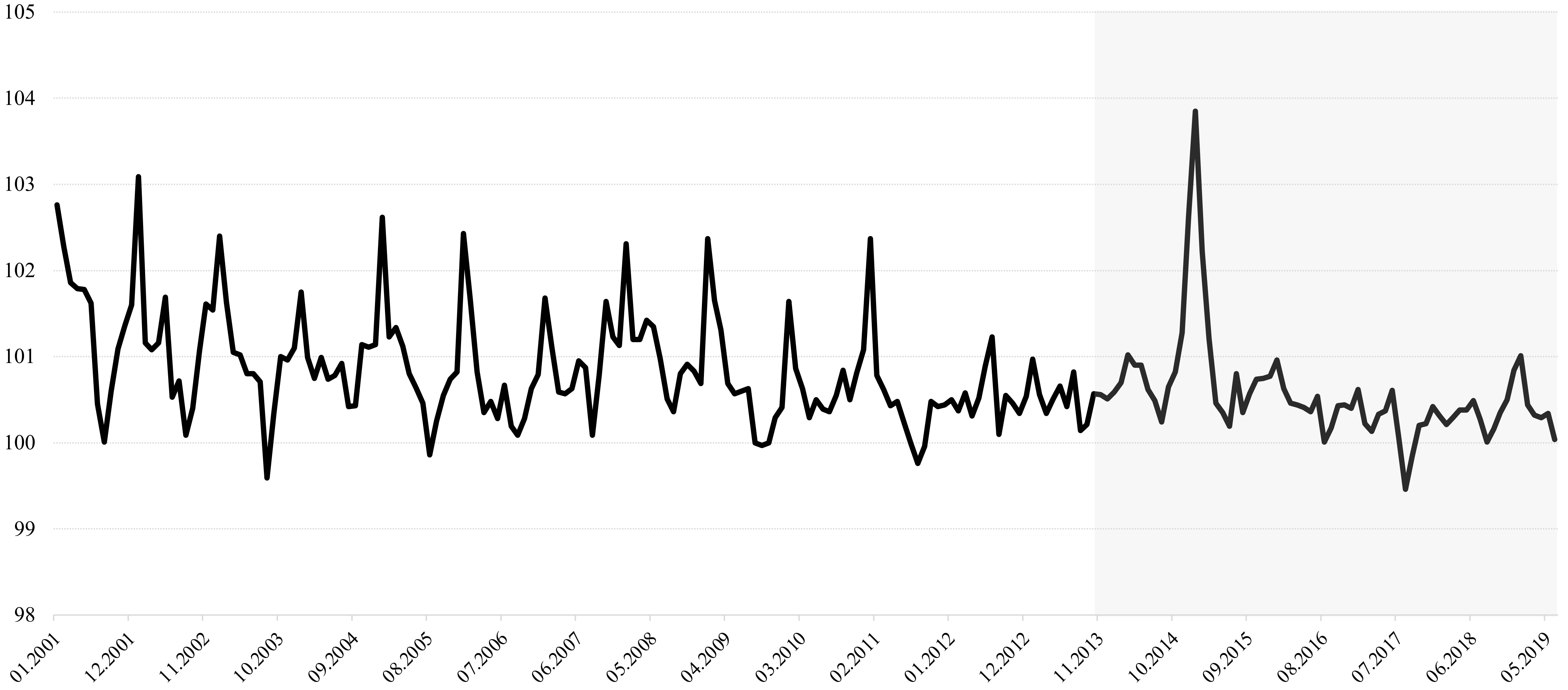
- Data availability;
- Data vintages;
- Seasonal adjustment.

Experiments:

Experiment	Starting lags	Vintages	Seasonality
1	$t-1/t-2$	vintages	NSA
2	$t-1$	vintages	NSA
3	$t-1$	regular	NSA
4	$t-1$	regular	SA

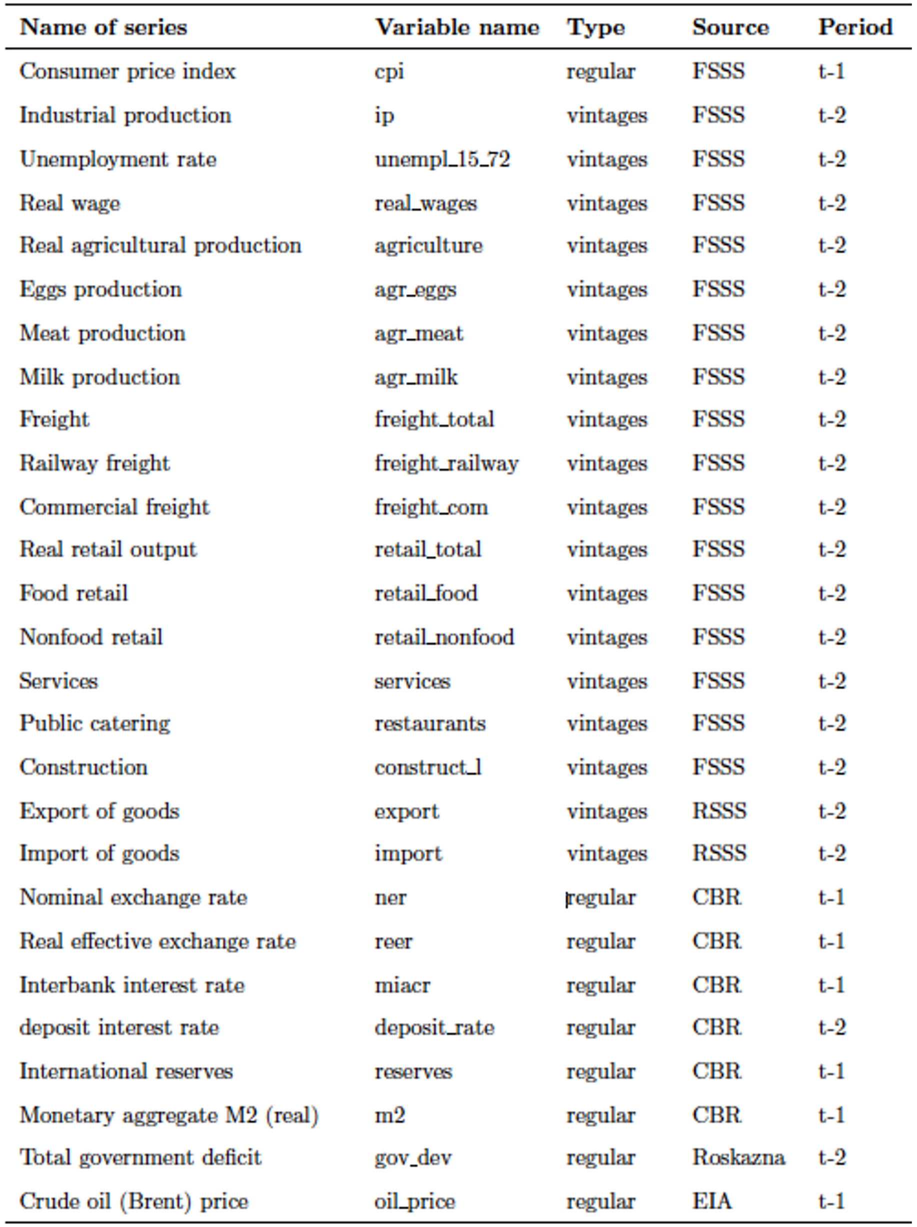
- 1st experiment is considered as benchmark;
- 2d and 3d experiments aimed at the replication of the forecast in real time;
- 4th experiment is devoted to the role of seasonal adjustment.
- The use of NSA data:
  - The forecast of an observed variable;
  - Is not revised (replicability);
  - Commonly used and accepted indicator.

CPI (NSA)


105
104
103
102
101
100
99
98
01.2001
12.2001
11.2002
10.2003
09.2004
08.2005
07.2006
06.2007
05.2008
04.2009
03.2010
02.2011
01.2012
12.2012
11.2013
10.2014
09.2015
08.2016
07.2017
06.2018
05.2019

Data:

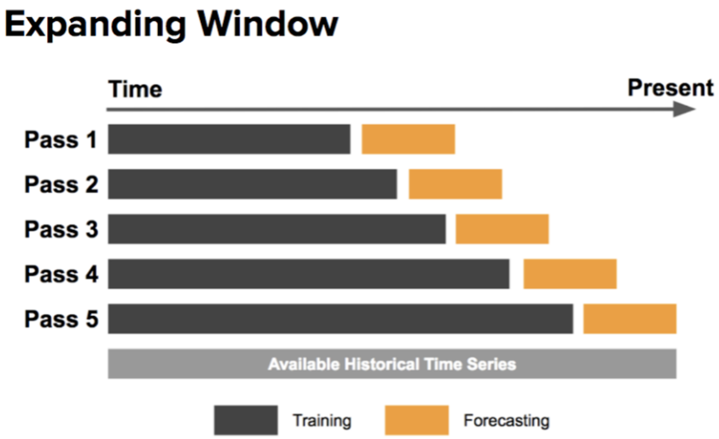
- Vintage data including revisions + other macro and financial variables;
- Data from January 2001 to June 2019;
- 27 variables + ‘dummy’ for each month;
- NSA, minimal series transformations;
- Forecast for 1, 3 and 6 months with 1, 3 and 6 lags each.



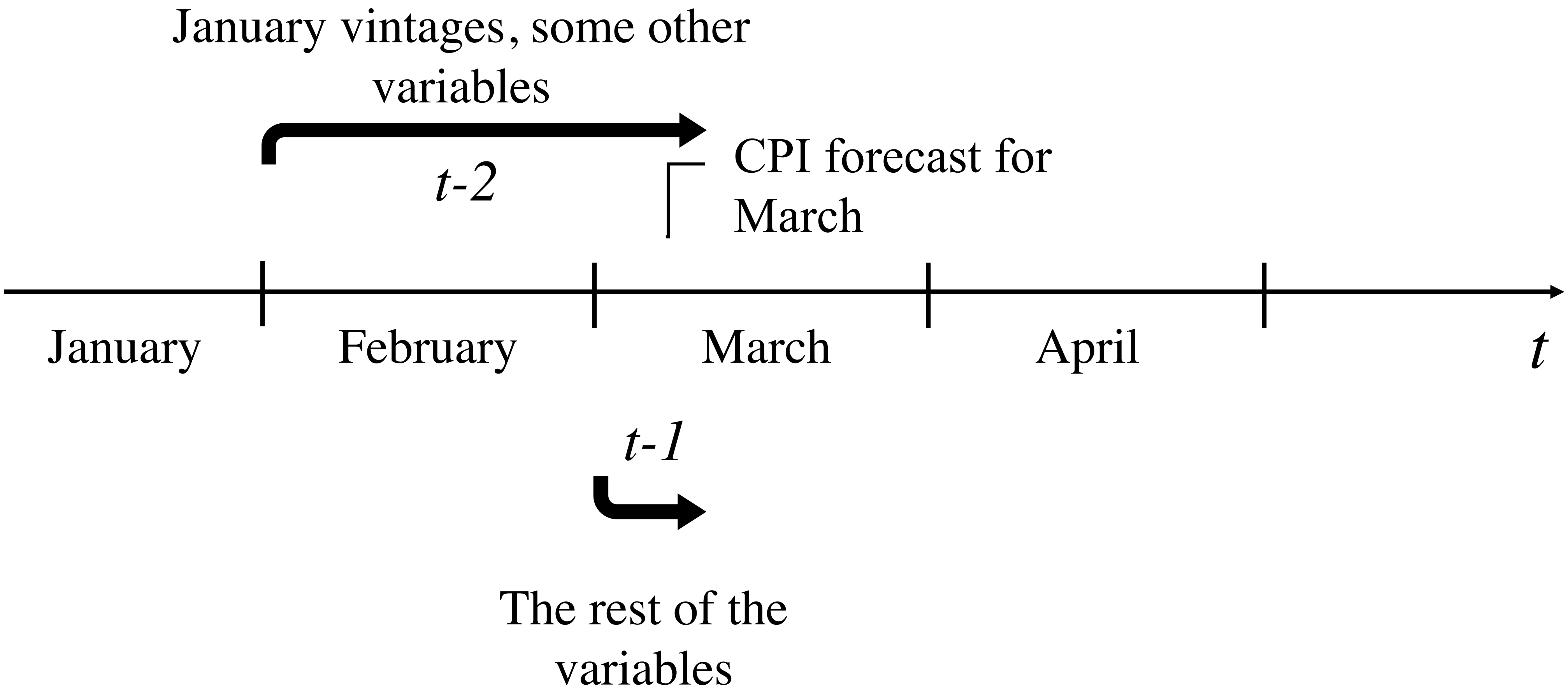
Name of series	Variable name	Type	Source	Period
Consumer price index	cpi	regular	FSSS	t-1
Industrial production	ip	vintages	FSSS	t-2
Unemployment rate	unempl_15_72	vintages	FSSS	t-2
Real wage	real_wages	vintages	FSSS	t-2
Real agricultural production	agriculture	vintages	FSSS	t-2
Eggs production	agr_eggs	vintages	FSSS	t-2
Meat production	agr_meat	vintages	FSSS	t-2
Milk production	agr_milk	vintages	FSSS	t-2
Freight	freight_total	vintages	FSSS	t-2
Railway freight	freight_railway	vintages	FSSS	t-2
Commercial freight	freight_com	vintages	FSSS	t-2
Real retail output	retail_total	vintages	FSSS	t-2
Food retail	retail_food	vintages	FSSS	t-2
Nonfood retail	retail_nonfood	vintages	FSSS	t-2
Services	services	vintages	FSSS	t-2
Public catering	restaurants	vintages	FSSS	t-2
Construction	construct_l	vintages	FSSS	t-2
Export of goods	export	vintages	RSSS	t-2
Import of goods	import	vintages	RSSS	t-2
Nominal exchange rate	ner	regular	CBR	t-1
Real effective exchange rate	reer	regular	CBR	t-1
Interbank interest rate	miacr	regular	CBR	t-1
deposit interest rate	deposit_rate	regular	CBR	t-2
International reserves	reserves	regular	CBR	t-1
Monetary aggregate M2 (real)	m2	regular	CBR	t-1
Total government deficit	gov_dev	regular	Roskazna	t-2
Crude oil (Brent) price	oil_price	regular	EIA	t-1

Cross-validation:

- Expending window



Data:



January vintages, some other variables

t-2

CPI forecast for March

January

February

March

April

t

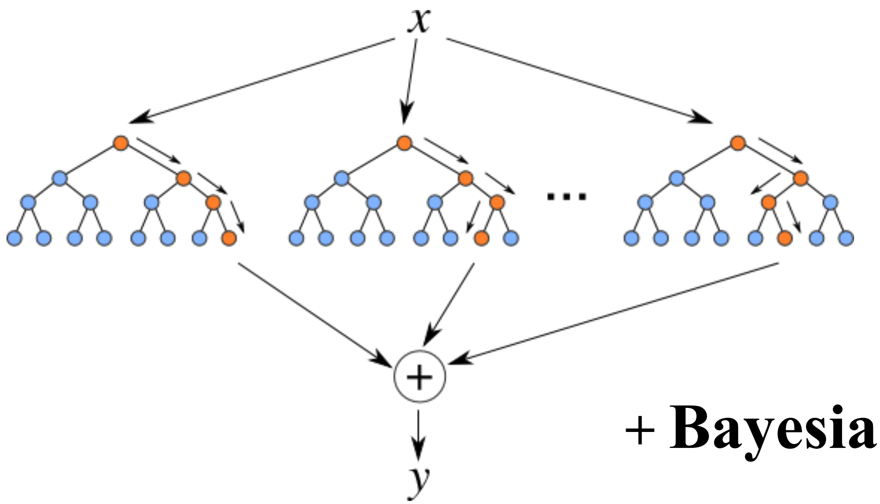
t-1

The rest of the variables

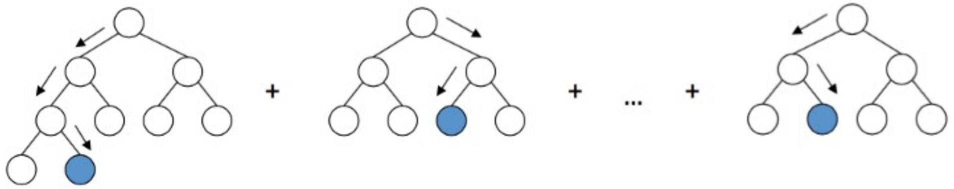
Models

Elastic net

Random forest



Gradient boosting



+ Bayesian optimization

Bayesian Neural Network

- BNN from Khabibullin, Seleznev (2020);
- Network size (30 and 10 neurons);
- Normal error distribution with Tanh activation function and tStudent distribution with ReLu activation function.

$$h_{1,t} = f(W_1 x_t + b_1),$$

$$h_{2,t} = f\left(W_2 h_{1,t} + b_2\right),$$

$$y_t = W_3 h_{2,t} + b_{y,t} + \varepsilon_t,$$

where x_t are input data, W_i are weights on the layers $i=1, 2$ and output layer, b_i and b_y are bias, $f(\cdot)$ is an activation function, h_i is the output of the hidden layer i, y_t is the output of the neural network at time t and ε_t is an error term.

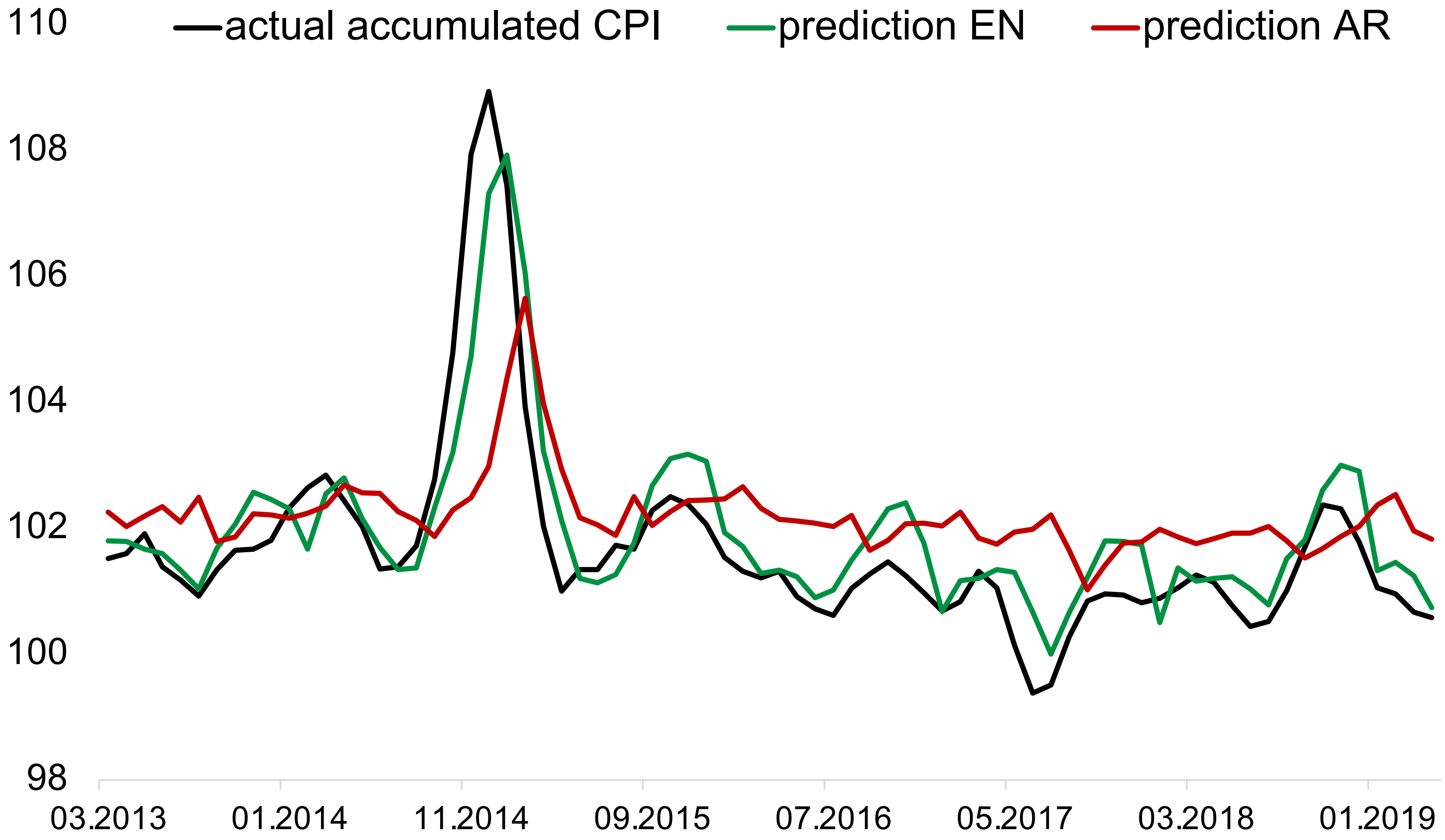
Model comparison, Exp.1:

RMSE levels for different models

lags	Elastic Net	Random Forest	Gradient Boosting	Neural network	AR
			1 month		
1	0,480	0,368	0,356	0,409	0,404
3	0,403	0,373	0,382	0,431	0,427
6	0,380	0,389	0,396	0,400	0,392
			3 months		
1	0,756	0,826	0,820	0,705	1,414
3	0,760	0,883	0,759	0,728	1,393
6	0,789	0,895	0,737	0,860	1,390
			6 months		
1	0,758	0,963	0,789	0,687	2,693
3	0,643	0,932	0,716	0,752	2,681
6	0,758	0,986	0,842	0,876	2,664

Note: AR model was estimated with seasonal dummies for proper model comparison. The lowest RMSE levels for each forecasting horizon are marked with grey.

Forecast comparison of EN and AR (for 3 month in advance with 1 lag):



The role of data availability:

Comparison of models RMSE, with and without taking into account data availability

	Elastic Net		**Random Forest**		**Gradient Boosting**		**Neural network**	
lags	$t-1$	$t-1/\,t-2$	$t-1$	$t-1/\,t-2$	$t-1$	$t-1/\,t-2$	$t-1$	$t-1/\,t-2$
1 month								
1	**0,359**	0,480	0,393	**0,368**	0,378	**0,356**	**0,338**	0,409
3	0,362	0,403	0,446	0,373	0,404	0,382	0,349	0,431
6	0,447	0,380	0,428	0,389	0,375	0,396	0,411	0,400
3 months								
1	**0,676**	0,756	**0,772**	0,826	**0,710**	0,820	**0,626**	0,705
3	0,697	0,760	0,816	0,883	0,789	0,759	0,787	0,728
6	0,720	0,789	0,814	0,895	0,790	0,737	0,765	0,860
6 months								
1	0,722	0,758	0,914	0,963	0,735	0,789	**0,685**	0,687
3	0,701	**0,643**	**0,903**	0,932	**0,716**	0,716	0,698	0,752
6	0,743	0,758	0,927	0,986	0,809	0,842	0,916	0,876

The role of data vintages:

Comparison of models RMSE, with and without taking into account data availability

	Elastic Net		**Random Forest**		**Gradient Boosting**		**Neural network**	
lags	vintages	no vintages	vintages	no vintages	vintages	no vintages	vintages	no vintages
1 month								
1	**0,359**	0,369	0,393	**0,373**	0,378	**0,334**	**0,338**	0,348
3	0,362	0,399	0,446	0,386	0,404	0,366	0,349	0,402
6	0,447	0,392	0,428	0,399	0,375	0,392	0,411	0,414
3 months								
1	**0,676**	0,697	0,772	**0,732**	0,710	**0,677**	0,626	**0,615**
3	0,697	0,711	0,816	0,824	0,789	0,803	0,787	0,766
6	0,720	0,751	0,814	0,881	0,790	0,848	0,765	0,822
6 months								
1	0,722	0,767	0,914	**0,901**	0,735	0,767	0,685	0,669
3	0,701	**0,690**	0,903	0,940	**0,716**	0,731	0,698	**0,657**
6	0,743	0,690	0,927	0,972	0,809	0,781	0,916	0,889

The role of seasonal adjustment:

Comparison of models RMSE, with and without seasonal adjustment

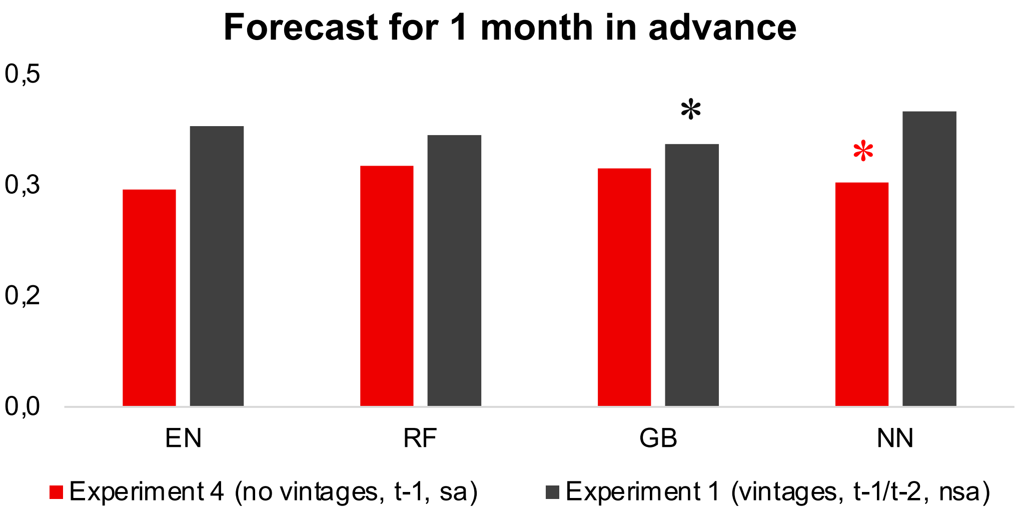
	Elastic Net		**Random Forest**		**Gradient Boosting**		**Neural network**	
lags	sa	nsa	sa	nsa	sa	nsa	sa	nsa
1 month								
1	0,348	0,369	**0,326**	0,373	**0,323**	0,334	**0,304**	0,348
3	**0,295**	0,399	0,346	0,386	0,344	0,366	0,347	0,402
6	0,377	0,392	0,353	0,399	0,344	0,392	0,391	0,414
3 months								
1	**0,622**	0,697	**0,686**	0,732	0,700	**0,677**	0,625	**0,615**
3	0,675	0,711	0,715	0,824	0,682	0,803	0,652	0,766
6	0,664	0,751	0,738	0,881	0,732	0,848	0,716	0,822
6 months								
1	0,666	0,767	**0,843**	0,901	**0,671**	0,767	0,837	0,669
3	**0,599**	0,690	0,846	0,940	0,791	0,731	0,804	**0,657**
6	0,658	0,690	0,880	0,972	0,827	0,781	0,921	0,889

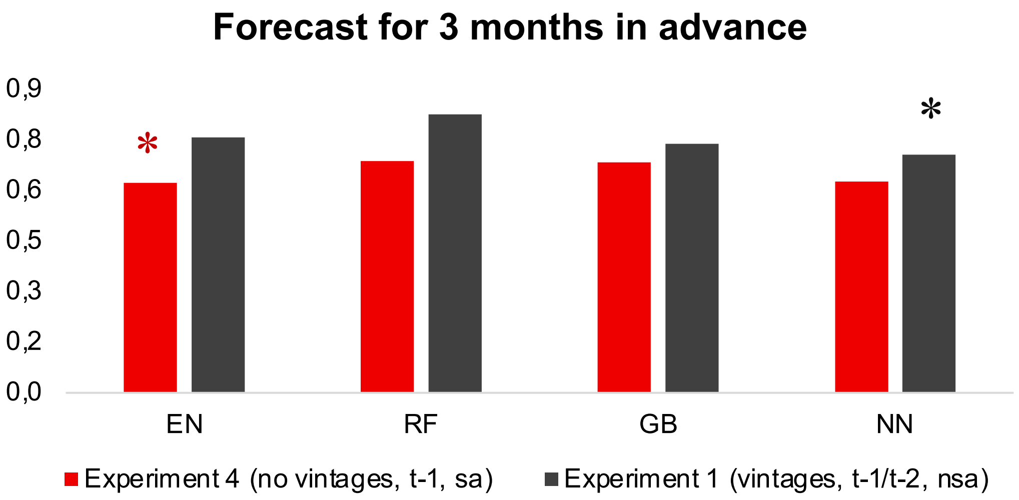
Joint effect of all data transformations:

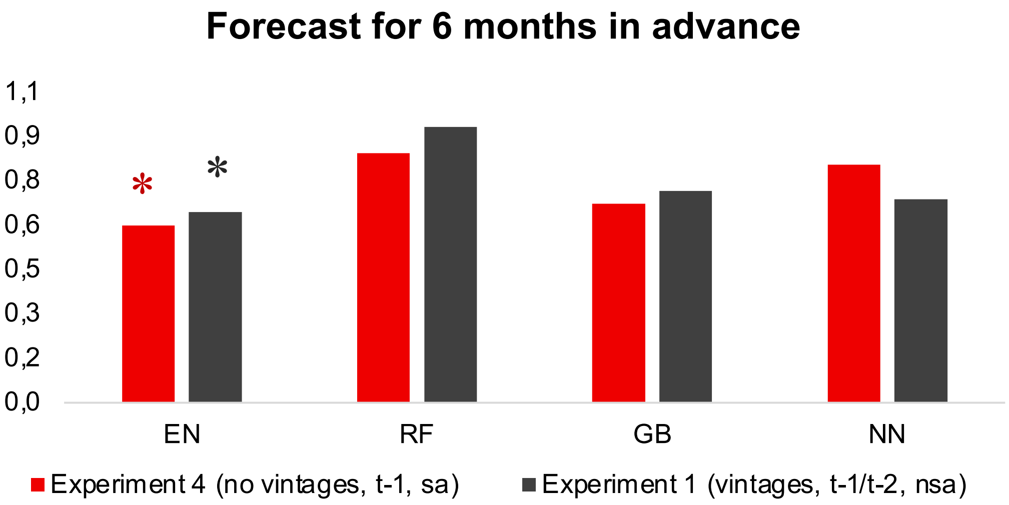
Comparison of models RMSE in two experiments

	Elastic Net		**Random Forest**		**Gradient Boosting**		**Neural network**	
lags	exp. 4	exp.1	exp. 4	exp.1	exp. 4	exp.1	exp. 4	exp.1
1 month								
1	0,348	0,480	**0,326**	0,368	**0,323**	0,356	**0,304**	0,409
3	**0,295**	0,403	0,346	0,373	0,344	0,382	0,347	0,431
6	0,377	0,380	0,353	0,389	0,344	0,396	0,391	0,400
3 months								
1	**0,622**	0,756	**0,686**	0,826	0,700	0,820	**0,625**	0,705
3	0,675	0,760	0,715	0,883	**0,682**	0,759	0,652	0,728
6	0,664	0,789	0,738	0,895	0,732	0,737	0,716	0,860
6 months								
1	0,666	0,758	**0,843**	0,963	**0,671**	0,789	0,837	**0,687**
3	**0,599**	0,643	0,846	0,932	0,791	0,716	0,804	0,752
6	0,658	0,758	0,880	0,986	0,827	0,842	0,921	0,876

Comparison of Exp.1 and Exp.4:







Results

- The use of all data (including unavailable) underestimate an error by 11%, 8% and 2% in average for the forecasting for 1, 3 and 6 months in advance correspondingly.
- Forecasts based on the final data have lower error on average (8%, 5% and 2%), yet the results are very sensitive to the model type, number of lags and forecast horizon.
- The use of seasonally adjusted data leads to the lower error (12%, 9% and 9%).
- The joint effect of all three transformations is on average 17%, 13% and 8% depending on the forecasted horizon.
- The results point in favor of GB (1 month), NN (3 months) and EN (6 months) comparing to AR.

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Getting insight of employment vulnerability from online news: a case study in Indonesia[1]

Nursidik Heru Praptono and Alvin Andhika Zulen,
Bank Indonesia

[1] This presentation 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 central banks and other institutions represented at the event.

Getting Insight of Employment Vulnerability from Online News: A case study in Indonesia

Nursidik Heru Praptono[1], Alvin Andhika Zulen[2]

Abstract

As mass media nowadays moves into online platform, exploring textual information related to economic condition from online news is becoming computationally straightforward and is an interesting opportunity. We develop an online news-based indicator in order to help getting insight on the condition of employment vulnerability. A large number of online high frequency news captured from various news websites are then utilised to construct such indicator. Our finding is that some simple inference models for text classification combined with index calculation gives a promising information that strongly reflects the rate or the risk of being unemployed, given a certain time and a certain sector. The indicator built confirms other related indicators such as consumer confidence index from our consumer survey, indicator of job vacancy in Indonesia, and national statistics office data, represented by some number of Pearson correlation values. In addition to that, we also demonstrate that during the pandemic of COVID-19 in 2020, this indicator showed a sharp inclining curve especially in the 2nd quarter, reflecting that during this period so many labours are in a very high risk. We suggest that this indicator can be potential to support central bank policies, either as a leading indicator or as a quick alternative of survey based indicator.

Keywords: text mining, machine learning, employment vulnerability, unemployment rate, online news.

JEL classification: C38, C55, C11, E24, E27

[1] Department of Statistics, Bank Indonesia. email: nursidik_hp@bi.go.id (corresponding author)

[2] Department of Statistics, Bank Indonesia, email: alvin_az@bi.go.id

Contents

1. Background

The importance of employment vulnerability becomes obvious as the condition of massive unemployment can affect many aspects related to the macro-economy as a whole (IMF (2010)). The systemic financial stability for example, can be affected by the increase of unemployment (Hada et al. (2020)). The ones' ability of purchasing becomes relatively lower if the unemployment rate is high. The risk of increasing non-performing loan (NPL) would also become unavoidable. It is thus a serious concern for macroprudential policy makers to consider pay attention on employment vulnerability as it can yield any further serious systemic risks.

The unemployment and its rate is still a global issue at least within the last 4 years, with various emphasises. The International Labour Organisation (ILO) in 2018, stated that global unemployment remains elevated by more 190 million, while the vulnerable employment was recorded still on the rise (ILO (2018)). In 2019 ILO also reported that decent work deficits were widespread (ILO (2019)). The trend in 2020 stated that the total labour underutilisation was doubled as high as unemployment and about 8.8 percent of total working hour were lost -- of the COVID-19 pandemic effect -- despite about 30 million of new jobs were created (ILO (2020)). Finally, in 2021 ILO (2021) stated that the huge impact of COVID-19 proliferated so many aspects including long term unemployment issue due to economy recovery. Given those evidences, monitoring the condition of unemployment and its potential is important to conduct in order to see the country's economic health.

In the recent years, we have also witnessed that the access for information has obviously becomes easy as we are in the digitalisation era. This phenomenon enables us to access some insight from high resolution of data. Online news textual data is not an exception. Research have been conducted to investigate some information about e.g. economic related from textual information such as news or social media.

Through this paper, we propose a methodology on investigating the employment vulnerability index utilising online news data. The organisation of this paper is as follows: Section 2 describes the literature review related to the employment vulnerability/unemployment rate and some research related to text analytics utilisation to support economic/macroeconomic related-use cases. Section 3 describes our proposed methodology, while in Section 4 we discuss the results of our proposed methodology. Finally, we conclude our works and discuss some further directions in Section 5.

2. Literature Review

The employment vulnerability, in general, is a condition when an employee (or a group of employees) has a tendency that they are in the lack of decent working condition, lack of adequate social security, and shut 'voice' out through a representative board or similar related organisation (Johnson (2010)). This kind of employee typically either belongs to own-account worker and/or contributing family worker. The employment vulnerability however has a close relation to the unemployment rate, although the latter is rather reflecting a condition/degree of unemployment. Identifying employment vulnerability could help anticipating massive unemployment rate.

Indonesia, as one of the most populous countries in the world is not without exception in having unemployment issue. The level of unemployment rate in Indonesia is still relatively higher than most ASEAN countries (OECD, (2020)). Our national statistics office (BPS) recorded that by February 2021, the open unemployment rate in Indonesia reached about 6,26%, lower than the previous 6 month (7,07% in Augsut 2020). The method conducted by BPS in order to produce such information is based on survey called Survei Angkatan Kerja Nasional (Sakernas) -- national labour force survey, whereas the respondents are in the productive age (BPS (2021)). However, as we may see, the obvious limitation on measuring the unemployment rate is that it is based on the 6-monthly basis. For e.g. macroprudential policy makers that need quicker analysis, an alternative way to gather the likely related information of unemployment in high frequency is thus required. One of the alternative data source to explore is online news, that is textual data that can be produced in near real time.

The research related to the utilisation of textual analytics on economic news have recently been conducted. Baker et al. (2016) in their work utilised news data from 10 leading US newspapers in order to construct the economic policy uncertainty index. The method the implemented is by providing a set of terms that reflect the economic uncertainty. The index is then constructed based on the classified news data compared to the available article data. Generally, the text analysis conducted in this case is rather of deterministic approach.

Another work on economic-related text analytics that utilising machine learning is the work of Tobback et al. (2017). Using textual data from media, their experiment result showed that a hawkish-dovish degree related to the central bank's communication measurement is better constructed by leveraging support vector machine. Moreover, in their report, Latent Dirichlet Allocation (LDA) is then utilised to identify the topic, on to which certain degree of communication measured.

The use of social media as the source of economic textual data has also been conducted, for example by Bollen et al. (2011). The stock markets are predicted by analysing the mood information inferred from twitter data. It is found that the prediction accuracy reached about 87,6% on the prediction given their prediction model setup.

Another experiments on social media data that is related to the issue is the work by Antenucci et al. (2014). The twitter data is used to see the labour market flows by creating jobs related indexes, ranging from "job loss", "job search", and "job posting". The result demonstrates that such constructed index, given the setting, can be used as the consideration related to insurance policies/support.

Recently, research work on more related to the employment vulnerability from the textual data has been conducted by Bailliu et al. (2018). They developed an index called Chinese Labour Market Conditions Index (LMCI) in order to measure the condition of labour market in China. This work utilises Support Vector Machine (SVM) to classify the newspaper article. Furthermore, their result also suggested that having setup as defined, the LMCI can be used in forecasting the labour market condition.

Having the discussion as above, we then generally summarise two main research questions to answer through our experiment:

1. How to classify the text with the limitation of training data while having prior knowledge?

Of the textual mentioned above, we however still have a big challenge when utilising textual data using either deterministic (rule based with some predefined keywords) or machine learning models. When using rule based method, the model's generalisation ability is solely based on human prior knowledge. On the other hand, machine learning model relies heavily on the data and thus can suffer from low performance when the quantity and the quality of the training data is insufficient.

2. Is there any better suggested method to construct employment vulnerability index with fine grained time basis (near realtime)?

To the best of our knowledge, there is currently no official survey based data that we can refer related to the employment vulnerability index in Indonesia. Even if it does exist -- say informally--, the cost could be very expensive. Moreover, it may suffer from non-trivial limitation such as subjective bias and less seamless information captured.

3. Methodology

3.1. Data

3.1.1. News Article

We utilise news articles for our main source of data, from more than 30 various domestic online news portals from January 1998 to August 2021. The average of total article is about 850 per day and about 27.000 per month. In order to demonstrate our methodology, we define some setups as the following:

1. **News filtering**: Of those the whole news, we filter the articles for where there is at least one sentence that contains any keywords related to unemployment. Thus we first perform sentence tokenisation onto each article. If an article does not contain any such keywords, then we simply flag this article as not indicating any employment vulnerability.

2. **Data Annotation**: Of the filtered articles as described in point (1) above, we subsample the data per month such that we obtain a 10% on each month. The time interval of this data ranges from January 2020 up to December 2020. This filtered data are then pooled out for annotation process by human, i.e. to flag whether an article is 0 or 1, based on the filtered sentence(s). Here label 1 means "the article indicates any information related to employment" where 0 means "the article does NOT indicate any information related to employment". The overall annotated data consists of 2979 of class 1 and 3167 of class 0. Of the number of records in class 1, we identify that there are 196 articles belong to manufacturing sector and 55 belong to service sector. We utilise this annotated data to construct and/or evaluate the text-classifier model.

3. **Full Data**: the full data for demonstrating the model on classifying the texts and constructing the index are of those articles from August 2018 to August 2021 (monthly).

3.1.2. Supporting Data for Index Evaluation

For the evaluation of our constructed index, we elaborate the job vacancy index data obtained by various online job bursaries. This data represents the number of job vacancy within a certain period. The higher job vacancy index, the higher job offered on the market. The index is available up to monthly. Another data that we have is consumer confidence index (CCI) provided monthly from our Consumer Survey. We also utilised the GDP data by specific sector, provided by BPS and available quarterly.

3.2. Inference Models

In this part we will discuss some alternatives of the inference models and their properties. Given an x (article), the model is to predict its corresponding category (class) $\hat{y} \in \{0, 1\}$. First we describe the deterministic model, then model-from-data and finally suggest the inference model that incorporating prior knowledge into model-from-data. The idea behind our experiment is that we would like to utilise our prior knowledge, expecting that it will help the inference process. This is as an alternative when we have any limitation to access the training data, while still be able to make any room for prior knowledge for inference process.

3.2.1. Deterministic Model

The deterministic model we apply here leverages the rule based model and some selected keywords. The objective of the rule is basically to find any pattern indicating the evidence of unemployment within the selected article.

$$\begin{aligned} r(x) = \exists(\mathrm{kw}_{\mathrm{unemployment}}) \text{ in } x \bigwedge \nexists\,(\mathrm{kw}_{\mathrm{neg1}}, \mathrm{kw}_{\mathrm{unemployment}}) \text{ in } x \\ \bigwedge \nexists(\mathrm{kw}_{\mathrm{unemployment}}, \mathrm{kw}_{\mathrm{neg2}}) \text{ in } x \end{aligned} \qquad (1)$$

Some keywords example used by Eq. 1 can be seen in Table 1 as the following.

Some keywords example and their English meanings

Table 1

Keywords Category	List of Keywords Example	Representative Meaning (English)	Notes
Keywords unemployment (kw_unemployment)	pemutusan hubungan kerja, phk, pemulangan, pengangguran, layoff, pemecatan, memecat, dipecat, pemulangan,...	≈ Fired-out, layoff	-
Keywords negation 1 (kw_neg1)	tidak	≈ no/there is/are/were no	Co-occurance: before kw_unemployment
Keywords negation 2 (kw_neg2)	menurun, turun, berkurang, melandai, menyusut, rendah,...	≈ decreasing, becomes lower.	Co-occurance: after kw_unemployment

For further utilisation in our experiment when elaborating with model-from-data (machine learning), we simply refer this deterministic model $r(x)$ as our "prior knowledge". This is because this model is simply as the representation of human knowledge.

This model in one hand is rather simple, rigid, and straightforward to apply. However, it is very sensitive to the change of characteristics of the data. The assumptions provided to construct the rule must capture all the possibilities of the whole patterns in order to obtain best performance, and so do the keywords. Otherwise, if the model is too much simple, it may result so many unexpected noises.

3.2.2. Inference Model from Data

Another approach to classify the filtered text with more flexible way is to build a model-from-data (or in general term we may say "machine learning"). In our case we use Logistic Regression to demonstrate the inference model. Thus we have:

$$p(y = 1|x) = \sigma\big(f(x)\big) = \frac{1}{1 + e^{-f(x)}} \tag{2}$$

where $f(x) = w\phi(x)$ represents our linear function, with w is the models parameter and ϕ as basis function. In our case, we simply use simple polynomial basis function with degree $m = 1$, thus $\phi(x): x \to x$.

Given the annotated dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ for $x \in \mathbb{R}^d$ and $y \in \{0,1\}$ the objective function is to minimise its negative log likelihood, that is

$$J = \sum_i \ln(1 + e^{\left(-(2y_i-1)f(x_i)\right)}) \tag{3}$$

Having the formulation as defined above, the goal of our optimisation problem is thus to find w that minimise J on such setting. In our experiment, we demonstrate the estimation of the parameter by quasi-newton approximation with Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. Note that before the article is used by the model, we perform feature extraction first by leveraging TFIDF[3]-bag of words feature extraction on the related sentences so that each article can be represented by a feature vector.

As it is of any general machine learning methods, this approach is more flexible in term of constructing knowledge as long as the samples obtained are representative enough to the whole population. However, when the number of the training data is rather limited or less representative to the population, the model learns insufficiently and perform badly on the prediction process.

3.2.3. Incorporating the Prior Knowledge

We finally tried to incorporate our prior knowledge into the model-from-data for inference process. While straightforward, incorporating prior knowledge needs concise modelling so that the it is properly blended into the model-from-data. We adopt the work of Schapire et al. (2002) for where they incorporate prior knowledge into AdaBoost. However, in our case we use the simpler model Logistic Regression to demonstrate the idea while applying into our text classification problem.

First, we introduce a probability distribution that enables to represent our prior knowledge. We refer this prior probability as π_+, that is the probability that any x would belong to class 1. In general, the form of the prior distribution indeed may vary depending on ones' choice. It may be justified either by human belief or by another mechanism that might rely on the data first before learning. We define our prior knowledge distribution in this case as in Eq. 4 follows:

[3] TF = Term Frequency; IDF = Inverse Document Frequency

$$\pi_+ = p(y = 1|x) = \begin{cases} 0.9, & \text{if } r(x) = 1 \text{ (True)} \\ 0.1, & \text{otherwise} \end{cases} \tag{4}$$

Here $r(x)$ is obtained by our deterministic model as described in 3.2.1 previously. Simply speaking, if any article follows the rule, then we inject the probability value as 0.9, while 0.1 otherwise.

This prior probability is then taken into the objective function when estimating the model's parameter. The objective function on Eq. 3 thus becomes:

$$J = \sum_i \ln\left(1 + e^{(-(2y_i-1)f(x_i))}\right) + \underbrace{\eta D_{KL}(\pi_+(x_i)||\sigma(f(x_i)))}_{\text{control on prior information}} \tag{5}$$

The second part of the equation represents the control on our prior knowledge. It is quantified by Kullback-Leibler Divergence between π_+ (our prior knowledge representation) and "information from training data", $f(x)$. The value of η controls the importance of our prior information. Here we set its value into 1 and we leave as it is since in this experiment, it is not our main focus. In estimating the parameters, we apply the quasi-newton approximation with BFGS algorithm to demonstrate the learning process. As in the previous section on model-from-data, we first convert our article into feature vector by leveraging TFIDF bag-of-words feature extraction.

Once the parameter is estimated, we then introduce the prior term h_0 in order to enable the blending process into the model-from-data's term. This h_0 is obtained by the inverse of our prior probability π_+, so that:

$$h_0(x) = \sigma^{-1}(\pi_+(x)) = \ln\left(\frac{\pi_+(x)}{1 - \pi_+(x)}\right) \tag{6}$$

The final prediction model is thus of the form of logistic function of the final blending $p(y = 1|x) = \sigma(f^*(x))$ where $f^*(x) = f(x) + h_0(x)$ represents our final blending term and $\sigma(.)$ is the logistic function.

3.3. Index Construction

Once the full data have been categorised whether belong to 1 (article indicates any contains related to employment vulnerability) or 0 (article does not indicate), the index is then constructed. The equation below formulates how we construct the employment vulnerability index.

$$\mathrm{idx}_t = \frac{\sum_i 1_{f(x_{i,t})=1}}{\mathrm{N}_t} x \log \mathrm{N}_t \tag{7}$$

Generally, the index is constructed based on the proportion of related articles among the overall articles within a time. Here $\sum_i \mathrm{I}_{f(x_{i,t})==1}$ indicating the total number of related articles, while N_t represents the total of articles within a time t.

The log term in the Eq. 7 represents the importance, or the magnitude of the information. We realise that there may be a difference of the total articles between time. Thus, the more article within a particular time, the more important ratio should be considered. We understand that this approach might not be a perfect formulation but we fundamentally argue that it is of the most reasonable approach that we can utilise.

4. Result and Discussion

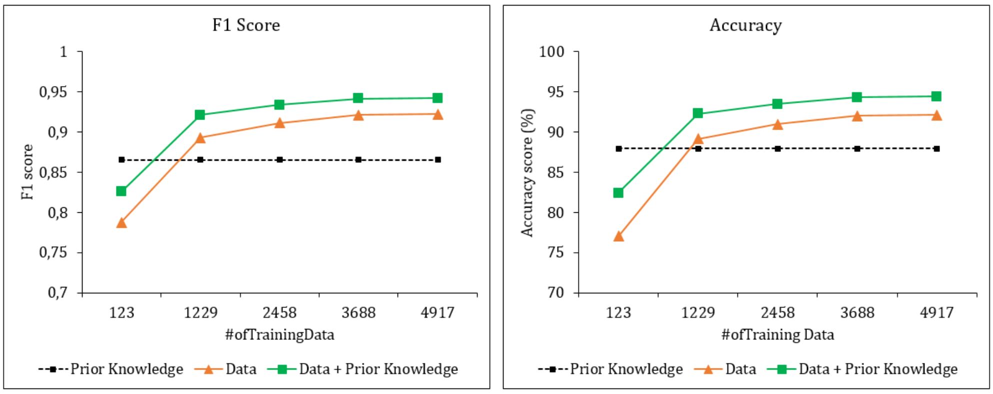
4.1. Evaluation on Text Classification Models

On experiment for text classification, we test each model into a fixed test data. As mentioned on the subsection 3.1.1 above, we have 6.146 annotated data points (2.979 of class 1 and 3.167 of class 0). Of those overall annotated data, we spare 20% as fixed testing data. The remaining portion is used for training experiments.

We compare three different approaches of inference models, which are deterministic/rule based (we note as "Prior Knowledge"), model-from-data/machine learning (we note as "data") and the incorporation of prior knowledge into model-from-data/machine learning (we note as "Data + Prior Knowledge"). We evaluate the model's performance with F1-score and accuracy score.

Experimental Results on Text Classification Model

Figure 1



a. F1 score of each model

b. Accuracy score of each model

The experiment for text classification can be seen as in Figure 1 above. Initially, when the size of training data is relatively small, the "data" model performed worst. We can see on Figure 1.a that our "prior knowledge"-- that is rule based -- is even still better than of "data" model on limited training data. As we increase gradually the size of training data, the F1-score is getting better. The "prior knowledge" only is constant, as it is a deterministic function.

It is can also clearly be seen that incorporating "prior knowledge" helps the performance of "data" models at any condition. On a small dataset, although below the rule based model, the performance when we incorporate prior knowledge help increase the model-from-data only. However, when there is sufficient amount of training data inference model that incorporating prior knowledge performs very well compared to model-from-data only or prior knowledge only.

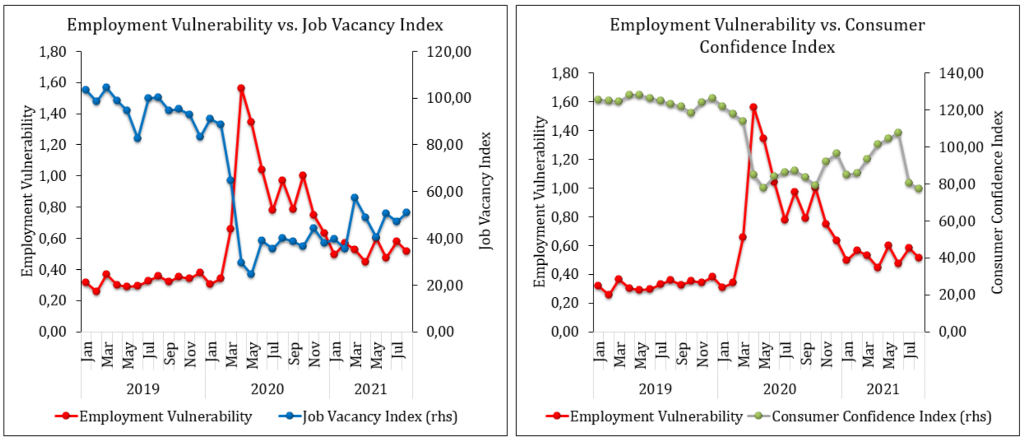
The accuracy of the model, on the other hand also show the similar trend, as can be seen in Figure 1.b. Initially, the model-from-data poses lowest accuracy on limited training data and while prior knowledge model is still the highest. Incorporating the prior knowledge, increase the accuracy although it is still lower than prior knowledge only. Once we have sufficient training data, the performance getting better when blending those both information for inference process.

4.2. Employment Vulnerability Index and its Evaluations

Once the articles have been classified, the employment vulnerability index is then straightforwardly constructed. Utilising Eq. 7, we plot our index based on timely basis. Figure 2 below shows the general monthly employment vulnerability index and its comparison with Job Vacancy Index and with consumer confidence index.

Employment vulnerability index vs. job vacancy index and consumer confidence index (CCI)

Figure 2



a. Employment vulnerability index vs. job vacancy index

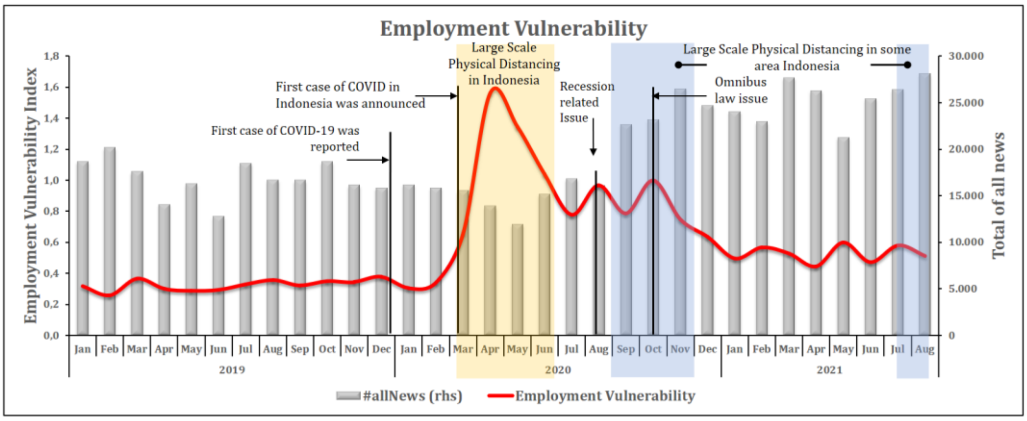
b. Employment vulnerability index vs. CCI

We found that generally the employment vulnerability index has the opposite direction to job vacancy index as we can see on Figure 2.a. above, having Pearson correlation of $\rho = -0.76$. In addition to that, its comparison to consumer confidence index also shows similar wise (as shown on Figure 2.b.), with the Pearson correlation of $\rho = -0.9$. The employment vulnerability index revealed the sharp inclining curve on the occurrence of the pandemic of COVID-19, especially since March 2020. During this period, so many labours are in a very high risk. This is due to the policies carried out by the government to anticipate the spread of COVID-19 by ruling strictly physical distancing related regulations on the whole country. The number of job vacancies drops dramatically in this period and so does the consumer's confidence.

We also can see that employment vulnerability index however is still relatively high at least until the end of 2020 compared to before. It is also found that there are some small fluctuating curves afterwards. We then performed the event analysis on the inferred articles as shown in Figure 3, mainly based on the content of the articles and related policies. We found that these small peaky curves indicate some events that can cause or have association with the employment vulnerability, such as the occurrence of recession issue, some demonstrations related to labour regularisation, and physical distancing on some particular area.

Event analysis on monthly employment vulnerability index

Figure 3

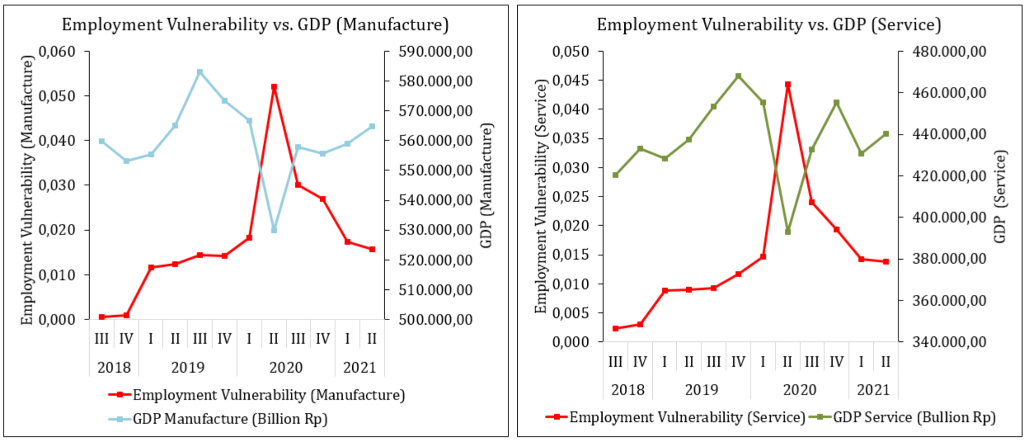


We then investigate the employment vulnerability on manufacturing sector and service sector, as they pose relatively significant sector on labour employment vulnerability. Most of the labours or employees are work on manufacture companies as well as service sector. We then compare such index with its GDP respectively.

The GDP of manufacturing sector consist of those GDP from manufacturing industries. The service sector consists of those GDP from some subsectors such as transportation and warehousing, finance and administration, company service, educational-related service, social/society-related service, and other services.

Employment vulnerability index on manufacturing sector and service sector

Figure 4



a. Employment vulnerability index on manufacturing sector and its GDP respectively

b. employment vulnerability index on service sector and its GDP respectively

We found that the employment vulnerability index on manufacturing sector has strong Pearson correlation with its GDP in the opposite direction ($\rho = -0.8$). The GDP drops dramatically on 2nd quarter of 2020 as well as the occurrence of pandemic COVID-19 (see Figure 4.a). It is then increasing gradually as the employment vulnerability index decreasing due to some regulation adjustments and economic policies for facing the new normal.

In the service sector, such trend applies similarly (see Figure 4.b.). The employment vulnerability index have strong opposite direction with its GDP, with the Pearson correlation $\rho = -0.72$. The GDP curve also dropped dramatically on the 2nd quarter of 2020, as the so many employees are in very high risk during such condition. It then increased gradually as the employment vulnerability index decreased also together with some regulation adjustment and economic policies on the new normal era.

We have shown that some macroeconomic-related indicators confirmed the employment vulnerability index. Having text classification based index construction above helps policy maker get insight on what would be the condition of employment's vulnerability. Although we demonstrated and discussed monthly and quarterly time basis, the employment vulnerability index can however straightforwardly be presented in the daily basis timeframe. Such finer grained representation is useful as it can provide the information in near real time.

5. Conclusion & Future Directions

5.1. Conclusion

Towards this research we conducted a methodology for assessing the employment vulnerability from online news. We demonstrated on how we enable our prior knowledge to be incorporated into the model-from-data (machine learning model). We found that given limitation of training data, the incorporating prior knowledge helps the inference models when classifying the texts. On the other hand, the use of model-from-data can also help the rule-based model as it enables the construction of knowledge representation from data. Thus, we suggest that incorporating prior knowledge for such inference task is important to consider.

The index constructed generally confirmed by related indicators, including e.g. job vacancy, CCI, and GDP-per sector. This has been shown by relatively strong Pearson correlation on the demonstrated time range, either on general unemployment vulnerability index, or per sector unemployment vulnerability index. The index constructed from online news is a form of very high frequency data in that it can be obtained by daily. Thus, we also suggest that the proposed methodology can be used as a leading indicator or as a quick alternative to survey based index of employment vulnerability.

5.2. Future Directions

We notice that there still some improvements to do for our works. We highlight some future directions:

- Stratify the level of employment vulnerability index

Instead of binary classification, we suggest that the classification task is possible to expand to multi label classification representing the level of vulnerability. This is also pointed out by survey based EVI (Baum, S., & Mithcell, W., (2020)) in that they stratify the level into high, medium, low risk. The higher level should obviously be more concern for policy makers to do some essential and important decisions to prevent any worse condition.

- Enhancement on Spatial Information

The proposed methodology is solely based on the overall area on a state. However, it also may be important to localise the employment into some specific area. This is to help seeing the distribution of vulnerability by spatial information, either on binary categorisation, or multi-level/stratified categorisation.

- Enhancement on more sectors.

The two sectors discussed in this works are based on the majority of workers. However, it is also important to expand the sectors horizontally, or vertically (detailing into subsectors). However, this also depends on the purpose on how detailed information one should investigate, by sectors/subsectors.

References

Antenucci, D., Cafarella, M., Levenstein, C., M., Re., C., & Shapiro, M., D., (2014). Using Social Media to Measure Labor Market Flows. *NBER Working Paper No. 20010*.

Bailliu, J., Han, X. &, Kruger, M., (2018). Can media and text analytics provide insights into labour market conditions in China?. *Bank of International Settlements working paper*. URL: https://www.bis.org/ifc/publ/ifcb49_44.pdf

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics, 131*(4), 1593-1636.

Baum, S., & Mithcell, W., (2020). Employment Vulnerability Index (EVI) 3.0. Retrived from EVI's official website. URL: http://www.fullemployment.net/publications/reports/2020/EVI_3.0_Final_Report.pdf.

Bollen, J., Mao, H., & Xiao-Jun, Z. (2011). Twitter Mood Predicts the Stock Market. *Journal of Computational Science, 2*(1), 1-8.

Hada, T., Barbuta-Misu, T., Iuga, I. C., & Wainberg, D., (2020). Macroeconomic Determinants of Nonperforming Loans of Romanian Banks. *Sustainability 12* (7533), 1-19.

International Labour Organisation (2018). *World Employment Social Outlook*. Retrieved from official ILO's website. URL: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_615594.pdf

International Labour Organisation (2019). *World Employment Social Outlook*. Retrieved from official ILO's website. URL: https://www.ilo.org/global/research/global-reports/weso/2019/lang--en/index.htm

International Labour Organisation (2020). *World Employment Social Outlook*. Retrieved from official ILO's website. URL: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_734455.pdf

International Labour Organisation (2021). *World Employment Social Outlook*. Retrieved from official ILO's website . URL: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_795453.pdf

International Monetary Fund (2010). Unemployment dynamics during recessions and recoveries: Okun's law and beyond in World Economic Outlook (Chapter 3). pp. 1-39. URL: https://www.imf.org/~/media/Websites/IMF/imported-flagship-issues/external/pubs/ft/weo/2010/01/pdf/_c3pdf.ashx

Johnson, J. L., (2010), ILO Online: How do you define 'vulnerable employment?'. Retrieved from official interview with ILO chief of employment trends unit. URL: https://www.ilo.org/global/about-the-ilo/mission-and-objectives/features/WCMS_120470/lang--en/index.htm

OECD, (2020). Promoting Stronger Local Employment in Indonesia. Retrieved from OECD's official website. URL: https://www.oecd-ilibrary.org/sites/1f8c39b2-en/index.html?itemId=/content/component/1f8c39b2-en#chapter-d1e9121

Schapire, R., E., Rochery, M., Rahim, M., & Gupta, N., (2002). Incorporating Prior Knowledge into Boosting. Proceeding of the Nineteenth International Conference on Machine Learning, (pp. 538-545).

Statistics Indonesia (BPS), (2021). Tingkat Pengangguran Terbuka. Retrieved from BPS' official website. URL: https://www.bps.go.id/indicator/6/543/1/unemployment-rate-by-province.html.

Tobback, E., Nardelli, S., & Martens, D. (2017). Between Hawks and Doves. *NBER Working Paper No. 2085*.

Getting Insight of Employment Vulnerability from Online News: a Case Study in Indonesia

[1]Nursidik Heru Praptono, [2]Alvin Andhika Zulen

[1]nursidik_hp@bi.go.id
[2]alvin_az@bi.go.id

October 2021

Outline

Background

Methodology

- Data & Prior Knowledge
- Inference Model
- Experimental Result

Employment Vulnerability Index

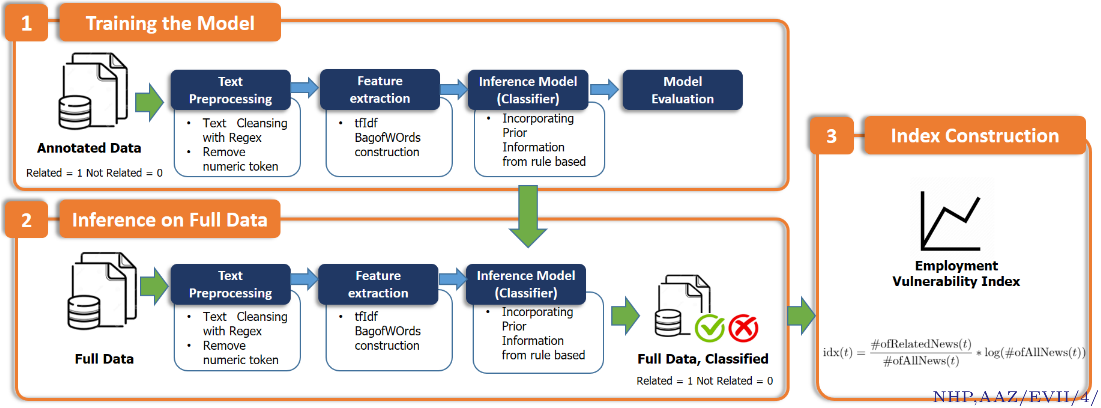
Conclusion and Future Directions

Background

- The unemployment rate is obviously an important factor that can reflects the health of the economy, e.g. systemic financial stability.
- Employment Vulnerability Index in Indonesia
  - Can have insight for macroprudential policy makers in order to make any further decision.
  - Currently there is no survey to construct the employment vulnerability index. Even if it exists, it may be expensive, subjective/biased and less seamless.
- Extraction economic-related information from news, e.g.:
  - Economic Policy Uncertainty (Baker et al, 2016)
  - Labour Market Condition Index, LMCI (Bailliu et al, 2018)
- Thus we introduce a methodology to enable employment vulnerability index, leveraging online news.
  - Intuitively, the more number of published articles related to e.g. unemployment, then there may potentially be more information that reflect the condition of employment vulnerability.

Overall Methodology

To construct the index, we first classify the text that possibly belongs to 1 (related) or 0 (not related). In general, our text classification methodology is rather straightforward as shown in part 1 and 2 of the diagram below. However we consider incorporating prior knowledge into our classifier in order to anticipate limited number of training data available. Once data is inferred, then the index is constructed (part 3).



1 Training the Model

Annotated Data

Related = 1 Not Related = 0

Text Preprocessing
- Text Cleansing with Regex
- Remove numeric token

Feature extraction
- tfIdf BagofWOrds construction

Inference Model (Classifier)
- Incorporating Prior Information from rule based

Model Evaluation

2 Inference on Full Data

Full Data

Text Preprocessing
- Text Cleansing with Regex
- Remove numeric token

Feature extraction
- tfIdf BagofWOrds construction

Inference Model (Classifier)
- Incorporating Prior Information from rule based

Full Data, Classified

Related = 1 Not Related = 0

3 Index Construction

Employment Vulnerability Index

$$\mathrm{idx}(t) = \frac{\#\mathrm{ofRelatedNews}(t)}{\#\mathrm{ofAllNews}(t)} * \log(\#\mathrm{ofAllNews}(t))$$

Data & Prior Knowledge

Data

Overall data: Online domestic news data from more than 30 various news portal, from Jan 1998 to Aug 2021. Total average 850 news per day, and about 27.000 news per month.

- Annotated Data: 6146 (2979 of class 1, 3167 class 0), filtered by keywords related to unemployment. It is of sampled data from 10% randomly per month from year 2000 to 2020. Of the 2979 records of class 1, we identify 196 articles belong to manufacture sector and 55 articles belong to service sector.
- Full Data: All articles from 2018 to Aug 2021.

Prior knowledge setup: Rule and Keywords $r(x)$

We introduce rule and predefined keywords as a representation of our prior knowledge to classify the text. This model will then be used as our basic deterministic model.

$$r(x) = \exists(\text{kw\_unemployment}) \text{ in } x \wedge \nexists(\text{kw\_neg1,kw\_unemployment}) \text{ in } x \wedge \nexists(\text{kw\_unemployment, kw\_neg2}) \text{ in } x$$

list of keywords (their meaning in English):
kw_unemployment = {unemployment, layoff, fireout ....}, kw_neg1 = {not, avoid, reduce,...}, kw_neg2 = {decrease, step down,...}

Inference Model for Text Classification

Applying machine learning is challenging when there is limitation on accessing the training data. On the other hand, quite often we have intuition about something. Why not we elaborate our prior knowledge for the better inference?

- We adopt the methodology incorporating prior knowledge described by Schapire et al, 2002, but in our case we demonstrate the simpler model, Logistic Regression.
- **Objective Function**: Given the training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, and $y \in \{0, 1\}$. The objective function is to minimise negative log likelihood, controlled by the prior information.

$$J = \sum_i [\ln(1 + e^{(-(2y_i - 1)f(x_i))}) + \underbrace{\eta D_{\mathrm{KL}}(\pi_+(x_i)||\sigma(f(x_i)))}_{\text{control on prior information}}]$$

where f is linear function, σ is the logistic function. Here π_+ is our "prior information" defined by:

$$\pi_+(x) = p(y = 1|x) = \begin{cases} 0.9; \text{ if } r(x) = 1 \text{ (related article, by rule-based model } r\text{)} \\ 0.1; \text{ otherwise} \end{cases}$$

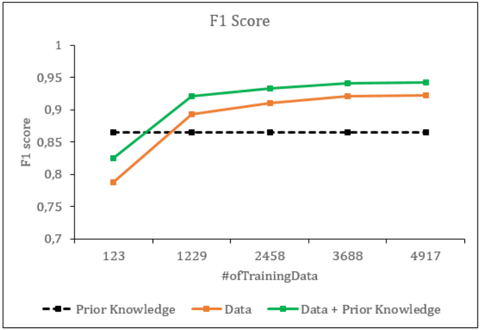
- **Inference Function**:

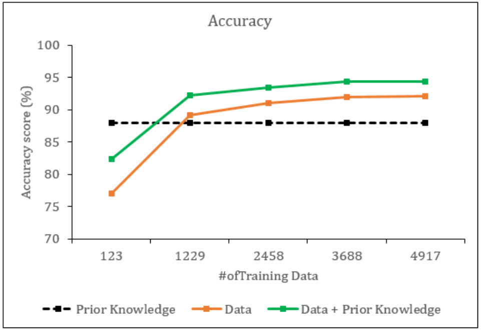
$$p(y = 1|x) = \sigma(f^*(x)); f^* = f + h_0$$

In this case, h_0 is the "prior term" defined as the inverse of logistic function of $\pi_+(x)$, that is $h_0(x) = \sigma^{-1}(\pi_+(x)) = \ln\left(\frac{\pi_+(x)}{1-\pi_+(x)}\right)$

Experimental Result

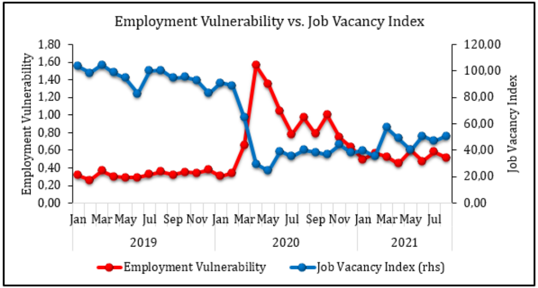
On a **fixed test data**, we compare three different approaches to classify the text. The model from data performs worst on small number of training data. As we increase the size of training data gradually the performance gets better. The model with prior knowledge only is constant, as it is a deterministic function leveraging rule based and predefined keywords. Incorporating prior knowledge into model-from-data's in this case helps in improving the performance, even when the number of training data is relatively small.

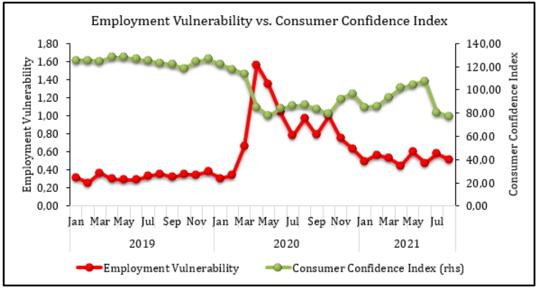




Employment Vulnerability Index

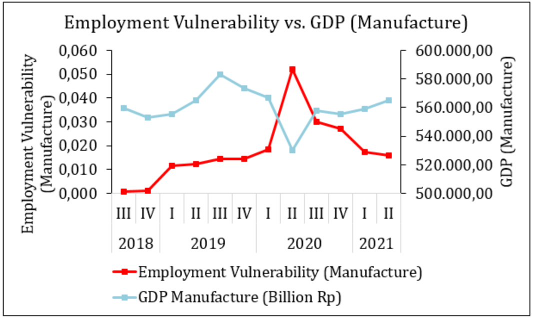
Once the articles are classified, we then construct the index. We found that generally the employment vulnerability index reveals the opposite direction with both the Job Vacancy Index ($\rho = -0.76$) and the Consumer Confidence Index ($\rho = -0.9$). The job vacancy index, describes the index of available job at a certain period. The consumer confidence index describes the consumer's confidence, obtained by survey. The curve shows an increase starting on March 2020 as in such period there were many actions to anticipate the COVID-19 pandemic.

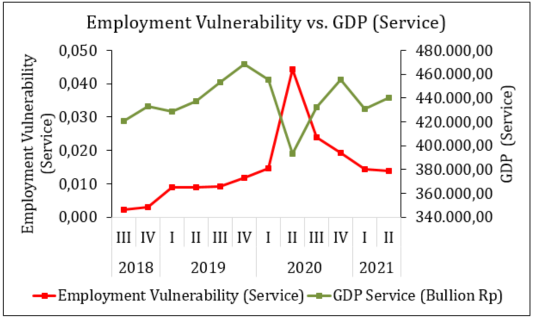




Employment Vulnerability Index - On Manufacture and Service Sector

Of the related news inferred by proposed approach, we also calculate the index per sector: manufacture and service. We found that there is relatively strong correlation (in opposite direction), on each sector with its GDP, $\rho = -0.8$ for Manufacture sector and $\rho = -0.72$ for Service sector respectively.





Conclusion and Future Directions

Conclusion

1. Incorporating prior knowledge into model-from-data can be helpful for text classification especially when the number of data is limited.
2. The proposed method shows that there is relatively strong Pearson correlation between the constructed employment vulnerability index and another related economic indicators. Thus we suggest that the method can be considered to use as an alternative way to survey based approach to monitor the employment vulnerability.

Future Directions

1. Instead of binary classification, we may stratify the employment vulnerability index by it's strength, e.g. high, medium, low risk – however, it depends on the purpose.
2. Enhancement on spatial information, e.g. specific area.
3. Enhancement on more sectors.

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Predicting foreign investors' behavior and flows projection in Indonesia government bonds market using machine learning[1]

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

[1] This presentation 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 central banks and other institutions represented at the event.

Predicting Foreign Investors' Behavior and Flows Projection in Indonesia Government Bonds Market Using Machine Learning

Anggraini Widjanarti[1], Arinda Dwi Okfantia[2], Muhammad Abdul Jabbar[3]

Abstract

Capital flows is one of the factors that greatly affect exchange rate stability in emerging markets. In Indonesia, the share of foreign investor ownership in government bonds market is large and increasing, signifying the importance of analyzing the behavior of foreign investors in government bonds market. This study aims to explain and predict the behavior and capital flows of individual foreign investors. We apply various machine learning techniques on daily data of government bonds transactions by foreign investors, combined with macroeconomic and market indicators. Investors are clustered using a data-driven algorithm based on their portfolio management, that is real money investors (long term) or traders (short term). For both investor groups as well as 35 investors with the largest share of government bond ownership, we build two additional sets of machine learning models: decision trees that explain each investor's behavior, and predictive models for the capital flows decision (net sale/net buy/hold) and the flows amount on a daily basis. The preliminary result show that the models have potential to support monetary operation strategy based on the direction of investors' decision.

Keywords: government securities, investors' behavior, flows projection, machine learning

JEL classification: C02, F34, E58

[1] Statistics Department – Bank Indonesia; E-mail: anggraini_widjanarti@bi.go.id

[2] Statistics Department – Bank Indonesia; E-mail: arinda_dwi@bi.go.id

[3] Statistics Department – Bank Indonesia; E-mail: muhammad_abdul@bi.go.id

Contents

1. Background

The movement of global capital flows has 2 main driving factors, namely: push and pull. Push factors are characterized by global macroeconomic conditions, central bank monetary policy, international financial market asset returns, and global liquidity. Meanwhile, pull factor capital flows from/to a country are determined by domestic macroeconomic conditions, perceived risk, and returns on domestic assets. In line with the conducive global push factor accompanied by the maintained Indonesian pull factor, the flow of foreign funds to Indonesia, especially to government bonds market (SBN) increased quite significantly in 2019. This condition aligns in many countries, which foreign holders make up the largest share of the investor base (Andritzky, 2012).

The increase in inflows to SBN on the one hand had a positive impact on the external balance in the context of deficit financing. The current account is getting wider. However, the increasing position of foreign investors in SBN has the potential to cause volatility in capital flows and exchange rates, which in turn disrupts economic stability (Agung & Darsono, 2012).

In line with high capital flows to Indonesia which will impact exchange rate volatility, it is necessary to understand the behavior of individual foreign investors in government bond market by classifying investors based on their portfolio management behavior, such as real money (long term) or trader (short term). By mapping foreign investors, we can see which groups of foreign investors are dominant and sensitive to financial market sentiment, so central bank can formulate more precise policy responses. In line with the increasing number and dynamic behavior of foreign investors in government bond market[1], it is necessary to calibrate the classification of investors with an enhanced methodology through data-driven techniques of Big Data Analytics. Big Data Analytics also have good potentials to predict the behavior of foreign individual investors in various scenarios of economic indicators or financial markets. The goal of this study is to develop methodology of Big Data Analytics that hopefully is able to strengthen the analysis of foreign investor behavior and to recalibrate foreign investor behavior classification.

2. Literature Review

2.1 Identification and Grouping of Foreign Investors

Identification of Unique Foreign Investors with Entity Resolution

One of the issue that we found when we explore the raw transactions that we have on our transaction database is that the investor name doesn’t fully represents a single entity of foreign investor that we need. Before we can do any analytics with the transactions data, we have to figure out how to identify different names as single entity of the investor of interests that we want to analyze.

[1] In 2019, several global bond indexes such as Bloomberg Barclays (BBGA), FTSE Russell (WGBI), and JP Morgan (GBI –EM) increased China's weight in the benchmark index, and the Norges Pension Fund which lowered its portfolio in EM debt which had an impact on capital flows to Indonesia.

Entity resolution (ER), the problem of extracting, matching and resolving entity mentions in structured and unstructured data, is a long-standing challenge in database management, information retrieval, machine learning, natural language processing and statistics (Getoor & Machanavajjhala, 2013). Entity resolution is necessary when there is clear indication that the entity that we have on our dataset doesn't necessarily reflect the needs of the big data analytics goals. Entity resolution can be done using several methodology from natural language processing to clustering. Measuring text similarity to group similar names together is one of the methodology to do entity resolution for text based entities. We experiment with 4 string similarity metrics in our study:

Jaro-winkler distance

Jaro–Winkler distance is a string metric measuring an edit distance between two sequences. It is a variant proposed in 1990 by William E. Winkler of the Jaro distance metric. Jaro-Winkler is computed by measuring Jaro distance and apply length of common prefix and constant scaling factor.

Normalized Levenshtein Distance

Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. It is named after the Soviet mathematician Vladimir Levenshtein, who considered this distance in 1965. Normalized Levenshtein Distance based on a study by Yujian & Bo in 2007 to make a normalized version of Levenshtein distance that can mathematically satisfy Triangle Inequality.

Weighted Levenshtein Distance

Weighted Levenshtein Distance or Damerau-Levenshtein Distance is a string metric for measuring the edit distance between two sequences. Informally, the Damerau–Levenshtein distance between two words is the minimum number of operations (consisting of insertions, deletions or substitutions of a single character, or transposition of two adjacent characters) required to change one word into the other. The Damerau–Levenshtein distance differs from the classical Levenshtein distance by including transpositions among its allowable operations in addition to the three classical single-character edit operations (insertions, deletions and substitutions) (Levenshtein, 1966).

Metric Longest Common Sub-sequences (MLCS)

The longest common subsequence (LCS) problem is the problem of finding the longest subsequence common to all sequences in a set of sequences (often just two sequences). MLCS is a metric based on the LCS problem proposed by Bakkelund in 2009. MLCS as a metric have the properties of Positive Definiteness, Symmetry, and Triangle Inequality.

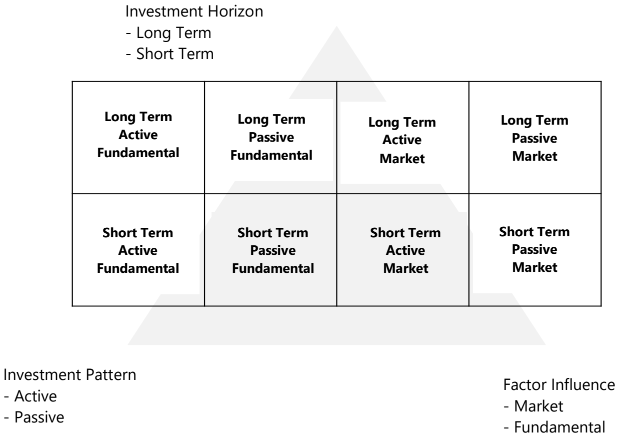
Grouping of Foreign Investors

Composite Index

The classification used previously divides investors into two groups, namely long term and short term investor group. The groupings are based on three different dimensions, namely: **investment horizon**, **transaction frequency**, and **transaction volume** (Indawan, Fitriani, Permata, & Karlina, 2013). In its use, foreign investors with short term classification are identified with investors who have a short investment horizon, high transaction frequency, high transaction volume, and tend to be influenced by market movement indicators (Hoffmann, Shefrin, & Pennings, 2010). Meanwhile, foreign investors with the long term classification are identified with investors who have a long investment horizon, low transaction frequency, low transaction volume, and tend to be influenced by economic fundamental indicators.

Foreign Investor Classification

Figure 1



Long Term Active Fundamental	**Long Term Passive Fundamental**	**Long Term Active Market**	**Long Term Passive Market**
Short Term Active Fundamental	**Short Term Passive Fundamental**	**Short Term Active Market**	**Short Term Passive Market**

Investment Horizon

Classification of investors based on the investment horizon is done by using the transaction ratio calculation approach derived from the calculations of Lakonishok, Shleifer, & Vishny (1992).

First Equation: Transaction Ratio

$$Transaction\ Ratio(t) = \frac{|Rpbuy(t) - Rpsell(t)|}{Rpbuy\ (t) + Rpsell\ (t)}$$

Where:

Rpbuy(t) = Buy Value (Rp) SBN on t period
Rpsell(t) = Sell Value (Rp) SBN on t period

A high ratio (close to 1.00 ratio) can explain that investors have a tendency to carry out all or most of their transactions to increase (net buy) or reduce (net sell) ownership of a financial asset in period (t). Conversely, a low ratio (close to the 0.00 ratio) can explain that transactions made by investors have little or no effect on changes in ownership of a financial asset in period (t) (buy for resale and vice versa).

Threshold First Equation:

1 – Transaction Ratio > 0,5: Short Term Investor
1 – Transaction Ratio < 0,5: Long Term Investor

The problem with this calculation is that if the investor does not make any transactions during the observation period, the number is zero. Zero transaction ratio number within the threshold will be categorized as Short Term Investor [1 – Transaction Ratio (0)] = 1. In the future, it is necessary to revisit the equation and threshold to determine the classification of investors based on the investment horizon.

Transaction Frequency:

In addition to the investment horizon, investor classification is also determined from the frequency of transactions which is calculated using the average transaction day per year with the following formula.

Second Equation: Average Transaction Day Per Year

$$Average\ transaction\ day\ per\ year = \frac{\Sigma_{i=1}^{n}\ transaction\ day\ (t)}{\Sigma_{i=1}^{n} year\ (t)}$$

Where:
transaction day (t) = transaction day on t period
year (t) = sum of year on t period

Threshold Second Equation:
Average transaction day per year > 0,2: Short Term Investor
Average transaction day per year < 0,2: Long Term Investor

The threshold assumption of 0.2 from equation 2 is taken from the calculation of 48 days/240 days. A 48-day approach is assumed with 4 primary SUN auctions per month and 12 months per year. Investors are expected to adjust their SUN portfolio at least 48 days for 240 working days. If the number of transactions is more than 48 days during a year, it is assumed that the investor is a short term investor. The problem with this calculation is the assumption of 48 times in 1 year based on professional judgment.

Transaction Volume:

In this study, the need for investor portfolio rebalancing is also added with the following formula.

Third Equation: Transaction Volume

$$Transaction\ Volume = \frac{\Sigma_{i=1}^{n} \{(Rp\ Buy(t) + Rp\ Sell(t))/Position(t)\}}{n}$$

Where:
Rpbuy (t) = Buy Value (Rp) SBN on t period
Rpsell (t) = Sell Value (Rp) SBN on t period
Position (t) = Ownership of SBN Position on t period
n = number of days

Threshold Third Equation:
Transaction Volume > 0,05: Short Term Investor
Transaction Volume < 0,05: Long Term Investor

The calculation assumption of equation 3 is the assumption of investors' need for asset rebalancing. The asset rebalancing threshold is 5% of the total portfolio. If the threshold is > 5%, it is assumed that the investor is short term in line with the frequent rebalancing of the portfolio to the total SUN portfolio in one transaction. The assumption of 0.5% rebalancing of the total portfolio is derived from professional judgment.

Composite Indicator for Investor's Classification (Short Term and Long Term):

Fourth Equation: Composite Index

$$Investor\ Type = \left[1 - \frac{\left|\sum_{t=1}^{T} Net\ Volume_t\right|}{\sum_{t=1}^{T}(Rp\ Buy + Rp\ Sell)_t}\right] x \left[\frac{\sum_{i=1}^{n} transaction\ day\ (t)}{\sum_{i=1}^{n} year\ (t)}\right] x \left[\frac{\sum_{i=1}^{n}\{(Rp\ Buy(t) + Rp\ Sell(t))/Position(t)\}}{n}\right] x100$$

Threshold Investor Type:
Investor Type > 0,05: Short Term Investor
Investor Type < 0,05: Long Term Investor

The total composite calculation of 3 indicators (horizon, frequency, and transaction volume) uses multiplication and is not weighted[2]. From the result of this composite index, the majority of individual investor composite values are close to zero so that the classification of investors is mostly long term. In the future, it is necessary to improve the threshold and calculation method to increase the number of short-term investors.

Clustering Methodology

We have the composite index as our benchmark to classify the investors. Next, in this study, we explore the use of clustering methodology to map investor behavior that is carried out without class information and annotations on which investors are short term and long term. Clustering can be done to group based on the proximity between the data according to various characteristics of the data.

Clustering is a machine learning (unsupervised learning) method that can group data points into groups based on the similarity and proximity of the data points. Clustering is usually used to see the group structure in the data without labelling or annotating the dataset. Das (2003) uses k-means clustering to classify hedge-fund investors based on their investment strategy and style. Validation for clustering can be done by using the Silhouette Coefficient[3] calculation to calculate the intra-cluster and inter-cluster distances and also using the average transaction frequency results from both short term and long term investor groups. In this study we use K-Means Clustering.

[2] The composite calculation of three indicators with multiplication actually produces a number that gets smaller and closer to zero as the value of each indicator is zero. The composite value that is getting closer to zero will result in a long term classification

[3] Silhouette coefficient is a measure that can be used to evaluate whether clustering results are good. The Silhouette Coefficient is calculated by using the proximity between points in one cluster and its cluster members and also calculating the proximity between points in one cluster and other cluster members. A well-separated cluster is a cluster with the distance between points in its adjacent cluster of members and the distance between the outer points of the cluster of members who are far apart.

K-Means Clustering

The k-means algorithm is an iterative algorithm that partitions the dataset into exclusive clusters with a predetermined number of clusters. This algorithm separates the cluster until it succeeds in finding a number of cluster centers that separate the clusters by the distance between the farthest cluster center and the distance between the closest cluster members with the closest distance until it becomes a single cluster containing all data points.

2.2 Behavior Modelling

There are many fundamental and market factors that may influence investors in making decisions. To select the factors/variables that most influence investors in making buying/selling/hold decisions, and to avoid overfitting the model, a feature selection methodology is necessary. By using feature selection methodology, we are able to identify those attributes that best describe how investors decide to buy or sell their positions in an objective and statistically correct manner (Silva, Tabak, & Ferreira, 2019).

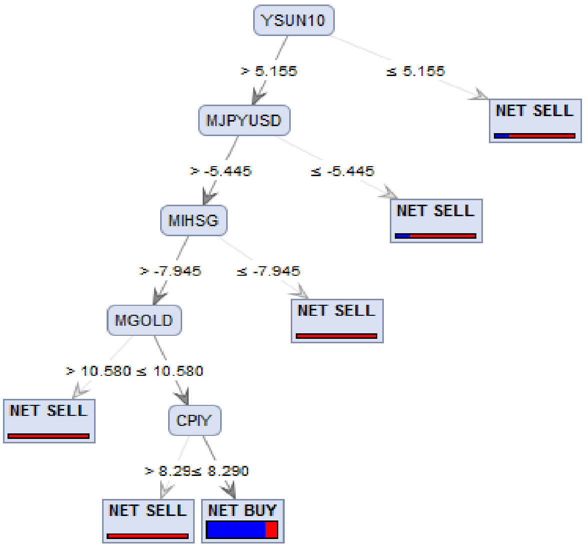
Our feature selection process gives us an objective way of identifying important variables that should be accounted to be put on forecasting model for flows projection. Every investor will have different important variables. With those important variables is the foundation for us to build model for flow projection.

Decision Tree

In this study, we use data-driven machine learning methods to do feature selection. We use decision tree algorithms for this part of the methodology. One of the great features of decision tree algorithms is that they inherently estimate a suitability of features for separation of objects representing different classes (Grabczewski, & Jankowski, 2005). The Decision Tree algorithm will sort the variables based on the largest information gain and will eliminate variables that have no effect on the investor decision (information gain is close to 0). We use 119 variables that consist of market and fundamental variables as an input for this process.

Decision Trees Illustration

Figure 2



We also apply Random forest and XGBoost algorithm as more advanced algorithms of decision trees algorithms that have feature importance calculation that we can use to measure the variable importance.

Random Forest

Random forest is a machine learning technique that can be used to solve regression and classification problems. Due to the random exploration of features, Random Forest lends itself to feature selection well and the measure of feature importance adopted here is the average information gain achieved during forest construction (Rogers, & Gunn, 2006). It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision.

XGBoost

XGBoost is a scalable tree boosting system that is widely used by data scientists and provides state-of-the-art results on many problems (Chen & Guestrin, 2016) .XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework that has some notable features:

- Clever penalization of trees
- A proportional shrinking of leaf nodes
- Newton Boosting
- Extra randomization parameter
- Implementation on single, distributed systems and out-of-core computation
- Automatic Feature selection

2.3 Flows Projection

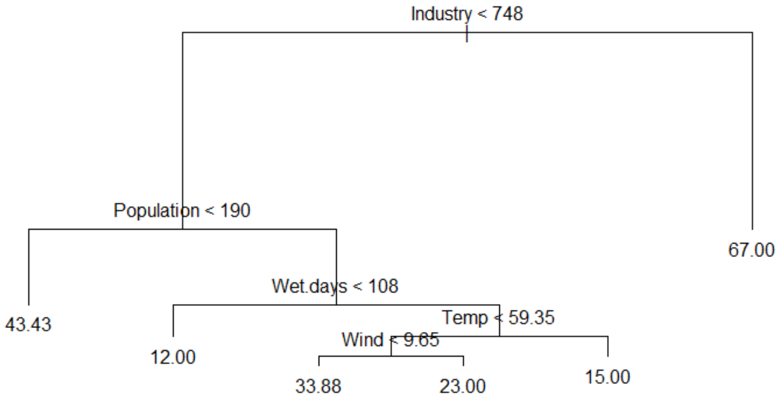
To build model for flow projections, we use Machine Learning and Big Data Analytics techniques can be used to generate capital flows projections with better precision. Here are some algorithms that can be used to perform time series projections:

Regression Tree

A regression tree is similarly a tree-structured solution in which a constant or a relatively simple regression model is fitted to the data in each partition (Loh, 2014). Regression tree is a class of decision tree algorithms that make decision trees made from observations of various variables with the final result in the form of predictive continuous numbers. Regression tree is created through a binary recursive partitioning process which divides data iteratively into partitions and branches based on the value of the existing data set.

Regression Trees Illustration

Figure 3

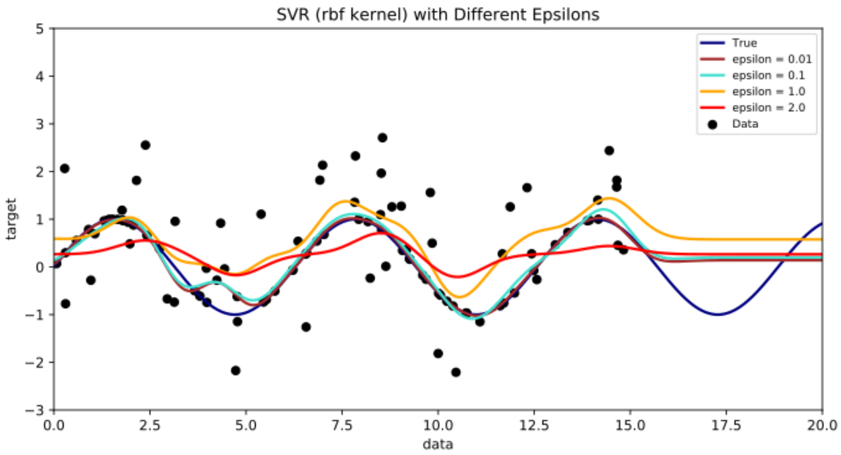


Support Vector Regression

Support vector regression is a variation of the support vector machine algorithm that creates a hyperplane function that can classify data in the function scope. Support-vector network implements the following idea: it maps the input vectors into some high dimensional feature space Z through some non-linear mapping chosen a priori (Vapnik & Cortes, 1995). Support vector regression can produce functions with non-linear constraints. Support Vector Regression also uses the parameter as a threshold value of how far the prediction result is from the original predicted value. The result of the support vector regression is in the form of a hyperplane function that gives a range where data values can be in the next data period and data projections for the next period.

Support Vector Regression Illustration

Figure 4

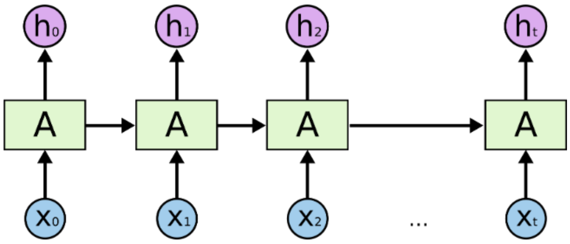


Long-Short Term Memory (LSTM)

LSTM is part of the Deep Learning algorithm class that can perform pattern and sequence recognition well. LSTM is a novel recurrent network architecture in conjunction with an appropriate gradient based learning algorithm (Hochreiter & Schmidhuber, 1997). LSTM can be used to study existing patterns in data, including patterns of changes in a value in time series data such as capital flows, stock prices, and others. LSTM as part of the Deep Learning algorithm contains activation functions that can recognize patterns from data points with high accuracy even though it requires large amounts of data and high computational complexity. The LSTM artificial neural network as illustrated below is given input in the form of time series values at a time, then learns the pattern of values based on the values in the previous sequences of time.

LSTM Illustration

Figure 5



2.4 Model Interpretation

Machine learning models are often considered as a black box. To understand the complexity of machine learning models, we need to apply model interpretability methodology to verify whether the model is in line with what our goal is.

In this study, we will use Local Interpretable Model-agnostic Explanations (LIME) as a method for interpreting machine learning models by using a simple model approach at a point of observation. Instead of training a global surrogate model, LIME focuses on training local surrogate models to explain individual predictions (Ribeiro, 2016). LIME learns what happens to predictions when given variations of data into a machine learning model. Then LIME generates a new dataset consisting of the correct and error samples according to the machine learning model results. On this new dataset, LIME then trains an interpretable model, which is weighted based on the proximity of the sample instance to the desired instance.

3. Methodology

3.1 Data

In this study, the daily net value of buy and sell transactions that indicate foreign investors' decision in the government bond market is used as the target variable. Specifically, the net transaction is flagged as “net buy” if it is positive (>0), “net sell” if it is negative (<0), and “hold” if it is equal to zero.

This study considers several market indicators collected from foreign exchange market, money market, bond market, commodity market, and stock market.

Market and Fundamental Indicators Example

Table 1

FX Market	Money Market	Bond Market	Commodity Market	Domestic Fundamental
1. Indonesia Rupiah Exchange Rate	1. Overnight Interbank Rate	1. Yield SUN Secondary Market	1. Oil WTI	1. BI Policy Rate
2. NDF 1 Month Exchange Rate	2. LIBOR USD	2. IDMA	2. Gold Spot Price	2. GDP
3. USD Index	3. LIBOR-OIS USD 1M Spread	3. HSBC Asia Local Bond Index	Stock Market	3. Inflation Rate
4. EURUSD	4. Bloomberg Financial Condition Index US	4. HSBC Asia Dollar Bond Index	1. DJI Index	4. Trade Balance
5. USDJPY	5. Bloomberg Financial Condition Index EU	5. JP Morgan Indonesia Total Bond Return Index	2. Stoxx 600 Index	5. CA Balance
6. ADXY	6. Bloomberg Financial Condition Index Asia	6. Yield UST-SUN 2Y	3. Nikkei Index	6. Fiscal Budget
7. REER		7. Yield Spread Bund PIIS	4. MSCI Asia Index	7. Foreign Reserve
		8. CDS PIIS	5. MSCI EM Index	Global Fundamental
		9. Yield Spread 10Y-2Y	6. IHSG	1. FFR
		10. CDS 5Y Indonesia		2. ECB Rate
				3. BOJ Rate

Moreover, the domestic and global fundamental indicators such as central bank policy rate, GDP, inflation rate, etc., also used to comprehend the effect of macroeconomic factors on investment decisions. The data used in this research was obtained from Bank Indonesia – Scripless Securities Settlement System (BI-SSSS) and Bloomberg from January 2016 to May 2020. BI-SSSS is a database that serves as both securities depository in the form of electronic registry and provider of securities settlement services and is directly connected between participants, operators, and the Bank Indonesia - Real Time Gross Settlement System (BI-RTGS System).

The data are split into two: training dataset (January 2016 - December 2019) and testing dataset (January - May 2020). In summary, this study uses 119 features, including 108 market indicators and 11 fundamental indicators as the independent variable.

We filter the transactions to include only the investor of interests which are:

1. Top 30 investors based on their ownership of government bonds in the market.
2. 5 investors in Indonesia investment focus group.

These investors accounted for 65% of the foreign investor's ownership in Indonesia government bonds market in December 2019.

We apply lag to the market and fundamental indicators from 1 day to 14 previous business days to factor the influence of previous day's indicators to the transaction settlement that captured in our database. We also feature engineer the variables value to get the periodical changes of the variables. The periodical changes that we feature engineered are as follows:

- Year to date (ytd)
- Quarter to date (qtd)
- Month to date (qtd)
- Daily changes (dΔ).

3.2 Entity Resolution

There is an issue where single investor entity can have multiple Single Identification Identifier (SID) in the database. Therefore, we need to group SID with similar names as single entity.

SID Investor Names Examples

Table 2

SID Investor Name	Investor Name
CSTDBK1 INVESTOR-A BOND FUND	INVESTOR-A
CSTDBK2 INVESTOR-A STRTEGY PLUS	INVESTOR-A
CSTDBK3 INVESTOR-A GL-MUL BND FUND	INVESTOR-A
CSTDBK2 INVESTOR-B LCL DEBT INDEX PRTF	INVESTOR-B
CSTDBK3 INVESTOR-B SBP	INVESTOR-B
INVESTOR-B GLBL ALLOC FND	INVESTOR-B

We develop entity resolution models using string similarity metrics to group investor names from multiple SIDs into a single entity. First, we study the pattern of names from a sample of the investor names in the database. Then we separate the raw names into the custodian bank names, investor names, and other texts. Then we label the raw investor names to the investor real names and produce a small database of investor names labelled with their real names. Finally, we do a horse race of 4 string similarity metrics to match the raw investor names to labelled investor names that we have in the database and use the best model as the entity resolution model to group similar SID names as single investor entities. The string similarity metrics that we experiment with are as follows:

1. Jaro-Winkler Distance
2. Normalized Levenshtein Distance
3. Weighted LevenShtein Distance
4. Metric Longest Common Subsequence (MLCS)

3.3 Grouping of Foreign Investors

One of our goals is to create a grouping of investors into long term (LT) investor or short term (ST) investor based on its yearly activities and behavior in Indonesia's government bonds market. Using the components of the composite index mentioned in 2.1.1 which are Investment Horizon, Transaction Frequency and Transaction Volume, then using the entity resolution result, and K-means clustering, we create clusters of investors based on their behavior and activities in the government bond market. We tried 2-6 number of clusters in our clustering methodology. Then, we calculate silhouette coefficients to measure the clusters quality with different number of clusters.

3.4 Behavior Modelling

For each individual investor we use decision tree algorithms feature importance to filter the important variables to be used for the investor decision model and projection flows model. This is done to avoid overfitting, minimize noises and redundant data, and improve the performance of the investor decision and flows projection model. We use decision dree algorithms feature importance which are Decision Tree, Random Forest and XGBoost models and information gain measures to calculate the important features.

We experiment with inherent lags in the model from 1 to 5 days to accommodate possible delays that may occurred from the time the investor get their information to the time of the transaction settled in the database. We evaluate the model using F1 scores with the variation of inherent lags and algorithms for each of the individual investors. We didn't use the model for prediction, but to help decide which of the lags and variables that produced the best F1 score. The lags and variables produced by the best model then used again as one of the input for investor decision prediction and flows projection model that used wider varieties of algorithms, pre-processing, and experiment.

3.5 Investor Decision Prediction and Flows Projection

We develop investor decision prediction model using classification machine learning algorithms to predict daily individual investor decision of whether the investor will have a net buy, net sell, or hold decision in the corresponding day. The algorithms that we use are Logistic Regression, SVM, KNN, Decision Tree, Random Forest, XGBoost, and LSTM. Before modelling we apply pre-processing techniques such as PCA, lag adjustment, and variables adjustment to see which option produced the best result from using all of the variables, only the important variables or using PCA transformed variables. The model then evaluated using F1 scores. The model that produces the best F1 scores for each individual investors then used to help the flows projection model.

The flows projection model experiment design is similar to the investor decision model with different algorithms and different dependent variable. We use regression machine learning models such as Logistic Regression, KNN, Regression Tree, SVR, LSTM, and XGBoost. In the flows projection model, we use daily transaction nominal capital flows of each investor as the dependent variable. We evaluate the flows projection model using Mean Average Error (MAE) regression error metric.

We use the investor decision model to help the flows projection model. If the investor decision model prediction is a hold, then the flows projection model will calculate the flows as 0 for the day. If it's net buy then the flows projection model will project positive values, and vice versa for net sell decision.

3.6 Model Interpretation

In order to conduct further analysis, it is very important to interpret the model to find out what influences investors in making decisions. However, the machine learning models are mostly black box models which tend to be either difficult or impossible to interpret.

We use LIME to interpret our machine learning model decision. We apply LIME on several random date on the testing period to see what variables affect the model to predict the direction of investor decisions. We chose top 10 (ten) most influential (highest weight) variables that drive the prediction of the model. We apply LIME to all the best models for each investor.

4. Result & Analysis

4.1 Entity Resolution

The models that have been trained in the previous steps need to be evaluated in order to measure their accuracy in predicting each target class. We use F1-score as the metric for entity resolution and prediction model evaluation, in order to get a balanced classification model with the optimal balance of recall and precision.

The results of evaluation for best entity resolution models using 4 string similarity metrics and a threshold from 0.90 to 1.00 shown in Table 2. The best model is obtained using the Jaro-Winkler metrics with threshold of 0.97, which produce F1 score of 87%. We also evaluate using 500 randomly taken out of sample data, and produced good accuracy results of 89%. Therefore, we believe that the model is robust enough to be implemented in all data in January 2014 - May 2020 period. From 4.215 unique SIDs and unique investor names in the Indonesia Government bond transaction data on period January 2014 – December 2020, we get 1.846 unique investors from the entity resolution results.

String Similarity Metrics Evaluation

Table 3

String similarity Metrics	Threshold	F1
Jaro Winkler	0.97	**87%**
Normalized Leveinshtein	0.96	79%
Weighted Leveinshtein	0.96	82%
MLCS	0.85	73%

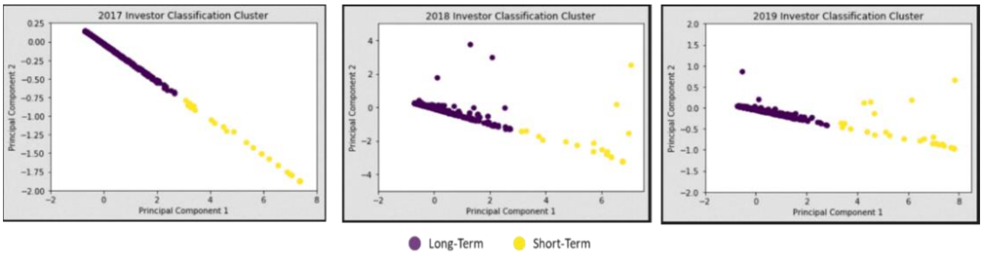
Note: Blue-shaded cells denote the best result for each model

4.2 Grouping of Foreign Investors

The clustering model is able to group investors well into short term (ST) investors and long term (LT) investors, with a Silhouette Coefficient of 0.89. These results are also in line with the grouping of investors with the expert judgement done by the Monetary Management Department using the Composite Index. Clustering result that we use as the grouping is the 2019 investor transactions data, with 1.075 investors grouped as long term investors and 35 investors grouped as short term investors.

Grouping of Foreign Investor Cluster Results

Figure 6



4.3 Behavior Modelling

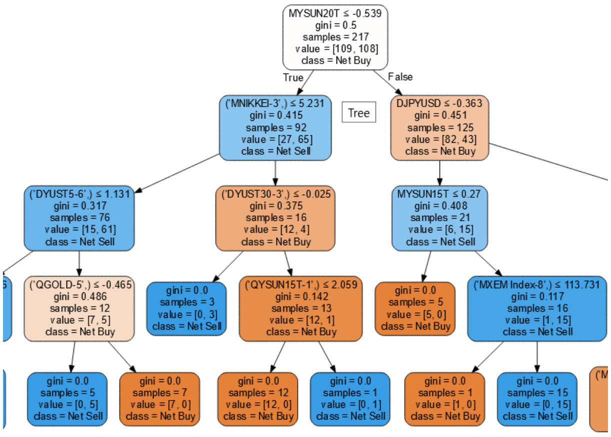
We use behavior analysis models to find the indicators that influence the decision of each investor of interests, and each the investor groups (short term and long term). Our findings are as follows:

1. We find that for both of the investors group, Indonesian Bonds Yields in different maturity are considered important with short term investors group considers shorter maturity of government bonds.
2. JAKCONS which is Jakarta Consumer Goods stock index is important for both of the investors group.
3. Short term investors group are affected by more daily and high frequency indices while long term investors group are highly affected by a fundamental indicator (Indonesia YoY Core Inflation).

Furthermore, the behavior model result will be used as selected important features for decision prediction and flows projection model.

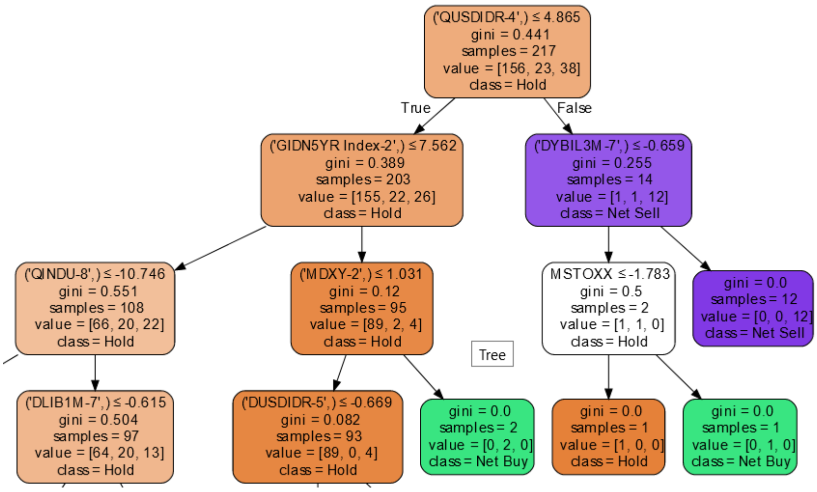
Behavior Model Single Decision Tree Short Term Investor Example

Figure 7



Behavior Model Single Decision Tree Longer Term Investor Example

Figure 8



Behavior Model Top 10 Important Variables for Investors Group

Table 4

Short Term Investors		Long Term Investors	
Ticker	Description	Ticker	Description
IDR Currency	Rupiah Currency	JASFXBAT Index	Indonesia Stock Capital Flows
HSITR Index	HSBC Asia Dollar Bond Index	FSSTI Index	Straits Times Index
SENSEX Index	S&P Bombay Stock Exchange	MGIY10Y Index	Malaysian 10-year Government Bond Yield
USGGBE10 Index	US Breakeven 10 Y	INR Currency	Indian Rupee Currency
GIDN12YR Index	Gov. Bonds Generic Yield 12 Years	GIDN30YR Index	Yield SUN Generic 30 Year
JAKINFR Index	Jakarta Infrastructure, Utilization and Transportation Stock Index	JAKCONS Index	Jakarta Consumer Goods Stock Index
IDPPON Index	PUAB o/n	IDIFCRIY index	Indonesia Yoy Core Inflation
SXXP Index	STOXX Europe 600 Index	IHNI1M Currency	Implied NDF 1M
JAKCONS Index	Jakarta Consumer Goods Stock Index	IBPRTRI Index	IBPA Government Bond Index
MXEM Index	MSCI Emerging Market Index	THB Currency	Thailand Baht Currency

Behavior Model Top 10 Important Variables for Individual Investors Example

Table 5

Example Short Term Investor		Example Long Term Investor	
Ticker	Description	Ticker	Description
EPUCNUSD Index	US Economic Policy Uncertainty Index	GIDN1YR Index	Generic Gov. Bonds Yield 5 Year
GIDN1YR Index	Generic IDN Gov. Bonds Yield 1 Year	IHN+1M Index	IDR NDF 1 Month
ADXY Index	Asian Dollar Index	USGG30YR Index	Generic US Treasury Yield 30 Years
GIDN15YR Index	Generic IDN Gov. Bonds Yield 15 Years	GIDN10YR Index	Generic IDN Gov. Bonds Yield 10 Years
GIDN5YR Index	Generic IDN Gov. Bonds Yield 5 Years	USGG3M Index	Generic US Treasury Bills Yield 3 Month
EURUSD Currency	Euro Currency	USSOA Index	USD Overnight Indexed Swap 1M
BFCIEU Index	Bloomberg Financial Condition Index EU	GIDN5YR Index	Generic IDN Gov. Bonds Yield 5 Years
GIDN10YR Index	Generic IDN Gov. Bonds Yield 10 Years	IDPPON Index	PUAB o/n
USGG3YR Index	Generic US Treasury Bills Yield 3 Years	JCI Index	IDX Stock Composite Index
DXY Index	Dollar Index	EURUSD Currency	Euro Currency

4.4 Investor Behavior Prediction and Projection Flows Model

In this study we develop 35 behavior prediction models for each investor. In addition, we also develop 1 model for the LT investor group and 1 model for the ST investor group. Based on the result, our investor behavior prediction models are able to predict the decision (buy, sell, or hold) made by 18 of 35 investors and the two investor groups with satisfying result (>60% F1 Score). The model for LT investors group has F1 score of 77%, while for ST investors group has F1 score of 64%. While for the individual investor prediction models, the best model has F1-score of 86%, while the lowest F1-score model get F1-score of 47%.

Investor Behavior Prediction Result

Table 6

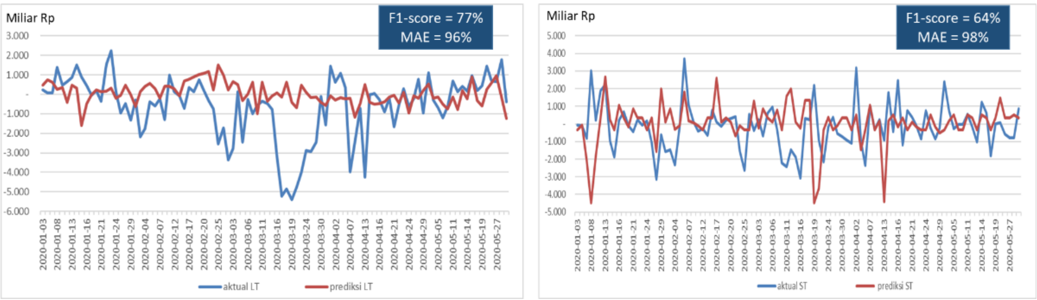
Investor	Algorithms	Historical data (days)	Using PCA	F1 Score
LT investors group	XGBoost	4	Yes	77%
ST investors group	XGBoost	6	Yes	64%
Best individual investor	Regression Tree	2	Yes	86%
Lowest individual investor	XGBoost	9	Yes	47%

Note: The model is each individual investor model with the highest f1 score.

However, the flows projection model's ability to predict the amount of flows still needs to be improved, considering the error (mean absolute error/MAE) is still quite large. The average MAE for the 37 investors is 93% of the average investor transaction, with the smallest MAE at 50% and the largest MAE being 101% of the average investor transaction.

Flows Projection of Long Term and Short Term Result Plot

Figure 9

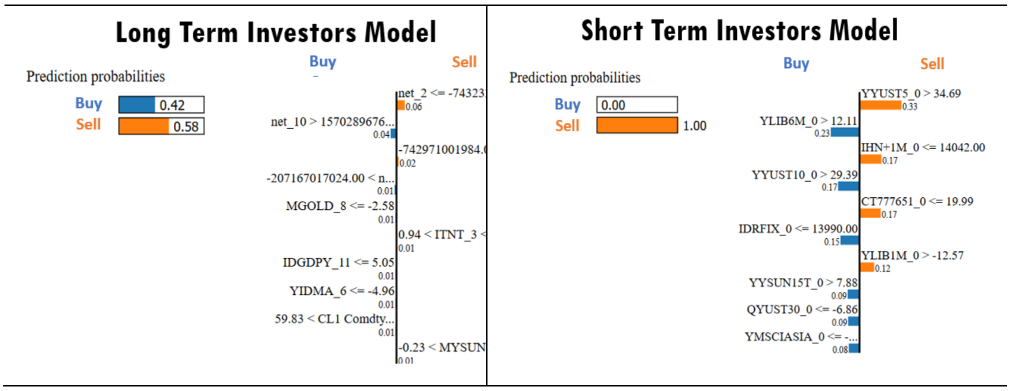


4.5 Model Interpretation

Based on LIME result on 3rd January 2020, we found out that the prediction model of long-term (LT) investors on that day is more influenced by the decision (net buy) of the investors on previous period. While for the prediction model of short-term (ST) investors, we found out that on the same day the model is more influenced by market variables, such as US Treasury yields and USD LIBOR.

LIME Top 10 Important Variables for Investors Group

Figure 10



5. Conclusion & Future Work

5.1 Conclusion

Firstly, to deal with the issue of the same investor entities that has many different SIDs in the database of government bond transaction that we have, we develop a methodology using string similarity metrics to match similar investor names as single entities. The result is able to match random out-of-sample investor names very well with accuracy of 89%.

Then, we develop clustering model to complete the analysis of the grouping foreign investors into real money investors (long term / LT) or traders (short term / ST), based on their portfolio management behavior with a data driven approach using machine learning. The cluster result has high silhouette coefficient in grouping the investors with similar activities in the government bonds market and matched the grouping using composite index done by Monetary Management Department.

Lastly, we develop prediction model for each investor decisions using market and fundamental data, namely 119 variables and their feature engineered and lag adjusted form with a total of approximately 2.000 variables. We use decision tree algorithms to do feature selection and filter the most influential variables for each individual investor. We then develop investor decision (buy, hold, or sell) prediction model and flows projection regression model using machine learning algorithms. The results are individual investor prediction decision models with 18 investors and 2 investor group's prediction decision models that are able to produce satisfying prediction power (>60% F1 Score). As for the flows projection model, the result is not yet good and still need to be improved.

5.2 Future Work

There are several improvements in the methodology that can be applied for future works.

- Improve the accuracy of the investor decision and flows projection models for all of the individual investors

For some investors, the prediction model that we have is not accurate enough, so it needs to be improved. Improvements can be done by adding longer data (perhaps from 2013) as well as adding other market or fundamentals variables that are considered influential in determining investor decisions.

Furthermore, we can try to do 2-stage classification, considering that for some investors the frequency of holding is much higher than the frequency of buying or selling. The first stage is to predict whether investors will hold or not hold. Furthermore, if the prediction results is not “hold”: then stage 2 predictions will predict whether investors will net buy/net sell.

- Develop investor investment prediction and flows projection model that can predict well during abnormal flows period (COVID 19 Pandemic)

During the pandemic period, around 2020 - 2021, the pattern of transactions in the government bonds market, both nominally and transaction frequency, is drastically different from the previous period. Investor behavior was also presumably different from the previous periods. Therefore, we need to include data from COVID-

19 pandemic period into our training data. As an alternative, it also worth to try to separate model in the abnormal period and the normal period. However, this could be quite challenging since the data for the abnormal COVID-19 period is limited to only 1 year of data.

- Develop model automation and dashboard for daily visualization of government bonds daily data and prediction

To use the prediction model to predict the foreign investors' decisions on daily basis, it is necessary to automate the process and disseminate the prediction results. We can try to automate the prediction process of the models and visualize it in a dashboard so that it can be used to analyze foreign investors' behavior in government bond market to support decision making related to monetary operations strategy in timelier manner.

- Develop model for foreign investors in stock and currency market

To complete the analysis of capital flows in Indonesia, the foreign investor behavior prediction models using machine learning can be applied to foreign investor in stock and currency market.

References

Agung, J., & Darsono. (2012). Post-Global Crisis Capital Inflows to Indonesia: Challenges and Policy Responses. SEACEN.

Andritzky, J. R. (2012). Government Bonds and Their Investors: What Are the Facts and Do They Matter? *IMF Working Paper WP/12/158*.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). Das, N. (2003). Hedge Fund Classification using K-means Clustering Method. *Computing in Economics and Finance 2003, 284*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. Hoffmann, A. O. I., Shefrin, H., & Pennings, J. M. E. (2010). Behavioral Portfolio Analysis of Individual Investors. *SSRN Electronic Journal*.

Indawan, F., Fitriani, S., Permata, M. I., & Karlina, I. (2013). Capital Flows in Indonesia: The Behavior, The Role, and Its Optimality Uses for The Economy. *Bulletin of Monetary, Economics and Banking*, 23-54.

Lakonishok, J., Shleifer, A., & Vishny, R. W. (1992). The Impact of Institutional Trading on Stock Prices. *Journal of Financial Economics, 32*, 23-43.

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, No. 8, pp. 707-710).

Loh W.-Y. (2014). Classification and Regression Tree Methods. Wiley StatsRef: Statistics Reference Online.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).

Rogers, J., & Gunn, S. (2005, February). Identifying feature relevance using a random forest. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection*" (pp. 173-184). Springer, Berlin, Heidelberg. Silva, T. C., Tabak, B. M., & Ferreira, I. M. (2019). Modeling Investor Behavior Using Machine Learning: Mean-Reversion and Momentum Trading Strategies. *Complexity*, 1-14.

BANK INDONESIA

PREDICTING FOREIGN INVESTORS' BEHAVIOR AND FLOWS PROJECTION IN INDONESIA GOVERNMENT BONDS MARKET USING MACHINE LEARNING

Anggraini Widjanarti, Muhammad Abdul Jabbar, Arinda Dwi Okfantia

Statistics Department – Bank Indonesia

Email: anggraini_widjanarti@bi.go.id, muhammad_abdul@bi.go.id, arinda_dwi@bi.go.id

The view expressed here are those of the authors and do not necessarily reflect the view of Bank Indonesia

OUTLINE

BACKGROUND AND GOALS

FRAMEWORK

METHODOLOGY

RESULTS AND ANALYSIS

CONCLUSION AND FUTURE WORKS

BACKGROUND AND GOALS

Analytical needs to **monitor individual foreign investors activities** in the government bonds market that potentially create currency volatility.

Increasing foreign investor ownership in Indonesia government bonds market

Utilization of data sources that BI has e.g Government Bonds transactions, fundamental and market indicators that can be used to predict foreign investor behavior

“**Predict foreign investor behavior on the Government Bonds Market by using various scenarios of macroeconomic and *market indicators* and machine learning methods that can produce a good level of accuracy.** “

FRAMEWORK

This study uses Machine Learning and Text Mining to predict the capital flows of foreign investors in the government bonds market. We use granular government bonds transactions data, fundamental and financial market indicators. The results are foreign investor behavior clusters, important variables that influence investor decision, and foreign investors flows projection.

Behaviour analysis of foreign investor in government bonds market includes:

- Classification of investor group cluster(*short term/long term*),
- Identification of important variables that influence individual investor decision
- Foreign investor decision prediction and flows projection.

REFERENCES:

- Hoffman, A. et al (2010). Behavioral Portfolio Analysis of Individual Investors
- Bontempi, G. et al. (2009). Machine Learning Strategies for Time Series Forecasting.
- Agung, J. and Darsono. (2012). Post-Global Crisis Capital Inflows to Indonesia: Challenges and Policy Responses.
- Mody, A. et al. (2001). Modelling Fundamentals for Forecasting Capital Flows to Emerging Markets.

COLLECT

PROCESS

ANALYTICS

DISSEMINATE

DATA CAPTURING

SI - BISSSS

FOREIGN GOVERNMENT BONDS TRANSACTION SETTLEMENT

BLOOMBERG

FUNDAMENTAL ECONOMY & MARKET DATA(AUTOMATIC RETRIEVAL)

ANALYTICS

PRE-PROCESSING

A

- *Data Cleansing*
- *Entity Resolution – String Similarity*

INVESTOR TYPE MAPPING & VALIDATION

B

- *Composite Index Formula*
- *K-Means & Hierarchical Clustering*

INDIVIDUAL INVESTOR BEHAVIOR ANALYSIS

C

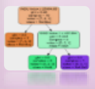
Decision Tree on market & fundamental variables

FLOWS PROJECTION, VALIDATION, & ANALYSIS

D

- *Machine Learning : SVR & Regression Tree*
- *Deep Learning : Long Short Term Memory (LSTM)*

RESULT

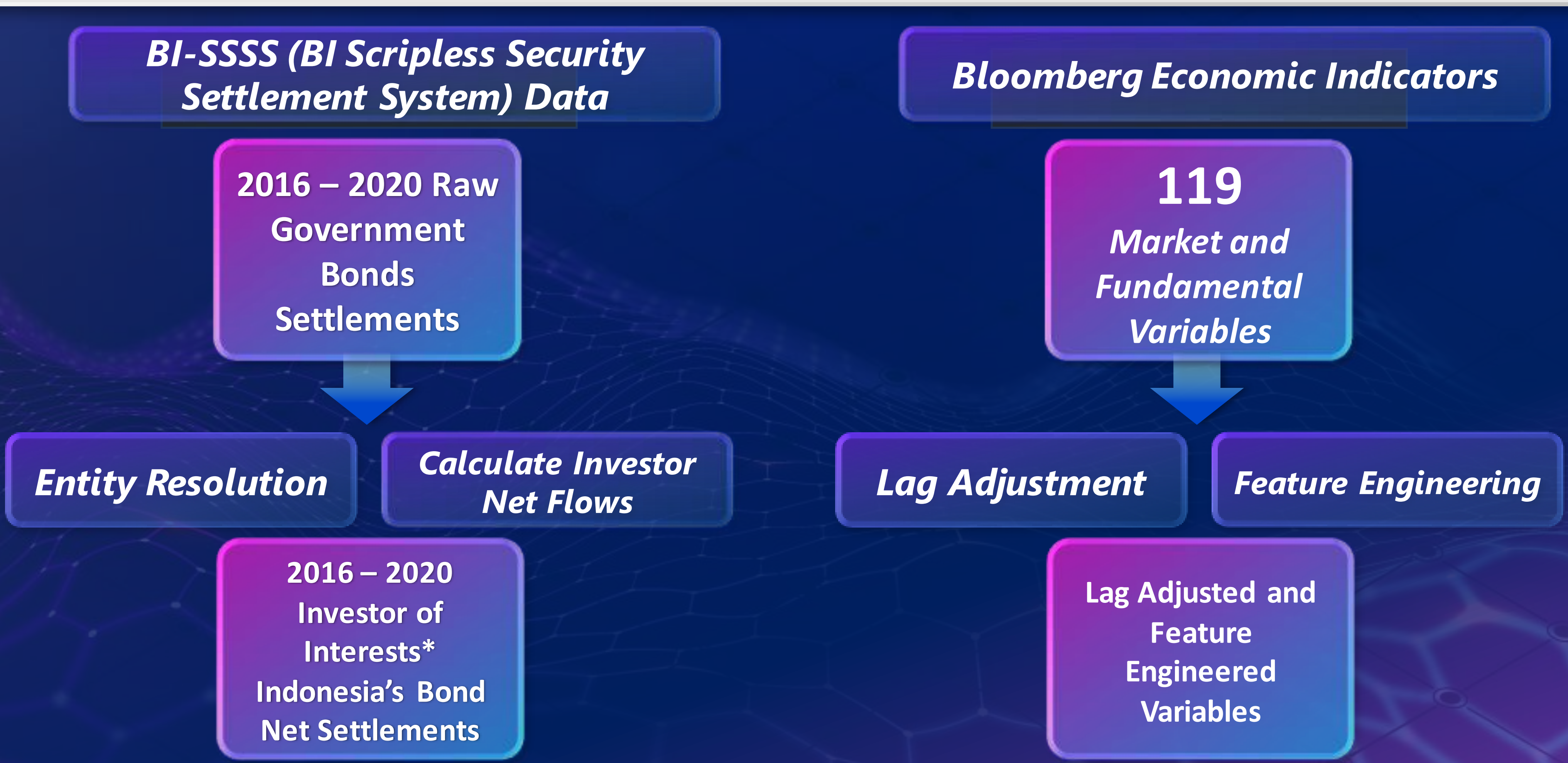


INDIVIDUAL INVESTOR BEHAVIOR & CLASSIFICATION

INVESTOR DECISION PREDICTION AND CAPITAL FLOWS PROJECTION

METHODOLOGY - DATA SOURCE AND PRE-PROCESSING

Bloomberg indicators data is pre-processed to have the lag adjusted version of the data, and feature engineered to produce dtd, mtd, qtd, and ytd changes. The SID (Single Investor Identification) are processed using text mining so that we can group each SID to its approximate investor name.



BI-SSSS (BI Scripless Security Settlement System) Data

2016 – 2020 Raw Government Bonds Settlements

Entity Resolution

Calculate Investor Net Flows

2016 – 2020 Investor of Interests* Indonesia's Bond Net Settlements

Bloomberg Economic Indicators

119 *Market and Fundamental Variables*

Lag Adjustment

Feature Engineering

Lag Adjusted and Feature Engineered Variables

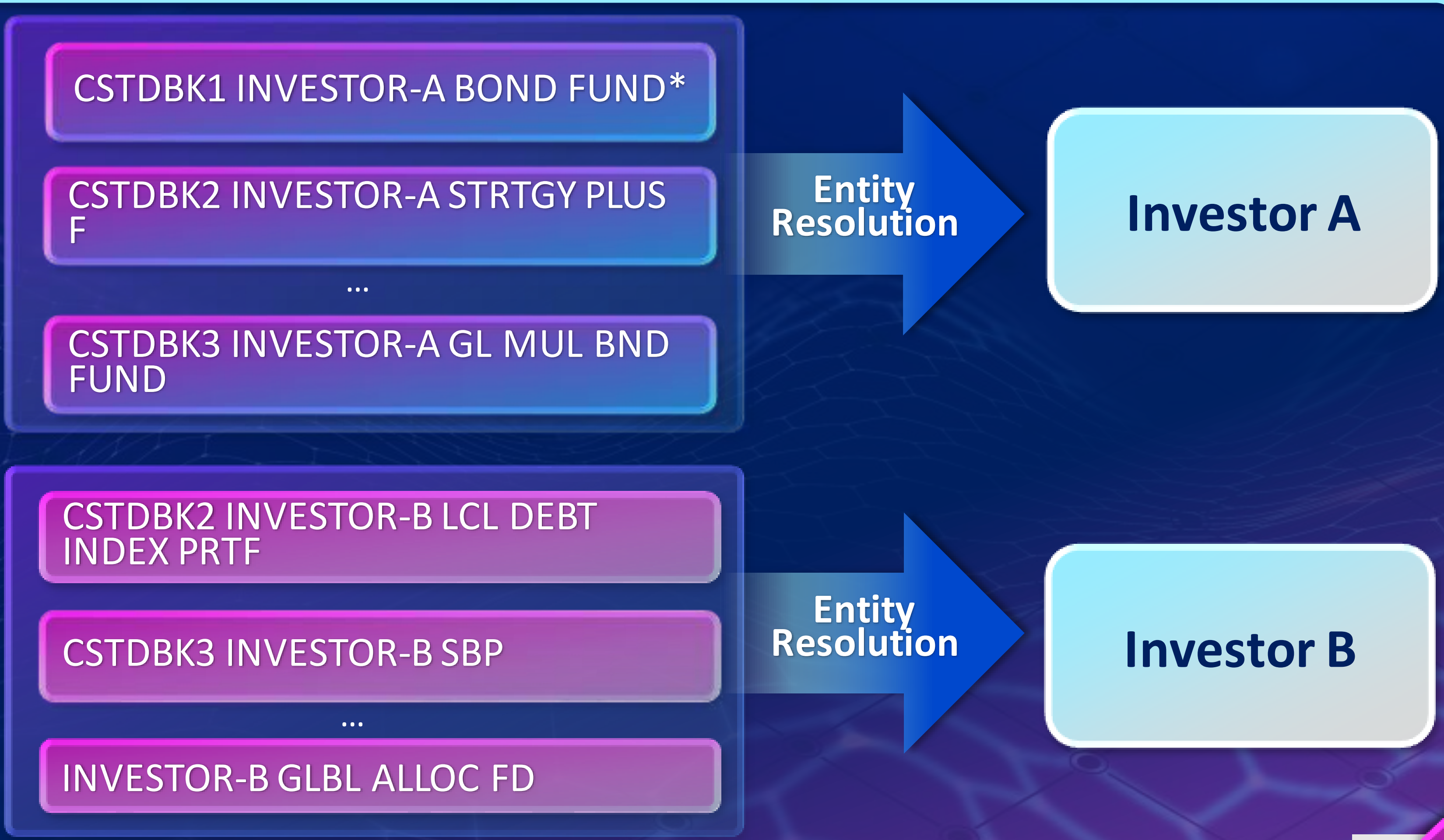
*Top 30 foreign investors of Indonesian Government Bonds + 5 foreign investment forums investors

METHODOLOGY - ENTITY RESOLUTION

Entity Resolution is done to obtain list of SID and and group foreign investors name into unique entities. Entity resolution is necessary because some investors have multiple SID in the government bonds transactional data.

Entity resolution model is developed using ***string similarity algorithm***, which is a class of algorithm that measures similarity between texts.

ALGORITHM	THRESHOLD	F1-SCORE
Jaro-Winkler	**0,97**	**0,87**
Normalized Leveinshtein	**0,96**	**0,79**
Weighted Leveinshtein	**0,96**	**0,82**
MLCS	**0,85**	**0,73**



CSTDBK1 INVESTOR-A BOND FUND*

CSTDBK2 INVESTOR-A STRTGY PLUS F

...

CSTDBK3 INVESTOR-A GL MUL BND FUND

Entity Resolution

Investor A

CSTDBK2 INVESTOR-B LCL DEBT INDEX PRTF

CSTDBK3 INVESTOR-B SBP

...

INVESTOR-B GLBL ALLOC FD

Entity Resolution

Investor B

*Names listed above are masked

METHODOLOGY - INVESTOR TYPE CLUSTERING

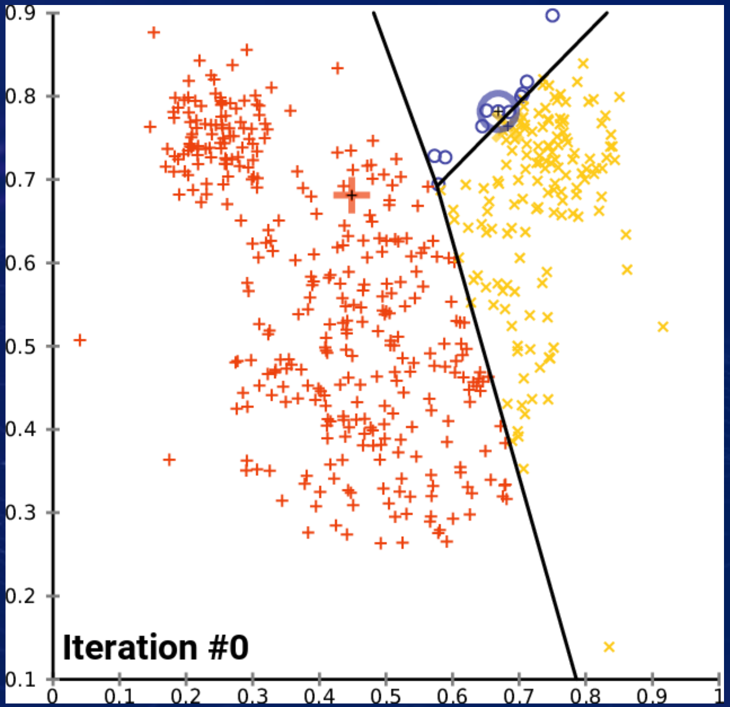
We measured *3 indicators to group investors*. Then, we use *K-means clustering* using the indicators to form foreign investor clusters as group of investors based on their yearly behavior in the government bonds market. The data-driven clustering result aligns with the expert judgement done by the Monetary Management Department and have a *high silhouette coefficient* as evaluation metrics.

Investment **Horizon**

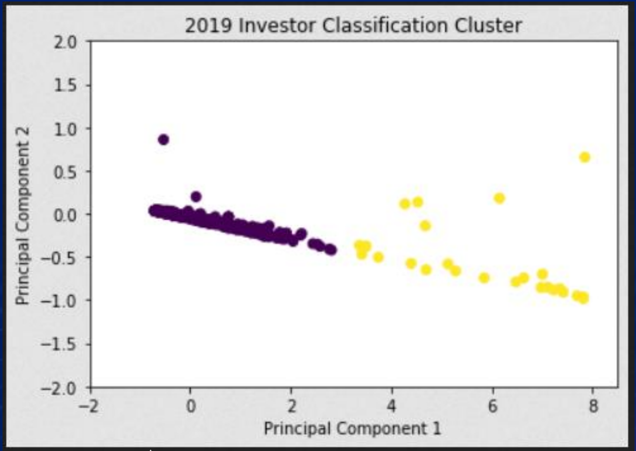
Transaction **Frequency**

Rebalancing ***Ratio***

K-Means Clustering



Cluster Result



● *Investor Long Term*

● *Investor Short Term*

Silhouette Coefficient*:

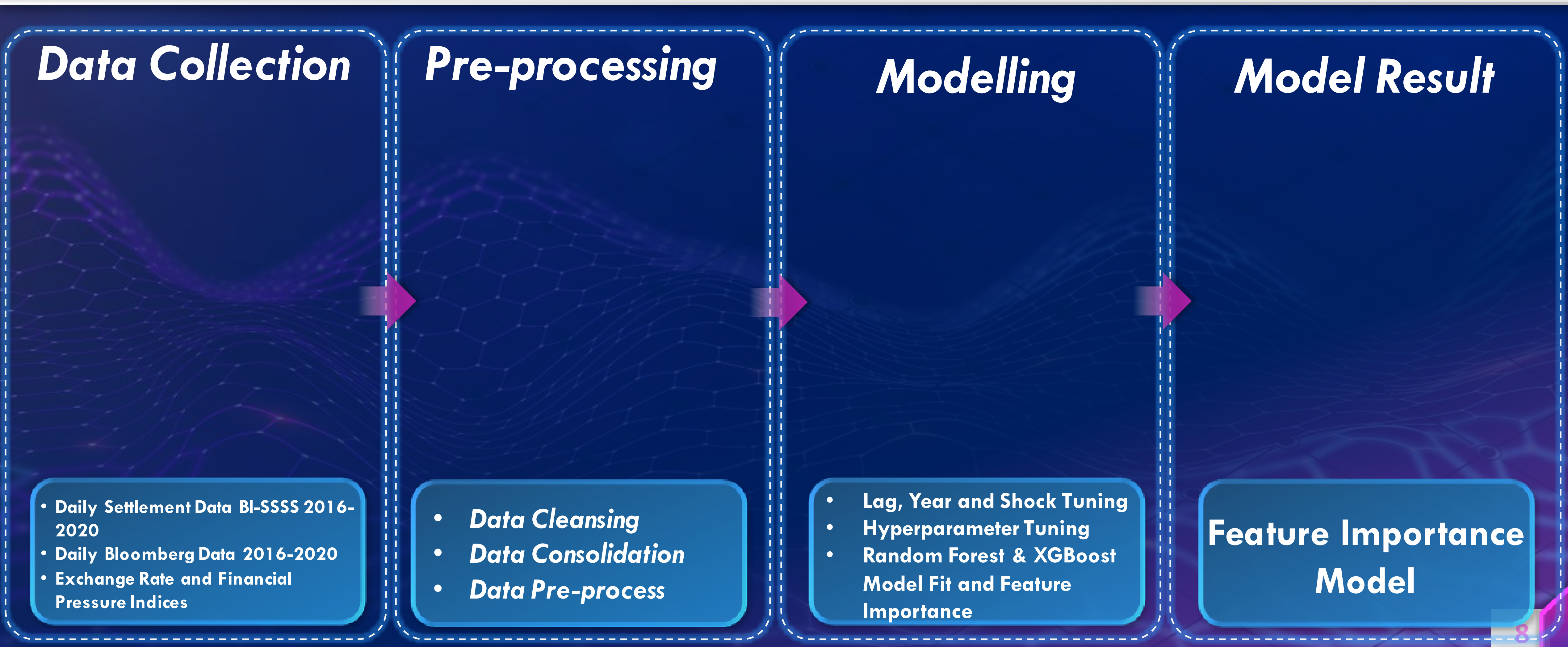
0.892

Number of investors ST/LT:

35/1075

METHODOLOGY - BEHAVIOR MODELLING

We built *Behavior Modelling* to get better understanding of the indicators that *influence the decision* of each investor of interests, and each the investor group. The behavior model result will be used as selected *important feature for decision prediction and flows projection model*.



Data Collection

- Daily Settlement Data BI-SSSS 2016-2020
- Daily Bloomberg Data 2016-2020
- Exchange Rate and Financial Pressure Indices

Pre-processing

- *Data Cleansing*
- *Data Consolidation*
- *Data Pre-process*

Modelling

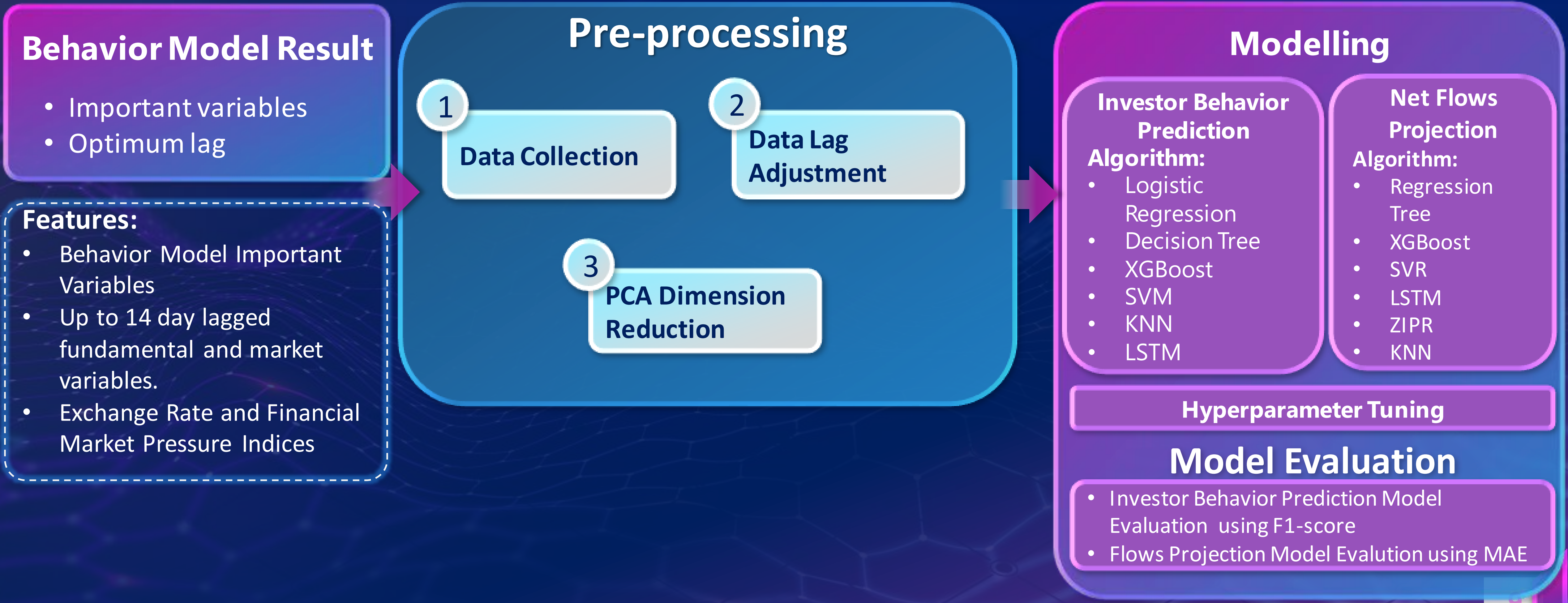
- Lag, Year and Shock Tuning
- Hyperparameter Tuning
- Random Forest & XGBoost Model Fit and Feature Importance

Model Result

Feature Importance Model

METHODOLOGY - FLOWS PROJECTION AND INVESTOR DECISION PREDICTION

Flows projection and investor behavior prediction experiment **is done by using machine learning algorithms from** *Logistic Regression to Deep Learning using LSTM*. **The model that produce the best result is used to predict the investor investment decision and project the net flows of each individual investor daily.**



Behavior Model Result

- Important variables
- Optimum lag

Features:

- Behavior Model Important Variables
- Up to 14 day lagged fundamental and market variables.
- Exchange Rate and Financial Market Pressure Indices

Pre-processing

1 Data Collection

2 Data Lag Adjustment

3 PCA Dimension Reduction

Modelling

Investor Behavior Prediction

Algorithm:

- Logistic Regression
- Decision Tree
- XGBoost
- SVM
- KNN
- LSTM

Net Flows Projection

Algorithm:

- Regression Tree
- XGBoost
- SVR
- LSTM
- ZIPR
- KNN

Hyperparameter Tuning

Model Evaluation

- Investor Behavior Prediction Model Evaluation using F1-score
- Flows Projection Model Evalution using MAE

BEHAVIOR MODELLING – LONG TERM AND SHORT TERM INVESTOR GROUP RESULT

- For both of the investors group, *Indonesian Bonds Yields in different maturity are considered important* with short term investors group considers shorter maturity of Government Bonds. JAKCONS Stock Index also important for both the groups.
- *Short term investors group* are affected by *more daily frequency indices*, while *long term investors group* are affected by a fundamental indicator which is the *Indonesia Yoy Core Inflation indicator* (IDIFCRIY Index)

Short Term Investors

Ticker	Description
IDR Currency	Rupiah Currency
HSITR Index	HSBC Asia Dollar Bond Index
SENSEX Index	S&P Bombay Stock Exchange
USSGGBE10 Index	US Breakeven 10 Y
GIDN12YR Index	Yield SUN Generic 12 Year
JAKINFR Index	Jakarta Infrastructure, Utilization and Transportation Stock Index
IDPPON Index	PUAB o/n
SXXP Index	STOXX Europe 600 Index
JAKCONS Index	Jakarta Consumer Goods Stock Index
MXEM Index	MSCI Emerging Market Index

Long Term Investors

Ticker	Description
JASXFBAT Index	Indonesia Stock Capital Flows
FSSTI Index	Straits Times Index STI
MGIY10Y Index	Malaysian 10-year Government Bond Yield
INR Currency	Indian Rupee Currency
GIDN30YR Index	Yield SUN Generic 30 Year
JAKCONS Index	Jakarta Consumer Goods Stock Index
IDIFCRIY index	Indonesia Yoy Core Inflation
IHNI1M Currency	Implied NDF 1M
IBPRTRI Index	IBPA Government Bond Index
THB Currency	Thailand Baht Currency

FLOWS PROJECTION AND INVESTOR BEHAVIOR PREDICTION – RESULT SUMMARY

The investor groups investor behavior prediction models are able to predict the decision made by 18 of 35 investor and the two investor groups with satisfying result (>60% F1 Score). As for the flows projection model the result is not good enough yet and still have to be improved in future works

Training Data
Jan 2016 – Des 2019

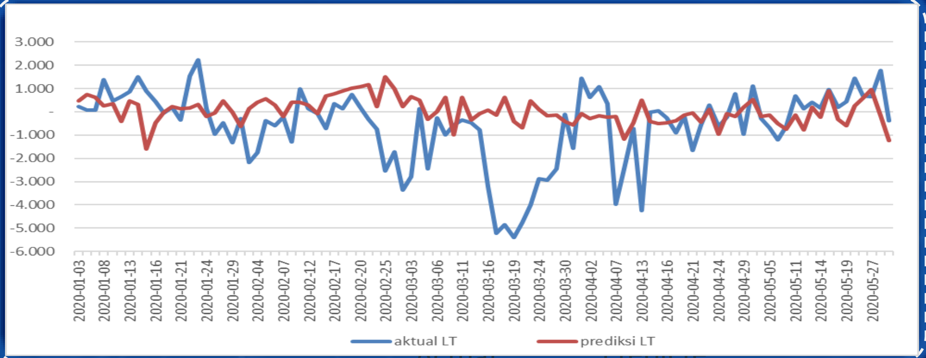
Testing Data
Jan 2020 – May 2020

Max F1-Score
86%

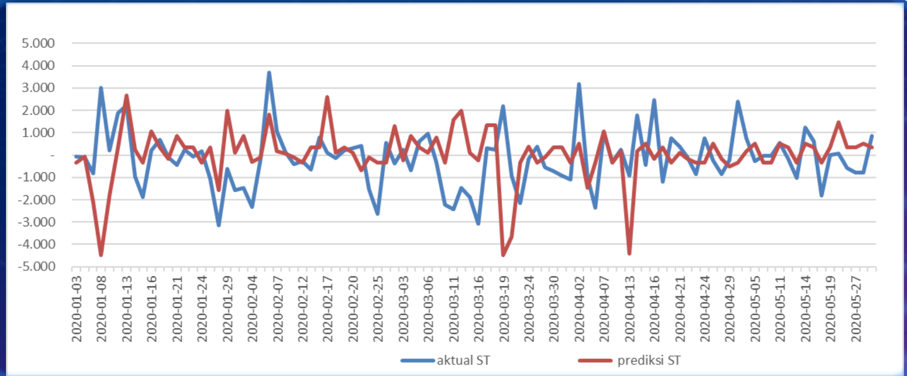
Min F1-Score
47%

Investor Model
with F1 Score > 60%

20 / 37



Long term Investor Flows Projection



Short term Investor Flows Projection

CONCLUSION AND FUTURE WORKS

CONCLUSION

With the result we are confident that machine learning methodology has been able to identify single investor based on its name similarity using string similarity method, cluster investor group using yearly behavior and predict investor decision on Government Bonds Transaction (buy, sell or hold). But there is still a lot of works to improve the prediction accuracy of the flows projection model.

FUTURE WORKS

1. *Improve the accuracy* of the investor decision and flows projection models for all of the individual investors.
2. Develop *investor investment prediction and flows projection model* that can predict well during ***abnormal flows*** period (COVID-19 Pandemic).
3. Develop *model automation* and *dashboard* for daily visualization of government bonds daily data and prediction.
4. Develop model for *foreign investors in stock and currency market*.

"The future belongs to those who prepare for it today."

Malcolm X

IFC-Bank of Italy Workshop on “Machine learning in central banking”

19-22 October 2021, Rome / virtual event

Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts[1]

Hector Carcel-Villanova,
International Monetary Fund

[1] This presentation 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 central banks and other institutions represented at the event.

Text Data Analysis using Latent Dirichlet Allocation, an Application to FOMC Transcripts[1]

Hali Edison (Williams College)

Hector Carcel-Villanova (International Monetary Fund)

Abstract

This short paper explains Latent Dirichlet Allocation (LDA), a machine learning algorithm, and uses it to analyse the content of U.S. Federal Open Market Committee (FOMC) transcripts covering the period 2003–2012, including 45,346 passages. The results of this exercise show that discussions on economic modelling were dominant during the Global Financial Crisis (GFC), with an increase in discussions on the banking system in the years following the GFC. Discussions on communication also gained relevance towards the end of the sample. LDA analysis could be further exploited by researchers at central banks and institutions to identify topic priorities in relevant documents such as the FOMC transcripts.

Keywords: FOMC, Text data analysis, Transcripts, Latent Dirichlet Allocation.

JEL classification: E52, E58, D78.

Contents

[1] This version is an adaptation from: Edison, H. and Carcel, H. (2021) Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts. *Applied Economics Letters* 28: 1, 38-42.

The views expressed in this paper are those of the author and do not necessarily represent the views of the IMF, its executive board or management. Special thanks to C. Schwarz for providing the Stata LDA codes.

1. Introduction

Text data analysis can be a useful tool to analyse the main topics addressed in central bank official documents. The aim of this note is to explain the LDA methodology as presented in Schwarz (2018), and show some results obtained using the ldagibbs Stata command. We analysed 45,346 entries of the Federal Open Market Committee (FOMC) during the period of 2003-2012, with the goal of detecting the evolution of the different topics discussed by the members of the FOMC.

Overall, we detected that discussions on economic modelling played an important role during the Great Financial Crisis (GFC), followed by a considerable increase in discussions on the banking system in the following years, and discussions on communication gained importance at the end of the sample.

2. FOMC Meetings

The FOMC decided in 1976 to release and make publicly available a detailed memorandum of all the discussions taking place at its meetings. The Federal Reserve Act states that the objectives of monetary policy enhanced by the FOMC shall "promote effectively the goals of maximum employment, stable prices and moderate long-term interest rates". There exists considerable debate among economists on how to translate these goals into a coherent description of U.S. monetary policy. This is the reason why a detailed and precise account on the discussions taking place during the FOMC meetings can result useful to understand the evolution in the conduct of U.S. monetary policy.

The FOMC meets eight times in a year to formulate monetary policy and determine other Federal Reserve policies. It is composed of nineteen members comprising seven Governors of the Federal Reserve Board located in Washington D.C., of whom one is the Chairperson of both the Board of Governors and the FOMC, and twelve Presidents of the Regional Federal Reserve Banks with the President of the New York Fed as Vice-Chairman of the FOMC.

The main policy variable of the FOMC is a target for the Federal Funds rate, as well as potential guidance on future monetary policy. At every meeting, all the seven governors have a vote, together with the president of the New York Fed and four of the remaining eleven Fed Presidents who vote on a rotating basis.

Most FOMC meetings last a single day except the meetings that take place before the Monetary Policy Report for the President, which last for two days. During each meeting, every member participates in the discussions independently from their voting right. In this note we analyse the transcripts of these meetings focusing on the conversations held between the FOMC members.

3. Methodology

As explained in Schwarz (2018), Latent Dirichlet Allocation (LDA) consists of two parts. The first is based on a probabilistic model describing the text data as a likelihood function. In the second part, given the unfeasibility of maximizing the likelihood function of text data, LDA utilizes an inference algorithm.

The probabilistic model of LDA considers that every document *d* of the *D* documents in the whole text can be assumed as a probabilistic mixture of *T* topics. These probabilities can be found in a document vector θ_d of length *T*. The value of T, that is, the number of topics, is decided by the user according to the preciseness required. The output of LDA is a $D \times T$ matrixθ containing the probabilities *P(*t_t*|*d_d*)*, of each document *d* belonging to topic *t*:

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_D \end{pmatrix} = \begin{pmatrix} P(t_1|d_1) & \cdots & P(t_T|d_1) \\ \vdots & \ddots & \vdots \\ P(t_1|d_D) & \cdots & P(t_T|d_D) \end{pmatrix}$$

Each topic $t \in T$ is defined by a probabilistic distribution over the vocabulary (the set of words in all documents) of size *V*. In this paper, documents related to forecasting will have a high probability of containing words such as "expectations" or "market-based", while in documents related to economic modelling there will be a higher probability of finding terms such as "model", "standard errors" or "shocks". The word probability vectors of the topics can be represented in a matrix φ of dimensions $V \times T$:

$$\varphi = (\varphi_1 \ldots\ldots, \varphi_T) = \begin{pmatrix} P(w_1|t_1) & \cdots & P(w_1|t_T) \\ \vdots & \ddots & \vdots \\ P(w_v|t_1) & \cdots & P(w_v|t_T) \end{pmatrix}$$

The probabilities $P(w_v|t_T)$ in φ_t describe how probable it is to observe word *w* from the vocabulary conditional on topic *t*. Hence, the φ_t vectors permit to decide the content of each topic and how each topic can eventually be named, since LDA does not produce concrete topic labels. These need to be decided by the users according to their knowledge on the subject.

With parameters θ and φ, the LDA probabilistic model infers that the whole data text is generated by the following process. First, a word probability distribution is drawn following $\varphi \sim Dir(\beta)$. For each document *d* in the text, topic proportions are drawn following $\theta_d \sim Dir(\alpha)$. For each of the N_d words w_d, a topic assignment is drawn such that $z_{d,n} \sim Mult(\theta_d)$ and each word $w_{d,n}$ is drawn from $p(w_{d,n}|z_{d,n}, \varphi)$. In this model, α and β are hyperparameters required for the Gibbs sampling process. The overall likelihood of the whole text with respect to the model parameters is:

$$\prod_{d=1}^{D} P(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{d,n}} P\big(z_{d,n}|\theta_d\big) P\big(w_{d,n}|z_{d,n}, \varphi\big) \right)$$

$P(\theta_d|\alpha)$ denotes how likely it is to obtain the topic distribution of θ_d of document d conditional on α. $P(z_{d,n}|\theta_d)$ determines how likely the topic assignment $z_{d,n}$ of word n in document d is conditional on the topic distribution of the document. Finally, $P(w_{d,n}|z_{d,n},\varphi)$ is the probability of having a concrete word conditional on the topic assignment of the word and the word probabilities of the given topics that are in φ. By calculating the sum over all possible topic assignments, the product over all words in a document and the product over all documents in the text, we obtain the likelihood of observing the texts in the documents.

The LDA procedure is based on finding the optimal topic assignment $z_{d,n}$ for every word in each document and the optimal word probabilities φ for each topic that maximizes this likelihood. Maximizing this likelihood would need adding over all possible topic assignment for all words in all documents, which would result in being computationally unfeasible. Thus, alternative methods such as the Gibbs sampler have been developed for this purpose. In this work, such method is used following Griffiths and Steyvers (2004), based on the ldagibbs Stata command introduced by Schwarz (2018).

Gibbs sampling consists of a Markov Chain Monte Carlo (MCMC) algorithm based on repeatedly drawing new samples conditional on all other data. In the case of LDA, the Gibbs sampler relies on iteratively updating the topic assignment of words conditional on the topic assignments of all other words. As Gibbs Sampling is a Bayesian technique, it requires priors for the values of the hyperparameters α and β, which lie within the unit interval. The prior for α is chosen based on the number of topics T while the prior for β depends on the size of the vocabulary. The higher the number of topics or the larger the vocabulary, the smaller the priors for α and β will be chosen. In general the choice of the priors will not influence the outcome of the sampling process.

Firstly, the ldagibbs algorithm splits the document into single words or word tokens. These are randomly assigned to one of the T topics with equal probability. This gives an initial assignment of words and thereby documents to topics for the sampling process. Later ldaggibs samples new topic assignments for each of the word tokens, with the probability of a word token being assigned to topic t being:

$$P(z_{d,n}=t|w_{d,n},\varphi) \propto P(w_{d,n}|z_{d,n}=t,\varphi)\cdot P(z_{d,n}=t)$$

The Gibbs Sampler makes use of the topic assignment of all other tokens in order to acquire approximate values for $P(z_{d,n}=t|w_{d,n},\varphi)$ and $P(z_{d,n}=t)$.$P(w_{d,n}|z_{d,n}=t,\varphi)$ is calculated by the number of of words which are identical to $w_{d,n}$ and assigned to topic t divided by the total number of words assigned to that topic.

4. Analysis and Results

We used a total of 80 FOMC meeting transcripts covering all the meetings that took place between 2003 and 2012. A full set of minutes for each FOMC meeting is published three weeks after each regular meeting but complete transcripts are published only five years after the meeting. It is precisely these complete transcripts that we used in our analysis. We introduced the text of the FOMC transcripts into the Stata software database dividing each of the transcripts into data text entries consisting of sentences or paragraphs mentioned by the Governors during the meetings.

A total of 45,346 discussion entries were analyzed covering all the conversations that took place between FOMC members. Staff explanations were not included, putting thus an emphasis on the predominant topics discussed by the Governors during the meetings. The LDA algorithm was then implemented with the goal of splitting the whole text data into 8 distinguishing topics. After a careful analysis of the data texts with highest probability of belonging to each topic, we decided that the topics corresponded to the following themes: Forecasting, Economic Modelling, Statement Language, Risks, Banking, Voting Decisions, Economic Activity and Communication.

The average evolution of the probability of each of the topics being addressed during this period at each of the meetings is graphically shown in Figure 1. Discussions on economic modelling played a major role during the GFC, followed by an increase in the discussion of the banking system in the following years, and in the most recent years discussions on communication have gained relevance. Figure 2 shows the evolution of the number of data entries assigned to each topic. A clear rising upward trend can be detected in the amount of text of the transcripts, showing that FOMC meetings have become more extensive.

Figure 1: Average Topic Share of FOMC Transcripts (2003-2012)

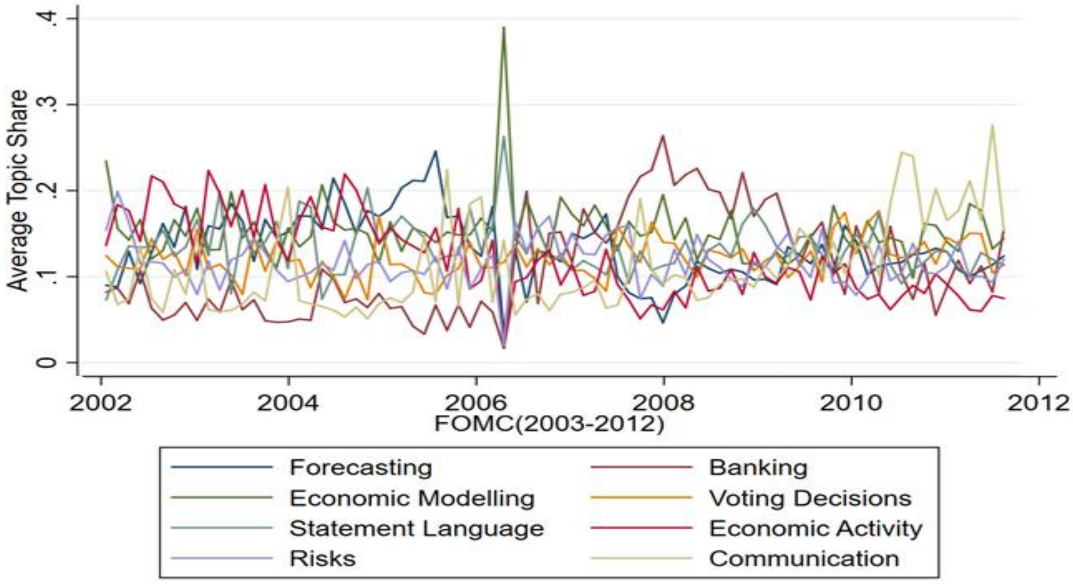
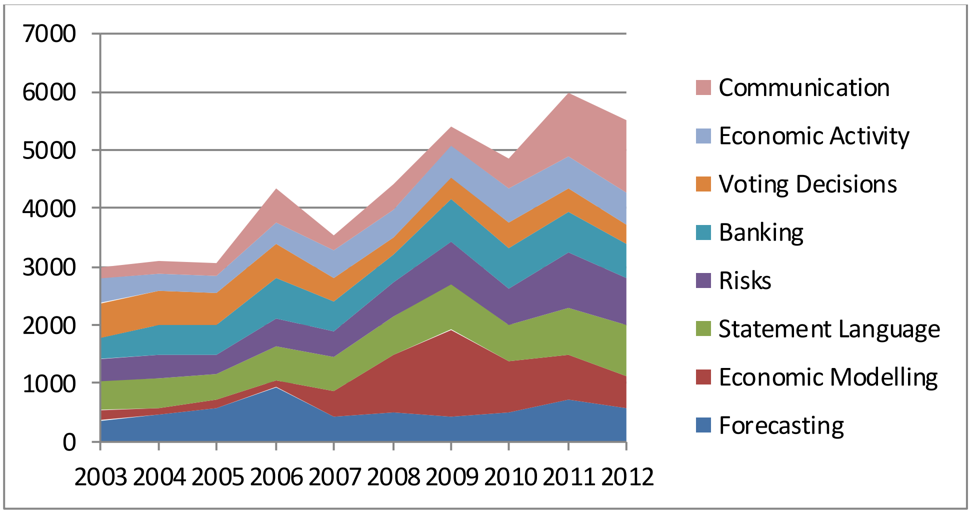


Figure 2: Annual number of data entries (sentences and paragraphs) in FOMC Transcripts (2003-2012)



5. Concluding Comments

The LDA algorithm can be easily implemented to analyze the different themes and their corresponding evolution in terms of use throughout time. In this note we have explained the algorithm, its implementation and estimation and we have provided an empirical example by analyzing the FOMC transcripts covering the meetings that took place during the period 2003-2012.

The use of the LDA algorithm and in particular the Stata command ldagibbs introduced by Schwarz (2018) can be easily implemented to detect which topics are addressed within an abundant number of documents. In this note, we have presented the case of the FOMC transcripts, applying the algorithm to more than 45,000 text data entries and obtaining the evolution of eight identified topics. We observed that discussions on economic modelling played a major role during the GFC, followed by an increase in the discussion of the banking system in the following years, with discussions on communication gaining relevance at the end of the sample. Such type of analysis could be further exploited and employed by researchers at central banks or institutions aiming at determining topic priorities in their official documents.

References

Griffiths, T.L. and M. Steyvers (2004) Finding scientific topics, *Proceedings of the National Academy of Sciences* 101, 5228-5235.

Schwarz, C. (2018) ldagibbs: A Command for Topic Modelling in Stata using Latent Dirichlet Allocation, *The Stata Journal* 18, 1, 101-117.

STATISTICS

Text Data Analysis using Latent Dirichlet Allocation: an Application to FOMC Transcripts

IFC WORKSHOP ON DATA SCIENCE IN CENTRAL BANKING
OCTOBER 21, 2021

Hector Carcel-Villanova

Financial Institutions Division, Statistics Department

The views expressed in this presentation are those of the author and do not necessarily represent the views of the IMF, its executive board or management.

What is Text Data Analysis and its goal

- A useful tool for disentangling and analyzing main topics in different kinds of documents.

- Detect changes in topics and subjects discussed in committees, public institutions, etc.

- Our contribution: a) Explain the LDA algorithm.
  b) Apply it to the analysis of the FOMC transcripts during the GFC.

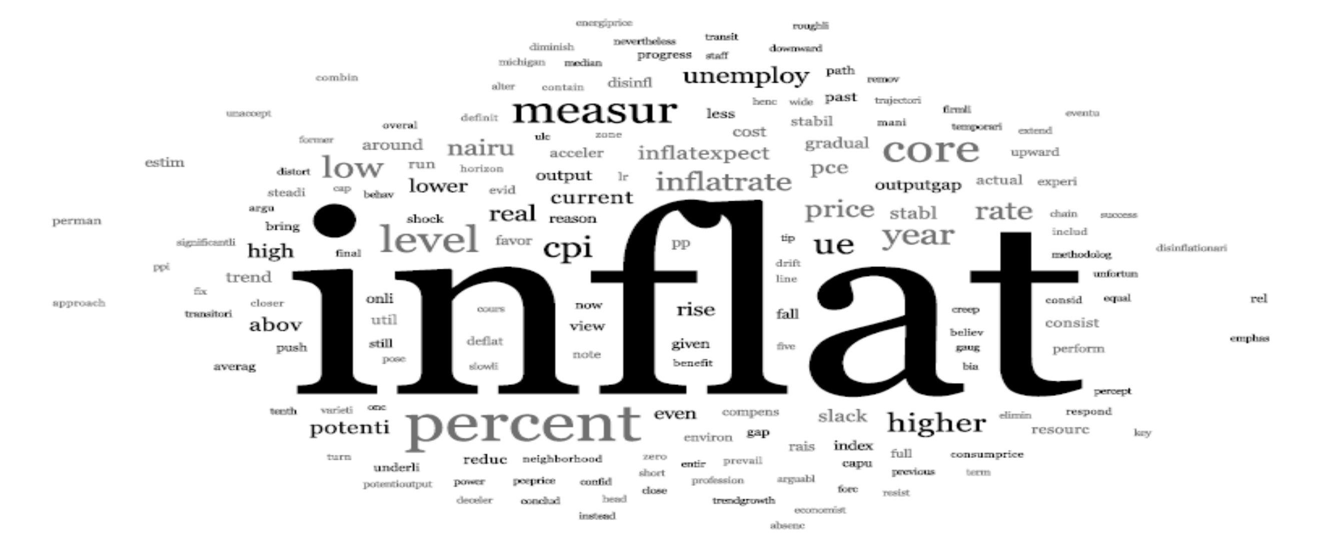


Figure 1: Topic 25—"Inflation"

Methodology

- Each document *d* of the *D* documents in the whole text can be described as a probabilistic combination of *T* topics.

- The outcome of LDA is a $D \times T$ matrix θ containing *P(*$t_t | d_d$*)*, with θ_1,.. θ_D being $1 \times T$ vectors, in such a way that the probability of document *d* belonging to topic *t* corresponds to:

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_D \end{pmatrix} = \begin{pmatrix} P(t_1|d_1) & \cdots & P(t_T|d_1) \\ \vdots & \ddots & \vdots \\ P(t_1|d_D) & \cdots & P(t_T|d_D) \end{pmatrix}$$

- Every topic $t \in T$ is determined by a probabilistic distribution over the vocabulary (the set of words in all documents) of size *V.*

- The word probability vectors of each of the topics can be represented in a matrix φ of dimensions $V \times T$:

$$\varphi = (\varphi_1 \ldots\ldots, \varphi_T) = \begin{pmatrix} P(w_1|t_1) & \cdots & P(w_1|t_T) \\ \vdots & \ddots & \vdots \\ P(w_v|t_1) & \cdots & P(w_v|t_T) \end{pmatrix}$$

Methodology

- Given the parameters θ and φ, the LDA probabilistic model considers that the whole data text is created by the following procedure:

  1. A word probability distribution is drawn following $\varphi \sim Dir(\beta)$.

  2. For each document *d* in the text, topic proportions are drawn following $\theta_d \sim Dir(\alpha)$.

  3. For each of the N_d words w_d, a topic assignment is drawn such that $z_{d,n} \sim Mult(\theta_d)$ and each word $w_{d,n}$ is drawn from $p(w_{d,n} | z_{d,n}, \varphi)$.

- The likelihood of the whole text with respect to the model parameters is:

$$\prod_{d=1}^{D} P(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{d,n}} P\left(z_{d,n} | \theta_d\right) P\left(w_{d,n} | z_{d,n}, \varphi\right) \right)$$

- LDA is based on finding the optimal topic assignment $z_{d,n}$ for each word in each document and the optimal word probabilities φ for each topic that maximizes this likelihood.

Methodology

- This would require adding up all possible topic assignments for all words in all documents, which is computationally impossible.

- Alternative methods such as the Gibbs sampler have been developed for this purpose.

- ldagibbs Stata command introduced by Schwarz (2018).

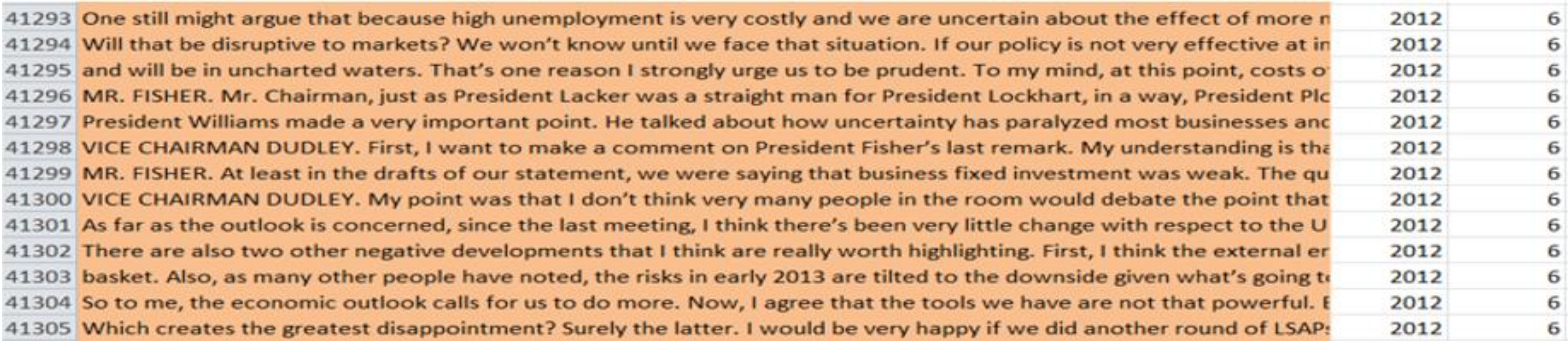
C. Schwarz (2018) *ldagibbs: A command for Topic Modelling in Stata using Latent Dirichlet Allocation. The Stata Journal 18,1, 101-117.*

Study set-up

- We used a total of 80 FOMC meeting transcripts, covering all the meetings between 2003 and 2012.
- How discussions at the FOMC meetings evolved leading to the GFC, at its height and thereafter.
- Transcripts divided into data text entries consisting of sentences or paragraphs stated by the governors during the meetings.
- A total of 45,346 discussion entries analyzed, covering all the conversations between FOMC members.
- Staff explanations were not included.

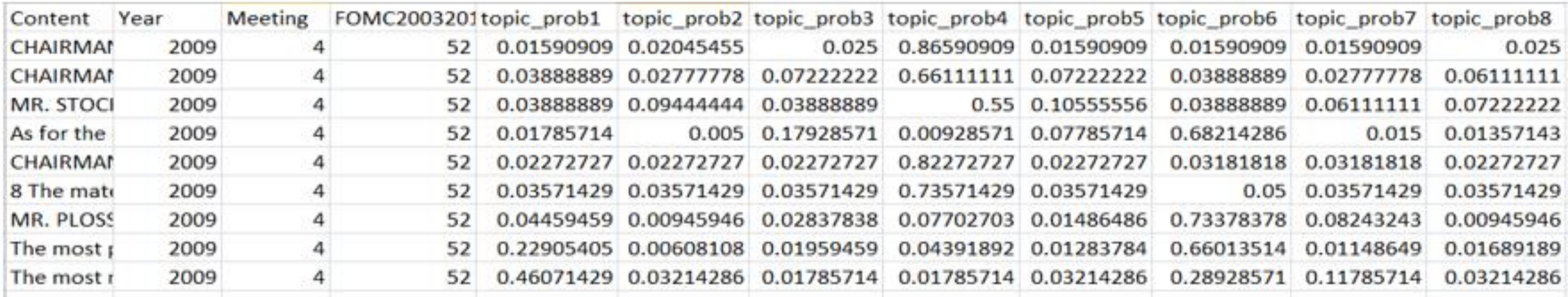
LDA Output

Data Text:



41293	One still might argue that because high unemployment is very costly and we are uncertain about the effect of more n	2012	6
41294	Will that be disruptive to markets? We won't know until we face that situation. If our policy is not very effective at in	2012	6
41295	and will be in uncharted waters. That's one reason I strongly urge us to be prudent. To my mind, at this point, costs o	2012	6
41296	MR. FISHER. Mr. Chairman, just as President Lacker was a straight man for President Lockhart, in a way, President Plc	2012	6
41297	President Williams made a very important point. He talked about how uncertainty has paralyzed most businesses anc	2012	6
41298	VICE CHAIRMAN DUDLEY. First, I want to make a comment on President Fisher's last remark. My understanding is tha	2012	6
41299	MR. FISHER. At least in the drafts of our statement, we were saying that business fixed investment was weak. The qu	2012	6
41300	VICE CHAIRMAN DUDLEY. My point was that I don't think very many people in the room would debate the point that	2012	6
41301	As far as the outlook is concerned, since the last meeting, I think there's been very little change with respect to the U	2012	6
41302	There are also two other negative developments that I think are really worth highlighting. First, I think the external er	2012	6
41303	basket. Also, as many other people have noted, the risks in early 2013 are tilted to the downside given what's going t	2012	6
41304	So to me, the economic outlook calls for us to do more. Now, I agree that the tools we have are not that powerful. E	2012	6
41305	Which creates the greatest disappointment? Surely the latter. I would be very happy if we did another round of LSAP:	2012	6

LDA Output:



Content	Year	Meeting	FOMC2003201	topic_prob1	topic_prob2	topic_prob3	topic_prob4	topic_prob5	topic_prob6	topic_prob7	topic_prob8
CHAIRMAI	2009	4	52	0.01590909	0.02045455	0.025	0.86590909	0.01590909	0.01590909	0.01590909	0.025
CHAIRMAI	2009	4	52	0.03888889	0.02777778	0.07222222	0.66111111	0.07222222	0.03888889	0.02777778	0.06111111
MR. STOCI	2009	4	52	0.03888889	0.09444444	0.03888889	0.55	0.10555556	0.03888889	0.06111111	0.07222222
As for the	2009	4	52	0.01785714	0.005	0.17928571	0.00928571	0.07785714	0.68214286	0.015	0.01357143
CHAIRMAI	2009	4	52	0.02272727	0.02272727	0.02272727	0.82272727	0.02272727	0.03181818	0.03181818	0.02272727
8 The mat	2009	4	52	0.03571429	0.03571429	0.03571429	0.73571429	0.03571429	0.05	0.03571429	0.03571429
MR. PLOSS	2009	4	52	0.04459459	0.00945946	0.02837838	0.07702703	0.01486486	0.73378378	0.08243243	0.00945946
The most p	2009	4	52	0.22905405	0.00608108	0.01959459	0.04391892	0.01283784	0.66013514	0.01148649	0.01689189
The most r	2009	4	52	0.46071429	0.03214286	0.01785714	0.01785714	0.03214286	0.28928571	0.11785714	0.03214286

Topic selection

TOPIC 1: Forecasting

-Turning to inflation, I have nudged my forecast for both core and headline PCE inflation down a little since April ...

-When I compare the Board staff's forecast with ours, I find that the Greenbook projection, even the most updated one ...

TOPIC 2: Banking System

-Wells, Goldman, Bank of New York, Sun Trust, and BB&T, for example—opted out. Whenever a fee is assessed on assets or ...

-The first thing is that if we had a floor system, there would be more reserves in the banking system, and that might ac …

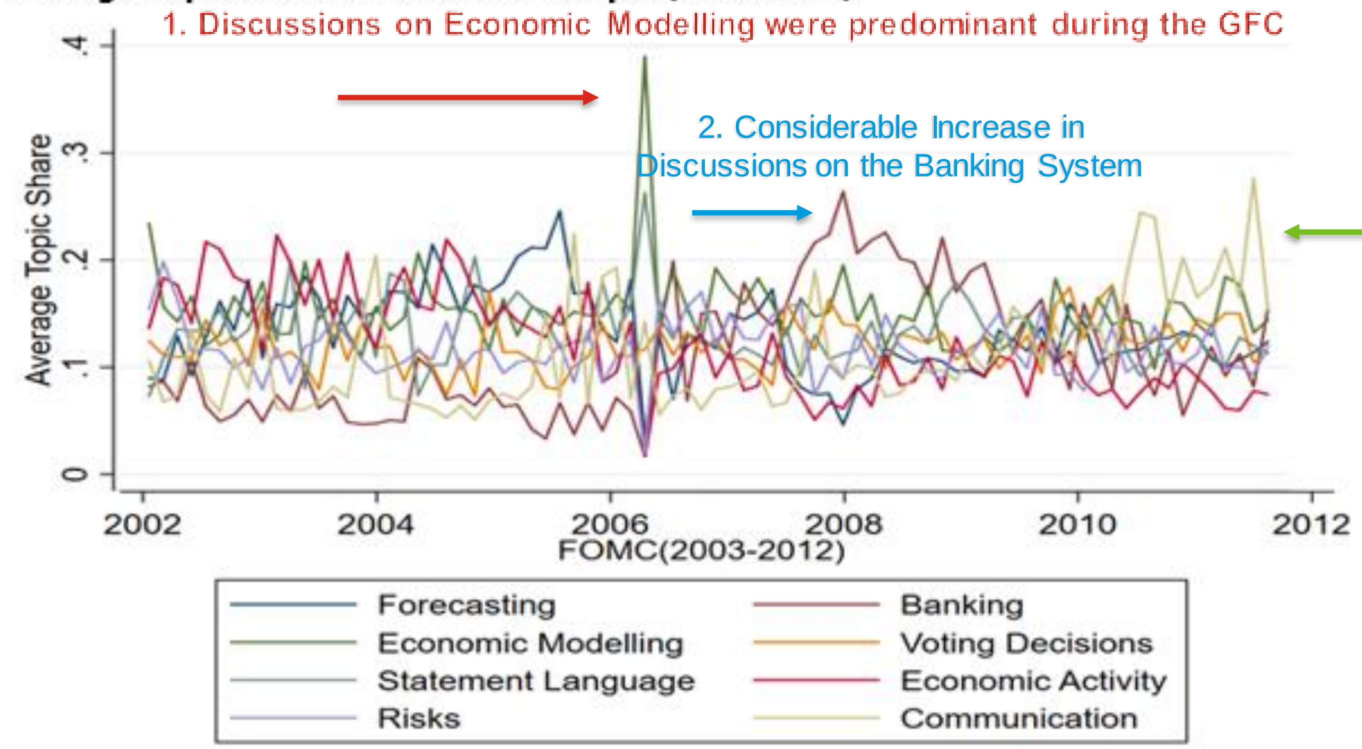
TOPIC 3: Economic Modelling

-Your second question about standard errors is a really good one, and it is hard. There are lots of different models that …

-If you ask whether a DSGE model would tell the story differently from, let's say, FRB/US, the answer is "maybe—it depends ...

Results: FOMC transcripts (2003-2012)

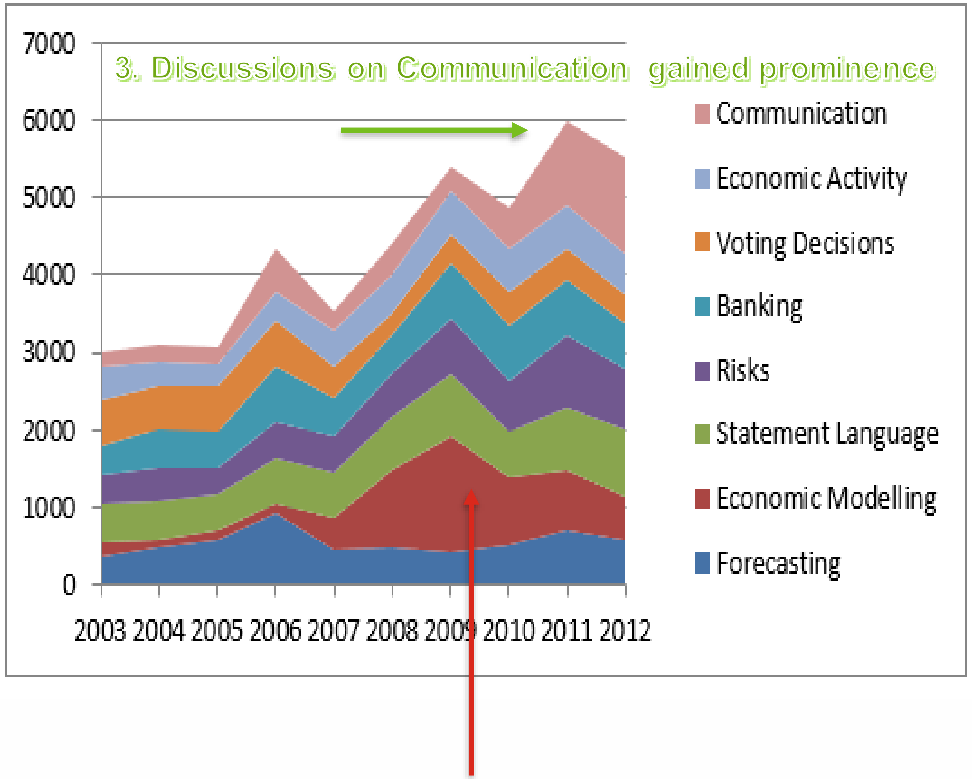
Average Topic Share of FOMC Transcripts (2003-2012)



1. Discussions on Economic Modelling were predominant during the GFC

2. Considerable Increase in Discussions on the Banking System

Annual number of data entries (sentences and paragraphs) in FOMC Transcripts (2003-2012)



3. Discussions on Communication gained prominence

- **LDA could be further used by researchers at central banks and institutions to determine topic priorities in relevant documents.**

- **Future aim: carry out further research to investigate the evolution of concrete economic models (e.g., Phillips curve, Taylor rule, etc.)**

Thank you!

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Estimating the effect of central bank independence on inflation using longitudinal targeted maximum likelihood estimation[1]

Philipp Baumann, ETH Zurich, KOF Swiss Economic Institute, Enzo Rossi, Swiss National Bank, and Michael Schomaker, UMIT University, Austria, and Institute of Statistics, LMU Munich, Munich, Germany

[1] This presentation 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 central banks and other institutions represented at the event.

Estimating the Effect of Central Bank Independence on Inflation Using Machine Learning

Modern doubly robust estimation techniques applied to an old macroeconomic question

Philipp F. M. Baumann, KOF Swiss Economic Institute, ETH Zurich. e-mail: baumann@kof.ethz.ch

Michael Schomaker, Institute of Statistics, LMU Munich, Munich, Germany; Institute of Public Health, Medical Decision Making and Health Technology Assessment, UMIT - University for Health Sciences, Medical Informatics and Technology, Hall in Tirol, Austria and Centre for Infectious Disease Epidemiology & Research, University of Cape Town, Cape Town, South Africa. e-mail: michael.schomaker@stat.uni-muenchen.de

Enzo Rossi, Swiss National Bank and University of Zurich. e-mail: enzo.rossi@snb.ch

Abstract

Does an independent central bank (CBI) reduce inflation? Despite numerous articles suggesting it does, this question has not been satisfactorily answered because the complex macroeconomic structure that gives rise to the data has not been adequately incorporated into economic analyses. We develop a causal model that summarizes the economic process of inflation and estimate the effect of CBI on inflation with modern doubly robust effect estimation techniques. We incorporate a large number of variables in our directed acyclic graph and overcome endogeneity issues of previous studies. Our approach includes machine learning algorithms which are tailored to the question of interest and reduce the chance of model misspecification. In this paper, we provide the motivation and give a short summary of our recent study (Baumann, 2021a).

Keywords: macroeconomics, monetary policy, causal inference, doubly robust

JEL classification: C13, C14, C21, C33, E3

Contents

1. Introduction

At least since David Ricardo's days the institutional relationship between central banks and governments has been a subject of considerable interest. Over the past three decades, following a series of events, such as financial liberalization, greater acceptance of long-run inflation as a monetary phenomenon, and rejection of the long-run tradeoff between inflation and unemployment, research on the impact of central bank independence (CBI) on economic outcomes has intensified (Cargill, 2013). It has been claimed that more than 9,000 works have been devoted to this question (Vuletin and Zhu, 2011). After the 2008-09 Global Financial Crisis, the debate on the optimal design of monetary policy authorities has become even more intense.

The traditional case for CBI rests on countering an inflation bias that may occur for various reasons in the absence of an independent central bank (de Haan et al., 2018). One reason for this bias is political pressure to boost output in the short run for electoral reasons. Another reason is the incentive for politicians to use the central bank's power to issue money as a means to finance government spending. Ricardo considered it a great danger to entrust the ministers with the power of issuing paper money. The third reason is the time-inconsistency problem of monetary policy making. When governments have discretionary control over monetary instruments, they can prioritize other policy goals over price stability. For instance, after nominal wages have been negotiated (or nominal bonds purchased), politicians may be tempted to create inflation to boost employment and output (or to devalue government debt).

To overcome an inflation bias, the literature stresses the benefits of enforced commitments (rules). In particular, Rogoff (1985) proposed delegating monetary policy to an independent and conservative central banker to reduce the tendency to produce high inflation. Once central bankers are insulated from political pressures, commitments to price stability can be credible, which helps to maintain low inflation.

Following these ideas, a policy consensus grew around the potential of having independent central banks to promote inflation stability (Bernhard et al., 2002; Kern et al., 2019). Numerous countries followed this policy advice. Between 1985 and 2012, and excluding the creation of regional central banks, there were 266 reforms to the statutory independence of central banks, 236 of which were implemented in developing countries. Most of these reforms (77%) strengthened CBI (Garriga, 2016).

Despite the broad impact of this policy advice, the empirical evidence in support of it remains controversial. Starting with (Bade and Parkin, 1978, 1982, 1988) and extended by (Cukierman et al., 1992), a substantial body of empirical literature evolved claiming the existence of a statistically significant inverse relationship between measures of CBI and inflation. However, the support for more independent central banks has often been based on correlations between CBI measures and inflation over time and across countries, frequently based on single-variable regression models or in models in which several economic and political variables were added as covariates. While many studies have found that an independent central bank may lower inflation (Alesina and Summers, 1993; Arnone and Romelli, 2013; Cukierman et al., 1992; Grilli et al., 1991; Klomp and De Haan, 2010b,a), other studies that have used a broader range of characteristics of a nation's economy have been unable to find such a relationship (Cargill, 1995b; Fuhrer, 1997; Oatley, 1999; Campillo and Miron, 1997; Fujiki, 1996). Other studies suggest that the effect of CBI on inflation can only be seen during specific time periods (Klomp and De Haan, 2010a) or only in developed countries (Alpanda and Honig, 2014; Klomp and De Haan, 2010b; Neyapti, 2012).

2. Motivation

Evaluating the effect of CBI on inflation based on simple cross-sectional regression approaches has some important weaknesses. First, a few researchers question the entire framework of measuring CBI and the statistical results obtained, arguing that these approaches provide misinformation about the fundamental relationship between the central bank and government. They emphasize the difficulty of measuring CBI and the predictive power of the estimated relationship for some countries (Cargill, 1989, 1995a). Second, there is a question as to the direction of causation implied by simple correlations between CBI and inflation (Posen, 1998). A third critique that advises caution in interpreting the results is the focus on de jure rather than de facto measures of independence (Pollard, 1993). Klomp and De Haan (2010a) combined 59 studies in a meta-regression analysis and concluded that the particular CBI measure used has little effect on the estimated effect and that there is indeed a negative and significant relationship between CBI and inflation. However, echoing the mixed evidence reported in the literature, Parkin (2013) notes that the meta-regression does not control for the amount of data mining undertaken. Nor does the conclusion sit well with the details of the 59 studies included in the analysis. From the 384 regressions included in the studies, 202 exhibit a significant negative relationship while 182 show either no relationship or a significant but “wrong” sign.

In our study (Baumann, 2021a) we offer a solution to two of the major problems encountered by previous empirical work. First, since the problem at hand is longitudinal in nature, only an appropriate panel setup may be suitable to estimate the (long-term) effect of CBI on inflation. Second, the abovementioned cross-sectional regression approaches do not incorporate any causal considerations into their analyses. We propose a novel framework which takes causality explicitly into account. Specifically, we ask what (average) inflation would we observe in 10 years’ time, if from now on each country had an independent central bank compared to a situation in which the central bank were not independent. The data set we use was created specifically for this purpose and extends the data set from Baumann et al. (2021b).

3. Causality in Complex Settings

While evaluating the effect of CBI on inflation requires a longitudinal causal estimation approach, it has been shown repeatedly that standard regression approaches are typically not suitable to answer causal questions, particularly when the setup is longitudinal and when the confounders of the outcome-intervention relationships are affected by previous intervention decisions (Daniel et al., 2013). There are at least three methods to evaluate the effect of longitudinal (multiple time-point) interventions on an outcome in such complex situations: 1) inverse probability of treatment weighted (IPTW) approaches (Robins et al., 2000); 2) standardization with respect to the time-dependent confounders (i.e., g-formula-type approaches (Robins, 1986; Bang and Robins, 2005)); and 3) doubly robust methods, such as targeted maximum-likelihood estimation (TMLE, Van der Laan and Rose, 2011), which can be seen as a combination and generalization of the other two approaches.

Using causal inference in economics has a long history, starting with path analyses and potential outcome language (Tinbergen, 1930; Wright, 1934) and continuing with regression discontinuity analyses (Hahn et al., 2001), instrumental variable designs (Imbens, 2014), and propensity score approaches in the context of the potential outcome framework (Rosenbaum and Rubin, 1983), among many other methods. More recently, there have been work advocating the use of doubly robust techniques in econometrics (Chernozhukov et al., 2018). From the perspective of statistical inference this is a very promising suggestion because the integration of modern machine learning methods in causal effect estimation is almost inevitable in areas with a large number of covariates and complex data-generating processes (Schomaker et al., 2019).

However, the application of doubly robust effect estimation can be challenging for (macro-)economic data. First, the causal model that summarizes the knowledge about the data-generating process is often more complex for economic than for epidemiological questions, where most successful implementations have been demonstrated so far (Kreif et al., 2017; Decker et al., 2014; Schnitzer, Moodie, van der Laan, Platt and Klein, 2014; Schnitzer, van der Laan, Moodie and Platt, 2014; Schnitzer, Lok and Bosch, 2016; Tran et al., 2016; Schomaker et al., 2019; Bell-Gorrod et al., 2019). The task of representing the causal model in a directed acyclic graph (DAG) becomes particularly challenging when considering how economic variables interact with each other over time. Thus, in order to build a DAG, a thorough review of literature is called for, and economic feedback loops need to be incorporated appropriately. Imbens (2019), who discusses different schools of causal inference and their use in statistics and econometrics, as well as different estimation techniques, emphasizes this point: "[...] a major challenge in causal inference is coming up with the causal model."

Second, even if a causal model has been developed, the identification of an estimand has been established, and the data have been collected, statistical estimation may be nontrivial given the complexity of a particular data set (Schomaker et al., 2019). If the sample size is small, potentially smaller than the number of (time- varying) covariates, recommended estimation techniques can fail, and the development of an appropriate set of learning and screening algorithms is important. The benefits of LTMLE, which is doubly robust effect estimation in conjunction with machine learning to reduce the chance of model misspecification, can be best utilized under a good and broad selection of learners that are tailored to the problem of interest.

3. Contributions of our Study

Estimating the effect of CBI on inflation is a typical example of a causal inference question that faces all of the challenges described above. Our paper makes five novel contributions to the literature. i) We discuss identification and estimation for our question of interest and estimate the effect of CBI on inflation; ii) develop a causal model that can be applied to other questions related to macroeconomics; iii) demonstrate that it is possible to develop a DAG for economic questions, which is important, as it has been argued that "the lack of adoption in economics is that the DAG literature has not shown much evidence of the benefits for empirical practice in settings that are important in economics." (Imbens, 2019); iv) demonstrate how to integrate machine learning into complex causal effect estimation, including how to define a successful learner set when the number of covariates is larger than the sample size and when there is time-dependent confounding with treatment-confounder feedback (Hernan and Robins, 2020); and v) use simulations to study the performance of doubly robust estimation techniques under the challenges described above.

4. Results

As discussed in Baumann et al. 2021a, our main analysis, PlainDAG, shows that if a country had legislated central bank independence for every year between 1998 and 2008, it would have had an average increase in inflation of 0.01 (95% confidence interval: -1.48, 1.50) percentage points in 2010. We conducted a further analysis where a central bank is made independent, if the corresponding country has experienced an inflation rate that is generally considered as too low or too high, respectively. That is, if a country had legislated an independent central bank for every year when the median of the past seven years of inflation had been above 5% or below 0% from 1998 to 2008, it would have led to an average reduction in inflation of -0.07 percentage points only (95% confidence interval: -1.29, 1.15) in 2010 compared to a dependent central bank for the same time span. We conducted two robustness checks with regard to the causal assumptions underlying our model. These two approaches are ScreenLearn and EconDAG. The results suggest somewhat stronger reductions of inflation caused by higher central bank independence (up to -0.61 percentage points). As suggested by the wideness of the confidence intervals, we can exclude neither a strong negative nor a strong positive average treatment effect.

References

Bernhard, W., Broz, J. L. and Clark, W. R. (2002), “The political economy of monetary institutions”, International Orga- nization 56(4), 693–723.

Cargill, T. F. (2013), “A critical assessment of measures of central bank independence”, Economic Inquiry 51(1), 260–272.

Vuletin, G. and Zhu, L. (2011), “Replacing a “disobedient” central bank governor with a “docile” one: A novel measure of central bank independence and its effect on inflation”, Journal of Money, Credit and Banking 43(6), 1185–1215.

de Haan, J., Bodea, C., Hicks, R. and Eijffinger, S. C. (2018), “Central bank independence before and after the crisis”, Comparative Economic Studies 60(2), 183–202.

Rogoff, K. (1985), “The optimal degree of commitment to an intermediate monetary target”, The Quarterly Journal of Economics 100(4), 1169–1189.

Kern, A., Reinsberg, B. and Rau-G ̈ohring, M. (2019), “Imf conditionality and central bank independence”, European Journal of Political Economy 59(C), 212–229.

Bade, R. and Parkin, M. (1978), "Central bank laws and monetary policies: A preliminary investigation", The Australian Monetary System in the 1970s, Monash University.

Bade, R. and Parkin, M. (1982), "Central bank laws and monetary policy", Unpublished Manuscript, University of Western Ontario London, ON.

Bade, R. and Parkin, M. (1988), "Central bank laws and monetary policy", Unpublished Manuscript, University of Western Ontario London, ON.

Cukierman, A., Web, S. B. and Neyapti, B. (1992), "Measuring the independence of central banks and its effect on policy outcomes", The world bank economic review 6(3), 353–398.

Alesina, A. and Summers, L. H. (1993), "Central bank independence and macroeconomic performance: some comparative evidence", Journal of Money, Credit and Banking 25(2), 151–162.

Arnone, M. and Romelli, D. (2013), "Dynamic central bank independence indices and inflation rate: A new empirical exploration", Journal of Financial Stability 9(3), 385–398.

Grilli, V., Masciandaro, D. and Tabellini, G. (1991), "Political and monetary institutions and public financial policies in the industrial countries", Economic Policy 6(13), 341–392.

Klomp, J. and De Haan, J. (2010a), "Central bank independence and inflation revisited", Public Choice 144(3-4), 445–457.

Klomp, J. and De Haan, J. (2010b), "Inflation and central bank independence: a meta-regression analysis", Journal of Economic Surveys 24(4), 593–621.

Cargill, T. (1995b), "The statistical association between central bank independence and inflation", BNL Quarterly Review 48(193), 159–172.

Fuhrer, J. C. (1997), "Central bank independence and inflation targeting: monetary policy paradigms for the next millennium?", New England Economic Review Jan/Feb, 19–36.

Oatley, T. (1999), "Central bank independence and inflation: Corporatism, partisanship, and alternative indices of central bank independence", Public Choice 98(3-4), 399–413.

Campillo, M. and Miron, J. A. (1997), Why does inflation differ across countries?, in "Reducing inflation: Motivation and strategy", University of Chicago Press, pp. 335–362.

Fujiki, H. (1996), "Central bank independence indexes in economic analysis: A reappraisal", Monetary and Economic Studies Bank of Japan(14), 79–101.

Alpanda, S. and Honig, A. (2014), "The impact of central bank independence on the performance of inflation targeting regimes", Journal of International Money and Finance 44, 118–135.

Neyapti, B. (2012), "Monetary institutions and inflation performance: cross-country evidence", Journal of Economic Policy Reform 15(4), 339–354.

Cargill, T. (1995a), "The bank of japan and the federal reserve: An essay on central bank independence", Monetarism and the Methodology of Economics, ed. K. Hoover and S. Sheffrin pp. 198–214.

Cargill, T. F. (1989), Central bank independence and regulatory responsibilities: the Bank of Japan and the Federal Reserve, number 2, Salomon Brothers Center for the Study of Financial Institutions.

Posen, A. S. (1998), "Do better institutions make better policy?", International finance 1(1), 173–205.

Pollard, P. S. (1993), "Central bank independence and economic performance", Review Federal Reserve Bank of St. Louis, 21– 36.

Parkin, M. (2013), Central bank laws and monetary policy outcomes: A three decade perspective, Technical report, EPRI Working Paper.

Baumann, P. F. M., Schomaker, M. and Rossi, E. (2021a) Estimating the effect of central bank independence on inflation using longitudinal targeted maximumlikelihood estimation. Journal of Causal Inference, 9, 109–146. URL: https://doi.org/10.1515/jci-2020-0016.

Baumann, P. F. M., Rossi, E. and Volkmann, A. (2021b) What drives inflationand how? evidence from additive mixed models selected by caic. Swiss National Bank Working Paper Series,12. URL: https://www.snb.ch/de/mmr/papers/id/working_paper_2021_12.

Daniel, R. M., Cousens, S. N., De Stavola, B. L., Kenward, M. G. and Sterne, J. A. (2013), "Methods for dealing with time-dependent confounding", Statistics in Medicine 32(9), 1584–618.

Robins, J. M., Hernan, M. A. and Brumback, B. (2000), "Marginal structural models and causal inference in epidemiology", Epidemiology 11(5), 550–560.

Robins, J. (1986), "A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect", Mathematical Modelling 7(9-12), 1393–1512.

Bang, H. and Robins, J. M. (2005), "Doubly robust estimation in missing data and causal inference models", Biometrics 64(2), 962–972.

Van der Laan, M. and Rose, S. (2011), Targeted Learning, Springer.

Tinbergen, J. (1930), "Determination and interpretation of supply curves: an example", Zeitschrift fu ̈r Nationalo ̈konomie 1, 669–679.

Wright, P. (1934), "The method of path coeffcients", The Annals of Mathematical Statistics 5, 161–215.

Hahn, J., Todd, P. and Van der Klaauw, W. (2001), "Identification and estimation of treatment effects with a regression- discontinuity design", Econometrica 69(1), 201–209.

Imbens, G. (2014), "Instrumental variables: An econometrician's perspective", Statistical Science 29(3), 323–358.

Rosenbaum, P. and Rubin, D. (1983), "The central role of the propensity score in observational studies for causal effects", Biometrika 70(1), 688–701.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), "Double/debiased machine learning for treatment and structural parameters", The Econometrics Journal 21(1), C1–C68.

Schomaker, M., Luque-Fernandez, M. A., Leroy, V. and Davies, M.-A. (2019), “Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions”, Statistics in Medicine 38(24), 4888–4911.

Kreif, N., Tran, L., Grieve, R., deStavola, B., Tasker, R. and Petersen, M. (2017), “Estimating the comparative effectiveness of feeding interventions in the paediatric intensive care unit: a demonstration of longitudinal targeted maximum likelihood estimation”, American Journal of Epidemiology 186, 1370–1379.

Decker, A., Hubbard, A., Crespi, C., Seto, E. and Wang, M. (2014), “Semiparametric estimation of the impacts of longitudinal interventions on adolescent obesity using targeted maximum-likelihood: Accessible estimation with the ltmle package”, Journal of Causal Inference 2(1), 95–108.

Schnitzer, M. E., Moodie, E. E., van der Laan, M. J., Platt, R. W. and Klein, M. B. (2014), “Modeling the impact of hepatitis C viral clearance on end-stage liver disease in an HIV co-infected cohort with targeted maximum likelihood estimation”, Biometrics 70(1), 144–52.

Schnitzer, M. E., van der Laan, M. J., Moodie, E. E. and Platt, R. W. (2014), “Effect of breastfeeding on gastrointestinal infection in infants: A targeted maximum likelihood approach for clustered longitudinal data”, Annals of Applied Statistics 8(2), 703–725.

Schnitzer, M. E., Lok, J. and Bosch, R. J. (2016), “Double robust and efficient estimation of a prognostic model for events in the presence of dependent censoring”, Biostatistics 17(1), 165–177.

Schomaker, M., Luque-Fernandez, M. A., Leroy, V. and Davies, M.-A. (2019), “Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions”, Statistics in Medicine 38(24), 4888–4911.

Bell-Gorrod, H., Fox, M. P., Boulle, A., Prozesky, H., Wood, R., Tanser, F., Davies, M.-A. and Schomaker, M. (2019), “The impact of delayed switch to second-line antiretroviral therapy on mortality, depending on failure time definition and cd4 count at failure”, bioRxiv .
URL: https://www.biorxiv.org/content/biorxiv/early/2019/06/07/661629.full.pdf

Imbens, G. (2019), Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics, Working Paper 26104, National Bureau of Economic Research, MA 02138, U.S.A.
URL: http://www.nber.org/papers/w26104

Hernan, M. and Robins, J. (2020), Causal Inference, Vol. forthcoming of Chapman & Hall/CRC Monographs on Statistics & Applied Probab, Taylor & Francis.
URL: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Estimating the Effect of Central Bank Independence on Inflation Using Longitudinal Targeted Maximum Likelihood Estimation

Philipp F. M. Baumann
(KOF Swiss Economic Institute, ETH Zurich)

joint work with
M. Schomaker (UMIT University Austria) and E. Rossi (Swiss National Bank)

@mf schomaker & @pfmbaumann

October 12, 2021

Introduction

Why a further study on Central Bank Independence (CBI)?

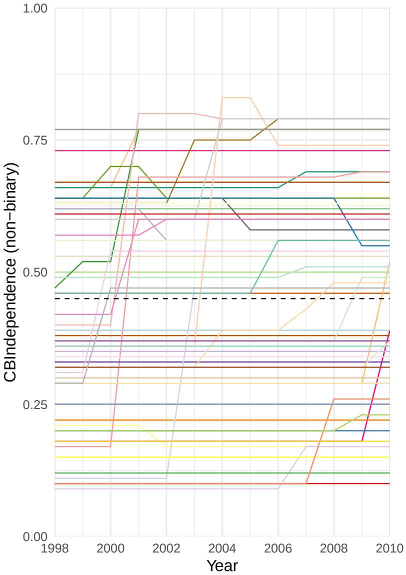
- Models in empirical studies often neglect a holistic causal framework which results in premature causal interpretation.
- Instrumental variable approaches have been proposed to tackle these problems but many authors have been unable to find strong instruments (e.g. Crowe and Meade; 2008).
- Effect estimation in practice: various practical challenges like small sample sizes, too many irrelevant covariates and too restrictive models lead to biased estimators for the causal effect.

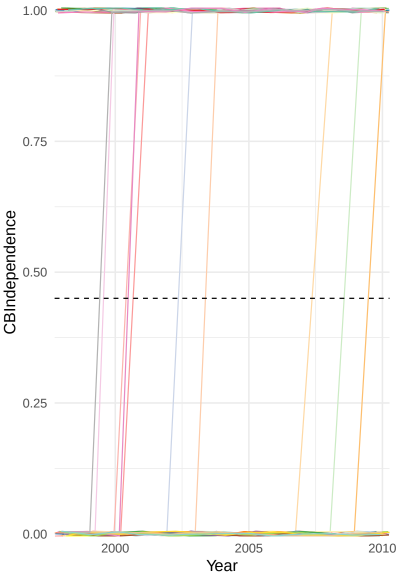
Question of our study: What (average) inflation would we observe in 10 years' time, if – from now on – each country's monetary institution had an independent central bank compared to the situation in which the central bank was not independent?

Data

- We accessed databases of the World Bank and the International Monetary Fund to collect annual data for economic, political, and institutional variables.
- Our aim was to include as many countries as possible in our analysis.
- Missing data (2.7%) lead to the use of multiple imputation.
- Finally, we obtained observations for 60 countries and 13 points in time (i.e., calendar years 1998–2010) for 19 measured variables.
- 20% of the 60 countries are low-income countries, 36% belong to the lower-middle-income category, 27% to the upper-middle-income category, and 17% belong to the high-income category.

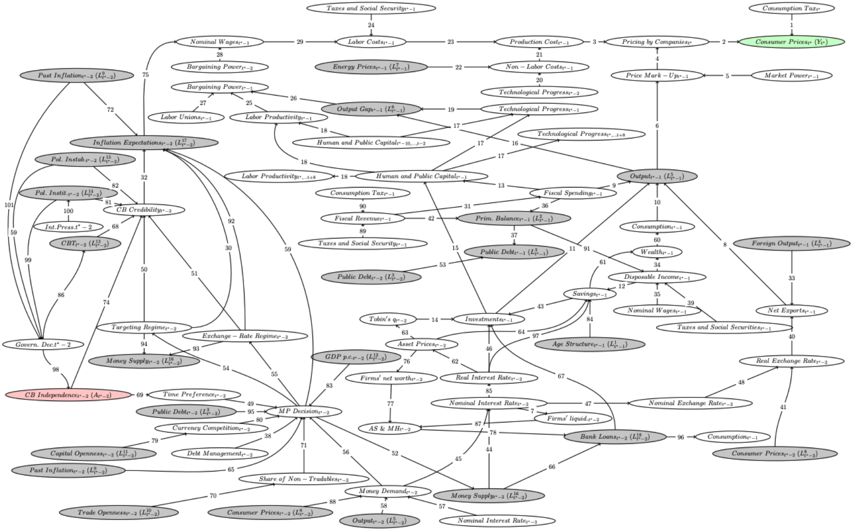
CBI Index: Dincer and Eichengreen (2014)





The Causal Analysis

The Economy as a DAG


$Taxes\ and\ Social\ Security_{t^*-1}$
$Consumption\ Tax_{t^*}$
$Nominal\ Wages_{t^*-1}$
$Labor\ Costs_{t^*-1}$
$Production\ Cost_{t^*-1}$
$Pricing\ by\ Companies_{t^*}$
$Consumer\ Prices_{t^*}\ (Y_{t^*})$
$Bargaining\ Power_{t^*-2}$
$Energy\ Prices_{t^*-1}\ (L^7_{t^*-1})$
$Non-Labor\ Costs_{t^*-1}$
$Past\ Inflation_{t^*-2}\ (L^9_{t^*-2})$
$Price\ Mark-Up_{t^*-1}$
$Market\ Power_{t^*-1}$
$Bargaining\ Power_{t^*-1}$
$Technological\ Progress_{t^*-2}$
$Output\ Gap_{t^*-1}\ (L^6_{t^*-1})$
$Technological\ Progress_{t^*-1}$
$Labor\ Unions_{t^*-1}$
$Labor\ Productivity_{t^*-1}$
$Technological\ Progress_{t^*,\ldots t+8}$
$Inflation\ Expectations_{t^*-2}\ (L^{17}_{t^*-2})$
$Human\ and\ Public\ Capital_{t^*-10,\ldots,t-2}$
$Pol.\ Instab._{t^*-2}\ (L^{15}_{t^*-2})$
$Labor\ Productivity_{t^*,\ldots t+8}$
$Human\ and\ Public\ Capital_{t^*-1}$
$Output_{t^*-1}\ (L^5_{t^*-1})$
$Pol.\ Instit._{t^*-2}\ (L^{14}_{t^*-2})$
$Consumption\ Tax_{t^*-1}$
$Fiscal\ Spending_{t^*-1}$
$CB\ Credibility_{t^*-2}$
$Fiscal\ Revenue_{t^*-1}$
$Prim.\ Balance_{t^*-1}\ (L^2_{t^*-1})$
$Int.Press.t^*-2$
$Consumption_{t^*-1}$
$CBT_{t^*-2}\ (L^{13}_{t^*-2})$
$Taxes\ and\ Social\ Security_{t^*-1}$
$Foreign\ Output_{t^*-1}\ (L^4_{t^*-1})$
$Public\ Debt_{t^*-1}\ (L^3_{t^*-1})$
$Wealth_{t^*-1}$
$Public\ Debt_{t^*-2}\ (L^3_{t^*-2})$
$Disposable\ Income_{t^*-1}$
$Savings_{t^*-1}$
$Tobin's\ q_{t^*-2}$
$Investments_{t^*-1}$
$Nominal\ Wages_{t^*-1}$
$Net\ Exports_{t^*-1}$
$Targeting\ Regime_{t^*-2}$
$Taxes\ and\ Social\ Securities_{t^*-1}$
$Exchange-Rate\ Regime_{t^*-2}$
$Govern.\ Dec.t^*-2$
$Asset\ Prices_{t^*-2}$
$Age\ Structure_{t^*-1}\ (L^1_{t^*-1})$
$Money\ Supply_{t^*-2}\ (L^{16}_{t^*-2})$
$GDP\ p.c._{t^*-2}\ (L^{12}_{t^*-2})$
$Real\ Exchange\ Rate_{t^*-2}$
$Firms'\ net\ worth_{t^*-2}$
$Real\ Interest\ Rate_{t^*-2}$
$CB\ Independence_{t^*-2}\ (A_{t^*-2})$
$Time\ Preference_{t^*-2}$
$Nominal\ Interest\ Rate_{t^*-2}$
$Nominal\ Exchange\ Rate_{t^*-2}$
$Public\ Debt_{t^*-2}\ (L^3_{t^*-2})$
$MP\ Decision_{t^*-2}$
$Firms'\ liquid._{t^*-2}$
$Currency\ Competition_{t^*-2}$
$AS\ \&\ MH_{t^*-2}$
$Bank\ Loans_{t^*-2}\ (L^{18}_{t^*-2})$
$Consumption_{t^*-1}$
$Capital\ Openness_{t^*-2}\ (L^{11}_{t^*-2})$
$Debt\ Management_{t^*-2}$
$Consumer\ Prices_{t^*-2}\ (L^8_{t^*-2})$
$Past\ Inflation_{t^*-2}\ (L^9_{t^*-2})$
$Share\ of\ Non-Tradables_{t^*-2}$
$Money\ Demand_{t^*-2}$
$Money\ Supply_{t^*-2}\ (L^{16}_{t^*-2})$
$Trade\ Openness_{t^*-2}\ (L^{10}_{t^*-2})$
$Consumer\ Prices_{t^*-2}\ (L^8_{t^*-2})$
$Output_{t^*-2}\ (L^5_{t^*-2})$
$Nominal\ Interest\ Rate_{t^*-3}$

Target Parameters

- Our target parameters are average treatment effects (ATEs)
- Three interventions. Two static and one dynamic. $\forall t^* \in \{1998, \ldots, 2008\}$ and $i \in \{1, \ldots, 60\}$

$\bar{d}^1_{t^*}$ = Set every CB i as "independent" for every t^*

$\bar{d}^2_{t^*}$ = Set a CB i as "independent" in t^* when inflation has exceeded 5% or was below 0% in the past seven years. Set "not independent" otherwise.

$\bar{d}^3_{t^*}$ = Set every CB i "not independent" for every t^*

$$\psi_{1,3} = \mathbb{E}(Y^{\bar{d}^1_{t^*}}_{2010}) - \mathbb{E}(Y^{\bar{d}^3_{t^*}}_{2010}), \tag{1}$$

$$\psi_{2,3} = \mathbb{E}(Y^{\bar{d}^2_{t^*}}_{2010}) - \mathbb{E}(Y^{\bar{d}^3_{t^*}}_{2010}). \tag{2}$$

Estimation Method

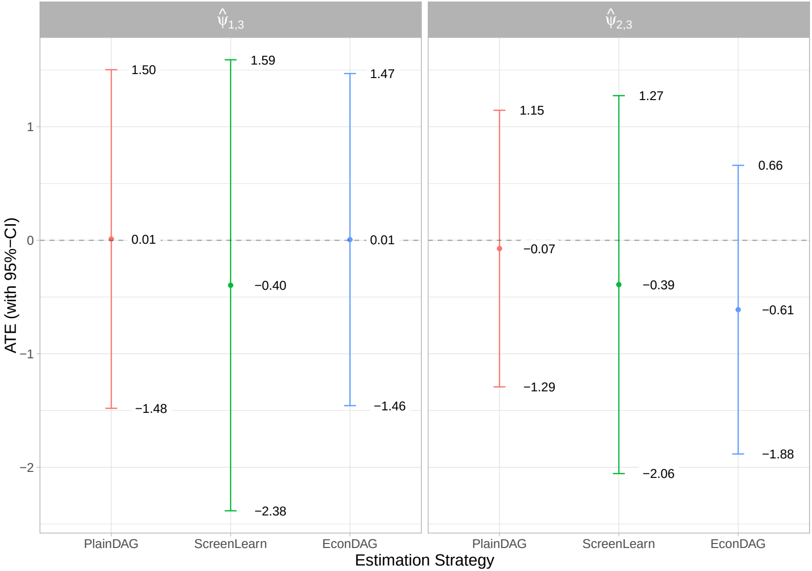
- Longitudinal Targeted Maximum Likelihood Estimation (LTMLE) has been mostly used in the field of bio statistics and epidemiology (van der Laan and Gruber; 2012).
- LTMLE is a doubly robust estimation technique that requires iteratively fitting models for the outcome and intervention mechanisms at each time point.
- LTMLE has the advantage that it can more readily incorporate **machine learning** methods while retaining valid statistical inference.
- Recent research has shown that this is important if correct model specification is difficult, such as when dealing with complex longitudinal data, potentially of small sample size, where relationships and interactions are most likely highly nonlinear and where the number of variables is large compared to the sample size (Tran et al.; 2019).

Which covariates need to be included?

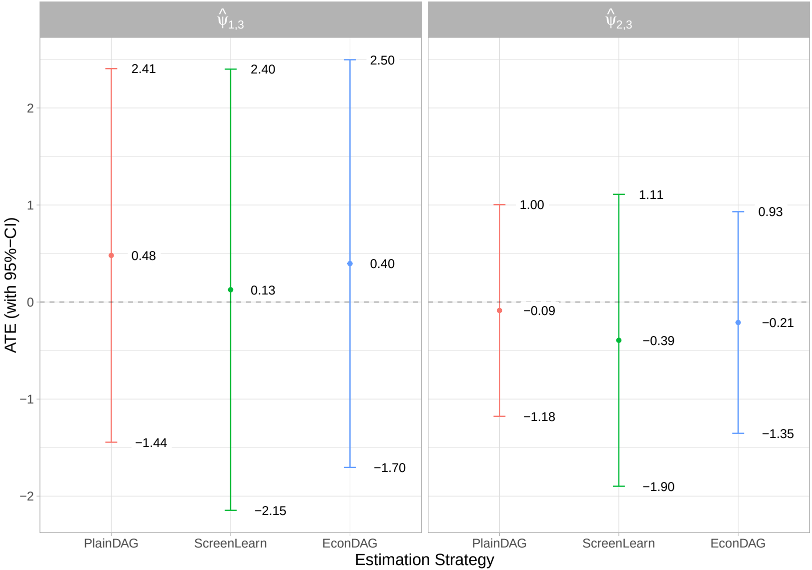
- **Main analysis – PlainDAG:** Models contain only the relevant baseline variables from 1998 that were measured prior to the first CBI intervention.
- Robustness check No. 1 – ScreenLearn: All measured variables are taken into account by the models with respect to the temporal ordering.
- Robustness check No. 2 – EconDAG: Models includes only variables that are measured during a particular 2-yearly transmission cycle, as defined by our DAG.

Results

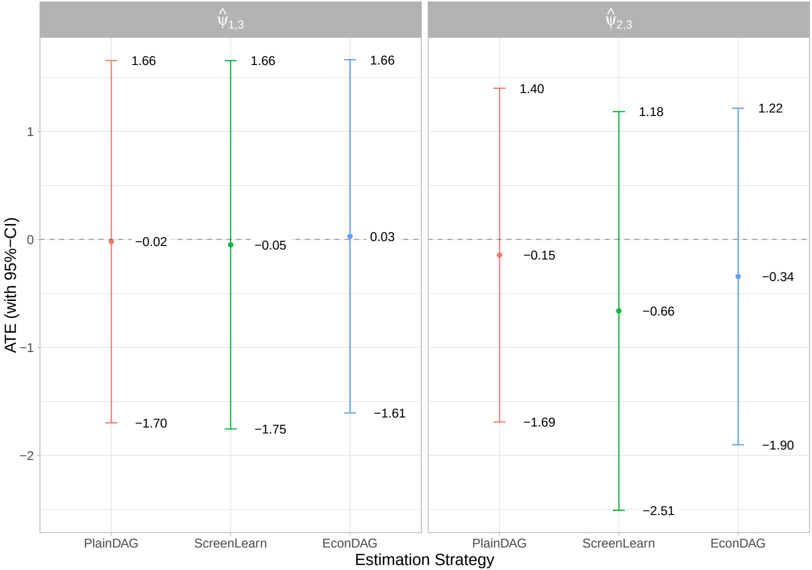
Results: Full Sample (n = 60)



Results: High income (n = 26)



Results: Low income (n = 34)



References

Crowe, C. and Meade, E. E. (2008). Central bank independence and transparency: Evolution and effectiveness, *European Journal of Political Economy* **24**(4): 763–777.

Dincer, N. N. and Eichengreen, B. (2014). Central bank transparency and independence: Updates and new measures, *International Journal of Central Banking* **10**(1): 189–259.

Tran, L., Yiannoutsos, C., Wools-Kaloustian, K., Siika, A., Van Der Laan, M. and Petersen, M. (2019). Double robust efficient estimators of longitudinal treatment effects: Comparative performance in simulations and a case study, *The international journal of biostatistics* **15**(2).

van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions, *The international journal of biostatistics* **8**(1).

IFC-Bank of Italy Workshop on “Machine learning in central banking”

19-22 October 2021, Rome / virtual event

An artificial intelligence application for accounting data cleansing[1]

Pablo Jiménez and Tello Serrano,
Bank of Spain

[1] This presentation 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 central banks and other institutions represented at the event.

Banco de España

AN ARTIFICIAL INTELLIGENCE APPLICATION FOR ACCOUNTING DATA CLEANSING

Central Balance Sheet Data Office (CBSO) Division. SMEs Unit.
Management Systems Division II. Economics and Statistics Systems Unit.
Collaboration: Instituto de Ingeniería del Conocimiento (IIC).

Abstract

The CBSO maintains, among others, a large annual accounting database of the Spanish non-financial corporations sector. By application of automatic rules, the CBSO considers the data of approximately 20% of companies as invalid for studies. Faced with this situation, the dual objective of this proof of concept (PoC) was to explore the application of machine learning techniques, to complement the detection of anomalous cases (outliers) and allow imputation in the absence of data (missing data). In this way, with the help of these new techniques, the CBSO seeks to maximise the number of companies with coherent information and minimise the risk of introducing anomalous questionnaires in the sample. This seems possible in view of the encouraging results of this study.

Keywords: isolation forest, outliers, anomaly detection, missing data, ERC, regression chains, accounting data, imputations, Shapley values, artificial intelligence (AI), machine learning.

JEL classification: C61 y C81

1. Introduction

The CBSO collects annual accounting data from non-financial corporations (NFCs) through two different sources that determine the creation of two very different databases: the Central Balance Sheet Data Office Annual Survey (CBA), which relies on voluntary contributions by reporting firms, where large corporations account for a larger share and where the nearly 11,000 questionnaires that make up the survey are manually refined (standardised questionnaire adapted to the Spanish General Chart of Accounts); and the CBB database which, based on a collaboration agreement between the Banco de España, the Ministry of Justice and the Spanish Association of Property and Mercantile Registrars (CORPME), receives the annual accounts deposited in the Mercantile Registers each year by Spanish companies in standardised models and where small and medium-sized enterprises (SMEs) are widely represented.

The data obtained from the Mercantile Registers are used to check the information held and provide information on a large sample of NFCs. They enable population totals to be inferred and make it possible to monitor NFCs which are underrepresented in the database built on the voluntary contributions of the CBSO reporting firms. The CBSO thus holds data on approximately one million firms for each financial year, of which more than 800,000 may ultimately be used for the preparation of studies, once the various automated data quality and consistency processes have been carried out.

There are two main reasons for the low quality of around 20% of the questionnaires determined by the automated validation system applied in the CBB sample:

- Data mismatches, due either to errors in the recording of values or to missing data.

- Data inconsistencies from a logical-accounting standpoint, such as, for example, high data variance between two consecutive years that cast doubt on their comparability.

Given that the data of approximately 20% of companies are being rejected as low quality, and to gain a more accurate picture of the population of NFCs, the PoC aims to meet two different objectives through the use of machine learning techniques applied to the CBB sample: (1) imputation of missing data; and (2) detection of anomalous questionnaires (outliers).

2. PoC objectives

2.1. Imputation of missing values

In each questionnaire there are certain information headings that are broken down in turn into detailed information; if these sections are not completed or are completed incorrectly the questionnaires are considered invalid. Appropriate imputation of the missing information would enhance the final quality.

For this test, four accounting items were selected whose amount is generally reported but whose breakdown is often incomplete. One of these items is "short-term debts", which consists of three **addends**: debts with credit institutions, finance lease creditors and other debts. The algorithm must fill in the missing addends, **subject to the restriction that their sum matches the total**.

There are numerous ways to impute values: imputing the mean, the median, regressions, moving averages, etc. The use of machine learning techniques seeks to ensure that **the imputation is neither linear nor pre-defined** or, in other words, that the imputed data are not biased by the aggregate chosen to obtain the supposedly analogous values, but that Artificial Intelligence evaluates the complete set of companies and learns from their characteristics to determine which imputation is correct. In this way, **no human decisions affect the data,**

introducing biases, pre-defined views of how each company should look or pre-defined functional forms to which the data must adapt. The main idea is not to make a priori assumptions.

The **ensemble of regressor chains** method (explained below) was chosen for the imputation of missing values.

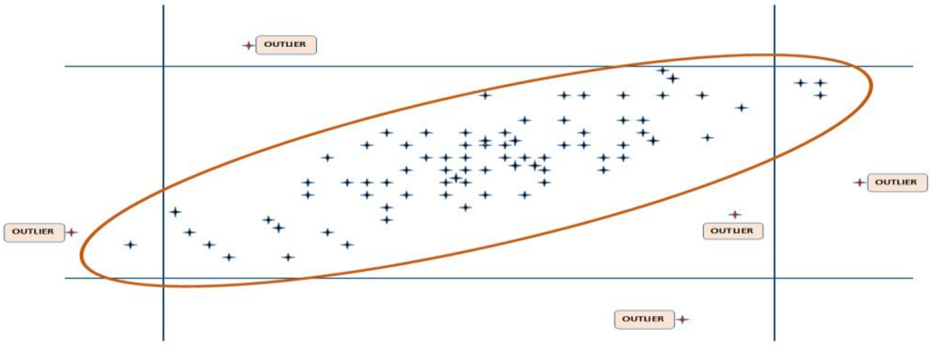
2.2. Detection of outliers

Value anomalies are commonly analysed independently of other variables, that is, from a one-dimensional standpoint. However, in this study an algorithm was used that captures the structure of the data. A value might be anomalous if considered individually, but it might not be when the **specific structure** of the data set is considered.

In Figure 1, the red points correspond to corporations that appear anomalous from a one-dimensional perspective because they lie outside the rectangle that marks the limits of that perspective. But they would not be outliers if observed **multidimensionally** because they remain within the normal structure of the point cloud. Similarly, there could be instances where, despite being within the limits set for the detection of outliers, the points are distant from the densely populated areas of the problem space and are, therefore, relatively anomalous.

The points that are outside the ellipse would be considered anomalous.

Figure 1: Multidimensional outliers



For the sake of clarity, the figure shows two dimensions, although the real problem addressed in the PoC has 97 dimensions (corresponding to the 97 variables used).

The method chosen for detection of outliers was an **Isolation Forest** (explained below).

As a by-product of the study, hidden patterns were discovered that are not visible using other more common statistical techniques. **Shapley values** or ratios**, the usual basis of the XAI (eXplainable AI)**, stand out for these purposes.

3. Characteristics of the microdata, selection and preparation of the sample

3.1. Data pre-processing

To enable the algorithms used to reach a solution, it is essential to work with as **many instances as possible**, use **comparable formats** and reduce the number of variables or **dimensional space** of the problem, without losing crucial information. This pre-processing of the data was carried out taking advantage of the CBSO's **expertise** in this field.

The CBB questionnaires from the Spanish Mercantile Registers may be one of two types, depending on the level of information they contain: normal or abbreviated. For this study only the abbreviated format was considered, since this is the one generally used by SMEs when they file their accounts. Accordingly, large corporations were excluded from the scope of the PoC. The abbreviated questionnaire format consists of four parts: some identification data, balance sheet data, profit and loss account data and model financial statements.

In order for the data to be comparable, the sample had to be **standardised**. This was done by dividing the Balance sheet variables by Total assets and the Profit and loss account variables by Net turnover. Likewise, the Employment variable was divided by Net turnover, in order to normalise it, although this makes no sense from an accounting standpoint.

The **number of dimensions** also had to be reduced: starting from more than 3,000 variables, linear relationships were eliminated (some being the result of the sum of others). Finally, the 97 most significant variables for studying a company, according to the accounting experts' criteria, were selected.

In addition, various **auxiliary variables** were generated to represent certain information that is useful when studying a company's accounts:

a. New variables containing the average values of each field in the last one, two, three, four and five years. These variables will provide the fundamental historical information when predicting a non-imputed value.

b. Age of the company, calculated from the date of the first questionnaire completed.

c. Number of different sectors reported by the company in its history.

d. Number of different large sectors reported.

Finally, in order to **classify the instances** according to their quality, null values owing to lack of information had to be distinguished from those that denote zero.

This process consisted of checking the values of the variables with the sum of their breakdowns: null values that participate in a correct sum were considered zero values; all others were considered missing. Instances were thus divided into three groups:

1. **Perfect** questionnaires: instances with no missing values, considered suitable for study according to the CBSO's automated debugging rules.
2. **Low quality** questionnaires: instances with no missing values, considered unsuitable for study.
3. **Missing data** questionnaires: instances with empty values.

3.2. Data selection criteria

In summary, the criteria adopted for selection of the questionnaires were:

a. Abbreviated subtype CBB questionnaires from 2008 to 2017 (the last complete year when the PoC started) from which 97 variables were selected:

   - 94 accounting items on the balance sheet and profit and loss account corresponding to the current year (the variables from the previous year were discarded).
   - Total average employment of each company.
   - Two sector variables: large sector of activity and 2-digit NACE code.

b. Both perfect and low quality questionnaires were included.

c. Non-standardisable instances were excluded, that is, instances where net turnover or total assets were equal to zero.

d. Other instances that the CBSO discards for different reasons (their main variables are blank, they have high negative values in positive variables or financial sector instances) were also eliminated.

Altogether, **more than 6.2 million questionnaires were included in the PoC**, of which, according to the groups explained above, 5.3 million were 'perfect', 0.5 million were 'low quality' and 0.5 million were ‘missing data’.

4. Methods applied

4.1. Methods applied to detect anomalous observations

To begin with, different methods were proposed to undertake the project. Several common techniques start by considering that the data come from normal distributions, others do not address the problem of the joint structure of the data, others require some type of initial assumption, others overlook the problem of data imbalance (by definition anomalies represent only a tiny percentage of the total data), etc.

Some of the techniques initially considered were: PCA (Principal Component Analysis), Mahalanobis distances, KNNs (K Nearest Neighbors), K-means, etc.

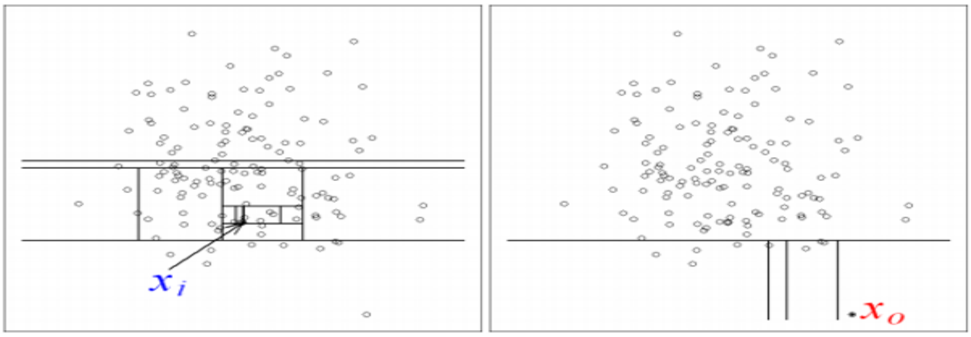
After applying some of these techniques to the data, analysing them and discussing the results, the **Isolation Forest** technique was finally chosen. This is an unsupervised algorithm that consists of "cutting" the space that houses the n-dimensional points by means of random secant planes of dimension n-1 (see Figure 2).

The main idea is that the more cuts it takes to isolate a point, the less anomalous that point is. From the opposite perspective, if a single cut is able to isolate a point, that point is far from the rest, therefore it is anomalous.

Figure 2 shows, in 2 dimensions, that isolating point x_i requires more cuts than isolating point x_0.

Image 2: Isolation Forest

Source: Fei Tony Liu, Kai Ming Ting and Zhi-Hua Zhou (2008), "Isolation-based Anomaly Detection" (page 4).



Since the cuts are random, many are executed and the average number of cuts needed to isolate each point in successive attempts is calculated. This mean, normalised between 0 and 1, is the anomaly score, with 1 being the maximum anomaly indicator and 0 the minimum anomaly indicator.

The formula for each point x is: **score = 2** ^$(-\boldsymbol{E}(\boldsymbol{cuts}(\boldsymbol{x})) * \boldsymbol{standarisation})$

The Isolation Forest technique has a problem when there are unreported variables. In those cases, the instances with unknown values in the variable by which the space is being subdivided cannot be classified in either of the two resulting subspaces. To deal with these cases, IIC created and implemented the **Missolation Forest** algorithm ("missing" + "isolation"). This algorithm takes into account the anomaly score that would be assigned to the instance if it were in each of the two subspaces, giving as a result the average of those values weighted by the probability of it being in one or the other subspace, according to the number of instances in each subspace.

4.2 Methods applied for imputation of values: ensemble of regressor chains (ERC)

After trying other methods such as self-similar neural networks (variational autoencoders) and MICE (multiple imputations) and obtaining a lower level of accuracy, an ERC was chosen.

Training this algorithm consists of randomly permuting the set of target variables (Y3, Y4, Y1, Y2) and constructing a regression for the first variable (Y3). In the next step, another regression is built for the next variable (Y4), but including as another regressor the result of the first regression (the estimate of Y3) as an added regressor for Y4. Thus, successively, **the regressions are "chained".**

In this particular case, the target variables are the fields to be completed: short-term debts with credit institutions, finance lease creditors and other debts.

Naturally, the results depend on the order obtained in the permutation. In our example, since Y3 is estimated before Y4, the effect that Y4 could have on Y3 does not appear. To avoid this problem, k groups of observations (bootstrap) and different permutations of the target variables are used, and the final prediction is computed as the **mean of the k individual predictions for each target variable**.

At the initiative of IIC, each regression was run through a **random forest** algorithm comprising a thousand trees and with randomly selected explanatory variables.

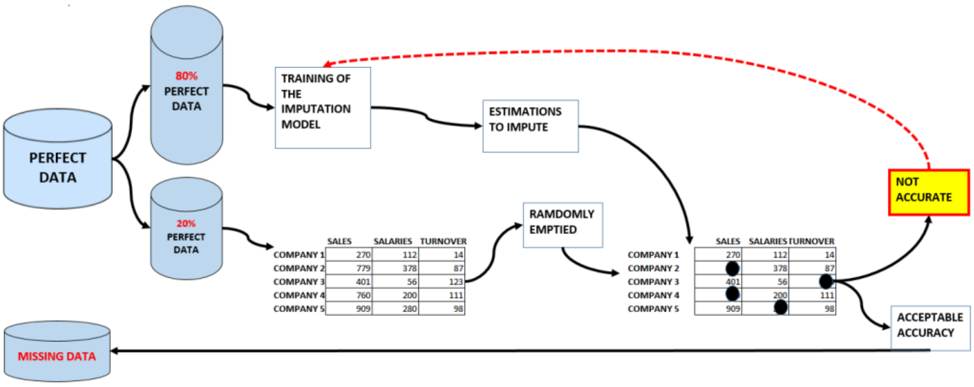
According to certain interpretations of this algorithm in the literature, it could be compared to a deep learning neural network in which the layers of the network are replaced by random regressions.

Since this algorithm is based on permutations, it is **computationally expensive**, especially in this case owing to the complexity added by the random forests used for each regression. Accordingly, it was considered appropriate to cut the number of observations to be used to 240,000 instances. Even so, it is a slow process to train.

To obtain an adequate prediction model, the algorithm was applied to 80% of the questionnaires which, according to the arithmetic logic procedures, were considered perfect ('training set’). The remaining 20% formed the test set, in which some of the data (perforation) were randomly emptied in order to evaluate the quality of the imputation carried out at a later stage. The trained prediction model was applied to the missing forms to estimate the relevant value to be imputed.

Figure 3: Training cycle and imputation

Source: Own elaboration.



	SALES	SALARIES	TURNOVER
COMPANY 1	270	112	14
COMPANY 2	779	378	87
COMPANY 3	401	56	123
COMPANY 4	760	200	111
COMPANY 5	909	280	98

	SALES	SALARIES	TURNOVER
COMPANY 1	270	112	14
COMPANY 2	●	378	87
COMPANY 3	401	56	●
COMPANY 4	●	200	111
COMPANY 5	909	●	98

5. Analysis of results

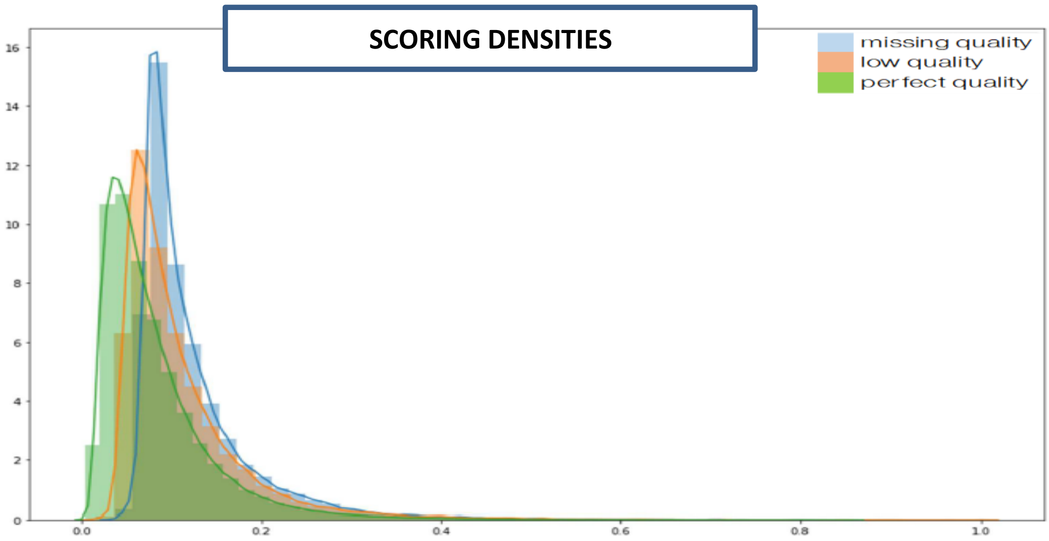
5.1. Results obtained in the detection of anomalous questionnaires

The **degree of anomaly** is indicated by a score between 0 (not anomalous) and 1 (highly anomalous). Since the concept of “anomaly” is not dichotomous, to put the algorithm into production a limit beyond which an observation is considered anomalous must be accepted. According to this distribution, a limit of 0.25 would render 5% of the observations anomalous.

The degree of anomaly estimated by the algorithm is similar, on average, to what the CBSO had been classing as "low quality". The companies which according to the CBSO had perfect quality show a low degree of anomaly (green histogram in Chart 1), those which according to the CBSO had low quality show slightly more anomaly (orange histogram) and those which lacked data show the highest degree of anomaly (blue histogram). This suggests that the algorithm captures the meaning of the task to be performed, at least on average, and that the path followed in the PoC may be the correct one.

Chart 1: Distribution of anomaly scores

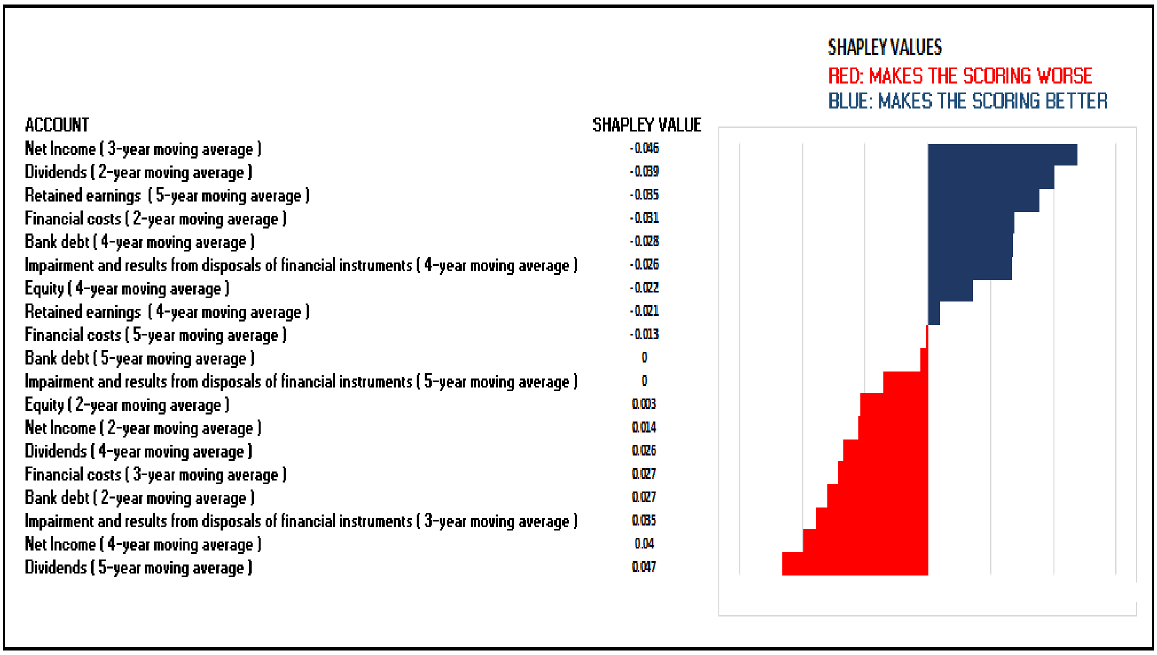
Source: IIC.



One way to dive into the origin of the anomalies is by using the **Shapley values**. These provide information on the extent to which each accounting item affects the anomaly score of each company. As Chart 2 shows, the "3-year moving average of dividens" is the item that most affects the degree of anomaly.

At the beginning of this PoC, the power of these values was not taken into consideration. Accordingly, from the start, the PoC was not designed to capture their explanatory power in an ideal way. For example, it was decided to use moving averages for each variable; this makes interpretability difficult because it disguises these effects within the moving averages themselves. To be more explicit, it could be the case that, for the same variable, the 3-year moving average was significantly positive but the 2-year moving average was significantly negative. They would thus be contradicting each other, and although they do not actually contradict each other, this complicates the interpretability.

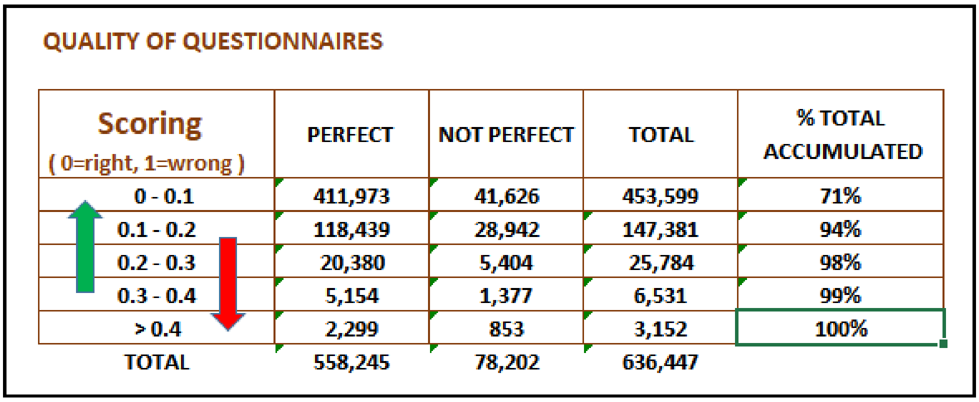
Chart 2: Shapley values for the group of companies with the highest anomaly scores
Source: IIC and own elaboration.



ACCOUNT	SHAPLEY VALUE
Net Income (3-year moving average)	-0.046
Dividends (2-year moving average)	-0.039
Retained earnings (5-year moving average)	-0.035
Financial costs (2-year moving average)	-0.031
Bank debt (4-year moving average)	-0.028
Impairment and results from disposals of financial instruments (4-year moving average)	-0.026
Equity (4-year moving average)	-0.022
Retained earnings (4-year moving average)	-0.021
Financial costs (5-year moving average)	-0.013
Bank debt (5-year moving average)	0
Impairment and results from disposals of financial instruments (5-year moving average)	0
Equity (2-year moving average)	0.003
Net Income (2-year moving average)	0.014
Dividends (4-year moving average)	0.026
Financial costs (3-year moving average)	0.027
Bank debt (2-year moving average)	0.027
Impairment and results from disposals of financial instruments (3-year moving average)	0.035
Net Income (4-year moving average)	0.04
Dividends (5-year moving average)	0.047

In practice, using this automatic automated algorithm to decide if a questionnaire is anomalous or not means deciding in accordance with Table 1, which states, for example, that 41,626 company questionnaires are considered "not perfect" by the CBSO and yet the algorithm considers them statistically normal.

Table 1: Quality of the questionnaires from two perspectives
Source: Banco de España and IIC. Own elaboration.



QUALITY OF QUESTIONNAIRES

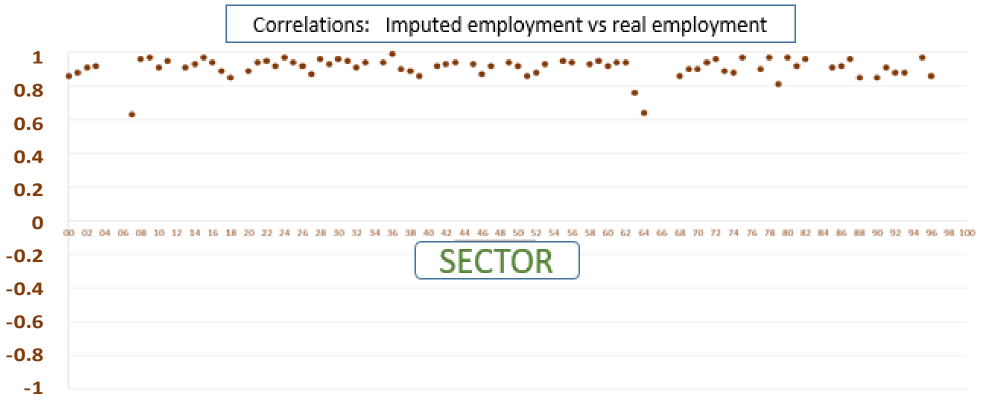
Scoring (0=right, 1=wrong)	PERFECT	NOT PERFECT	TOTAL	% TOTAL ACCUMULATED
0 - 0.1	411,973	41,626	453,599	71%
0.1 - 0.2	118,439	28,942	147,381	94%
0.2 - 0.3	20,380	5,404	25,784	98%
0.3 - 0.4	5,154	1,377	6,531	99%
> 0.4	2,299	853	3,152	100%
TOTAL	558,245	78,202	636,447	

5.2 Results obtained in the imputation of missing values

As a first approximation to the similarity between the real and the imputed values, the correlation between the two was calculated (grouped by each 2-digit NACE code) for each of the imputed variables.

A correlation close to 1 indicates a high degree of similarity. Chart 3 shows the employment correlation.

Chart 3: Correlations between real and imputed data
Source: Banco de España and IIC. Own elaboration.



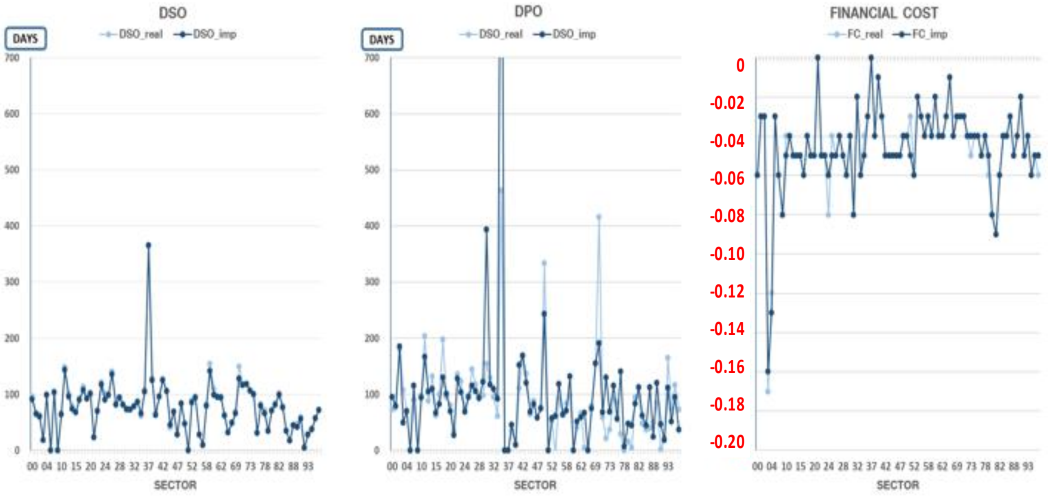
The degree of success is high when the variable considered is "Clients", but it is somewhat lower for "Suppliers" (an item on which there are fewer data).

In general, the worst ¿least successful? imputations correspond to those variables on which there are fewer data. The extreme case is "Called-up share capital", which is usually non-existent in the accounts; the correlation obtained between the real and the imputed data is zero, that is, practically at random. **Our interpretation is that when there are few data, the algorithm cannot learn**. On the contrary, it follows that if more data had been used, the imputation would have been more accurate (it should be remembered at this point that the number of observations had to be limited for the ERC because of the high computational cost).

A simulation of how some accounting ratios would stand if these imputations were used can be seen in Chart 4. For the DSO (days sales outstanding = 365*clients / sales) the adjustment is very good. It is slightly less good for the DPO (days payable outstanding = 365*suppliers / purchases) on account of the problem indicated above with suppliers. For financial costs it is reasonably acceptable.

Chart 4: Comparison of ratios calculated with real vs imputed data

Source: Banco de España and IIC. Own elaboration.



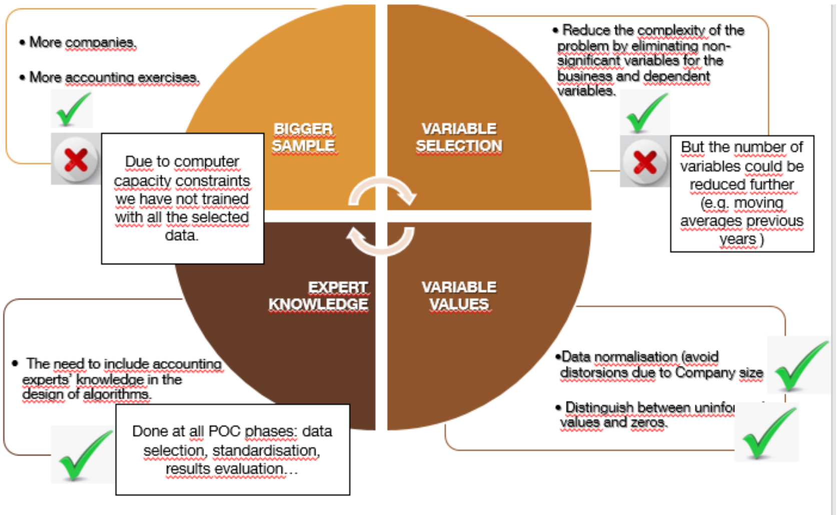
6. Conclusions and lessons learned

- Emphasis should be placed on the **feature selection**. This is normally a primary issue in the entire field of artificial intelligence and there are no magic bullets. In addition, accounting has its own difficulties, including, notably, the high degree of interrelation between the items. After all, the accounts reflect the large balance sheet items on which data have been built up over the years. That particular aspect of the accounting data affects the algorithms, but it can also provide clues that can contribute to their correct formation.
- The problem of the **distinction between empty** (non-existent) **values and values saved as zero** in the database was relevant here because the aim was precisely to impute missing values.
- The failure to normalise the data in the first tests resulted in all large corporations appearing as anomalous. A statistical normalisation would not have solved this problem because the relative scale does not vary. For subsequent analysis usual **“accounting standardisation”** was used: dividing balance sheet items by total assets and income statement items by sales. In consequence, some two million companies were dispensed with because they lacked sales figures, but the results improved substantially.
- The **computational cost** is a major problem. The cost varies significantly **depending on the algorithm chosen**, although this effect is generally exclusive to the training part of the model because, once trained, the application is usually fast. In general, the more data available the greater the accuracy, but that in turn requires more computation. IIC's access to computational resources was essential.

- **Expert accounting knowledge** was key at certain times during the process. Applying mathematically correct solutions that do not take accounting into consideration can produce strange effects that do not go unnoticed by an accountant. For example, in some instances, after imputing the addends of a summation, the summation did not match, meaning that the imbalance had to be distributed among the addends. Initially a linear distribution was made, which gave data that were clearly illogical for accounting purposes but which the system accepted with no problem. Finally the issue was corrected by distributing the mismatch proportionally to the addends. Closer collaboration can save a lot of time in such cases.

Chart 5: Lessons learned

Source: Own elaboration.



- More companies.
- More accounting exercises.

✓ ✗

Due to computer capacity constraints we have not trained with all the selected data.

- Reduce the complexity of the problem by eliminating non-significant variables for the business and dependent variables.

✓ ✗

But the number of variables could be reduced further (e.g. moving averages previous years)

- The need to include accounting experts' knowledge in the design of algorithms.

✓

Done at all POC phases: data selection, standardisation, results evaluation…

- Data normalisation (avoid distorsions due to Company size ✓
- Distinguish between uninfo values and zeros. ✓

BANCO DE ESPAÑA
Eurosistema

INTERNAL USE

AI TOOLS IN OUTLIER DETECTION AND MISSING DATA IMPUTATION

POC DEVELOPED BY BANCO DE ESPAÑA´S CBSO

IFC WORKSHOP ON "DATA SCIENCE IN CENTRAL BANKING"

October 19th-22th, 2021

Pablo Jiménez
Tello Serrano

STATISTICS AND INFORMATION SYSTEMS DEPARTMENTS

INDEX

1. **Introduction. Scope of the 2019 initiative**
2. **ML models**
   - I. Anomalies detection
   - II. Missing value imputation
3. **Analysis of results**
   - I. Anomalies
   - II. Missing value imputation
4. **Lessons learned and next steps**

1. INTRODUCTION. SCOPE OF THE 2019 INITIATIVE

Central Balance Sheet Data Office case study

BANCO DE ESPAÑA
Eurosistema

Questionnaires with accounting information of Spanish non-financial corporations:

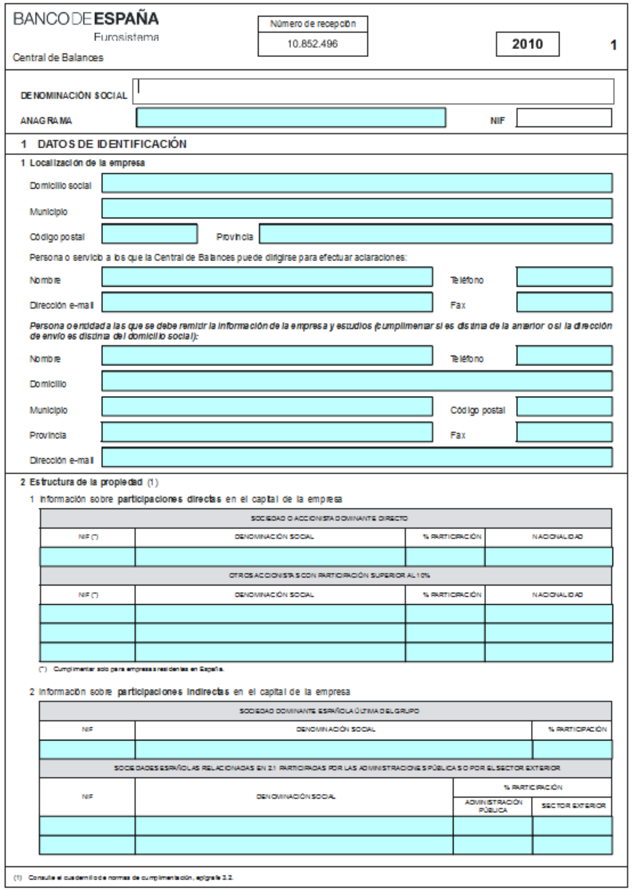
10 exercises x 900,000 companies x 3,000 data

Treated and classified by automatic processes

20% are classified as unsuitable for study

Can AI help us to improve these processes?

- Find alternative patterns to classify the questionnaires: *Case I. Anomaly detection*
- Complete the omitted information: *Case II. Value imputation*



BANCO DE ESPAÑA
Eurosistema
Central de Balances

Número de recepción
10.852.496

2010

1

DENOMINACIÓN SOCIAL

ANAGRAMA

NIF

1 DATOS DE IDENTIFICACIÓN

1 Localización de la empresa

Domicilio social

Municipio

Código postal

Provincia

Persona o servicio a los que la Central de Balances puede dirigirse para efectuar aclaraciones:

Nombre

Teléfono

Dirección e-mail

Fax

Persona o entidad a las que se debe remitir la información de la empresa y estudios (cumplimentar si es distinta de la anterior o si la dirección de envío es distinta del domicilio social):

Nombre

Teléfono

Domicilio

Municipio

Código postal

Provincia

Fax

Dirección e-mail

2 Estructura de la propiedad (1)

1 Información sobre **participaciones directas** en el capital de la empresa

SOCIEDAD O ACCIONISTA DOMINANTE DIRECTO			
NIF (*)	DENOMINACIÓN SOCIAL	% PARTICIPACIÓN	NACIONALIDAD

OTROS ACCIONISTAS CON PARTICIPACIÓN SUPERIOR AL 10%			
NIF (*)	DENOMINACIÓN SOCIAL	% PARTICIPACIÓN	NACIONALIDAD

(*) Cumplimentar solo para empresas residentes en España.

2 Información sobre **participaciones indirectas** en el capital de la empresa

SOCIEDAD DOMINANTE ESPAÑOLA ÚLTIMA DEL GRUPO		
NIF	DENOMINACIÓN SOCIAL	% PARTICIPACIÓN

SOCIEDADES ESPAÑOLAS RELACIONADAS EN 2.1 PARTICIPADAS POR LAS ADMINISTRACIONES PÚBLICAS O POR EL SECTOR EXTERIOR			
NIF	DENOMINACIÓN SOCIAL	% PARTICIPACIÓN	
		ADMINISTRACIÓN PÚBLICA	SECTOR EXTERIOR

(1) Consulte el cuadernillo de normas de cumplimentación, epígrafe 2.2.

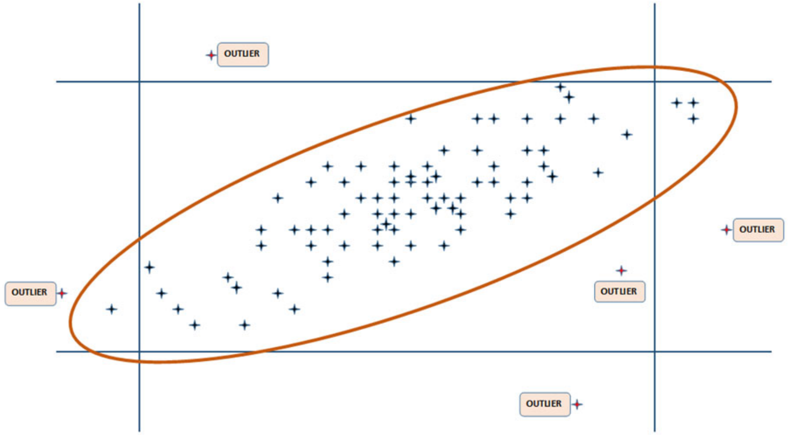
1. INTRODUCTION. SCOPE OF THE 2019 INITIATIVE

POC objective

RECOVER QUESTIONNAIRES FOR STUDY

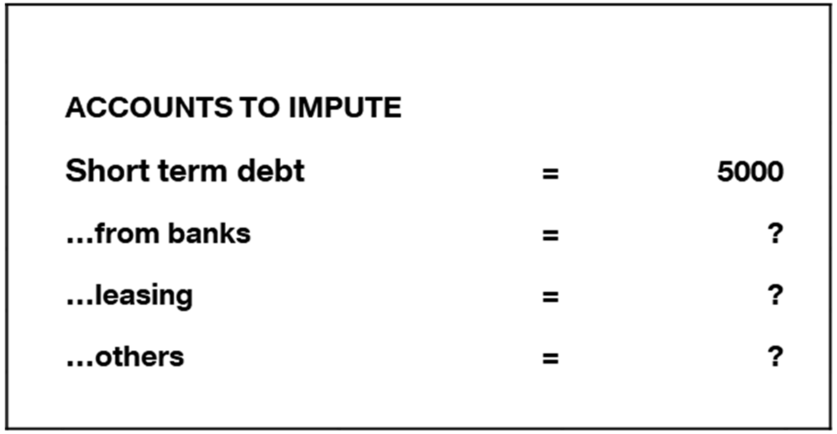
ANOMALY SCORE

Anomaly index valuing n dimensions



VALUE IMPUTATION

(i) Most common imbalances and (ii) employment



ACCOUNTS TO IMPUTE

Short term debt	**=**	**5000**
...from banks	**=**	**?**
...leasing	**=**	**?**
...others	**=**	**?**

1. INTRODUCTION. SCOPE OF THE 2019 INITIATIVE

Data pre-processing

- **Variable selection: 94 accounting keys + employment + activity sector**

- **Accounting standardisation**:

  Divide the P&L fields by net revenues

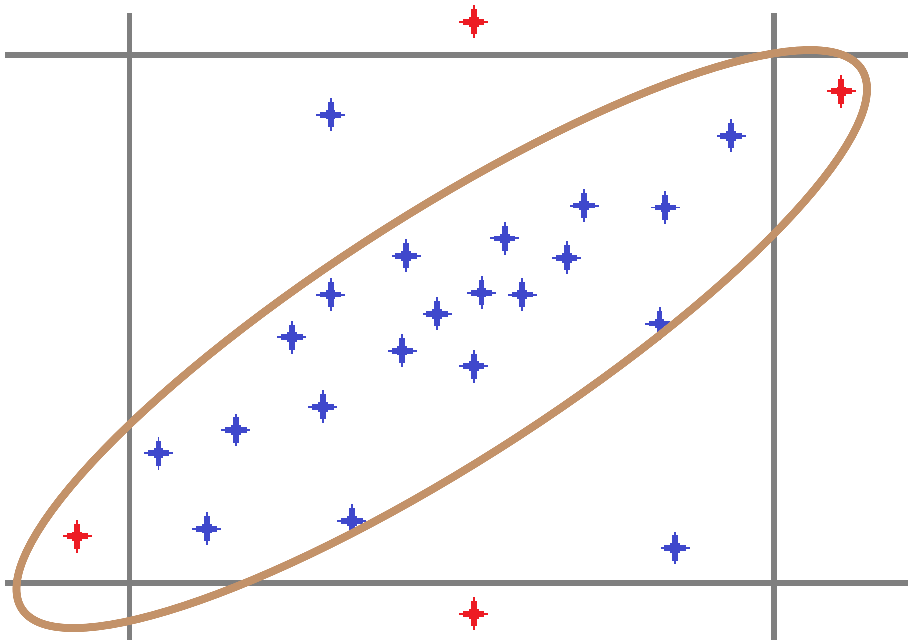
  Divide the *Balance fields* between *Total Assets*

- **Generate new variables**: **Averages of each value in the last 2-5 years, number of declared sectors, company age…**

- **Separate questionnaires according to their quality:**

  - Perfect (5.323.000)
  - Low quality (476.000)
  - *Missing* (469.000)

1) ANOMALIES DETECTION



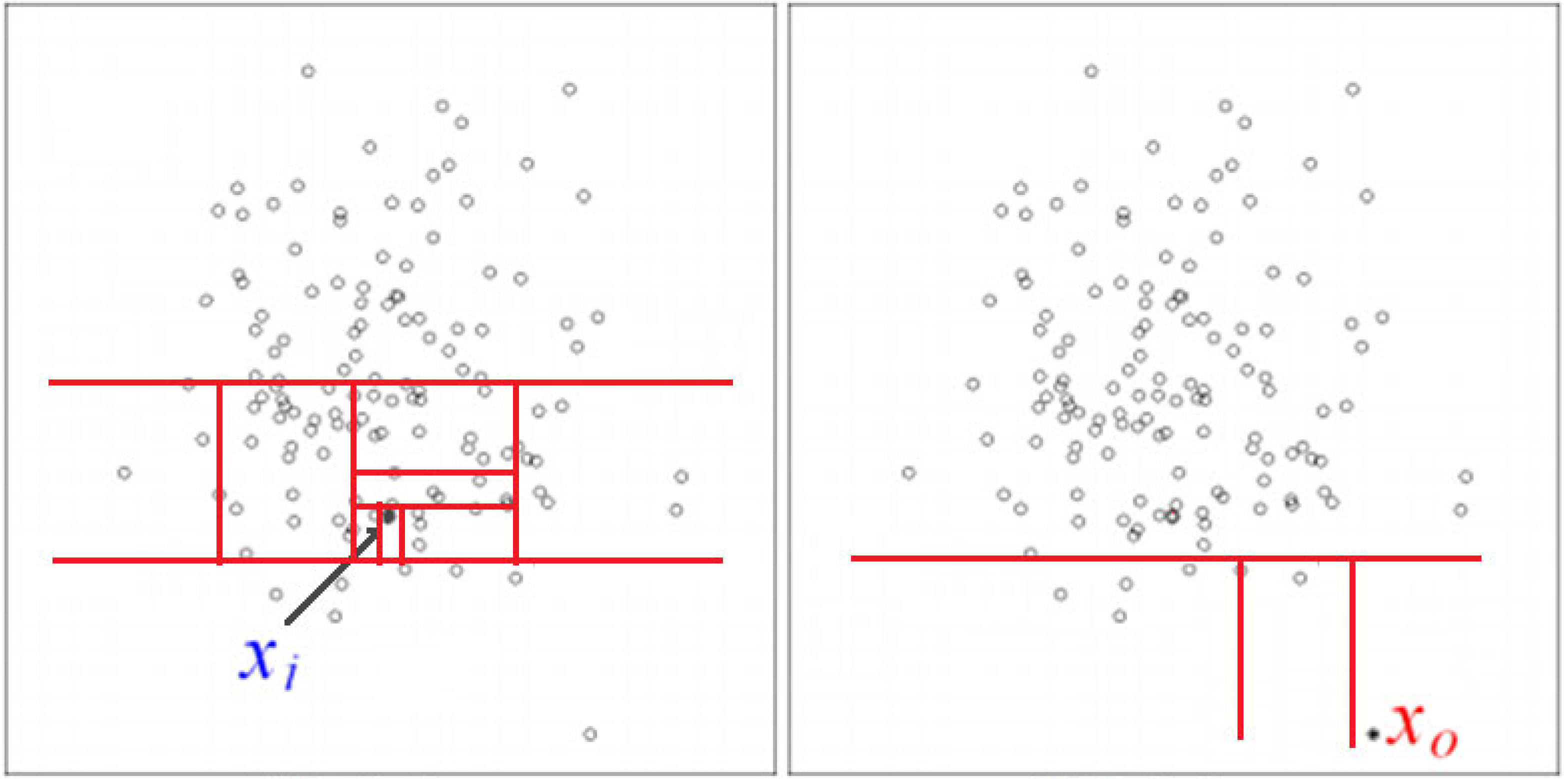
Anomaly score calculation *[0,1]* vs.
Outlier detection *(Yes/No)*

Unsupervised learning

Algorithm: **Isolation Forest**

ISOLATION FOREST

Anomalous instances are easily isolated by random divisions of space



(a) Isolating x_i (b) Isolating x_o

MissolationForest: custom modification of Isolation Forest algorithm to allow estimation of anomaly score when missing values are present in the data

Liu et al – Isolation Forest

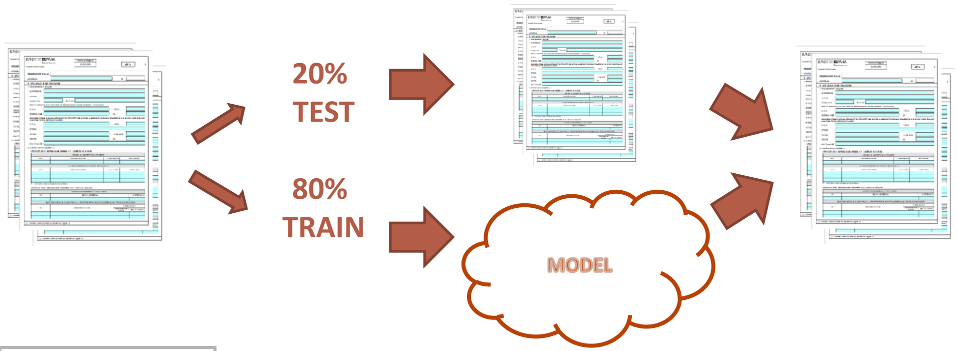
2. ML MODELS

Anomalies detection

2) VALUE IMPUTATION

Supervised learning

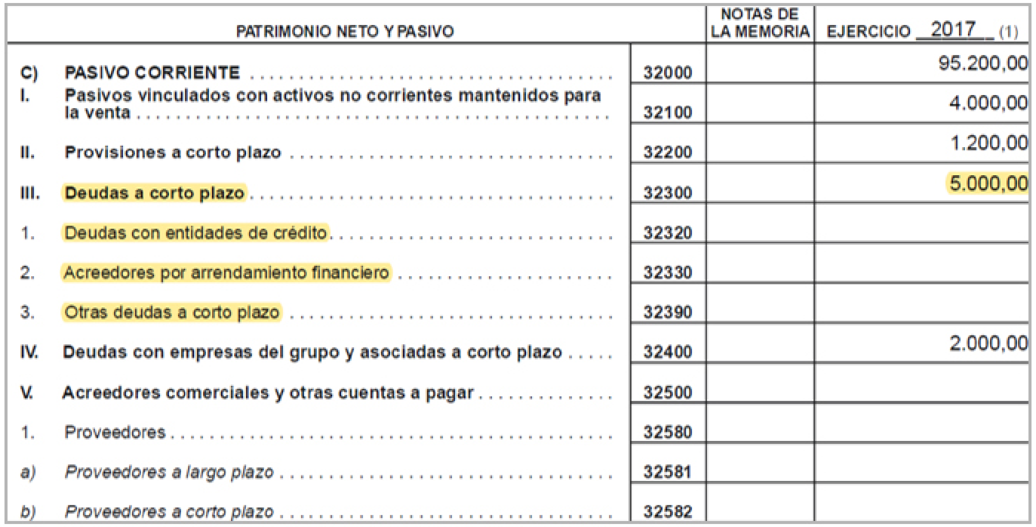
Most common imbalances and employment variables



20% TEST

80% TRAIN

MODEL



	PATRIMONIO NETO Y PASIVO		NOTAS DE LA MEMORIA	EJERCICIO 2017 (1)
C)	**PASIVO CORRIENTE**	**32000**		95.200,00
I.	**Pasivos vinculados con activos no corrientes mantenidos para la venta**	**32100**		4.000,00
II.	**Provisiones a corto plazo**	**32200**		1.200,00
III.	**Deudas a corto plazo**	**32300**		5.000,00
1.	Deudas con entidades de crédito	**32320**		
2.	Acreedores por arrendamiento financiero	**32330**		
3.	Otras deudas a corto plazo	**32390**		
IV.	**Deudas con empresas del grupo y asociadas a corto plazo**	**32400**		2.000,00
V.	**Acreedores comerciales y otras cuentas a pagar**	**32500**		
1.	Proveedores	**32580**		
a)	*Proveedores a largo plazo*	**32581**		
b)	*Proveedores a corto plazo*	**32582**		

Tested algorithms:

- ❑ Variational AutoEncoder (VAE)
- ❑ Multivariate Imputation by Chained Equations (MICE)
- ❑ **Ensamble of Regressor Chains (ERC)**

3.I. ANALYSIS OF RESULTS

ANOMALIES. Scoring vs CB quality

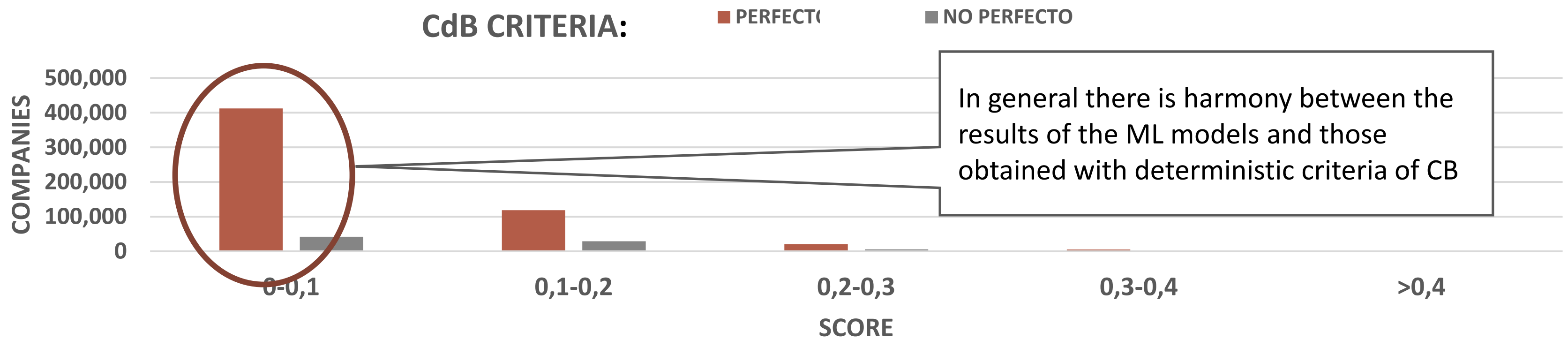
False positives? Analyze to detect possible improvements in our filtering systems

False negatives? Analyse whether or not it is necessary to relax our filtering systems

94% of the questionnaires are concentrated in a range of anomaly between 0 and 0.2

QUALITY OF CBB QUESTIONNAIRES 2017

Scoring IIC (0=Right; 1=Wrong)	PERFECT	NOT PERFECT	TOTAL	% Total accumulated
0-0,1	411.973	41.626	453.599	71,3%
0,1-0,2	118.439	28.942	147.381	94,4%
0,2-0,3	20.380	5.404	25.784	98,5%
0,3-0,4	5.154	1.377	6.531	99,5%
>0,4	2.299	853	3.152	100,0%
TOTAL	**558.245**	**78.202**	**636.447**	



In general there is harmony between the results of the ML models and those obtained with deterministic criteria of CB

3.I. ANALYSIS OF RESULTS

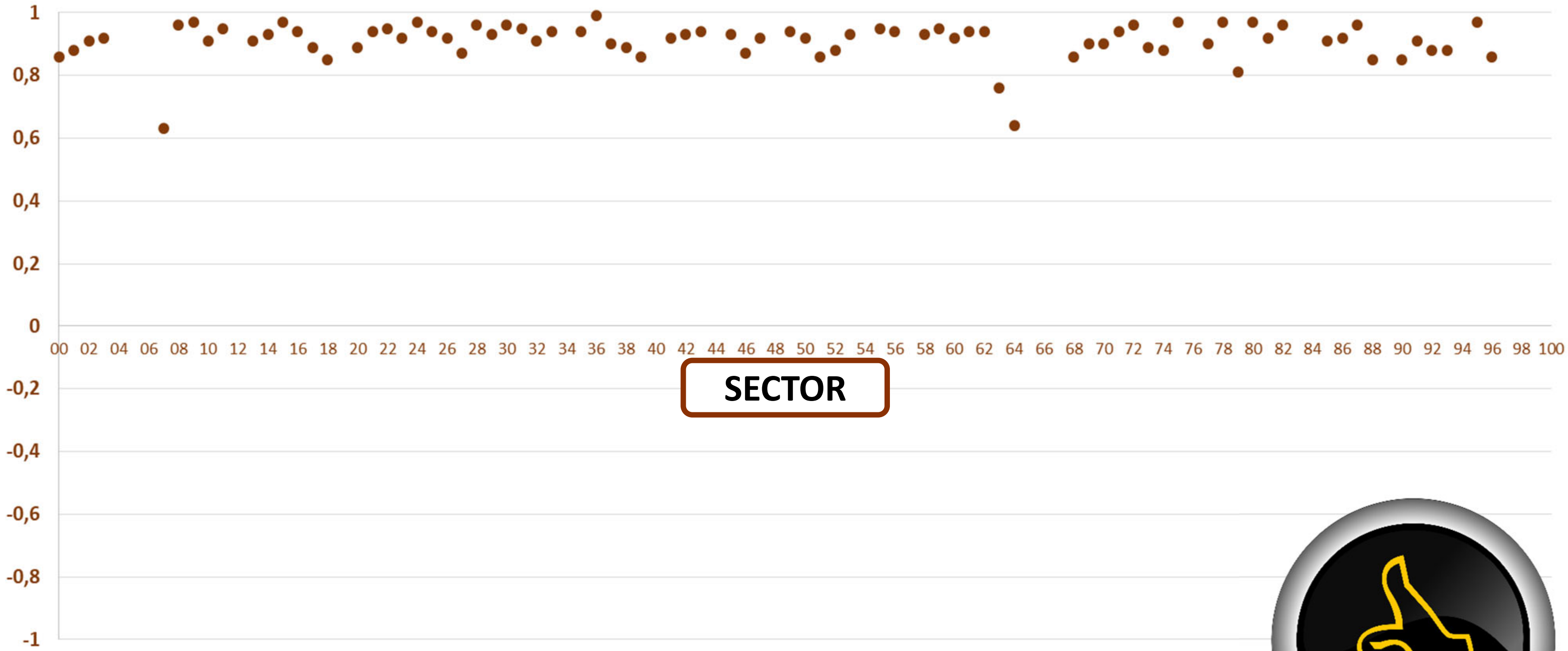
Anomalies. Why should we trust the score?

In summary:

Accepting this score…	…we add or lose these companies	…giving up on these…	…and including these…
0.1	-104,646	-146,272	41,626
0.2	42,735	-27,833	70,568
0.3	68,519	-7,453	75,972
0.4	75,050	-2,299	77,349

3.II. ANALYSIS OF RESULTS

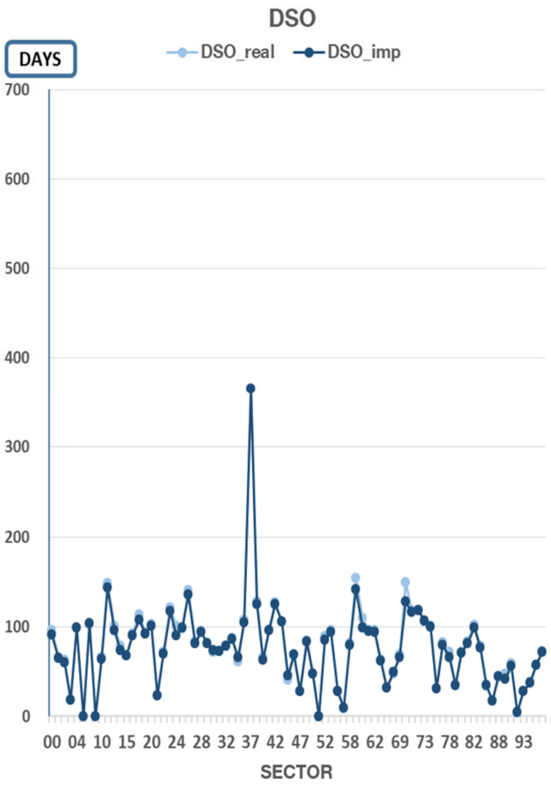
IMPUTATION: Correlations: Imputed employment vs real employment

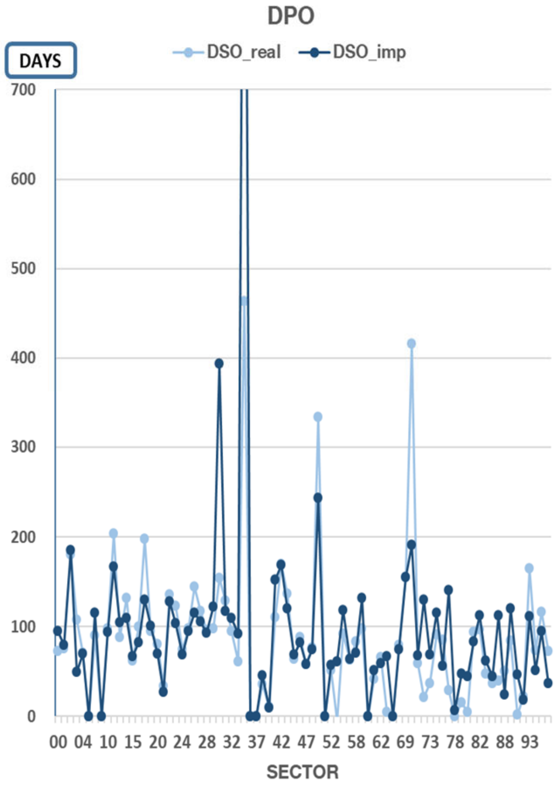


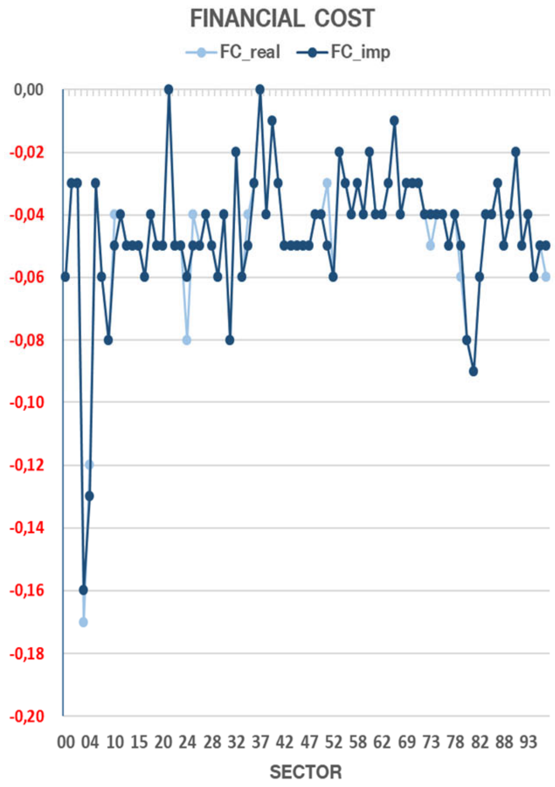
3.II. ANALYSIS OF RESULTS

IMPUTATIONS: Days payable and sales outstanding by activity sector

BANCO DE ESPAÑA
Eurosistema





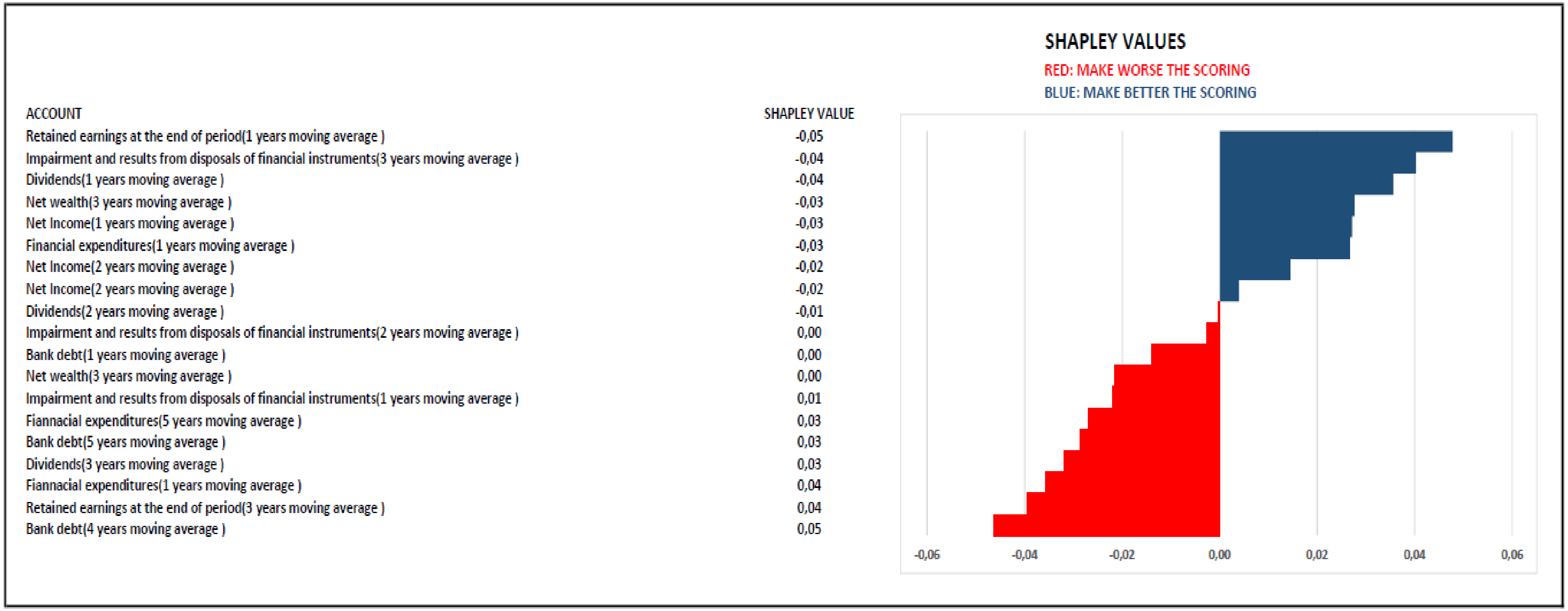


The correlations are acceptable for DSO and financial cost, but are lower for DPO, perhaps because fewer imputations have been made in the supplier key

2.I. ANOMALIES SCORE

Explainability

Shapley values are additive, which allow us to compute the global influence of a variable for the whole dataset or for a subset of the data



ACCOUNT	SHAPLEY VALUE
Retained earnings at the end of period(1 years moving average)	-0,05
Impairment and results from disposals of financial instruments(3 years moving average)	-0,04
Dividends(1 years moving average)	-0,04
Net wealth(3 years moving average)	-0,03
Net Income(1 years moving average)	-0,03
Financial expenditures(1 years moving average)	-0,03
Net Income(2 years moving average)	-0,02
Net Income(2 years moving average)	-0,02
Dividends(2 years moving average)	-0,01
Impairment and results from disposals of financial instruments(2 years moving average)	0,00
Bank debt(1 years moving average)	0,00
Net wealth(3 years moving average)	0,00
Impairment and results from disposals of financial instruments(1 years moving average)	0,01
Fiannacial expenditures(5 years moving average)	0,03
Bank debt(5 years moving average)	0,03
Dividends(3 years moving average)	0,03
Fiannacial expenditures(1 years moving average)	0,04
Retained earnings at the end of period(3 years moving average)	0,04
Bank debt(4 years moving average)	0,05

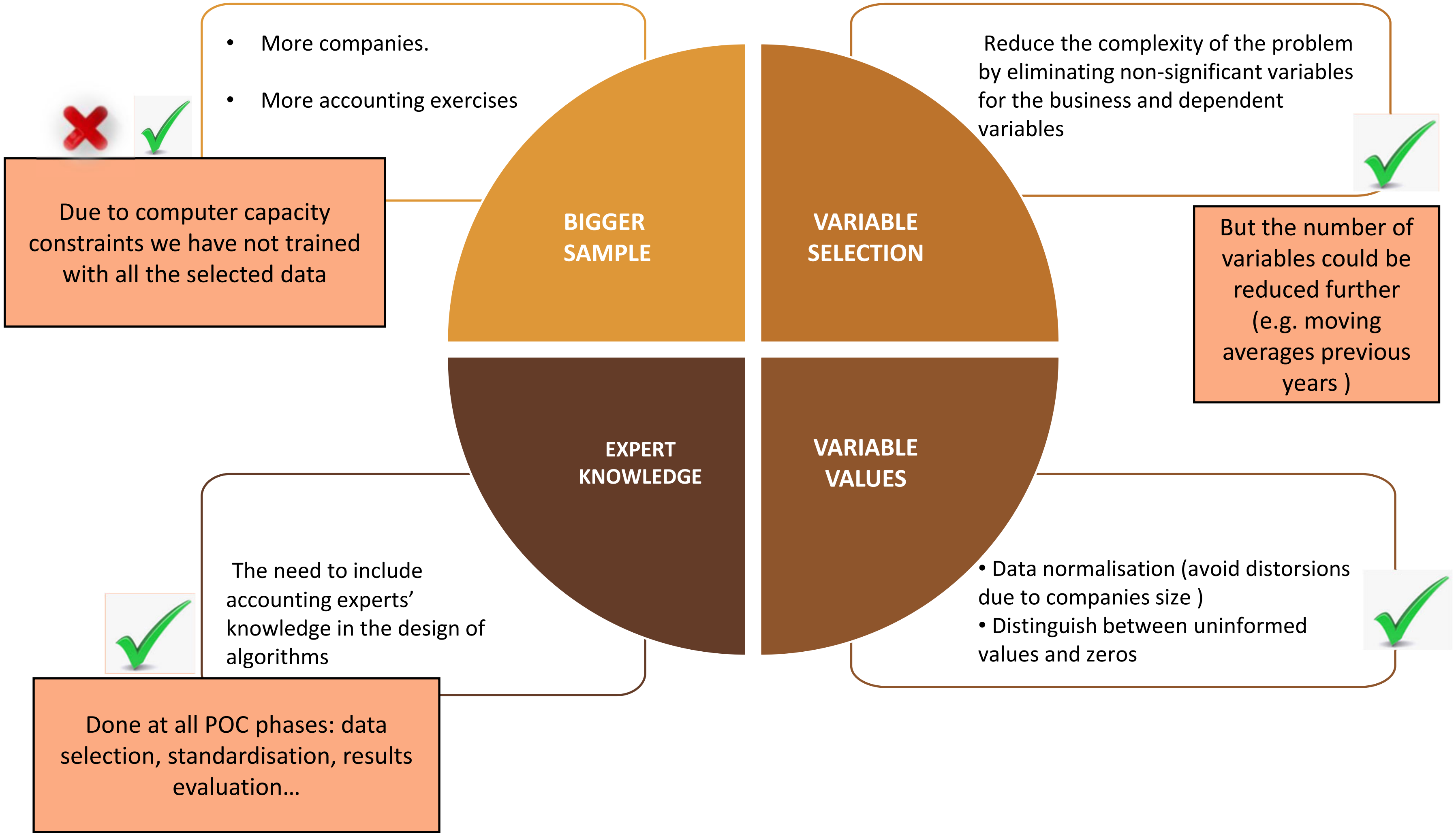
SHAPLEY VALUES

RED: MAKE WORSE THE SCORING

BLUE: MAKE BETTER THE SCORING

4. LESSONS LEARNED AND NEXT STEPS

BANCO DE ESPAÑA
Eurosistema



BIGGER SAMPLE

- More companies.
- More accounting exercises

✗ ✓

Due to computer capacity constraints we have not trained with all the selected data

VARIABLE SELECTION

Reduce the complexity of the problem by eliminating non-significant variables for the business and dependent variables

✓

But the number of variables could be reduced further (e.g. moving averages previous years)

EXPERT KNOWLEDGE

The need to include accounting experts' knowledge in the design of algorithms

✓

Done at all POC phases: data selection, standardisation, results evaluation…

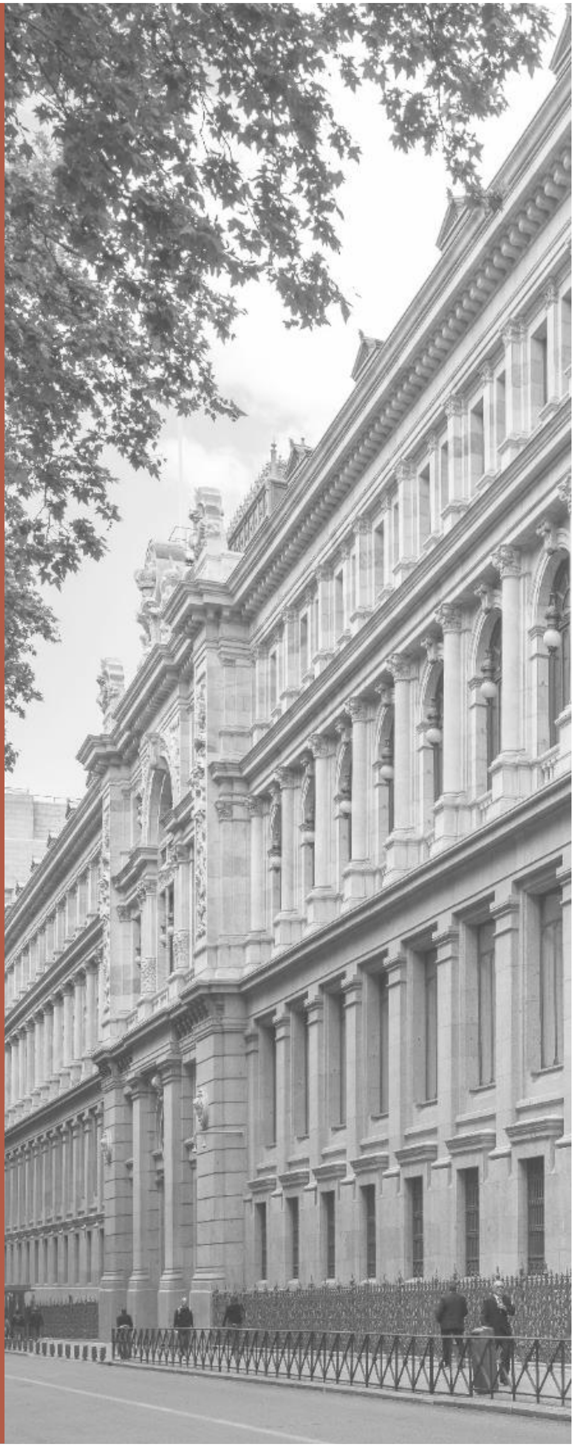
VARIABLE VALUES

- Data normalisation (avoid distorsions due to companies size)
- Distinguish between uninformed values and zeros

✓

BANCO DE ESPAÑA
Eurosistema

THANK YOU FOR YOUR ATTENTION



STATISTICS AND INFORMATION SYSTEMS

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Machine learning for anomaly detection in financial regulatory data[1]

Maryam Haghighi, Colin Jones and James Younker,
Bank of Canada

[1] This presentation 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 central banks and other institutions represented at the event.

Machine Learning for Anomaly Detection on bank regulatory data

October 2021

Maryam Haghighi, Colin Jones and James Younker, Bank of Canada

The views expressed in this presentation are solely those of the authors/presenters and may differ from official Bank of Canada views. No responsibility for them should be attributed to the Bank.

Context

- Rapid growth in data, tools, and analytic techniques to leverage data.

- Opportunity to drive efficiency in our internal operations using non-traditional techniques such as machine learning.

- Transforming to exploit these opportunities.

Key Messages from this presentation

1. We developed a novel method based on machined learning, for detecting anomalies in data from financial institutions.

2. We have operationalized this method to gain efficiencies and better detect anomalies.

3. A robust pipeline built for business users, currently in-use on a daily basis.

Power of Operationalized Machine Learning

60-80 FIs filing returns

From 100's to 10,000's data points/return

26 returns (out of 40) operationalized so far

Millions of datapoints every month

Impossible to examine every single one with traditional ways.
Instead, analysts focused on only a few variables. Even then, *it took significant time*.
Risk that critical anomalies may be missed.

Need high quality data used daily in sensitive economic policy analysis
Over the past year, developed a ML model for anomaly detection.
Now operationalized, running daily @1 am
Detecting anomalies we couldn't detect before + Saving significant time

Formalized with modern Data Science standards: reliable, scalable, fully explainable.

Moving over to cloud environment- data lake

The challenge:

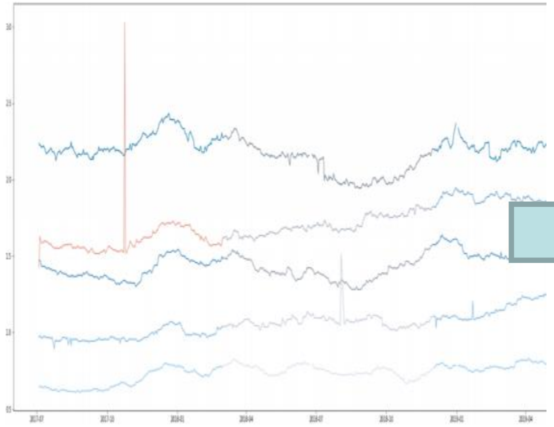
- Financial institutions send us the data (weekly, biweekly, monthly etc.)
- Not every anomaly is an error. How do we know what is an error and what is not?
- Even small numbers of errors can undermine the usefulness of the data
- Traditionally, we used rule-based approaches for anomaly detection.

Our minimum requirements:

1. Function with little labelled data
2. Manage incorrect labelling
3. Avoid reputational risk of false positives
4. Detect anomalies conditional on the activities of that bank
5. Handle regime changes
6. Be useable by non-data scientists
7. Be reliable, automated and scalable

Binary Classification

- Tries to predict if a given time series has at least one anomaly
- Imbalance issue
- Accuracy is not a good performance metric
- Instead use precision (control number of false positive to true positives), recall (identification of true positives) and F_β statistics



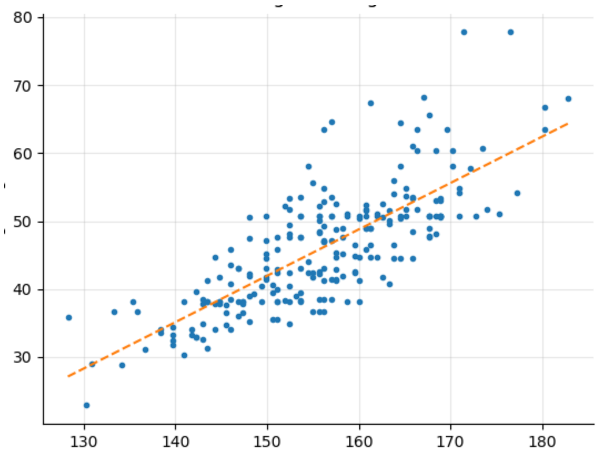
$\{0,1\}$

Questions:

- Null values may or may not be an anomaly
  - Is the bank known to be involved in the specific category, for example in collateral swap market?
- Usually large/small volumes
  - How large is large? How small is small? May vary across financial institutions
- Volatility: How volatile is volatile?
- Spikes: How big of a spike is unusual?

Use of Correlation

- We use correlations between banks and inside the return itself
- Use the data from other FIs that have similar behaviour
  - This multivariate approach allows us to leverage more information as compared to just considering one FI on its own.



- Use the correlation or similarity structure of time series inside the return or across returns.
- Suppose that time series X and time series Y are known historically to be very correlated.
- If one time series behaves significantly differently then the other, we may conclude that the time series in question may be an anomaly.

Two step procedure

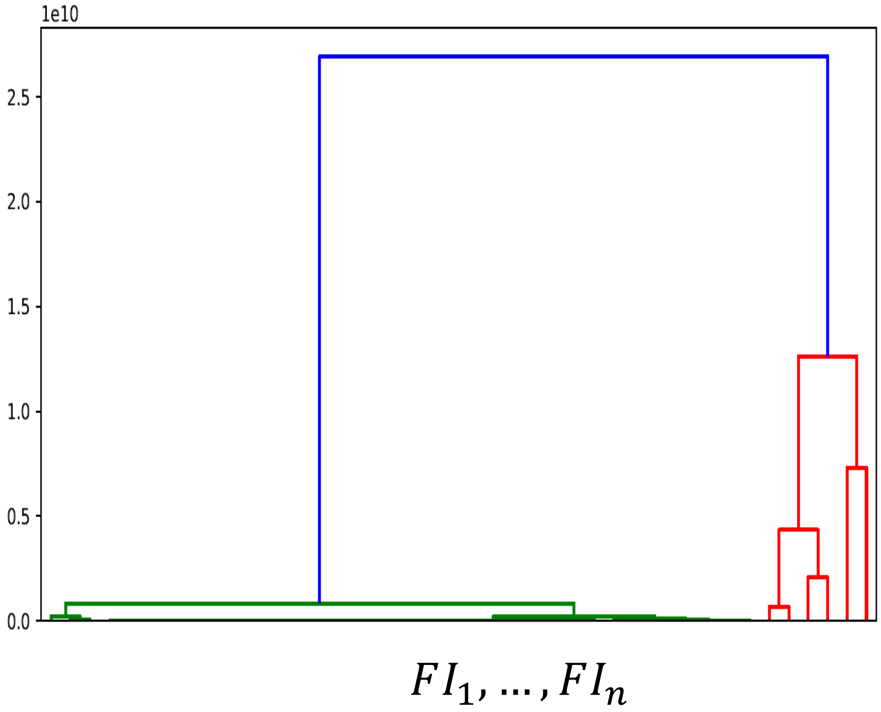
Step 1

- Cluster the Financial Institutions (FIs) based on the raw time series

Step 2

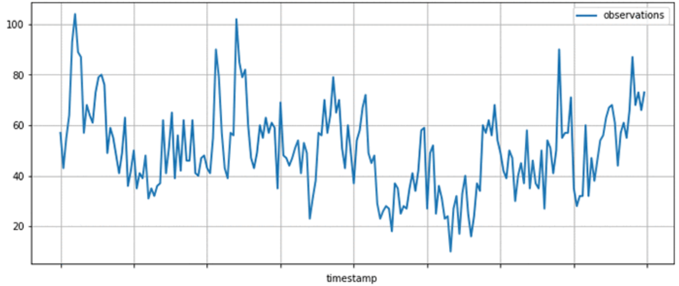
- Implement a supervised ML algorithm using time series features that include covariates from all FIs in the cluster

Clustering (Dendrogram)



Time series features

- Using the time series features approach we considered many features. These include
  - Mean, max of a rolling standard deviation, max first difference, proportion of zeros; for all banks
  - Mean and standard deviation of correlated time series



$$Features = \{F_1, \dots, F_n\}$$

In production

- The model needs to be maintained, documented, governed and managed to ensure that the model is still performing at its best over time.
- Code requirements (run on several computers, and documented packages required to run ML code), troubleshooting and error management, step-by-step analysis, use of Git and other version control technologies, production schedules, and a centralized tracking log for communication with the FIs.
- Periodic review of training sets, metrics of performance need to be continually examined to further ensure the efficacy of the model.
- Moving the project over cloud environment- ML Ops

Running the model on each FI individually

Worse Performance

Running the two-step model

Better Performance

Conclusion

- Novel method of multivariate anomaly detection
  - Clustering time series by banks that have a strong correlation and then applying a supervised classification ML algorithm.
  - Used correlation among time series within a given return to enhance the detection algorithm.

  These uses of correlation amplified our detection power compared to evaluating each time series on its own.

- This model is actively in production mode. Running daily at BoC, Detecting anomalies otherwise not detectable before + Saving significant time
- reliable, explainable, scalable and robust.
- Moving the model over to cloud environment via MLOps framework.

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Supervisory letter writing app: expediting letter drafting and ensuring tone consistency[1]

Joshua Tan, Chi Ken Shum and Mohd Akmal Amri,
Central Bank of Malaysia

[1] This presentation 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 central banks and other institutions represented at the event.

Supervisory Letter Writing App: Expediting Letter Drafting and Ensuring Tone Consistency[1]

Joshua Tan[2], Chi Ken Shum[3] and Mohd Akmal Amri[4]

Abstract

The Central Bank of Malaysia issues supervisory letters to all financial institutions under its purview on an annual basis. The supervisory letters communicate the supervisory ratings assigned to the respective institutions, the supervisory concerns as well as the remediation actions required. With more than 120 institutions under purview, writing and reviewing every supervisory letter consumes hundreds of costly work hours. In an effort to improve both the efficiency in the writing process as well as the consistency in the communication tone of supervisory letters, we develop a web application that comprises two main features: i) Tone Analysis; ii) Sentence Search. With Tone Analysis, we empower supervisors to better understand and calibrate the tone of their drafted letters commensurate with the intended corrective measures to be taken by financial institutions. We utilize a Transformer-based Natural Language Processing (NLP) model called DistilBERT coupled with optimizations using ONNX Runtime and quantization to perform multi-class text classification on every sentence in a letter. Each sentence is classified into one of four ordinal tone classes — "Neutral", "Cautious", "Concerned", and "Forceful" with tone ranging from no tone ("Neutral") to the most severe tone ("Forceful"). Using the weighted-F1 metric, we obtain a score of 77.6% for our test data set. With Sentence Search, we enable supervisors to search for sentences extracted from previously issued letters by tone class to expedite the writing process. Supervisors can search for sentences by keywords or by semantic similarity with the latter utilizing an NLP model called Sentence-BERT.

Keywords: Central bank communication, prudential supervision, supervisory letter, natural language processing, sentiment, deep learning, web application.

JEL classification: C55, C80, E58, G28, Y80.

[1] This paper was prepared for the IFC and Bank of Italy Workshop on "Data Science in Central Banking" held virtually on 19-22 October 2021. 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 or of anyone else associated with the Central Bank of Malaysia. The authors are grateful to Daniel Chin Shen Li, Intan Zainura Ahmad and Syairulzila Mohd Yusof for their assistance and useful comments on the paper.

[2] Bank Negara Malaysia; Email: joshua@bnm.gov.my

[3] Bank Negara Malaysia; Email: shum@bnm.gov.my

[4] Bank Negara Malaysia; Email: akmalamri@bnm.gov.my

Contents

1. Introduction

As part of the Central Bank of Malaysia's mandate to preserve the country's financial stability, all licensed financial institutions (FIs) operating in Malaysia are subject to annual supervisory review by the Central Bank of Malaysia, where the safety and soundness of each FI is assessed and determined by their respective supervisors. The supervisors are guided by a robust risk-based supervisory framework where each FI is evaluated based on its risk profile, quality of risk management, sustainability of earnings, as well as strength of capital and liquidity before being assigned with a composite supervisory rating. At the conclusion of their review, supervisors will issue supervisory letters to the respective FIs to communicate their assigned supervisory ratings, supervisory concerns as well as the remediation actions required.

Currently, there are more than 120 FIs under the central bank's purview and therefore, at minimum 120 supervisory letters need to be issued every year. Each supervisory letter is drafted and reviewed thoroughly by the respective supervisors, and it may take weeks before the letter is fit for issuance as it goes through the necessary editing and governance processes. Based on our engagement with supervisors, most supervisors who are relatively new to the position find drafting supervisory letters laborious due to the numerous revisions that occur during the review process. This stringent review process is necessary due to the subjective nature of adjusting the tone of the supervisory letter when detailing supervisory concerns and the required remediation actions. For example, the Central Bank of Malaysia's management would expect a strongly worded letter when conveying regulatory breaches or recurring lapses in control functions. However, having a good grasp of the appropriate words and sentences to use under specific situations when drafting a supervisory letter is predominantly a result of tacit knowledge gained through years of experience. Therefore, inexperienced supervisors often find this aspect challenging.

In addition, as supervisory letters are issued by multiple supervisory departments, there is no effective nor systematic way to streamline the tones used in these letters across the departments. There are a variety of factors that influence the tones used in the supervisory letters. These include the severity of issues highlighted, their sentiment regarding the competency of personnel to whom they address the issue and their personal writing style. While variety in tone is necessary in some cases to address situations unique to certain institutions, having a tool to streamline the way supervisors write letters will help to reduce the wide range of individual writing styles and adopt a more consistent communication strategy. Such communication tone is critical as supervisory letters reflect the central bank's overall assessment of an FI and its human capital. Thus, achieving a precise tone in the letters will establish clear expectations on the subsequent actions the FIs must undertake.

As such, we develop a web application with two main features to address the challenges mentioned above. The first feature of the web application is Tone Analysis. Tone Analysis enables supervisors and management to gauge the tone employed in a letter and determine if it is proportionate and appropriate to the risk profile and severity of the issues highlighted to the FI. The second feature, Sentence Search, expedites the letter drafting process as supervisors can easily query a database of sentences from historical supervisory letters while drafting letters. When used together, these features enhance the supervisory process by improving supervisors' efficiency in drafting supervisory letters and by maintaining a consistent communication approach for the Central Bank of Malaysia.

2. Related Work

2.1 Central Bank and Supervision Domain

Much effort in analyzing central bank communications using Natural Language Processing (NLP) have revolved around sentiment analysis and topic analysis in the context of public-facing monetary policy and financial stability communications. The sentiments are often summarized into an index that is then studied in relation to a set of economic or financial indicators. Examples of such studies include Correa et al. (2017), Jegadeesh and Wu (2015) and Born et al. (2014). At the same time, there has also been increased interest in the broader application of NLP within the supervision domain. Many supervisory authorities, including central banks, have embarked on studies, projects and proof-of-concepts applying new technologies in the realm of supervision, engendering the frequently dubbed term "SupTech". Financial Stability Board (2020) highlights this fact with the numerous case studies of applied NLP for content extraction, risk identification and news-based sentiment analysis in relation to supervisory matters.

Our project focuses on analyzing the sentiment or tone of central bank communications with regulated FIs — a private but crucial stakeholder. In this respect, Bholat et al. (2017) is one notable example that resembles our initiative as they studied the linguistic features of the letters sent by the Bank of England's Prudential Regulation Authority (PRA) to banks and building societies under its supervision. However, unlike Bholat et al. (2017), this paper does not discuss the characteristics of supervisory letters such as their directiveness and formality in depth but focuses on the methodology applied in developing a SupTech tool to facilitate the letter drafting process.

2.2 Data Science Domain

In our web application, the main techniques we employ are text classification (more specifically, sentiment analysis) and sentence embedding. Text classification is an NLP technique used to categorize a sequence of text into groups. It has made significant progress in recent years with state-of-the-art deep learning models leapfrogging traditional machine learning models in performance (Minaee et al., 2021). This advancement is in part due to the groundbreaking Transformers architecture proposed by Vaswani et al. (2017) coupled with transfer learning techniques proposed by Howard and Ruder (2018). While there has been some effort in utilizing text classification to analyze sentiment in central bank communications (Rybinski, 2019), most of the literature studied dictionary-based approaches as seen in Shapiro and Wilson (2019), Hubert and Labondance (2017) and Fraccaroli et al. (2020).

Meanwhile, in the broader data science domain, text classification is commonly used for sentiment analysis. This is evidenced by the various sentiment-related data sets used as benchmarks for the evaluation of state-of-the-art text classification models. At present, a variety of models have achieved top performance across a broad array of tasks (binary or multiclass classification) and data sets (IMDb, Yelp etc.). Some noteworthy models are XLNet (Yang et al., 2019), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), Universal Language Model Fine-tuning (ULMFiT) (Howard & Ruder, 2018) along with their variations.

On the other hand, word and sentence embeddings are dense vectors or numeric representations of a text that, when produced using machine learning models, can capture its semantic meaning. These representations are commonly generated as an intermediate step for text classification models. While they are key components for many NLP models, they can also be used standalone to serve other purposes such as computing cosine similarities for the search of similar sentences. Modern embedding models have evolved significantly, with popular models like word2vec (Mikolov et al., 2013) and Global Vectors for Word Representation (GloVe) (Pennington et al., 2014) being able to create static word embeddings that combine all the different meanings of a word into one vector. Nonetheless, newer models like Embedding Language Model (ELMo) (Peters et al., 2018) and BERT have advanced to generate dynamic word embeddings that change its representation according to the context of a sentence, producing superior results in most cases.

3. Web Application

The Django web application we develop has two main features: i) Tone Analysis; and ii) Sentence Search. The technical details of their implementations are provided in the Methodology section below. In this section, we describe their business use cases.

With Tone Analysis, we empower supervisors to better understand and calibrate the tone of their drafted letters commensurate with the intended corrective measures to be taken by FIs. Supervisors can upload a draft letter and obtain a sentence level tone prediction that is color-coded according to tone for easier analysis. For further insight, they can choose to interpret the predicted tone of a sentence which will display the confidence score of the prediction and highlight keywords that contribute most to the prediction. If a supervisor intends to draw attention to a key risk faced by an FI via a strong directive message but receives a soft predicted tone from the model, he or she may want to re-evaluate the language of the sentence. For example, specific words that are deemed to significantly influence the predicted tone may be edited to improve the intended message severity level. Conversely, if the predicted tone is consistent with the supervisor’s intention, the application may serve as a form of validation before letter issuance. Supervisors can also view the tone distribution of all predicted sentences summarized in a chart. These sentences classified into their respective tones each contribute a specific score that is aggregated and normalized using a formula to arrive at a compound score for the entire letter. The compound score represents the overall tone of the letter and is assigned to a category according to predefined thresholds for document-level comparisons. Regularly using the application to run tone analyses over the course of drafting supervisory letters ensures that the Central Bank of Malaysia, across its multiple departments, communicates to FIs in a consistent manner.

With Sentence Search, we enable supervisors to search for sentences extracted from previously issued letters, categorized by tone to expedite the writing process. As supervisors draft letters, the need to reference past supervisory letters issued to FIs with similar issues or similar ratings often arises. This reference process is necessary to provide context for the issue at hand to supervisors and to subsequently prevent FIs from overreacting or underreacting to sudden changes in vocabulary, writing style and in turn, communication tone in the supervisory letters. Supervisors can either search by keywords, e.g., ”liquidity risk” or by a full sentence. When searching by sentence, the application will extract sentences that are semantically similar to the query sentence even though they may not be identical. For example, given a query “Loan growth has declined due to poor GDP growth”, a possible search result is “Due to worsening economic condition, loan growth has shrunk.”. To perform a search, supervisors simply need to type in the word or sentence of interest in an input box to query the database. An alternative workflow would be the following — after a supervisor has uploaded a letter for tone analysis, he or she may identify specific sentences that do not convey the intended tone and seek to amend them. The supervisor can directly highlight words or sentences from the tone analysis page and click a button to query the database. Once the most similar sentences are presented, supervisors can also expand each search result to read the original letter content for a better understanding of the context in which the sentence was used. Ultimately, these search features enable supervisors to effectively query and reference historical letters, leading to stronger draft letters, thereby reducing the total time spent on amendments as they alternate between author and reviewer.

While these two features aim to improve the letter drafting process for supervisors, we also developed a privileged access dashboard in the web application for administrators. The dashboard provides easy retrieval of past letters and a holistic visualization of overall letter tones alongside simple filtering options for time series and cross-sectional analyses. This view is valuable for monitoring and spotting anomalies in letter tones across FIs of similar risk ratings or within a particular FI across time.

4. Methodology

4.1 Tone Analysis

We perform multi-class text classification on sentences extracted from past supervisory letters into four ordinal tone classes — “Neutral”, “Cautious”, “Concerned” and “Forceful” with tone severity ranging from none (“Neutral”) to the most severe (“Forceful”). We choose the weighted-F1 score as our metric due to an imbalanced class distribution and the relatively equal importance of precision and recall for our use case.

4.1.1 Data Preprocessing and Labeling

Our data set is derived from a collection of confidential supervisory letters issued over a period of four years between 2013 and 2016. The letters are first processed into 15,000 individual sentences using a combination of regular expression and custom rules, then anonymized using a custom Named-entity recognition (NER) model. The NER model identifies sensitive information such as financial institution names, corporation names, and individual names that are then masked for data security reasons. Further, these anonymized entities may prevent the trained model from associating certain words with certain financial institutions, potentially reducing model bias.

Once data preprocessing is complete, the sentences are manually labeled as one of the four aforementioned tone classes by experienced supervisors. This serves as training data for our text classification model and as ground truth for model evaluation. Every sentence is labeled independently by three supervisors according to a labeling guide with the majority vote for each sentence selected as the final label. In addition, the labeled data is reviewed by a fourth supervisor on a sampling basis as quality check to ensure consistency of the labels. The entire labeling process is divided into six separate rounds, with supervisors reconvening after each round to discuss difficulties that they encounter, especially for sentences that are labeled differently by all three supervisors. Further, the periodic recalibrations also provide the opportunity for supervisors to jointly update the labeling guide with new insights, retroactively amend labeling from earlier rounds and evaluate intermediate model performance.

In our context of supervisory communication, labeling for sentence tone is challenging due to the nuanced language used in supervisory letters. Generally, supervisors need to identify parts of the sentence that best reflect the intended tone without accounting for the severity of the supervisory issue at hand. For example, in the sentence "inadequate details presented in credit risk reports", the word "inadequate" should convey tone, but not the credit risk report reference. This approach is necessary as the range of issues that concern an FI can vary tremendously. Therefore, if all possible types of issues are taken into consideration during labeling, the machine learning model would likely underperform considering the multiple tone classes and small data size. Nonetheless, there are exceptions provided for specific issues like money laundering and terrorist financing risks that are taken very seriously by the Central Bank of Malaysia. These sentences warrant a deviation in tone classification due to the severity of the issue.

4.1.2 Model and Performance

As mentioned earlier, deep learning models have been making breakthrough performances in text classification. However, these models are typically large and require significant computing power for training and inferencing. Due to compute limitations, we restrict our model search to smaller and lighter models that are able to train quickly and perform inference effectively. Hence, we choose DistilBERT (Sanh et al., 2020), a smaller and faster version of the original BERT model proposed by Google, that retains much of BERT's performance with a significantly lower number of parameters (66 million vs 110 million).

Similar to BERT, DistilBERT employs transfer learning, a technique of pre-training a model on vast amounts of general-domain internet text data before transferring that knowledge to a downstream task. This process gives the language model a significant boost in performance as it learns a diverse range of word usage in various linguistic structures. We first download the pre-trained DistilBERT model provided by Hugging Face. However, instead of directly finetuning DistilBERT on a downstream task like text classification, we implement the target task language model fine-tuning technique proposed by Howard and Ruder (2018) for ULMFiT and demonstrated by Sun et al. (2019) for BERT. This entails additional pre-training on both the training and test data set to update the parameters in DistilBERT's masked language model in order to better reflect information from a prudential supervision domain. This is also necessary since supervisory letters may contain words that are only used in a local context, such as those derived from Malaysian regulatory requirements. We then integrate the fastai library (Howard & Gugger, 2020) with Hugging Face's transformers library (Wolf et al., 2020) as shown by Roberti (2019) to leverage on discriminative fine-tuning, slanted triangular learning rates and gradual unfreezing for training after attaching the classifier layer. These additional features provided by the fastai library stem from an Idea derived from the sub-field of computer vision, whereby earlier layers of a neural network contain more general features, while later layers learn features specific to the task (Yosinski et al., 2014). Therefore, we use these techniques to achieve granular control over the training extent for each layer — later layers of DistilBERT are trained more than earlier layers for better classification ability. Due to the small data size, we limit the number of training epochs to a small number to prevent the model from overfitting. We find that overall, this additional fine-tuning step improves the weighted-F1 score by 1.4%, resulting in a final weighted-F1 score of 77.6% as shown in Table 1.

Model performance on test set

In percentage (%) Table 1

Model	Accuracy	F1 (weighted)
Bag-of-Words + Logistic Regression	66.0	65.4
Bag-of-Words + XGBoost	67.0	65.3
SBERT + Logistic Regression	70.2	69.4
SBERT + XGBoost	69.1	67.7
DistilBERT-FiT	77.8	77.3
DistilBERT-FiT (Quantized)	74.6	73.0
DistilBERT-FiT + ONNX Runtime (Quantized)	76.8	76.2
DistilBERT-ITPT-FiT	**78.8**	**78.4**
DistilBERT-ITPT-FiT (Quantized)	74.8	74.4
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	78.0	77.6

Note: SBERT is "Sentence-BERT". DistilBERT-FiT is "DistilBERT + downstream FineTuning". DistilBERT-ITPT-FiT is "DistilBERT +8nto8sn-Task Pre-Training + downstream Fine-Tuning". Our final selected model is DistilBERT-ITPT-FiT + ONNX Runtime (Quantized).

At the initial project exploration stage, we applied a simple lexicon-based approach by identifying and grouping domain-specific words that are commonly used in supervisory letters. Each of these words were given a tone score according to its group to indicate tone severity which could then be used to automatically analyze and score a full-length supervisory letter. Not surprisingly, this approach proved difficult to account for the various word forms, combinations, semantics, and grammatical structure within the letter. For example, if a section in a letter contained one word with a severe tone but many words with softer tones, its overall tone could be ambiguous. Given that the web application serves to aid supervisors in validating an intended tone or recalibrating an unintended tone, this situation may be difficult for supervisors to interpret.

Hence, we decided to proceed with supervised machine learning where we performed multi-class text classification on every sentence in a letter. Despite the loss of some tonal information when a text sequence is evaluated on a sentence level without full context from the original paragraph, the trade-off for an overall simpler and more interpretable tone was worthwhile. We first experimented with computationally cheaper models, starting with a Bag-of-Words feature representation for sentences which were then piped in as input to various traditional machine learning models like logistic regression and Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016). While this overcame some of the issues mentioned earlier, these vectors still lacked semantic meaning and contextual information. Thus, we quickly progressed to a Transformer-based NLP model called

Sentence-BERT (SBERT). SBERT is a modification of the pre-trained BERT model that uses9nto9see and triplet network structures to derive dense vectors that can capture semantic meaning and contextual information of words in a sentence (Reimers & Gurevych, 2019). These generated sentence embeddings were then fed into a downstream machine learning classification model.

Despite yielding performance gain, the sentence embeddings were created from a model trained on a general-domain text corpus which may not have been suitable for the prudential supervision domain. In order to achieve better performance, we required a model embedded with a classifier which could update its weights for a text classification task in the prudential supervision domain, leading to our final selected model — DistilBERT.

4.1.3 Model Inference Optimizations

Inference time per sentence

In milliseconds (ms)

Table 2

Framework	Min	Mean	Median	Max
DistilBERT-ITPT-FiT	13.78	30.70	28.67	96.57
DistilBERT-ITPT-FiT + ONNX Runtime	6.59	20.52	19.65	173.64
DistilBERT-ITPT-FiT (Quantized)	5.85	18.70	17.02	145.49
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	**2.67**	**12.69**	**11.89**	**52.70**

Inference time per letter

In seconds (s)

Table 3

Framework	Min	Mean	Median	Max
DistilBERT-ITPT-FiT	0.43	2.98	2.42	9.16
DistilBERT-ITPT-FiT + ONNX Runtime	0.28	1.99	1.62	5.78
DistilBERT-ITPT-FiT (Quantized)	0.23	1.81	1.38	5.28
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	**0.17**	**1.23**	**1.02**	**3.64**

Note: Distributions of inference times for a sentence and a letter are presented in Figure 1 and Figure 2 respectively in Appendix I.

To optimize inference (text classification) speed, we use ONNX Runtime and quantization on DistilBERT to reduce the median inference time for a sentence by 2.4 times from 28.67 milliseconds to 11.89 milliseconds with minimal impact to the weighted-F1 score (decreased by 0.8%). ONNX Runtime is an optimization library that boosts inference speed while quantization is a technique that approximates floating-point numbers with integers to reduce memory use and accelerate performance (Functowicz and Li, 2020). As supervisors typically need to repeatedly upload different versions of their letters when drafting, it is imperative that we maintain acceptable inference time for a full-length letter while maximizing the weighted-F1 metric. Even though a baseline DistilBERT without optimizations ("DistilBERT-ITPT-FiT") would perform inferencing faster than a larger baseline BERT, limited computational resources would still dampen user experience of the web application as users could potentially wait up to 9.16 seconds for a letter to be analyzed. This necessitates the use of ONNX Runtime and quantization optimizations, performing inference on a letter with a median of 1.02 seconds and a maximum of 3.64 seconds.

4.1.4 Model Prediction Interpretation

In order to interpret a predicted sentence tone, we utilize integrated gradients (Sundararajan et al., 2017) implemented in Captum (Kokhlikyan et al., 2020), a model interpretability library for PyTorch. We also display the probability of the model's prediction when presenting the integrated gradients output. Integrated gradients is a method that attributes a model's prediction on an input to features of the input. Th10ntouition is the following — integrated gradients generates a linear interpolation from a baseline vector to a final vector that represents our predicted sentence, effectively producing multiple vectors that are progressively closer to the final vector. It then acquires gradients which significantly influence the prediction probability across the interpolated vectors and averages them. For our use case, integrated gradients is able to attribute the predicted tone of a sentence to every word in it with varying contribution levels. Words with positive attribution scores nudge the sentence towards the predicted tone while those with negative attribution scores pull the sentence away from it. In practice, this often results in specific keywords visibly highlighted for their positive attributions, reflecting the specific vocabulary used by supervisors when communicating issues of different severities (Appendix II). For example, in the sentence "inadequate compliance review and limited audit coverage scope", integrated gradients may highlight the words "inadequate" and "limited" as keywords that contribute most to the predicted sentence tone. When paired with the prediction probability score, supervisors can assess the reliability of this prediction and interpretation wherein a lower probability indicates greater model prediction uncertainty and therefore may not be relied on.

4.1.5 Machine Learning Workflow

To ensure a sustainable machine learning lifecycle that continuously produces relevant results, we incorporate a machine learning workflow component to support intuitive model training, tracking and deployment. This is accomplished in part via the integration of MLflow (Zaharia et al., 2018), a tool that greatly simplifies these operations. The simplicity of MLflow allows us to extend its features to enable supervisory administrators to manage the web application without significant input from the technical team.

The machine learning workflow begins with supervisors uploading finalized supervisory letters that are ready for storage into the web application. The application preprocesses the letters into a list of sentences, masks the sensitive information using a custom NER model, then stores them in Elasticsearch. Thereafter, the data labeling team can generate the list of preprocessed sentences for labeling. Once the sentences are labeled, the file is uploaded back into the web application. This action triggers a model training run that produces a new version of the model alongside performance metrics and hyperparameter configurations. Supervisors can then easily compare the model's performance across different training runs and select the best model for deployment.

In addition to the periodic data labeling by a dedicated team, we supplement the labeled dataset with input from the community of supervisors who utilize the Tone Analysis feature to analyze their draft letters. As they analyze the tone of their draft supervisory letters, supervisors can provide input on the model's predicted sentence tone by suggesting a different tone classification via the user interface. These user feedbacks are then aggregated and reviewed by an administrator for potential inclusion in the next cycle of model retraining.

4.2 Sentence Search

Sentence Search enables search by keywords or by semantically similar sentences extracted from historical supervisory letters stored in Elasticsearch, a database optimized for text queries. With keyword matching, all sentences that contain the query will be extracted with options to filter search results by tone.

To implement sentence search by semantic similarity, we utilize a pre-trained SBERT ("all-mpnet-base-v2"), fine-tuned from Microsoft's MPNet model (Song et al., 2020). Similar to DistilBERT, SBERT can provide semantically meaningful sentence embeddings, but does not include a classifier layer required for text classification and is optimized for tasks like semantic textual similarity. We attempted to further train the SBERT model using existing supervisory sentences to adapt it for the supervision domain, but the fine-tuned model performed worse than the original model due to limited data (insufficient examples of highly similar sentences).

As a result, we directly use the pre-trained model to encode the sentences from all historical supervisory letters into sentence embeddings that we then store in Elasticsearch alongside the original sentences. When a supervisor inputs a query sentence through the web application, the query is converted to an embedding using the same SBERT model. With both the query and historical sentences embedded in the same vector space, we can calculate the cosine similarity scores to determine how similar the sentences are. Subsequently, we rank the sentences by similarity scores and present the relevant results. Once again, the search results can then be filtered by tone.

5. Conclusion

In this project, we develop a web application to facilitate the drafting of supervisory letters and to ensure consistency in their communication tone. We accomplish this via text classification and semantic textual similarity in the tone analysis and sentence search features respectively. For future work, we seek to re-evaluate the labeled data to better distinguish the tone classes with the aim of improving the weighted-F1 score. This may be performed in parallel with the labeling of additional data derived from more recent supervisory letters to update the model for changes in language and supervisory issues. Further, we intend to develop additional tools surrounding supervisory issues and remediation actions detailed in the letters that may be integrated with this Supervisory Letter Writing App to form a holistic web application.

References

Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). Text mining for central banks. *Available at SSRN 2624811*.

Bholat, D., Brookes, J., Cai, C., Grundy, K., & Lund, J. (2017). Sending Firm Messages: Text Mining Letters from PRA Supervisors to Banks and Building Societies They Regulate. Bank of England Working Paper, 688. https://doi.org/10.2139/ssrn.3066809

Born, B., Ehrmann, M., & Fratzscher, M. (2014). Central Bank Communication on Financial Stability. The Economic Journal, 124(577), 701–734. https://doi.org/10.1111/ecoj.12039

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 785–794. https://doi.org/10.1145/2939672.2939785

Correa, R., Garud, K., Londono, J. M., & Mislang, N. (2017). Sentiment in Central Bank's Financial Stability Reports. International Finance Discussion Paper, 1203, 1–46. https://doi.org/10.17016/ifdp.2017.1203

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171–4186. https://10.18653/v1/N19-1423

Fraccaroli, N., Giovannini, A., & Jamet, J.-F. (2020). Central banks in parliaments: a text analysis of the parliamentary hearings of the Bank of England, the European Central Bank and the Federal Reserve. ECB Working Paper, 20202442.

FSB. (2020). The Use of Supervisory and Regulatory Technology by Authorities and Regulated Institutions: Market developments and financial stability implications. https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2442~e78be127c0.en.pdf

Howard, J., & Gugger, S. (2020). Fastai: A Layered API for Deep Learning. Information, 11(2), 108. https://doi.org/10.3390/info11020108

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1, 328–339. https://doi.org/10.18653/v1/p18-1031

Hubert, P., & Labondance, F. (2018). Central Bank Sentiment and Policy Expectations. Bank of Englang Working Paper, 648. https://doi.org/10.2139/ssrn.2920496

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-richardson, O. (2020). Captum : A unified and generic model interpretability library for PyTorch An Overview of the Algorithms. ArXiv, abs/2009.07896

Li, Y. (2020, September 1). Faster and smaller quantized NLP with Hugging Face and ONNX Runtime. Medium. https://medium.com/microsoftazure/ faster-and-smaller-quantized-nlp-with-hugging-face-and-onnx-runtime-ec5525473bb7

Jegadeesh, N., & Wu, D. (2015). Deciphering Fedspeak: The Information Content of FOMC Minutes. Working Paper, University of Pennsylvania.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR), 1–12.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning Based Text Classification: A Comprehensive Review. ArXiv, abs/2004.03705

Pennington, P., Socher, R., & Manning C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532– 1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 2227–2237. https://doi.org/10.18653/v1/n18-1202

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992. https://doi.org/10.18653/v1/d19-1410

Roberti, M. (2019, November 27). Fastai with Transformers (BERT, RoBERTa, XLNet, XLM, DistilBERT). Towards Data Science. https://towardsdatascience.com/fastai-with-transformers-bert-roberta-xlnet-xlm-distilbert-4f41ee18ecb2

Rybinski, K. (2019). A machine learning framework for automated analysis of central bank communication and media discourse. The case of Narodowy Bank Polski. Bank i Kredyt, 50(1), 1–19.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108

Shapiro, A. H., & Wilson, D. (2019). Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis. Federal Reserve Bank of San Francisco Working Paper, 2. https://doi.org/10.24148/wp2019-02

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. NeurIPS, 1–14. https://arxiv.org/abs/2004.09297

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? Lecture Notes in Computer Science, 11856, 194–206. https://doi.org/10.1007/978-3-030-32381-3\16

Sundararajan, M., Taly, A., & Yan, Qiqi. (2017). Axiomatic Attribution for Deep Networks. ArXiv, abs/1703.01365v2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Advances in Neural Information Processing Systems, 6000-6010.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in

Natural Language Processing: System Demonstrations, 38-45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

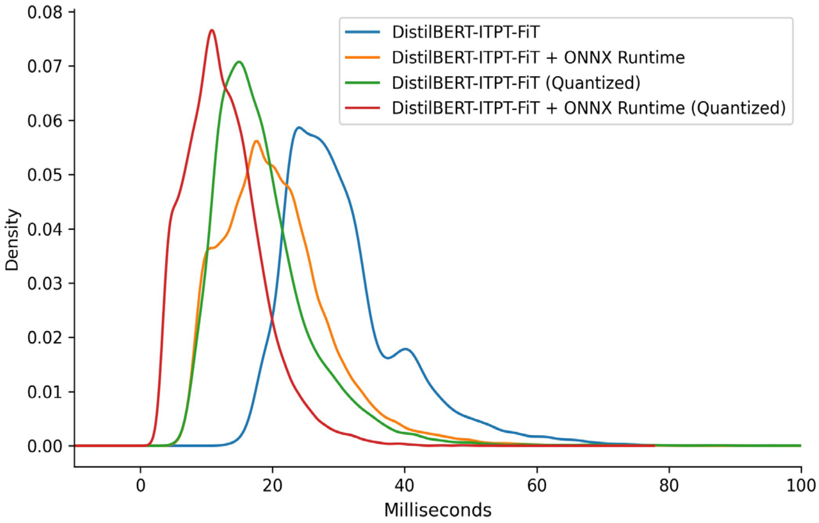
Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Aautoregressive Pretraining for Language Understanding. NeurIPS.

Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., ... & Zumar, C. (2018). Accelerating the machine learning lifecycle with MLflow. *IEEE Data Eng. Bull.*, *41*(4), 39-45.

Appendix I: Inference Time Distribution Graphs

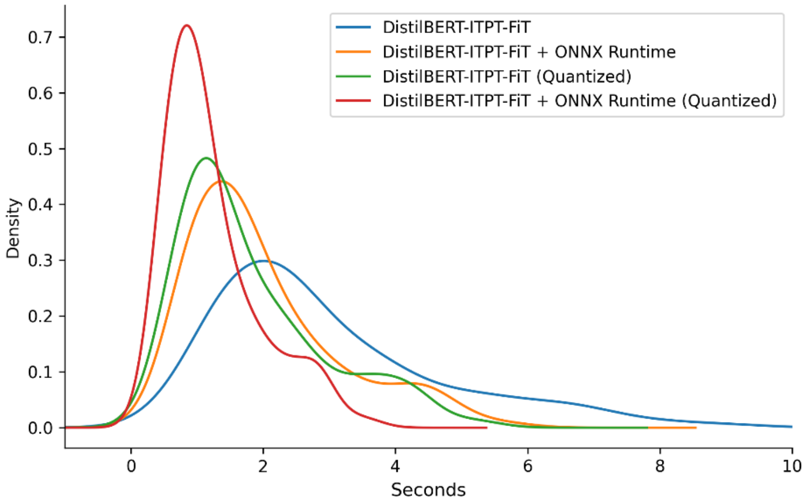
Distribution of inference time for a sentence

Graph 1



Distribution of inference time for a letter

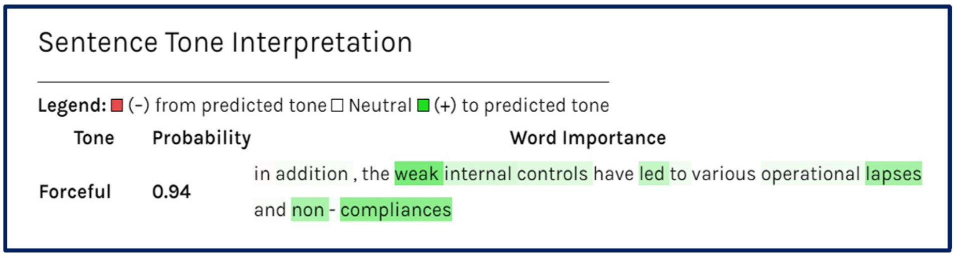
Graph 2



Appendix II: Model Prediction Interpretation

Interpretation for a sentence with predicted “forceful” tone

Figure 1



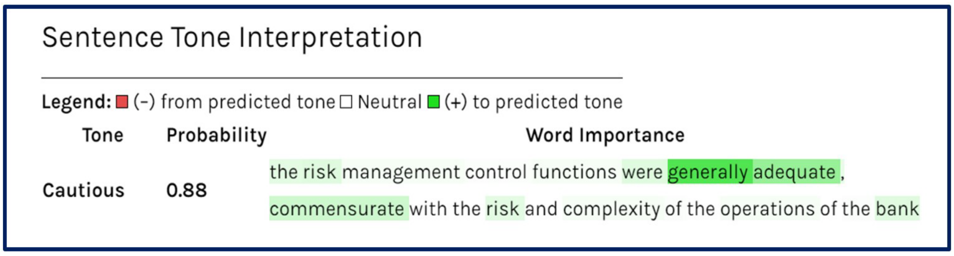
Sentence Tone Interpretation

Legend: 🟥 (–) from predicted tone □ Neutral 🟩 (+) to predicted tone

Tone	Probability	Word Importance
Forceful	0.94	in addition , the weak internal controls have led to various operational lapses and non - compliances

Interpretation for a sentence with predicted “cautious” tone

Figure 2



Sentence Tone Interpretation

Legend: 🟥 (–) from predicted tone □ Neutral 🟩 (+) to predicted tone

Tone	Probability	Word Importance
Cautious	0.88	the risk management control functions were generally adequate , commensurate with the risk and complexity of the operations of the bank

Supervisory Letter Writing App

Expediting Letter Drafting & Ensuring Tone Consistency

IFC-Bank of Italy Workshop on Data Science in Central Banking

*Authors: Joshua Tan, **Chi Ken Shum**, Mohd. Akmal*

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

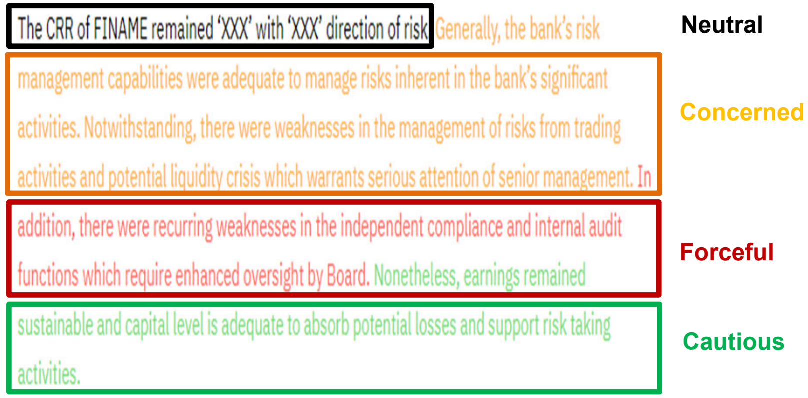
Background

Background	• Financial institutions (FIs) operating in Malaysia are subject to annual supervisory examinations by Bank Negara Malaysia, where the safety and soundness of each FI is assessed. • At the conclusion of their review, supervisors will issue supervisory letters to the respective FIs to communicate their assigned supervisory rating, highlight supervisory concerns and recommend remediation actions required.
Problem Statement	• Most supervisors who are relatively new to the position find drafting supervisory letters is laborious due to the numerous revisions that occur during the review process. • Supervisory letters are issued by multiple departments. Hence, there is no effective nor systematic way to gauge the tones used in these letters.

Web Application (1)

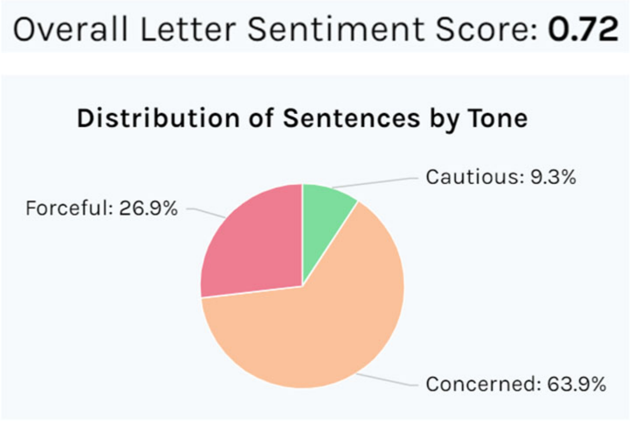
Module 1: Tone Analysis

- Facilitate supervisors to better understand and calibrate the tone of their drafted letters to commensurate with the supervisory concerns and intended corrective measures to be taken by FIs.
- Supervisors can upload a draft letter and obtain sentence level tone predictions that are color-coded according to tone for easier analysis.
- Supervisors can also view the overall letter sentiment score and the tone distribution of all predicted sentences summarized in a chart.



The CRR of FINAME remained 'XXX' with 'XXX' direction of risk Generally, the bank's risk management capabilities were adequate to manage risks inherent in the bank's significant activities. Notwithstanding, there were weaknesses in the management of risks from trading activities and potential liquidity crisis which warrants serious attention of senior management. In addition, there were recurring weaknesses in the independent compliance and internal audit functions which require enhanced oversight by Board. Nonetheless, earnings remained sustainable and capital level is adequate to absorb potential losses and support risk taking activities.

Supervisory letter with predicted sentence tones



Overall Letter Sentiment Score: **0.72**

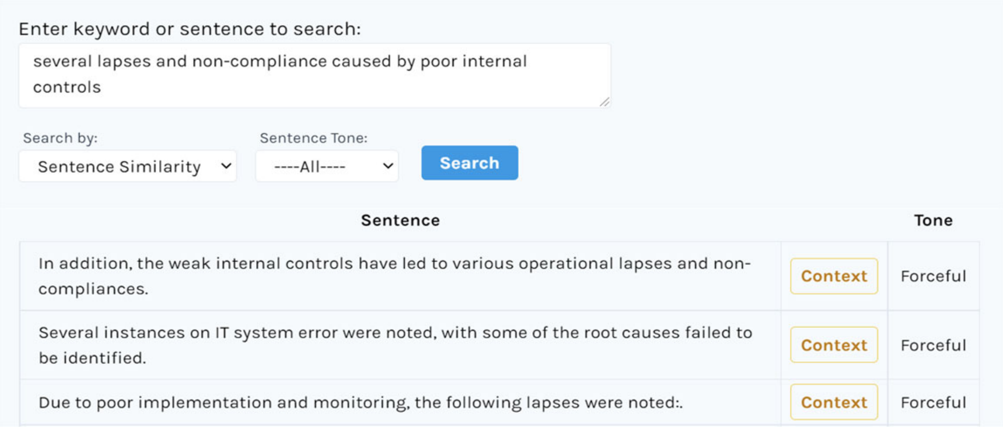
Distribution of Sentences by Tone

Overall letter score and sentence tone distribution

Web Application (2)

Module 2: Sentence Search

- Enable supervisors to search for sentences extracted from previously issued letters by tone to expedite the writing process.
  - – Important for context understanding, consistency in vocabulary.
- How to perform a search?
  1. Navigate to search page → Type query in input box
  2. After uploading the supervisory letter for tone analysis → Highlight word/ sentence of interest and query the database directly from the tone analysis page
- Can search by keywords or semantically similar sentences.



Enter keyword or sentence to search:

several lapses and non-compliance caused by poor internal controls

Search by: Sentence Similarity

Sentence Tone: ----All----

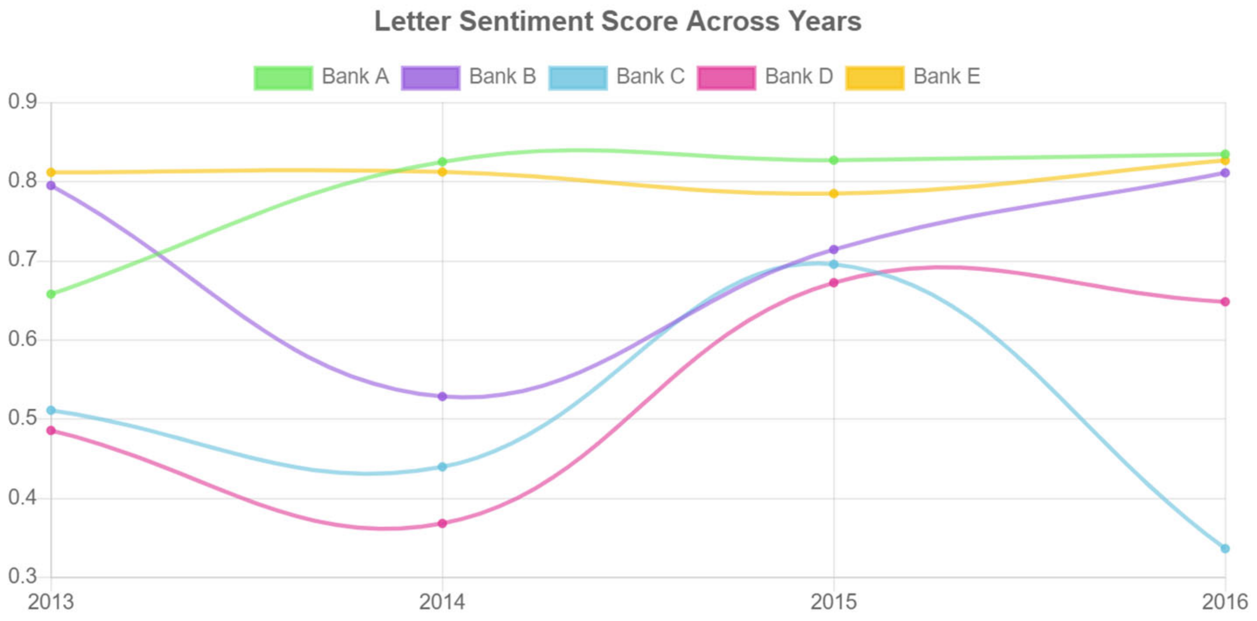
Sentence		Tone
In addition, the weak internal controls have led to various operational lapses and non-compliances.	Context	Forceful
Several instances on IT system error were noted, with some of the root causes failed to be identified.	Context	Forceful
Due to poor implementation and monitoring, the following lapses were noted:.	Context	Forceful

Search result of a sentence by semantic similarity

Web Application (3)

Admin Dashboard

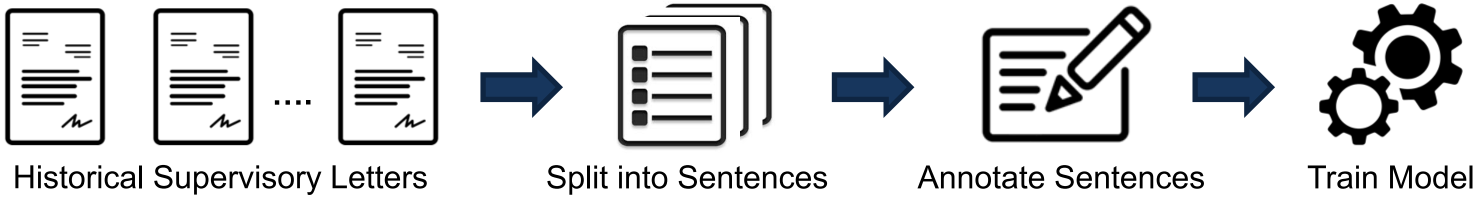
- Privileged access dashboard for easy retrieval of past letters and holistic visualization of letter tones across time and across FIs.
- Able to monitor and spot anomalies in letter tones across FIs of similar risk ratings or of a particular FI across time.



Time series of document-level sentiment scores

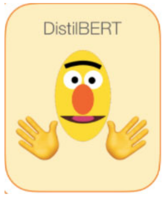
Tone Analysis (1): Methodology

- Multi-class text classification on each sentence in a draft supervisory letter.
- Each sentence is classified into one of 4 ordinal tone classes: ‘Neutral’, ‘Cautious’, ‘Concerned’, ‘Forceful’.
- 15K sentences, labelled by 12 independent supervisors in 4 groups of 3 with quality checks by 3 supervisors on a sampling basis.
- Each sentence contributes a score corresponding to their predicted tone that is aggregated and normalized to arrive at a compound score representing the overall tone of the letter.
- Letters are then assigned to different categories based on overall tone scores for document-level comparisons.

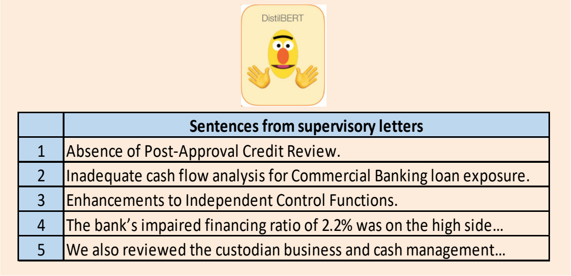


Tone Analysis (2): Model Training and Performance

- Leverage transfer learning to improve model performance.

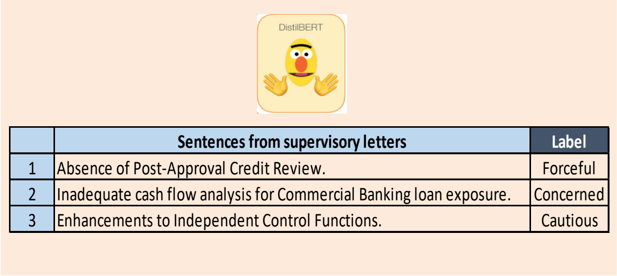


Download pre-trained DistilBERT from HuggingFace.



	Sentences from supervisory letters
1	Absence of Post-Approval Credit Review.
2	Inadequate cash flow analysis for Commercial Banking loan exposure.
3	Enhancements to Independent Control Functions.
4	The bank's impaired financing ratio of 2.2% was on the high side…
5	We also reviewed the custodian business and cash management…

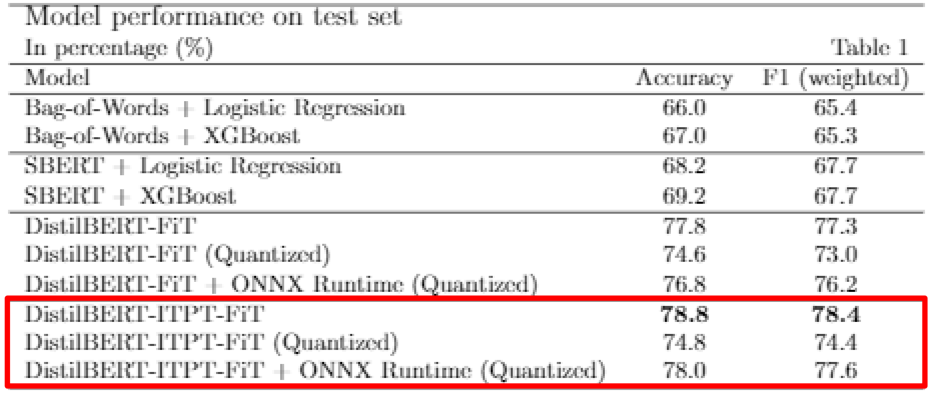
Further train the model with all unlabelled sentences.



	Sentences from supervisory letters	Label
1	Absence of Post-Approval Credit Review.	Forceful
2	Inadequate cash flow analysis for Commercial Banking loan exposure.	Concerned
3	Enhancements to Independent Control Functions.	Cautious

Fine-tune the model on labelled sentences.

- The fined-tuned transformer-based model performs the best with a weighted-F1 score of 78.4.



Model performance on test set

In percentage (%) Table 1

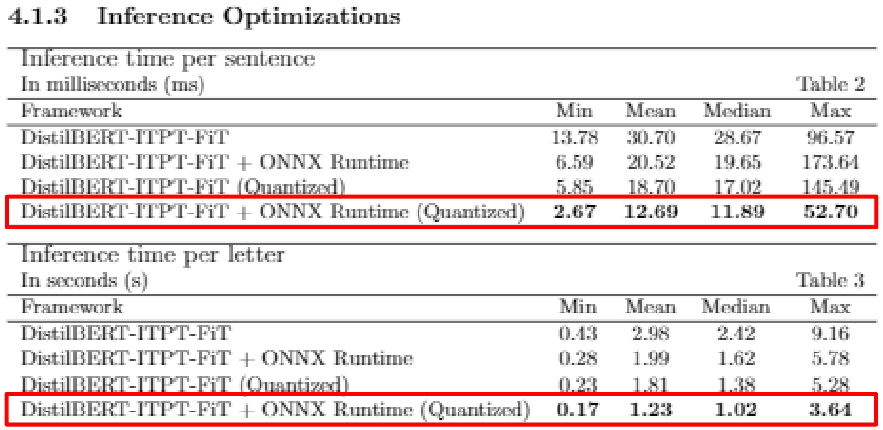
Model	Accuracy	F1 (weighted)
Bag-of-Words + Logistic Regression	66.0	65.4
Bag-of-Words + XGBoost	67.0	65.3
SBERT + Logistic Regression	68.2	67.7
SBERT + XGBoost	69.2	67.7
DistilBERT-FiT	77.8	77.3
DistilBERT-FiT (Quantized)	74.6	73.0
DistilBERT-FiT + ONNX Runtime (Quantized)	76.8	76.2
DistilBERT-ITPT-FiT	**78.8**	**78.4**
DistilBERT-ITPT-FiT (Quantized)	74.8	74.4
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	78.0	77.6

Accuracy and F1 score on test data

Image source: https://jalammar.github.io/illustrated-transformer/

Tone Analysis (3): Inference Optimizations

- Utilize ONNX Runtime and quantization to optimize inference speed.
- Reduced median inference time for a sentence by 2.4 times from 28.67 milliseconds to **11.89 milliseconds**.
- Reduced maximum inference time for a letter by 2.5 times from 9.16 seconds to **3.64 seconds**.
- Minimal impact to the weighted-F1 score (decreased by 0.8%).



4.1.3 Inference Optimizations

Inference time per sentence
In milliseconds (ms)

Table 2

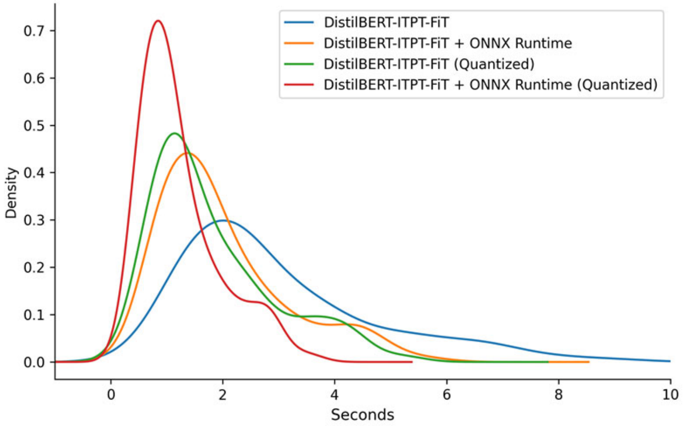
Framework	Min	Mean	Median	Max
DistilBERT-ITPT-FiT	13.78	30.70	28.67	96.57
DistilBERT-ITPT-FiT + ONNX Runtime	6.59	20.52	19.65	173.64
DistilBERT-ITPT-FiT (Quantized)	5.85	18.70	17.02	145.49
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	**2.67**	**12.69**	**11.89**	**52.70**

Inference time per letter
In seconds (s)

Table 3

Framework	Min	Mean	Median	Max
DistilBERT-ITPT-FiT	0.43	2.98	2.42	9.16
DistilBERT-ITPT-FiT + ONNX Runtime	0.28	1.99	1.62	5.78
DistilBERT-ITPT-FiT (Quantized)	0.23	1.81	1.38	5.28
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	**0.17**	**1.23**	**1.02**	**3.64**

Model inference time

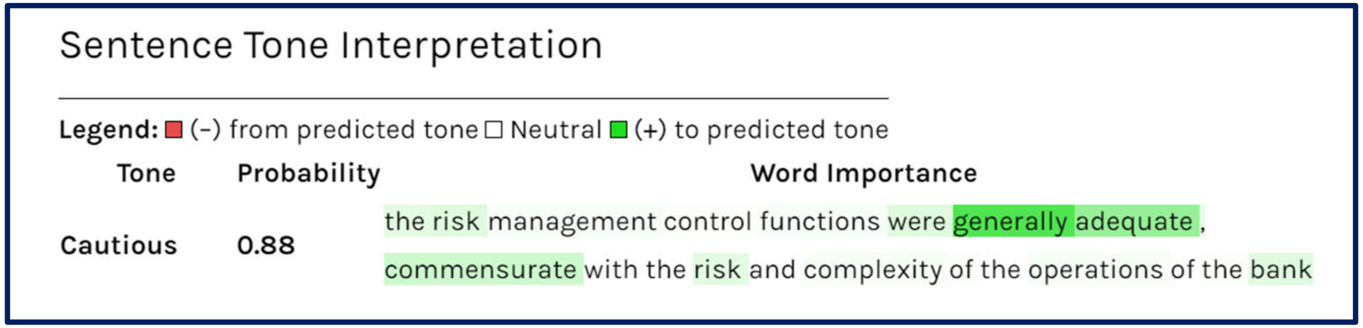


Inference time distribution for a letter

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

Tone Analysis (4): Model Prediction Interpretation

- Utilize integrated gradients in Captum to identify words (in green) that significantly contribute towards the predicted sentence tone. Higher color intensity indicates higher contribution and vice versa.
- Display probability as a confidence score for the model prediction.

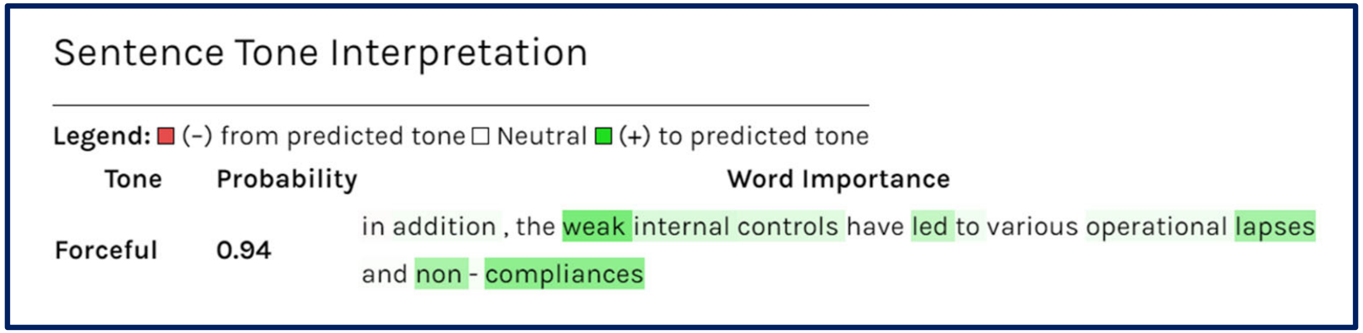


Sentence Tone Interpretation

Legend: 🟥 (–) from predicted tone □ Neutral 🟩 (+) to predicted tone

Tone	Probability	Word Importance
Cautious	0.88	the risk management control functions were generally adequate , commensurate with the risk and complexity of the operations of the bank

Interpretation for a sentence with predicted "cautious" tone



Sentence Tone Interpretation

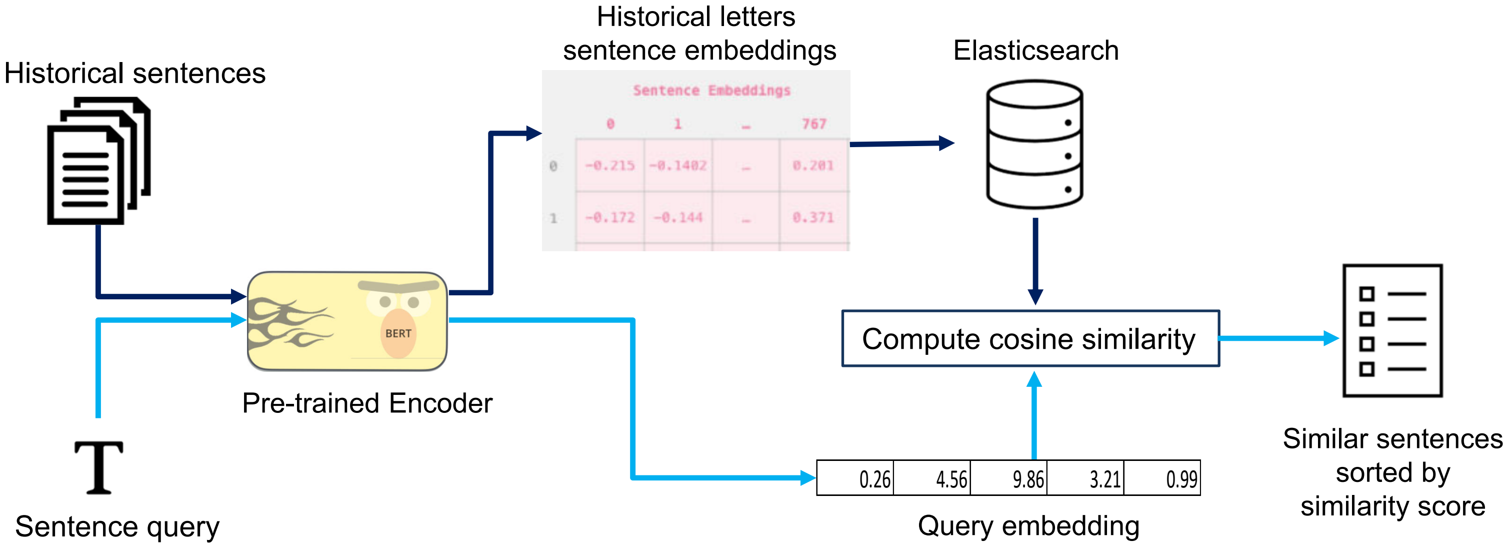
Legend: 🟥 (–) from predicted tone □ Neutral 🟩 (+) to predicted tone

Tone	Probability	Word Importance
Forceful	0.94	in addition , the weak internal controls have led to various operational lapses and non - compliances

Interpretation for a sentence with predicted "forceful" tone

Sentence Search: Approach

- Utilize a pre-trained Sentence-BERT model ("all-mpnet-base-v2").
- Encode sentences from all historical supervisory letters into sentence embeddings, then store in Elasticsearch.
- Upon receiving a sentence query, the model encodes the raw text into an embedding then calculates the cosine similarities between the query embedding and the sentence embeddings from historical supervisory letters.
- The results are then presented in descending order by similarity scores.


Historical sentences
Historical letters sentence embeddings
Elasticsearch
BERT
Pre-trained Encoder
Compute cosine similarity
T
Sentence query
0.26
4.56
9.86
3.21
0.99
Query embedding
Similar sentences sorted by similarity score

Sentence Embeddings

	0	1	…	767
0	-0.215	-0.1402	…	0.201
1	-0.172	-0.144	…	0.371

Image source: https://jalammar.github.io/illustrated-transformer/

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Disagreement between human and machine predictions[1]

Daisuke Miyakawa, Hitotsubashi University Business School, Japan,
and Kohei Shintani, Bank of Japan

[1] This presentation 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 central banks and other institutions represented at the event.

Disagreement between Human and Machine Predictions

By DAISUKE MIYAKAWA AND KOHEI SHINTANI*

In this study, we show that human predictions of firm exits disagree with machine predictions. First, human predictions generally underperform machine predictions. Second, the performance of human relative to machine predictions improves for firms with less observable information that is possibly due to the unstructured information that only humans can use. Specifically, under the environment where the number of exiting firms is much smaller than that of non-exiting firms, the reduction in type I errors from reallocating prediction tasks to humans instead of machines for opaque firms leads to better performance of predictions. (93 words)

Keywords: Machine Prediction; Human Prediction; Disagreement

JEL classification: C10, C55, G33

* Miyakawa (corresponding author): Associate Professor, Hitotsubashi University Business School, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8439 JAPAN. E-mail: dmiyakawa@hub.hit-u.ac.jp. Shintani: Director and Senior Economist, Institute for Monetary and Economic Studies, Bank of Japan, 2-1-1 Nihombashi-Hongokucho, Chuo-ku, Tokyo 103-8660 JAPAN. E-mail: kouhei.shintani@boj.or.jp.

I. Introduction

Prediction is an important task in both private business and public policy. Recent advances in prediction techniques, such as machine learning, have helped make the performance of prediction tasks more reliable than those dependent upon human judgment and classical parametric models. The practical application of these new prediction techniques has been the focus of recent academic, policy, and business discussions (Varian 2014; Mullainathan and Spiess 2017; Athey 2019). A number of fields have already reported successful applications of these techniques such as labor markets (Chalfin et al. 2016), public services (Kleinberg et al. 2018; Bazzi et al. 2019; Lin et al. 2020), medical services (Patel et al. 2019; Mei et al. 2020), and the financial industry (Agrawal et al. 2018).

The growing employment of these powerful prediction techniques naturally raises the question of the ways in which machine predictions disagree with and outperform human predictions. This question is particularly relevant given the number of recent studies which argue that technological advances will lead either to the replacement of human labor with machines in certain types of jobs (e.g., Frey and Osborne 2017) or to the reallocation of human resources to other types of jobs (e.g., Autor et al. 2003; Acemoglu and Autor 2011; Acemoglu and Restrepo 2018). To understand the ways in which machines outperform humans in predictions, we identify the cases in which human predictions outperform machine predictions.

While this question has started to be examined in fields like medical studies (e.g., Raghu et al. 2019), it has not yet been investigated in the context of social sciences.

The goal of this study is to use the context of firm exits to show the patterns of disagreement between human predictions and machine predictions and each predictor's relative performance. First, following the medical studies, we test the relative performance of predictions based on machine learning techniques and those based on human judgment for two modes of firm exits: corporate default and voluntary closure. Second, we identify the systematic patterns of disagreements between human and machine predictions for those events. The disagreement between them is measured by the performance of the machine relative to that of the human. Thus, we can see not only whether humans and machines disagree but also, more importantly, the ways in which they disagree. Suppose a firm is actually found to default ex post. Ex-ante human and machine predictions could differ. As reported by Kleinberg et al. (2018) in the context of judicial bail decisions, machine predictions outperform human predictions more often. Nonetheless, the relative performance of human predictions may be better in specific circumstances, such as default predictions for informationally opaque firms. Given this conjecture, we find that the relative performances of human and machine predictions are conditional on the characteristics of their prediction targets: firms. Third, after confirming the conjecture, we implement a set of counterfactual exercises that reallocate the predictions for firms with specific characteristics to humans instead of machines and see how overall performance of predictions varies.

To the best of our knowledge, this study is the first to explicitly examine the systematic patterns of disagreement between human and machine predictions in the context of social science and to use these systematic patterns to improve the overall performance of predictions.[1] We take advantage of our access to a huge volume of firm-level high-dimension panel data collected by one of the largest Japanese credit reporting agencies, together with the prediction results of anonymous professional analysts who work for the agency. These comprehensive datasets provide us with an ideal research ground on which we can construct a machine prediction model to compare its predictions with human predictions and show how they disagree and perform.

The empirical findings are summarized as follows: First, machines have better average performance in predicting firm exits than humans have, which is consistent with the results reported by the studies in another field (e.g., Kleinberg et al. 2018). Second, the performance of human predictions relative to that of machine predictions improves as the availability of information on firm characteristics declines. This improvement could be the case when human predictions effectively

[1] Anderson et al. (2017) report in the domain of chess that human decisions tend to be wrong for more difficult instances of chess. Their study shares the motivation with ours in the sense that both characterize the determinants of the performance of human decisions. The difference is that we compare human predictions not only with the ground truth (i.e., firm exits which we observe ex post), which is done in Anderson et al. (2017), but also with machine predictions.

use unstructured information in their predictions. The research has referred to this kind of unstructured information as “soft information” (e.g., Liberti and Petersen 2019). Examples of soft information include CEO’s management ability, the prospects of future product development, and so on. It is difficult to record all of this highly qualitative information as structured (i.e., “hard”) information in, for example, firms’ financial statements or other documents.

Therefore, we compare the human predictions recorded in our dataset not only to machine predictions but also to the part of the human predictions solely correlated with structured information.[2] As the latter structured human predictions do not rely on unstructured information, the comparison between them and the original identifies to what extent humans used unstructured information in their predictions. Similar to the comparison between the original human predictions and machine predictions, we find that the performance of human predictions relative to that of structured human predictions improves as the availability of information on firm characteristics declines. We also separately regress the performance of human and machine predictions on various characteristics including firm attributes and confirm that the negative marginal effects associated with lower availability of information is more sizable for machine predictions than for human predictions.

[2] A similar attempt to replicate human decisions was done in the context of chess (e.g., McIlroy-Young et al. 2020).

Given the empirical finding that the availability of observable information is a key driver in the disagreement between human and machine predictions and their relative performance, we implement a set of counterfactual exercises that reallocate predictions to professional analysts from machines that depends on how much information is available for each firm. The “improvement” in the relative performance of human predictions along with the change in specific firm characteristics does not guarantee that the “level” of conditional performance of human predictions is higher than that of machine predictions. In this sense, our counterfactual exercises are useful in confirming whether there could be any cases in which humans outperform machines when making predictions in the level of prediction performance.

As a main characteristics of firms, we pay attention to the number of available variables for each firm, which is closely related to the opaqueness of the firms. We orthogonalize the number of available variables to other firm characteristics such as size, past growth trend, and industry fixed effects so that we can extract the variation in the information opaqueness independent of those characteristics. Using this orthogonalized variable accounting for the information opaqueness, we classify firms into five categories that range from firms with the least information, little information, average information, more information, and the most information. For most of the cases except for firms with the least information, machine predictions outperform human predictions in terms of both type I and type II errors.

Nonetheless, we also find that reallocating predictions on firms with the least information to humans instead of machines leads to a sizable reduction in the type I error. To illustrate, for firms with the least information, the number of actual non-exiting firms predicted as “exit” by machines but “non-exit” by humans is larger than the number of actual non-exiting firms predicted as “non-exit” by machines but “exit” by humans. Thus, reallocating predictions on those firms to humans instead of machines reduces the number of false-positives, and the type I error becomes smaller. However, the reallocation of the predictions on these firms is also accompanied by a larger type II error; that is, the number of actual exiting firms predicted as “exit” by machines but “non-exit” by humans is larger than the number of actual exiting firms predicted as “non-exit” by machines but “exit” by humans. These results mean that reallocating predictions to humans instead of machines also reduces the number of true-positives, and thus the type II error increases. As the number of exit firms are much smaller than that of non-exit firms, the reduction in the type I error achieved by reallocating predictions on those opaque firms to humans instead of machines overwhelms the increase in the type II error. This is the mechanics in which the relative performance of human predictions to that of machine predictions improves as the availability of information on firm characteristics declines.

These results jointly show the usefulness of powerful machine prediction techniques for practical purposes and highlight a subtle feature of human prediction in the context of exit prediction. Overall, most of the prediction work for firm exits can be assigned to machines. Nonetheless, under specific circumstances, such as when the prediction targets are informationally opaque and the user of the resulting predictions is more concerned about the type I error than the type II error due to, for example, the imbalance between the numbers of exit and non-exit firms, then there is still room for human predictions to outperform machine predictions. Although we are not dealing with individuals as the subjects of predictions in the present study, these results support Gebru (2020) who reports that automated facial analysis systems tend to have lower prediction power for individuals with specific characteristics (e.g., dark-skinned women). Regardless of what types of subject that are the targets of the prediction, understanding under which cases machines could be wrong is useful.

The rest of the study proceeds as follows: Section II presents the theoretical underpinning of our empirical study, which follows Raghu et al. (2019). In Section III, we explain our empirical methodology and give a brief account of the institutional background related to the prediction of firm exits. Section IV gives details on the data used for our study. In Section V, we present and discuss the empirical results. Section VI concludes.

II. Conceptual Framework

In this section, we present the conceptual framework that represents the disagreement between human and machine predictions and their relative performance. Suppose there is a prediction f for a specific outcome. We set predictions for firms' default or voluntary closure as our prediction f. The f is accompanied by a set of attributes. It consists of, for example, the amount of available information associated with the firms as well as other firm characteristics in their financial statement. The f has the actual outcome $a(f)$ that we refer to as a ground truth. This ground truth only exists ex post when we observe whether the firm defaults or not within specific periods of time. For f, a prediction machine has its own prediction denoted by $m(f)$. Similarly, a professional analyst i with a set of individual attributes has its own prediction for f. We name this analyst's prediction $h(f,i)$. Using these items, we can first define the prediction error $\Theta(f)$ of machines for an f as follows:

(1) $$\Theta(f) = L(a(f), m(f)).$$

Second, we can define the prediction error $\Omega(f,i)$ of humans for f by an analyst i as follows:

(2) $$\Omega(f,i) = L(a(f), h(f,i)).$$

Suppose we have a set of predictions U. What we ultimately want to solve is an allocation problem of U for machines (i.e., S) or analysts (i.e., T). Such an optimization problem can be formulated as follows:

$$\min_{S,T} \sum_{f \in S} \Theta(f) + \sum_{f \in T} \Omega(f,i) \text{ s.t. } S \cup T = U; S \cap T = \emptyset. \tag{3}$$

This problem is called "an algorithmic triage" in Raghu et al. (2019). To solve this problem, we obtain the best assignment (S^*, T^*) as a function of (f, i). This optimal assignment function tells us whether we should assign a specific prediction f to the machine or to an analyst i.[3] In this paper, we specifically aim to identify $\Theta(f)$ and $\Omega(f,i)$ so that we can understand the sources of the disagreement and further solve the algorithmic triage problem as a counterfactual exercise.

For this purpose, we define an additional function $Proxy_{f,i}$ as follows:

$$Proxy_{f,i} = \Omega(f,i) - \Theta(f). \tag{4}$$

As $\Theta(f)$ and $\Omega(f,i)$ denote the prediction errors of the machine and the analyst, the relative performance of the human prediction becomes higher as $Proxy_{f,i}$

[3] Although the current setup does not contain any constraints for the optimization problem, realistic constraints such as a maximum number of instances a professional analyst can take care of could be introduced to the problem. Such a problem is a classic example of a matching problem.

becomes smaller. As we explicitly demonstrate in the following sections, this $Proxy_{f,i}$ accounts not only for the disagreement between human and machine predictions but also for their relative performance.

While the current setup suffices to study the systematic disagreement between human and machine predictions, further decomposition of $\Omega(f,i)$ into those correlated with structured information and the rest of the components is useful for understanding the source of the disagreement between human and machine predictions. Let $\Omega_h(f)$ account for the error component of the human prediction correlated with structured observable attributes of f. Using this decomposition, we can define another measure for disagreement between the human prediction and the structured human prediction that relies solely on hard information.

$$Proxy'_{f,i} = \Omega(f,i) - \Omega_h(f). \tag{5}$$

Suppose $Proxy'_{f,i}$ becomes smaller as the change in an attribute of the instance f (e.g., the amount of available information decreases). This change means the relative performance of the human prediction to the structured human prediction becomes better due to the change in the attribute. In this illustration, the volume of structured information becomes smaller, the room for analysts to effectively utilize unstructured information for prediction becomes larger. This comparison between human and structured human predictions highlights the reason why human predictions can surpass machine predictions, with the latter relying only on structured information.

III. Empirical Strategies

This section first presents the way that we construct a machine learning prediction model for firm exits. Then, we explain how to identify the determinants of disagreement and the relative performance of human and machine predictions.

A. Machine Prediction

To obtain machine predictions, we construct a standard machine learning method. Our particular problem with predicting relatively rare firm exits falls into the class of "imbalanced label predictions." Following the literature, we apply a weighted random forest and a minority-class oversampling method.[4] Random forest models aggregate many individual decision tree models that are each trained on a randomly selected samples and predictors from the training data. To predict rare events, Chen et al. (2004) develop an extension of the random forest, called a weighted random forest. Logically, the method weighs data corresponding to a minority event (e.g., a firm exit) much more heavily than that corresponding to a majority event (e.g., non-exit).

[4] We also use other machine learning techniques such as LASSO and extreme gradient boost to construct prediction models and confirm the robustness of our results. All the results are in the online appendix.

In our baseline exercise, we train models by using outcome variables from the end of year $t-1$ to the end of year t and the predictors available for the periods from year $t-3$ to $t-1$. Then, we conduct out-of-sample predictions of the realization of the outcome variables from the end of year t to the end of year $t+1$ by using the information available over the periods from year $t-2$ to t.

We use the receiver operating characteristic (ROC) curve to evaluate the predictive performance of the model. To implement the prediction of a binary exit outcome, we need a specific threshold. When a predicted score surpasses the threshold, it indicates a positive binary outcome. For a given trained model, the ROC curve plots the true and false positive rates that correspond to the variation in this threshold value. Without any predictors (i.e., random guesses), the curve should follow a 45-degree line, and curves that are closer to the top-left corner are desirable (maximize true positive rate and minimize false positive rate). Following convention, we summarize the ROC curve with the area under the curve (AUC).

B. Human Prediction

"fscore"—Credit reporting agencies examine and predict firm exits as these outcomes are of great interest to business and government entities. Examples of such credit reporting agencies include Dunn and Bradstreet in the US, Experian in European countries, and Tokyo Shoko Research (TSR) in Japan. By providing structured information such as financial statements to their clients, credit reporting

agencies typically calculate and publish a credit rating score, which we call "$fscore$", to summarize the overall performance of a firm. This score is typically constructed from both structured information on firm characteristics and from the contents of in-depth interviews on firm's characteristics, reputation, growth opportunity, and so on (i.e., unstructured information). The score is constructed by a professional analyst and assigned to each firm in each year. As in financial institutions such as banks, the agency evaluates each analyst on the performance of their predictions of this $fscore$. Thus, analysts have a reasonable incentive to produce good predictions.

These agencies typically rely on their own (often confidential) algorithm to construct the scores. While a part of the score systematically depends on structured information, a large part of the score reflects professional analysts' subjective evaluation of the targeted firm. To illustrate, according to the publicly available information, a score given by TSR (max: 100 points) is the summation of (i) the capability of the firm (max: 20 points) based on business attitude, experience, and asset condition; (ii) the growth possibility (max: 25 points) based on past sales growth, growth of profits, and characteristics of the products; (iii) stability (max: 45 points) based on the firm's age, stated-capital, financial statement information, room for collateral provision, and real and financial transaction relationships; and (iv) the firm's reputation (max 10 points) based on the level of disclosure and overall reputation. Most of these items are rarely recorded as structured information but largely as unstructured information. Given this institutional background, we use the *fscore* assigned by TSR as the output of human predictions.

We use this score and the ex-post record of exit to run a weighted Probit estimation that has the exit indicator on the left hand-side and only $fscore$ on the right hand-side of the estimated equation. Through this equation, we transform the $fscore$ into a value between 0 and100 as the score associated with the occurrence of the firm exit and use it as the result of human prediction.[5]

Can we really use fscore as a human prediction? There could be several immediate concerns about using the *fscore* as the output of human predictions. First, this score might also be constructed by some machine algorithms. If this is the case, the comparison between human and machine predictions becomes merely a comparison of two algorithms. However, we also try to separate out the analysts' predictions correlated with structured information from the original *fscore*. Using this framework, we can explicitly study the difference between predictions based on structured information and those based on unstructured information, the latter of which can be handled only by human analysts.

[5] We should note that due to the weighting procedure for a minority-class oversampling, the output obtained by WRF and this Probit estimation are not exactly the exit probability in the data. Instead it is the probability of exits in the balanced sample consisting of equal numbers of exits and non-exits. Given there is no problem for us to use these probabilities as far as the machine outputs are constructed in the comparable way, we use them in the following empirical analyses. We also construct a ranking based on the outputs obtained by WRF and the Probit estimation and use it for our empirical analysis. The results of which are reported in the online appendix.

Second, machine predictions can take into full account higher dimensions of information than human analysts can. When this is the case, the comparison between *fscore* and machine prediction might account only for the difference between the two different datasets used by humans and machines. Although we think the ability to handle different volumes of information itself is one aspect of the difference between humans and machines and thus worth examining, we also try to compare human and machine predictions on an equal footing in terms of the volume of structured information.

Third, the target of machine and human predictions might not be exactly the same. This issue is called an omitted payoff bias in the literature (Chalfin et al. 2016). As we will detail in the next section, we construct machine learning-based prediction models explicitly targeting one of the two modes of firm exits (i.e., default and voluntary closure), while the *fscore* summarizes the overall performance of a firm. Although the *fscore* is typically used in credit risk management and thus largely accounts for the prospects of firm exits, it is better to have human predictions more directly connected to firm exits.[6] For this purpose, we employ not only the overall firm performance score but also the sub-scores corresponding to the financial stability of firms as human predictions.

[6] TSR guidelines provide the following categorization of *fscore* ranges: (a) caution required (scores 29 and under), (b) medium caution required (scores between 30 and 49), (c) little caution required (scores between 50 and 64), (d) no specific concern (scores between 65 and 79), and (e) no concern at all (scores 80 and above).

Apart from these concerns, the external validity of the results is also important. Disagreements between human and machine predictions may be important in other situations, such as the comparison between machines and investors who put more emphasis on the "upside" of a firm's performance rather than the downside. To address these concerns, we implement the same set of analyses for firms' sales growth and assess the robustness of our results regarding firm exits.

Structured human prediction—As already noted, *fscore* is likely to account for both structured and unstructured information. While it is still informative to compare the original *fscore* with the machine score, we also extract the component of *fscore* associated only with the unstructured information. For this purpose, we construct a machine learning prediction model for *fscore* by using the same right hand-side variables as we use to construct the machine prediction model. This "structured" *fscore* accounts only for the part of *fscore* correlated with the structured information. We use this predicted score and the actual record of exits to run a weighted Probit estimation to transform the structured *fscore* into the probability that is associated with the occurrence of the firm exits.

C. Measurement of "disagreement"

We measure the disagreement between human and machine predictions for a specific exit mode of firm f in year t. We standardize the machine scores of exits, the calibrated *fscore* by a weighted Probit estimation, and the calibrated structured *fscore* as a mean zero and the standard deviation as one. By using these standardized scores for machines (ML), analysts (H), and structured humans (*SH*) that are denoted by $Outcome$, we compute a variable $Proxy$ for a firm (f), analyst (i), and a time (t), which was conceptualized in the previous section, as the following definition:

$$(6) \qquad Proxy_{f,i,t} = Outcome^{ML}_{f,t} - Outcome^{H}_{f,i,t} \text{ for exit firms,}$$

$$= Outcome^{H}_{f,i,t} - Outcome^{ML}_{f,t} \text{ for non-exit firms,}$$

$$(7) \qquad Proxy'_{f,i,t} = Outcome^{SH}_{f,t} - Outcome^{H}_{f,i,t} \text{ for exit firms,}$$

$$= Outcome^{H}_{f,i,t} - Outcome^{SH}_{f,t} \text{ for non-exit firms.}$$

Due to the way we compute $Proxy$, this measure of the disagreement becomes larger when the machine or structured human produces better predictions than the human does.

D. Identifying the determinants of "disagreement"

Once we obtain a measurement of $Proxy$, we can estimate the relationship between $Proxy$ and a linear function $G(\cdot)$ of various explanatory variables that consist of informational opaqueness of firms ($\boldsymbol{O}_{f,t}$), their attributes ($\boldsymbol{F}_{f,t}$), analyst attributes

($\boldsymbol{I}_{i,t}$), and team attributes ($\boldsymbol{Z}_{i,t}$) as well as various configurations of fixed effects ($\boldsymbol{\eta}_{f,i,t}$):

$$(8)\ Proxy_{f,i,t} = G\left(\boldsymbol{O}_{f,t}, \boldsymbol{F}_{f,t}, \boldsymbol{I}_{i,t}, \boldsymbol{Z}_{i,t}\right) + \boldsymbol{\eta}_{f,i,t} + \varepsilon_{f,i,t} \text{ for } t = 2013, \cdots, 2016.$$

In the baseline estimation, we use a firm-level fixed effect, analyst-level fixed effect, and a year-level fixed effect for $\boldsymbol{\eta}_{f,i,t}$, while alternative configurations of fixed effects are also used for the robustness check.

IV. Data

In this section, we provide the details of the data used in our empirical analysis. All the data were obtained from TSR through its joint research contract with Hitotsubashi University. We use multiple datasets to construct a machine prediction model for firm exits to estimate the determinants of $Proxy_{f,i,t}$ and to implement counterfactual exercises.

A. Firm-level panel data

One of our main data sources is an annual-frequency panel of Japanese firm data from t=2010 to 2016 that provide information on firms’ financial statements and basic details such as industry classification, firm characteristics, precise geographic location, and age. The year identifier t accounts for the timing of collection and means that t consists of the data extracted as of the end of December of the year t from TSR. Given a large portion of Japanese firms use an accounting period that ends on March 31, the file labeled t =2012, for example, consists of a large amount of firm information that corresponds to the accounting period up to the end of March 2012. The original data cover around three million firms per year. We use the data that cover around one million firms which provide the information we need for our empirical analysis such as the latest financial statement. According to the Japanese Small and Medium Size Enterprises Agency, there are around three to four million active companies in Japan. The TSR data account for around one-third of that firm population. One point of note is that the sample selection is tilted toward some specific industries, such as construction companies.

These firm-level panel data are accompanied by three types of relational information regarding real and financial partners. First, this information contains a list of up to 10 lender banks. Second, the information also covers firm-to-firm trade. It lists up to 48 customer and supplier firms for each company. In addition to the list of each target firm’s trade partners, we also use the trade relationship reported by those trade partners. As there are many trade relationships not reported by the targeted firms but only by their trade partners, this operation significantly extends the list of trade partners. Third, the data also contain the list of shareholders.

B. Predictions

We consider the two exit outcomes to be predicted one-year ahead: firm default and voluntary closure. The explanatory variables and outcome variable used in constructing a machine prediction model are defined for separate time intervals; explanatory variables from 2010 to 2012 to predict the outcome for the one-year window from the end of 2012 to the end of 2013, explanatory variables from 2011 to 2013 to predict the outcome from the end of 2013 to the end of 2014, and so on. The latest data are the explanatory variables from 2014 to 2016 that are used to predict the outcome from the end of 2016 to the end of 2017.

We measure defaults and voluntary closures as the firms that exited the market for these reasons during the one-year window. Then, we separately prepare two dummy variables that equal one if firms exited through either default or voluntary closure.

C. Firm attributes

To construct a machine prediction model of firm exits, we use the following six categories of attributes of firms: basic characteristics (***firm own***), detailed financial statement information (***financial statement***), geography and industry-related variables (***geo/ind***), firm-bank borrowing relationship variables (***bank***), supply chain network variables (***network***), and shareholder-subsidiary relationship variables (***shareholder***). All the variables categorized in each group are summarized in the online appendix.

We set up the two prediction models for each one of the exit modes using these six groups of firm attributes together with the differenced and double-differenced variables of those variables.[7] We create a set of dummy variables to deal with missing variables that equals one if the corresponding variable is missing for a firm and zero otherwise. When a missing dummy variable equals one, we use zero for the original missing record.

D. Potential determinants of disagreement

To estimate the determinants of the disagreement between human and machine predictions, we set up the following three groups of variables: the amount of available information, firm attributes, and analyst/team attributes.

Number of available variables—As the most important potential determinant in our analysis, which is denoted by $\boldsymbol{O}_{f,t}$, we use the number of variables available (#(*available variables*)) for each firm in the dataset. This number accounts for the opaqueness of firms. When this number is small, both humans and machines can use only a limited amount of structured information. As humans can also utilize

[7] In our data, the predictors and the ex-post outcomes accounting for firm exits are observable. In this sense, our analysis does not suffer from the selective label problem that some of the ex-post outcomes cannot be observed due to selection (Lakkaraju et al. 2017).

soft information, the estimated coefficient associated with #(*available variables*) shows how effectively humans use such soft information in their predictions.

Firm attributes—We use a subset of variables that we used to construct the machine prediction model as the potential determinants, which we denote $\boldsymbol{F}_f$. The list consists of the logarithm of firm sales, its difference, the dummy variable for listed status, and the number of industries the targeted firms operate in. We use this list of variables as they are less prone to missing data.[8] In addition to these variables, we also use the information that relates to the task priority of each firm (*priority*) inside the credit reporting agency that is denoted by a number, with a larger number corresponding to a higher priority. The dataset includes the firm-level panel data of *fscore*. The number is computed as the sum of the four sub-scores that represent the ability of the firm, growth possibility, stability, and reputation. In the following empirical analysis, we use both the *fscore* and the decomposition of each component.

Analyst/Team attributes—We also use the attributes $\boldsymbol{I}_i$ of the analysts. To measure $\boldsymbol{I}_i$, at each data point, we use the attributes of the analysts working for TSR as stored in their anonymized background information. As analysts enter and exit the pool of

[8] Note that the existence of missing data in specific variables can be taken care of by introducing dummy variables that account for the missing data in the non-parametric model such as the random forest we use for constructing the prediction model. Contrary to this, the parametric model such as the panel estimation used for identifying the determinants of the disagreement cannot take care of the missing variables well.

TSR employees, the data become unbalanced panel data. This dataset is accompanied by a table that lists the firms assigned to each analyst at each data point that we use to relate analysts to firms. The dataset allows us to identify the list of assigned firms in each year and the tenure of each analyst. The former information allows us to calculate the number of firms assigned to each analyst and any previous exposure of an analyst to other firms in the industry of the targeted firms, which can be interpreted as the industry expertise of the analyst.

The dataset also allows us to measure the characteristics associated with the team each analyst belongs to, which is denoted by $\boldsymbol{Z}_{i,t}$. First, we measure the size of the team by counting the number of analysts in each department. Second, we measure the average tenure of all members of the team. Third, we measure the average number of firms assigned to the analysts in the team. Fourth, we also measure the average industry expertise of all the analysts in each team.

We understand that this analyst and team information is endogenous as the assignments of analysts to teams and to targeted firms are not random. Thus, we treat these variables simply as control variables in the regression of the determinants for $Proxy_{f,i,t}$ and do not intend to establish any causal relation between these variables and $Proxy_{f,i,t}$.

Table 1 summarizes the variables used to estimate the determinants of the disagreement between human and machine predictions, together with the *fscore*, structured *fscore*, and $Proxy_{f,i,t}$.

Table 1: Summary statistics

Variable	Definition	#samples	min.	25%tile	median	mean	75%tile	max	sd
Disagreement									
Proxy $_{f,\,i,\,t}$	Relative performance of machine predictions for firm *f* . The larger (smaller) value means that machine (analyst *i*) can predict outcome better.	3,983,158	-5.066	-0.95	-0.09	0.00	0.89	5.62	1.29
structured fscore $_{f,\,t}$	Firm *f* 's hypothetical *fscore* considered as analysts could use only hard information for predictions. It is calculated as a replication of *fscore* by machine prediction method.	3,983,158	19.300	43.27	46.19	46.82	49.66	80.95	5.26
Number of available variables									
#(*available variables*) $_{f,\,t}$	The number of firm *f* 's hard information available for predictions.	3,983,158	10	38.00	80.00	91.02	132.00	276	60.42
Firm Characteristics									
log(*sales* $_{f,\,t}$)	The logarithm of firm *f* 's gross sales.	3,983,158	0.000	10.29	11.29	11.37	12.41	23.92	1.86
log(*sales* $_{f,\,t}$)-log(*sales* $_{f,\,t\text{-}1}$)	Log change in firm *f* 's gross sales.	3,983,158	-14.230	-0.06	0.00	0.01	0.07	12.73	0.36
#(*industry*) $_{f,\,t}$	The number of industry codes which are assigned to firm *f* . It takes values from 1 to 3.	3,983,158	1	1.00	2.00	1.92	3.00	3	0.85
priority $_{f,\,t}$	Firm *f* 's relative importance for analysts.	3,810,937	0	0.00	2.00	14.76	8.00	41,290	75.80
fscore $_{f,\,t}$	A score that summarizes an overall performance of firm *f* provided by TSR. It takes values from 0 to 100.	3,983,158	0	43.00	46.00	46.82	50.00	88	5.91
Analyst Characteristics									
#(*tenure years*) $_{i,\,t}$	Analyst *i* 's length of serveice.	3,503,183	0.003	3.59	8.05	10.51	15.38	43.620	8.67
#(*assigned companies*) $_{i,\,t}$	The number of companies for which analyst *i* is responsible to make *fscore* .	3,810,987	1	610	939	1,516	1,862	11,570	1,684.70
industry experience $_{f,\,i,\,t}$	The number of companies (1) having the same industry codes as firm *f* , and (2) having been responsible for analyst *i* to make *fscore* for recent 3 years.	3,810,987	1	27.00	85.00	263.60	271.00	6,241	515.25
Team Characteristics									
#(*team members*) $_{i,\,t}$	The number of colleagues belonging to the same division as analyst *i* .	3,495,647	0	8.00	13.00	15.02	20.00	119	9.70
Average # (*tenure years*) $_{i,\,t}$	Average length of service across team members including analyst *i* .	3,466,648	0.504	7.50	9.76	10.35	12.72	37.19	4.18
Average *industry experience* $_{f,\,i,\,t}$	Average industry experience across team members including analyst *i* .	3,466,648	0	25.67	60.33	117.60	162.30	883.00	136.57
Average #(*assigned companies*) $_{i,\,t}$	Average number of assigned companies across the team members including analyst *i* .	3,466,648	1	920.20	1,230.00	1,407.00	1,877.00	3,543	679.30

V. Empirical Results

A. Prediction performance

The following four panels in Table 2 show the AUCs and their standard errors of the five prediction models for the years 2013 to 2016. The first and second rows show the performance of human predictions and machine predictions, respectively. The third row is for the structured human predictions. The fourth and fifth rows show the performances of machine predictions with different sets of independent variables. The fourth row is the case where we add *fscore* to the list of independent variables used to construct a machine prediction model. The fifth row corresponds to the case where we use only a small set of independent variables to construct a machine prediction model.[9] This smaller set is used to compare human and machine predictions on an equal footing in terms of the volume of structured information.

Table 2: AUC

Test data: *t* = 2013

Model	default	voluntary closure
Human	0.634 (0.0049)	0.719 (0.0030)
Machine	0.793 (0.0041)	0.828 (0.0024)
Structured human	0.617 (0.0046)	0.749 (0.0027)
Machine & *fscore*	0.807 (0.0040)	0.829 (0.0023)
Machine with small information	0.777 (0.0044)	0.829 (0.0024)

Test data: *t* = 2014

Model	default	voluntary closure
Human	0.639 (0.0052)	0.729 (0.0031)
Machine	0.780 (0.0045)	0.828 (0.0024)
Structured human	0.622 (0.0049)	0.757 (0.0028)
Machine & *fscore*	0.794 (0.0043)	0.830 (0.0024)
Machine with small information	0.765 (0.0048)	0.829 (0.0024)

Test data: *t* = 2015

Model	default	voluntary closure
Human	0.653 (0.0055)	0.737 (0.0031)
Machine	0.786 (0.0045)	0.833 (0.0024)
Structured human	0.638 (0.0052)	0.766 (0.0028)
Machine & *fscore*	0.799 (0.0044)	0.835 (0.0024)

Test data: *t* = 2016

Model	default	voluntary closure
Human	0.663 (0.0053)	0.748 (0.0031)
Machine	0.773 (0.0045)	0.841 (0.0025)
Structured human	0.648 (0.0050)	0.776 (0.0027)
Machine & *fscore*	0.789 (0.0044)	0.843 (0.0025)

[9] As the smaller set of variables, we use all the ***firm own*** variables except for dividend-related variables; ***financial statement*** variables that represent total assets, profit, and EBITDA all the ***bank*** variables; ***network*** variables that represent only customers and suppliers with direct links; and ***shareholder*** variables in direct shareholding relations.

Machine with small information	0.768 (0.0050)	0.834 (0.0025)

Machine with small information	0.758 (0.0049)	0.843 (0.0024)

Note: Each number represents the AUC and the number in the parentheses is its standard error.

First, the tables show that the AUC of machine predictions (the second row) is significantly higher than that of human predictions (the first row) given the size of the standard errors of those AUCs. This is the case even when we use the smaller set of independent variables (the fifth row). Thus, machine predictions outperform human predictions on average.

Second, in the case of predicting default, humans outperform structured humans (the first and third rows). We also find that *fscore* makes an additional contribution to the overall performance of the machine predictions (the second and fourth rows). These results contrast with the findings of Kleinberg et al. (2018). In their empirical analysis of judicial bail decisions, they report that the structured human does a better job of predicting risky criminals than the judge. They claim that the "psychologist's view" in which humans make noisy predictions overwhelms the "economist's view" in which humans can use soft information to make a better prediction. Our result shows that at least in our setup for default predictions, the economist's view should be more reliable. Furthermore, as for predicting voluntary closure, the structured human does a better job than the human does, which is consistent with the psychologist's view.[10]

B. Determinants of disagreement

Table 3 summarizes the results of the panel estimation associated with default and voluntary closure. All the coefficients are shown in the percent point (i.e., the estimated coefficients times 100).

Table 3: Baseline estimation

	default				***voluntary closure***			
	Machine vs. Human		***SH vs. Human***		***Machine vs. Human***		***SH vs. Human***	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Number of available variables								
#(*available variables*) $_{f,t}$	0.566	0.001	0.041	0.000	0.485	0.001	0.031	0.000
Firm characteristics								
log(*sales* $_{f,t}$)	-18.545	0.127	3.987	0.028	-8.511	0.111	5.036	0.030
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	13.015	0.097	-0.618	0.022	5.205	0.086	-0.521	0.023
listed $_{f,t}$	-2.105	2.758	0.605	0.621	-18.931	2.429	-6.351	0.662
#(*industry*) $_{f,t}$	-3.009	0.159	-0.084	0.036	0.097	0.140	-0.129	0.038
priority $_{f,t}$	0.001	0.000	0.000	0.000	0.002	0.000	-0.000	0.000
Analyst characterstics								
#(*assigned companies*) $_{i,t}$	-0.001	0.000	-0.000	0.000	-0.001	0.000	-0.000	0.000
industry experience $_{f,i,t}$	-0.004	0.000	0.000	0.000	-0.001	0.000	0.001	0.000
Team characteristics								
#(team members) $_{i,t}$	0.081	0.012	-0.001	0.003	0.106	0.010	-0.001	0.003
Average #(*tenure years*) $_{i,t}$	0.136	0.016	-0.008	0.004	-0.008	0.014	-0.006	0.004
Average industry experience $_{f,i,t}$	0.014	0.001	0.000	0.000	0.001	0.001	0.000	0.000
Average #(*assigned companies*) $_{i,t}$	-0.001	0.000	-0.000	0.000	-0.002	0.000	-0.000	0.000
Constant	152.997	1.512	-49.111	0.340	54.692	1.331	-59.965	0.363
Firm fixed-effect	yes		yes		yes		yes	
Analyst fixed-effect	yes		yes		yes		yes	
Year fixed-effect	yes		yes		yes		yes	
#(obs)	3,238,817		3,238,817		3,238,817		3,238,817	
F	14,314.100		3,591.740		12,417.240		3,908.300	
Adj. R-squared	0.879		0.789		0.831		0.777	
Within R-squared	0.071		0.019		0.062		0.020	

[10] In the online appendix, we examine the recall and precision measures for machine, human, and structured human predictions over different thresholds of prediction.

From the columns labeled as "*Machine vs. Human*", regardless of whether we use default or voluntary closure as the prediction target, we find that the prediction performance of humans relative to machines becomes better for firms with less observable information on their attributes (i.e., lower values for #(*available variables*)). Thus, for more opaque firms, the relative performance of human predictions to machine predictions improves.

Why do analysts perform better in the case of opaque firms? One conjecture is that analysts are using unstructured information that by definition, cannot be used in machine predictions. To confirm this conjecture, we also run the panel regression for $Proxy'_{f,i,t}$ that is defined by replacing $Outcome^{ML}_{f,t}$ with $Outcome^{SH}_{f,i,t}$. This regression characterizes under what conditions human predictions outperform those of the structured humans. The results in the columns labeled as "*SH vs. Human*" show a similar pattern to that in "*Machine vs. Human*", that is, the relative power of human predictions compared with structured human predictions becomes higher as the amount of available information becomes smaller.[11]

We also separately regress the performance of human and machine predictions on the same set of attributes. From the estimation results (reported in the online appendix), we confirm that the negative marginal effect associated with lower

[11] We also find that the marginal effect of the available information on the relative performance of human predictions compared to that of structured human predictions is much smaller than that for humans vs. machines. This difference means that the sensitivity of the structured human predictions to the level of available information is much lower than that of machine predictions.

availability of information is greater for machine predictions than for human predictions. This effect could be the case again when humans effectively use unstructured information to make predictions.

To check the robustness of the results and address the concerns we raised in the previous section, we first use alternative methods of measuring the disagreement between human and machine predictions. As detailed above, we are using the ex-post record of firm exits to obtain the probabilities of exit that are measured by *fscore* and the structured *fscore*. As the transformation of *fscore* to the probability is simply a monotonic transformation and does not change the order of the score, it does not affect the comparison of human and machine predictions. Nonetheless, in reality, such an ex-post record of exit that is used to calibrate *fscore* to probability is not attainable in the process of human predictions. Thus, we also construct a set of "rankings" based on the machine prediction, *fscore*, and the structured *fscore*. In this ranking of prediction outcomes, we do not need to refer to the ex-post default records for the purpose of calibration. Second, we also define a dummy variable that is equal to one if $Proxy_{f,i,t}$ is positive and zero otherwise. We use this dummy variable and run a linear probability model with the abovementioned fixed effects and conditional logit model with firm-level fixed effects. We also set 1 to 10 variables depending on the level of $Proxy_{f,i,t}$ and run an ordered-logit estimation without fixed effects. Third, we replace the analyst-level fixed effect with the analyst-year-level fixed effect so that we can take complete account of analyst-level unobservable factors that vary over time and that subsume team-level time-variant unobservable factors. These are not likely to be captured by the limited number of explanatory variables $\boldsymbol{I}_i$ and $\boldsymbol{Z}_{i,t}$. Fourth, we use one of the sub-scores of *fscore*, which represents the stability of a firm, instead of the total *fscore*, so that the target of human predictions becomes more comparable to that of machine predictions. Fifth, instead of weighted random forest, we use LASSO or extreme gradient boost for producing machine predictions. All the results are shown in the online appendix and are consistent with the results in Table 3.

C. Counterfactual exercises

Can we use the empirical findings presented in the previous section to improve the overall performance of predictions on firm exits? Given that the performance of humans relative to machines improves for more opaque firms, then agencies will naturally assign these firms to humans and firms with greater information to machines.

Based on this conjecture, we split the sample into five subsamples according to the number of observable variables. We aim at setting up multiple groups for which the relative performance of humans differs from that of machines. To construct subgroups purely tied up to the number of observable variables, we regress #(available variables) to a firm's sales, growth, and industry classification that are significant in the estimation of $Proxy_{f,i,t}$ and take out the residual. Then, we use this residual to sort the firms and construct five subsamples so that we can set up five groups of firms depending on the level of #(available variables) that is orthogonal to other firm attributes.

In each subsample, we evaluate the performances of human and machine predictions. By comparing, for example, the number of false negatives based on machine predictions *(ML)* to those based on human predictions *(H)* for the same set of firms, we can describe what happens to the prediction performance for the subsample by reallocating predictions to humans instead of machines.

Table 4: Reallocation of predictions instances

(a) Firms actually do *NOT* exit ex post

	Prediction for default			*Prediction for voluntary closure*		
	ML = default *H* = not default (1)	*ML* = not default *H* = default (2)	(2)/(1)	*ML* = closure *H* = not closure (1)	*ML* = not closure *H* = closure (2)	(2)/(1)
~20 %tile	49,117	23,068	0.47	25,206	19,453	0.77
20~40 %tile	36,094	54,446	1.51	28,326	23,667	0.84
40~60 %tile	37,362	46,368	1.24	28,370	28,134	0.99
60~80 %tile	33,409	39,218	1.17	20,249	30,962	1.53
80 %tile~	11,652	30,608	2.63	8,026	34,406	4.29

(b) Firms actually do exit ex post

	Prediction for default			*Prediction for voluntary closure*		
	ML = default *H* = not default (3)	*ML* = not default *H* = default (4)	(3)/(4)	*ML* = closure *H* = not closure (3)	*ML* = not closure *H* = closure (4)	(3)/(4)
~20 %tile	88	21	4.19	140	51	2.75
20~40 %tile	82	40	2.05	195	42	4.64
40~60 %tile	86	37	2.32	231	43	5.37
60~80 %tile	74	37	2.00	174	54	3.22
80 %tile~	38	27	1.41	72	45	1.60

Note: ML and *H* denote the predictions of machines and humans, respectively.

The two panels in Table 4 summarize the number of false-positive, false-negative, true-positive, and true-negative cases for the five subsamples. We treat the top 30% of firms in terms of the prediction score as the firms predicted to exit.[12]

For example, the columns marked (1) in panel (a), show the number of false-positives for machine predictions and true-negatives for human predictions, as these columns show the number of firms that do *not* exit ex post. Conversely, the columns marked (2) in panel (a) show the number of true-negatives for machine predictions and false-positives for human predictions for firms that do not exit ex

[12] For robustness check, we vary this prediction threshold (i.e., the top 30% in this baseline exercise) from the top 50% to the top 20% and confirm the results do not change.

post. Panel (b) in Table 4 summarizes the number in the same manner but for the firms that actually *do* exit ex post.

Comparing the numbers in each column, we can see how type I and type II errors vary depending on whether the predictions are allocated to machines or to humans. In six out of the 10 rows in Panel (a), the number in column (1) is smaller than that in column (2), while in Panel (b), all the numbers in column (3) are larger than those in column (4).

First, these results mean that the type II error is always smaller in machine predictions than in human predictions regardless of the level of available information. Even for the firms with the least information, human predictions cannot outperform machine predictions. Second, in the case of the firms with the least information (i.e., the first raw labeled as "~20%tile"), it is still possible to reduce the number of false-positives, and thus reduce the type I error, by reallocating the default predictions to humans instead of to machines (i.e., the number of false-positives is reduced from 49,117 to 23,068). In the case of voluntary closure, we can also achieve a smaller type I error for firms with the least, little, and average amounts of information (i.e., the first, second, and third raws labeled "~20%tile", "20~40%tile", and "40~60%tile") by reallocating the default predictions to humans instead of machines.

However, a reallocation of predictions is accompanied by a larger type II error, as shown above. The numbers in column (3) are always larger than those in column

(4) that indicates the reallocation of predictions always increases the number of false-negatives. As one interesting result, we also find that in the case of default predictions, the ratio is larger as we move from the subsample with the least information to that with the largest amount. This pattern is inconsistent with the positive coefficient obtained in our estimation of $Proxy_{f,i,t}$. This is the case simply because, in our data, the number of exits is much smaller than that of non-exits. In other words, the performance of human predictions relative to machine predictions with respect to the level of available information is driven by human predictions correctly predicting non-exit firms.

These results reconfirm the usefulness of machine prediction techniques in the context of exit predictions. There is however room for human predictions to outperform machine predictions under specific circumstances, such as when the prediction targets are informationally opaque or when the user of the prediction results is more concerned with a type I error than a type II error due to, for example, the imbalance between the numbers of exit and non-exit firms.

D. Growth prediction

We have so far focused on exit predictions. What happens if we focus on the upside of firm dynamics instead? We repeat the same analyses by considering firm growth as the target of our predictions. We define growth in sales as a rate of one standard deviation higher than the industry average defined in two digits over the one-year window used to measure the outcome. Then, we prepare a dummy variable that equals one if firms experience a growth rate higher than these criteria.

As predictions for upside events are the opposites of downside predictions, we conjecture that while overall performance is still higher for machine predictions than human predictions, and the relative performance of human predictions also improves when the available information is smaller as we have described, the source of this better performance is not from a lower type I error but from a lower type II error. In other words, analysts more correctly predict growth for actually growing firms based on less information. As presented in the online appendix, this is indeed the case. Although the levels of type I and type II errors are always higher in the case of human predictions, relative prediction performance of analyst to machine improves for actually growing firms as available information becomes smaller.

VI. Conclusion

We empirically examine the relative performance of machine and human subjective predictions for firm exits. Using a huge volume of firm-level high-dimension panel data, we find that human predictions are not as accurate as machine predictions on average. As for predicting the exits of informationally opaque firms, the relative performance of human predictions improves.

One important point is that when using machine predictions in practice, Luca et al. (2016) claim that they cannot ensure automated decision-making as it is necessary to take into account the various dimensions of the problems under consideration. This study provides evidence that accounting for the conditions under which a prediction is to be assigned to a machine is also necessary. Our findings cast light on the circumstances and the extent to which tasks should be allocated either to machines or to humans.

Future extensions of the present study may benefit from the inclusion of additional explanatory variables as determinants of $Proxy$. A large-sized aggregate-level shock, such as a market downturn or a natural disaster, could have a marginal effect on each determinant of $Proxy$. Understanding potentially relevant shocks is useful in considering how we should allocate prediction tasks to machines and humans under specific circumstances. Such an additional analysis will help us to understand both the nature of human error and how humans and machines can work together to provide accurate predictions.

REFERENCES

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics* Vol. 4B, edited by Ashenfelter, Orley, and David Card, 1043-1171. Amsterdam: North-Holland.

Acemoglu, Daron, and Pascual Restrepo. 2018. "Artificial Intelligence, Automation and Work." NBER Working Paper No. 24196.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press.

Anderson, Ashton, Jon Kleinberg, and Sendhil Mullainathan. 2017. "Assessing Human Error Against a Benchmark of Perfection." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11 (4), 45:1-25.

Athey, Susan. 2019. "The Impact of Machine Learning on Economics." In *The Economics of Artificial Intelligence: An Agenda*, edited by Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, chapter 21. University of Chicago Press.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118 (4): 1279-1333.

Bazzi, Samuel, Robert A. Blair, Christopher Blattman, Oeindrila Dube, Matthew Gudgeon, and Richard Merton Peck. 2019. "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia." NBER Working Paper No. 25980.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review* 106 (5): 124-27.

Chen, Chao, Andy Liaw, and Leo Breiman. 2004. “Using Random Forest to Learn Imbalanced Data.” Technical Report 666 Statistics Department of University of California at Berkley.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. “The Future of Employment: How Susceptible Are Jobs to Computerisation? *Technological Forecasting and Social Change* 114: 254-280.

Gebru, Timnit. 2020. "Race and Gender." In *The Oxford Handbook of Ethics of AI* ch. 13, edited by Dubber, Markus D., Frank Pasquale, and Sunit Das, 251–269. Oxford University Press.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human Decisions and Machine Predictions.” *Quarterly Journal of Economics* 133 (1): 237-293.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. “The Selective Labels Problem: Evaluating Algorithmic.” *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* August 2017: 275-284.

Liberti, José María, and Mitchell A. Petersen. 2019. “Information: Hard and Soft.” *Review of Corporate Finance Studies* 8 (1): 1-41.

Lin, Zhiyuan “Jerry”, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. “The Limits of Human Predictions of Recidivism.” *Science Advances* 6 (7).

Luca, Michael, Jon Kleinberg, and Sendhil Mullainathan. 2016. “Algorithms Need Managers, Too.” *Harvard Business Review* 94 (1/2): 96-101.

McIloroy-Young, Reid, Siddhartha sen, Jon Kleinberg, and Ashton Anderson. 2020. “Aligning Superhuman AI with Human Behavior: Chess as a Model System.” *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* August 2020: 1677-1687.

Mei, Xueyan, Hao-Chih Lee, Kai-yue Diao, Mingqian Huang, Bin Lin, Chenyu Liu, Zongyu Xie, Yixuan Ma, Phillip M. Robson, Michael Chung, Adam Bernheim, Venkatesh Mani, Claudia Calcagno, Kunwei Li, Shaolin Li, Hong Shan, Jian Lv, Tongtong Zhao, Junli Xia, Qihua Long, Sharon Steinberger, Adam Jacobi, Timothy Deyer, Marta Luksza, Fang Liu, Brent P. Little, Zahi A. Fayad, and Yang. 2020. "Artificial Intelligence-enabled Rapid Diagnosis of Patients with COVID-19." *Nature Medicine*.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87-106.

Patel, Bhavik N., Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, Curtis Langlotz, Edward Lo, Joseph Mammarappallil, A. J. Mariano, Geoffrey Riley, Jayne Seekins, Luyao Shen, Evan Zucker, and Matthew P. Lungren. 2019. "Human-machine Partnership with Artificial Intelligence for Chest Radiograph Diagnosis." *npj Digital Medicine* 2.

Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. "The Algorithmic Automation Problem: Prediction, Triage, and Human Effort." arXiv:1903.12220.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspective* 28 (2): 3-28.

The research data for this article

The data used in this study are proprietary, and we gained access to the data through a joint research contract between Miyakawa’s institute and TSR. Therefore, we cannot provide the data to the journal.

Acknowledgments

This study was conducted as a part of a joint research project of Hitotsubashi University and Tokyo Shoko Research Ltd. (TSR). We thank Hiroshi Fujiki, Kaoru Hosono, Kentaro Iwatsubo, Kiyoshi Izumi, Koichi Kume, Makoto Nirei, Kazuhiko Ohashi, Arito Ono, Shigenori Shiratsuka, Daisuke Tsuruta, Hirofumi Uchida, Iichiro Uesugi, Toshiaki Watanabe, Yasutora Watanabe, and Isamu Yamamoto. We also want to thank the seminar participants at SigFin, Hitotsubashi University, and RIETI; the participants at the 2019 Finance Workshop held by the Institute for Monetary and Economic Studies; the Bank of Japan; 2020 JEA spring meeting; 28th NFA annual meeting; and the staff of the Bank of Japan for helpful suggestions. This study was supported by the Grant-in-Aid for Scientific Research (B) of the Japan Society for the Promotion of Science [grant number 17H02526]. Further, it was prepared in part while Miyakawa was a visiting scholar at the Institute for Monetary and Economic Studies, Bank of Japan. The views expressed in this study are those of the authors and do not necessarily reflect the official views of the Bank of Japan.

For Online Publication

Online Appendix A

The list of variables we use to construct the machine learning prediction model is as follows:

Firm-own characteristics (*firm own*): As variables that represent the firms' own characteristics, we use size as measured by the logarithm of sales and the change in sales from the previous period, profit-to-sales ratio and any change from the previous period, the status of dividend payments (paid or not) and any change from the previous period, whether the firm is listed or not, the number of employees, the logarithm of stated capital, and dummy variables that represent industry classification (note: multiple industry codes are recorded). We also use firm age, owner age, and the number of establishments.

Firms' financial statement information (*financial statement*): We set up a number of financial variables used in the literature to represent firms' detailed financial statement information.[13]

[13] The list of "***financial statement***" variables consists of the following items: Logarithm of total assets, cash-to-total assets ratio, liquid assets-to-total assets ratio, tangible assets-to-total assets ratio, receivables turn-over, inventory turn-over, total liability-to-total assets ratio, liquid liability-to-total assets ratio, bond-to-total liability ratio, bank borrowing-to-total liability ratio, bank short borrowing-to-total bank borrowing ratio, payables turn-over, interest coverage ratio, liquid assets-to-liquid liability ratio, fixed compliance ratio, fixed ratio, working capital turn-over, gross profit-to-sales ratio, operating profit-to-sales ratio, ordinary profit-to-sales ratio, net profit before tax-to-sales ratio, logarithm of EBITDA, logarithm of EBITDA-to-sales ratio, special income-to-sales ratio, special expenses-to-sales ratio, and labor productivity.

Industry and geographical information (*geo/ind*): We set up the following two groups of variables to represent the industry and area to which the firms belong. First, we construct the variables measuring the average sales growth of firms located in the same city as the targeted firms. Second, we compute the average sales growth of firms belonging to the same industry that are classified at the 2-digit level.

Lender banks information (*bank*): As variables that represent the firms' borrowing relationships with lender banks, we construct a dummy variable to represent a change in main lenders (i.e., top lender bank) or in the number of lender banks.

Supply-chain linkage information (*network*): We construct the following two groups of variables to represent the supply chain network. First, we compute widely used network metrics for each firm by using the network information on the supply chain. The metrics consist of degree centrality; eigenvector centrality; egonet eigenvalue; co-transaction; and the number of transaction partners, both direct (i.e., customers and suppliers) and indirect (e.g., suppliers' suppliers, and customers' suppliers). Second, we construct a number of variables that represent the characteristics of transaction partners. To summarize this information, we use the average, maximum, minimum, and the sum of *fscore* associated with each transaction partner. Note that while the network metrics cover both direct and indirect transaction partners, the transaction partners’ characteristics only cover direct transaction partners.

Shareholder linkage information (*shareholder*): We set up similar variables to those for the supply chain network as predictors of shareholder information.

Online Appendix B

We list the tables and figures referred to in the study for the robustness check. First, we show an alternative way to compare the prediction power of machines, humans, and structured humans (Figure A1). We can confirm that machine predictions outperform human predictions on average. Regarding the comparison between human predictions and those of the structured human predictions, human predictions are more precise in the case of default predictions, while the structured human predictions are better in terms of voluntary closure. Second, instead of estimating the determinants of $Proxy_{f,i,t}$, we estimate separately the determinants of $Proxy^{m}_{f,t}$ and $Proxy^{h}_{f,i,t}$, that represent the prediction performances of machines and humans, respectively. Comparing the estimated coefficients associated with the independent variables, we can see how the respective prediction powers of machines and humans vary according to the change in determinants (Table A1).

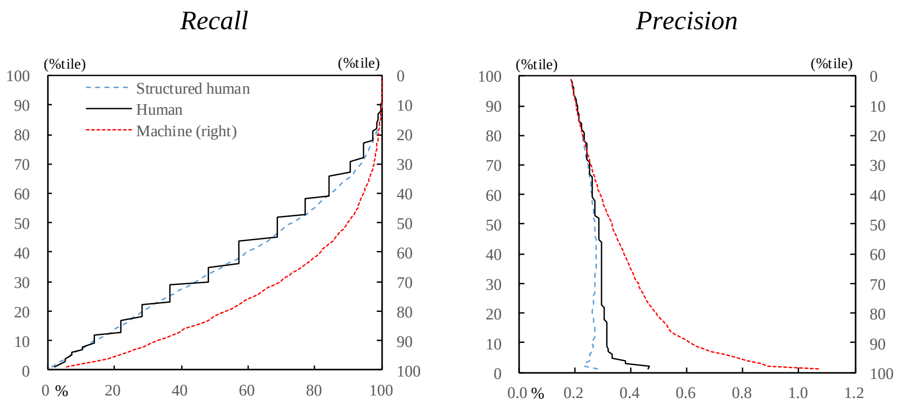
(A1) $$Proxy^{m}_{f,t} = Outcome^{ML}_{f,t} - 1 \text{ for exit firms,}$$
$$= 1 - Outcome^{ML}_{f,t} \text{ for non-exit firms,}$$

(A2) $$Proxy^{h}_{f,i,t} = Outcome^{H}_{f,i,t} - 1 \text{ for exit firms,}$$
$$= 1 - Outcome^{H}_{f,i,t} \text{ for non-exit firms.}$$

Third, we construct a set of rankings based on the machine prediction, *fscore*, and structured *fscore* and repeat the same estimation for the disagreement (Table A2). Fourth, we also define a dummy variable that equals one if $Proxy_{f,i,t}$ is positive and zero otherwise. Then we run a linear probability model and conditional logit model (Table A3). We also set 1 to 10 variables, which depend on the level of $Proxy_{f,i,t}$, and run an ordered-logit estimation (Table A4). Fifth, we replace the analyst-level fixed effect with the analyst-year-level fixed effect (Table A5). Sixth, we use one of the sub-scores of *fscore*, which represents the stability of each firm, instead of the total *fscore*, so that the target of human predictions becomes plausibly more comparable to that of machine predictions (Table A6). Seventh, we summarize the results of the proxy estimation and counterfactual exercise representing firm growth (Table A7). Eighth, we repeat the AUC estimation and proxy estimation based on the two alternative methods (i.e., LASSO and extreme gradient boost) (Table A8, A9). All the results are consistent with the ones we presented in the study.

Figure A1: Recall and precision measures over different thresholds

Default (test year: *t*=2016)



Voluntary closure (test year: *t*=2016)

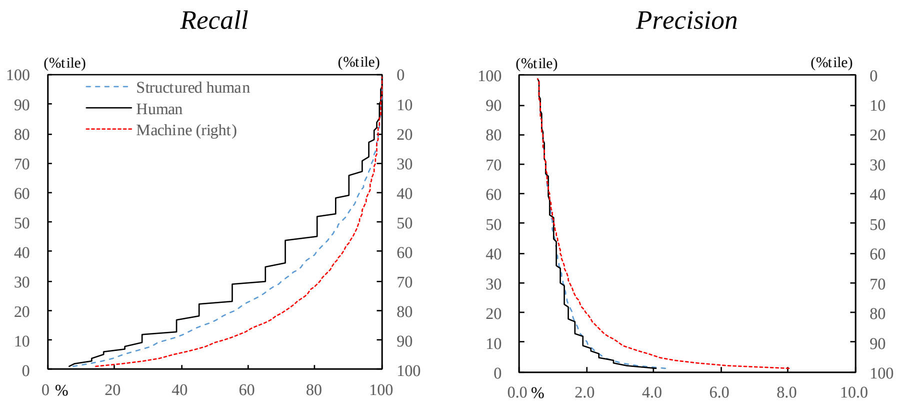


Table A1: Prediction performance of machines and humans

	default				***voluntary closure***			
	Machine		***Human***		***Machine***		***Human***	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Number of available variables								
#(*available variables*) $_{f,t}$	0.102	0.000	0.008	0.000	0.118	0.000	0.012	0.000
Firm characteristics								
log(*sales* $_{f,t}$)	2.318	0.020	5.024	0.014	6.461	0.021	7.493	0.021
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	1.701	0.015	-0.440	0.011	0.231	0.017	-0.760	0.016
listed $_{f,t}$	2.477	0.443	2.621	0.303	-1.838	0.481	2.168	0.467
#(*industry*) $_{f,t}$	-0.502	0.025	0.099	0.017	0.244	0.027	0.202	0.027
priority $_{f,t}$			0.000	0.000			0.000	0.000
Analyst characterstics								
#(*assigned companies*) $_{i,t}$			0.000	0.000			0.000	0.000
industry experience $_{f,i,t}$			-0.000	0.000			-0.000	0.000
Team characteristics								
#(team members) $_{i,t}$			0.002	0.001			-0.005	0.002
Average #(*tenure years*) $_{i,t}$			0.014	0.002			0.016	0.003
Average industry experience $_{f,i,t}$			-0.000	0.000			0.000	0.000
Average #(*assigned companies*) $_{i,t}$			0.000	0.000			0.000	0.000
Constant	29.191	0.226	-4.012	0.166	-19.798	0.245	-28.631	0.256
Firm fixed-effect	yes		yes		yes		yes	
Analyst fixed-effect	yes		yes		yes		yes	
Year fixed-effect	yes		yes		yes		yes	
#(obs)	3,756,803		3,238,817		3,756,803		3,238,817	
F	53,485.400		15,304.020		78,182.190		14,025.710	
Adj R-squared	0.815		0.897		0.876		0.866	
Within R-squared	0.092		0.075		0.129		0.069	

Table A2: Rank-based disagreement estimation

	Machine vs. Human			
	default		***voluntary closure***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	1,607.929	4.271	1,527.788	3.784
Firm characteristics				
log(*sales* $_{f,t}$)	-58,115.530	374.526	-25,088.000	331.840
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	37,273.310	287.922	16,041.170	255.107
listed $_{f,t}$	27,956.380	8,164.855	-34,210.110	7,234.288
#(*industry*) $_{f,t}$	-8,595.519	471.108	620.723	417.415
priority $_{f,t}$	5.258	1.144	8.109	1.013
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	-1.894	0.313	-3.357	0.277
industry experience $_{f,i,t}$	-11.528	0.604	-6.217	0.535
Team characteristics				
#(team members) $_{i,t}$	268.315	34.572	346.771	30.632
Average #(*tenure years*) $_{i,t}$	384.545	48.371	-63.242	42.858
Average industry experience $_{f,i,t}$	39.630	2.346	-2.152	2.079
Average #(*assigned companies*)	-2.936	0.437	-5.742	0.387
Constant	470,115.500	4,475.366	125,805.500	3,965.298
Firm fixed-effect	yes		yes	
Analyst fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,238,817		3,238,817	
F	13,426.970		13,873.310	
Adj. R-squared	0.876		0.820	
Within R-squared	0.067		0.069	

Table A3: Dummy variable measure for disagreement

(1) Linear probability model

	Machine vs. Human			
	default		***voluntary closure***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	0.157	0.001	0.265	0.001
Firm characteristics				
log(*sales* $_{f,t}$)	-5.664	0.076	-3.578	0.085
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	4.064	0.059	2.315	0.065
listed $_{f,t}$	2.856	1.664	-7.332	1.849
#(*industry*) $_{f,t}$	-1.350	0.096	0.042	0.107
priority $_{f,t}$	0.001	0.000	0.002	0.000
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	-0.000	0.000	-0.001	0.000
industry experience $_{f,i,t}$	-0.001	0.000	-0.000	0.000
Team characteristics				
#(team members) $_{i,t}$	0.041	0.007	0.041	0.008
Average #(*tenure years*) $_{i,t}$	0.005	0.010	0.005	0.011
Average industry experience $_{f,i,t}$	0.006	0.000	0.000	0.001
Average #(*assigned companies*) $_{i,t}$	-0.001	0.000	-0.001	0.000
Constant	93.738	0.912	59.737	1.014
Firm fixed-effect	yes		yes	
Analyst fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,238,817		3,238,817	
F	3,135.790		6,343.690	
Adj. R-squared	0.721		0.659	
Within R-squared	0.016		0.033	

(2) Conditional logit model

	Machine vs. Human			
	default		***voluntary closure***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	1.942	0.013	2.587	0.012
Firm characteristics				
log(*sales* $_{f,t}$)	-87.264	1.207	-42.894	1.011
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	65.887	0.962	28.807	0.783
listed $_{f,t}$	45.617	25.010	-82.705	20.077
#(*industry*) $_{f,t}$	-20.860	1.326	-6.271	1.235
priority $_{f,t}$	0.095	0.014	0.072	0.008
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	0.000	0.001	0.000	0.000
industry experience $_{f,i,t}$	0.006	0.001	-0.002	0.001
Team characteristics				
#(team members) $_{i,t}$	0.425	0.071	0.409	0.065
Average #(*tenure years*) $_{i,t}$	-0.241	0.114	-0.067	0.104
Average industry experience $_{f,i,t}$	0.022	0.006	-0.104	0.005
Average #(*assigned companies*) $_{i,t}$	-0.003	0.001	-0.002	0.001
Constant				
Firm fixed-effect	yes		yes	
Analyst fixed-effect	no		no	
Year fixed-effect	no		no	
#(obs)	736,498		922,303	
Log-likelihood	-259,176.670		-315,385.000	
χ-squared	30,953.570		57,174.730	

Table A4: Ordered logit estimation

	Machine vs. Human			
	default		***voluntary closure***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	1.214	0.005	2.262	0.005
Firm characteristics				
log(*sales* $_{f,t}$)	-171.686	0.244	-22.596	0.210
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	103.072	0.390	26.065	0.366
listed $_{f,t}$	542.157	6.472	-103.528	5.877
#(*industry*) $_{f,t}$	-48.697	0.389	-1.500	0.385
priority $_{f,t}$	0.086	0.003	0.010	0.002
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	0.001	0.000	-0.001	0.000
industry experience $_{f,i,t}$	0.047	0.001	0.032	0.001
Team characteristics				
#(team members) $_{i,t}$	2.314	0.028	2.805	0.028
Average #(*tenure years*) $_{i,t}$	-0.375	0.049	-0.498	0.049
Average industry experience $_{f,i,t}$	0.255	0.002	0.297	0.002
Average #(*assigned companies*) $_{i,t}$	-0.030	0.000	-0.041	0.000
Constant				
Firm fixed-effect	no		no	
Analyst fixed-effect	no		no	
Year fixed-effect	no		no	
#(obs)	3,466,611		3,466,611	
Log-likelihood	-6,008,220.100		-6,508,573.100	
χ-squared	621,072.400		253,758.480	

Table A5: Alternative fixed-effects specification

	Machine vs. Human			
	default		***voluntary closure***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	0.571	0.001	0.482	0.001
Firm characteristics				
log(*sales* $_{f,t}$)	-19.063	0.125	-8.293	0.111
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	13.213	0.096	5.074	0.085
listed $_{f,t}$	-4.449	2.732	-19.247	2.412
#(*industry*) $_{f,t}$	-3.538	0.158	0.002	0.140
priority $_{f,t}$	0.000	0.000	0.002	0.000
Analyst characterstics				
#(*assigned companies*) $_{i,t}$				
industry experience $_{f,i,t}$	0.001	0.000	0.000	0.000
Team characteristics				
#(team members) $_{i,t}$				
Average #(*tenure years*) $_{i,t}$				
Average industry experience $_{f,i,t}$	0.017	0.001	0.000	0.001
Average #(*assigned companies*) $_{i,t}$				
Constant	157.847	1.465	49.298	1.293
Firm fixed-effect	yes		yes	
Analyst-Year fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,238,266		3,238,266	
F	22,197.050		18,409.250	
Adj. R-squared	0.882		0.834	
Within R-squared	0.073		0.061	

Table A6: Using sub-score as human predictions

	default				***voluntary closure***			
	Machine vs. Human		***SH vs. Human***		***Machine vs. Human***		***SH vs. Human***	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Number of available variables								
#(*available variables*) $_{f,t}$	0.637	0.002	0.018	0.000	0.519	0.002	0.018	0.000
Firm characteristics								
log(*sales* $_{f,t}$)	5.178	0.191	3.120	0.044	13.864	0.166	3.240	0.044
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	17.783	0.142	-2.203	0.033	13.444	0.123	-2.283	0.033
listed $_{f,t}$	8.962	3.434	4.606	0.787	-9.880	2.974	4.304	0.787
#(*industry*) $_{f,t}$	-2.132	0.227	0.090	0.052	1.092	0.197	0.086	0.052
priority $_{f,t}$	0.000	0.000	0.000	0.000	0.001	0.000	-0.000	0.000
Analyst characterstics								
#(*assigned companies*) $_{i,t}$	-0.002	0.000	0.000	0.000	0.000	0.000	0.001	0.000
industry experience $_{f,i,t}$	-0.003	0.000	0.001	0.000	0.002	0.000	0.001	0.000
Team characteristics								
#(team members) $_{i,t}$	0.028	0.019	-0.017	0.004	0.026	0.017	-0.018	0.004
Average #(*tenure years*) $_{i,t}$	0.080	0.026	-0.046	0.006	-0.078	0.022	-0.047	0.006
Average industry experience $_{f,i,t}$	0.026	0.001	-0.002	0.000	-0.005	0.001	-0.002	0.000
Average #(*assigned companies*) $_{i,t}$	0.001	0.000	0.000	0.000	-0.001	0.000	0.000	0.000
Constant	-132.004	2.359	-38.266	0.540	-212.930	2.044	-39.522	0.540
Firm fixed-effect	yes		yes		yes		yes	
Analyst fixed-effect	yes		yes		yes		yes	
Year fixed-effect	yes		yes		yes		yes	
#(obs)	2,199,518		2,199,518		2,199,518		2,199,518	
F	10,515.140		719.200		11,101.810		752.040	
Adj. R-squared	0.825		0.712		0.830		0.718	
Within R-squared	0.081		0.006		0.085		0.006	

Table A7: Growth prediction

(1) $Proxy$ estimation

	Machine vs. Human		***SH vs. Human***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	0.196	0.003	0.037	0.000
Firm characteristics				
log(*sales* $_{f,t}$)	-50.833	0.229	-0.166	0.039
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	14.032	0.174	-0.439	0.030
listed $_{f,t}$	-24.028	4.837	3.056	0.830
#(*industry*) $_{f,t}$	-1.239	0.281	0.036	0.048
priority $_{f,t}$	0.005	0.001	0.000	0.000
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	-0.000	0.000	-0.000	0.000
industry experience $_{f,i,t}$	0.003	0.000	0.000	0.000
Team characteristics				
#(team members) $_{i,t}$	-0.167	0.021	-0.008	0.004
Average #(*tenure years*) $_{i,t}$	-0.357	0.029	-0.014	0.005
Average industry experience $_{f,i,t}$	-0.017	0.001	0.000	0.000
Average #(*assigned companies*) $_{i,t}$	0.001	0.000	-0.000	0.000
Constant	574.761	2.737	-0.627	0.470
Firm fixed-effect	yes		yes	
Analyst fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,037,588		3,037,588	
F	4,799.540		650.920	
Adj. R-squared	0.590		0.639	
Within R-squared	0.026		0.004	

(2) Counterfactual exercise

(a) Firms that actually do not grow ex post

	M = growth H = not growth (1)	M = not growth H = growth (2)	(2)/(1)
~20 %tile	12,799	30,678	2.40
20~40 %tile	15,822	38,401	2.43
40~60 %tile	18,513	31,610	1.71
60~80 %tile	25,171	22,727	0.90
80 %tile~	34,835	11,263	0.32

(b) Firms that actually grow ex post

	M = growth H = not growth (3)	M = not growth H = growth (4)	(3)/(4)
~20 %tile	1765	791	2.23
20~40 %tile	2170	978	2.22
40~60 %tile	2660	883	3.01
60~80 %tile	3599	760	4.74
80 %tile~	5308	401	13.24

Table A8: AUCs of alternative prediction models for default

Test data: t = 2013

Model	LASSO	XGBoost
Human	0.634 (0.0049)	
Machine	0.783 (0.0042)	0.807 (0.0039)
Structured human	0.529 (0.0047)	0.598 (0.0046)
Machine & *fscore*	0.806 (0.0040)	0.823 (0.0037)
Machine with small information	0.746 (0.0046)	0.783 (0.0043)

Test data: t = 2014

Model	LASSO	XGBoost
Human	0.639 (0.0052)	
Machine	0.774 (0.0047)	0.787 (0.0044)
Structured human	0.537 (0.0051)	0.558 (0.0096)
Machine & *fscore*	0.798 (0.0044)	0.815 (0.0042)
Machine with small information	0.740 (0.0051)	0.768 (0.0049)

Test data: t = 2015

Model	LASSO	XGBoost
Human	0.653 (0.0055)	
Machine	0.774 (0.0049)	0.804 (0.0044)
Structured human	0.547 (0.0053)	0.500 (0.0115)
Machine & *fscore*	0.804 (0.0046)	0.818 (0.0044)
Machine with small information	0.735 (0.0054)	0.772 (0.0050)

Test data: t = 2016

Model	LASSO	XGBoost
Human	0.663 (0.0053)	
Machine	0.779 (0.0049)	0.786 (0.0046)
Structured human	0.563 (0.0054)	0.516 (0.0111)
Machine & *fscore*	0.803 (0.0046)	0.810 (0.0045)
Machine with small information	0.738 (0.0054)	0.767 (0.0049)

Note: Each number represents the AUC, and the number in the parentheses is its standard error.

Table A9: Proxy estimation based on alternative prediction models

(1) LASSO

	Machine vs. Human		***SH vs. Human***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	0.495	0.002	0.150	0.001
Firm characteristics				
log(*sales* $_{f,t}$)	-12.859	0.146	10.266	0.082
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	17.666	0.113	-1.179	0.063
listed $_{f,t}$	59.775	3.193	4.973	1.792
#(*industry*) $_{f,t}$	-4.934	0.184	-0.769	0.103
priority $_{f,t}$	0.007	0.000	0.001	0.000
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	-0.001	0.000	-0.001	0.000
industry experience $_{f,i,t}$	-0.001	0.000	-0.000	0.000
Team characteristics				
#(team members) $_{i,t}$	0.112	0.014	0.009	0.008
Average #(*tenure years*) $_{i,t}$	0.123	0.019	0.016	0.011
Average industry experience $_{f,i,t}$	0.009	0.001	-0.005	0.001
Average #(*assigned companies*) $_{i,t}$	-0.001	0.000	-0.001	0.000
Constant	97.460	1.750	-130.928	0.982
Firm fixed-effect	yes		yes	
Analyst fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,238,817		3,238,817	
F	9,181.380		4,103.740	
Adj. R-squared	0.841		0.832	
Within R-squared	0.047		0.021	

(2) Extreme gradient boost

	Machine vs. Human		***SH vs. Human***	
	Coef.	S.E.	Coef.	S.E.
Number of available variables				
#(*available variables*) $_{f,t}$	0.449	0.003	0.075	0.004
Firm characteristics				
log(*sales* $_{f,t}$)	0.298	0.264	2.947	0.348
log(*sales* $_{f,t}$) - log(*sales* $_{f,t-1}$)	12.878	0.203	-0.930	0.268
listed $_{f,t}$	-5.342	5.763	-24.407	7.592
#(*industry*) $_{f,t}$	-3.276	0.333	-5.364	0.438
priority $_{f,t}$	-0.051	0.001	-0.123	0.001
Analyst characterstics				
#(*assigned companies*) $_{i,t}$	0.002	0.000	-0.001	0.000
industry experience $_{f,i,t}$	-0.008	0.000	0.010	0.001
Team characteristics				
#(team members) $_{i,t}$	0.768	0.024	0.392	0.032
Average #(*tenure years*) $_{i,t}$	0.508	0.034	0.139	0.045
Average industry experience $_{f,i,t}$	-0.035	0.002	-0.020	0.002
Average #(*assigned companies*) $_{i,t}$	-0.005	0.000	-0.006	0.000
Constant	-52.916	3.159	-27.909	4.161
Firm fixed-effect	yes		yes	
Analyst fixed-effect	yes		yes	
Year fixed-effect	yes		yes	
#(obs)	3,238,817		3,238,817	
F	2,886.910		1,230.400	
Adj. R-squared	0.506		-0.042	
Within R-squared	0.015		0.007	

Disagreement between Human and Machine Predictions

Oct 21st, 2021

IFC and Bank of Italy Workshop on "Data Science in Central Banking"

Daisuke Miyakawa (Hitotsubashi)

Kohei Shintani (Bank of Japan)

*This is the joint work with TOKYO SHOKO RESEARCH, LTD. (TSR) and conducted during one of the authors staying at BOJ-IMES. Usual disclaimer is applied.

Background

- Prediction tasks
  - E.g., firm exit, financial markets, macro, etc.
  - Better prediction ⇒ Better decision
- Machine learning (ML) methods
  - Using high dimensional information “mainly” for prediction
  - Varian ’14, Mullainathan & Spiess ’17, Athey ‘19
- Use ML for prediction
  - Successful in general
    - Labor: Chalfin et al. ‘16
    - Public: Kleinberg et al. ’18, Bazzi et al. ’19, Lin et al. ‘20
    - Medical: Patel et al. ’19, Mei et al. ‘20
    - Financial: Agrawal et al. ‘18
  - “ML ≻ Human” on average (⇔ They disagree)

Our research question

□ Any **systematic pattern** in the **disagreement**?

■ Informative to understand human **AND** machine errors

- E.g., informational opaqueness
- Can "ML $\prec$ Human" be the case?

  ⇒ ***Yes*** (economist view): Signal extraction from soft info

  ⇒ ***No*** (psychologist view): Noisy prediction

  ⇔ Kleinberg et al. '18: ML $\succ$ "Predicted" judge $\succ$ Judge

■ Useful for **task allocation**

- General computerization: Frey & Osborne '13
- Automation: Acemoglu & Restrepo '18

What we do

A) Construct a ML-based prediction model

- Massive size of firm-level data w/ high dimension information
- Various outcomes (default + voluntary exit + sales growth)

B) Measure the disagreement b/w ML & Human

- Human = Credit rating made by analysts
- Vs. Machine or "Structured" human
- *"Proxy"* ↑ (↓) ⇔ ML works better (worse)

C) Examine how opaqueness works as its determinants

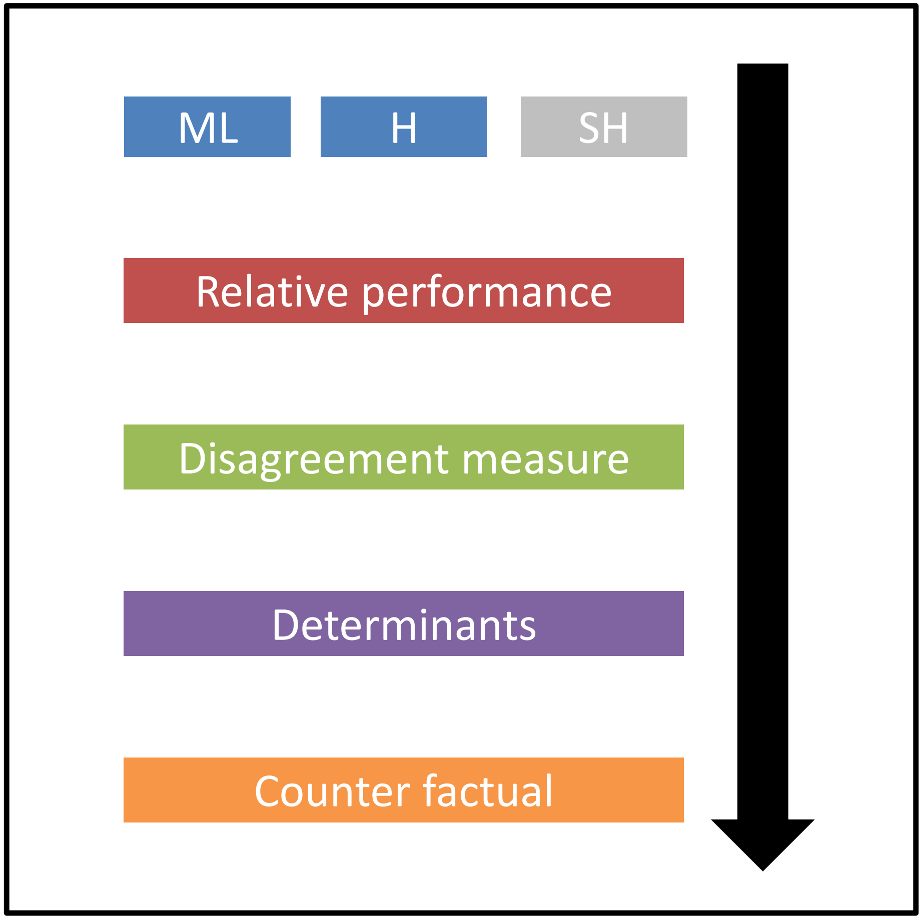
- Firms' informational opaqueness
- Controlling for various attributes as much as possible

D) Do a counterfactual exercise for task allocation

- Improve prediction power by allocating tasks to M & H

Organization of the paper

1. Theoretical illustration
2. Methodology
3. Data
4. Results
5. Summary



ML H SH

Relative performance

Disagreement measure

Determinants

Counter factual

Result: ML $\succ$ Human?

Relative performance

- □ Default & Closure
- □ Economist vs. psychologist
  - ■ Default: Econ
  - ■ Closure: Psy

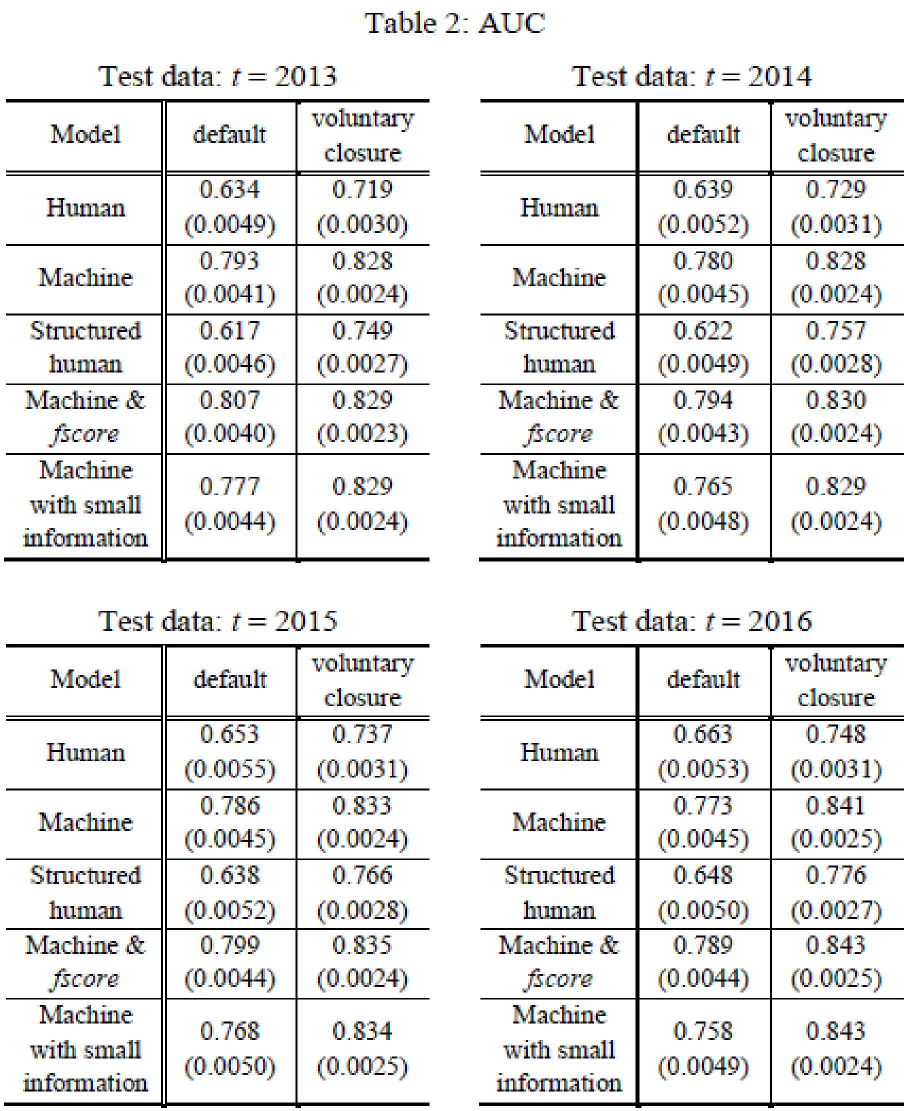


Table 2: AUC

Test data: $t = 2013$

Model	default	voluntary closure
Human	0.634 (0.0049)	0.719 (0.0030)
Machine	0.793 (0.0041)	0.828 (0.0024)
Structured human	0.617 (0.0046)	0.749 (0.0027)
Machine & *fscore*	0.807 (0.0040)	0.829 (0.0023)
Machine with small information	0.777 (0.0044)	0.829 (0.0024)

Test data: $t = 2014$

Model	default	voluntary closure
Human	0.639 (0.0052)	0.729 (0.0031)
Machine	0.780 (0.0045)	0.828 (0.0024)
Structured human	0.622 (0.0049)	0.757 (0.0028)
Machine & *fscore*	0.794 (0.0043)	0.830 (0.0024)
Machine with small information	0.765 (0.0048)	0.829 (0.0024)

Test data: $t = 2015$

Model	default	voluntary closure
Human	0.653 (0.0055)	0.737 (0.0031)
Machine	0.786 (0.0045)	0.833 (0.0024)
Structured human	0.638 (0.0052)	0.766 (0.0028)
Machine & *fscore*	0.799 (0.0044)	0.835 (0.0024)
Machine with small information	0.768 (0.0050)	0.834 (0.0025)

Test data: $t = 2016$

Model	default	voluntary closure
Human	0.663 (0.0053)	0.748 (0.0031)
Machine	0.773 (0.0045)	0.841 (0.0025)
Structured human	0.648 (0.0050)	0.776 (0.0027)
Machine & *fscore*	0.789 (0.0044)	0.843 (0.0025)
Machine with small information	0.758 (0.0049)	0.843 (0.0024)

Method: Disagreement

Disagreement measure

- □ *Proxy*: Measuring the "disagreement"
  - ■ Predict firms' outcome with test data by M & H & Structured H
    - Predicted outcomes for each company (between 0 and 1)
    - Larger means the company is more likely to face an event
    - *"t"* is addeted to the subscript
  - ■ Normalize predicted outcomes for each model

$$Outcome_{f,t}^{ML} \ \& \ Outcome_{f,i,t}^{H} \ \& \ Outcome_{f,t}^{SH}$$

Method: Disagreement

Disagreement measure

- ☐ *Proxy*: Measure the disagreement

  - ■ Large ⇔ M or SH $\succ$ H

  - ■ M vs H

$$Proxy_{f,i,t} = Outcome^{ML}_{f,t} - Outcome^{H}_{f,i,t} \text{ for exit firms}$$

$$= Outcome^{H}_{f,i,t} - Outcome^{ML}_{f,t} \text{ for non-exit firms}$$

  - ■ Structured H vs H

$$Proxy'_{f,i,t} = Outcome^{SH}_{f,t} - Outcome^{H}_{f,i,t} \text{ for exit firms}$$

$$= Outcome^{H}_{f,i,t} - Outcome^{SH}_{f,t} \text{ for non-exit firms}$$

Method: Determinants

- □ Identifying the determinants

  - ■ Firm-Analyst-time level Panel estimation:

$$Proxy_{f,i,t} = G\left(\boldsymbol{O}_{f,t}, \boldsymbol{F}_{f,t}, \boldsymbol{I}_{i,t}, \boldsymbol{Z}_{i,t}\right) + \boldsymbol{\eta}_{f,i,t} + \varepsilon_{f,i,t}$$

where

$\boldsymbol{O}_{f,t}$: Firm (i.e., target of scoring)’ informational opaqueness

$\boldsymbol{F}_{f,t}$: Firm (i.e., target of scoring)-attribute

$\boldsymbol{I}_{i,t}$: Analyst (i.e., human making score)- attribute

$\boldsymbol{Z}_{i,t}$: Team- attribute

$\boldsymbol{\eta}_{f,i,t}$: Fixed-effects

4-3. Result: Determinants

□ Higher opaqueness ⇒ M $\prec$ H

□ Same pattern for SH $\prec$ H

	default						***voluntary closure***					
	Machine vs. Human			***SH vs. Human***			***Machine vs. Human***			***SH vs. Human***		
	Coef.	S.E.		Coef.	S.E.		Coef.	S.E.		Coef.	S.E.	
Number of available variables												
#(*available variables*) $_{f,t}$	0.566	0.001	***	0.041	0.000	***	0.485	0.001	***	0.031	0.000	***
Firm fixed-effect	yes			yes			yes			yes		
Analyst fixed-effect	yes			yes			yes			yes		
Year fixed-effect	yes			yes			yes			yes		
#(obs)	3,238,817			3,238,817			3,238,817			3,238,817		
F	14,314.100			3,591.740			12,417.240			3,908.300		
Adj. R-squared	0.879			0.789			0.831			0.777		
Within R-squared	0.071			0.019			0.062			0.020		

(All the attributes $\boldsymbol{F}_{f,t}, \boldsymbol{I}_{i,t}, \boldsymbol{Z}_{i,t}$ are controlled)

Key takeaways

- □ "ML ≻ Human" on average
  - ■ Highly robust against many concerns

- □ "ML ≻ Human ≻ Predicted human"
  - ■ ≠ Kleinberg et al. (***QJE*** '18) and supporting economists' view

- □ Relative performance of H/M ↑ as firms opaqueness↑
  - ■ Highly robust against many concerns

- □ "ML ≺ Human" could be the case when…
  i. Firms are very opaque
  ii. Type I error is more concerned (than Type II error is)

Contribution

- First to study H-M disagreement in social science
  - Raghu et al. ’19: Algorithmic triage for diabetic retinopathy
    (≠ Anderson et al. ‘17, McIlroy-Young ’20 for “chess”)

- This is mainly because…
  - Data limitation on human prediction
  - Data limitation on target attributes
  - Data limitation on “human” (⇒ severe omitted variable issues)
    ⇔ E.g., Kleinberg et al. ‘18: No judge attributes
  - Selection label problem
    ⇒ Not the case in our data

⇒ **When we should/shouldn’t use ML?** (≠ Luca et al. ‘16)

Thank you and comments are welcome!

<Contact Information>

Daisuke Miyakawa:
Associate Professor
Hitotsubashi University Business School (HUB)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8439 Japan
E-mail: dmiyakawa@hub.hit-u.ac.jp
Web: https://sites.google.com/site/daisukemiyakawaphd/

Kohei Shintani:
Director
Bank of Japan
2-1-1 Nihombashi-Hongokucho, Chuo-ku, Tokyo 103-8660 Japan
E-mail: kouhei.shintani@boj.or.jp

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on “Machine learning in central banking”

19-22 October 2021, Rome / virtual event

Probability of default model with transactional data of Russian companies [1]

Gleb Buzanov and Andrey Shevelev,
Central Bank of the Russian Federation

[1] This presentation 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 central banks and other institutions represented at the event.

Bank of Russia

PROBABILITY OF DEFAULT MODEL WITH TRANSACTIONAL DATA OF RUSSIAN COMPANIES

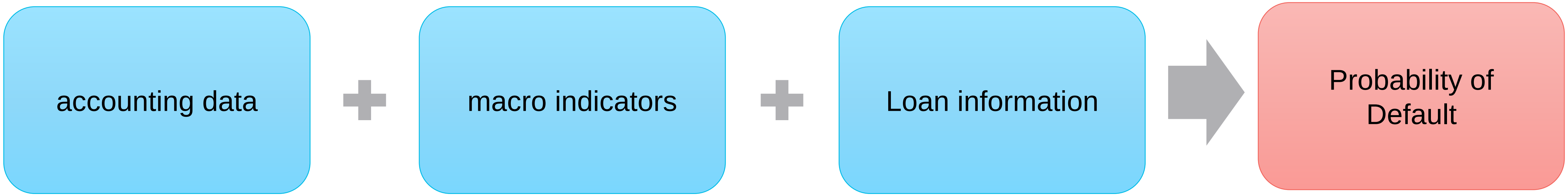
Andrey Shevelev[1]

Gleb Buzanov[2]

(1) Research and Forecasting Department
(2) Financial Stability Department

September 29, 2021

Overview of standard PD models & Our Goals



Cons:

- Rare publication of data
- Data is published with lags
- There is no connection between agents

***The global goal* of this project is to *improve* the existing models of the Bank of Russia for predicting the Probability of Default of Russian companies in terms of *quality* and decision *frequency* using transactional data of the Bank of Russia Payment System (BRPS).**

***The purpose of this work* is to study the *usefulness* of the data of the Bank of Russia Payment System (BRPS) for improving the existing probability of default models of Russian companies.**

Bank of Russia Payment System (BRPS) data

A firm's Payment Data can be used as information about a change in its state:

- VAT as a proxy for revenue
- Income tax as a proxy for profit
- Personal income tax and insurance payments as a proxy for payroll
- Payment graph for counterparty risk assessment
- and so on

Key points:

- Cover most of firms' payments
- Payment Graph
- Daily data

Literature reviews (with transactional data)

A. Khandani, A. Kim, A. Lo (2010): Consumer Credit Risk Models via Machine-Learning Algorithms

- Monthly aggregated transactions for credit risk assessment

H. Kvammea, N. Sellereiteb, K. Aasb, S. Sjursen (2018): Predicting Mortgage Default using Convolutional Neural Networks

- Daily transactions for mortgage defaults prediction

D. Babaev, M. Savchenko, A. Tuzhilin, and D. Umerenkov (2019): E.T.-RNN: Applying Deep Learning to Credit Loan Applications

V. Shumovskaia, K. Fedyanin, I. Sukharev, D. Berestnev, and M. Panov (2020): Linking Bank Clients using Graph Neural Networks Powered by Rich Transactional Data

- Transaction based model for fraud and scoring SBER clients

Definition of Default

2018												2019												Class in
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	2018
▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	not def
				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	>=90									def
													▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	X
▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	>=90	▨	▨	▨	▨	>=90					def
▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	>=90	▨	▨	▨	▨	▨	▨	def
		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	✓				not def
				▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	>=90	▨	▨	▨	✓		def

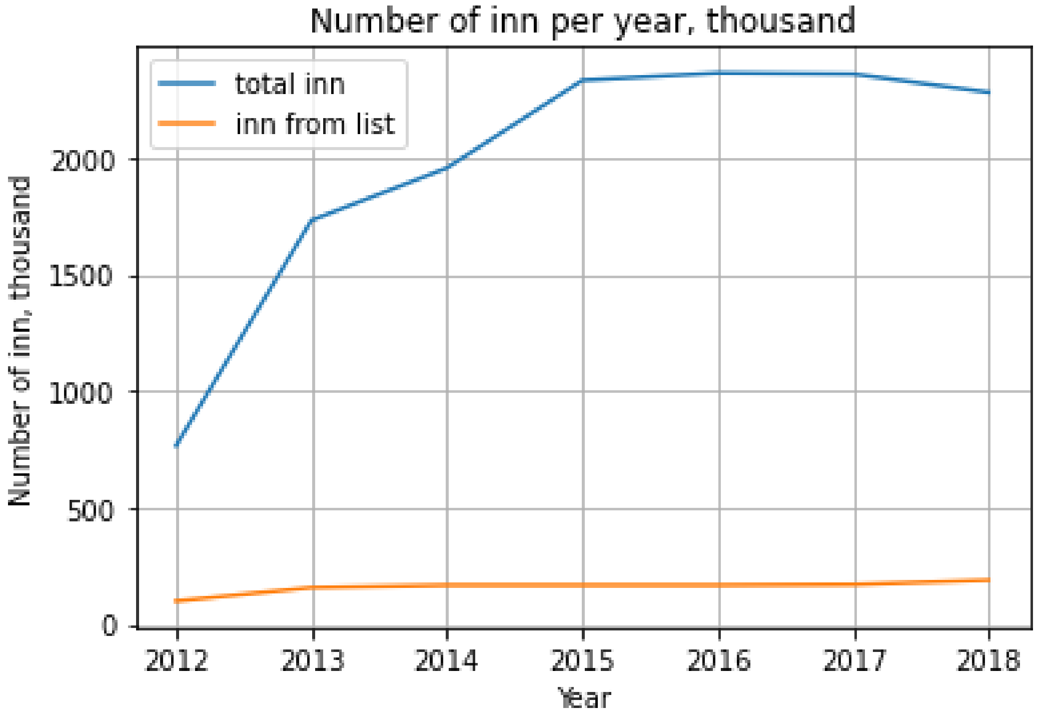
>= 90 days payment overdue

✓ successful loan

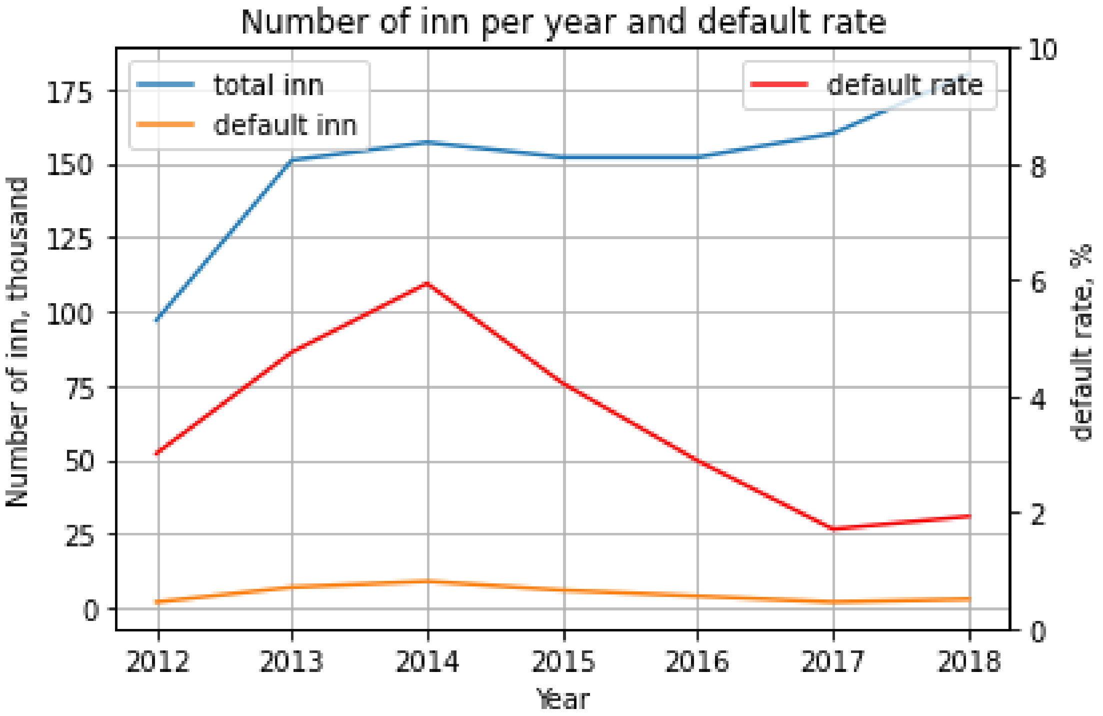
- Payment overdue >= 90 days [Basel II (III) IRB methodology]

Data & Default Rate

- Accounting data 2012-2018
- Default date 2013-2019
- Over 2 million unique companies per year
- About 200 thousand companies has a loan in next year after reported year
- 100-170k total unique companies
- 2-6% default rate



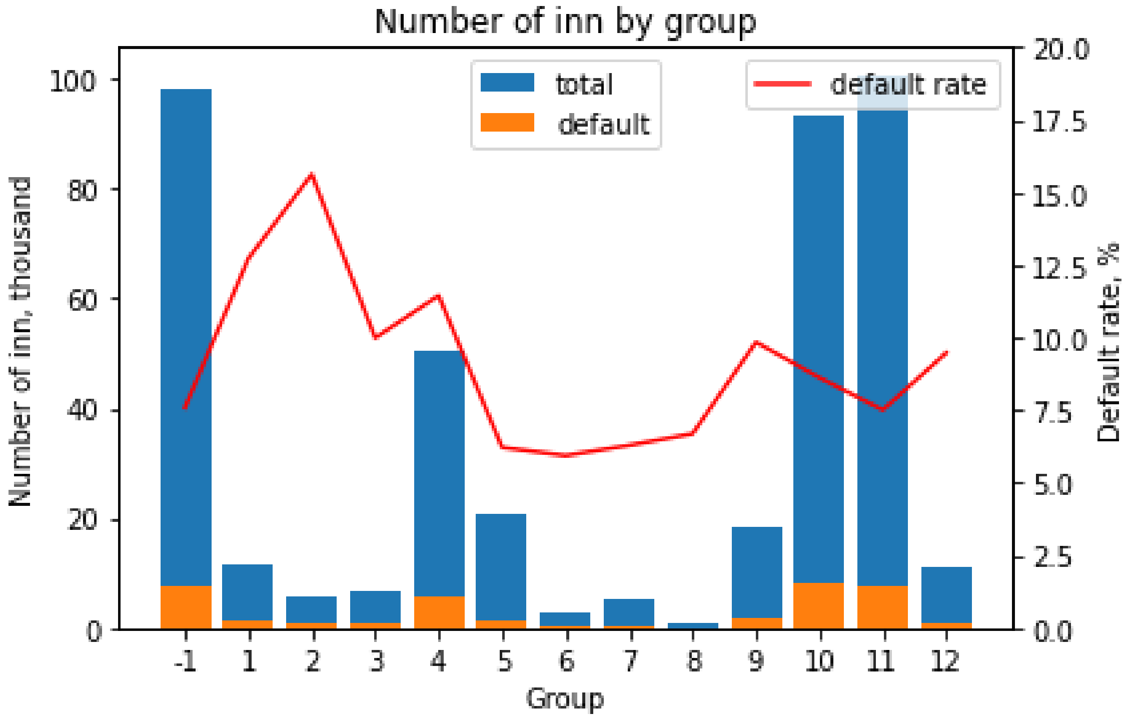
Number of inn per year, thousand



Number of inn per year and default rate

Default rate by industry, 2012-2018 year

Group	Industry name
1	Agro-industrial complex plant growing
2	Agro-industrial complex animal husbandry
3	Oil and gas industry
4	Construction
5	Real estate operations
6	Electricity and utilities sector
7	Chemical industry
8	Automotive
9	Consumer sector
10	Services sector
11	Wholesale trade (excluding fuel and minerals)
12	Metallurgy and mining
-1	Other



The Bank of Russia Payment System data

- From 2015 year
- ~ 10 million transaction per day

pd_main

- oper_dt
- pmt_payment_doc_hk
- inn_in
- inn_out
- amt
- acc_in
- acc_out
- kbk_cd
- pmt_type_kor_cd

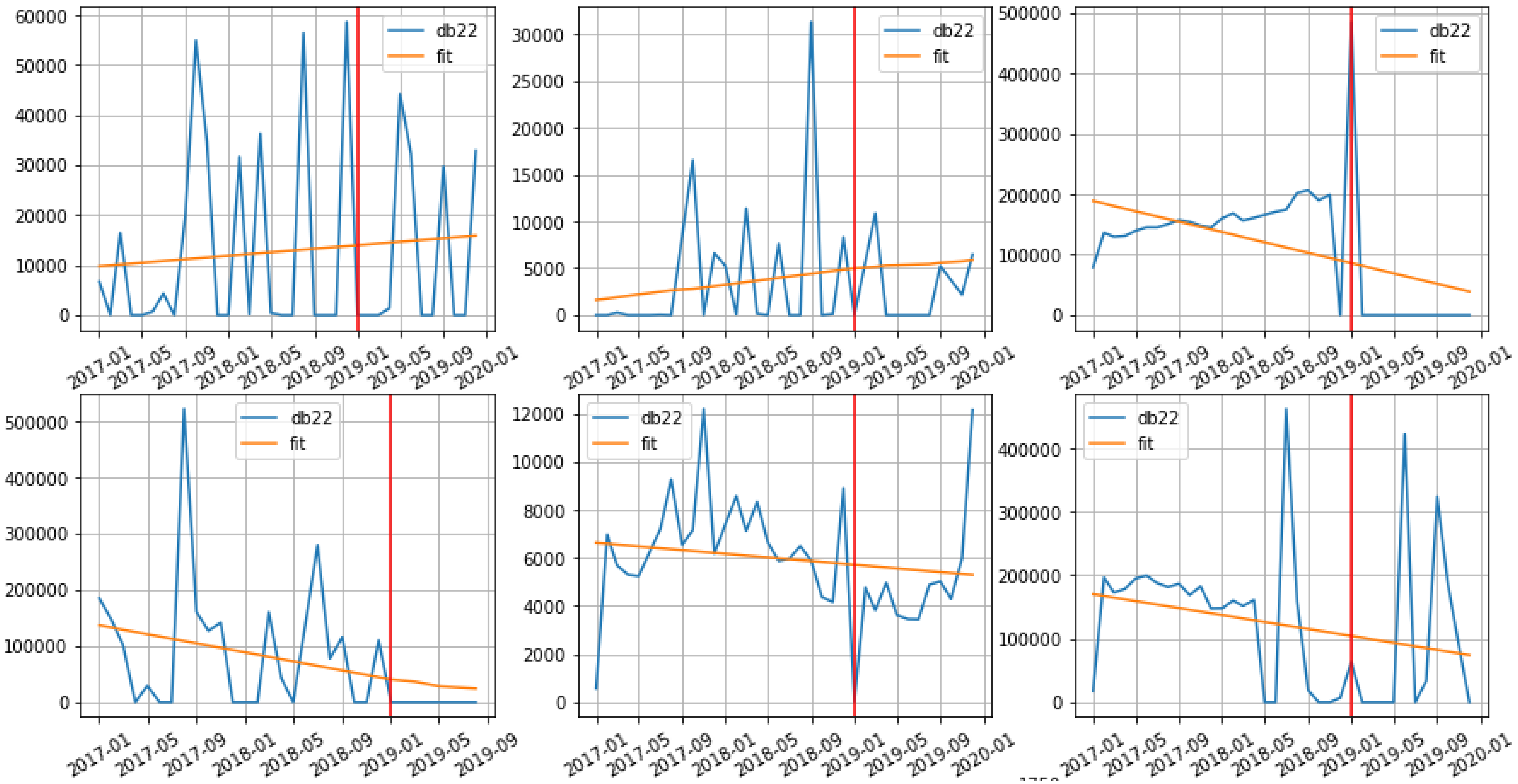
pd_reestr

- oper_dt
- reestr_sqn
- pmt_payment_doc_hk
- inn_in
- inn_out
- trx_amt
- acc_in
- acc_out
- trx_nm

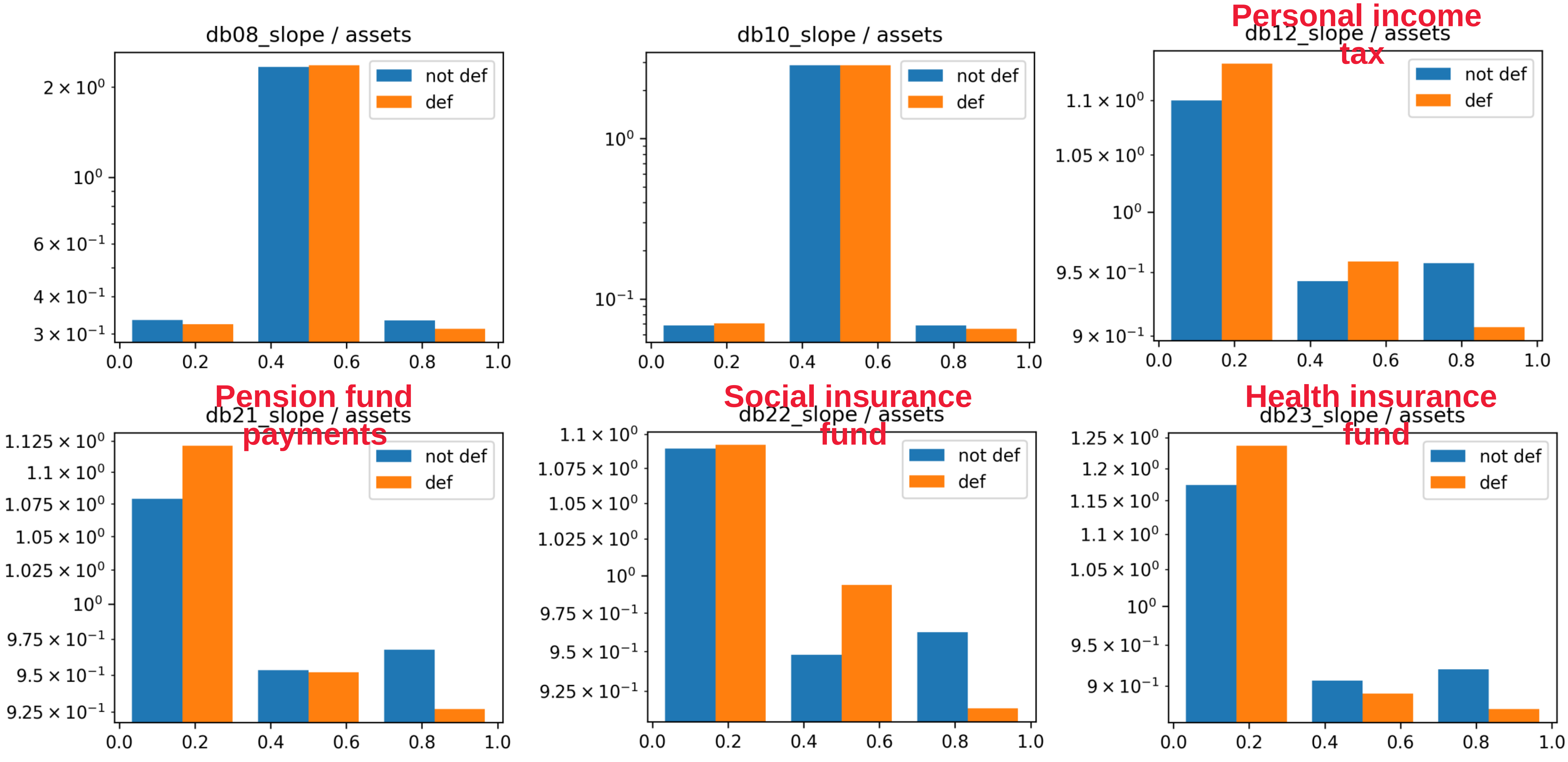
pd_status

- oper_dt
- pmt_payment_doc_hk
- pd_status_cd

Example of BRPS data: social security payments



Histograms of some BRPS data slopes



db08_slope / assets

db10_slope / assets

Personal income tax

db12_slope / assets

Pension fund payments

db21_slope / assets

Social insurance fund

db22_slope / assets

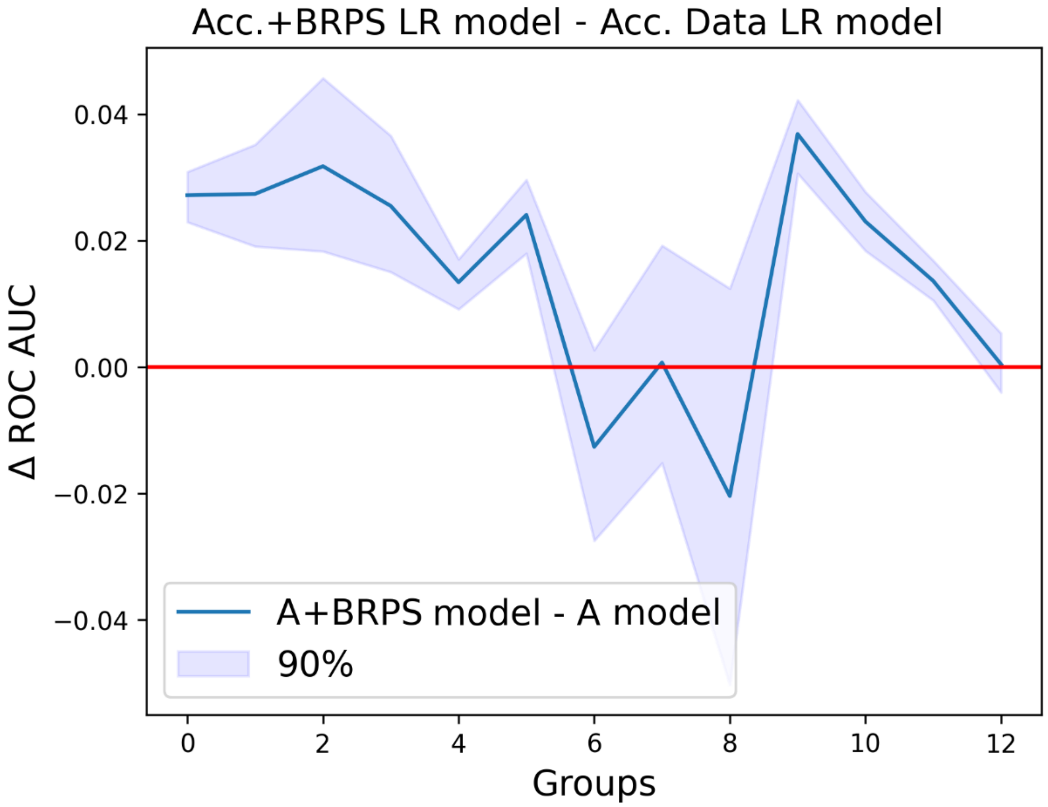
Health insurance fund

db23_slope / assets

*All slopes are ranked, normalized, $\sum \text{def} = 1, \sum \text{not def} = 1$, logscale

Logistic Regression on AD vs A+BRPS data

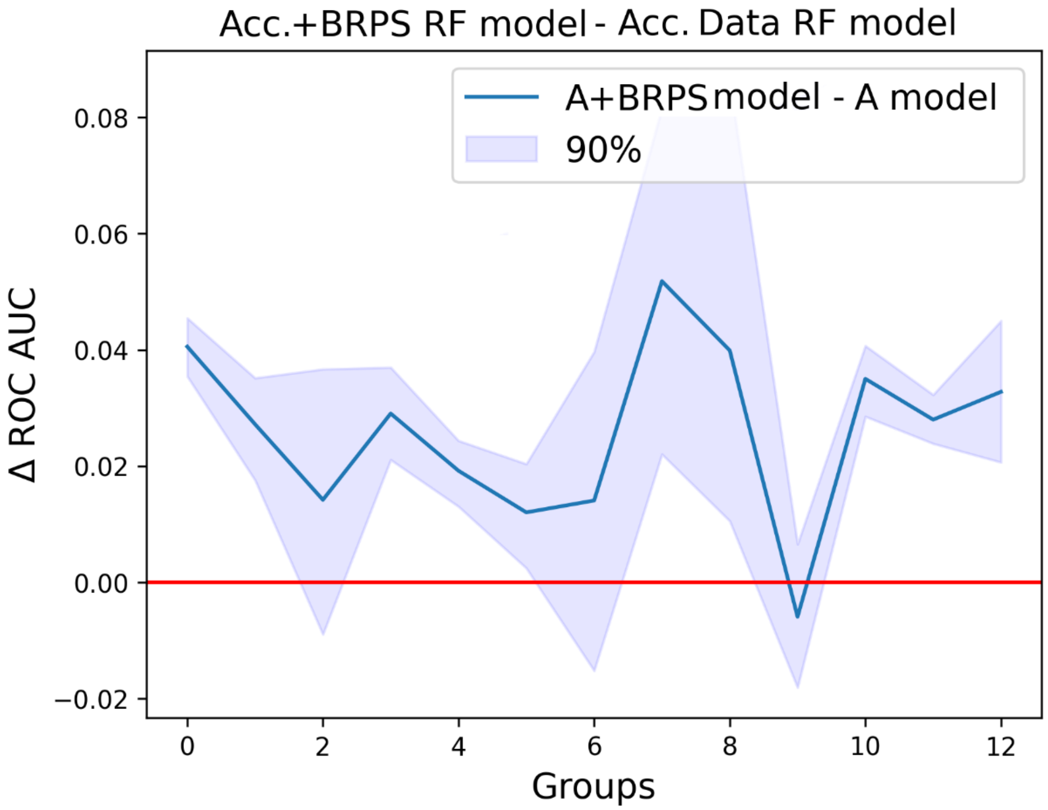
- L2 regularization
- Weighted likelihood



Combined score:
0.74 vs 0.76

Random Forest on AD vs A+BRPS data

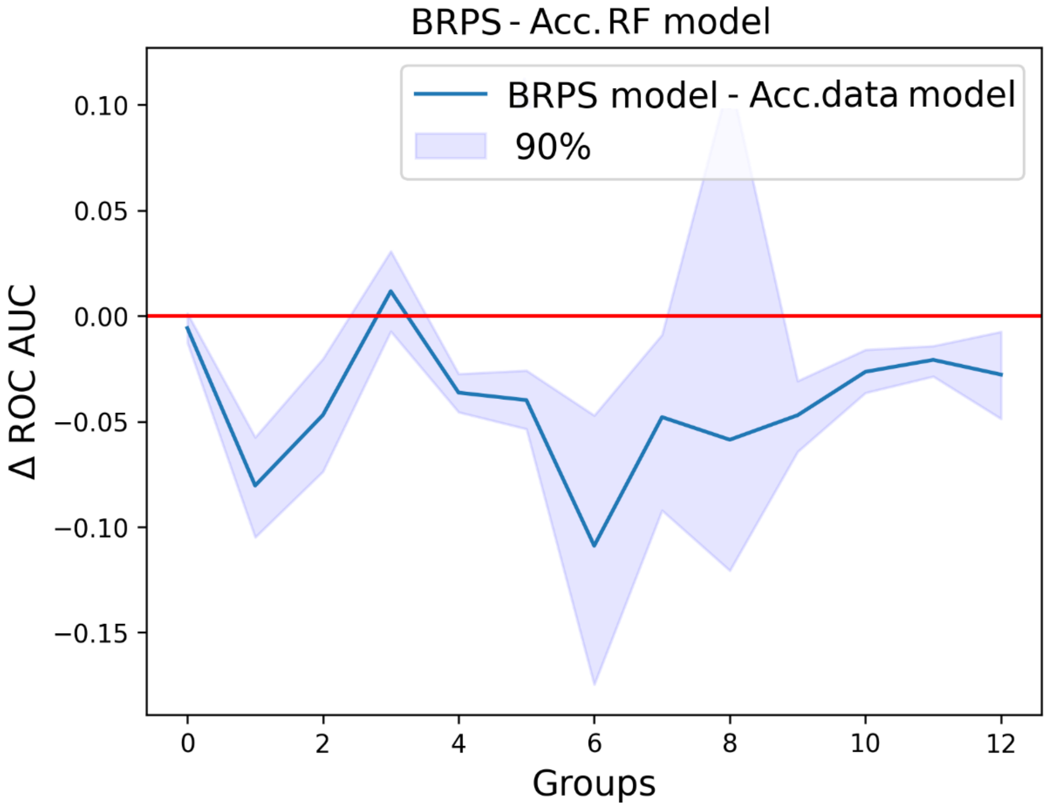
- Hyper parameters optimization
- Weighted Gini criterion for finding splits



Combined score:
0.72 vs 0.76

Random Forest on Accounting data vs BRPS data

- BRPS data only model is a useful for obtaining early estimates



Combined score:
0.72 vs 0.70

Conclusion

Results:

- BRPS data improve forecast quality
- Model with BRPS data only is a useful for getting early estimates

Next steps:

- Extending dataset up to date
- Neural Network model
- Graph Neural Network model
- Higher frequency
- Time series based

Bank of Russia

shevelevaa@mail.cbr.ru

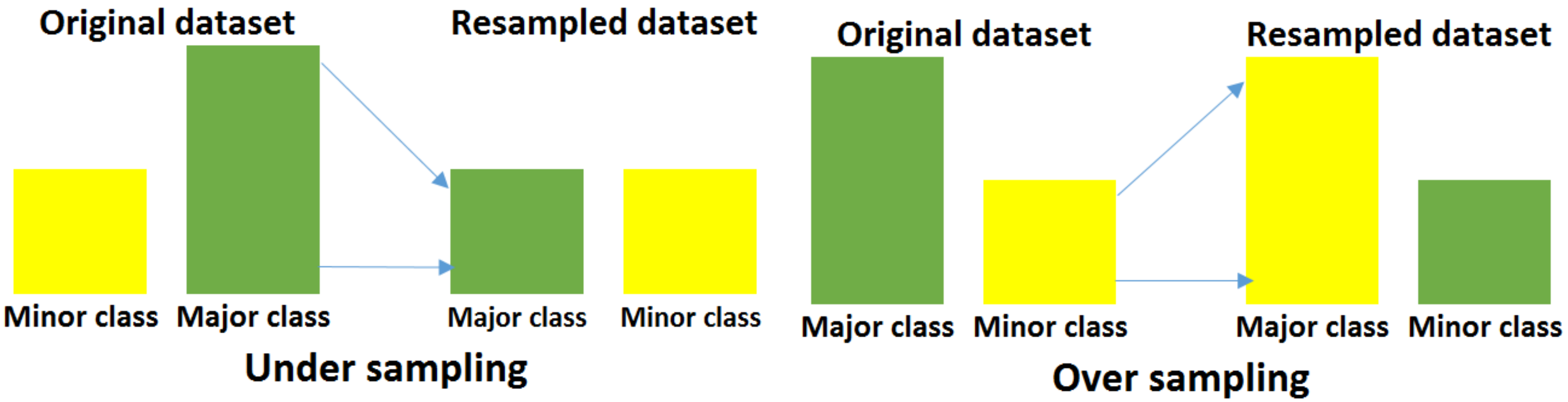
BACKUP SLIDES

Accounting data features

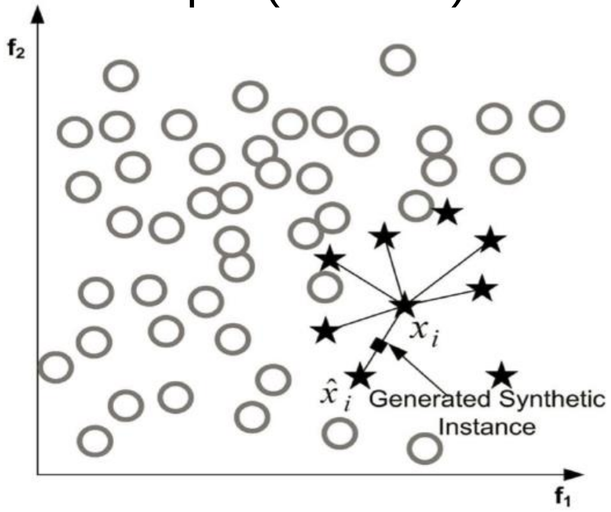
Name	Formula	Label
K1	'12003' / '15003'	Current liquidity ratio
K2	'22003' / '16003'	Return on assets
K3	'21003' / '21103'	Gross margin
K4	'24003' / '21103'	Net profit margin
K5	'22003' / '21103'	Operating margin
K6	'13003' / '16003'	Equity-to-asset ratio
K7	('14103'+'15103') / ('22003')	Debt/earnings from sales
K8	'22003' / '23303'	Interest coverage ratio
K9	'22003' / '23003'	Interest burden ratio
K10	('13003') / ('14003'+'15003')	Borrowed funds/Equity
K11	('13003'-'11003') / '12003'	Working capital to current assets ratio
K12	'24003' / '23003'	Tax burden ratio
K13	('21103' / '15003')	Short-term debt/Revenue
K14	'21103' / '12303'	Receivables turnover ratio
K15	'21203' / '15203'	Accounts payable turnover ratio

Imbalanced Data

- Random Under Sampling
- Random Over Sampling
- SMOTE
- Weighted likelihood estimation

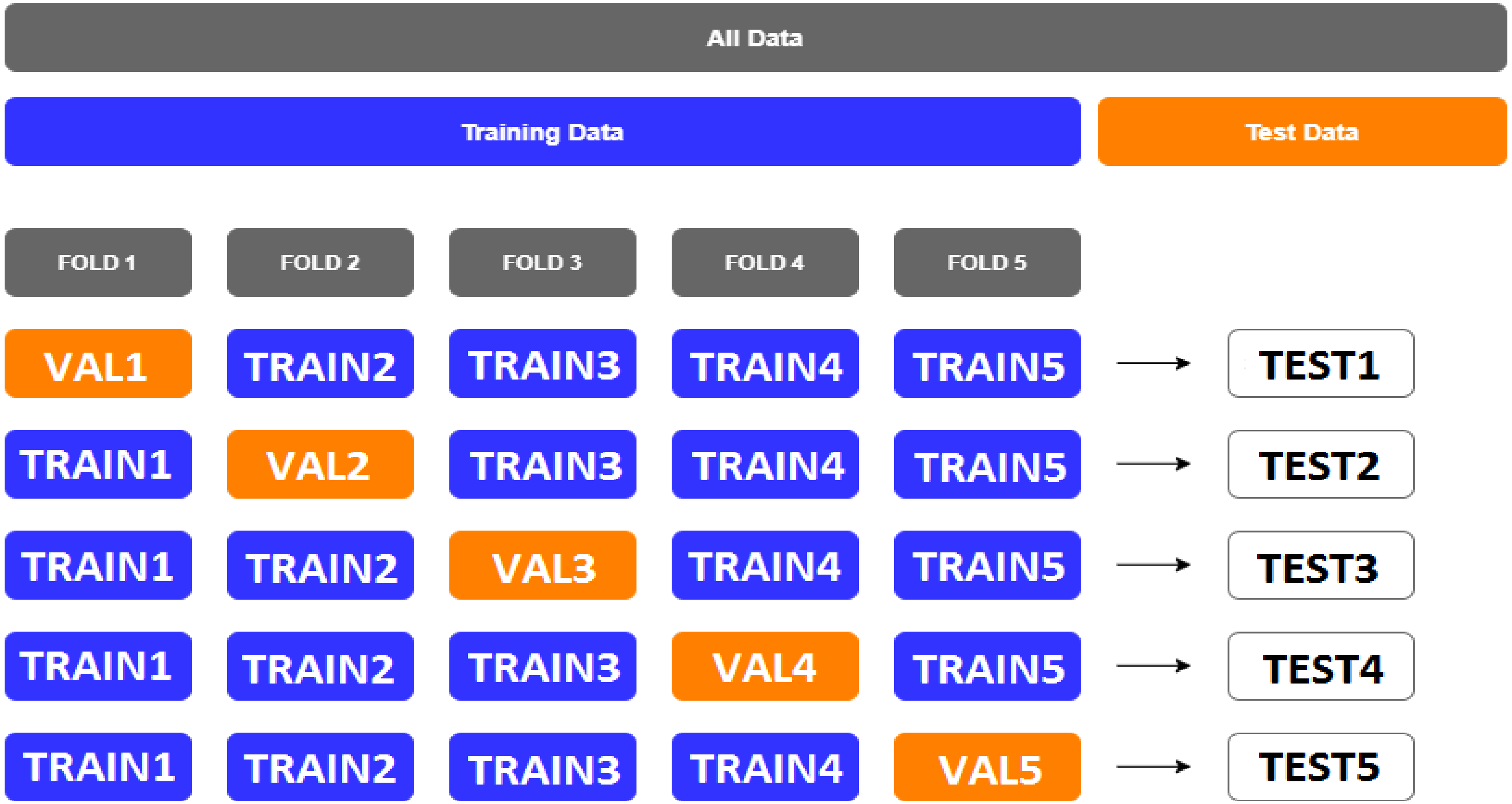


Synthetic Minority Oversampling Technique (SMOTE)



Cross validation

- K-Fold cross-validation
- Repeated k-Fold cross-validation



All Data

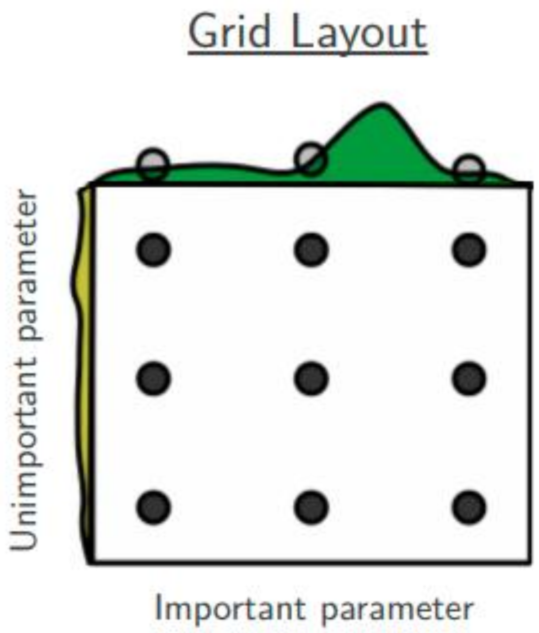
Training Data

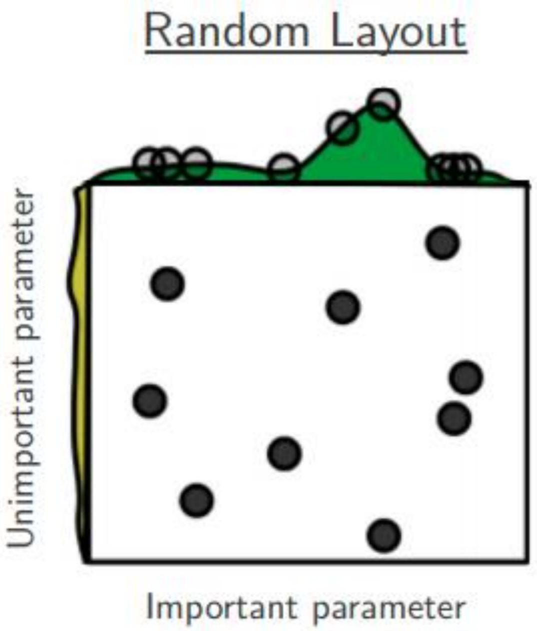
Test Data

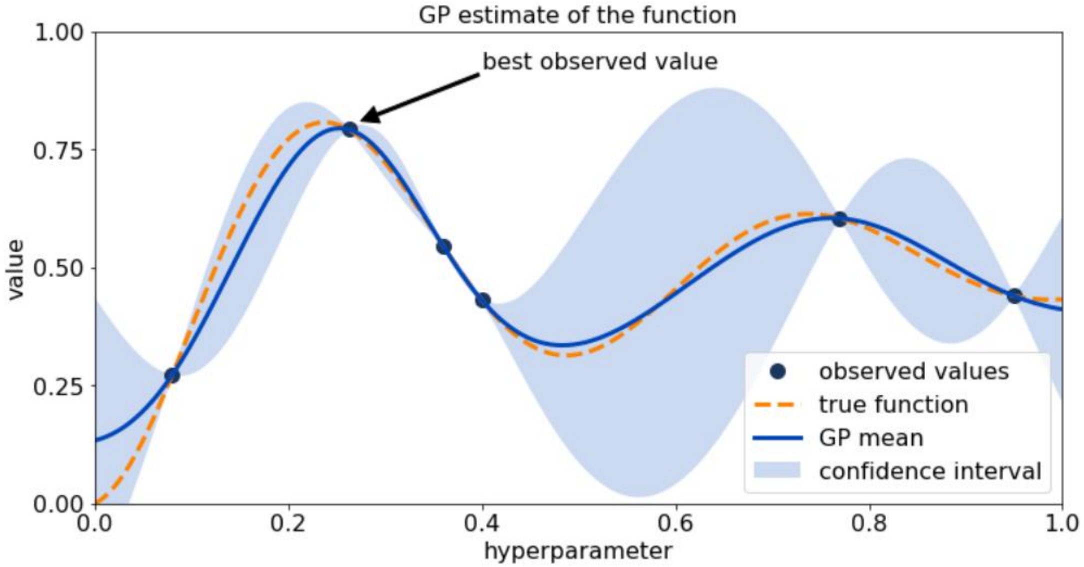
FOLD 1	FOLD 2	FOLD 3	FOLD 4	FOLD 5		
VAL1	TRAIN2	TRAIN3	TRAIN4	TRAIN5	→	TEST1
TRAIN1	VAL2	TRAIN3	TRAIN4	TRAIN5	→	TEST2
TRAIN1	TRAIN2	VAL3	TRAIN4	TRAIN5	→	TEST3
TRAIN1	TRAIN2	TRAIN3	VAL4	TRAIN5	→	TEST4
TRAIN1	TRAIN2	TRAIN3	TRAIN4	VAL5	→	TEST5

Optimization of Hyper Parameters

- Grid Search
- Random Search
- Bayesian optimization



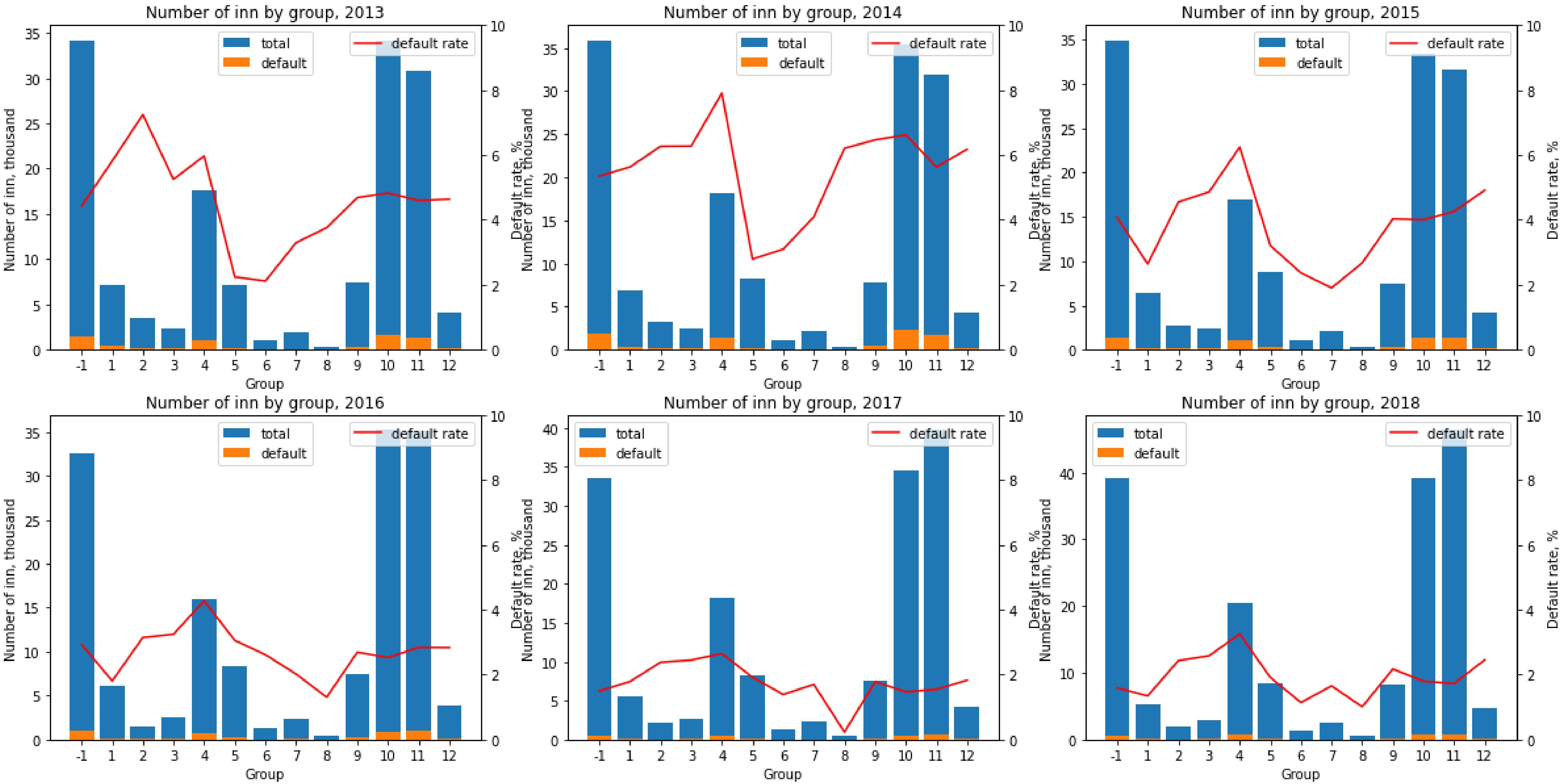




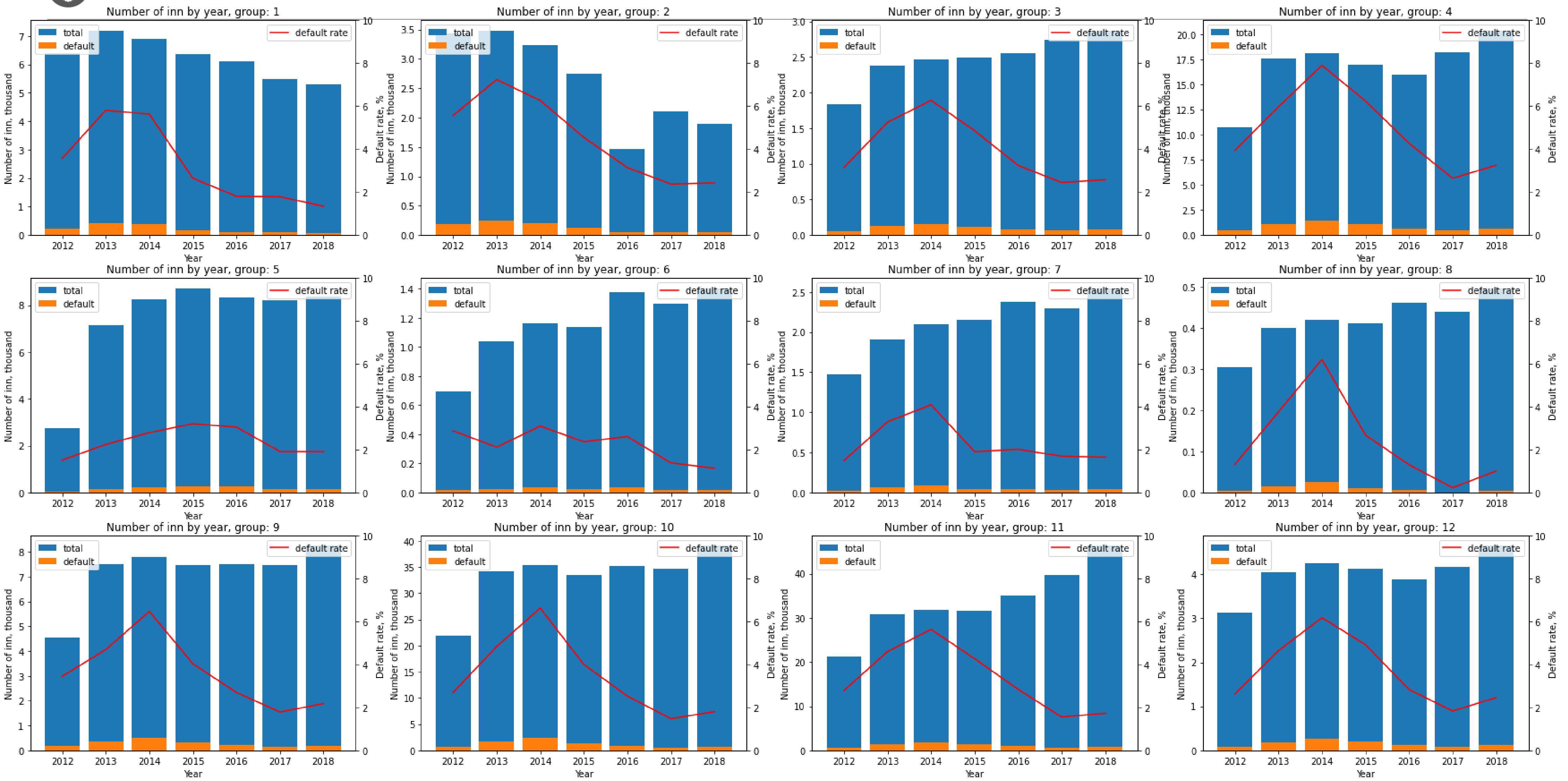
Monthly aggregate BRPS data (transcript)

N	Name	Transcript
1	DT	Date of report
2	INN	Taxpayer Identification Numbers
3	OKVED_CODE	Russian Economic Activities Classification System Code
4	CNT_FI_ID_DB	Number of outgoing transactions
5	CNT_FI_ID_CR	Number of incoming transactions
6	PRC_MAX_DB	The share of the maximum outgoing turnover
7	PRC_MAX_CR	The share of the maximum ingoing turnover
8	REGNUM_MAX_DB	Number of maximum outgoing turnovers
9	REGNUM_MAX_CR	Number of maximum ingoing turnovers
10	TOT_DB	Outgoing payments value
11	TOT_CR	Incoming payments value
12	CNT_DB	Number of outgoing payments
13	CNT_CR	Number of incoming payments
14	DB06	Land tax
15	DB07	Gambling tax
16	DB08	Property tax
17	DB09	Transport tax
18	DB10	Other taxes
19	DB11	Corporate income tax
20	DB12	Personal income tax
21	DB13	Value added tax on goods imported to Russia
22	DB14	Value added tax on goods sold in Russia
23	DB15	Simplified tax
24	DB16	Single tax on imputed income for certain types of activities
25	DB17	Agricultural tax
26	DB18	Tax levied in connection with the application of the patent taxation system
27	DB20	Other payments for social needs
28	DB21	Pension fund payments
29	DB22	Social insurance fund payments
30	DB23	Health insurance fund payments
31	DB24	Federal customs service payments
32	DB30	Other budget payments
33	DB31	Client payments
34	DB32	Payments to non-residents
35	DB33	Settlements with the exchange
36	DB35	Write-off from deposits
37	DB36	Payments to non-resident banks
38	DB37	Foreign currency purchase
39	DB38	Loan repayment
40	DB00	Other payments to be debited
41	CR39	VAT refund
42	CR40	Payment from the budget
43	CR41	Clients payments
44	CR42	Payments from non-residents
45	CR43	Settlements with the exchange
46	CR45	Deposit credits
47	CR46	Payments from non-resident banks
48	CR47	Foreign currency sale
49	CR48	Getting a loan
50	CR00	Other credit operations

BRPS data: default rate by industry, 2012-2018 year



BRPS data: default rate by industry



Model

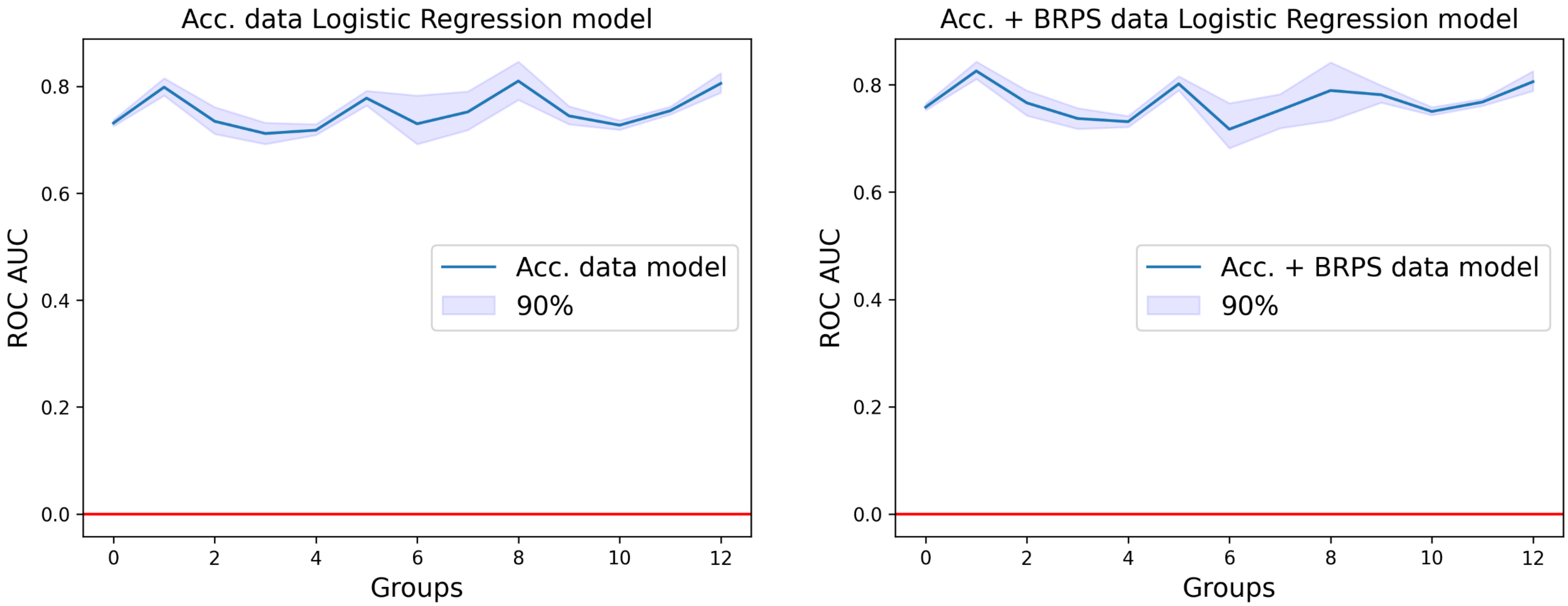
Baseline model includes:

- Accountant data

Extended model includes:

- Accountant data
- Annual aggregated BRPS data normed by assets
- Slopes of monthly aggregated BRPS data
- Slopes of monthly aggregated BRPS data normed by assets

Logistic Regression results



Literature reviews

Babaev D., Savchenko M., Tuzhilin A., and Umerenkov D. (2019): E.T.-RNN: Applying Deep Learning to Credit Loan Applications

Bergstra J., Bengio Y. Random Search for Hyper-Parameter Optimization // Journal of Machine Learning Research 13 (2012) 281-305

Breiman L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Demeshev B., Tikhonova A. (2014). Default prediction for Russian companies: Intersectoral comparison // HSE Economic Journal, Vol. 18, No. 3, pp. 359—386. (In Russian)

Jackson R., Wood A. The performance of insolvency prediction and credit risk models in the UK: A comparative study // The British Accounting Review Volume 45, Issue 3, September 2013, Pages 183-202

Khandani A., Kim A, Lo A.(2010): Consumer Credit Risk Models via Machine-Learning Algorithms

Kvammea H., Sellereiteb N., Aasb K., Sjursen S. (2018): Predicting Mortgage Default using Convolutional Neural Networks

Louppe G., “Understanding Random Forests: From Theory to Practice”, PhD Thesis, U. of Liege, 2014.

Mogilat A. Bankruptcy in Russian real sector: Basic tendencies and financial indicators of a typical bankrupt // Nauchnye Trudy INP RAN, No. 13. 2015. pp. 156—186. (In Russian)

Mogilat A. Modelling financial distress of Russian industrial companies, or What bankruptcy analysis can tell // Voprosy Ekonomiki. 2019;(3):101-118. (In Russ.) https://doi.org/10.32609/0042-8736-2019-3-101-118

Shibitov, D., & Mamedli, M. (2019, August). The finer points of model comparison in machine learning: forecasting based on Russian banks' data. Retrieved from Bank of Russia Working Paper Series No. 43: http://www.cbr.ru/content/document/file/87572/wp43_e.pdf

Shirata C. Predictors of Bankruptcy after Bubble Economy in Japan: What can we learn from Japan Case? // the 15th Asia-Pacific Conference on International Accounting Issues, November 24, 2003

Shumovskaia V., Fedyanin K., Sukharev I., Berestnev D., and Panov M. (2020): Linking Bank Clients using Graph Neural Networks Powered by Rich Transactional Data

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

The use of AI for company data gathering
Finding and monitoring fintechs in Germany and France[1]

Elisabeth Devys, Bank of France,
and Ulf von Kalckreuth, Deutsche Bundesbank

[1] This presentation 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 central banks and other institutions represented at the event.

The use of AI for company data gathering

Finding and monitoring fintechs in Germany and France

Guillaume Belly, Banque de France
Andy Bosyi, Neusinger
Elisabeth Devys, Banque de France
Ulf von Kalckreuth, Deutsche Bundesbank

Abstract

In dealing with fintechs, we have to do without the cornerstones of traditional statistics. There are few if any standardised reporting requirements, no developed taxonomy and no established set of quantitative measures. By definition, innovation involves new activities, and this is intrinsically difficult for traditional statistics, which need stable classifications. Company registers are mostly useless here. The business environment and market structures are changing rapidly. The segment is characterised by a high rate of "metabolism": entries, mergers and acquisitions, exits. Any list of fintech companies is rapidly outdated.

Following the guidance of the final report of the IFC Working Group on Fintech Data Issues, we investigate non-standard ways of collecting and organising information. Specifically, we explore the use of AI to find new fintechs and to monitor changes in the characteristics of known entities. The Banque de France and the Deutsche Bundesbank follow two different, but complementary approaches. The first involves collecting data from publicly available or third-party proprietary databases and submitting them to classification algorithms, while the second involves a graph approach that looks at where a company is placed within a network of nodes and edges of named entities. The initial results are incomplete, but encouraging.

The methodologies developed here can potentially be applied to many other issues in company level statistics. It may also show a way to more informative and timely statistics using information in the public domain, without onerous new reporting requirements.

Keywords: fintech; master data; artificial intelligence; non-administrative approach; classification algorithms

JEL classification: C81, G23, O30

Contents

1. Introduction

Fintech happens where innovation takes place in the financial sector, where new methods and products emerge, are tested and made ready for the market. The results are shaping the financial industry as a whole. Central banks need to identify, describe and understand fintech activities.

In dealing with fintechs, we have to do without the cornerstones of traditional statistics. There are few if any standardised reporting requirements, no developed taxonomy and no established set of quantitative measures. By definition, innovation involves new activities, and this is intrinsically difficult for traditional statistics, which need stable classifications. Company registers are mostly useless here. The business environment and market structures are changing rapidly. The segment is characterised by a high rate of "metabolism": entries, mergers and acquisitions, exits. Any list of fintech companies is rapidly outdated. To make things even worse, there is not even an accepted general definition of what "fintech" means.[1]

In order to prepare a multi-purpose fintech monitoring system for statistics, regulation and financial stability, we need to find new ways of collecting data and new tools to identify this type of firm. This should involve mostly publicly available information, plus information that is shared voluntarily. The project described here concentrates on the aspect of gathering information on the activity of fintech firms and identifying both known and unknown firms. The task is perfectly general – the toolbox developed for this project will thus be well-suited for monitoring fintechs all over the world, and any other rapidly evolving sector.

2. Project objective

We present a project in its infancy. However, we think that the approach is straightforward and generic. The project aims to create AI tools for gathering the master data needed for fintech monitoring in the first place, and for monitoring them over time. The task has two major aspects: (a) finding new fintechs and characterising their activity, and (b) recording major changes in activity among known market participants. As a co-operative effort of most of the major central banks in the world, the Irving Fisher Committee (IFC) Working Group on Fintech Data Issues, with the close co-operation of the BIS, has reviewed the state of affairs and outlined a targeted roadmap for constructing fintech statistics[2] (see footnote for a link to the report). Developing non-administrative AI methods is part of this roadmap. Activities in the fintech industry are notoriously cross-border, and purely national or regional data collection is of limited value. Data collection on the fintech industry should be co-ordinated worldwide. A solution for gathering master data on fintechs that can be implemented globally would be an important first step.

[1] Schueffel (2016), after a painstaking search through the available literature, arrives at three common features: 'technology', 'innovation' and 'finance'. However, what 'technology' and 'innovation' mean will always depend on time and context. Von Kalckreuth and Wilson (2020) argue that these terms cannot be part of an operational statistical classification system.

[2] See Irving Fischer Committee (2020), the final report of the working group.

3. Problem statement

By definition, innovation involves new activities. The business environment and market structures are changing rapidly. This is intrinsically difficult for traditional statistics, which need stable classifications. In addition, there are hardly any reporting requirements. Essentially, this will also remain the case in the future, because reporting requirements can only be legislated for activities which have been known for quite some time, not for innovative activities which are as yet unknown.

New and innovative methods are required in order to consolidate data and find new fintech entities, to characterise their activity, and to record major changes in the activities of existing market participants. The market activity of fintech companies is almost exclusively web-based, and it is well documented on their websites and in the dedicated online economic media. We can make use of this fact in several ways. Web scraping in connection with AI methods (text mining and graph theory) make it possible to find new fintechs and track the activities of known companies. With an operational list of fintechs at hand, we can use supervised or unsupervised learning on a large set of websites of IT-related companies. Concerning external information, we first concentrate on data from specialised information services. Later on, the websites of known fintechs may be added, as well as information from official registers, the websites of fintech associations, and consulting, venture capital and recruiting companies. Finally, internal data within central banks may further improve data quality. Charting the links between websites can yield a machine-based representation of conglomerates and company networks. This type of methodology is used at the Organisation for Economic Co-operation and Development (OECD) for its Analytical Database on Individual Multinationals and Affiliates (ADIMA), but not yet by central banks.

Using AI-based methods of probabilistic matching, it is possible to map websites to universal registers such as the ESCB Register of Institutions and Affiliates Database (RIAD) or the registers of statistical firms. This links information on economic activity to information on legal entities.

4. The two aspects of the task

The system needs a robust infrastructure to organise information. Once a sufficient amount of granular information on fintechs is gathered, classification methods (e.g. based on cluster analysis or graph mining to structure, partition and chart interactions within this ecosystem) can help provide a better understanding of the fintech landscape. The challenge has two dimensions, each interesting in its own right: monitoring known fintechs and finding new, hitherto unknown fintechs.

4.1. Monitoring known fintechs

Known fintechs can be monitored by tracking information brokers, lists and web fora on fintech. Websites can be monitored for relevant changes in activity in order to aid classification. This can help identify events such as mergers, acquisitions, insolvencies, and exits for a large number of firms. The downloading of balance sheets and other financial statements can be partly automatised. Key positions can be fed directly into information systems. This is very much analogous to existing "know your client" solutions on the market.

4.2. Finding new fintechs

New fintechs can be found by periodically running classification algorithms over large databases of private companies. Additionally, data from specialised information brokers, lists and web fora on fintech can be fed into the system. A rather free approach that does not rely on specific data structures is to scrape websites from companies in a larger list of IT companies in order to detect patterns that are characteristic of fintech companies.

5. Towards a non-administrative solution

The solution envisaged by the Bundesbank and the Banque de France is non-administrative and does not rely on the cooperation of the entities being monitored. It mostly uses publicly available data,[3] which is both an advantage and a limitation. Both central banks are performing preparatory pilot studies concerning feasibility and the prerequisites of the information structure, and are working together with fintechs. Thus far, the teams have concentrated on the classification problem, i.e. how to tell fintechs and non-fintechs apart. There is no closed-form and time-invariant definition of fintech, so the concept of fintech we use is implicit. We start from lists of firms which are considered to be fintechs based on the fintech monitoring in both central banks. This list embodies (implicit) information on the characteristics of innovative firms and technology oriented firms which offer or enable financial services that are of interest in the various areas of central bank work: supervision, financial stability, payments, etc.

The Banque de France has developed preliminary AI algorithms to identify fintechs in a large database of French firms, using a training data set of firms which the Banque de France has classified as fintechs. This is a well understood, fast and reproducible type of data collection, with natural connections to traditional statistics. The Bundesbank is exploring a graph approach that allows a rather open type of data gathering. Exploiting news data or the websites of companies, graphs of named entities (locations, organisations and persons) are created and enriched with additional information. Graph information can be processed by NLP techniques (such as named entity recognition, bag of words, word distances, etc.), and firms are classified by the nature of their links to the nodes. This has the advantage of being very up-to-date, as it draws directly on real-time information in news media or on

[3] Most of the data are publicly available or "readable", but sometimes restrictions concerning use or the ability to apply scraping techniques – for example with regard to data from free online newspapers, blogs or social networks – make them almost proprietary. This is the case if, for example, fees are charged to collect them, depending on the business model and volume. In the case of using a third-party data broker, the data become proprietary even though most of the data are publicly available, since the collecting, aggregating, cleaning and formatting steps are charged.

websites.[4] Potentially, the approach may actually recognise new types of fintechs hitherto unknown to classifiers.

The authors believe that these two approaches are complementary – both are needed to keep track of developments in fast evolving, innovative markets.

6. A graph-based approach

The first and principal step of the approach followed by the Bundesbank team is to extract relevant text information from publicly accessible websites or news databases for classification purposes. The name of a company to be classified is assumed to be known. In the following, we emphasise web scraping, as in the case of structured news information databases the task is similar, but simpler. Two tools are needed, a search engine and an article extraction engine.

The task of the search engine consists in searching for a term in the web or news database and return relevant information. The team used the Google search engine to search for terms. This service is provided by Rapid Search API. Returned results will depend on the search term, but search engines typically also allow the specification of search parameters, such as regular websites, news, images or places. The proper use of the parameters that would provide the best resulting data distribution is itself a matter for extensive research and analysis.

The article extraction engine is needed for properly extracting content from a web page. Modern HTML language provides different structural components (such as the <article> tag) for specifying the exact location of an article within the page. Still, this task is complex since:

- not all pages are structured conveniently – even if there is an article within the page, it may not be separated from the rest of the page;
- not all pages contain proper articles – the relevant content may be placed as pieces of text on different parts of the page;
- a lot of extra HTML code may have to be cleaned up and the parts containing relevant text have to be preserved, which needs to be done carefully.

Setting up an infrastructure for web scraping is a complex and time-consuming task. The team is using a readily available tool called Zyte (formerly ScrapingHub). When provided with a link to an article, it is capable of returning much useful information: article body, article HTML, article author list and main author, publication date, etc.

The data collection process for a single search term is summarised in Figure 1. Below we provide a short description of the steps.

[4] In order to do so, the Deutsche Bundesbank collaborates with Neusinger, an independent IT consultant firm specialised in AI and machine learning. Neusinger was one of the winners of a Bundesbank innovation challenge conducted in 2020 and 2021.

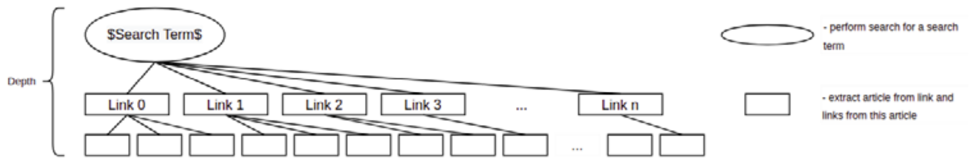

$Search Term$
Depth
Link 0
Link 1
Link 2
Link 3
...
Link n
...
- perform search for a search term
- extract article from link and links from this article

Figure 1: Data collection process

6.1. Performing search

The process starts with a search on a specific API, specifying the search terms. The initial search term for each company should be its name. Search term optimisation may be used here. For example, adding keywords like “company”, “fintech”, “finances”, etc. may improve the relevance of results. Using search query punctuation may lead to certain unwanted results being filtered out. Again, the modalities of such optimisation is a matter of research.

6.2. Processing links

For each link, the same actions are performed:

- An article or articles are scraped from the link using Zyte or a similar tool. The returned information is stored for future use and analysis.
- All named entities are extracted using Spacy, an article extraction tool. Although other entities may be used for modelling, the named entities with the highest potential are PERsons, LOCations and other ORGanisations.
- Links within the extracted article are extracted using tools for identifying regular expressions.

This process leads to a tree of links. The results need to be stored properly, preserving the information on hierarchical relations.

6.3. Proceeding for greater depth

The procedure on links described above is iterated until a desired depth level is reached. Judging from experiments, a depth of two levels (search links and next level links) may be sufficient.

6.4. Extending the graph

If more information is required, one may use extracted named entities as search terms and perform the same procedure from the beginning. In order to limit the number of links to be processed, it may be advisable not to go beyond one additional level of depth (search results links) for collected named entities.

6.5. Data collection results

To test the approach, the team started with a list of 1,190 companies, 390 of which were classified as fintech companies in the Bundesbank statistics. We performed a search for each one of them without search term optimisation, but querying for both regular google search results and specific news items (50 for each). This yielded 39K links on the first level and 518K links on the second level. A total of 39GB of data was retrieved, with 6.3M named entities in total. Among those, 1.1M were different, unique named entities.

6.5. Graph embeddings and DNN results

In order to transform the collected graph to a statistical model for predicting whether a firm is fintech or not, it is required to decrease the amount of data by cleaning procedures. Filtering nodes by the number of connections reduces the number of graph nodes to 1/60 of the original and filtering by the signal content of nodes cut the number of nodes by another 4/5th. The latter was achieved by retaining those nodes that had either a high or a low fraction of edges leading to fintech entities, thus being informative. At the same time, the cleaning procedure reduces the number of edges, and thus the graph size, from 54M to 260K.

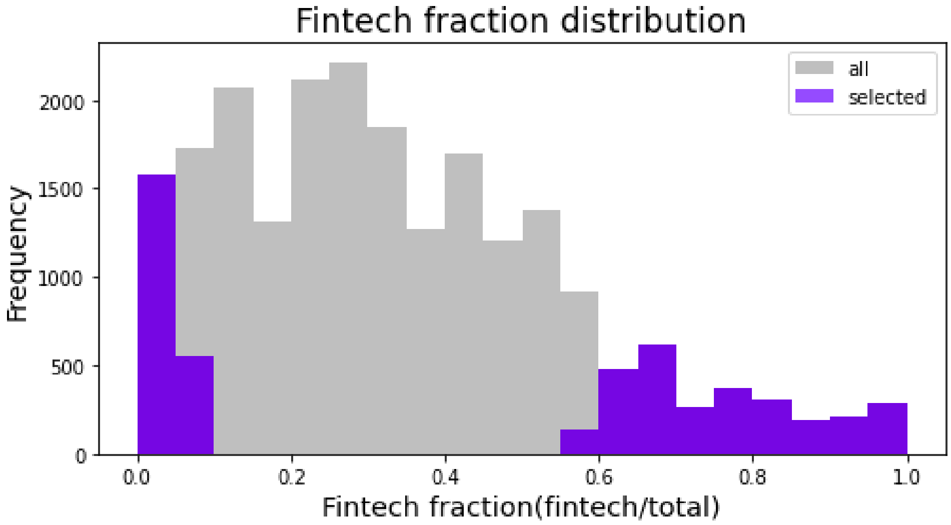


Figure 2: Selecting nodes based on their connections to fintechs and non-fintechs

Transformation from graph structure to a flat table is called node embedding and is carried out using the node2vec approach. Based on the deepwalk algorithm, the routine learns the graph structure distribution and trains a skip-gram encoder-decoder that can predict surrounding nodes from the characteristics of a given node. Taking the encoder hidden layer yields a trained neural network that can convert every node to a vector in the latent z-space. Figure 3 is a visual representation of the embedding process.

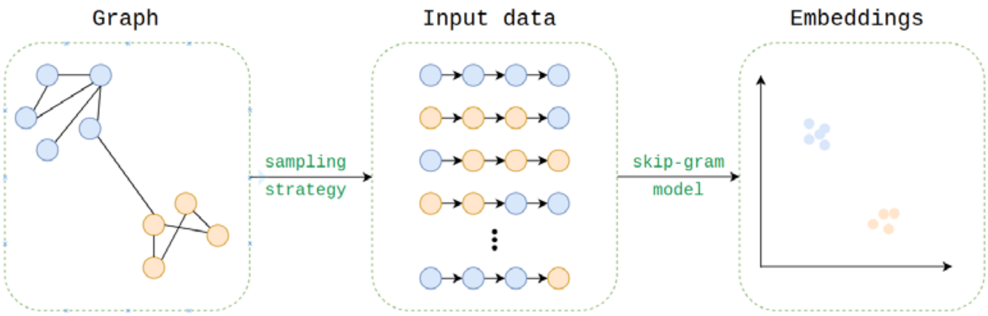


Figure 3: Embedding

The result in form of a 64 feature vector for every company is passed to a multilayer perceptron (DNN) of 64x16x1 layers with sigmoid activation function to determine the probability of the company being a fintech. Figure 3 is a t-SNE visualisation of data points by subset and label.

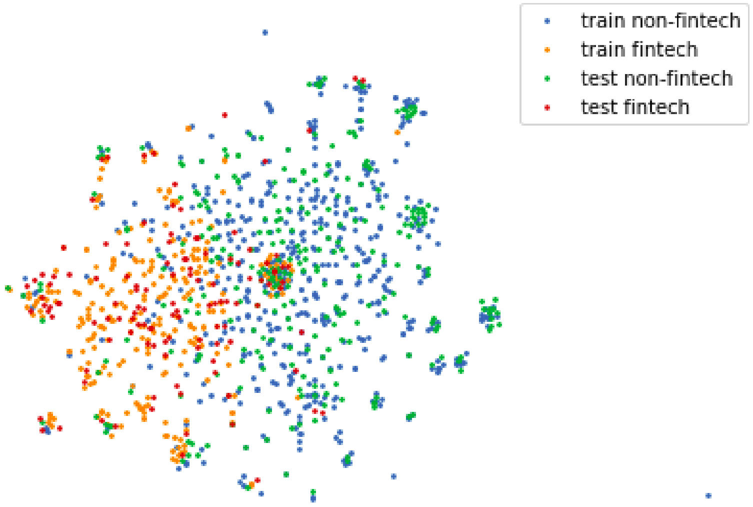


Figure 4: Visualisation of data point by subset and label

The evaluation of the model is carried out using k-fold cross validation. The dataset is split in three parts. In each round, the entire set of data is used for modelling, but only two of the three parts with the true label. In each turn, one of the three parts was used for evaluation, the other two for training. Ultimately, the metrics are averaged in the result table.

Graph approach: Algorithm performance

Table 1

Accuracy	Recall	Precision
0.87	0.73	0.88

The experiment has shown that public web data (news, webpages, articles) on companies can be used to classify them as fintech or non-fintech with relatively good accuracy. More graph and model tuning will improve the results to the production level.

7. Semi-supervised AI database filtering

Complementary to the Bundesbank work, the Banque de France approach has focussed on classifying/detecting companies based on their characteristics, predicting whether a company is a fintech or not. Those characteristics were obtained by means of a company information database provided by a third-party data broker and a dedicated semi-supervised classification algorithm that was built and trained on it.

The use case that the algorithm has to solve is: "Among a very large set of companies, filter out the few of them that can be considered as fintechs and extract the features that contribute the most."

7.1 Data quality and featuring

The third-party database contains mostly publicly available data such as web-scraped data from social networks, legal data from public registers, news and economic press articles as well as financial data. This database covers only French companies and contains more than 10 million referenced companies and more than 1,000 features.

As a proof of concept, it has been decided to limit our investigations to 10,000+ companies extracted from the database. To keep the strong imbalanced structure of the problem, the 10,000+ set of data has been built as follows:

- Around 350 companies have been selected and pre-tagged by the Banque de France experts as "potential fintechs".
- 10,000 "non-fintechs" extracted randomly from the database.

Data types are mostly categorical (e.g. economic sector, kind of newspaper, etc.), text-based (company description, title or text of news articles, employees job titles, etc.), numerical (e.g. turnover, number of employees, etc.) or dates (company establishment date, article publication date) and as a matter of consequence fully semi-structured.

As for most of these "big data" sources, a preliminary study has been conducted to get rid of the sparsity of the initial features (missing values), to select both the densest and the most business-oriented initial features. This has led to an initial draft set of 147 features including non-financial and financial data. In addition to this preliminary study, the iterative process of maximising the performance of the classification algorithm helped us to reduce this number to 84 features. Most of the financial features have been discarded because of their high level of sparsity, which creates too many false positives or false negatives in the classification results.

Once selected, the initial features were processed in the form of numerical categorical vectors (for categorical features), word embeddings or vectors of TF-IDFs, combining term frequency with inverse document frequencies (for text features) and scaled numerical features (for numerical features and dates) to be learnt or predicted by the algorithm.

7.2 Algorithm training and performance

The algorithm devised to classify an entity as a fintech or a non-fintech is semi-supervised. It undertakes one-sided learning of fintechs' characteristics, depicting non-fintechs as anomalies to the fintechs, and is based on a technique called "isolation forest".[5] The algorithm has been trained on a subset of the fintechs tagged by the experts of the Banque de France. The remaining fintechs and the non-fintechs have been kept for performance (prediction) evaluation, as could be done in real use. The semi-supervised nature of the algorithm helps to deal with the strong imbalance of the data (fintechs vs. non-fintechs). However, to promote better training, performance computation and thus model selection (involving the optimisation of hyperparameters), additional "synthetic fintechs" based on the "real" fintechs have been generated.

The best model obtained for our algorithm exhibits very low false negative and false positive rates. Accuracy, recall and precision are given in Table 2.

Semi supervised database filtering: Algorithm performance

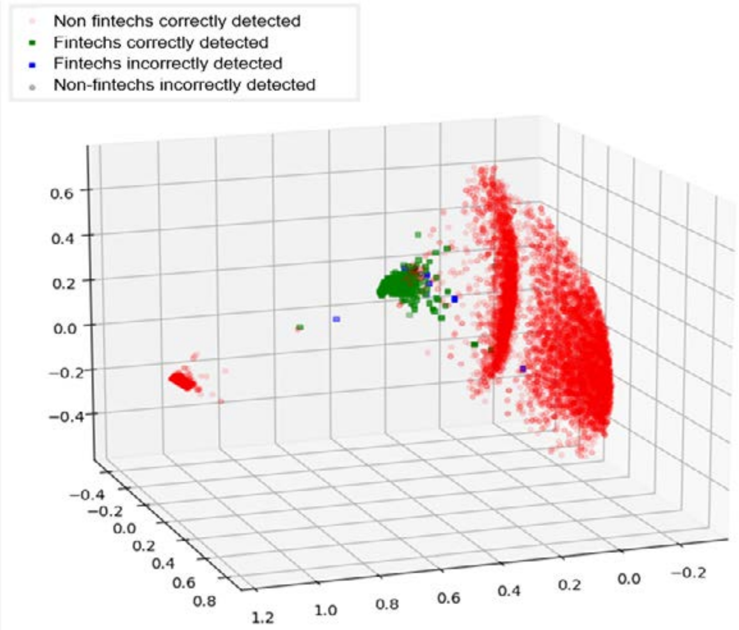
Table 2

	Accuracy	Recall	Precision
	0.995	0.991	0.980

Another way to visualise the result is to plot and colour the data points as functions of the algorithm results. Using a nonlinear kernel-based transformation (cosine based), the data can be projected and drawn in a 3D space. Each of the data points in this space corresponds to a company. The 10,000+ set of companies are plotted in Figure 5.

It is important to note the dense ball-like structure defined by the data points representing fintechs, which contrasts with the data points of non-fintechs, which inhabit a more elongated space. In addition to this, the distance between the fintechs and the non-fintech clusters is large enough to limit any overlapping between the two domains, improving the quality of partitioning and, as a consequence, the quality of the classification as fintech or non-fintech. These two observations highlight the relevance of the data preparation (the numerical featurisation step and the appropriate quality level of the data for our classification task).

[5] Liu, F.T., Ting, K.M., Zhou, Z.-H. (2008)



Non-fintechs correctly detected: true negatives

Fintechs correctly detected: true positives

Fintechs incorrectly detected: false positives

Non-Fintechs incorrectly detected: false negatives

Figure 5: Fintechs and non-fintechs plot (3D embedding axis)

7.3 Explainability of the results and the importance of features importance

The ten features that globally contribute the most to the decision to classify an entity as a fintech or a non-fintech include:

- news articles: topic, source of a paper/journal;
- administrative: activity code, description, description of goods/services;
- job titles of several employees in the companies;
- name of the executives/funds/people on the boards;
- sector of registered trade marks.

7.4 Lessons learnt and way forward

From our work, we have proven the ability to detect fintechs with a very high level of confidence in a time snapshot and subset of French companies.

The main challenge has been to deal with the large data volume, as well as the variety and quality of data. This is a very time-consuming task, requiring a lot of expertise and a dedicated infrastructure to accommodate big data, known as a data lake. Scaling up from 10,000+ companies to 10 million (French perimeter), and from

10 million to several hundred million (worldwide perimeter) is a challenge even for third-party database providers.

To go further, a first step would be to confirm the ability of the algorithm to scale up to the full French perimeter and detect new fintechs over time and to monitor the efficiency over time over learning cycles

A second step would be to scale up this algorithm to at least the European perimeter (or even larger) and check the volume, variety, quality and availability of the data at this scale. As part of this second step, it could be assessed how and to what extent the graph-based approach and "algorithm training" could be combined to optimize the overall performance of an identification and monitoring framework.

A third step would be to assess the typology of the fintechs, figuring out whether or not some of them could be grouped into distinct classes.

Conclusion

In the process of preparing a multi-purpose fintech monitoring system for statistics, regulation and financial stability, the Banque de France and the Deutsche Bundesbank have started to investigate two complementary approaches based on AI. Both of them consider a wide range of non-structured or semi-structured data directly crawled from the web or in a more traditional way from third-party data brokers – "big data like" company information databases. While the Bundesbank approach is focussed on data gathering and extracting the properties of the subsequent graph to find unknown fintechs, the Banque de France approach concentrates on learning the characteristics of fintechs in a semi-supervised way to find the unknown fintechs from the whole pool of firms within a country.

In both approaches, the main challenges are to address the scaling up of the AI tools, and to address the amount of data to gather, process and feature at the level of a country, an economic area (eg the euro area) or worldwide with a robust and dedicated "big data like" infrastructure/data lake.

Literature

Irving Fisher Committee on Central Bank Statistics, Towards monitoring financial innovation in central bank statistics, IFC Report 12, July 2020

Schueffel, P., "Taming the beast: a scientific definition of fintech, *Journal of Innovation Management,* Vol. 4 (4) (2016), pp. 32-54.

von Kalckreuth, U., Wilson N., *Fintech and statistics – the challenge of classifying something that hasn't existed before*, in Irving Fisher Committee on Central Bank Statistics, Towards monitoring financial innovation in central bank statistics, IFC Report 12, July 2020, pp. 126-136.

Liu, F.T., Ting, K.M., Zhou, Z.-H. *Isolation Forest*, 2008 Eighth IEEE International Conference on Data Mining, IEEE, January 2009, 978-0-7695-3502-9.

The Use of AI for Data Gathering - Finding and Monitoring Fintechs

IFC and Bank of Italy Workshop on “Data Science in Central Banking”

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

19-22 October 2021, virtual event hosted by the Bank of Italy

Elisabeth Devys, Banque de France

Ulf von Kalckreuth, Deutsche Bundesbank

Statistics and Fintech

- Fintech is the place where **innovation takes place in the financial sector**, where **new methods and products emerge, are tested and made ready for the market.**

- **As yet, fintech is not a separately defined field of statistical activity**, neither in the Bundesbank nor in most other central banks.

- We got **three key messages for you:**

  - Monitoring fintech is **important**
  - Monitoring fintech is **difficult**
  - Monitoring fintech is **feasible**

Monitoring fintech is important!

We need to **monitor fintech**:

- Much of innovation activity in the financial sector takes place in Fintechs. Fintechs are **essential for the growth dynamics** of the financial sector.

- Fintechs are of increasing importance for **financial stability**, **supervision**, **payment** and **monetary policy** – the key business areas of central banks.

- **IFC Working Group on Fintech Data Issues**, see final report of July 2020: https://www.bis.org/ifc/publ/ifc_report_monitoring_financial_innovation.pdf

Monitoring fintech is difficult!

- **Few standardised reporting requirements**
- No developed taxonomy and no established set of quantitative measures.
- By definition, innovation involves **new activities**. Intrinsically difficult for traditional statistics, which **needs stable classifications.**
- **Company registers mostly useless**.
- Business environment and market structures are **changing rapidly.**
- High rate of “metabolism” in the sector: **entry, merger & acquisition, exit**
- Any list of fintech companies is rapidly outdated.

We have to do without all the cornerstones of traditional company level statistics

Monitoring fintech is feasible!

- New ways of collecting data need to be explored. **Ways that use mostly publicly available information plus information that is shared voluntarily.**
- Market activity of fintechs are **almost exclusively web-based**. We can **make use of this!**
- Fintech monitoring **needs to be transnational –** firms often do not reside in the country where they are most active.
- To cope with innovation and market dynamics, we **need to become faster!**
- The **IFC WG on fintech data issues** mentions AI-supported data gathering tools as an **important solution element and recomments their inclusion in the BISH.**

Solving this issue will be very important also for statistics on other innovative areas!

Monitoring fintech is feasible!

Web based avenues:

- **Web-scraping** in connection with AI methods (mostly NLP) may enable finding new fintechs as well as tracking the activities of known companies. With an operational list of fintechs at hand, we can use supervised or unsupervised learning on company information. Eg information on startups from business intelligence services or company websites.
- **AI augmented search** of **organised information platforms.**
- Using **AI based methods of probabilistic matching**, mappings of websites and product labels to universal registers such as RIAD or the statistical firm registers can be accomplished. This links information on economic activity to information on legal units.

Monitoring fintech is feasible!

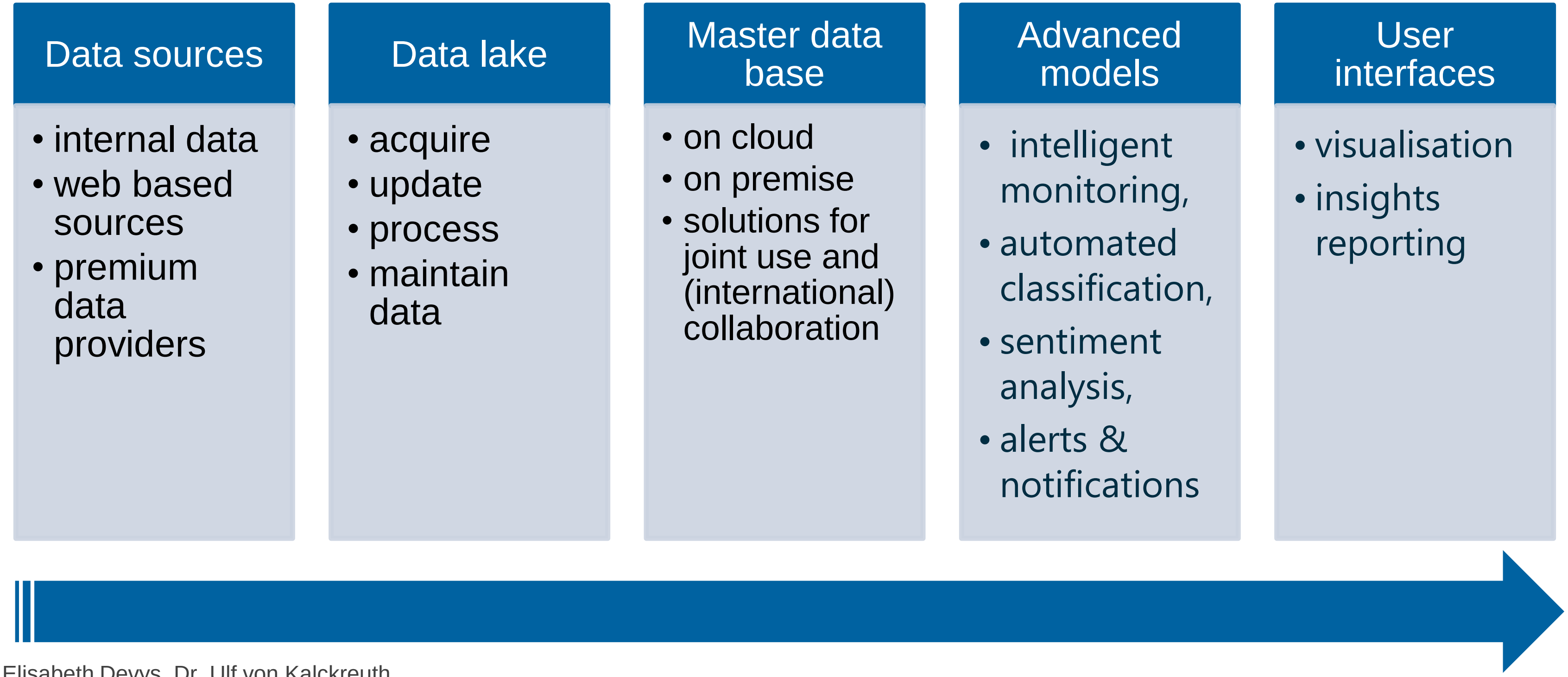
The challenge has **two dimensions**, each interesting in its own right:

- Monitoring **known fintechs:**
  - Tracking information brokers, lists and web fora on fintech
  - Monitoring websites for relevant changes in activity – help classification
  - Report events: mergers, acquisitions, insolvencies, exits
  - Download balance sheets and other financial statements,
  - Feed key positions into information systems
  - „Know your client“ systems
- Finding **new fintechs:**
  - Tracking information brokers, lists and web fora on fintech
  - Possibly scraping websites from companies in a larger list of IT companies.

Monitoring must be set up **internationally**

Solution elements

Some elements of a fully fledged solution



Data sources

- internal data
- web based sources
- premium data providers

Data lake

- acquire
- update
- process
- maintain data

Master data base

- on cloud
- on premise
- solutions for joint use and (international) collaboration

Advanced models

- intelligent monitoring,
- automated classification,
- sentiment analysis,
- alerts & notifications

User interfaces

- visualisation
- insights reporting

Responding to the challenge

Banque de France activities: creation of a pilot project by the Lab & DDSA teams

Step 1: getting a master database

- we contracted with a data provider:

→ database of 10 millions companies (among which 354 are tagged as fintechs, and a subset of 10,000 non-fintechs selected randomly for the challenge), > 1400 variables

→ webscrapped data, legal data from public registers, data from press articles, from public tenders, etc…

→ financial data, for which the density was low, were discarded (it increased the number of false positives and degraded the performance)

→ we isolated the 84 most useful variables

Step 2: preparation of the data

- joining the data tables
- featurisation of the variables (textual, categorical variables,…)

Responding to the challenge

Banque de France activities: creation of a pilot project by the Lab & DDSA teams

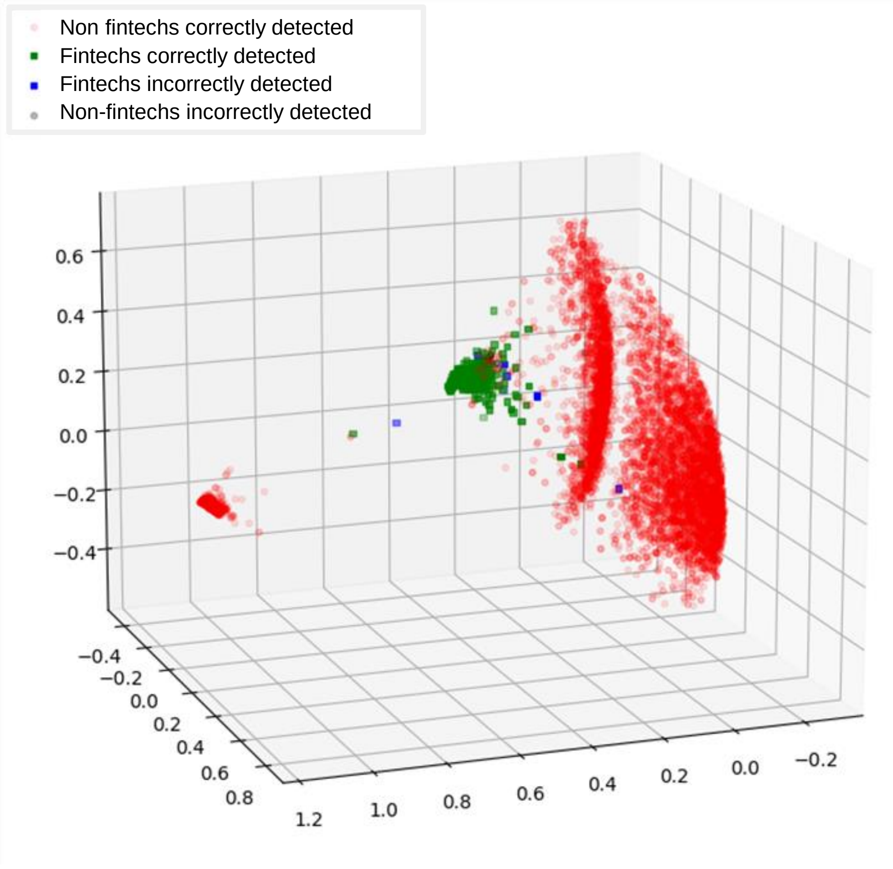
Step 3: algorithm of identification of the fintechs

- Semi-supervised classification algorithm (isolation forest)

- Use of a dedicated synthetic data generation algorithm to tackle the imbalance problem between fintechs and non-fintechs, allowing a better training and model selection.

→ High accuracy and very low false negative and false positive rates.

Accuracy	99,5%
Recall	99,1%
Precision	98%



Responding to the challenge

Banque de France activities: creation of a pilot project by the Lab & DDSA teams

First results:

→Among the ten variables that contribute most to explainability, we found:

- **Variables from press articles**: theme, description, source of the articles
- **Administrative variables**: activity code, size, denomination
- Variables linked to **the type of positions** in the company
- Names of the **people / funds in the Board**
- Sector of the **registered trademark**

Lessons learnt:

- One of the main challenges is the volume, variety and quality of data
- Getting all the data for the Master data base is a necessary, preliminary, but very time-consuming task
- Robust infrastructures will be required, especially if the analysis is expanded on a worldwide perimeter. Hundreds of millions of companies should potentially be analyzed

Way forward:

1) **Dynamic actualization** → is the model able to detect new fintechs over time?
2) **Geographical extension** to Europe, or to an even larger perimeter
3) **Classification** → would it be possible to create a model that would classify fintechs (among the different types of fintechs)?

Responding to the challenge

Bundesbank activities:

- Integrated three information sets on fintech companies -- with each other and with RIAD, the company master data base of the Eurosystem
- Fintech Monitoring in the Bundesbank as a rather singular inter-departamental co-operative initiative
- Making the project part of the Bundesbank Innovation Challenge in November/ December 2020: rich and interesting results
- Tech-scouting with Tech-Quartier, a startup hub, in Dec 20 - Feb 21 to learn about feasibility options
- Small scale pilot as joint work fintech company in 2021, starting in September 1st
- Approach: Using graph analysis for identifying and classifying fintech companies
- PoC expected by summer 2022

Responding to the challenge

Many thanks!

Points of contact at the Banque de France:
elisabeth.devys@banque-france.fr and guillaume.belly@banque-france.fr

Points of contact at the Bundesbank:
ulf.von-kalckreuth@bundesbank.de and maximilian.koenig@bundesbank.de

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Novel methodologies for data quality management
Anomaly detection in the Portuguese central credit register[1]

André Faria da Costa, Francisco Fonseca and Susana Maurício,
Bank of Portugal

[1] This presentation 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 central banks and other institutions represented at the event.

Novel Methodologies for Data Quality Management
Anomaly Detection in the Portuguese Central Credit Register[1][2]

André Faria da Costa

Banco de Portugal, Lisboa, Portugal – anfcosta@bportugal.pt

Francisco Fonseca

Banco de Portugal, Lisboa, Portugal – ffonseca@bportugal.pt

Susana Maurício

Banco de Portugal, Lisboa, Portugal – scmauricio@bportugal.pt

Abstract

Since 2018, with the launch of the new Central Credit Register (CCR) covering, on a loan-by-loan basis, all loans to legal and natural persons and the corresponding credit risk, the granularity and volume of data collected by Banco de Portugal in this context have increased tremendously. The CCR receives roughly 20 million highly granular records every month, spread across more than 200 attributes, corresponding to individual instruments. The collection of such granular data has led to new challenges concerning data quality management (DQM), in particular how to detect in an efficient and effective way possible outliers in particular instruments or in a specific attribute of a particular instrument.

Due to the complexity of the CCR database, a single model would likely be unable to detect the majority of potentially anomalous data. In this paper, we will showcase the implementation of a set of methodologies which seek to cover as much potentially anomalous data as possible. The starting point of this process was the exploration of the Isolation Forest algorithm, which is based upon random forests and specifically designed for anomaly detection in large data sets. Afterwards, we have developed additional methodologies which complement this algorithm. Finally, we have implemented a Power BI dashboard where the results are presented, taking advantage of its visuals. This approach has allowed for the integration of all the different outputs from the DQM processes, and it notoriously increased the usability of results by the analysts.

Keywords: Banco de Portugal; Central Credit Register; Data Quality Management; Outlier Detection; Anomaly Detection; Isolation Forest

JEL classification: C80

[1] The views expressed are those of the authors and do not necessarily reflect those of Banco de Portugal.

[2] Acknowledgements: We would like to express our gratitude to Homero Gonçalves and Marta Veloso for all their valuable contributions and insights during this project and also to António Simões for helping us set up the IT Infrastructure. We would also like to thank André Fernandes, Daniel Peixeiro, Pedro Cordeiro, Rita Pisco and Rui Purificação for their collaboration in the team that provided the foundation for the development of these new tools. And last, but not least, we are thankful to Mário Lourenço for his revision and valuable suggestions.

Contents

1. Introduction

The use of machine-learning (ML henceforth) based methods for the purpose of anomaly detection has gained significant traction since the beginning of the 21st century. This increase in popularity is in part due to technological advancements in the field of ML as a whole, but also due to a massive increase in data granularity and complexity. Faced with massive and complex databases, more traditional outlier detection methods will (in general) perform poorly, especially from the computational point of view.

In the context of this paradigm shift regarding data volume and complexity, Banco de Portugal is not an exception. This is particularly apparent since the launch of the new Central Credit Register (CCR henceforth) system, in September 2018, which operates on an instrument-by-instrument logic, rather than a debtor-by-debtor one. Additionally, this new system follows a single service desk approach, which resulted in a widening of the set of variables reported. These changes resulted in a huge increase in the volume of information to be analysed - over 20 million monthly records, each characterized by more than 200 possible attributes.

Furthermore, as a central bank, data quality is of the utmost importance for Banco de Portugal. As such, the development and implementation of new data quality control methodologies, ML-based or otherwise, that support the management of this large and complex database becomes a necessity.

To this end, in late 2019, a team at the Statistics Department began researching novel methodologies that could potentially be used for data quality control on the CCR database, some of which were ML based outlier detection methods. The main objective of the team was the development and implementation of methodologies which were effective at identifying anomalies, with a low false negative rate and a reasonable false positive rate, while also being efficient from a computational point of view. Furthermore, our data is unlabelled, and for this reason, we looked mostly at unsupervised learning ML algorithms.

In this paper, we will highlight the three methodologies that were put into production as a result of the research conducted by this team:

- A pattern based test, which aims to detect potential inconsistencies in the reporting of any given instrument|debtor pairing in the database;
- A concentration check to oversee the evolution of the reporting of the categorical variables that are harder to handle for most outlier detection algorithms, including those based on ML models;
- An application of the Isolation Forest (IF henceforth) algorithm, which as the name indicates is a type of random forest based algorithm specifically designed for the purpose of isolating potentially anomalous observations. In our case, we use the IF algorithm to identify potentially anomalous behaviour in the evolution of reported outstanding amounts.

The remainder of this paper is organized in the following way: in section 2, we briefly describe the Portuguese CCR. In section 3, we discuss the implemented methodologies in detail. In section 4, we describe the way each test was implemented and the way the results are presented to the analysts. In section 5, we present our main conclusions and discuss the effectiveness of these algorithms across the full year that they have been used in production.

2. The Portuguese Central Credit Register

The Portuguese CCR is a system managed by Banco de Portugal, which gathers on a monthly basis a wide range of information provided by the observed agents (credit-granting institutions) associated with actual and potential credit liabilities of their costumers (natural or legal persons).

The main purpose of the CCR is to provide support to the credit-granting institutions in their assessment of counterparty risk. The registered entities have access to all the credit liabilities of their (actual and potential) customers, vis-à-vis the entirety of the resident financial system.

The data reported to the CCR, which is rich in both volume and complexity, is used for a variety of purposes – compiling statistics, banking supervision, financial stability analysis, informing monetary policy decisions, among many other possible uses.

Due to the multitude of purposes for which CCR data is used, data quality assurance is of the utmost importance. Prior to the present work, there were already multiple tests in place to ensure the consistency and coherence of the reported information, including:

- A set of validation tests that ensure that the reporting institutions abide by the established reporting rules;
- Cross-validation with other (internal and external) data sources that have some overlap with information reported to the CCR (like data from the Central Balance Sheet database, from Securities Statistics and from Statistics Portugal, among others);
- Analysis of the most significant (absolute and relative) month-on-month and homologous variations in the reported amounts, with the intent of identifying potentially anomalous variations that may be an indication of reporting errors.

As was previously stated, the CCR is large both in volume, as can be seen in Figure 1, but also in complexity – each record is characterized by more than 200 possible attributes, both quantitative (for example, outstanding amounts) and categorical (for example, the purpose of the credit). For this reason, and due to the importance of the CCR as a data source for both internal and external users, the need for new methodologies that complement the above-mentioned tests became evident. In the next sections, we will describe some of the methodologies that were developed and implemented with the intent of increasing the efficiency and the effectiveness of the data quality assessment process.

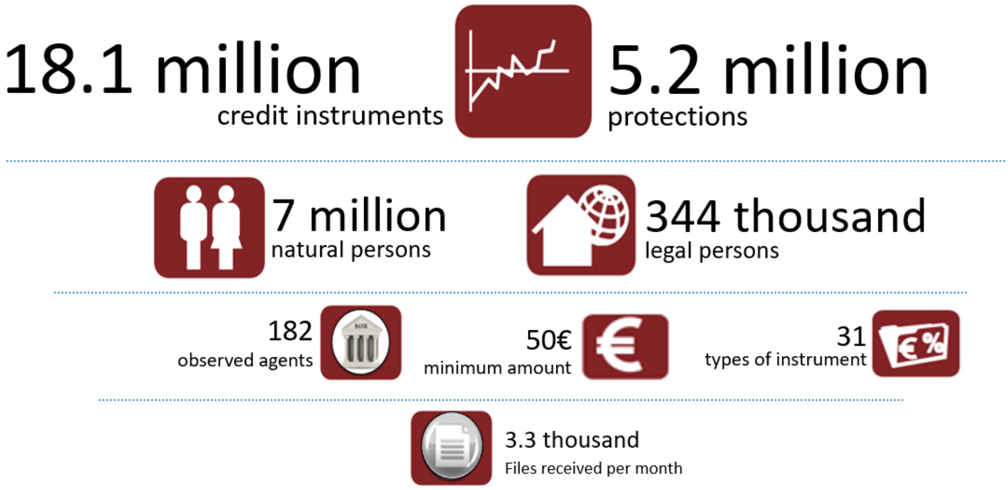


Figure 1 CCR indicators infographic

3. Selected Methodologies

Reporting Consistency Test

As mentioned in the previous section, the CCR receives information on a monthly basis, and as such it is highly relevant to ensure that each instrument is reported consistently until maturity. To this end, we developed a pattern-based test with the purpose of evaluating the reporting stability on an instrument-by-instrument and debtor-by-debtor basis.

The test looks at the last six months of a given instrument|debtor's lifetime and assigns a number between 0 and 2 based on the instrument|debtor's status in a given month:

- A value of 0 indicates that the instrument|debtor is not present in the database;
- A value of 1 indicates that the instrument|debtor is present in the database as active;
- A value of 2 indicates that the instrument|debtor is present in the database as finalised.

Taking into account the very large volume of information associated with this test (6 months of information, with over 20 million instrument|debtor pairs each month), we further define five types of patterns that are likely a sign of a reporting error, and as such should be analysed further:

- Instruments that are reported as finalised with non-zero outstanding amounts;
- Instruments reported as finalised and then as active in a following period;
- Instruments with reporting gaps (for example reported at time n, not reported at time n+1, and reported at time n+2);
- Instruments that cease to be reported without being finalised;
- Instruments that are reported as finalised for multiple periods (an instrument should be reported as finalised only once, and then it should not be reported in subsequent months).

Furthermore, to provide further context when presenting the results of this test, we also present the corresponding amounts. This allows us to quantify the impact that these potentially anomalous instruments have on the stock of credit as a whole.

Concentration Check

The concentration check test was developed with the purpose of tackling categorical variables that are reported to the CCR, since they are generally harder to handle for most anomaly detection algorithms.

For each combination of variable/observed agent/type of instrument, and for the latest four months of information reported, we compute the percentage of instruments that are reported with each possible element for the categorical variable. We also compute the percentage when considering the totality of instruments reported to the CCR (global averages).

As an example, if we are analysing the variable "Type of negotiation", for which the possible elements are "New operation", "Automatic renewal", "Regular renegotiation" and "Renegotiation due to default", we will compute the percentage of instruments reported with each of these possible elements, for both every individual institution and for the totality of the system, in the last four months.

The purpose of this test is two-fold:

- We can see if, for a given variable, the weight of any given element, within a given observed agent, has changed significantly in the last four months;
- We can see if, for a given variable, the weight of a given element, within a given observed agent, differs significantly from the weight when considering the totality of the system.

Furthermore, it is relevant to note that some of the variables which are utilized in the IF model are categorical in nature, meaning that this test also helps to ensure the quality of the data that is fed into the IF algorithm.

Isolation Forest

In order to select the ML model to be implemented, we had to take into account that, due to the volume and complexity of the database, and due to the way it was designed, we do not have a variable that classifies a given observation as being correctly or incorrectly reported. Hence, the use of supervised methods was excluded *a priori*, since we do not have a target variable.

From the pool of unsupervised methods, we considered two main model branches – clustering models (in particular, DBSCAN) and density-based models (in particular the IF algorithm).

In the case of clustering models, the DBSCAN algorithm performed well in testing. However, its complexity is $O(n^2)$, and since we are dealing with a very large dataset, the algorithm didn't perform well from a computational standpoint. This is particularly relevant when taking into consideration the fact that our intention was for the model to be ran frequently. Furthermore, estimating parameters is challenging, since DBSCAN is quite sensitive to small changes in its parameters.

Concerning density-based models, we tested the IF algorithm and found that it had many characteristics that we found advantageous:

- It performed well in testing;
- It has $O(n \log n)$ complexity (quasilinear);
- It is a scoring model, allowing us to establish a priority list that will let the analysts focus on the observations which have a higher likelihood of being anomalous;
- The task of parameter estimation is simpler than in the case of DBSCAN, since the algorithm is much less sensitive to small changes in its parameters.

This algorithm was initially proposed and described in detail in (Liu, Ting, & Zhou, 2009).

Summarily, the IF algorithm identifies anomalies through a process the authors refer to as a process of isolation – we select a subsample of our dataset and perform successive random partitioning, by first randomly selecting a variable and then randomly selecting a split value between the maximum and minimum values for that variable.

This process of recursive partitioning results in a binary tree structure. A node in the tree is considered a terminal node if it contains a single observation, or if all the observations contained within it have the same values for all attributes.

The partitioning process is carried out recursively until all nodes are terminal nodes. To improve performance, the process can also be stopped prematurely when we reach a pre-defined maximal tree depth.

Rather intuitively, the smaller the number of partitions needed for an observation to end up in a terminal node, the more isolated it is, and hence the likelier it is to be an anomaly. The model thus attributes a higher anomaly score to the observations that are more easily isolated.

Concerning model specification, it is relevant to establish that the main purpose of this model is the detection of unusual variations in the amounts reported to the CCR.

As was already mentioned in section 2, the CCR incorporates a wide variety of highly heterogeneous types of instrument; hence, we decided early on to calibrate a separate model for each type of instrument. Furthermore, we focused only on types of instrument that should have a regular payment, and for which that communication is mandatory.

Due to the nature of the IF algorithm, we also had to carefully choose the variables to be included – the model works through a process of isolation, so we had to pick variables where values (or combinations of values) that deviate from the norm are more than likely anomalies. For this reason, rather than using the variations of the raw reported amounts as model variables, we combined a set of variations that, generally, should cancel each other out into a new variable (henceforth referred to as "Residual Variation"):

$$\text{Residual Variation} = \text{Abs}(\Delta\text{Outstanding nominal amount} + \Delta\text{Accumulated write-offs} + \text{Payment} \\ + \Delta\text{Off-balance sheet amount} + \text{Early repayment})$$

The construction of this new variable not only ensures that large values will more than likely be an indication of an anomaly in the reporting, but also reduces the number of variables to be included in the model, as it combines five numerical variables, without significant loss of information. However, to provide further context when analysing the results, we also present the raw amount variations to the analysts.

Adding to this numerical variable, we also include a set of categorical variables that characterize the loan (numerically encoded due to high cardinality):

- Purpose of the credit;
- Type of negotiation;
- Residual maturity (in five year steps).

It is also relevant to note that we need to be concerned about the special case of instruments that have a payment frequency that is not monthly. For this kind of instrument, although a due payment is reported on a monthly basis, this payment will only be reflected in a variation of the outstanding amounts when it is due (once every three, six or twelve months, in general). Thus, in months where a payment is not due, we will have (simplifying) $Residual\,Variation = Payment$. To avoid these instruments from showing up as potential anomalies, we filter out instruments where a payment is reported but no variation is recorded in any of the remaining amounts in a given month. This significantly reduced the false positive rate.

We use five consecutive months of reported information, meaning we have four deltas (variations) – the first three are used to train the model, and the last one corresponds to the production month.

Finally, it should also be noted that, since our numerical variable is a variation, the model will not be able to evaluate instruments, which are reported for the first time in the production month (since we are unable to calculate variations).

4. Implementation and Main Results

Workflow and Technologies Used

After defining the methodologies we wished to implement, we had to outline an operational workflow and to settle on which tools were to be used in the implementation process.

The reporting consistency and concentration check tests were implemented through the use of SQL procedures, and the IF test was implemented in Python, where we use the IF algorithm implemented in the Scikit-learn library (Pedregosa, et al., 2011). The test results are stored in a SQL Server database, for all three tests. The use of SQL, whenever possible, is a natural consequence of the fact that all of the data that is reported to the CCR is stored in SQL databases. For this reason, using SQL (or, in the case of Python, connecting to it via ODBC) minimizes the time required to transfer large volumes of data, making the process as efficient as possible.

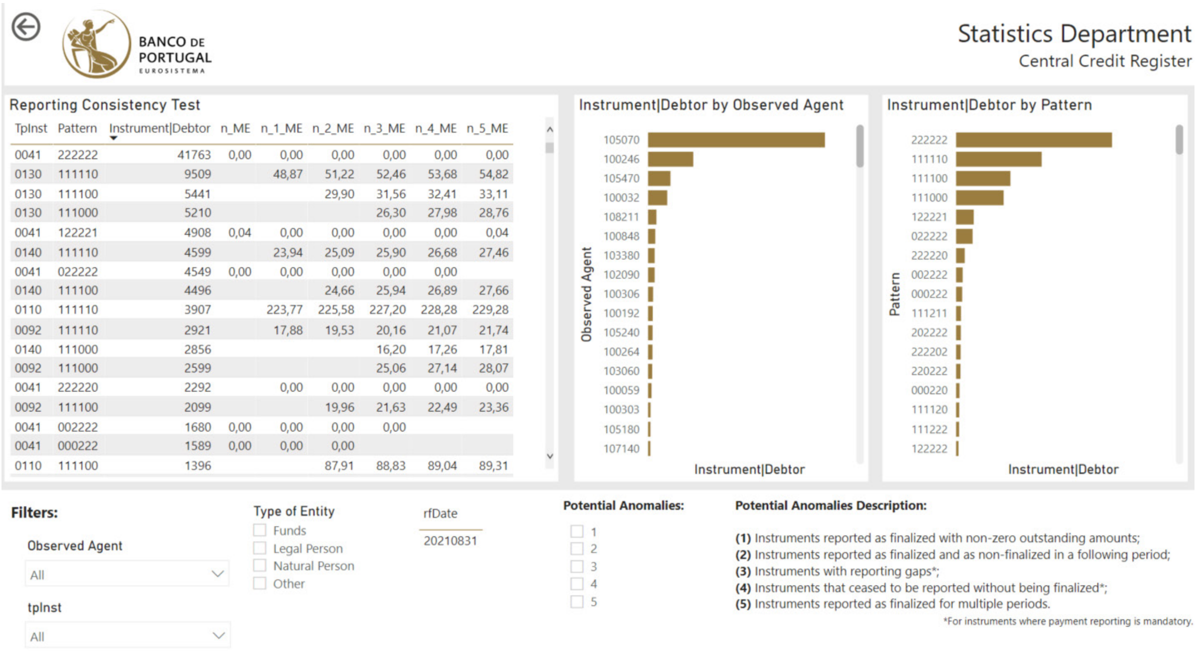
As a final step, for data exploration and visualization, we implemented a Power BI dashboard, that allows the analysts to easily access and interact with the results. The flexibility of the dashboard lets the analysts look at the results in the way that best suits the task at hand and select the cases that should be sent to the reporting institutions for further clarifications.

In the remainder of this section, we will provide real examples of anomalies that were detected and corrected through the use of each of the three tests we described previously. For confidentiality reasons, all identifying characteristics (observed agent, instrument and instrument identifiers) are anonymised.

Reporting Consistency Test

As described in the previous section, this test aims to evaluate the reporting consistency of the reporting institutions, through the construction of a pattern that characterizes the last six months of information for a given instrument|debtor pairing. The pattern is thus composed of six digits, where the last digit (rightmost) corresponds to the most recent month.

The results are made available in a Power BI dashboard, where we display the potentially anomalous patterns for each reporting institution and type of instrument. The use of Power BI makes the task of analysing the results more intuitive and far more flexible. Using filters, we can either look at the system as a whole, or select a single observed agent or type of instrument to narrow the scope of the analysis. The final dashboard (without any filtering) is shown in Figure 2.



Statistics Department
Central Credit Register

Reporting Consistency Test

TpInst	Pattern	Instrument|Debtor	n_ME	n_1_ME	n_2_ME	n_3_ME	n_4_ME	n_5_ME
0041	222222	41763	0,00	0,00	0,00	0,00	0,00	0,00
0130	111110	9509		48,87	51,22	52,46	53,68	54,82
0130	111100	5441			29,90	31,56	32,41	33,11
0130	111000	5210				26,30	27,98	28,76
0041	122221	4908	0,04	0,00	0,00	0,00	0,00	0,04
0140	111110	4599		23,94	25,09	25,90	26,68	27,46
0041	022222	4549	0,00	0,00	0,00	0,00	0,00	
0140	111100	4496			24,66	25,94	26,89	27,66
0110	111110	3907		223,77	225,58	227,20	228,28	229,28
0092	111110	2921		17,88	19,53	20,16	21,07	21,74
0140	111000	2856				16,20	17,26	17,81
0092	111000	2599				25,06	27,14	28,07
0041	222220	2292		0,00	0,00	0,00	0,00	0,00
0092	111100	2099			19,96	21,63	22,49	23,36
0041	002222	1680	0,00	0,00	0,00	0,00		
0041	000222	1589	0,00	0,00	0,00			
0110	111100	1396			87,91	88,83	89,04	89,31

Filters:

Observed Agent
All

tpInst
All

Type of Entity
□ Funds
□ Legal Person
□ Natural Person
□ Other

rfDate
20210831

Potential Anomalies:
□ 1
□ 2
□ 3
□ 4
□ 5

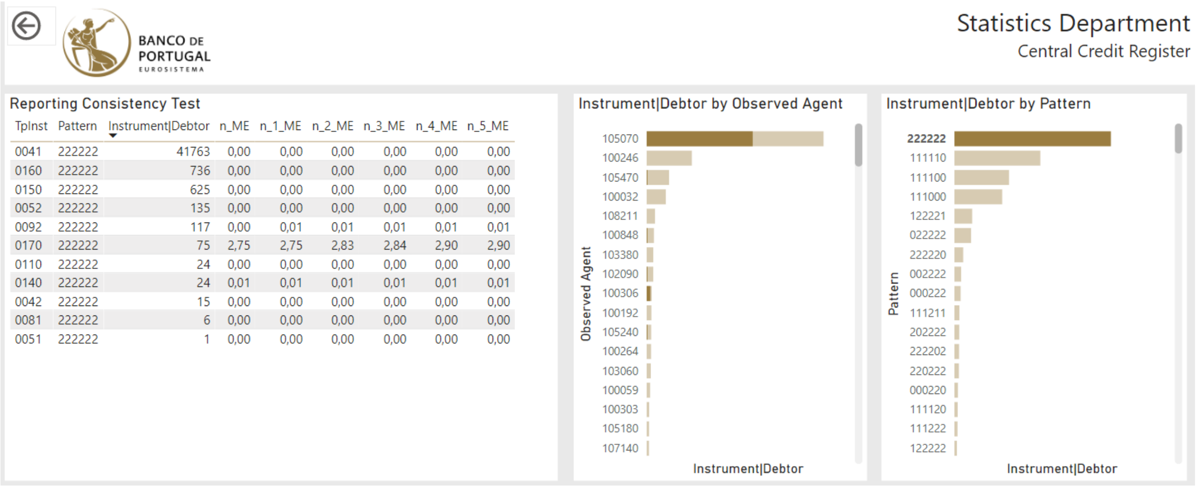
Potential Anomalies Description:

(1) Instruments reported as finalized with non-zero outstanding amounts;
(2) Instruments reported as finalized and as non-finalized in a following period;
(3) Instruments with reporting gaps*;
(4) Instruments that ceased to be reported without being finalized*;
(5) Instruments reported as finalized for multiple periods.
*For instruments where payment reporting is mandatory.

Figure 2 Dashboard view of the Reporting Consistency Test

On the right hand side, we can easily see the frequency with which each pattern shows up in our data, and the way these patterns are distributed amongst the observed agents. On the left hand side, we have a table where, for a given type of instrument and anomalous pattern, we show the number of instruments|debtors that display such a pattern, and their corresponding amounts in each of the six dates that make up the pattern. We can also perform a drill-down in order to drop to the level of each individual instrument|debtor, which often proves useful when contacting the reporting institutions to request clarification or correction.

As an example of how the use of the presented filters can be helpful, if we filter for the pattern "222222" (Figure 3), we immediately see that institution 105070 is responsible for the vast majority of instruments that display this pattern. We can also see the corresponding number of instrument|debtor pairs and the amounts involved. In this case, as can be inferred in Figure 3, we have a few type 1 (Instruments reported as finalised with non-zero outstanding amounts) and multiple type 5 (Instruments reported as finalised for multiple periods) patterns. Even though we have detected just a few type 1 patterns, those cases are always analysed carefully since they indicate the existence of liabilities that are associated with terminated instruments, which is in general a sign of a reporting error with impact in the amounts.



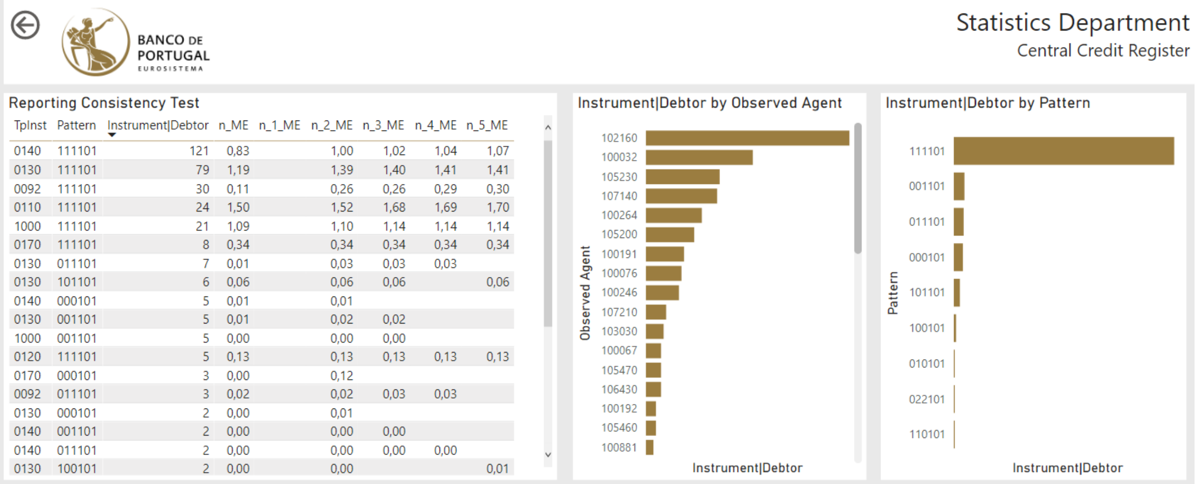
Statistics Department
Central Credit Register

Reporting Consistency Test

TpInst	Pattern	Instrument|Debtor	n_ME	n_1_ME	n_2_ME	n_3_ME	n_4_ME	n_5_ME
0041	222222	41763	0,00	0,00	0,00	0,00	0,00	0,00
0160	222222	736	0,00	0,00	0,00	0,00	0,00	0,00
0150	222222	625	0,00	0,00	0,00	0,00	0,00	0,00
0052	222222	135	0,00	0,00	0,00	0,00	0,00	0,00
0092	222222	117	0,00	0,01	0,01	0,01	0,01	0,01
0170	222222	75	2,75	2,75	2,83	2,84	2,90	2,90
0110	222222	24	0,00	0,00	0,00	0,00	0,00	0,00
0140	222222	24	0,01	0,01	0,01	0,01	0,01	0,01
0042	222222	15	0,00	0,00	0,00	0,00	0,00	0,00
0081	222222	6	0,00	0,00	0,00	0,00	0,00	0,00
0051	222222	1	0,00	0,00	0,00	0,00	0,00	0,00

Figure 3 Reporting Consistency Test – filtering by a specific pattern

Type 3 patterns are also prioritised, since they show the existence of reporting gaps in the instrument|debtor's lifetime. As can be seen in Figure 4, we have a very small number of records that present a type 3 pattern and they are mostly associated with a gap at time n-1.



Statistics Department
Central Credit Register

Reporting Consistency Test

TpInst	Pattern	Instrument|Debtor	n_ME	n_1_ME	n_2_ME	n_3_ME	n_4_ME	n_5_ME
0140	111101	121	0,83		1,00	1,02	1,04	1,07
0130	111101	79	1,19		1,39	1,40	1,41	1,41
0092	111101	30	0,11		0,26	0,26	0,29	0,30
0110	111101	24	1,50		1,52	1,68	1,69	1,70
1000	111101	21	1,09		1,10	1,14	1,14	1,14
0170	111101	8	0,34		0,34	0,34	0,34	0,34
0130	011101	7	0,01		0,03	0,03	0,03	
0130	101101	6	0,06		0,06	0,06		0,06
0140	000101	5	0,01		0,01			
0130	001101	5	0,01		0,02	0,02		
1000	001101	5	0,00		0,00	0,00		
0120	111101	5	0,13		0,13	0,13	0,13	0,13
0170	000101	3	0,00		0,12			
0092	011101	3	0,02		0,02	0,03	0,03	
0130	000101	2	0,00		0,01			
0140	001101	2	0,00		0,00	0,00		
0140	011101	2	0,00		0,00	0,00	0,00	
0130	100101	2	0,00		0,00			0,01

Figure 4 Reporting Consistency Test – filtering by a specific anomaly type

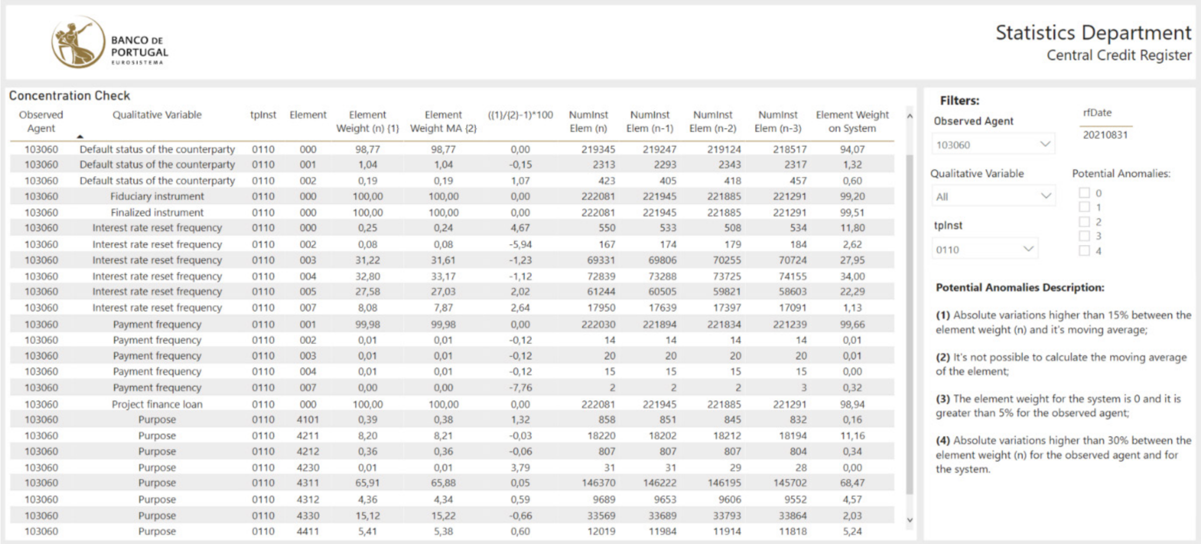
As was shown in the preceding examples, the set of potentially anomalous instrument|debtor pairs is quite low when taking into consideration the volume of the database. We can also see that, looking at the amounts involved, they are most significant in the case of type 4 patterns, that is, the observed agent ceased to report the instrument without first reporting it as finalised. Although this pattern may result from gaps in reporting, in most cases it simply means that the observed agent did not report the contract as finalised before ceasing to report it. Thus, this situation, while anomalous, was not our primary focus since it generally does not affect the conclusions we can derive from the CCR data, as finalised instruments do not have amounts different from 0.

Although the anomalies detected in this test may not result in a huge effect on the aggregates, they may have a significant impact for each individual debtor. Hence identifying them and requesting clarification or correction is highly relevant to the CCR's main purpose: provide the registered entities with information on all the credit liabilities of their (actual and potential) clients, vis-à-vis the resident financial system, in order to aid in their assessment of counterparty risk.

Concentration Check

As we have previously stated, the concentration check test allows us to evaluate the quality of the reporting of categorical variables to the CCR at a given moment, and it lets us assess how it has evolved over time. As with the remaining tests, this analysis is usually performed by looking at a single combination of observed agent and type of instrument, but we can also look at the state of the system as a whole. To aid in this analysis, we define four filters that are intended to highlight potential anomalies of different types.

As with the previous test, we display the results in a Power BI dashboard:



Statistics Department
Central Credit Register

Concentration Check

Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Default status of the counterparty	0110	000	98,77	98,77	0,00	219345	219247	219124	218517	94,07
103060	Default status of the counterparty	0110	001	1,04	1,04	-0,15	2313	2293	2343	2317	1,32
103060	Default status of the counterparty	0110	002	0,19	0,19	1,07	423	405	418	457	0,60
103060	Fiduciary instrument	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	99,20
103060	Finalized instrument	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	99,51
103060	Interest rate reset frequency	0110	000	0,25	0,24	4,67	550	533	508	534	11,80
103060	Interest rate reset frequency	0110	002	0,08	0,08	-5,94	167	174	179	184	2,62
103060	Interest rate reset frequency	0110	003	31,22	31,61	-1,23	69331	69806	70255	70724	27,95
103060	Interest rate reset frequency	0110	004	32,80	33,17	-1,12	72839	73288	73725	74155	34,00
103060	Interest rate reset frequency	0110	005	27,58	27,03	2,02	61244	60505	59821	58603	22,29
103060	Interest rate reset frequency	0110	007	8,08	7,87	2,64	17950	17639	17397	17091	1,13
103060	Payment frequency	0110	001	99,98	99,98	0,00	222030	221894	221834	221239	99,66
103060	Payment frequency	0110	002	0,01	0,01	-0,12	14	14	14	14	0,01
103060	Payment frequency	0110	003	0,01	0,01	-0,12	20	20	20	20	0,01
103060	Payment frequency	0110	004	0,01	0,01	-0,12	15	15	15	15	0,00
103060	Payment frequency	0110	007	0,00	0,00	-7,76	2	2	2	3	0,32
103060	Project finance loan	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	98,94
103060	Purpose	0110	4101	0,39	0,38	1,32	858	851	845	832	0,16
103060	Purpose	0110	4211	8,20	8,21	-0,03	18220	18202	18212	18194	11,16
103060	Purpose	0110	4212	0,36	0,36	-0,06	807	807	807	804	0,34
103060	Purpose	0110	4230	0,01	0,01	3,79	31	31	29	28	0,00
103060	Purpose	0110	4311	65,91	65,88	0,05	146370	146222	146195	145702	68,47
103060	Purpose	0110	4312	4,36	4,34	0,59	9689	9653	9606	9552	4,57
103060	Purpose	0110	4330	15,12	15,22	-0,66	33569	33689	33793	33864	2,03
103060	Purpose	0110	4411	5,41	5,38	0,60	12019	11984	11914	11818	5,24

Filters:

Observed Agent: 103060

rfDate: 20210831

Qualitative Variable: All

tpInst: 0110

Potential Anomalies:
□ 0
□ 1
□ 2
□ 3
□ 4

Potential Anomalies Description:

(1) Absolute variations higher than 15% between the element weight (n) and it's moving average;

(2) It's not possible to calculate the moving average of the element;

(3) The element weight for the system is 0 and it is greater than 5% for the observed agent;

(4) Absolute variations higher than 30% between the element weight (n) for the observed agent and for the system.

Figure 5 Dashboard view of the Concentration Check

In Figure 5, on the right hand side, we have a set of filters that allow us to select both the observed agent (103060, in this case) and the type of instrument we want to focus on (0110, housing credit). In the case of this particular agent, we can see that for the displayed variables, the reporting is quite stable from either perspective:

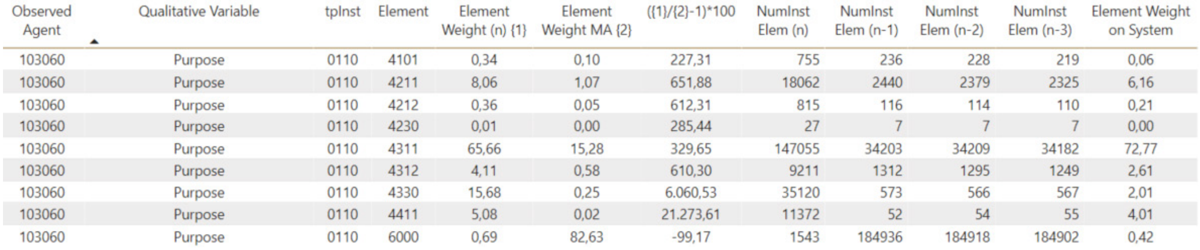
- When looking at the way the agent's reporting evolves over time, by comparing the "Element Weight (n) {1}"[3] and "Element Weight MA {2}"[4] columns or looking at the variation (as displayed in the column labeled $\left(\frac{\{1\}}{\{2\}} - 1\right) * 100$), we find no significant changes;
- When looking at how the agent's reporting compares to the system as a whole, by comparing the "Element Weight (n)" and "Element Weight on System" columns, we see that the agent conforms to system trends.

The small number of potentially anomalous situations detected by this test, while also related to the fact that it deals with aggregate values (which in general are far more stable than individual records), is fundamentally a result of the arduous work carried out by the analysts at the Banco de Portugal CCR. Their work and subsequent interactions with the observed agents has resulted in numerous corrections since the test has been in production, which have improved the quality of the reported data very significantly.

In Figure 6 we can see an example of such a correction, concerning the agent/instrument pairing 103060/0110, which was corrected in August 2020.

[3] "Element Weight (n)" is the weight of the element on the reporting of the corresponding categorical variable, for a given observed agent, in the most recent reporting period. It is computed by dividing the number of contracts reported with the element in question (column "NumInst Elem (n)") by the total number of contracts reported by the agent for the type of instrument being considered.

[4] "Element Weight MA" is a weighted average of the three months prior to the current month, where the weights are 3/6 for period (n-1), 2/6 for period (n-2) and 1/6 for period (n-3). It is computed in the same way as "Element Weight (n)".

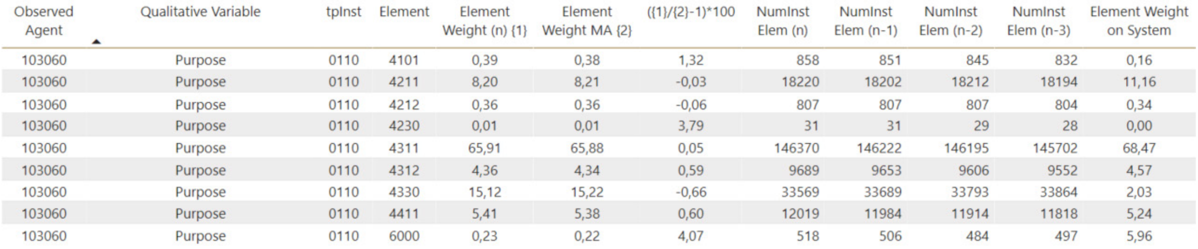


Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Purpose	0110	4101	0,34	0,10	227,31	755	236	228	219	0,06
103060	Purpose	0110	4211	8,06	1,07	651,88	18062	2440	2379	2325	6,16
103060	Purpose	0110	4212	0,36	0,05	612,31	815	116	114	110	0,21
103060	Purpose	0110	4230	0,01	0,00	285,44	27	7	7	7	0,00
103060	Purpose	0110	4311	65,66	15,28	329,65	147055	34203	34209	34182	72,77
103060	Purpose	0110	4312	4,11	0,58	610,30	9211	1312	1295	1249	2,61
103060	Purpose	0110	4330	15,68	0,25	6.060,53	35120	573	566	567	2,01
103060	Purpose	0110	4411	5,08	0,02	21.273,61	11372	52	54	55	4,01
103060	Purpose	0110	6000	0,69	82,63	-99,17	1543	184936	184918	184902	0,42

Figure 6 Concentration Check – purpose of loan (August 2020)

As we can observe, up to July 2020, this agent presented a concentration of over 80% in the element "6000 - Other purposes" for the variable "Purpose" (seen in column "Element Weight MA"), which is meant to be a residual element. This was corrected in the following month of August 2020, as can be seen in the "Element Weight (n)" column, and the instruments that presented this element were mostly reallocated across the remaining available elements, in particular many were moved to element "4311 - Residential real estate purchase – permanent residential property". The element weights displayed as of August 2020 for this variable also fell into line with the behaviour displayed by the system as a whole, which was not the case in July, as should be expected in general. Nevertheless, this correction triggered multiple type 1 anomalies alerting the analyst for the significant changes that took place. If we had instead focused on the picture as of July 2020, the very high concentration in the element "6000 - Other purposes", when compared with the system weights, would give rise instead to a type 4 anomaly.

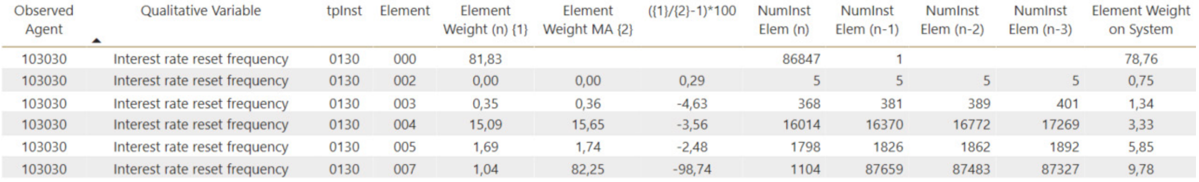
After this correction the situation has remained quite stable, as can be seen in Figure 7 showing August 2021:



Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Purpose	0110	4101	0,39	0,38	1,32	858	851	845	832	0,16
103060	Purpose	0110	4211	8,20	8,21	-0,03	18220	18202	18212	18194	11,16
103060	Purpose	0110	4212	0,36	0,36	-0,06	807	807	807	804	0,34
103060	Purpose	0110	4230	0,01	0,01	3,79	31	31	29	28	0,00
103060	Purpose	0110	4311	65,91	65,88	0,05	146370	146222	146195	145702	68,47
103060	Purpose	0110	4312	4,36	4,34	0,59	9689	9653	9606	9552	4,57
103060	Purpose	0110	4330	15,12	15,22	-0,66	33569	33689	33793	33864	2,03
103060	Purpose	0110	4411	5,41	5,38	0,60	12019	11984	11914	11818	5,24
103060	Purpose	0110	6000	0,23	0,22	4,07	518	506	484	497	5,96

Figure 7 Concentration Check - purpose of loan (August 2021)

As a final example, in Figure 8, we present a recent case where a correction took place for the agent/type of instrument pair 103030/0130 (consumer credit), for the variable "Interest rate reset frequency".



Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103030	Interest rate reset frequency	0130	000	81,83			86847	1			78,76
103030	Interest rate reset frequency	0130	002	0,00	0,00	0,29	5	5	5	5	0,75
103030	Interest rate reset frequency	0130	003	0,35	0,36	-4,63	368	381	389	401	1,34
103030	Interest rate reset frequency	0130	004	15,09	15,65	-3,56	16014	16370	16772	17269	3,33
103030	Interest rate reset frequency	0130	005	1,69	1,74	-2,48	1798	1826	1862	1892	5,85
103030	Interest rate reset frequency	0130	007	1,04	82,25	-98,74	1104	87659	87483	87327	9,78

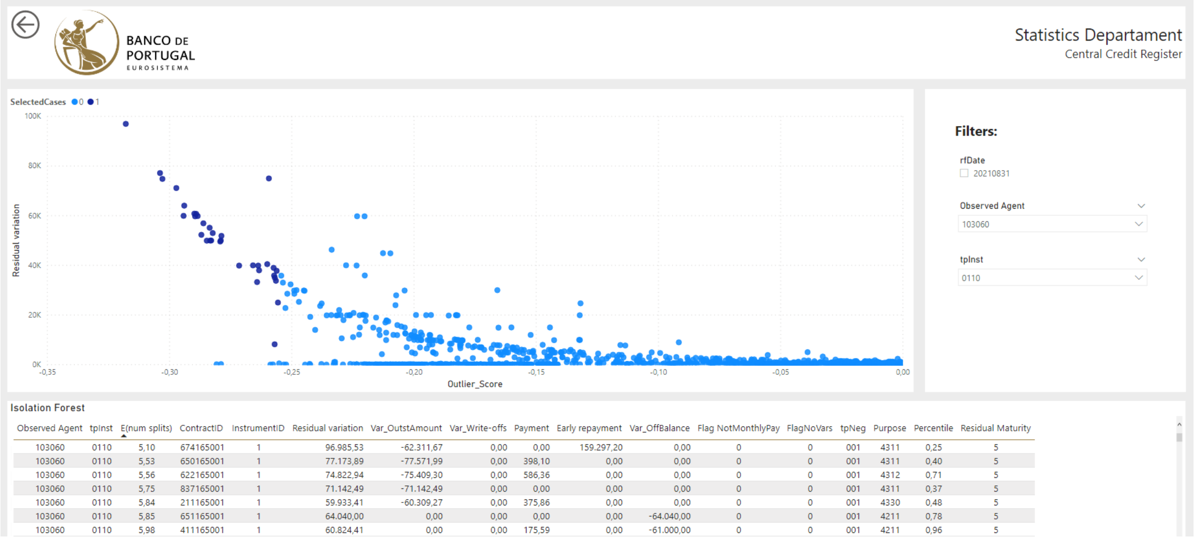
Figure 8 Concentration Check – interest rate reset frequency (August 2021)

This situation is quite similar to the previous one. The agent was reporting over 80% of instruments with element "007 – Other frequencies" (as can be seen in the "Element Weight MA" column). After further investigation, we concluded that this issue was the result of a mapping error for the instruments with fixed interest rate. After contacting the reporting agent, a correction took place and the instruments that were previously classified with element "007" were reclassified as "000 - Rate cannot be reset" which is now the element with the most significant weight, in line with the system weights.

Isolation Forest

As was stated in the previous section, IF is a scoring algorithm, meaning that it provides us with a metric of how likely it is for a given observation to be an anomaly. This means we can easily provide the analysts with an ordered list of potential anomalies, according to the score attributed by the model.

As with the previous tests, the results are provided to the analysts using a Power BI dashboard, with a visual and an accompanying table, as displayed in Figure 9:



Statistics Departament
Central Credit Register

Filters:

rfDate
□ 20210831

Observed Agent
103060

tpInst
0110

Isolation Forest

Observed Agent	tpInst	E(num splits)	ContractID	InstrumentID	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance	Flag NotMonthlyPay	FlagNoVars	tpNeg	Purpose	Percentile	Residual Maturity
103060	0110	5,10	674165001	1	96.985,53	-62.311,67	0,00	0,00	159.297,20	0,00	0	0	001	4311	0,25	5
103060	0110	5,53	650165001	1	77.173,89	-77.571,99	0,00	398,10	0,00	0,00	0	0	001	4311	0,40	5
103060	0110	5,56	622165001	1	74.822,94	-75.409,30	0,00	586,36	0,00	0,00	0	0	001	4312	0,71	5
103060	0110	5,75	837165001	1	71.142,49	-71.142,49	0,00	0,00	0,00	0,00	0	0	001	4311	0,37	5
103060	0110	5,84	211165001	1	59.933,41	-60.309,27	0,00	375,86	0,00	0,00	0	0	001	4330	0,48	5
103060	0110	5,85	651165001	1	64.040,00	0,00	0,00	0,00	0,00	-64.040,00	0	0	001	4211	0,78	5
103060	0110	5,98	411165001	1	60.824,41	0,00	0,00	175,59	0,00	-61.000,00	0	0	001	4211	0,96	5

Figure 9 Dashboard view of the Isolation Forest results

In the accompanying table, along with the model variables ("residual variation", "purpose of the loan", "type of negotiation" and "residual maturity"), we also present a set of other variables to provide further context that could help the analysts determine if a given observation is, in fact, an anomaly:

- Identifiers of the observed agent and of the contract/instrument ("Observed Agent", "ContractID", "InstrumentID");
- Type of instrument ("tpInst");
- Expected value of the number of splits needed to isolate the data point in the IF ("E(num splits)");
- Variations of the amounts that, when combined, result in the value of the "residual variation";
- Flag that indicates if the instrument has a payment frequency which is not monthly ("Flag NotMonthlyPay");
- Flag that indicates if the instrument did not experience variations in any of the reported amounts (excluding the payment) in the current month ("FlagNoVars");
- Percentile of the outstanding nominal amount, considering all instruments reported for a given type of instrument, for a given observed agent, in the reference period.

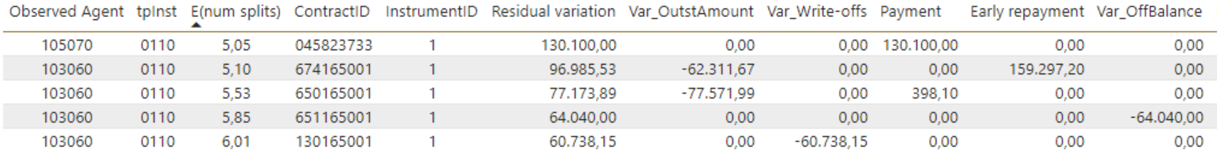
In the visual, we show the distribution of the residual variations related to the outlier score, where the lower the score, the greater the probability of the observation being an anomaly.

All the observations that the test detects (regardless of anomaly score) are displayed in the visual. In dark blue, we differentiate the observations that should be prioritised when questioning the observed agents. An instrument is selected as being a priority if it meets all of the following criteria:

- Outlier score belongs to the 1st percentile of outlier scores, within the corresponding type of instrument/observed agent combination;
- Outlier score less than or equal to -0.25;
- Residual variation greater than or equal to €1,000.

As stated above, multiple situations can generate significant outlier scores, either due to the high value of the residual variation or due to the uncommon combination of the three categorical variables.

In Figure 10 we present a few examples where the residual variation is clearly the factor that determines the small "E(num splits)"[5]. The descriptions are presented in the order they are shown in the table:



Observed Agent	tpInst	E(num splits)	ContractID	InstrumentID	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance
105070	0110	5,05	045823733	1	130.100,00	0,00	0,00	130.100,00	0,00	0,00
103060	0110	5,10	674165001	1	96.985,53	-62.311,67	0,00	0,00	159.297,20	0,00
103060	0110	5,53	650165001	1	77.173,89	-77.571,99	0,00	398,10	0,00	0,00
103060	0110	5,85	651165001	1	64.040,00	0,00	0,00	0,00	0,00	-64.040,00
103060	0110	6,01	130165001	1	60.738,15	0,00	-60.738,15	0,00	0,00	0,00

Figure 10 Isolation Forest – quantitative variables impact

- Monthly payment is reported but the outstanding amounts do not decrease;
- Early payment is reported but the outstanding amounts do not decrease accordingly;
- The monthly payment significantly differs from the variation observed in the outstanding amounts;
- The off-balance sheet amount decreases but the outstanding amount does not increase;
- Decrease in the amount allocated to write-offs with no increase in the outstanding amounts.

In Figure 11, we highlight two data points that show that despite the fact that the residual variation seems to be the determining factor for the value of the outlier score, the categorical variables also have a significant effect. The highlighted data points have a significant outlier score, despite the fact that they have residual variations that are not very high.

[5] Expected value of the number of splits required to isolate a given observation. This value conveys the exact same information as the outlier score, i.e., the lower the number of splits required to isolate a data point, the higher the likelihood of the point being an anomaly.

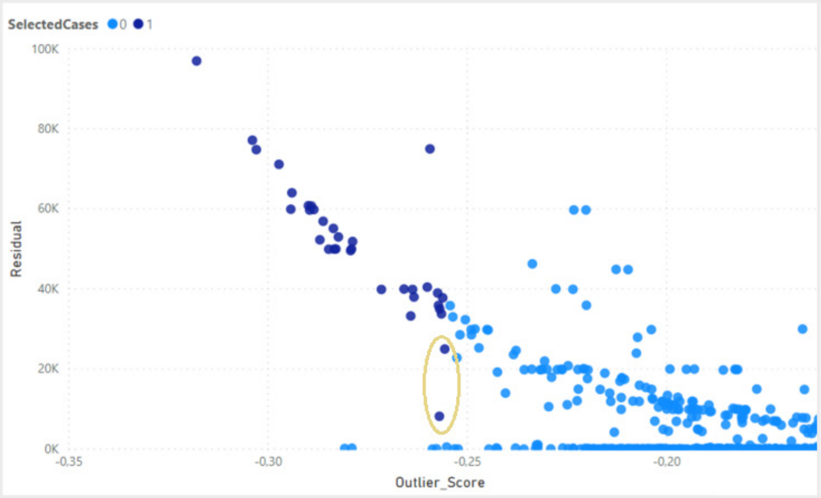
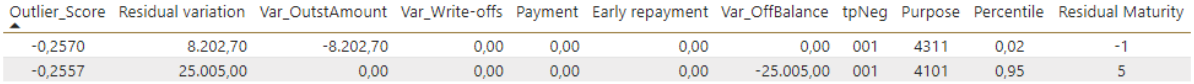


Figure 11 Isolation Forest – categorical variables impact (visual)

Furthermore, we can see that the data point with the lower "residual variation" actually has a more significant outlier score. If we analyse both data points in detail (Figure 12), we can see that this is because the data point with a higher outlier score has a "residual maturity" of -1 (passed maturity), which is likely an unusual element for this agent/instrument pairing, whereas the element 5 (Over 20 years) is far more common.



Outlier_Score	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance	tpNeg	Purpose	Percentile	Residual Maturity
-0,2570	8.202,70	-8.202,70	0,00	0,00	0,00	0,00	001	4311	0,02	-1
-0,2557	25.005,00	0,00	0,00	0,00	0,00	-25.005,00	001	4101	0,95	5

Figure 12 Isolation Forest – categorical variables impact (table)

5. Conclusion and Final Remarks

As we transited in September 2018 from a debtor-by-debtor logic to an instrument-by-instrument one with the new CCR, we experienced a huge increase in the volume of information to be processed and analysed. In addition, the new approach of a single service desk represented a widening in the set of variables related with the credit concession, with the addition of multiple new attributes associated.

This evolution represented a new challenge and it became of critical importance to find new ways of looking at data efficiently to detect and control for a multitude of anomalies arising from either the evolution or the interaction of the variables reported. This was the context that motivated the development of the three new tests we have presented: the reporting consistency test, concentration check and IF.

After a full year of usage in production, on a monthly basis, the new tests developed have shown to be very useful and we have received very encouraging feedback. They represent a valuable addition to the quality control process, focusing on dynamics that complement the other existing processes, allowing for the identification of a set of anomalies that previously would not be detected or would require complex and time-consuming *ad hoc* analyses. Examples of this kind of abnormal evolutions include:

- The detection of reporting gaps and strange patterns, even subtle ones that only affect a few instruments;
- Oversee the evolution of the reporting of categorical variables and to detect structural changes;
- Monitor the evolution of the amounts reported for the instrument, taking into account the categorical variables that characterize them and ranking them by the degree of severity for further questioning.

Hence, these new tools have contributed unquestionably to an increase in both the effectiveness and efficiency of the data quality assessment process and are an important enhancement to the analysts' tool set.

References

Liu, F. T., Ting, K., & Zhou, Z.-H. (2009). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, (pp. 413 - 422).

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.

BANCO DE PORTUGAL
EUROSISTEMA

Anomaly Detection in the Portuguese Central Credit Register (CCR)

André Costa

Francisco Fonseca

22 October 2021

Introduction

❑ The Portuguese CCR is a system which gathers on a monthly basis information associated with actual and potential credit liabilities of natural and legal persons.

❑ The main purpose is to provide support to the credit institutions in their assessment of counterparty credit risk.

❑ The data reported to the CCR, which is rich in both volume and complexity, is used for a multitude of different purposes – compiling statistics, banking supervision, financial stability analysis, support on monetary policy decisions…

18.1 million
credit instruments

5.2 million
protections

7 million
natural persons

344 thousand
legal persons

182
observed agents

50€
minimum amount

31
instrument types

3.3 thousand
Files received per month

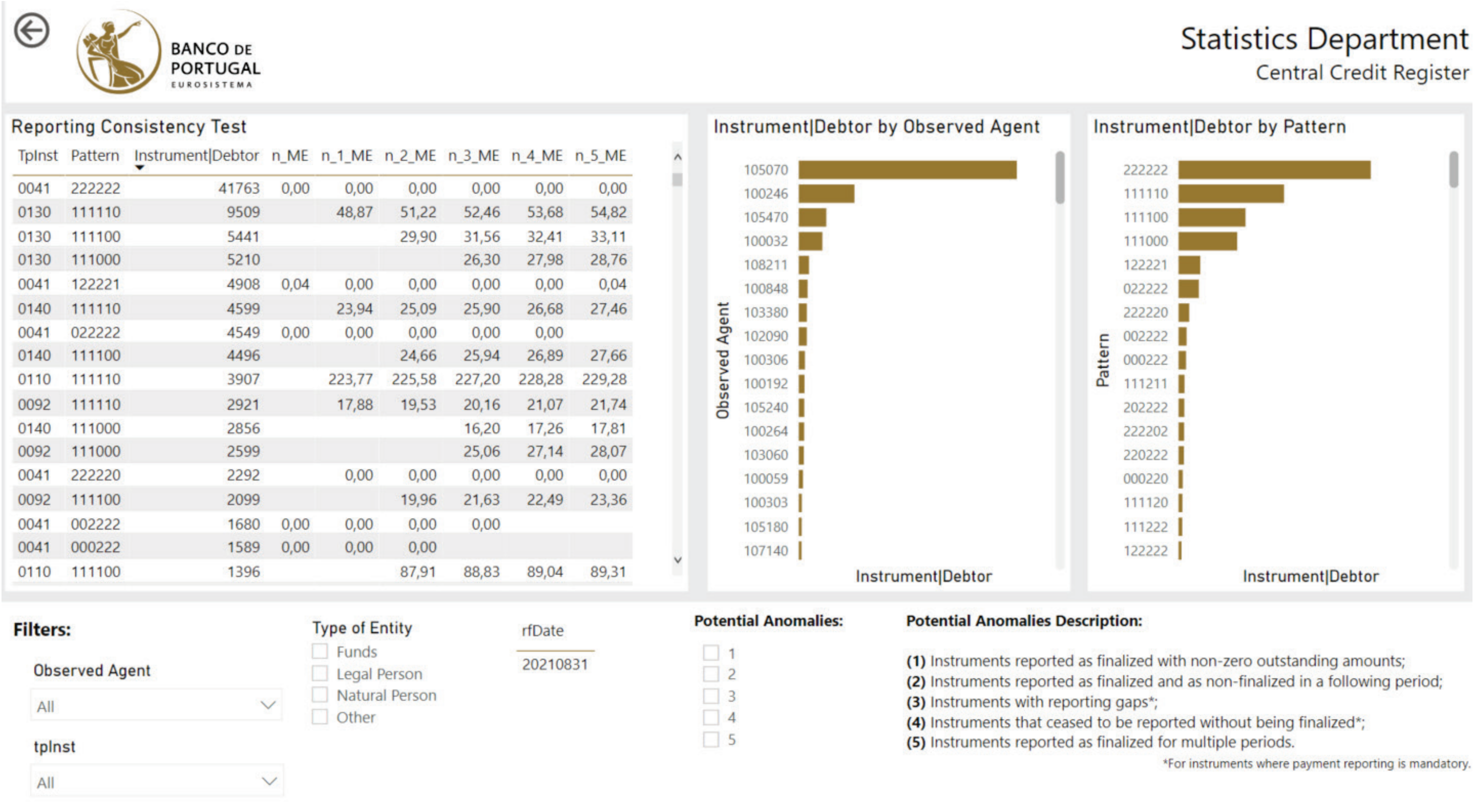
❑ It's quite challenging to perform the quality control of a database with such a level of granularity and variety of attributes (over 200). This was the main motivation that lead to the development of these new tests: increase the efficiency of the data quality controls, detect more subtle evolutions, create automatic filters for a range of potential anomalies and ensure the coherence of the process across the different observed agents.

Reporting Consistency Test

- ❑ The goal is to evaluate if all instruments are reported consistently until maturity;

- ❑ A pattern is created with the last six months of information reported (the rightmost digit is the most recent period);

- ❑ Pattern possible values for each month:
  - ➢ 0 – the instrument/entity was not reported;
  - ➢ 1 - the instrument/entity is active;
  - ➢ 2 – the instrument/entity was finalized.[1]



BANCO DE PORTUGAL
EUROSISTEMA

Statistics Department
Central Credit Register

Reporting Consistency Test

TpInst	Pattern	Instrument|Debtor	n_ME	n_1_ME	n_2_ME	n_3_ME	n_4_ME	n_5_ME
0041	222222	41763	0,00	0,00	0,00	0,00	0,00	0,00
0130	111110	9509		48,87	51,22	52,46	53,68	54,82
0130	111100	5441			29,90	31,56	32,41	33,11
0130	111000	5210				26,30	27,98	28,76
0041	122221	4908	0,04	0,00	0,00	0,00	0,00	0,04
0140	111110	4599		23,94	25,09	25,90	26,68	27,46
0041	022222	4549	0,00	0,00	0,00	0,00	0,00	
0140	111100	4496			24,66	25,94	26,89	27,66
0110	111110	3907		223,77	225,58	227,20	228,28	229,28
0092	111110	2921		17,88	19,53	20,16	21,07	21,74
0140	111000	2856				16,20	17,26	17,81
0092	111000	2599				25,06	27,14	28,07
0041	222220	2292		0,00	0,00	0,00	0,00	0,00
0092	111100	2099			19,96	21,63	22,49	23,36
0041	002222	1680	0,00	0,00	0,00	0,00		
0041	000222	1589	0,00	0,00	0,00			
0110	111100	1396			87,91	88,83	89,04	89,31

Filters:

Observed Agent
All

tplnst
All

Type of Entity
- □ Funds
- □ Legal Person
- □ Natural Person
- □ Other

rfDate
20210831

Potential Anomalies:
- □ 1
- □ 2
- □ 3
- □ 4
- □ 5

Potential Anomalies Description:

(1) Instruments reported as finalized with non-zero outstanding amounts;
(2) Instruments reported as finalized and as non-finalized in a following period;
(3) Instruments with reporting gaps*;
(4) Instruments that ceased to be reported without being finalized*;
(5) Instruments reported as finalized for multiple periods.

*For instruments where payment reporting is mandatory.

[1] Finalized instruments should be reported only once and with all amounts set to 0.

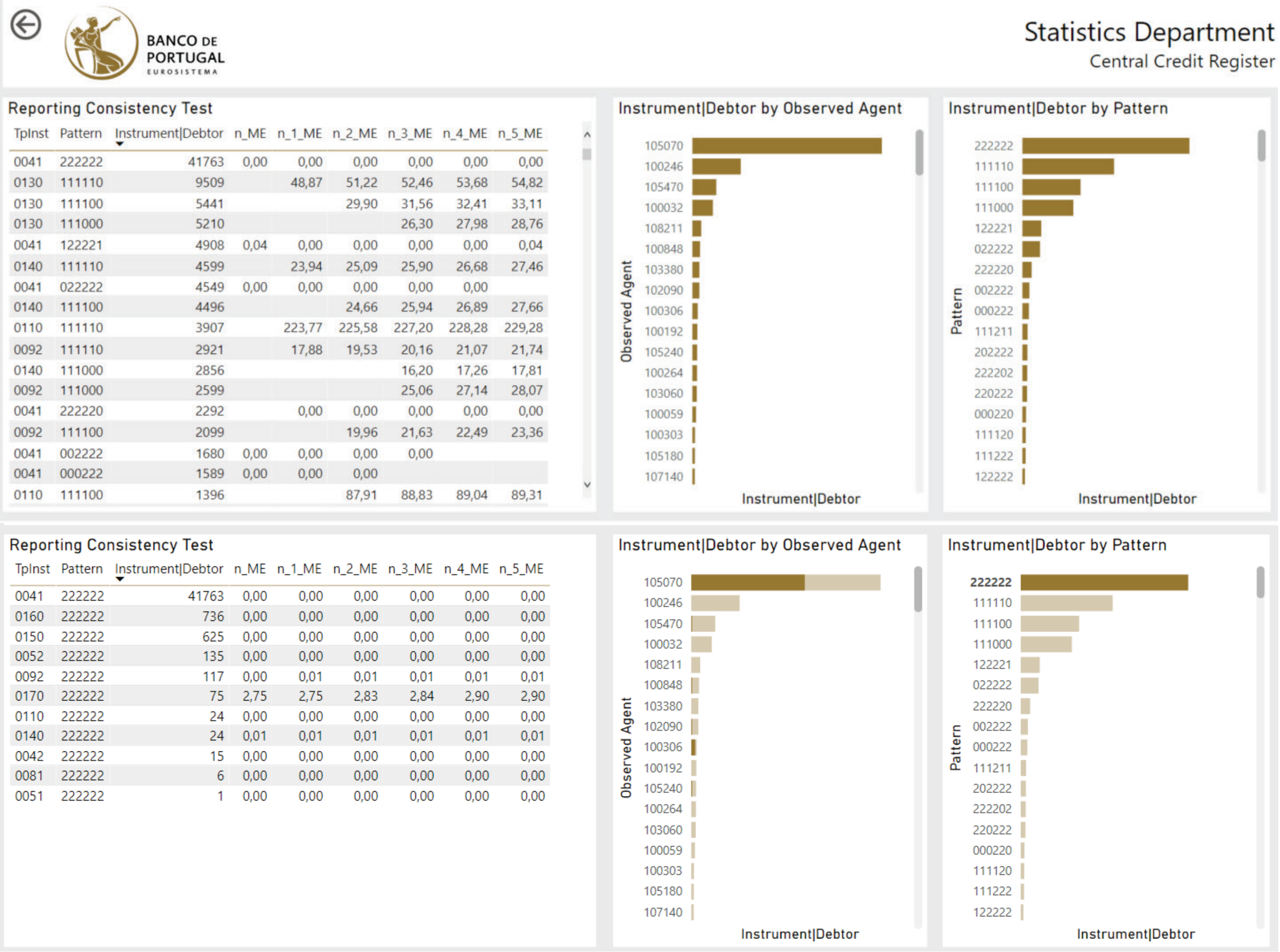
Reporting Consistency Test

- ❑ Only instruments with potentially anomalous patterns are displayed;

- ❑ The amounts associated with each instrument|debtor are presented to allow for measuring the overall impact of potential anomalies;

- ❑ Automatic filters were defined to allow the detection of potential anomalies such as:

  - ➢ Reporting gaps;
  - ➢ Lack of / inconsistent finalization.



BANCO DE PORTUGAL EUROSISTEMA

Statistics Department
Central Credit Register

Reporting Consistency Test

TpInst	Pattern	Instrument|Debtor	n_ME	n_1_ME	n_2_ME	n_3_ME	n_4_ME	n_5_ME
0041	222222	41763	0,00	0,00	0,00	0,00	0,00	0,00
0130	111110	9509		48,87	51,22	52,46	53,68	54,82
0130	111100	5441			29,90	31,56	32,41	33,11
0130	111000	5210				26,30	27,98	28,76
0041	122221	4908	0,04	0,00	0,00	0,00	0,00	0,04
0140	111110	4599		23,94	25,09	25,90	26,68	27,46
0041	022222	4549	0,00	0,00	0,00	0,00	0,00	
0140	111100	4496			24,66	25,94	26,89	27,66
0110	111110	3907		223,77	225,58	227,20	228,28	229,28
0092	111110	2921		17,88	19,53	20,16	21,07	21,74
0140	111000	2856				16,20	17,26	17,81
0092	111000	2599				25,06	27,14	28,07
0041	222220	2292		0,00	0,00	0,00	0,00	0,00
0092	111100	2099			19,96	21,63	22,49	23,36
0041	002222	1680	0,00	0,00	0,00	0,00		
0041	000222	1589	0,00	0,00	0,00			
0110	111100	1396			87,91	88,83	89,04	89,31

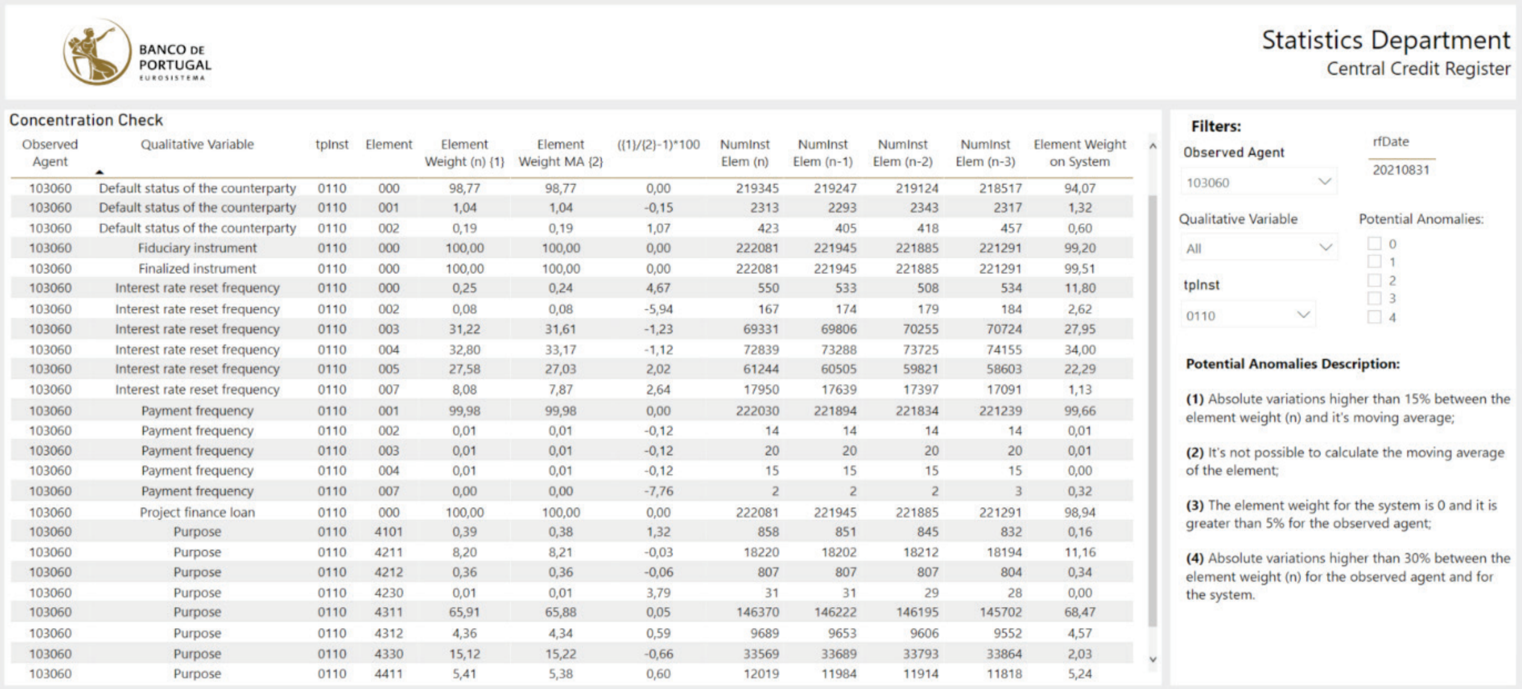
Reporting Consistency Test

TpInst	Pattern	Instrument|Debtor	n_ME	n_1_ME	n_2_ME	n_3_ME	n_4_ME	n_5_ME
0041	222222	41763	0,00	0,00	0,00	0,00	0,00	0,00
0160	222222	736	0,00	0,00	0,00	0,00	0,00	0,00
0150	222222	625	0,00	0,00	0,00	0,00	0,00	0,00
0052	222222	135	0,00	0,00	0,00	0,00	0,00	0,00
0092	222222	117	0,00	0,01	0,01	0,01	0,01	0,01
0170	222222	75	2,75	2,75	2,83	2,84	2,90	2,90
0110	222222	24	0,00	0,00	0,00	0,00	0,00	0,00
0140	222222	24	0,01	0,01	0,01	0,01	0,01	0,01
0042	222222	15	0,00	0,00	0,00	0,00	0,00	0,00
0081	222222	6	0,00	0,00	0,00	0,00	0,00	0,00
0051	222222	1	0,00	0,00	0,00	0,00	0,00	0,00

Concentration Check

- ❑ The goal of this test is to evaluate the reporting of categorical variables;

- ❑ For each element we display:
  - ➢ The weight in the most recent period;
  - ➢ The weighted average of the previous three months;
  - ➢ The number of instruments for the last four months;
  - ➢ The average for the entire system in the most recent period.



BANCO DE PORTUGAL
EUROSISTEMA

Statistics Department
Central Credit Register

Concentration Check

Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Default status of the counterparty	0110	000	98,77	98,77	0,00	219345	219247	219124	218517	94,07
103060	Default status of the counterparty	0110	001	1,04	1,04	-0,15	2313	2293	2343	2317	1,32
103060	Default status of the counterparty	0110	002	0,19	0,19	1,07	423	405	418	457	0,60
103060	Fiduciary instrument	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	99,20
103060	Finalized instrument	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	99,51
103060	Interest rate reset frequency	0110	000	0,25	0,24	4,67	550	533	508	534	11,80
103060	Interest rate reset frequency	0110	002	0,08	0,08	-5,94	167	174	179	184	2,62
103060	Interest rate reset frequency	0110	003	31,22	31,61	-1,23	69331	69806	70255	70724	27,95
103060	Interest rate reset frequency	0110	004	32,80	33,17	-1,12	72839	73288	73725	74155	34,00
103060	Interest rate reset frequency	0110	005	27,58	27,03	2,02	61244	60505	59821	58603	22,29
103060	Interest rate reset frequency	0110	007	8,08	7,87	2,64	17950	17639	17397	17091	1,13
103060	Payment frequency	0110	001	99,98	99,98	0,00	222030	221894	221834	221239	99,66
103060	Payment frequency	0110	002	0,01	0,01	-0,12	14	14	14	14	0,01
103060	Payment frequency	0110	003	0,01	0,01	-0,12	20	20	20	20	0,01
103060	Payment frequency	0110	004	0,01	0,01	-0,12	15	15	15	15	0,00
103060	Payment frequency	0110	007	0,00	0,00	-7,76	2	2	2	3	0,32
103060	Project finance loan	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	98,94
103060	Purpose	0110	4101	0,39	0,38	1,32	858	851	845	832	0,16
103060	Purpose	0110	4211	8,20	8,21	-0,03	18220	18202	18212	18194	11,16
103060	Purpose	0110	4212	0,36	0,36	-0,06	807	807	807	804	0,34
103060	Purpose	0110	4230	0,01	0,01	3,79	31	31	29	28	0,00
103060	Purpose	0110	4311	65,91	65,88	0,05	146370	146222	146195	145702	68,47
103060	Purpose	0110	4312	4,36	4,34	0,59	9689	9653	9606	9552	4,57
103060	Purpose	0110	4330	15,12	15,22	-0,66	33569	33689	33793	33864	2,03
103060	Purpose	0110	4411	5,41	5,38	0,60	12019	11984	11914	11818	5,24

Filters:

Observed Agent: 103060

rfDate: 20210831

Qualitative Variable: All

Potential Anomalies:
□ 0
□ 1
□ 2
□ 3
□ 4

tpInst: 0110

Potential Anomalies Description:

(1) Absolute variations higher than 15% between the element weight (n) and it's moving average;

(2) It's not possible to calculate the moving average of the element;

(3) The element weight for the system is 0 and it is greater than 5% for the observed agent;

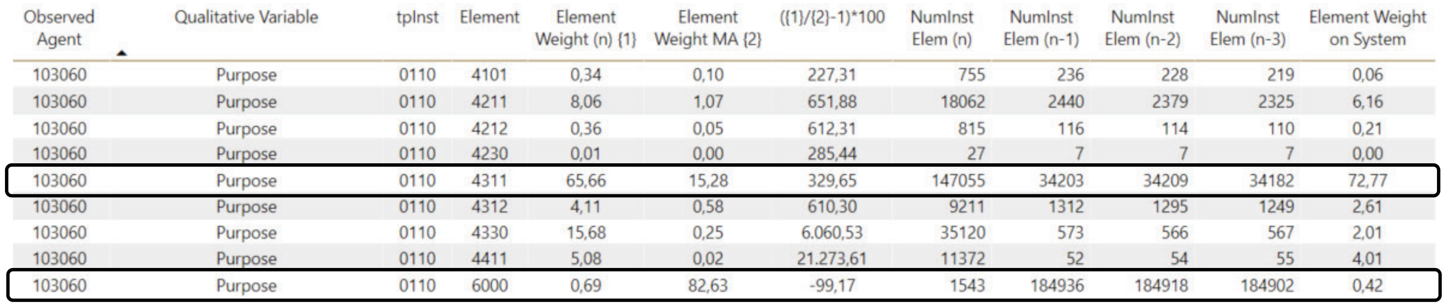
(4) Absolute variations higher than 30% between the element weight (n) for the observed agent and for the system.

- ❑ The purpose of the test is two-fold:
  - ➢ Checking the reporting consistency at the observed agent level;
  - ➢ Checking if there are significant differences in the reporting between the observed agent and the system as a whole.

Concentration Check - an example

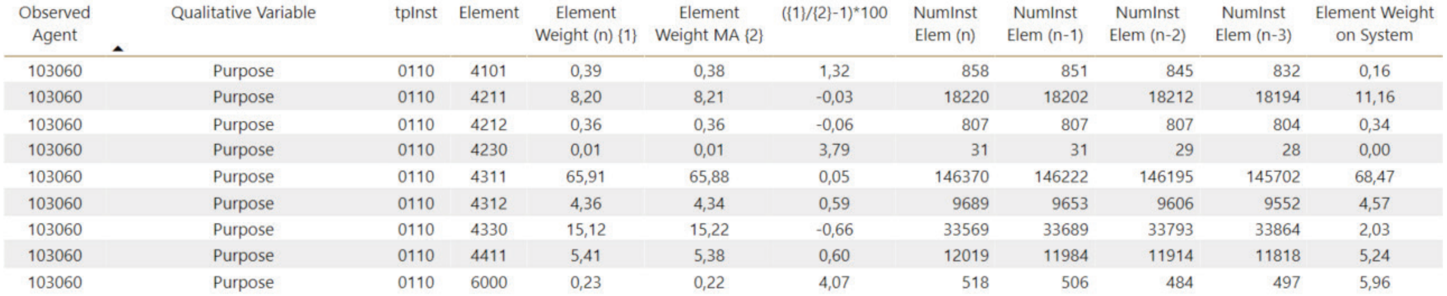
- ❑ Up to July 2020:
  - ➢ Over 80% in the element “Other purposes” (6000) for housing credit (0110), usually a residual element.
- ❑ In August 2020:
  - ➢ The majority of the instruments were reallocated across the remaining elements, in particular to “Residential real estate purchase – permanent residential property” (4311).
  - ➢ The new weights for the observed agent are now in line with the system.
- ❑ In August 2021:
  - ➢ After this correction the reporting has been quite stable.

August 2020



Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Purpose	0110	4101	0,34	0,10	227,31	755	236	228	219	0,06
103060	Purpose	0110	4211	8,06	1,07	651,88	18062	2440	2379	2325	6,16
103060	Purpose	0110	4212	0,36	0,05	612,31	815	116	114	110	0,21
103060	Purpose	0110	4230	0,01	0,00	285,44	27	7	7	7	0,00
103060	Purpose	0110	4311	65,66	15,28	329,65	147055	34203	34209	34182	72,77
103060	Purpose	0110	4312	4,11	0,58	610,30	9211	1312	1295	1249	2,61
103060	Purpose	0110	4330	15,68	0,25	6.060,53	35120	573	566	567	2,01
103060	Purpose	0110	4411	5,08	0,02	21.273,61	11372	52	54	55	4,01
103060	Purpose	0110	6000	0,69	82,63	-99,17	1543	184936	184918	184902	0,42

August 2021



Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Purpose	0110	4101	0,39	0,38	1,32	858	851	845	832	0,16
103060	Purpose	0110	4211	8,20	8,21	-0,03	18220	18202	18212	18194	11,16
103060	Purpose	0110	4212	0,36	0,36	-0,06	807	807	807	804	0,34
103060	Purpose	0110	4230	0,01	0,01	3,79	31	31	29	28	0,00
103060	Purpose	0110	4311	65,91	65,88	0,05	146370	146222	146195	145702	68,47
103060	Purpose	0110	4312	4,36	4,34	0,59	9689	9653	9606	9552	4,57
103060	Purpose	0110	4330	15,12	15,22	-0,66	33569	33689	33793	33864	2,03
103060	Purpose	0110	4411	5,41	5,38	0,60	12019	11984	11914	11818	5,24
103060	Purpose	0110	6000	0,23	0,22	4,07	518	506	484	497	5,96

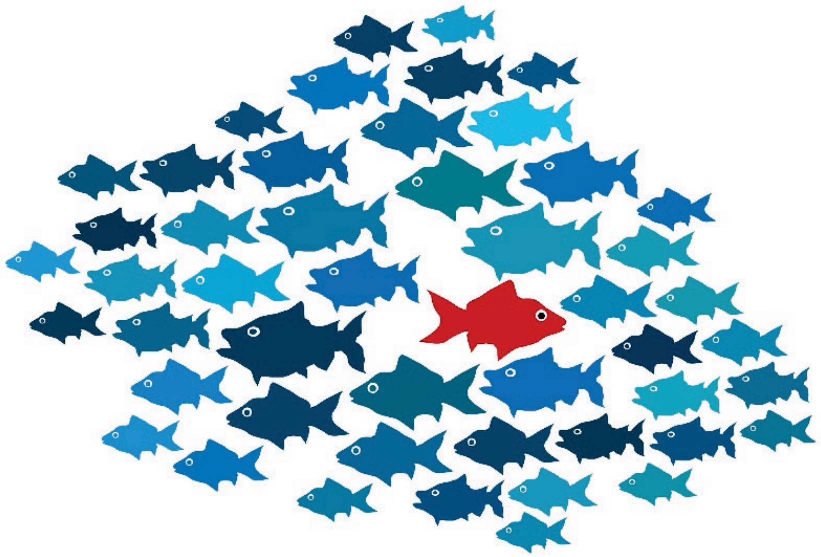
Isolation Forest

- ❑ The Isolation forest is an unsupervised learning algorithm that works on the principle of isolating anomalies;

- ❑ Main advantages:

  - ➢ It has quasilinear complexity which makes it viable to integrate it in our daily routines;

  - ➢ It is a scoring algorithm allowing us to filter the most probable outliers when we are analyzing the results.



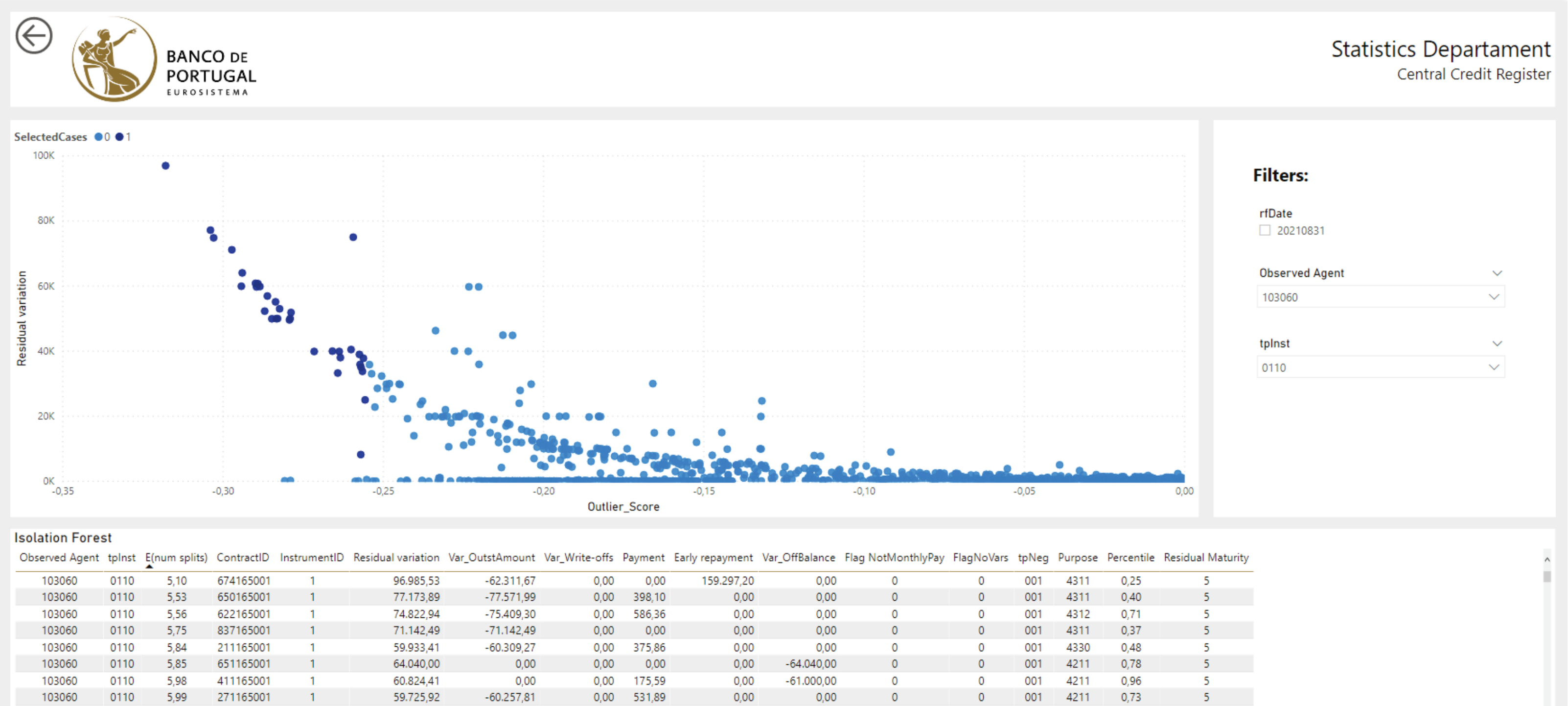
- ❑ The observations that require the least number of steps to be isolated will have a greater probability of being anomalies;

- ❑ The isolation forest is built using information from the four months prior to the most recent period to generate the training set.

Isolation Forest

- ❑ The goal is to detect abnormal evolutions of the outstanding amounts and associated qualitative variables;

- ❑ Model specification:
  - ➢ Residual variation[1];
  - ➢ Purpose of the loan;
  - ➢ Type of negotiation;
  - ➢ Residual maturity.



BANCO DE PORTUGAL
EUROSISTEMA

Statistics Department
Central Credit Register

Filters:

rfDate
□ 20210831

Observed Agent
103060

tpInst
0110

Isolation Forest

Observed Agent	tpInst	E(num splits)	ContractID	InstrumentID	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance	Flag NotMonthlyPay	FlagNoVars	tpNeg	Purpose	Percentile	Residual Maturity
103060	0110	5,10	674165001	1	96.985,53	-62.311,67	0,00	0,00	159.297,20	0,00	0	0	001	4311	0,25	5
103060	0110	5,53	650165001	1	77.173,89	-77.571,99	0,00	398,10	0,00	0,00	0	0	001	4311	0,40	5
103060	0110	5,56	622165001	1	74.822,94	-75.409,30	0,00	586,36	0,00	0,00	0	0	001	4312	0,71	5
103060	0110	5,75	837165001	1	71.142,49	-71.142,49	0,00	0,00	0,00	0,00	0	0	001	4311	0,37	5
103060	0110	5,84	211165001	1	59.933,41	-60.309,27	0,00	375,86	0,00	0,00	0	0	001	4330	0,48	5
103060	0110	5,85	651165001	1	64.040,00	0,00	0,00	0,00	0,00	-64.040,00	0	0	001	4211	0,78	5
103060	0110	5,98	411165001	1	60.824,41	0,00	0,00	175,59	0,00	-61.000,00	0	0	001	4211	0,96	5
103060	0110	5,99	271165001	1	59.725,92	-60.257,81	0,00	531,89	0,00	0,00	0	0	001	4211	0,73	5

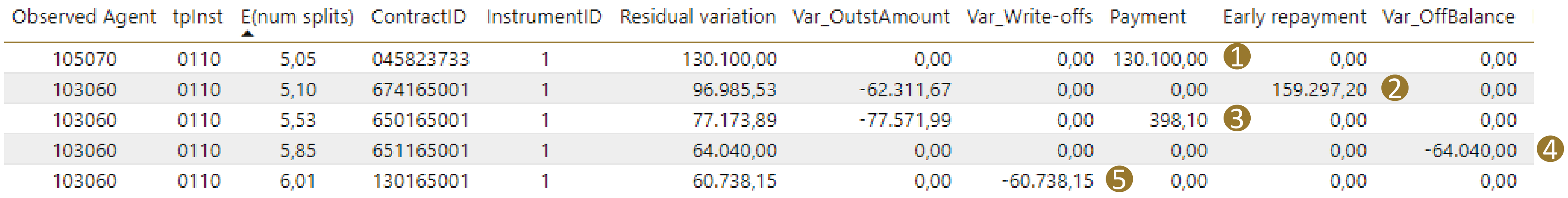
- ❑ The criteria to select the observations in dark blue consisted in three cumulative conditions:

Be from the 1% more negative outlier scores | Have an outlier score of -0.25 or less | Have a residual of 1,000€ or higher.

[1] Residual Variation = $Abs(\Delta\text{Outstanding nominal amount} + \Delta\text{Accumulated write-offs} + \text{Payment} + \Delta\text{Off-balance-sheet amount} + \text{Early repayment})$

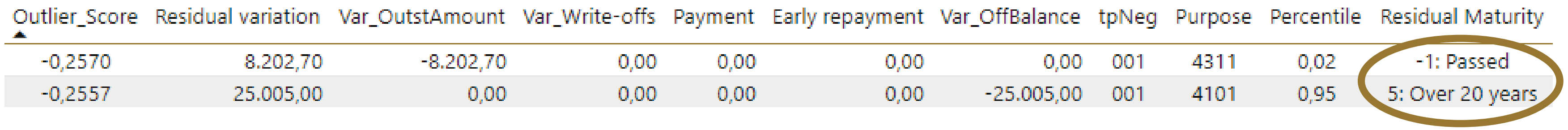
Isolation Forest

- ❑ Multiple situations can generate significant outlier scores, either due to the residual variation:
  - ➢ Monthly payment without a decrease of the outstanding amounts (1) or they are significantly different (3);
  - ➢ Early repayment without a similar decrease in the outstanding amounts (2);
  - ➢ The off-balance sheet amount decreases without an increase of the outstanding amounts (4);
  - ➢ Decrease in write-offs, with no increase in the outstanding amounts (5).



Observed Agent	tpInst	E(num splits)	ContractID	InstrumentID	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance
105070	0110	5,05	045823733	1	130.100,00	0,00	0,00	130.100,00 ❶	0,00	0,00
103060	0110	5,10	674165001	1	96.985,53	-62.311,67	0,00	0,00	159.297,20 ❷	0,00
103060	0110	5,53	650165001	1	77.173,89	-77.571,99	0,00	398,10 ❸	0,00	0,00
103060	0110	5,85	651165001	1	64.040,00	0,00	0,00	0,00	0,00	-64.040,00 ❹
103060	0110	6,01	130165001	1	60.738,15	0,00	-60.738,15 ❺	0,00	0,00	0,00

- ❑ Or due to the uncommon combination of the three qualitative variables



Outlier_Score	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance	tpNeg	Purpose	Percentile	Residual Maturity
-0,2570	8.202,70	-8.202,70	0,00	0,00	0,00	0,00	001	4311	0,02	-1: Passed
-0,2557	25.005,00	0,00	0,00	0,00	0,00	-25.005,00	001	4101	0,95	5: Over 20 years

Main conclusions

❑ The new tests have been used in production, on a monthly basis, for over a year with encouraging results;

❑ They have allowed us to:

- ➢ Detect reporting gaps and strange patterns, even subtle ones that only affect a few instruments;
- ➢ Oversee the evolution of the reporting of qualitative variables and to detect structural changes;
- ➢ Monitor the evolution of the amounts reported instrument by instrument, taking into consideration the qualitative variables that characterize them and ranking them by the degree of severity for further questioning.

❑ Hence, these new tools have contributed to an increase both in the effectiveness and efficiency of the data quality assessment process and so far we have received very positive feedback from the analysts.

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Monitoring at scale[1]

Enrico Apicella, Marco D'Errico and Pedro Marques, European Central Bank;
Antonio Ciullo, Deloitte; and Caroline Übelhör, Google

[1] This presentation 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 central banks and other institutions represented at the event.

Monitoring at Scale

Enrico Apicella,[1] Antonio Ciullo,[3] Pedro Marques,[2] Caroline Übelhör,[4] and Marco D'Errico[1]

[1] European Central Bank (European Systemic Risk Board Secretariat)

[2] European Central Bank (Directorate General Information Systems)

[3] Deloitte

[4] Google

IFC and Bank of Italy Workshop on "Data science in central banking"

21 Oct 2021

Disclaimer and acknowledgements

The views expressed in this presentation are our own and do not necessarily reflect those of the European Central Bank, the European Systemic Risk Board or its member institutions.

Thanks to Roberto Stok, Grzegorz Skrzypczynski, Oana Furtuna, Marco Canu, Giulia Di Rienzo, Marco Tizzoni, Umberto Coda, Eva Abend, Sylwia Kret, Mario Picano and Teresa Felici for input, comments and suggestions.

Overview

- Background: the ESRB’s monitoring mandate and analytical strategy
- The scaling problem in Systemic Risk Monitoring (SRM)
- Monitoring “at Scale”:
  - developing a general SRM framework
  - building blocks and additional elements
- The framework at work:
  - the March 2020 market turmoil and the largest margin call in history
- Conclusions and way forward

The ESRB's mandate and strategy rely on data and technologies

Mandate

*"The ESRB's task should be **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 and enhancing the financial system's resilience to shocks."*

*"The **interconnectedness of financial institutions and markets** implies that the monitoring and assessment of potential systemic risks should be based **on a broad set of relevant macroeconomic and micro-financial data and indicators**"*

Regulation (EU) No 1092/2010 establishing the ESRB

Strategy

*"Monitoring an interconnected financial system involves the availability of **detailed and granular transactions data**. But **in order to have the full picture, it is vitally important to be able to link data across markets, instruments and counterparties**." "Only a holistic view of the system will allow potential contagion channels to be identified and modelled. And that requires **investing in new technologies for data analytics and enhancing the capacity for authorities to link and share data and technical knowledge**."*

Former ESRB Chair Mario Draghi

Welcome remarks at the third and fourth annual conference of the ESRB

27 September 2018 and 26 September 2019

“*Data across markets*, *instruments and counterparties*” (examples)

Several granular data collections stemming from post-crisis reforms...

Dataset	Topic	Frequency	Level	Size / Complexity
EMIR (ESMA)	Derivatives	daily	counterparty / transaction	300k entities reporting 100 mln / day approx 160 bln records in total
SFTR (ESMA)	Repos/SFTs/other	daily	counterparty / transaction	9-10 mln / day nested and complex data 15k+ entities
AIFMD (ESMA)	AIF managers	quarterly	holdings, risk profiles, counterparties	60k funds
STSS (ESMA)	Securitisation	variable	variable	TBA / In progress
AnaCredit (ECB)	Loans	monthly	loan level	~29 mln / month

The two dimensions of Systemic Risk Monitoring (SRM)

Analytical	Organisational
Descriptive / distributional (monitor risk distribution & clusters)	**Research and Development** (enable full cycle: from ideas to usable products)
Interconnectedness / contagion (monitor and quantify potential channels)	**User eXperience** (translation of results, dissemination, etc.)
What if / scenario / simulations (compare against potential developments in the system)	**Knowledge Management** (replicability, sharing, understand, knowledge transfer, training)

The scaling problem in SRM: the *analytical* dimension

Analytical scaling: dimensionality problem

the number of analyses, indicators, processes and aggregations increases exponentially

- **Complexity & size**:
  the financial system is inherently complex, interconnected and adaptive
- **Linking** datasets
- **Speed**: as markets shifts rapidly, analytical outputs must be produced at a faster pace
- **Aggregation layers**: seamlessly/iteratively move across several layers of aggregation
  - Micro ↔ meso ↔ macro
    counterparty, groups, sector, countries, areas
- *[Deal with the (alas substantial!)* ***data quality issues*** *in granular supervisory data (not in this talk)]*

The scaling problem in SRM: the *organisational* dimension

Organisational scaling: operationalising the full cycle in the organisation

- **Process:**
  - from research to production (full R&D cycle)
  - possibly language agnostic
  - reproducible, reusable

- **Policy**: development, calibration

- **Communication & collaboration**: across the whole organisation and at different technical levels (data scientist, researcher, expert, management)

- **Knowledge management:**
  - share, explain, reuse, replicate
  - skill set & education

Solution (step 1): generalising the SRM framework

A scalable data model…

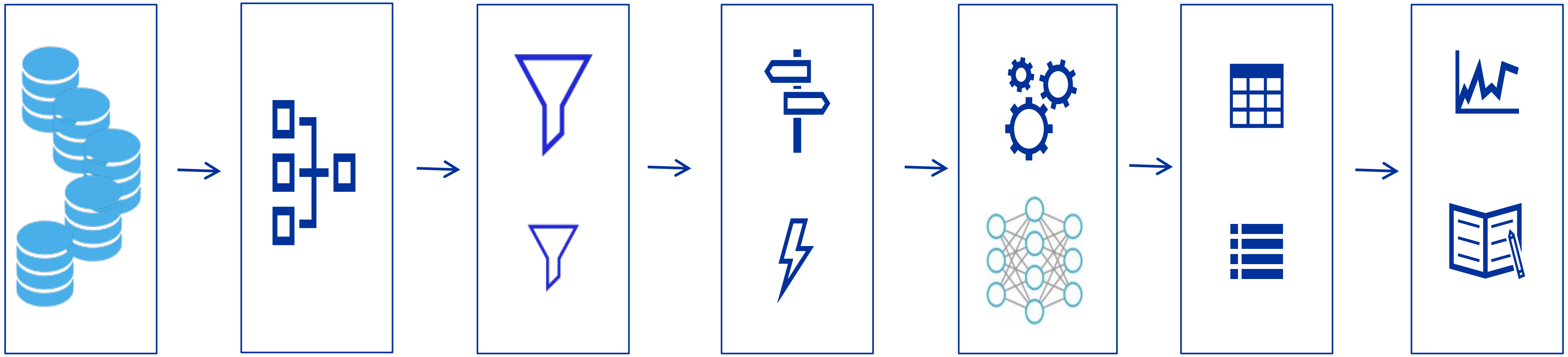
- model the **entire financial system at any level** as a *dynamic multilayer* network (multigraph)
- filters / aggregations / modifications ***preserve*** the same data model

Nodes	micro (counterparty sets), meso (groups), macro (sectors / countries)
Edges	(directed) balance sheet relationships (contracts, holdings, etc.)
Weights	any *measure* on the individual edges

…enabling scalable operations on the data

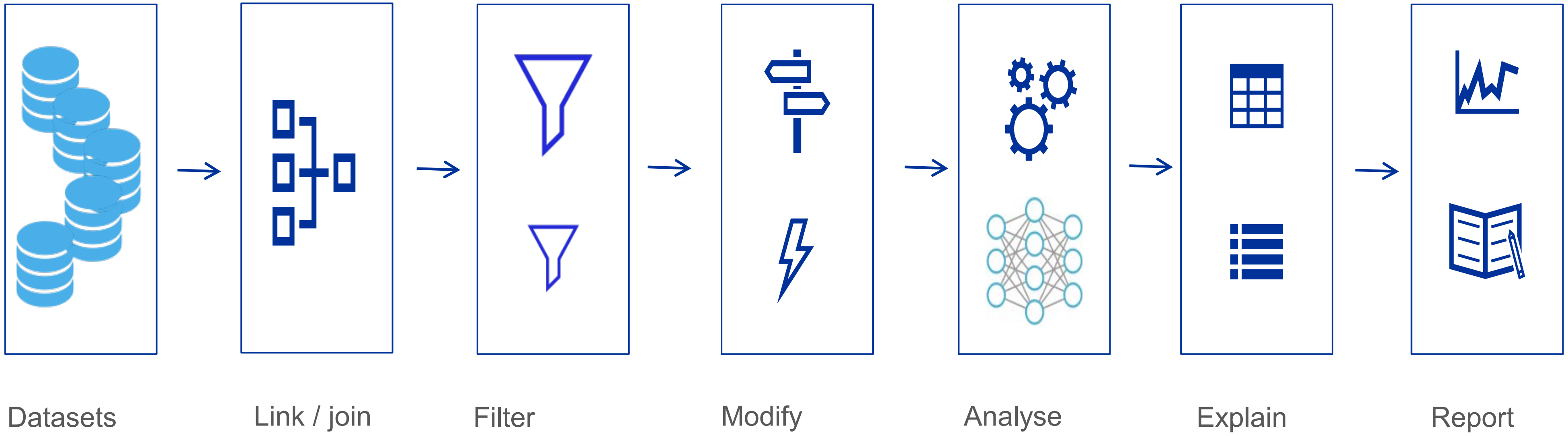
Descriptive / distributional	descriptive statistics on the multilayer network
Interconnectedness / contagion	dynamics on the multilayer network
What if / scenario / simulations	operations on a transformed network

Solution (step 2): generalising the SRM framework



Datasets → Link / join → Filter → Modify → Analyse → Explain → Report

Solution (step 2): generalising the SRM framework



indicators	
conceptualisation	**implementation**
concepts	parameters
operations	functions
processes	compositions

→

Combining:
*- **operations (functions)** and*
*- **concepts (parameters)***

*we **obtain all indicators at every level***
with a precise explanation of the process

Solution (step 3): generalising the SRM framework

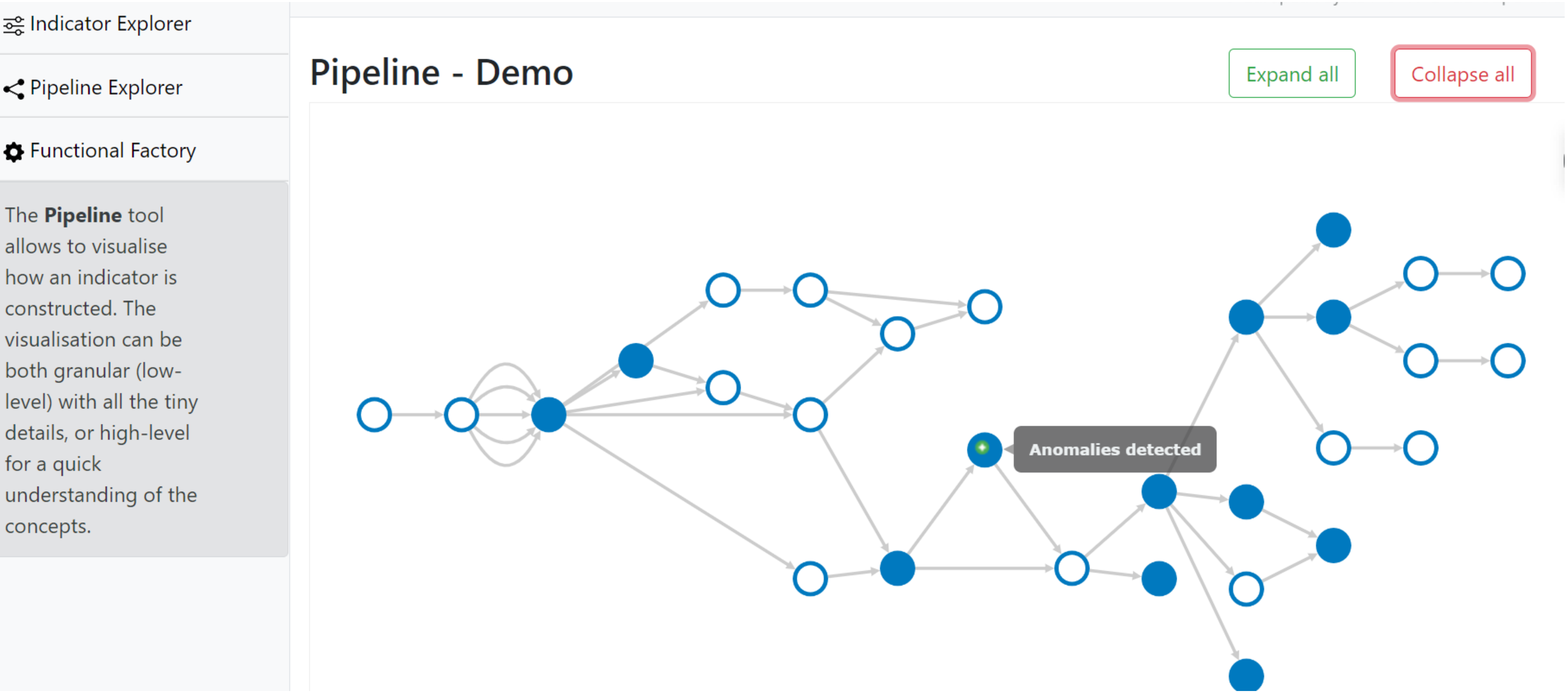
Building blocks (based on *functional* principles)

- **Functions** as operations defined for given **parameters** (concepts)
    - specialisation matters: first-class, higher-order, closures

- **Directed Acyclic Graphs** (DAG): encapsulating relationships between functions
    - sequential → composition (pipeline)
    - parallel → hierarchical / multivariable functions
      → adds expressivity via merging and branching, also increases complexity
      → counterfactuals / scenario analyses obtained via branching and merging
    - a DAG *becomes itself a function* → carries the description of the process (self documentation and rich semantics)
    - contagion models as fixed point iterations

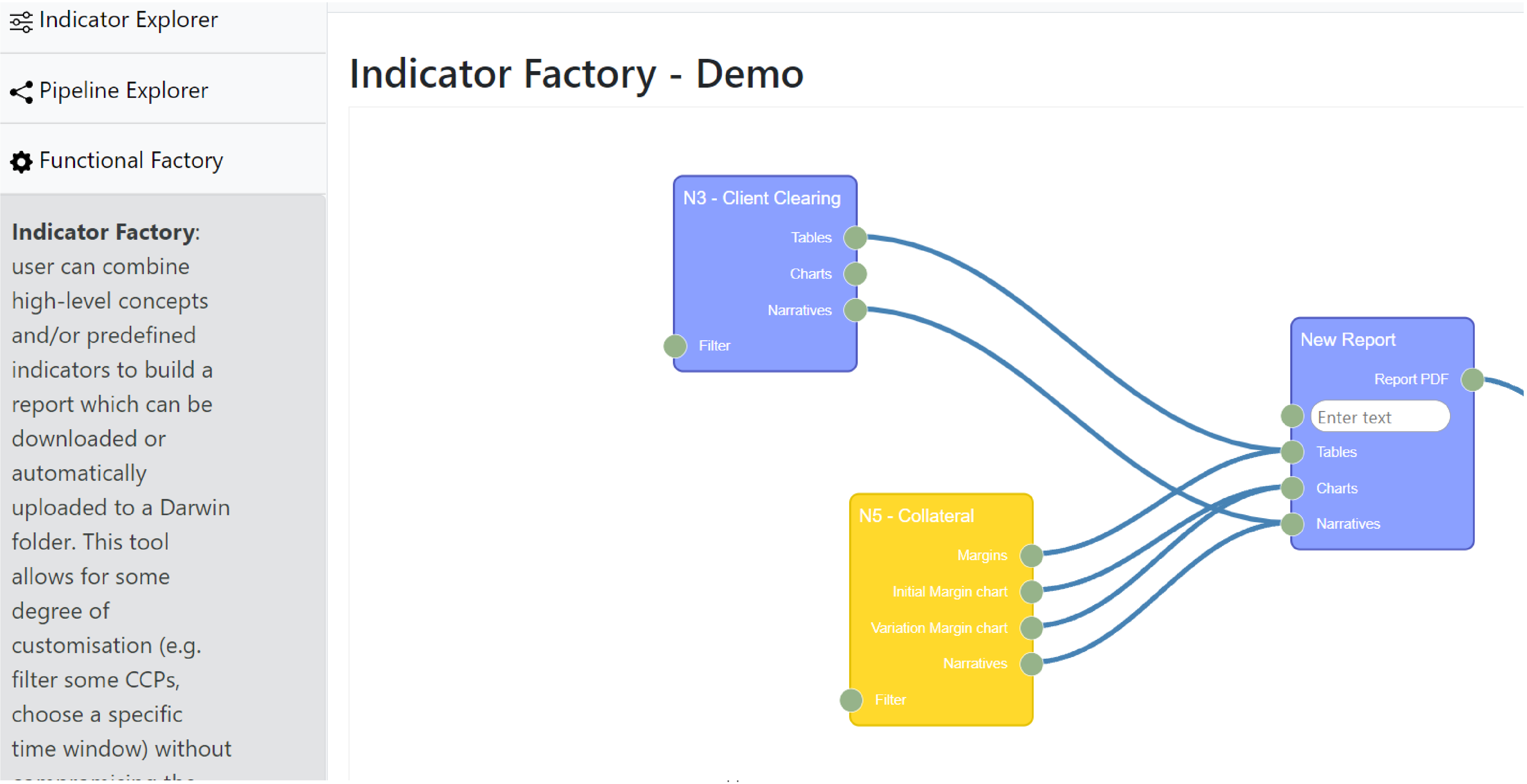
- **Factories:** a function returning a (set of) function(s) from other ones or from DAGs
- **SRM indicators**: factory, pipelines or elementary functions bound to specified parameters

Solution (step 3): generalising the SRM framework



The **Pipeline** tool allows to visualise how an indicator is constructed. The visualisation can be both granular (low-level) with all the tiny details, or high-level for a quick understanding of the concepts.

Solution (step 3): generalising the SRM framework



Indicator Factory: user can combine high-level concepts and/or predefined indicators to build a report which can be downloaded or automatically uploaded to a Darwin folder. This tool allows for some degree of customisation (e.g. filter some CCPs, choose a specific time window) without

Indicator Factory - Demo

Solution (step 4): generalising the SRM framework

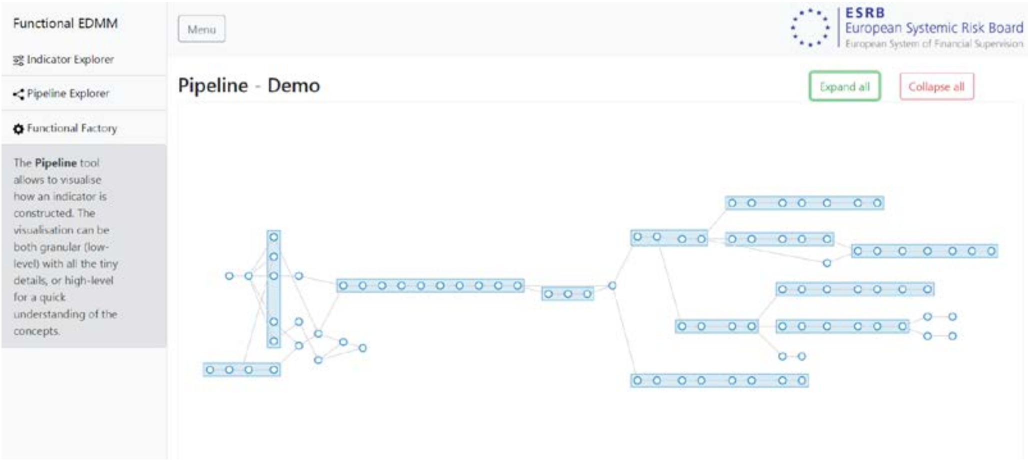
Building blocks - process

- Granular functions can be semantically framed by higher level ones and convey information about the hierarchical structure in the operations
  - ***filter*** *by* EU banks THEN ***group by*** bank country THEN ***compute*** max(margin call)
  - granularity tailored to each analysis

- Branching / merging conveys information on the reachability between operations:
  - compare two different scenarios (embeds causal relations)
  - reverse: find parameters at origin that would lead to given outcome (e.g. reverse stress testing)

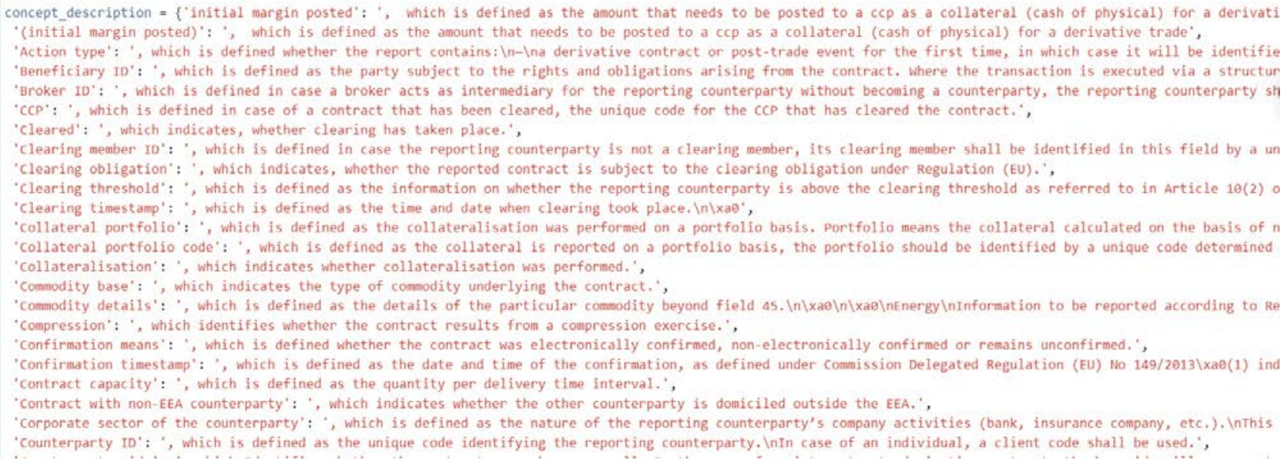
- Parameters can be linked to Knowledge Graphs: e.g. "*an **initial margin call** value of 10*, ***which is defined as****…*"

Solution (step 5): artefacts

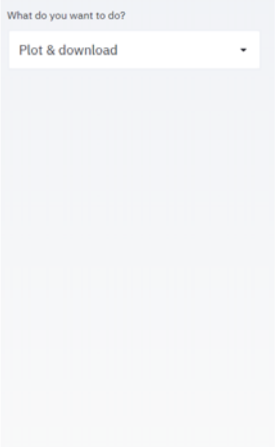


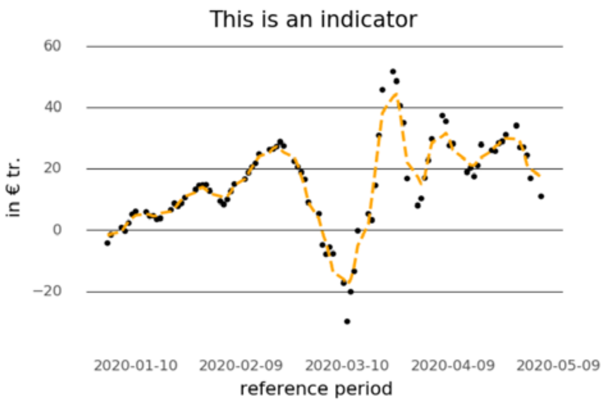
Pipeline - Demo

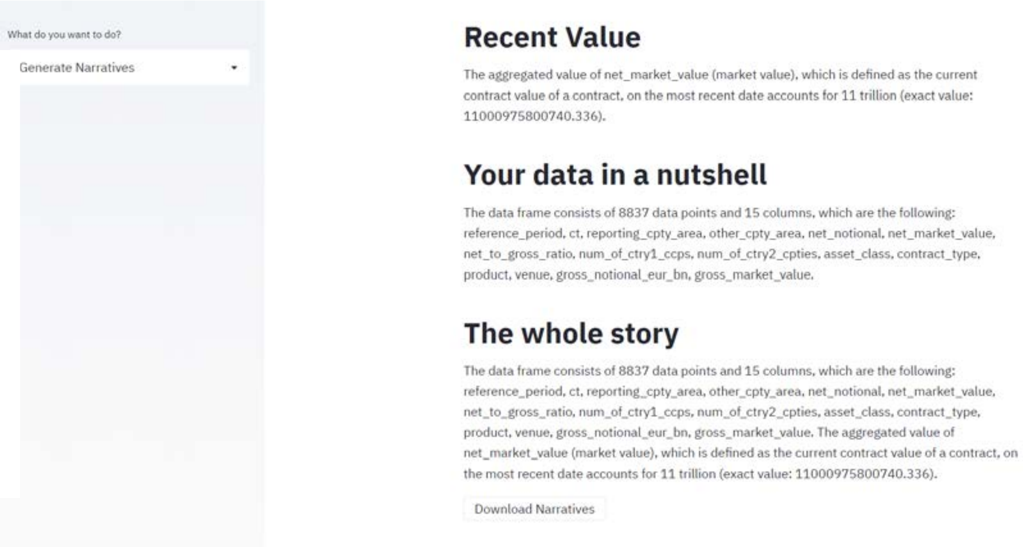
The **Pipeline** tool allows to visualise how an indicator is constructed. The visualisation can be both granular (low-level) with all the tiny details, or high-level for a quick understanding of the concepts.



```
concept_description = {'initial margin posted': ',  which is defined as the amount that needs to be posted to a ccp as a collateral (cash of physical) for a derivati
 '(initial margin posted)': ',  which is defined as the amount that needs to be posted to a ccp as a collateral (cash of physical) for a derivative trade',
 'Action type': ', which is defined whether the report contains:\n–\na derivative contract or post-trade event for the first time, in which case it will be identifie
 'Beneficiary ID': ', which is defined as the party subject to the rights and obligations arising from the contract. Where the transaction is executed via a structur
 'Broker ID': ', which is defined in case a broker acts as intermediary for the reporting counterparty without becoming a counterparty, the reporting counterparty sh
 'CCP': ', which is defined in case of a contract that has been cleared, the unique code for the CCP that has cleared the contract.',
 'Cleared': ', which indicates, whether clearing has taken place.',
 'Clearing member ID': ', which is defined in case the reporting counterparty is not a clearing member, its clearing member shall be identified in this field by a un
 'Clearing obligation': ', which indicates, whether the reported contract is subject to the clearing obligation under Regulation (EU).',
 'Clearing threshold': ', which is defined as the information on whether the reporting counterparty is above the clearing threshold as referred to in Article 10(2) o
 'Clearing timestamp': ', which is defined as the time and date when clearing took place.\n\xa0',
 'Collateral portfolio': ', which is defined as the collateralisation was performed on a portfolio basis. Portfolio means the collateral calculated on the basis of n
 'Collateral portfolio code': ', which is defined as the collateral is reported on a portfolio basis, the portfolio should be identified by a unique code determined
 'Collateralisation': ', which indicates whether collateralisation was performed.',
 'Commodity base': ', which indicates the type of commodity underlying the contract.',
 'Commodity details': ', which is defined as the details of the particular commodity beyond field 45.\n\xa0\n\xa0\nEnergy\nInformation to be reported according to Re
 'Compression': ', which identifies whether the contract results from a compression exercise.',
 'Confirmation means': ', which is defined whether the contract was electronically confirmed, non-electronically confirmed or remains unconfirmed.',
 'Confirmation timestamp': ', which is defined as the date and time of the confirmation, as defined under Commission Delegated Regulation (EU) No 149/2013\xa0(1) ind
 'Contract capacity': ', which is defined as the quantity per delivery time interval.',
 'Contract with non-EEA counterparty': ', which indicates whether the other counterparty is domiciled outside the EEA.',
 'Corporate sector of the counterparty': ', which is defined as the nature of the reporting counterparty's company activities (bank, insurance company, etc.).\nThis
 'Counterparty ID': ', which is defined as the unique code identifying the reporting counterparty.\nIn case of an individual, a client code shall be used.',
```







Recent Value

The aggregated value of net_market_value (market value), which is defined as the current contract value of a contract, on the most recent date accounts for 11 trillion (exact value: 11000975800740.336).

Your data in a nutshell

The data frame consists of 8837 data points and 15 columns, which are the following: reference_period, ct, reporting_cpty_area, other_cpty_area, net_notional, net_market_value, net_to_gross_ratio, num_of_ctry1_ccps, num_of_ctry2_cpties, asset_class, contract_type, product, venue, gross_notional_eur_bn, gross_market_value.

The whole story

The data frame consists of 8837 data points and 15 columns, which are the following: reference_period, ct, reporting_cpty_area, other_cpty_area, net_notional, net_market_value, net_to_gross_ratio, num_of_ctry1_ccps, num_of_ctry2_cpties, asset_class, contract_type, product, venue, gross_notional_eur_bn, gross_market_value. The aggregated value of net_market_value (market value), which is defined as the current contract value of a contract, on the most recent date accounts for 11 trillion (exact value: 11000975800740.336).

Download Narratives

Solution (step 5): testing

Scaling issue: adding functions to the repository and/or an existing DAG → potential blind spots

Solution: key principle → tests as functions (can compose and added to a DAG)

Properties

- *Left to right inheritance*: a test on a given function is inherited by any other function that composes with it → a test on an indicator (factory) is inherited by any specification of the indicator
- *Right to left inheritance*: tests on a composed function hold (and do so recursively) if and only if all tests the argument functions hold
- Implication: integration testing corresponds to unit testing on composed functions / DAGs

Applications

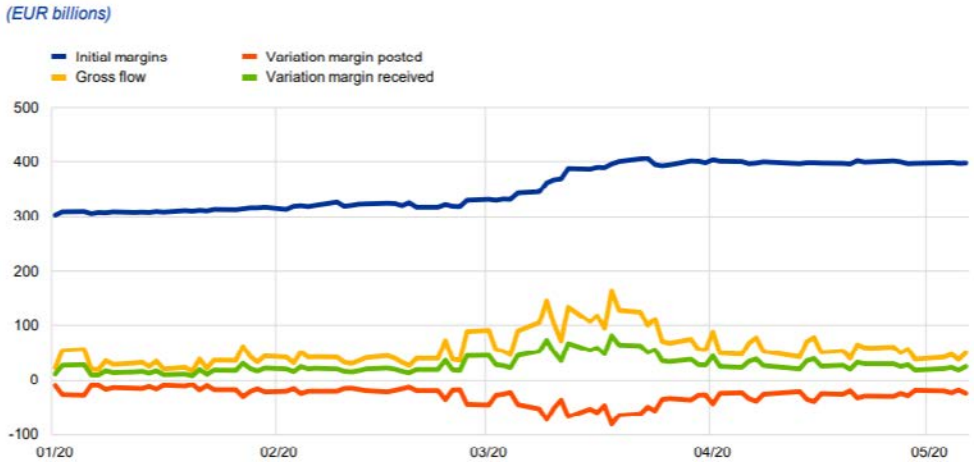
- A number of different test families can be applied as ***assertions on data***
  - Quality checks
  - Probabilistic testing (e.g. tests on the distribution or as r.v. transformations)
  - Business: testing on scenarios

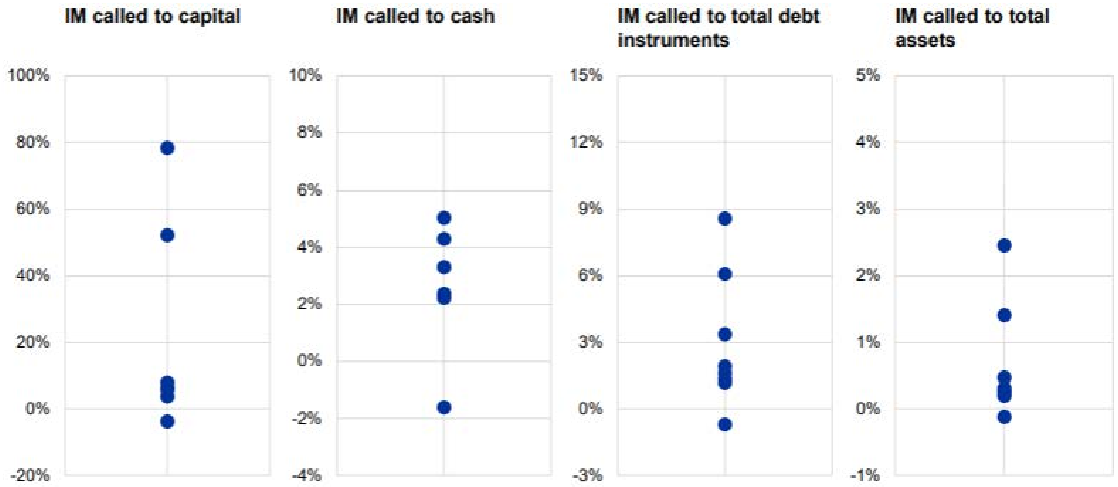
Monitoring at scale: the March 2020 margin call

- **March 2020**: sharp drop in asset prices and increased volatility, resulting in the **largest margin call in history**

- First ever experience of use of granular data in "**crisis mode**"

- ESRB tasks:
  - monitoring macro / meso / micro level
  - understanding the ***causes*** (what asset class? what dynamics?)
  - counterfactual analysis / reverse stress testing (depletion of liquid resources)
  - develop formal recommendation on liquidity risks


(EUR billions)


IM called to capital

IM called to cash

IM called to total debt instruments

IM called to total assets

Conclusions and way forward

- Systemic Risk Monitoring as a core task of ESRB requires data, technologies and analytical frameworks

- Contribution to solving the scaling problem by developing a general SRM framework inspired by functional principles

- Such framework allows to “monitor at scale” the financial system and has been adopted in several instances (also for policy development)

- Next steps will focus on:
  - enhancing the indicators set (in particular: constrained simulations)
  - provide a richer user experience, linking with existing knowledge

References (a few…)

- *Functional principles in data science*:
  - filter-map-reduce / split-apply-combine (Wickham, 2011)
  - tidyverse (Wickham et al, 2019)
  - grammar of graphics (Wilkinson, 1999)
  - lambda calculus

- *Data engineering*:
  - Google Dataflow
  - Facebook Prophet (scaling issues in forecasting)
  - Airflow, Luigi, Dask

- *Systemic Risk Monitoring*:
  - Rethinking the financial network (Haldane, 2009)
  - Shedding light on dark markets (ESRB, 2016)
  - Mapping exposures (Abad et al., 2021)
  - Network Valuation (Barucca et al. 2020)

- *Statistics / data science*:
  - Causality (Pearl, 2000, 2009, 2018)
  - Functions of learning, Learning from data (Strang, 2019 and 2020)

IFC-Bank of Italy Workshop on “Machine learning in central banking”

19-22 October 2021, Rome / virtual event

The impairment costs of traditional non-quantitative retail banking practices during residential real estate foreclosure sales and their effect on National, Central & Reserve bank(s) policy[1]

Emmanuel Blonkowski and James N Nicol,
Quant Property Solutions Australia

[1] This presentation 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 central banks and other institutions represented at the event.

The impairment costs of traditional non-quantitative retail banking practices during residential real estate foreclosure sales and their effect on National, Central & Reserve bank(s) policy

…with focus on improved supervised machine learning applications in micro-prudential processes within the product life cycle of residential mortgages.

James N. Nicol
Emmanuel Blonkowski

Quantifying a qualitative micro-prudential process during residential real estate foreclosures

Residential real estate is the world's most commonly transacted financial asset.

To facilitate this transaction, every country has its own unique sovereign prudential guidelines and requirements governing retail banking standards on residential mortgage lending.

Since the global financial crisis of mid-2007 to early 2009, many countries' Central/National/Reserve banks and monetary authorities have increased supervisory measures to ensure responsible lending standards, and amplified consumer protection advocacy to defend against improper lending practices and improve market stability within their retail banking systems.

And while the retail lenders' requirements for mortgage loan approval and servicing are backed by standardised regulatory processes and approved prudential procedures, there is an opportunity for quantitative improvement within the product lifecycle of residential mortgages, specifically when non-performance necessitates repossession, and a foreclosure sale.

With the creation of the Statistical Data and Metadata eXchange (SDMX) and ongoing updates to data analytics platforms and tools, there is an inference that Central/National/Reserve banks and Monetary Authorities can improve guidance and governance by incorporating quantitative data analytics tools, platforms and applications within qualitative risk management frameworks for better data-driven outcomes in their retail banking systems.

To aid in that transition, we have built a micro-prudential application that improves transactional optimisation by benchmarking and indexing a retail bank's choice of selling agent (market participant), facilitating a foreclosure sale.

The application utilises a supervised, machine-learning, geo-spatial, geo-specific, market-augmented, time series regression algorithm, with predictive and prescriptive attributes, that eliminates the qualitative selection bias that impairs sale prices of bank-owned, repossessed and foreclosed residential real estate assets.

JEL classification: C5 C81 D23 D46 D63 D91 R3

Keywords: #micro-prudential #residential #mortgage #foreclosure #repossession #quantitative #modelling #forecasting #transactioncosts #impairmentcosts #organizationlbehaviour

Contents

Introduction

Since the first version of the SDMX technical standard (1.0) was finalised in 2004 and approved in 2005 by the International Organization for Standardization (ISO), use of and interest in the SDMX by Central, National, Reserve banks and monetary authorities within the BIS membership has grown dramatically[1].

Though the use of the SDMX standards is concentrated in a few statistical areas[2], it has improved the regulatory reporting capabilities of Central, National and Reserve banks. However, as prudential supervisors of their retail banking systems, they face issues in the implementation of the SDMX standards and use of analytical applications to overcome the qualitative status quo[3]. This is due to the different interests of various stakeholders, and the challenge of replacing legacy reporting systems, processes, procedures, and practices[4] that utilize information at the level of granular, operational data.

For example, retail banking practices, procedures and processes guiding the product life cycle of residential real estate mortgages – specifically the qualitative processes and legacy procedures retail banks follow when mortgages become non-performing and default – have been slow to adopt, integrate or evolve alongside the SDMX or data analytic platforms. This has led to an ongoing impairment of prices and realised values when retail banks sell bank-owned, mortgage default, foreclosed and repossessed real estate[5].

Establishing reference points and benchmarks

The residential real estate sector plays an important role in financial and macroeconomic stability given its tight links with both the real economy and the financial system.[6]

A review of working papers and policy statements from global organizations, as well as guides and guidelines from sovereign prudential and supervisory bodies, establishes a common reference point for benchmarking residential mortgage underwriting standards and for management practices of immovable property within retail banking systems.

[1] Tissot B (2018) Journal for Mathematics and Statistical Science Volume 2018, Issue 1 SDMX, a Key Standard for Central Banks' Statistics

[2] Ehrmann H, Tissot B, Basşr E and Hülagü T (2015) IFC Report No 4 Central banks' use of the SDMX Standard - 2015 Survey, conducted by the SDMX Global Conference Organising Committee

[3] Cristano J, Kienecker K, Prenio J and Tan E (2020) FSI Insights on policy implementation No 29

[4] *Ibid*

[5] Allen F and E Carletti (2011a), 'Systemic Risk from Real Estate and Macro-Prudential Regulation', Paper presented at the Federal Reserve Board and *Journal of Money, Credit and Banking* Conference 'The Regulation of Systemic Risk', Washington DC, September 15–16, 2013

[6] European Systematic Risk Board (2021) European A Review of Macroprudential Policy in the EU in 2020

Analysing and tracing these reference points to their point of origin establishes the framework of the principles and standards within retail banking systems, and allows scope to empirically review, compare and test those principles and standards.

Overview of countries, supervisory bodies, financial authorities, working papers, policy statements, guides and guidelines for retail residential mortgage lending covered in this paper

Table1

Country	Financial Authority	Document	Date
Australia	Australian Prudential Regulatory Authority	Prudential Practice Guide APG 223 Residential Mortgage Lending	July 2019
Canada	Office of the Superintendent of Financial Institutions	Residential Mortgage Underwriting Practices and Procedures B-20	October 2017
England	The Prudential Regulatory Authority The Financial Conduct Authority	The PRA Rule Book The FCA Handbook	April 2013 April 2014
France	The French Prudential Supervision and Resolution Authority High Council for Financial Stability	Compendium of regulations relating to the exercise of banking and financial activities R-2021 - 1	October 2020 December 2019
Germany	The Federal Financial Supervisory Authority	Section 48u – the Banking Act	June 2017
Ireland	Central Bank of Ireland	The Irish Statute Book	February 2015
Spain	Bank of Spain	Regulatory Changes in Prudential Supervision	May 2013
United States of America	Consumer Financial Protection Bureau	The 2016 Mortgage Servicing Rule	August 2016
International Supervisory Body	Financial Stability Board	Thematic Review on Mortgage Underwriting and Origination Practices Principles for Sound Residential Mortgage Underwriting Practices	March 2011 April 2012
European Supervisory Body	European Central Bank	Trends and risks in credit underwriting standards of significant institutions in the Single Supervisory Mechanism	June 2020
European Supervisory Body	European Systematic Risk Board	Recommendation of the European Systemic Risk Board of 31 October 2016 on closing real estate data gaps as amended by Recommendation ESRB/2019/3 (ESRB/2016/14)	June 2021

In Australia, lending secured by mortgages over residential property constitutes the largest credit exposure in the Australian banking system[7]. Prudential practice guides, (PPG) provide guidance on the Australian Prudential Regulatory Authority's (APRA) view of sound practice in particular areas. The PPG APG 223 focus is Residential Mortgage Lending (July 2019) which in turns sites the Financial Stability Board's (FSB) Principles for Sound Residential Mortgage Underwriting Practices (April 2012) as a Rosetta Stone of sorts which sets out minimum underwriting standards

[7] APRA Prudential Practice Guide APG 223 Residential Mortgage Lending

that the FSB encourages supervisors of authorised deposit-taking institutions (ADI), like APRA, to implement.

In March 2011, the Financial Stability Board (FSB) published a thematic review of residential mortgage underwriting and origination practices[8]. One of the recommendations was that the FSB set out to develop a principles-based framework for sound under writing practices. Of the seven principles identified in the FSB Principles for Sound Residential Mortgage Underwriting Practices (April 2012), principle four; Effective collateral management, has several sub points of process requirements on the professionalism, diligence and independence of the appraisers and valuers of residential real estate.[9]

Understandably, the Principles refer to consumer protection issues that contribute to efforts to improve financial stability and prudential standards, the Principles are not intended to be a statement of consumer protection standards[10]. Allowing jurisdictions to adopt consumer protection standards that are appropriate to them.

Absent is guidance and guidelines on the selection process of market participants should non-performance of the mortgage necessitate the involuntary liquidation and foreclosure sale of the collateral by retail banks.

Likewise, while PPG APG 223 Residential Mortgage Lending (July 2019) offers ADI's guidance and guidelines in several key areas of the residential mortgage product life cycle, i.e. risk management framework, loan origination, stress testing, reliance on automated valuation models (AVMs) in security valuation as well as requirements on the professionalism, diligence and independence of the appraisers and valuers of residential real estate, absent is guidance and guidelines on the selection of market participants, should non-performance of the mortgage necessitate the involuntary liquidation and foreclosure sale of the collateral by retail banks.

As real estate and financial markets are differential and dynamic, we reviewed the documents in table 1 to confirm homogenous absences in guidance and guidelines on the selection process of market participants should non-performance of the mortgage necessitate the involuntary liquidation and foreclosure sale of the collateral by retail banks.

There were absences.

In that absence, we analysed a random sample group of mortgagee residential real estate sale transactions from Australia, Canada, England, France, Germany, Ireland, Spain, and the United States of America applying quantitative processes to predict and prescriptively index the retail banks choice of market participant's sale result against their markets average (arithmetic mean) sale result as a guide.

[8] http://www.financialstabilityboard.org/publications/r_11031a.pdf

[9] FSB (2012) Principles for Sound Residential Mortgage Underwriting Practices

[10] FSB (2012) Principles for Sound Residential Mortgage Underwriting Practices

Research, Observations and Comparisons

Our research was done in three parts:

Part One

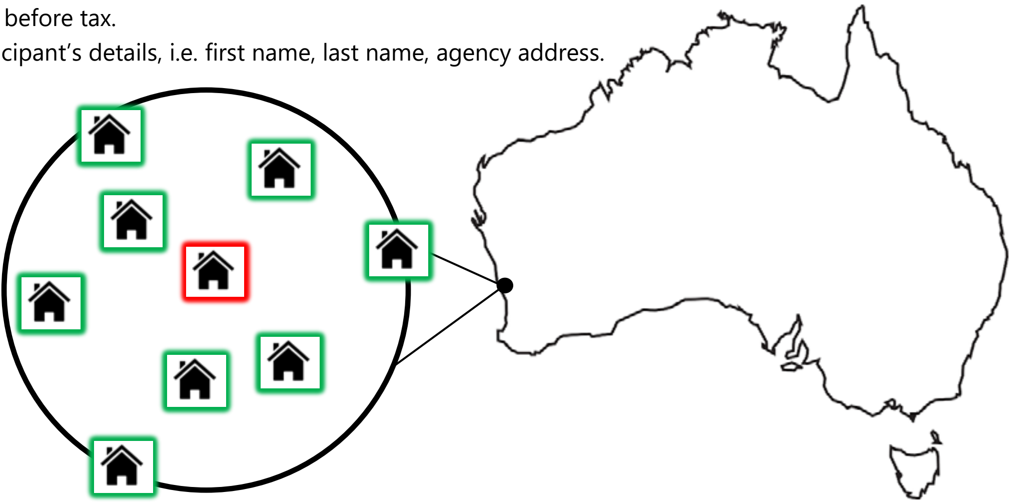
A random sample group of 500 Australian mortgagee residential real estate sales transacted between June 2018 and March 2020 (21 months) was analysed using newly developed analytical software. This analysis aggregated metrics which had previously not been analysed together – unique key result indicators of market participants – using API feeds from multiple data suppliers to ensure data accuracy and veracity.

In this way we predicted and prescriptively indexed the transaction metrics of the sale results of a retail bank's choice of market participants to be either above or below their market's average sale result.

Rather than over indexing, the observations of the key result indicators used to evaluate the transaction econometrics of the 'market average' were basic:

Box 1

- Property attributes, such as land size, number of bedrooms, bathrooms, car spaces correlating with the repossessed/foreclosed residential real estate sale.
- Property location, within 1,000 meters of a bank managed/bank owned/mortgage default residential real estate sale.
- Time series, within 180 days, 90 days prior and post bank managed/bank owned/mortgage default residential real estate sale transaction.
- Sale Price, $ Per Square Meter (Land Size).
- Sale Price, $ before tax.
- Market participant's details, i.e. first name, last name, agency address.



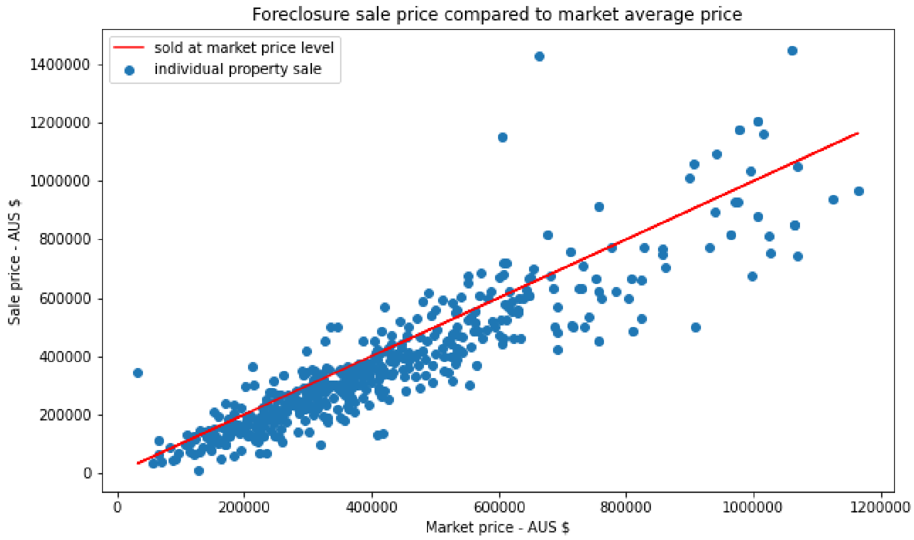
Of the 500 properties sold, within 3 months of being transacted, 94.3% were correctly predicted in having sold above or below their market average.

Within 6 months of being transacted, 91.1% were correctly predicted in having sold above or below their market average and 89.9% were correctly predicted in having sold above or below their market average within 12 months of being transacted.

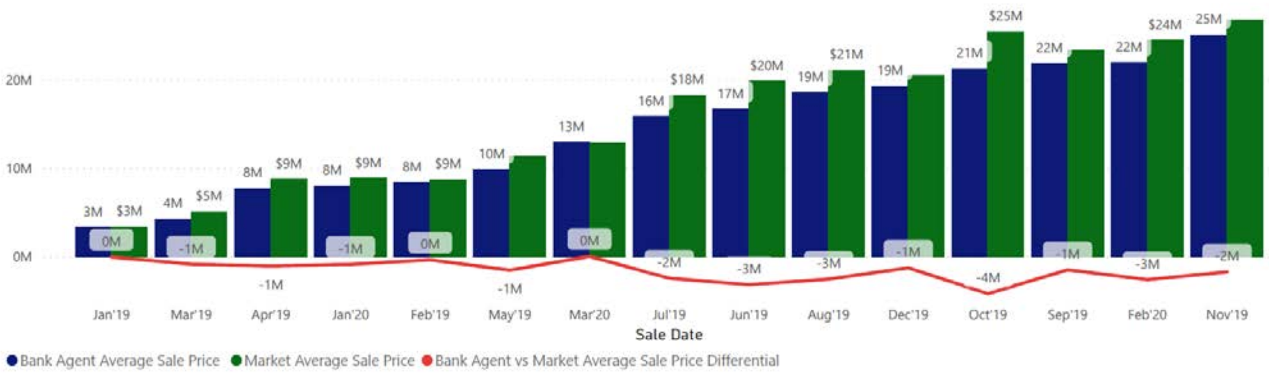
At the time of sale, the indexing revealed 76.8% of the bank selected market participants sold the foreclosed property below their market's average sale price.

Month by month, over the 21 months, the banks choice of market participant, under-performed when compared against their market's average.

The average difference between the bank selected market participant's foreclosure result and their market's average sale price, per the sample group, was -$46,607 at the time of sale.



Box 2



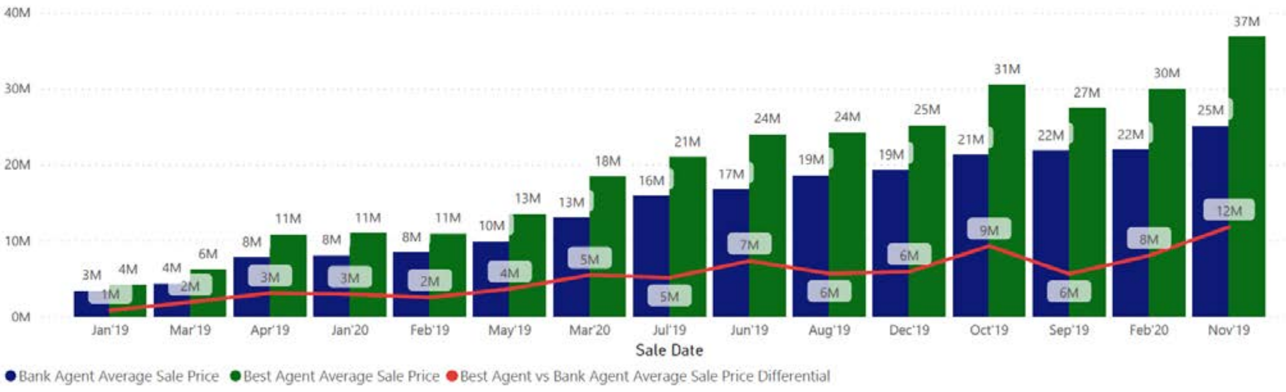
Box 3

Part Two

When benchmarking the performance of the bank selected market participant of the foreclosed property against their market's average sale price results, we indexed market participants who would have been a more ideal or 'BestAgent' to ensure transactional optimisation of the foreclosed property.

The gap between the bank selected market participant's sale price result and the 'BestAgent' market participant's average sale price result is even more significant when compared with their market's average sale price result, $157,907.

Box 4



Part Three

Repeating the process outlined in Part One to include bank owned, mortgage default, repossessed and foreclosed real estate sales in Canada, England, France, Germany, Ireland, Spain, and the United States of America, we prescriptively indexed the bank's selected market participants results to be either above or below their market averages and against other market participants.

Table 2

Country	Residential real estate foreclosure sales	% of foreclosure sales below market average	Differential average foreclosure sale price to market average sale price	Differential bank selected agent average sale price to BestAgent average sale price
Australia	500[1]	77	-$46,607	$157,907
Canada	750[2]	78	-$43,899	$87,455
England	1120[3]	68	-£28,540	£79,887
France	1340[4]	71	-€42,115	€87,766
Germany	1660[5]	71	-€29,009	€67,882
Ireland	98[6]	66	-€33,150	€64,078
Spain	940[7]	69	-€47,441	€82,369
The United States of America	6560[8]	79	-$52,572	$104,741

[1] Corelogic, Domain and SQM [2] Realtor.ca and ForeclosureSearch.ca [3] UKAuctionList, Rightmove and Zoopla [4] LeBonCoin.fr, SeLoger.com and Pap.fr [5] Immobilienscout24.de, Immowelt.de and Govesta.co [6] Daft.ie and MyHome.ie [7] Fotocasa.es and BankBargainsSpain.com [8] RealtyTrac, Zillow and RedFin

Algorithm and modelling

The regression algorithm used to index the market participants operates within a narrow band of temporal and spatial scales, nominated by the user.

The best fit, a binomial distribution model, allowed us to figure the probability of observing a specific number of ‘best’ outcomes when the process was repeated against the market participants with the outcome for a given market participant is either success (above market) or failure (below market) benchmark result against the market’s average and other market participants.

Per below, "n" denotes the number of observations or agent performance metrics i.e. Days on Market, List Price to Sale Price Differential, $ per SQM, and "x" denotes the number of "successes" or events of interest occurring during "n" observations. The probability of "success" or occurrence of the outcome of interest is indicated by "p".

With this notation in mind, the binomial distribution model is defined as:

$$P\left(X\,\text{"}\mathbf{successes}\text{"}\right)=\frac{n!}{x!\,(n-x)!}p^{x}\left(1-p\right)^{(n-x)}$$

This model allows for the proprietary indexing of market participants, geospatially on a property-by-property basis. In use and function as a micro-prudential tool, the software utilises supervised machine learning, attributes. Identifying and isolating outlier results, dimensionally reducing and classifying input variables with each software update.

As a time series, geospatial, geospecific, market augmented proprietary indexing system with predictive and prescriptive attributes to quantitatively improve upon legacy qualitative retail banking practices during the involuntary liquidation of residential real estate, The ‘BestAgent’ Index ensures greater risk transparency to facilitate better data driven decisions when selecting market participants during an involuntary liquidation event at a micro-prudential level.

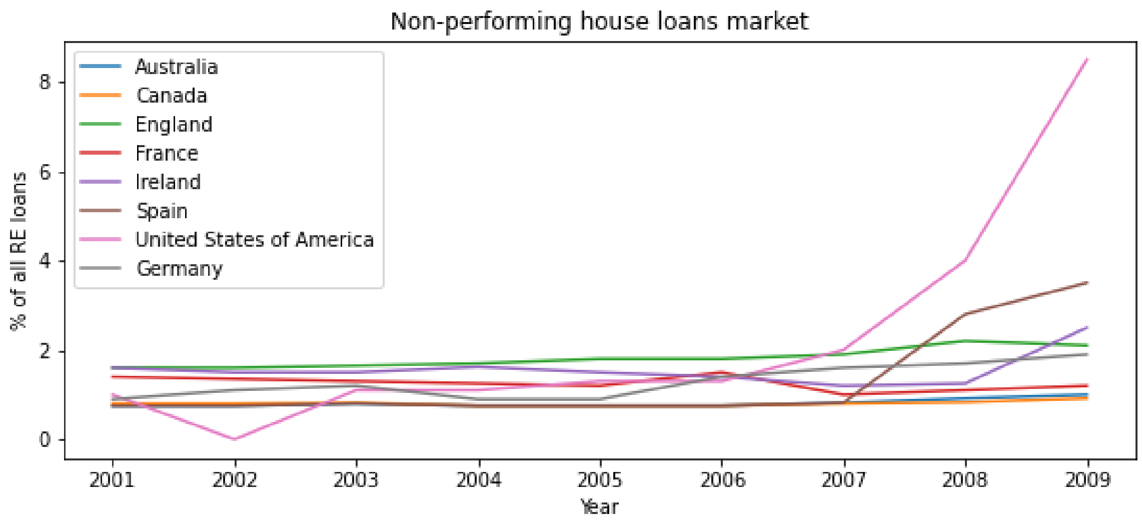
Residential foreclosures and their effect on Central/National/ Reserve Bank(s) and Monetary Authorities Policies

In comparison, globally, losses during and after the global financial crisis were minimal in Australia[11] (Rogers D 2015).

The global financial crisis highlighted involuntary liquidation value vs market value of residential real estate from impatient creditors and cautionary markets. In healthy developed residential real estate markets; prior to 2007; foreclosure sales and non-performing housing loans annually represented less than 2%[12] of their respected markets.

[11] Rogers D Credit Losses at Australian Banks: 1980–2013

[12] Lea M (2010) International Comparison of Mortgage Product Offerings



① Lea M (2010) International Comparison of Mortgage Product Offerings

Post global financial crisis, the increased regulation and oversight of residential mortgage underwriting practices by many Central/National/Reserve banks, monetary authorities, prudential regulatory authorities, and various global supervisory economic bodies focused on loan origination, lending documentation standards and definitions of best practices for residential valuers and appraisers.

Absent from the commentary, conclusions and recommendations was standards and best practices on the selection process(es) of the administrators for the mechanism of involuntary liquidation: the market participants.

And while it can be difficult to disentangle the triggers of a downturn event from its amplifiers, default, foreclosure and repossession sales of residential real estate were amplifier events of the global financial crisis.[13]

The future challenge for Central/National/Reserve banks, monetary and prudential regulatory authorities is setting quantitative standards within a qualitative principles-based framework to improve outcomes and reduce the gap between involuntary liquidation value vs market value.

Current frameworks contain flexibility for induvial jurisdictions to adopt their own standards according to their own circumstances[14] as real estate markets are dynamic with underlying risk differing greatly across jurisdictions and within countries.[15]

[13] Calhoun M (2018) Lessons from the financial crisis: The central importance of a sustainable, affordable and inclusive housing market

[14] Financial Stability Board (2011) Thematic Review on Mortgage Underwriting and Origination Practices Peer Review Report

[15] *Ibid*

Appreciating those factors, the most universal of statistical and market metrics is the arithmetic mean of a market.

Outcomes are either above or below average when supplied temporal and spatial context.

The use of supervised machine learning software by Central/National/Reserve Banks and Monetary Authorities would aid in effective collateral management by lowering a threshold of action triggered by current frameworks and would advance consumer finance protection by identifying market participants that outperform their market's average.

During involuntary liquidation, a market's average can be used as a predictive benchmark to quantitatively index a market participant's past outcomes in that market as an indication of future results.

In summary

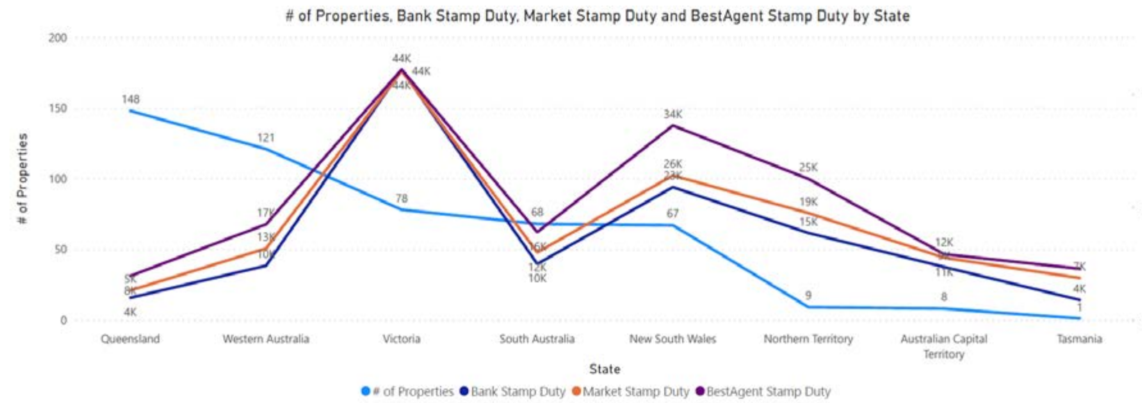
With Australia as point of reference, within the sample group, when retail banks sell bank owned property, 3 out of 4 transactions sell for below their market's average sale price while other market participants transact homes of identical basic hedonic attributes for above their market's average.

What has been identified is a quantitative opportunity for improvement in the guidance that Central, National, Reserve Banks, Monetary Authorities and their prudential bodies give to their retail banking sectors to lessen the impairment costs of traditional qualitative retail banking practices within the product lifecycle of residential mortgages when non-performance necessitates involuntary liquidation.

This is not a theoretical concept or an abstract financial exercise. It is a real-world data science solution use of a supervised machine learning algorithm that recognises the difference in unrealised funds not as a simple redistribution of wealth but as a net loss to a collective markets value.

The difference, in people terms: tens of thousands of dollars, pounds and euros in unrealised funds for the defaulted mortgager's repayment capacity to creditors.

Exponentially, again using Australia as a point of reference, the gain of hundreds of thousands, quarterly, in unrealised taxes and stamp duty, at state level, from the causal inference of impaired correlated sales.



Box 6

Next steps

As mentioned in the introduction, as prudential supervisors, introducing new quantitative SDMX standards and analytical applications to replace qualitative legacy, processes, procedures, and practices[16] within retail banking systems can be a challenge due to different interests of various stakeholders.

To overcome this challenge, we envision as a litmus test to ensure quantitative best practice the use of innovation labs and hubs within Central/National/Reserve banks and monetary, prudential, and financial supervisory authorities, to pool resources and to test new analytical software by running comparative testing parallel to the processes, procedures and practices currently used in their retail banking sectors.

[16] Cristano J, Kienecker K, Prenio J and Tan E (2020) FSI Insights on policy implementation No 29

References

Allen F and E Carletti (2011a), 'Systemic Risk from Real Estate and Macro-Prudential Regulation', Paper presented at the Federal Reserve Board and *Journal of Money, Credit and Banking* Conference 'The Regulation of Systemic Risk', Washington DC, September 15–16.

APRA Prudential Practice Guide APG 223 Residential Mortgage Lending

Calhoun M (2018) Lessons from the financial crisis: The central importance of a sustainable, affordable and inclusive housing market

Cristano J, Kienecker K, Prenio J and Tan E (2020) FSI Insights on policy implementation No 29 From data reporting to data sharing; how far can suptech and other innovations challenge the status quo of regulatory reporting.

Ehrmann H, Tissot B, Basşr E and Hülagü T (2015) IFC Report No 4 Central banks' use of the SDMX Standard - 2015 Survey, conducted by the SDMX Global Conference Organising Committee

European Systematic Risk Board (2021) A Review of Macroprudential Policy in the EU in 2020

Financial Stability Board (2012) Principles for Sound Residential Mortgage Underwriting Practices

Lea M (2010) International Comparison of Mortgage Product Offerings

Rogers D (2015) Credit Losses at Australian Banks: 1980–2013

Tissot B (2018) Journal for Mathematics and Statistical Science Volume 2018, Issue 1 SDMX, a Key Standard for Central Banks' Statistics

QUANT
PROPERTY SOLUTIONS
AUSTRALIA

Who are we and what do we do?

We’re a RegTech SaaS developer specialising in creating new and innovative micro-prudential tools for Central, National and Reserve Banks to aide their retail banking sectors in better data driven decisions around stress testing, risk management and transactional optimisation during the involuntary liquidation of distressed, repossessed and foreclosed residential real estate.

QUANT
PROPERTY SOLUTIONS
AUSTRALIA

What’s the problem?

What are the impairment costs of traditional non-quantitative retail banking practices during residential real estate foreclosure sales and what is their effect on Central bank policy?

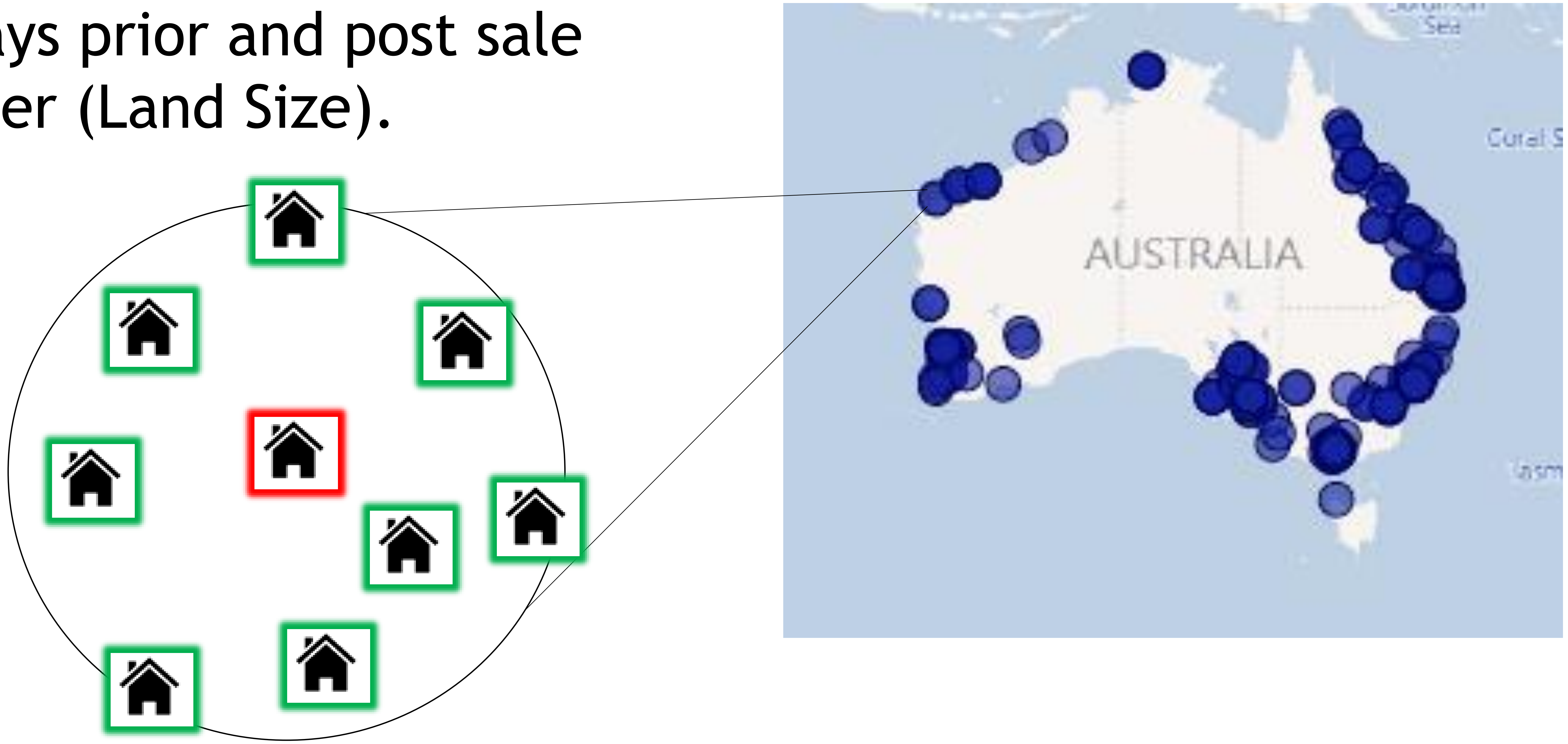
And how could the use of supervised machine learning applications during micro-prudential processes within the product life cycle of residential mortgages, lessen that impairment.

Overview of countries, supervisory bodies, financial authorities, working papers, policy statements, guides and guidelines for retail residential mortgage lending covered in this presentation

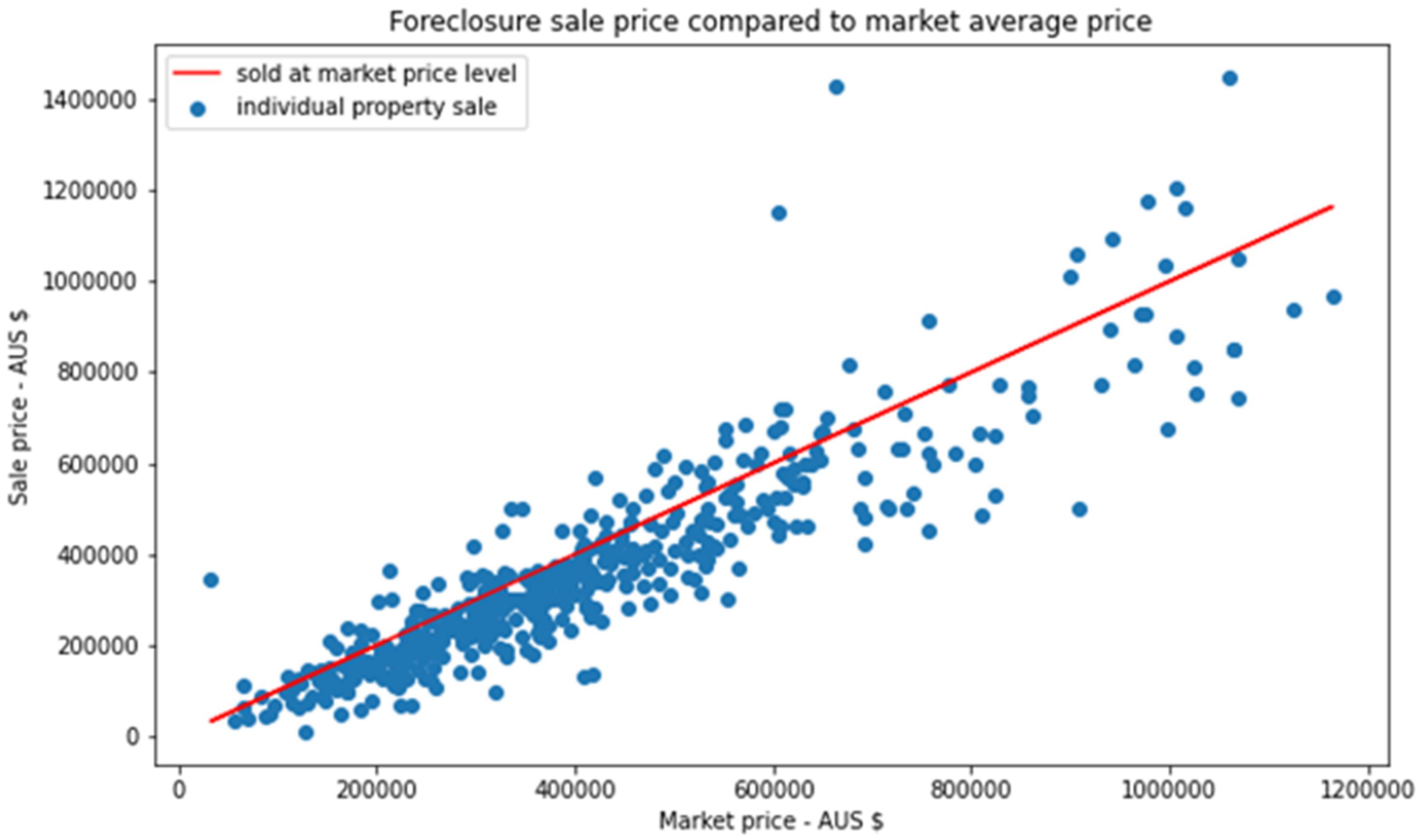
Country / Body	Financial Authority	Document	Date
Australia	Australian Prudential Regulatory Authority	Prudential Practice Guide APG 223 Residential Mortgage Lending	July 2019
Canada	Office of the Superintendent of Financial Institutions	Residential Mortgage Underwriting Practices & Procedures B-20	October 2017
England	The Prudential Regulatory Authority The Financial Conduct Authority	The PRA Rule Book The FCA Handbook	April 2013 April 2014
France	The French Prudential Supervision and Resolution Authority High Council for Financial Stability	Compendium of regulations relating to the exercise of banking and financial activities R-2021 - 1	October 2020 December 2019
Germany	The Federal Financial Supervisory Authority	Section 48u – the Banking Act	June 2017
Ireland	Central Bank of Ireland	The Irish Statute Book	February 2015
Spain	Bank of Spain	Regulatory Changes in Prudential Supervision	May 2013
United States of America	Consumer Financial Protection Bureau	The 2016 Mortgage Servicing Rule	August 2016
International Supervisory Body	Financial Stability Board	Thematic Review on Mortgage Underwriting & Origination Practices Principles for Sound Residential Mortgage Underwriting Practices	March 2011 April 2012
European Supervisory Body	European Central Bank	Trends and risks in credit underwriting standards of significant institutions in the Single Supervisory Mechanism	June 2020
European Supervisory Body	European Systematic Risk Board	Recommendation of the European Systemic Risk Board of 31 October 2016 on closing real estate data gaps as amended by Recommendation ESRB/2019/3 (ESRB/2016/14)	June 2021

Research, Observations and Comparisons

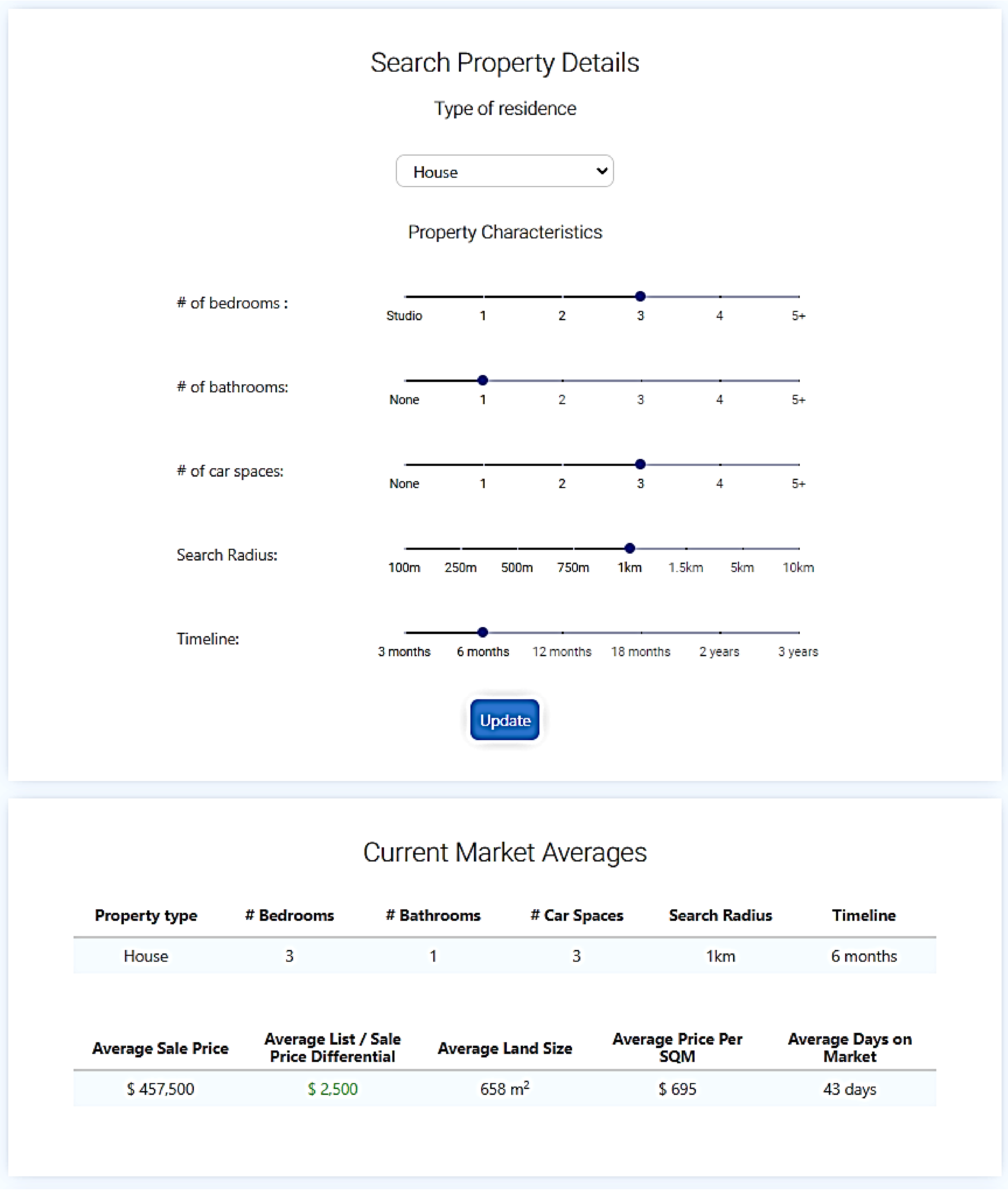
- A random sample group of 500 Australian mortgagee sales - June 2018 to March 2020
- Basic hedonic property attributes: # bedrooms, # bathrooms, # car spaces, land size
- Property location: within 1000 sqm
- Time series: 180 days, 90 days prior and post sale
- Sale Price: $ Per Square Meter (Land Size).
- Sale Price: $ before tax.
- Market participant's details



Research, Observations and Comparisons



76.8% sold below market



Search Property Details

Type of residence

House

Property Characteristics

# of bedrooms : Studio 1 2 [✓ 3] 4 5+

# of bathrooms: None [✓ 1] 2 3 4 5+

# of car spaces: None 1 2 [✓ 3] 4 5+

Search Radius: 100m 250m 500m 750m [✓ 1km] 1.5km 5km 10km

Timeline: 3 months [✓ 6 months] 12 months 18 months 2 years 3 years

Current Market Averages

Property type	# Bedrooms	# Bathrooms	# Car Spaces	Search Radius	Timeline
House	3	1	3	1km	6 months

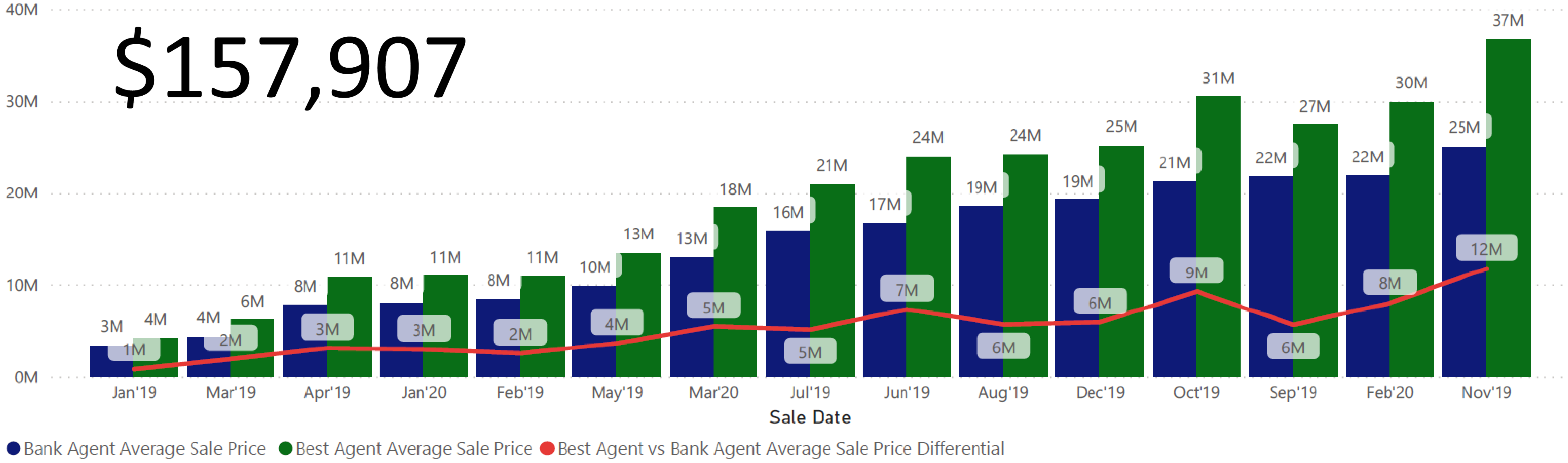
Average Sale Price	Average List / Sale Price Differential	Average Land Size	Average Price Per SQM	Average Days on Market
$ 457,500	$ 2,500	658 m^2	$ 695	43 days

Research, Observations and Comparisons

-$46,607

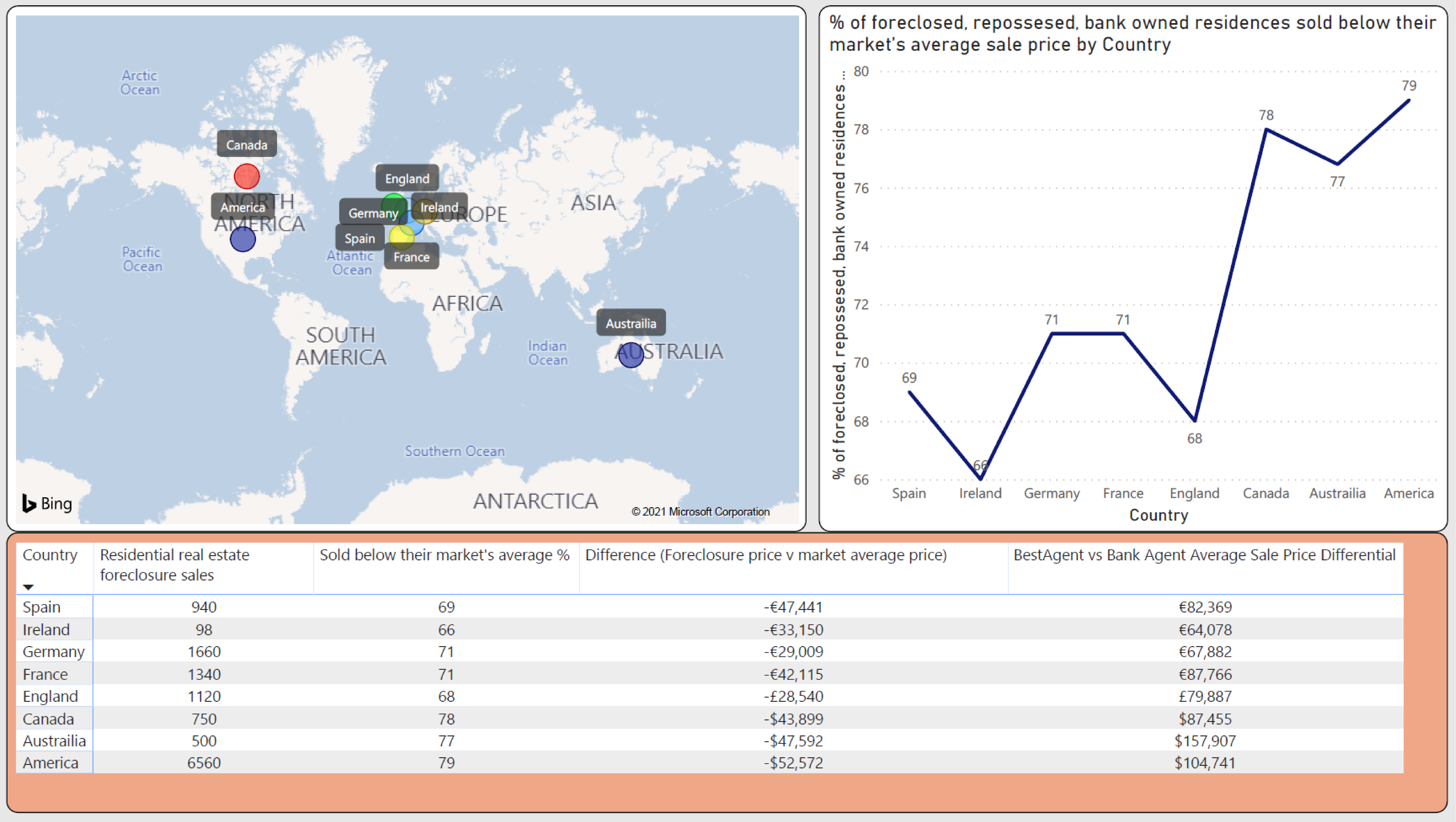
20M
10M
0M
3M
$3M
0M
4M
$5M
-1M
8M
$9M
-1M
8M
$9M
-1M
8M
$9M
0M
10M
-1M
13M
0M
16M
$18M
-2M
17M
$20M
-3M
19M
$21M
-3M
19M
-1M
21M
$25M
-4M
22M
-1M
22M
$24M
-3M
25M
-2M
Jan'19
Mar'19
Apr'19
Jan'20
Feb'19
May'19
Mar'20
Jul'19
Jun'19
Aug'19
Dec'19
Oct'19
Sep'19
Feb'20
Nov'19
Sale Date
● Bank Agent Average Sale Price
● Market Average Sale Price
● Bank Agent vs Market Average Sale Price Differential

Research, Observations and Comparisons



$157,907

Research, Observations and Comparisons



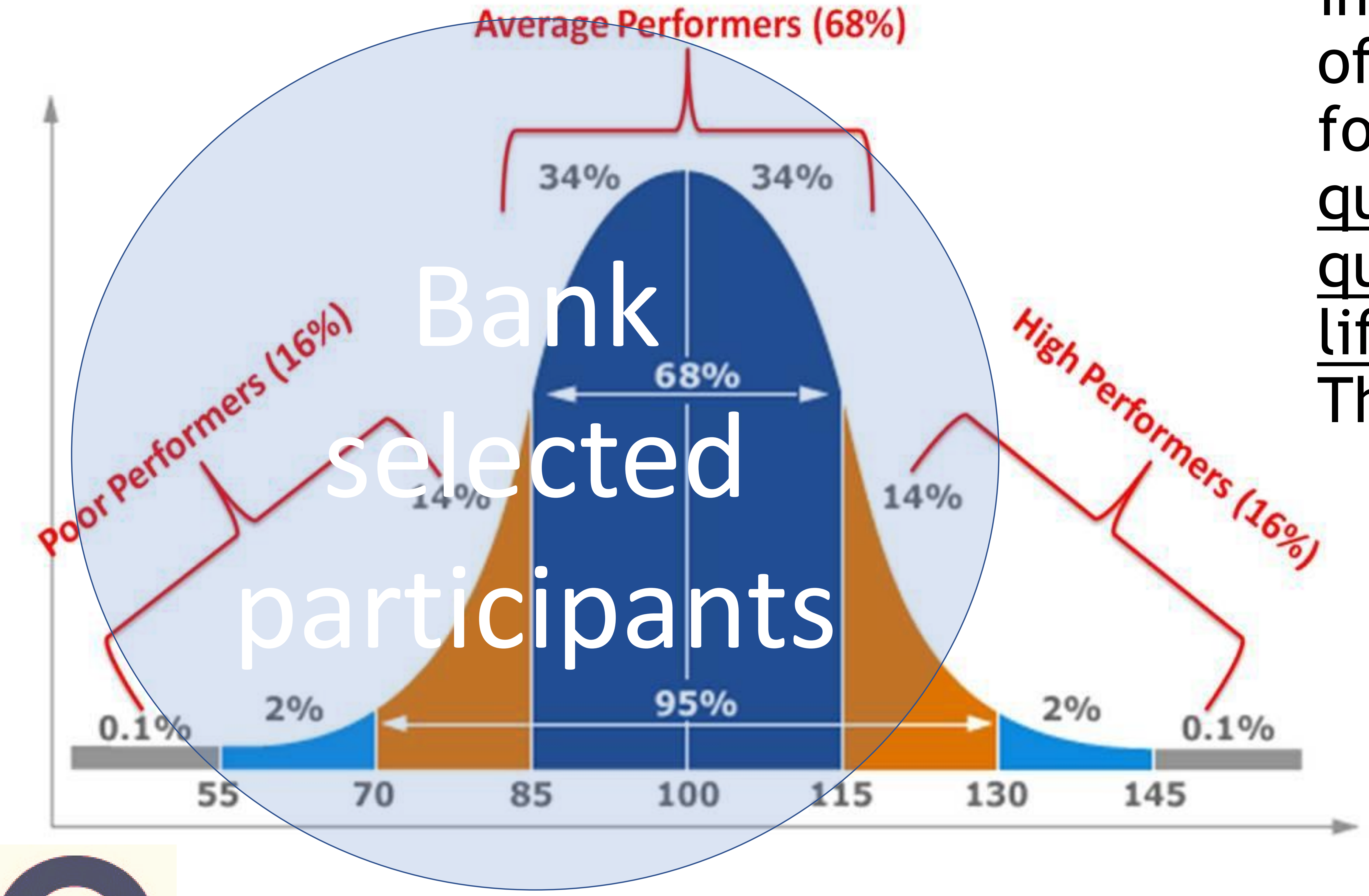
Country	Residential real estate foreclosure sales	Sold below their market's average %	Difference (Foreclosure price v market average price)	BestAgent vs Bank Agent Average Sale Price Differential
Spain	940	69	-€47,441	€82,369
Ireland	98	66	-€33,150	€64,078
Germany	1660	71	-€29,009	€67,882
France	1340	71	-€42,115	€87,766
England	1120	68	-£28,540	£79,887
Canada	750	78	-$43,899	$87,455
Austrailia	500	77	-$47,592	$157,907
America	6560	79	-$52,572	$104,741

The best fit, a binomial distribution model, allowed us to figure the probability of observing a specific number of ‘best’ outcomes when the process was repeated against the selling agent with the outcome for a given selling agent is either success (above market) or failure (below market) benchmark result against the market’s average and other market participants.

$$P\ (X\ \text{“successes”}) = \frac{n!}{x!\,(n-x)!}\,p^{x}\ (1-p)^{(n-x)}$$

"n" denotes the number of observations or agent performance econometrics, of which we discussed previously, Days on Market, List Price to Sale Price Differential, $ per SQM, and "x" denotes the number of "successes" or events of interest occurring during "n" observations. The probability of "success" or occurrence of the outcome of interest is indicated by "p".

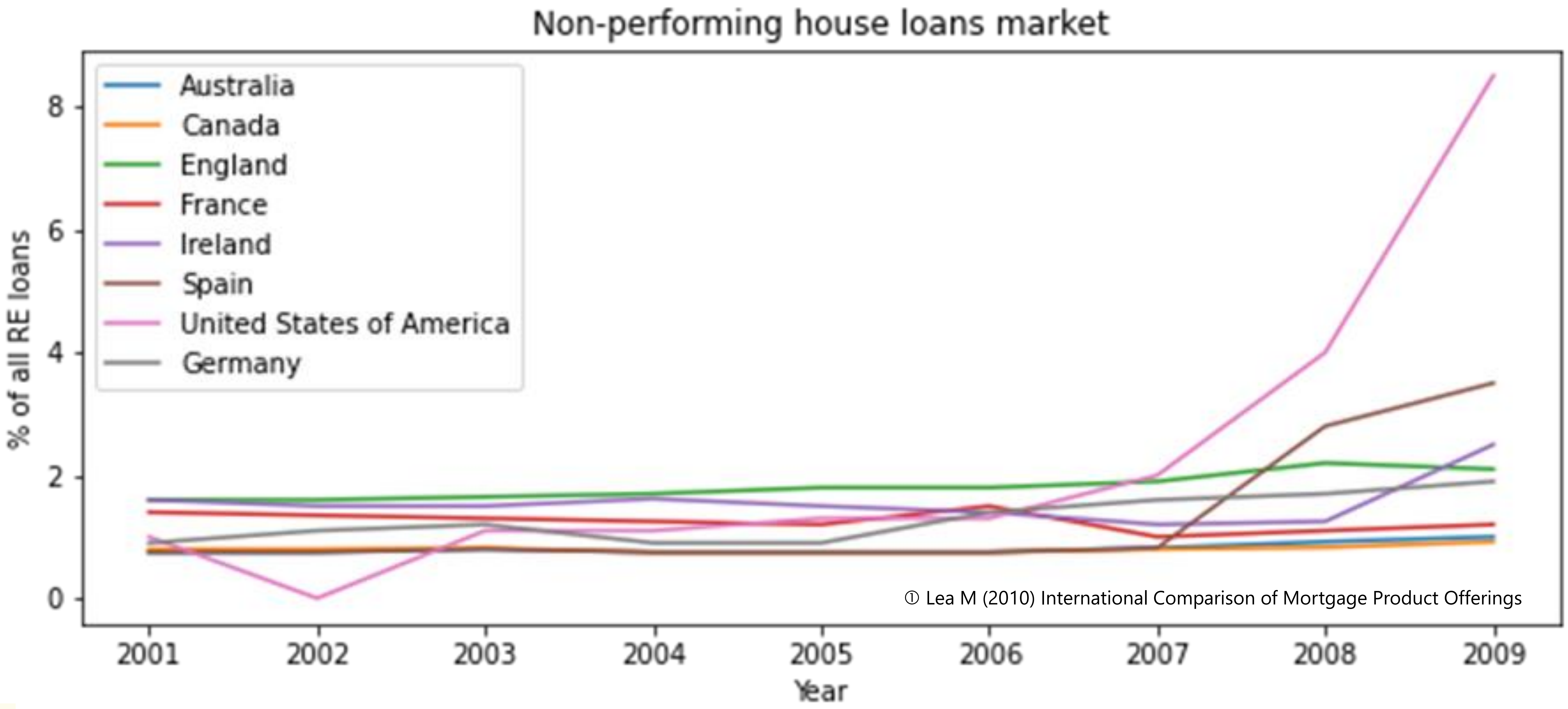
Problem Identified,Isolated and Solved



The opportunity for improvement is in the retail banks selection process of the market participant selling the foreclosed property. <u>It is the only qualitative part of an entirely quantitative process in the product lifecycle of a residential mortgage.</u> The selection process is:

Unsystematic.
Biased.
Nonreplicable.
Inefficient.
Achieves results below market.

Research, Observations and Comparisons



① Lea M (2010) International Comparison of Mortgage Product Offerings

Research, Observations and Comparisons

# of Properties, Bank Stamp Duty, Market Stamp Duty and BestAgent Stamp Duty by State

200
150
100
50
0
# of Properties
40K
20K
0K
148
121
78
68
67
9
8
1
44K
44K
44K
5K
8K
4K
17K
13K
10K
16K
12K
10K
34K
26K
23K
25K
19K
15K
12K
9K
11K
7K
4K
Queensland
Western Australia
Victoria
South Australia
New South Wales
Northern Territory
Australian Capital Territory
Tasmania
State
● # of Properties
● Bank Stamp Duty
● Market Stamp Duty
● BestAgent Stamp Duty

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Machine learning-based approaches for automatic data validation and outlier control of loan microdata in the Bank of Russia[1]

Dmitrii Diachkov,
Central Bank of the Russian Federation

[1] This presentation 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 central banks and other institutions represented at the event.

Machine learning-based approaches for automatic data validation and outlier control of loan microdata in the Bank of Russia

Dmitrii Diachkov[1]

Abstract

Validation of loan microdata is an outstanding issue in central bank statistics, and the challenge is magnified by increasing variability and heterogeneity of the underlying data. In this work, an application of machine learning approaches in large loan datasets is discussed. We identified a set of tools, such as gradient boosting, neural networks, random forests, that can be used to enhance the quality of microdata on loans available in the Bank of Russia. Ensemble methods and pre-processing techniques in RStudio are used to explore, analyse and determine outliers and potentially erroneous clusters of data on loans and borrowers. Based on the ensemble of machine learning algorithms, the toolkit efficiently reduces the variance of predictions and, in some cases, outperforming base classifiers (logistic regression) is expected to be very useful for quality control. The results reveal that highly atypical groups may be identified, providing additional insight for further scrutiny, methodological research, and the development of statistical indicators.

Keywords: machine learning, data validation, outliers, bank loans, data quality

JEL classification: E51, E58, G18, G21, C81

[1] Dmitrii Diachkov is an Expert Team Leader in Statistics Department at the Bank of Russia and a Statistical Systems student of NOVA IMS, Lisbon, Portugal. I would like to thank my colleagues from the Bank of Russia's Statistics Department for their valuable ideas, support and assistance. The views expressed are those of the author and not necessarily those of the Bank of Russia or NOVA IMS.

Contents

Introduction

Loan-level information is one of the primary data sources for a wide variety of analytical tasks performed in central banks (Morandi, Nicoletti, 2017). The use of highly granular data (microdata) opens up new frontiers for analyzing the dynamics and structure of the credit market (Santos, 2013).

Microdata on loans can be used for various purposes like a compilation of monetary statistics (Dyachkov, Nurimanova, 2017), sectoral credit risks analysis, or supporting decisions on the countercyclical capital buffer (Carstens, 2016) in order to analyze actual distributions of indicators, and not their distorted representations by aggregation (Aaron & Hogg, 2005).

Despite all the advantages of microdata, some factors hinder their full use (Livraga, 2019), including possible quality problems (Osiewicz, Fache-Rousova & Kulmala, 2015). Even though the data obtained from administrative sources are, in fact, of relatively higher quality (Crato & Paruolo, 2018), in order to maximize their usefulness, verification procedures of the incoming data should be built. The use of machine learning will reduce the cost of producing statistics and improve its quality (Tam & Clarke, 2015).

The paper analyzes existing approaches to classification in large volumes of microdata using machine learning methods, proposes new ways to solve applied problems for analyzing large volumes of data to improve their quality and search for atypical values. The main advantages and disadvantages of using decision trees, regression trees, and random forests in classification problems are considered, several models for the practical application of machine learning methods are proposed, including gradient boosting methods. Particular attention is paid to the problems of balancing the training samples of microdata.

The microdata sets on loans available to the Bank of Russia contain more than 150 attributes and cover 100% of all loans to legal entities and individual entrepreneurs provided by banks in Russia.

Due to limited resource capacity and rapidly increasing data volumes, analysis of the reliability of a large and diverse set of data can become a tricky task; hence it simultaneously allows the application of machine learning methods to amplify the efficiency of data quality control and decisions made on their basis.

Interpretability, controllability, the possibility of automatic selection of informative features of decision trees, and regressions were the reason for their use as the primary tool for efficient classification of processing large amounts of data, searching for atypical values for subsequent filtering and identifying erroneous values in categorical variables. Current research is made to improve data quality and is based on a variety of disciplines, and represents a rich set of scientific and technological tasks for statisticians.

Literature review

Data availability alone does not guarantee that assumptions, derived from this data, are correct as well as it does not guarantee that data management functions are efficient (Manjunath, Hegadi, Ravikumar 2010).

The issue is not new, but it would be safe to estimate that a large-scale numeric database without critical judgment can have an error rate of 2-5% (Dong et al., 2002); hence it is not clear what can be considered as acceptable accuracy. According to Karr et al. (2006), some time ago, data quality was just a scientific problem rooted in measurement errors and research uncertainty.

Nevertheless, in today's world of high dimensional data and complex economic policy decisions, data quality problems can create significant economic losses (Madhikermi et al., 2016) and short-term fixes (Lee et al., 2006), and organizations tend to underestimate the consequences that in fact may vary "from significant to catastrophic" (Gudivada, Apon & Ding, 2015). Errors in microdata increase variance and create biased results (Eurostat, 2020).

Consumers can evaluate data quality based on their objectives (Lee et al., 2006). The same datasets may be used for multiple tasks that need different quality characteristics (Batini et al., 2015 and Aljumaili, 2016). In addition, if task requirements change over time, some quality characteristics might change (Eurostat, 2007). Variables in large microdata databases that are not of particular interest for data owners may be of lower quality (Crato & Paruolo, 2019). Therefore, providing high-quality data means tracking a constantly moving target. Perrella & Catz (2020) conclude that IT tools should not only provide consistency checks but assess the plausibility of the data.

Gomolka et al. (2021) conclude that it is essential to track the results of data quality checks to make sure that any modifications conducted to enhance data quality for researchers do not affect the contents of data. Maintaining the high quality of a microdata register might become a challenging task because of small but frequent objects, such as taxpayers' data (Gavin, 2021).

All these challenges can be resolved with appropriate work-process organization and modern computer science methods (Crato & Paruolo, 2019), taking into account a context-based nature of data quality (Batini et al., 2015).

The main goal of modern microdata quality control is protection from incorrect or invalid information (Crato & Paruolo, 2019) and missing data (Smith et al., 2018). Eliminating and rectifying these quality gaps should be one of the primary concerns for statisticians (Perrella & Catz, 2020), and internal researchers need microdata to be as precise as possible to achieve reliable results (Domingo-Ferrer & Blanco-Justicia, 2021).

We live in the "era of big data" and collecting such data requires tremendous effort, and publication is often delayed. However, there has been an explosive growth in the amount of data available to use (Doerr, Gambacorta & Garralda 2021). New data collection and dissemination models enable real-time analysis of massive amounts of data.

The growing use of artificial intelligence (AI) and machine learning applications makes it even more difficult to ensure data quality in organizations (Janssen et al., 2020) as well as the introduction of real-time streaming platforms that continuously transmit large amounts of data to corporate systems. In addition, data quality now often needs to be managed in combination with on-premises and cloud systems, and hence data quality enhancement tools should be embedded in the process of quality management (Kropf, 2020).

Usually, microdata quality control is designed by implementing automatic checking procedures (logical or mathematical rules) during the data collection (Zambuto et al., 2020).

The literature on data quality has not yet paid proper regard to the design of methodologies that can provide automatized verification of large multivariate datasets (Farnè & Vouldis, 2018). However, effective ML applications require high-quality datasets. For example, data that is distorted by outliers can result in non-convergence in ensemble learning and a dramatic reduction of quality in prediction (Gudivada, Apon & Ding, 2015).

Coeuré (2017) states that carrying out automatic checks, based on machine learning techniques, and AI is one of the ways to ensure that data remain of high quality. According to Zambuto et al. (2020), machine-learning application to microdata will be more efficient when designed models are backed up by pre-determined relationships, such as accounting rules. The predictive power of all ML models improves when the length of the relationship between attributes and objects increases, and the non-linear relationship between variables in a dataset is better mined my machine learning (Gambacorta et al., 2019).

Using data on proprietary loan transactions from a leading fintech company in China, Gambacorta et al. (2019) show that while regular models perform well in ordinary times, machine learning models are way better able to predict extraordinary events. It is interesting to note that a possible reason for that behavior is that machine learning can better react to the non-linear relationship between factors. Lukauskas & Ruzgas (2021) used various techniques like artificial neural networks, XGBoost, LightGBM, Catboost to predict borrowers' default taking into account computational time. The recent research by Severino & Peng (2021) revealed that ensemble-based random forests, gradient boosting, and neural networks achieve the most effective results, overcoming base classifiers such as logistic regression. Likewise, Dou et al. (2019) analyzed online fraud and applied the XGBoost model to predict fraud using a variety of feature sets, achieving more than 99% level of accuracy, taking into account the prediction class balance. Odegua (2020) has successfully used the XGBoost for bank loan default prediction.

Regular decision trees are intuitive because the model visualizes a decision scheme (Wang et al., 2020). Many researchers support the idea because of the high interpretability and flexibility in feature selection of decision trees (Lee et al., 2006; Kao et al., 2012).

Divakar & Chitharanjan (2019) explored credit card fraud data using several boosting methods and concluded that the XGBoost algorithm is the best model among others considered, like AdaBoost and Gradient Boost. Raju (2021) shows that the method may outperform random forests and OLightGBM models, while Trisanto et al. (2021) showed that imbalanced data might be considered an issue to this method.

Manjeet et al. (2018) compared traditional classification models with neural networks concluded that the neural network other models to predict loan default. Li Ying (2018) compared three model families: random forest, logistic regression, and SVM on bank loan data. The results show that random forests generally perform better. This result corresponds with a study of COŞER, Maer-matei & ALBU (2019) that compared LightGBM, XGBoost, Logistic Regression, random forest to evaluate loan probability default and recognized random forest as an optimal classifier for the task. The learning rate is fast and can be applied to large-scale datasets (Wang et al., 2020)

According to Cheng (2021), XGBoost training requires cumbersome setting and adjustment; hence Koduru et al. (2020) offer even more complicated ML tools such as random forest + XGBoost for loan application scoring.

Motivated by all presented research, we explore microdata on loans in the Bank of Russia and seek possible application of ML methods to enhance data quality of the essential attributes.

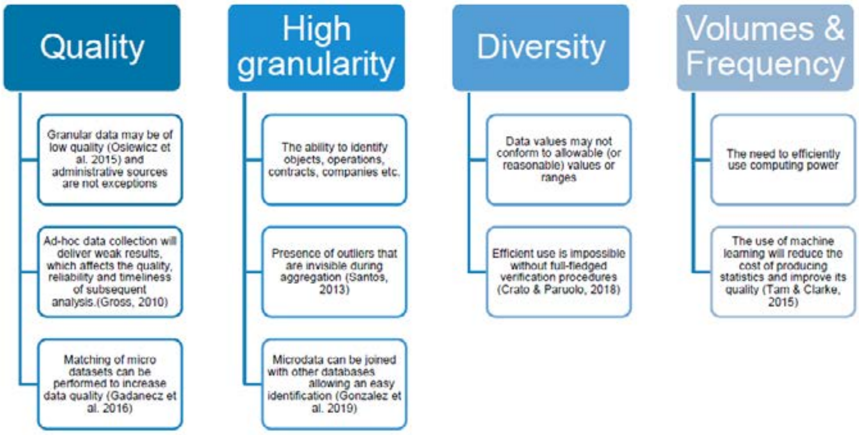
Loan microdata quality control in the Bank of Russia

Magnified by variability and heterogeneity of the underlying data, validation of loan microdata is an outstanding issue in the Bank of Russia.

Practical difficulties in collecting and processing microdata, non-transparent methodology, lack of quality assessments, and instability of "big data" do not allow putting these new sources of information on a par with data used in normal statistical business processes. This situation requires conceptual comprehension and evolutionary development with the approbation of approaches on individual projects.

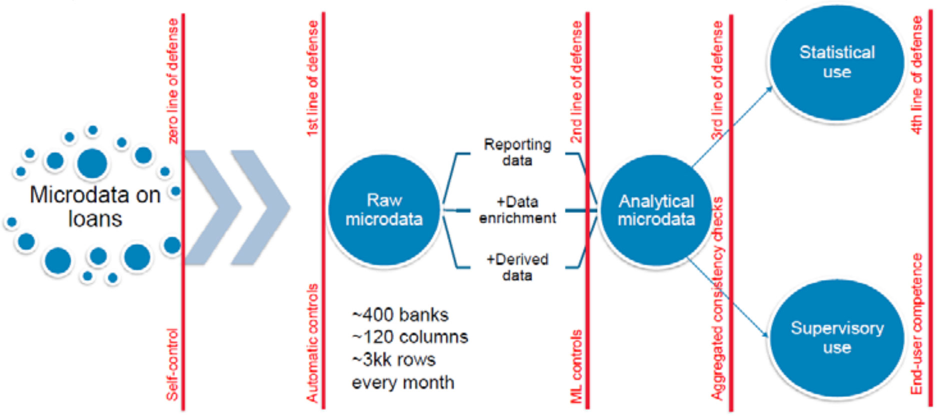
The accumulated experience of working with loan microdata in the Bank of Russia's Statistics Department is devised based on four main data quality components presented in Figure 1.

Figure 1: Accumulated experience of working with loan microdata


Quality
Granular data may be of low quality (Osiewicz et al. 2015) and administrative sources are not exceptions
Ad-hoc data collection will deliver weak results, which affects the quality, reliability and timeliness of subsequent analysis.(Gross, 2010)
Matching of micro datasets can be performed to increase data quality (Gadanecz et al. 2016)
High granularity
The ability to identify objects, operations, contracts, companies etc.
Presence of outliers that are invisible during aggregation (Santos, 2013)
Microdata can be joined with other databases allowing an easy identification (Gonzalez et al. 2019)
Diversity
Data values may not conform to allowable (or reasonable) values or ranges
Efficient use is impossible without full-fledged verification procedures (Crato & Paruolo, 2018)
Volumes & Frequency
The need to efficiently use computing power
The use of machine learning will reduce the cost of producing statistics and improve its quality (Tam & Clarke, 2015)

The conceptional scheme of loan microdata pipeline in the Bank of Russia is presented in Figure 2. The best way to apply ML for data quality is before or when forming analytical microdata. All data items should be enriched at this stage, and derived columns calculated and applying ML controls may potentially form the 2nd line of defence (after automatic controls, but before aggregated consistency checks).

Figure 2: Possible implementation of ML controls to data quality control scheme

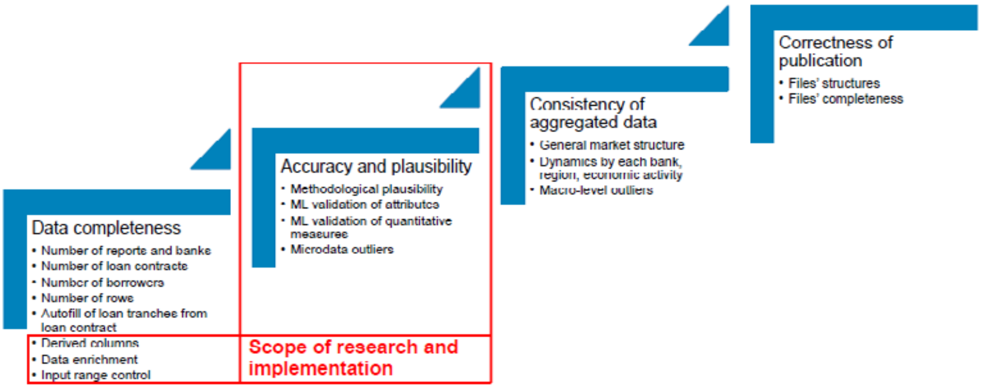


Examples of errors that arise in the process of collecting information are the following:

- mistyping errors;
- technical errors (i.e., numerical field that contains 0s, that can be economically interpreted, for missing values);
- errors of database merges (i.e., a big bank with comprehensive branch coverage collects data separately from each branch and unite that data in its' warehouse)
- unit conversion errors (i.e., thousand of units versus units);
- errors of source data interpretation (i.e., misreading of the original documents);
- errors in the compilation of the metadata (i.e., the supporting information is incorrectly entered or interpreted);
- errors in the underlying data (i.e., the results presented in the original data source are incorrect or misleading).

All the error types mentioned above may be found at any stage of data processing. The possible niche for implementing first ML models and their role in verification workflow is presented in Figure 3 below.

Figure 3: Scope of first possible ML controls in a context of data verification


Data completeness
• Number of reports and banks
• Number of loan contracts
• Number of borrowers
• Number of rows
• Autofill of loan tranches from loan contract
• Derived columns
• Data enrichment
• Input range control
Accuracy and plausibility
• Methodological plausibility
• ML validation of attributes
• ML validation of quantitative measures
• Microdata outliers
Scope of research and implementation
Consistency of aggregated data
• General market structure
• Dynamics by each bank, region, economic activity
• Macro-level outliers
Correctness of publication
• Files' structures
• Files' completeness

The main goal of adding the ML component to this workflow is to explore the possibility of applying relatively simple machine learning methods to improve the overall quality of microdata and decisions based on them in the Bank of Russia.

Limited human abilities (analysts or statisticians) to extensively analyze the reliability of large and complicated datasets, furthermore changing over time, inevitably leads to ML applications. The effective application should meet the following criteria:

Criteria for effective ML application

Table 1

Criterion	Purpose
High interpretability of results	All participants of data quality management and end-users should be able to understand the principles of control and resulting outputs.
Moderate ease of implementation	Applications should be relatively simple so staff members without strong IT knowledge could implement them in the field of competence
Moderate process control	Underlying banking data is constantly changing so applied techniques should be amendable and provide relatively robust results when methodology of data collection changes or new products emerge
High scalability and reusability	Good models or model patterns should be re-used (if appropriate) for similar data types as well as be independent of training dataset size.
High automation capability	Quality control needs to be designed that way so it can be executed on server instance by timetable or by request.

Empirical applications of ML quality controls

In order to create a new line of data quality validation, the following datasets were used as material for approbation: Banks' reporting form 0409303 "Information on loans granted to legal entities and individual entrepreneurs" with microdata on loans, annual accounting data, Statistical Registry on companies provided by State Statistics Service, State Registry of SMEs provided by Federal Tax Service.

For proper validation of underlying data for loan statistics in the Bank of Russia, it is vital to pay close attention to the industrial classification of borrowers as one of the main focal points as well as SME classification and main loan parameters. The design of the first models should be targeted at main grouping variables or attributes used to form input data arrays.

1. Validation of balance sheet codes for borrowers with decision trees, XGBoost, and logistic regression

Balance sheet code is a simple and understandable attribute for economists that can be used to filter borrowers by type and industry. Information about the balance sheet codes is reported to the Bank of Russia by banks directly. Balance sheet codes may be determined by banks based on irrelevant data or just by mistake. Another type of mistake is technical errors during report preparation or submission.

The bank should assign the balance sheet code to its borrower based on a limited list of parameters, such as type of loan, business entity forms, and economic activity, so according to the factors mentioned earlier, we can verify balance sheet codes with the data of public registers or other data sources. Generally, we would like to have a probability score of correct balance sheet code for each item as an output from our ML model.

Building on Wang et al. (2020), Lee et al. (2006), Kao et al. (2012), we developed a set of regular decision trees to promote interpretable and easily visualized models for binary classification. The rationale behind these models was based on the idea that a combination of loan type, economic activity, business entity form, and debt sum may be used to answer the question of whether the reported balance sheet code is valid or not.

Standard balance sheet codes that are used for loan statistics and hence to be validated in a dataset with 1056k observations:

«452» - Loans to non-financial companies;

«454» - Loans to individual entrepreneurs;

«451» - Loans to financial companies.

Both «452» and «454» sub-samples did not require additional transmutations, and the dependent variable was evenly distributed in the training and test samples. We developed a set of 15 models (solutions for account «452» presented in Figure 4 in Appendix) with different specifications to seek the best dependence between response and explanatory variables.

Since the training and test samples are balanced (predictable classes are evenly distributed), standard measure like accuracy (1) is an excellent metric of model quality.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{ (1)}$$

As many models have led to relatively comparable results with high accuracy, we decided to introduce computational time as a penalty for efficiency, normalized by the number of explanatory variables (Figure 5a in Appendix). Hence, relative cost-efficiency for the evaluation purpose of several models with balanced classes could be formulated as (2):

$$\boldsymbol{Relative\ Efficiency} = \frac{\boldsymbol{E(Accuracy)}}{\boldsymbol{E(Time)}} \ (2),$$

We also tested XGBoost (Figure 5b in Appendix) and logistic regression models to predict the same balanced classes, and it achieved the same results in terms of efficiency but with a longer computational time. Results are presented in Table 2.

Summary of results for decision tree, XGBoost, and logistic regression approaches to classification task for account code “452”

Table 2

Model type	Mean accuracy	Mean precision	Mean recall	Mean time of fit and predict (per 1kk rows), seconds	Relative efficiency score
Decision tree	0,989	0,987	0,990	17	1,939
XGBoost	0,988	0,993	0,984	45	0,731
Logistic regression	0,989	0,99	0,989	113	0,291

During the analysis of the model output, we did a deviation analysis. It showed that the decision tree classified 0.66% or 1331 loans as loans of non-financial companies (code «452»), but in fact, they are reported as other codes (FP, for example, «453» loans to non-residents or «451» loans to financial companies). Additionally, 0.44% or 983 loans were classified by the decision tree as loans that should be recorded as non «452», but in fact, they are recorded in «452» accounts (FN, mainly financial companies, which should be reflected with code «451»). These deviations should be considered as rare events or errors and subsequently scrutinized. The reason behind this atypical coding may be one of the following factors:

- Methodological features, which should lead to the methodology refinement;
- Reporting errors, which leads to informing the bank about the error and asking for report re-submission;
- Accounting errors, which require to inform the bank about the error and elimination in the future.

Luckily, we have not faced significant challenges handling imbalanced classes. While codes «452», «454» represent almost a half of all observations each, «451», corresponding to financial companies, was presented only in 1-2% of observations. The dependent variable is not evenly distributed in the training and test samples.

A similar set of 15 models were derived, and the same computational limitations were implemented as well as in the case of balanced samples, but they were adapted for imbalanced samples and normalized by the number of explanatory variables. Hence, relative cost-efficiency for the evaluation purpose of several models with imbalanced classes could be formulated as (3):

$$\boldsymbol{Relative\ Efficiency2} = \frac{\boldsymbol{E(F1\_score)}}{\boldsymbol{E(Time)}}\ (3),$$

Precision (4) and Recall (5) metrics do not depend on the ratio of classes and therefore are applicable in conditions of imbalanced samples. The F1 score is a good balance between these two metrics.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}\ (4)\quad \boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}\ (5)$$

$$\boldsymbol{F1\_score} = \frac{\boldsymbol{2*Precision*Recall}}{\boldsymbol{Precision+Recall}}\ (6)$$

The affiliation of the legal entity to the finance industry «K» and specific loan types both determine the belonging to the code «451». Despite the high accuracy (99.1%), the confidence in the model was significantly lower since the training and test samples were unbalanced. Accuracy in this situation is an incorrect metric. However, the disproportion of classes in the task does not influence the overall performance presented in Table 3 below.

Summary of results for decision tree, XGBoost, and logistic regression approaches to classification task for account code "451"

Table 3

Train data type	Model type	Accuracy, %	Fit and predict time (seconds per 1kk rows)	F1_score, %	Relative efficiency score 2
Downsampled (30k rows)	Decision tree	98,98%	9,976	99,48%	0,100
	Logistic regression	99,35%	62,34	99,67%	0,016
	XGBoost	99,35%	50,606	99,67%	0,020
Original - imbalanced (845k rows)	Decision tree	99,43%	14,081	99,71%	0,071
	Logistic regression	99,44%	136,62	99,71%	0,007
	XGBoost	99,32%	38,783	99,65%	0,026
Upsampled (1658k rows)	Decision tree	98,99%	15,331	99,48%	0,065
	Logistic regression	99,39%	125,16	99,69%	0,008
	XGBoost	99,35%	70,38	99,67%	0,014

We can conclude that in this particular case, class imbalance does not affect model performance. In such cases, we should shift towards down-sampled versions because of fewer computational expenditures.

2. Validation of interest rates data with XGBoost

Interest rate statistics require only high-quality data because a single outlier can dramatically change weighted average rates. To verify outliers, we decided to develop an algorithm to define a pattern of ordinary data items. XGBoost provided a quick and effective solution.

Using data on 421k loans as a training dataset and 105k loans as a test dataset, the XGBoost algorithm was implemented and benchmarked with neural net and regular linear regression (Table 4).

Comparison results

Table 4

Model	RMSE
XGBoost	2,04
Neural net (caret and nnet)	3,25
Linear Regression	3,85

Nine variables were used as an input (Figure 6 in Appendix), but subsequently, the number was reduced due to little importance of some variables.

XGBoost model showed better results, dynamics of RMSE of each training repetition, and testing presented in Figure 7 (Appendix). It can be used to restore omissions in interest rates and search for outliers. The corresponding actual VS XGBoost predicted values for interest rates by loan type (as a top-importance variable) are presented in Figure 8 in Appendix.

Another critical task for interest rate statistics is a breakdown of loans by the size of the borrower's business. Hence, XGBoost multiclass classifier was used to predict the size of SME companies: micro, small, medium, or non-SME.

Two approaches were used: XGBoost on balanced train data and imbalanced. The results were very similar, but the balanced sample provided more insight into the importance of features. However, if the importance of features does not matter, one can apply the technique without up-sampling (to save computational resources) or down-sampling (because it lessens the model's exposure to non-standard objects). Confusion matrices and feature importance structures are presented in Figure 9 in Appendix.

Finally, we conclude that XGBoost is a powerful tool with high speed and very high accuracy on millions of rows. Performance on the imbalanced sample is 2% less accurate than on the upsampled and balanced (90% VS 92%).

3. Validation of SME status with neural networks, random forests and logistic regression

Belonging to the SME Registry (provided by Federal Tax Service) defines a borrower's business size. This attribute is crucial for the analysis of the economic situation. However, small and insignificant companies may be excluded from the Registry for various reasons. We need to be able to check the validity of any given status.

Large borrowers have more assets and, accordingly, apply for more significant amounts of loans. SME borrowers are usually smaller in business size and balance sheet.

To solve this task, we have built several simple neural networks that classify companies as SMEs and non-SMEs based on various quantitative indicators characterizing the loan and the borrower. The dependent variable is not evenly distributed in the training and test samples (75% of borrowers are SMEs), but this issue was solved with down-sampling. Down-sampling was the only option because of already vast amounts of data items. We compared 20 different compositions of neural networks. The best neural network consisted of 2 hidden layers (with 4 and 3 neurons respectively – figure 11) and predicted SME status with 90,4% accuracy and F1_score of 80% (figure 12). This model cannot be considered ready for productional usage in data quality control because of the time-consuming training process and lack of precision and recall, which results in a low F1_score with many false-positive SME-statuses, so the model needs to be re-designed.

We approached the same task with randomForest and logistic regression and achieved the same accuracy and F1_score results (randomForest confusion matrix is presented in Figure 13 of the Appendix), but faster. Additional implementations were made with six explanatory variables on random sampling from up-sampled (balanced) training data. In terms of computational speed, the efficiency of random forest is two times higher, while results are even slightly better. Results are presented in Table 5 below. Traditional classifier logistic regression has beaten neural networks and random forests in terms of result/speed ratio.

Comparison results

Table 5

Model	Accuracy	F1_score	Computation time (per 1kk rows), seconds
Neural net	90,4%	80,1%	757
Random forest	92,8%	81,4%	318
Log regression	93,3%	83,1%	108

Conclusions

The main lessons are following:

Interpretability, controllability, the possibility of automatic selection of informative features of decision trees, and regressions were the reason for their use as the primary tool for efficient classification of processing large amounts of data, searching for atypical values for subsequent filtering, and identifying erroneous values in categorical variables.

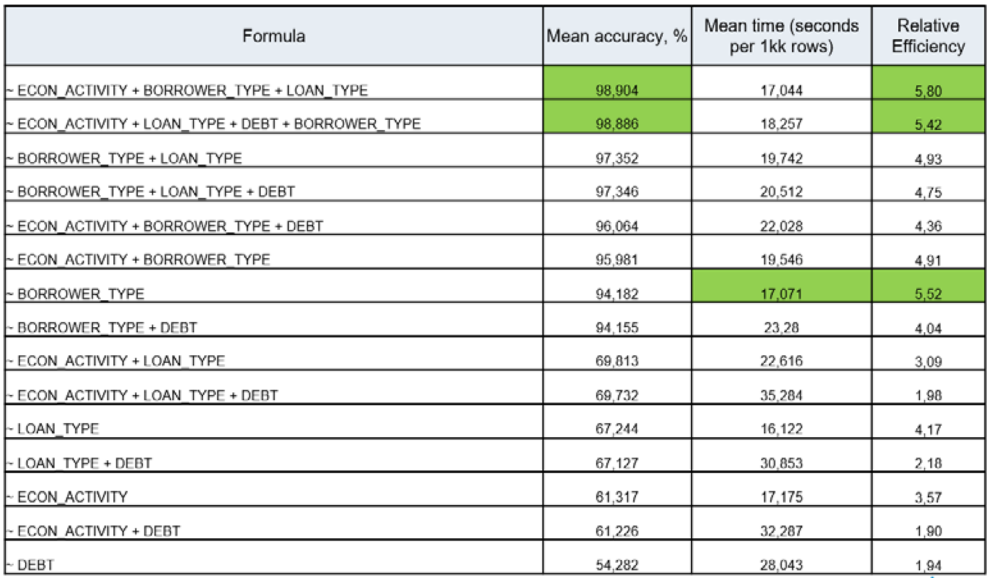
Due to human disabilities, to analyze the reliability of a large and diverse set of data, expanding the field of applied machine learning methods will increase the quality of data and decisions made on their basis.

When solving classification problems, metrics should be monitored carefully and problems solved under business logic. With unequal classes, metrics should be selected carefully, and up-sampling or down-sampling applied when necessary.

Simpler models often give more balanced and correct results during cross-validation on test data.

Appendix

Figure 4. Efficiency of 15 different specifications for decision trees, predicting the class «452» (non-financial companies) after 100 repetitions (green highlight – best models).



Formula	Mean accuracy, %	Mean time (seconds per 1kk rows)	Relative Efficiency
~ ECON_ACTIVITY + BORROWER_TYPE + LOAN_TYPE	98,904	17,044	5,80
~ ECON_ACTIVITY + LOAN_TYPE + DEBT + BORROWER_TYPE	98,886	18,257	5,42
~ BORROWER_TYPE + LOAN_TYPE	97,352	19,742	4,93
~ BORROWER_TYPE + LOAN_TYPE + DEBT	97,346	20,512	4,75
~ ECON_ACTIVITY + BORROWER_TYPE + DEBT	96,064	22,028	4,36
~ ECON_ACTIVITY + BORROWER_TYPE	95,981	19,546	4,91
~ BORROWER_TYPE	94,182	17,071	5,52
~ BORROWER_TYPE + DEBT	94,155	23,28	4,04
~ ECON_ACTIVITY + LOAN_TYPE	69,813	22,616	3,09
~ ECON_ACTIVITY + LOAN_TYPE + DEBT	69,732	35,284	1,98
~ LOAN_TYPE	67,244	16,122	4,17
~ LOAN_TYPE + DEBT	67,127	30,853	2,18
~ ECON_ACTIVITY	61,317	17,175	3,57
~ ECON_ACTIVITY + DEBT	61,226	32,287	1,90
~ DEBT	54,282	28,043	1,94

Figure 5a. Best decision tree structures and confusion matrices for predictions of class «452» (non-financial companies)

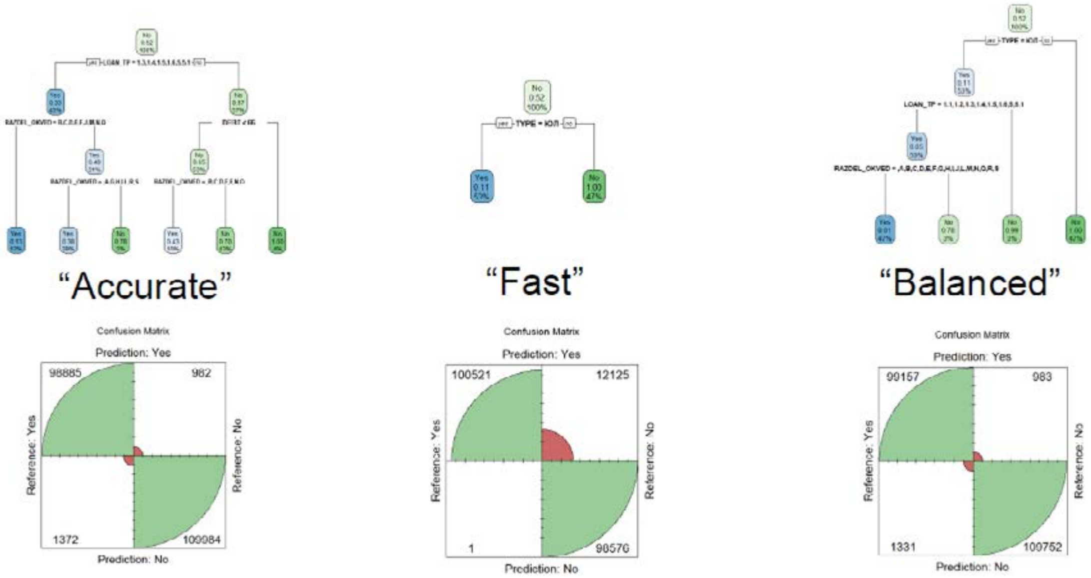


Figure 5b. Confusion matrix (left) and feature importance for XGBoost model for predictions of class «452» (non-financial companies)

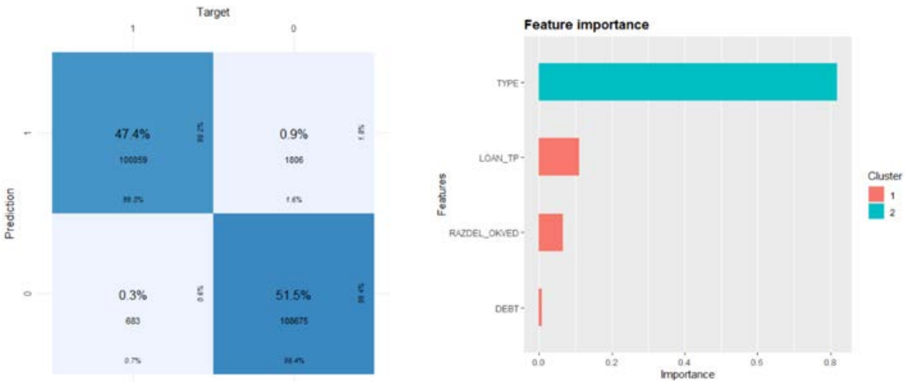


Figure 6. Importance matrix of XGboost models for interest rate prediction

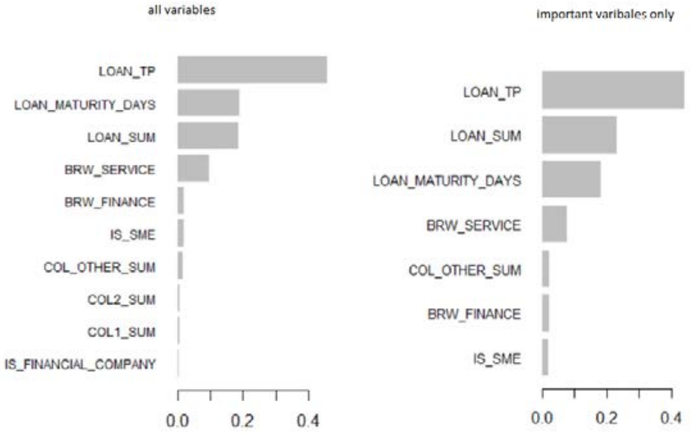


Figure 7. RMSE for XGboost interest rate prediction (train and test)

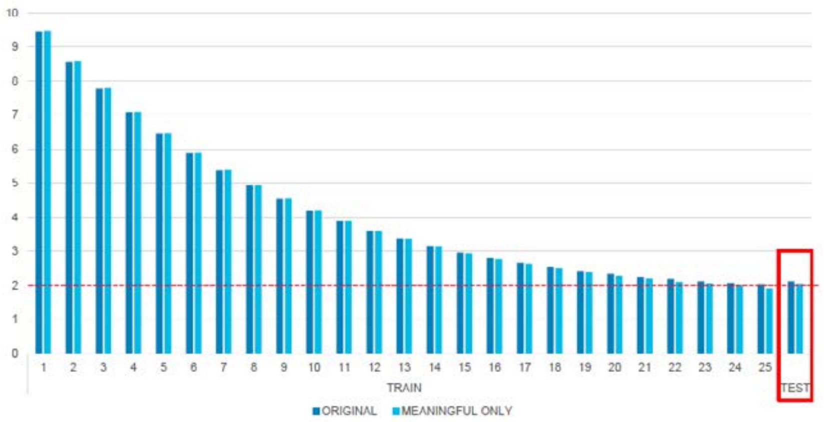


Figure 8. Correspondence of XGBoost predicted (y_pred) and actual (y) values for interest rates by loan type (as top-importance variable)

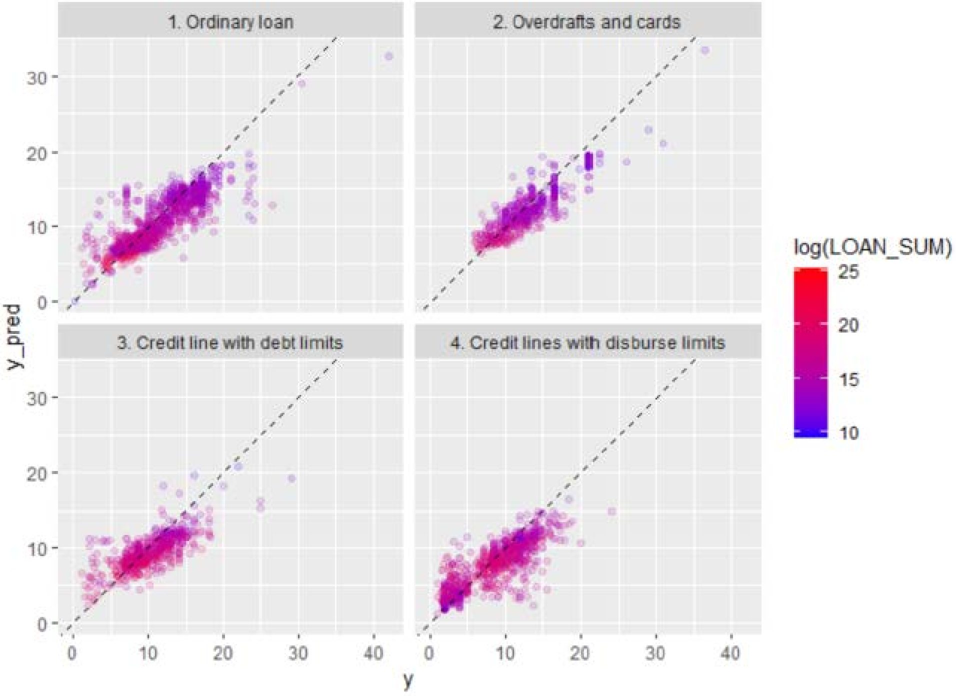


Figure 9. Confusion matrices for XGBoost predictions of a company's size (where 1 – is micro, 2 – small, 3 – medium, 4 – non-SME)

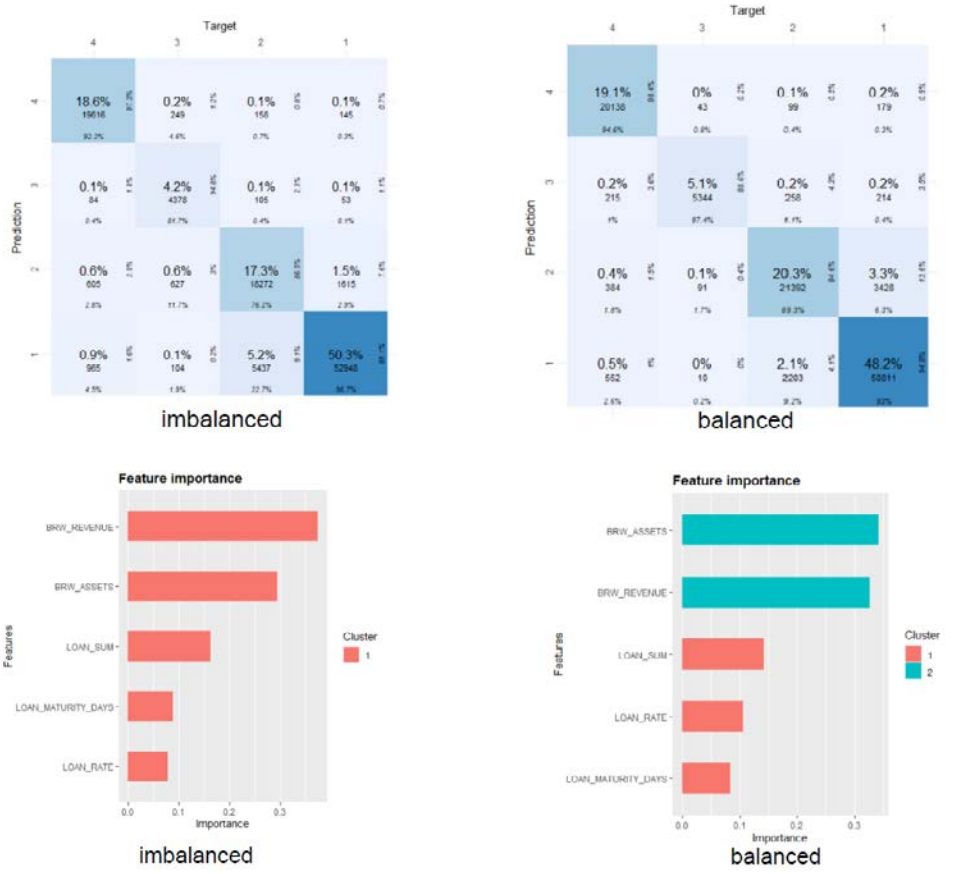
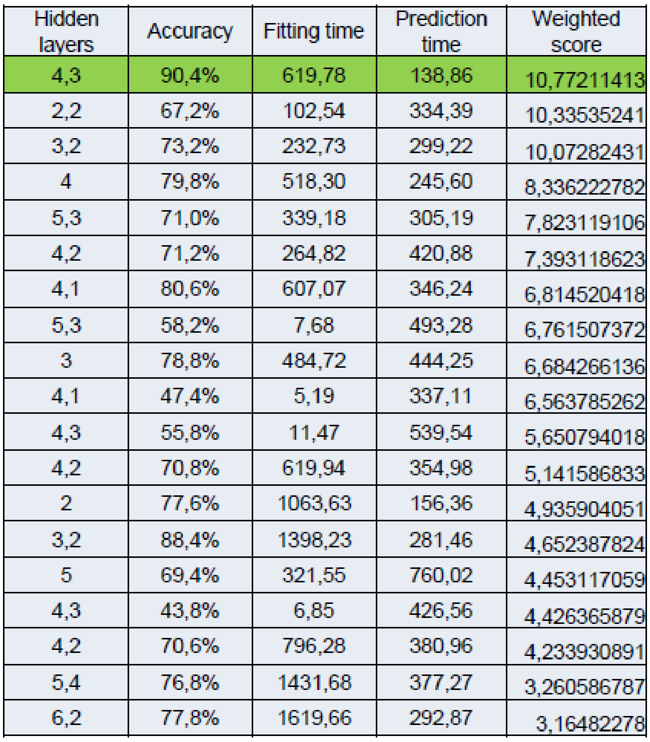


Figure 10. Accuracy and fitting time of tested neural networks with different composition of layer and neurons to predict SME status (green highlight – the best model).



Hidden layers	Accuracy	Fitting time	Prediction time	Weighted score
4,3	90,4%	619,78	138,86	10,77211413
2,2	67,2%	102,54	334,39	10,33535241
3,2	73,2%	232,73	299,22	10,07282431
4	79,8%	518,30	245,60	8,336222782
5,3	71,0%	339,18	305,19	7,823119106
4,2	71,2%	264,82	420,88	7,393118623
4,1	80,6%	607,07	346,24	6,814520418
5,3	58,2%	7,68	493,28	6,761507372
3	78,8%	484,72	444,25	6,684266136
4,1	47,4%	5,19	337,11	6,563785262
4,3	55,8%	11,47	539,54	5,650794018
4,2	70,8%	619,94	354,98	5,141586833
2	77,6%	1063,63	156,36	4,935904051
3,2	88,4%	1398,23	281,46	4,652387824
5	69,4%	321,55	760,02	4,453117059
4,3	43,8%	6,85	426,56	4,426365879
4,2	70,6%	796,28	380,96	4,233930891
5,4	76,8%	1431,68	377,27	3,260586787
6,2	77,8%	1619,66	292,87	3,16482278

Figure 11. Best neural network for SME size prediction

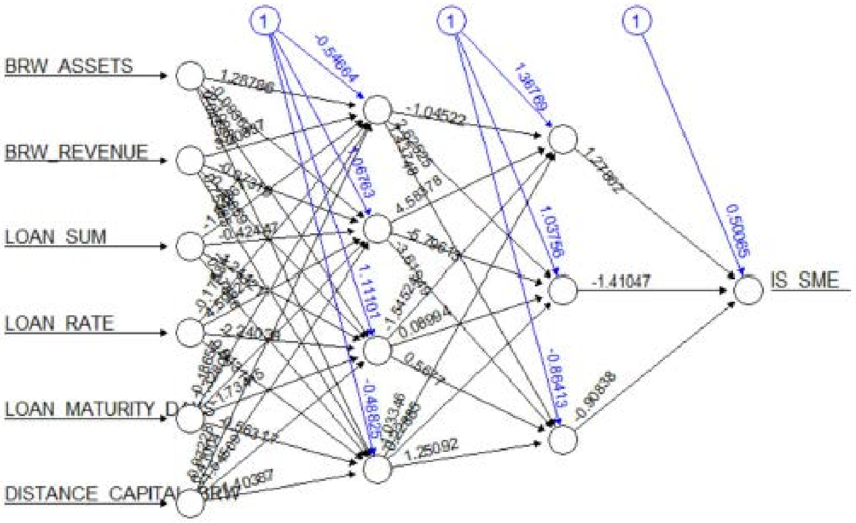


Figure 12. Confusion matrix for the neural network for SME-status prediction

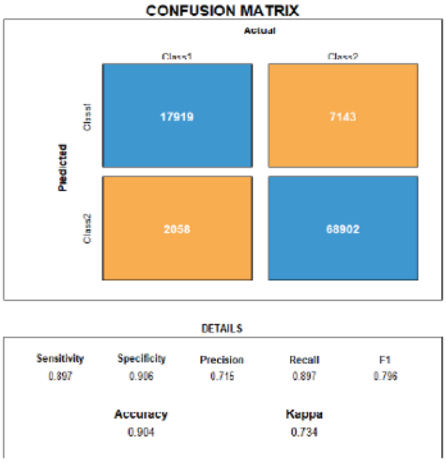
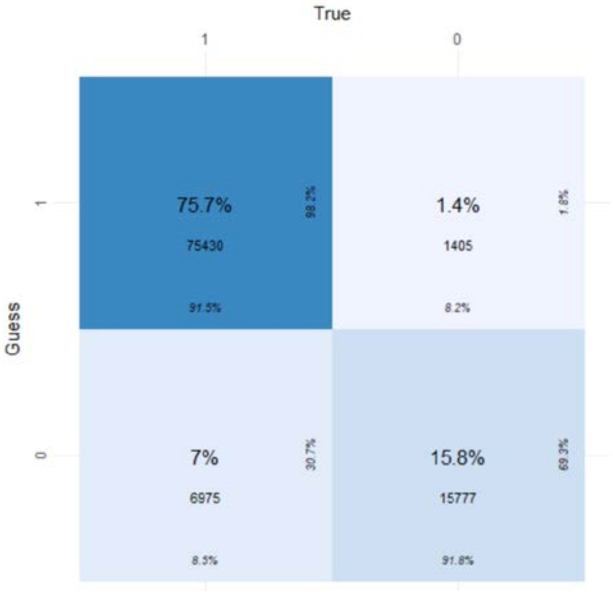


Figure 13. Confusion matrix for random forest solution for SME prediction



References

1. Aaron M., Hogg D. (2005) The use of microdata to assess risks in the non-financial corporate sector, Financial System Review, December, Bank of Canada.

2. Aljumaili, M. (2016). Data quality assessment: Applied in maintenance (Doctoral dissertation, Luleå tekniska universitet).

3. Batini, Carlo & Rula, Anisa & Scannapieco, Monica & Viscusi, Gianluigi. (2015). From Data Quality to Big Data Quality. Journal of Database Management. 26. 60-82. 10.4018/JDM.2015010103.

4. Blaise Gadanecz & Bruno Tissot & Mariagnese Branchi & Mario Ascolese, 2016. "The sharing of micro data – a central bank perspective," IFC Reports 6, Bank for International Settlements.

5. Carstens A. (2016) Micro-data as a Key Input to Designing Macro-prudential Policy: The Mexican Experience // Eighth European Central Bank Conference on Statistics, p.1-18

6. Carstens, A. (2016). Micro-data as a Key Input to Designing Macro-prudential Policy: The Mexican Experience. Remarks at the Eighth European Central Bank Conference on Statistics

7. Cheng, Y. (2021, April). Research on Credit Strategy Based on XGBoost Algorithm and Optimization Problem. In Journal of Physics: Conference Series (Vol. 1865, No. 4, p. 042137). IOP Publishing.

8. Coeuré, B., "Setting standards for granular data", opening remarks at the Third OFR-ECB-Bank of England workshop on "Setting Global Standards for Granular Data: Sharing the Challenge", Frankfurt am Main, March 2017.

9. COŞER, A., Maer-matei, M. M., & ALBU, C. (2019). PREDICTIVE MODELS FOR LOAN DEFAULT RISK ASSESSMENT. Economic Computation & Economic Cybernetics Studies & Research, 53(2).

10. Crato, N. & Paruolo, P. (Eds.). (2018). Data-driven policy impact evaluation: How Access to microdata is transforming policy design. Springer.

11. Crato, N., & Paruolo, P. (2019). The Power of Microdata: An Introduction. In Data-Driven Policy Impact Evaluation (pp. 1-14). Springer, Cham.

12. Divakar, K., & Chitharanjan, K. (2019). Performance evaluation of credit card fraud transactions using boosting algorithms. Int. J. Electron. Commun. Comput. Eng. IJECCE, 10(6), 262-270.

13. Domingo-Ferrer, J., & Blanco-Justicia, A. (2021, September). Towards Machine Learning-Assisted Output Checking for Statistical Disclosure Control. In International Conference on Modeling Decisions for Artificial Intelligence (pp. 335-345). Springer, Cham.

14. Dong, Q., Yan, X., Wilhoit, R. C., Hong, X., Chirico, R. D., Diky, V. V., & Frenkel, M. (2002). Data Quality Assurance for Thermophysical Property Databases Applications to the TRC SOURCE Data System. Journal of chemical information and computer sciences, 42(3), 473-480.

15. Dou, Y., Li, W., Liu, Z., Dong, Z., Luo, J., & Philip, S. Y. (2019, August). Uncovering download fraud activities in mobile app markets. In 2019 IEEE/ACM International

Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 671-678). IEEE.

16. Dyachkov, D.V, Nurimanova, I.F. (2017) Specifics of microdata-based statistics of interest rates on lending to non-financial sector // Russian Journal of Money and Finance (Money and Credit). 2017 Issue 12 – p. 64-72.

17. Eurostat (2007), Handbook on Data Quality Assessment Methods and Tools. Editors: Manfred Ehling and Thomas Körner.

18. Eurostat (2020). European Statistical System (ESS) handbook for quality and metadata reports — 2020 edition.

19. Farnè, Matteo; Vouldis, Angelos T. (2018) : A methodology for automatised outlier detection in high-dimensional datasets: An application to euro area banks' supervisory data, ECB Working Paper, No. 2171, ISBN 978-92-899-3276-9, European Central Bank (ECB), Frankfurt a. M.

20. Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2019). How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm.

21. Gavin, E. (2021). How to Collaborate Effectively to Improve Data Quality and Use in Revenue Administration and Official Statistics. IMF How To Notes, 2021(005).

22. Gomolka, M., Blaschke, J., Brîncoveanu, C., Hirsch, C., & Yalcin, E. Data Orchestration Blueprint Based on YAML {dobby} Research data pipelines in R.

23. González A.G. & Valadez M.S. & Cerecero M.R, 2019. "Sharing and using financial micro-data," IFC Bulletins chapters, in: Bank for International Settlements (ed.), Are post-crisis statistical initiatives completed?, volume 49, Bank for International Settlements.

24. Gross, F. (2010). Micro-data as a necessary infrastructure–standardisation of reference data on instruments and entities as a starting point: need for a Reference Data Utility. IFC Bulletin, 25, 334.

25. Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly, 37(3), 101493.

26. Kao, L.-J., Chiu, C.-C., and Chiu, F.-Y. (2012). A bayesian latent variable model with classification and regression tree approach for behavior and credit scoring. Knowledge-Based Systems, 36:245–252.

27. Karr, Alan F., Ashish P. Sanil, and David L. Banks. "Data quality: A statistical perspective." Statistical Methodology 3.2 (2006): 137-173.

28. Koduru, M., Pranati C., M., Phanidhar, M., & Srinivas, D. K. (2020). RF-XGBoost Model for Loan Application Scoring in Non Banking Financial Institutions. International Journal of Engineering Research & Technology (IJERT) ISSN, 2278-0181.

29. Kropf, S. L. (2020, November). ENHANCING DATA QUALITY FOR DATA ANALYTICS THROUGH MACHINE LEARNING. In European Scientific Conference of Doctoral Students (p. 97).

30. Lee, Y. W., Pipino, L., Funk, J. D., & Wang, R. Y. (2006). Journey to data quality (pp. 137-150). Cambridge: MIT press.

31. Li Ying. (2018). Research on bank credit default prediction based on data mining algorithm. The International Journal of Social Science and Humanities Invention 5(06): 4820-4820, ISSN: 2349-2031.

32. Livraga, G. (2019). Privacy in microdata release: Challenges, techniques, and approaches. In Data-Driven Policy Impact Evaluation (pp. 67-83). Springer, Cham.

33. Lukauskas, M., & Ruzgas, T. (2021). Bank credit card default classification based on clustering using machine learning algorithms. In 9th world sustainability forum, virtual, Switzerland, 13–15 September 2021: program and abstract book. MDPI.

34. Madhikermi, M., Kubler, S., Robert, J., Buda, A., & Främling, K. (2016). Data quality assessment of maintenance reporting procedures. Expert Systems with Applications, 63, 145-164.

35. Manjeet K., Vishesh G., Tarun J., Sahil S., DR. Lalit M. G. (2018). Neural Network Approach To Loan Default Prediction, International Research Journal of Engineering and Technology (IRJET) , p-ISSN: 2395-0072

36. Manjunath, T. N., Hegadi, R. S., & Ravikumar, G. K. (2010). Analysis of data quality aspects in datawarehouse systems. International Journal of Computer Science and Information Technologies, 2(1), 477-485.

37. Morandi, G., & Nicoletti, G. (2017). Using microdata from monetary statistics to understand intra-group transactions and their implication in financial stability issues. IFC Bulletins chapters, 46.

38. Odegua, R. (2020). Predicting Bank Loan Default with Extreme Gradient Boosting. arXiv preprint arXiv:2002.02011.

39. Osiewicz, M., Fache-Rousova, L., & Kulmala, K. M. (2015) Reporting of derivatives transactions in Europe–Exploring the potential of EMIR micro data against the challenges of aggregation across six trade repositories. BIS Report.

40. Perrella, A., & Catz, J. (2020). Integrating microdata for policy needs: the ESCB experience (No. 33). ECB Statistics Paper.

41. RAJU, O. (2021) CREDIT CARD FRAUD DETECTION USING XGBOOST CLASSIFIER.

42. Santos, C. (2013). Bank interest rates on new loans to non-financial corporafions–one first look at a new set of micro data. Economic Bulletin and Financial Stability Report Articles and Banco de Portugal Economic Studies.

43. Sebastian Doerr & Leonardo Gambacorta & José María Serena Garralda, 2021. "Big data and machine learning in central banking," BIS Working Papers 930, Bank for International Settlements.

44. Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. Machine Learning with Applications, 100074.

45. Smith, M., Lix, L. M., Azimaee, M., Enns, J. E., Orr, J., Hong, S., & Roos, L. L. (2018). Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. Journal of the American Medical Informatics Association, 25(3), 224-229.

46. Tam, S. M., & Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. International Statistical Review, 83(3), 436-448.

47. Trisanto, D., Rismawati, N., Mulya, M. F., & Kurniadi, F. I. (2021) Modified Focal Loss in Imbalanced XGBoost for Credit Card Fraud Detection.

48. V. Gudivada, A. Apon, and J. Ding. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations". In: International Journal on Advances in Software 10.1 (2017), pp. 1 - 20.

49. Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning - a case study of bank loan data. Procedia Computer Science, 174, 141-149.

50. Zambuto, F., Buzzi, M. R., Costanzo, G., Di Lucido, M., La Ganga, B., Maddaloni, P., ... & Svezia, E. (2020). Quality checks on granular banking data: an experimental approach based on machine learning?. Bank of Italy Occasional Paper, (547).

Bank of Russia

MACHINE LEARNING-BASED APPROACHES FOR AUTOMATIC DATA VALIDATION AND OUTLIER CONTROL OF LOAN MICRODATA IN THE BANK OF RUSSIA

Dmitrii Diachkov
STATISTICS DEPARTMENT

Introduction

Goal

Explore the possibility of applying relatively **simple machine learning methods** to improve an **overall quality** of microdata and decisions based on them in the Bank of Russia.

Motivation

Limited human ability (analyst or statistician) to analyze the reliability of a large and complicated datasets that change over time

Field of application

Attributes of loan microdata:

- Borrower’s economic activity;
- Bank’s accounting
- Borrower’s business size;
- Loan type
- Interest rates

Goal achievement criteria

High interpretability of results;

Moderate ease of implementation;

Moderate process control;

High scalability;

High automation capability.

Data sources

- Banks’ reporting form 0409303 "Information on loans granted to legal entities and individual entrepreneurs" with microdata on loans;

- Annual accounting data;

- Statistical Registry on companies provided by State Statistics Service;

- State Registry of SMEs provided by Federal Tax Service.

Stack of technologies – Oracle R Advanced Analytics

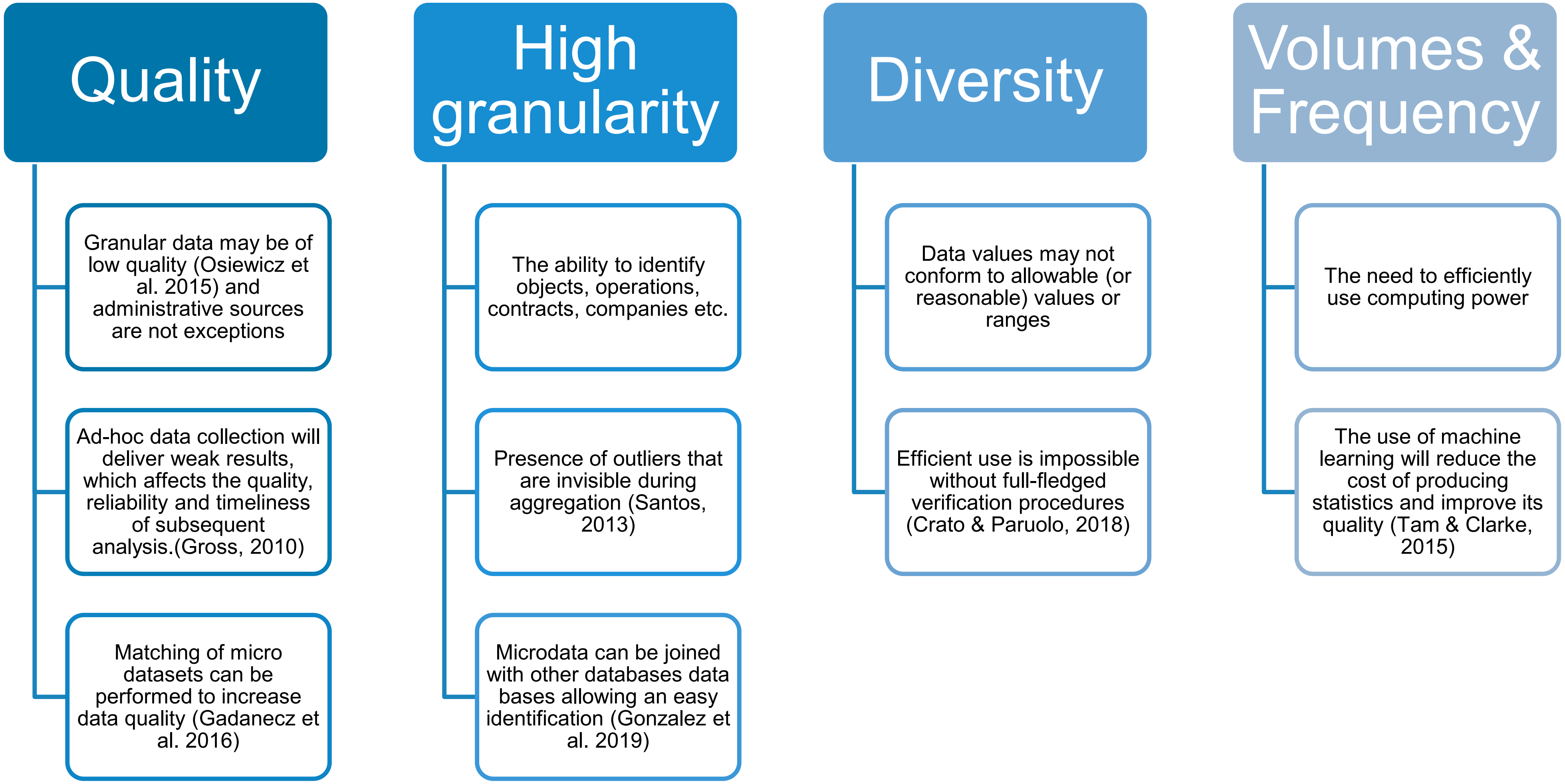
ORACLE

Highly reliable organized collection of structured data

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management

Packages: caret, rpart, neuralnet, nnet, caTools, randomForest, class, cvms

Accumulated experience of working with loan microdata



Quality

- Granular data may be of low quality (Osiewicz et al. 2015) and administrative sources are not exceptions
- Ad-hoc data collection will deliver weak results, which affects the quality, reliability and timeliness of subsequent analysis.(Gross, 2010)
- Matching of micro datasets can be performed to increase data quality (Gadanecz et al. 2016)

High granularity

- The ability to identify objects, operations, contracts, companies etc.
- Presence of outliers that are invisible during aggregation (Santos, 2013)
- Microdata can be joined with other databases data bases allowing an easy identification (Gonzalez et al. 2019)

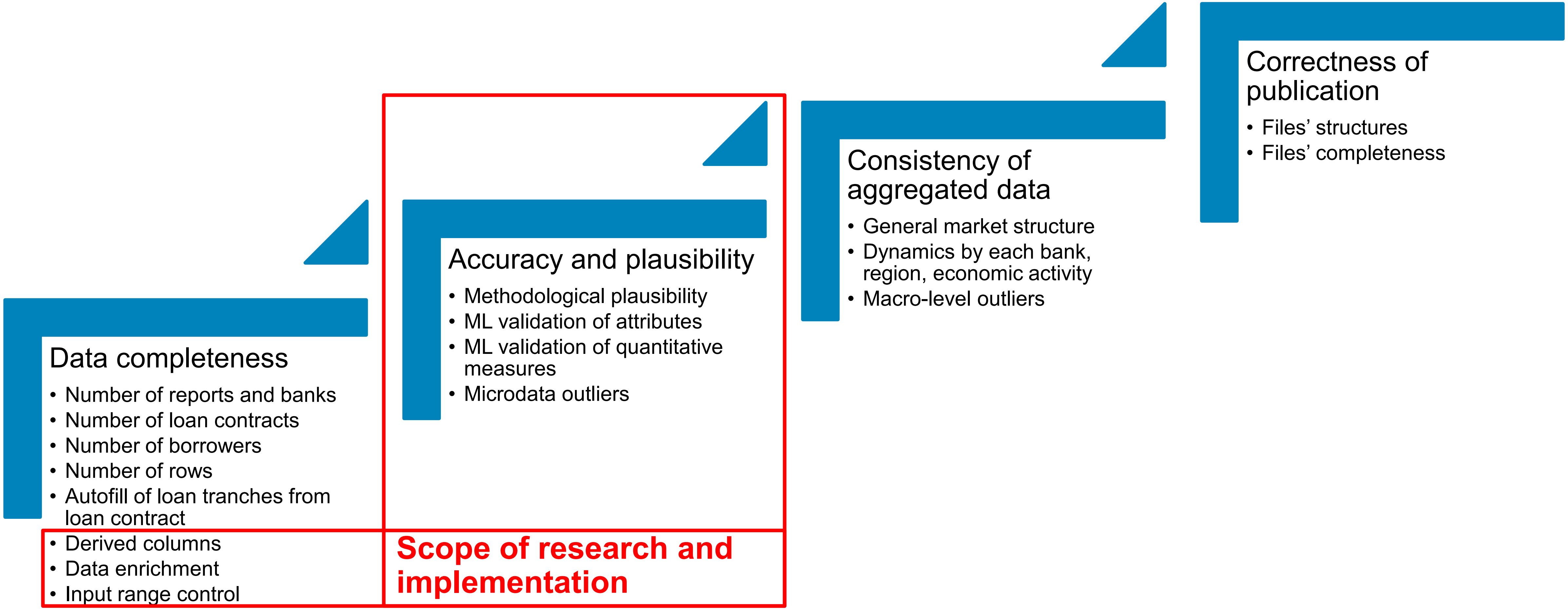
Diversity

- Data values may not conform to allowable (or reasonable) values or ranges
- Efficient use is impossible without full-fledged verification procedures (Crato & Paruolo, 2018)

Volumes & Frequency

- The need to efficiently use computing power
- The use of machine learning will reduce the cost of producing statistics and improve its quality (Tam & Clarke, 2015)

Control procedures for loan microdata in the Bank of Russia



Data completeness

- Number of reports and banks
- Number of loan contracts
- Number of borrowers
- Number of rows
- Autofill of loan tranches from loan contract
- Derived columns
- Data enrichment
- Input range control

Accuracy and plausibility

- Methodological plausibility
- ML validation of attributes
- ML validation of quantitative measures
- Microdata outliers

Scope of research and implementation

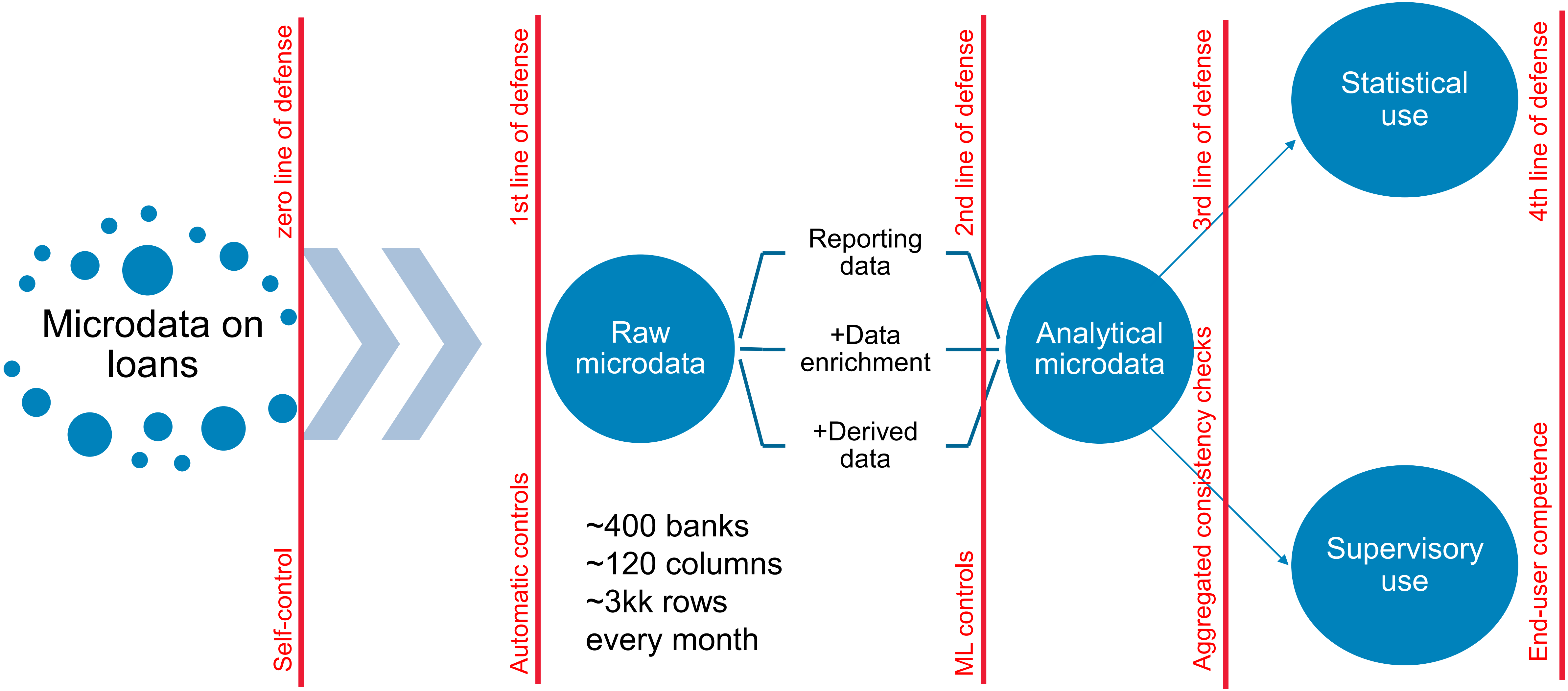
Consistency of aggregated data

- General market structure
- Dynamics by each bank, region, economic activity
- Macro-level outliers

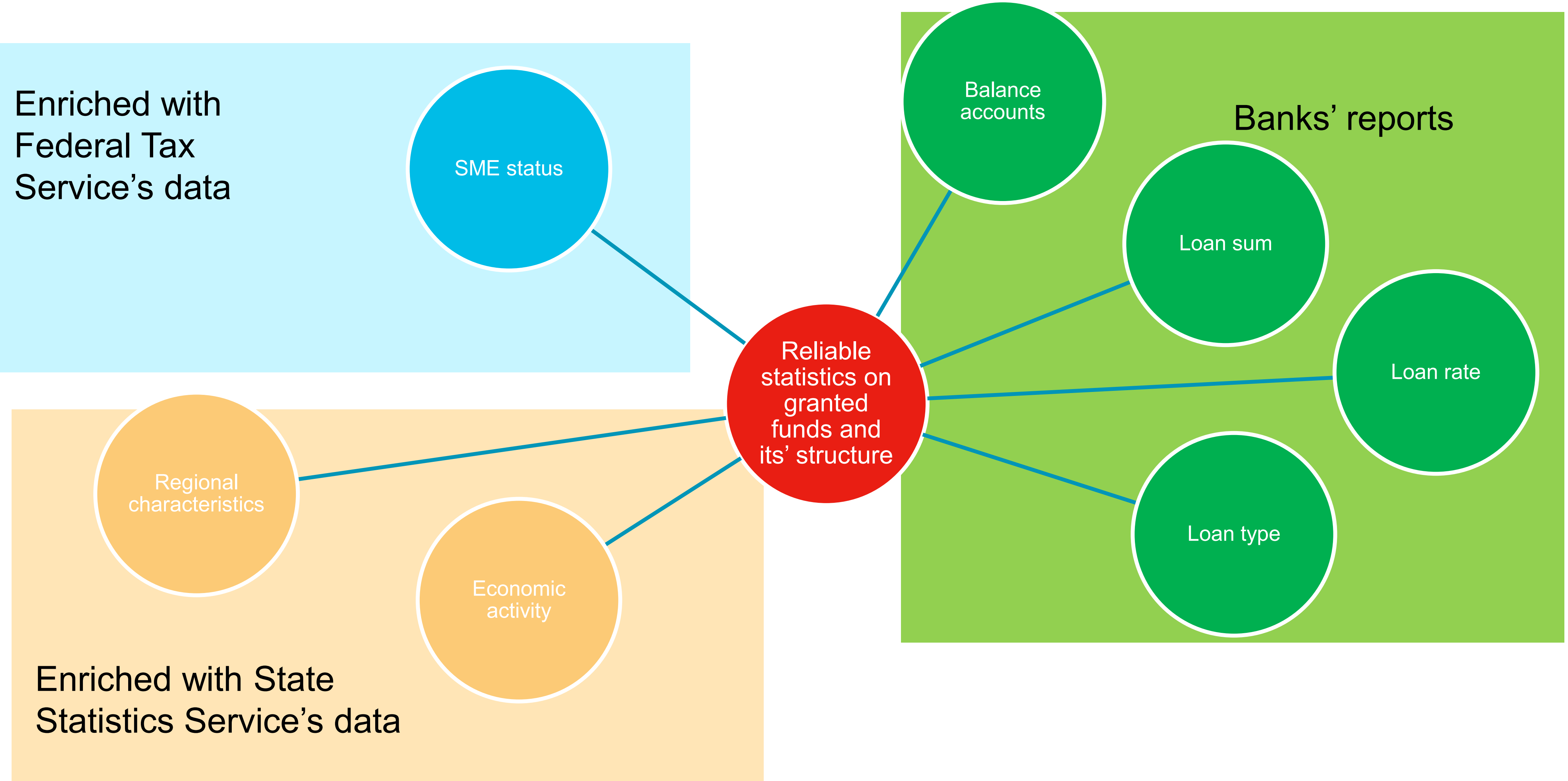
Correctness of publication

- Files’ structures
- Files’ completeness

Loan microdata pipeline in the Bank of Russia



Most beneficial ways to apply ML validation (as of today)


Enriched with Federal Tax Service's data
SME status
Banks' reports
Balance accounts
Loan sum
Loan rate
Loan type
Reliable statistics on granted funds and its' structure
Regional characteristics
Economic activity
Enriched with State Statistics Service's data

Case 1. Validation of balance sheet codes for non-financial companies

Problem

Balance sheet account is a simple and understandable attribute for economists that can be used to filter borrowers by type and industry. Information about the balance sheet account is reported to the Bank of Russia by banks.

Balance sheet code can be determined by bank based on irrelevant data or just by mistake. Another type of mistake is technical errors during report preparation or submission.

Intuition

The balance sheet account should be assigned by bank to it's borrower based on limited list of parameters, such as type of loan, business entity form and economic activity …

…so there must be a way to cross-check balance sheet codes with the data of public registers or other data…

Implementation

A set of decision trees, that establish dependencies between balance sheets codes:

- Loan type
- Economic activity
- Business entity form
- Debt sum

Main balance sheet codes that are used for loan statistics and to be validated:

452 – Loans to non-financial companies

454 – Loans to individual entrepreneurs

451 – Loans to financial companies

Measuring the efficiency of 15 models: balancing accuracy VS speed

TARGET:

BALANCE SHEET CODE == “452”

Train data: 845k loans;

Test data: 211k loans (20%).

The dependent variable is evenly distributed in the training and test samples.

Data type	Yes	No
Train	47,56%	52,44%
Test	47,54%	52,46%

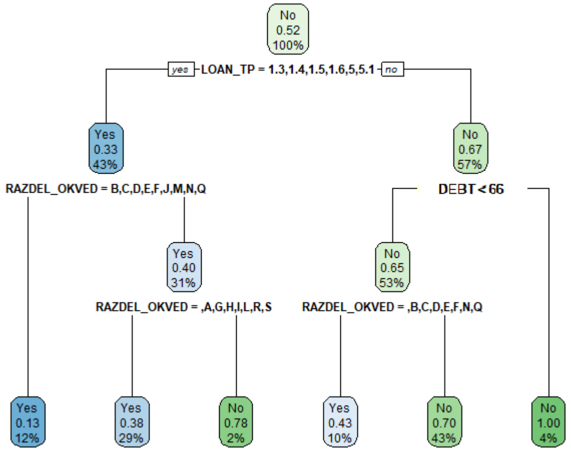
Since the training and test samples are balanced (predictable classes are evenly distributed) -> Accuracy (1) is an excellent metric of model quality.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

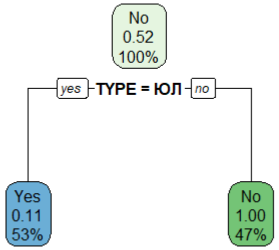
Formula	Mean accuracy, %	Mean time (seconds per 1kk rows)	Relative Efficiency
~ ECON_ACTIVITY + BORROWER_TYPE + LOAN_TYPE	98,904	17,044	5,80
~ ECON_ACTIVITY + LOAN_TYPE + DEBT + BORROWER_TYPE	98,886	18,257	5,42
~ BORROWER_TYPE + LOAN_TYPE	97,352	19,742	4,93
~ BORROWER_TYPE + LOAN_TYPE + DEBT	97,346	20,512	4,75
~ ECON_ACTIVITY + BORROWER_TYPE + DEBT	96,064	22,028	4,36
~ ECON_ACTIVITY + BORROWER_TYPE	95,981	19,546	4,91
~ BORROWER_TYPE	94,182	17,071	5,52
~ BORROWER_TYPE + DEBT	94,155	23,28	4,04
~ ECON_ACTIVITY + LOAN_TYPE	69,813	22,616	3,09
~ ECON_ACTIVITY + LOAN_TYPE + DEBT	69,732	35,284	1,98
~ LOAN_TYPE	67,244	16,122	4,17
~ LOAN_TYPE + DEBT	67,127	30,853	2,18
~ ECON_ACTIVITY	61,317	17,175	3,57
~ ECON_ACTIVITY + DEBT	61,226	32,287	1,90
~ DEBT	54,282	28,043	1,94

$$Relative\ Efficiency = \frac{E(Accuracy)}{E(Time)} \ (2),$$

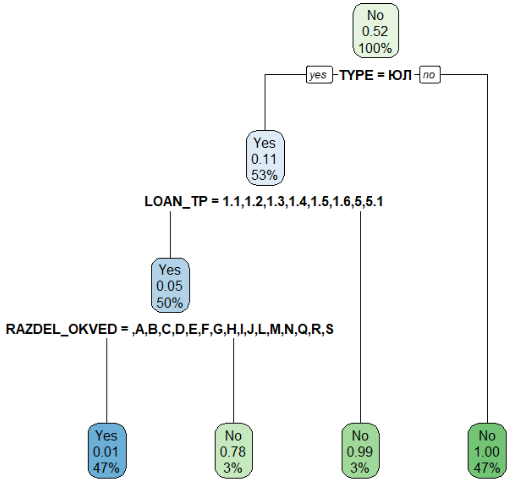
Top-3 models for Case 1: measuring the efficiency


No
0.52
100%
yes
LOAN_TP = 1.3,1.4,1.5,1.6,5,5.1
no
Yes
0.33
43%
No
0.67
57%
RAZDEL_OKVED = B,C,D,E,F,J,M,N,Q
DEBT < 66
Yes
0.40
31%
No
0.65
53%
RAZDEL_OKVED = ,A,G,H,I,L,R,S
RAZDEL_OKVED = ,B,C,D,E,F,N,Q
Yes
0.13
12%
Yes
0.38
29%
No
0.78
2%
Yes
0.43
10%
No
0.70
43%
No
1.00
4%

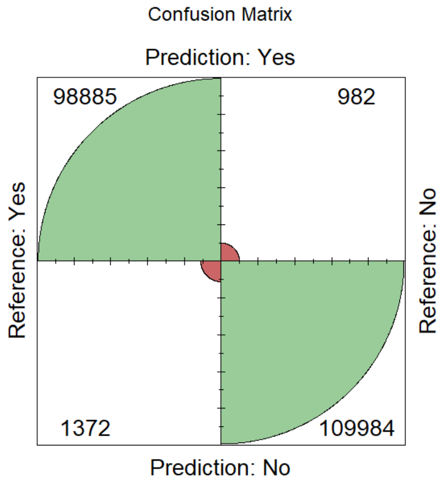
“Accurate”

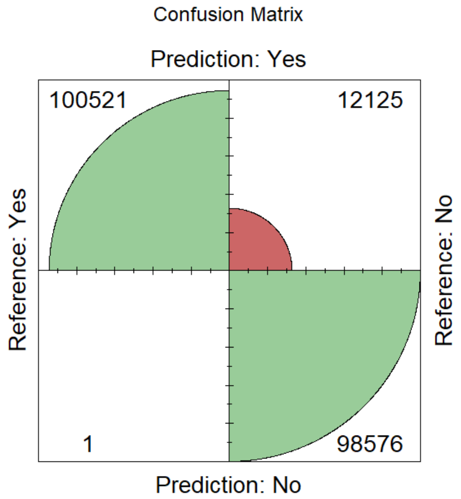


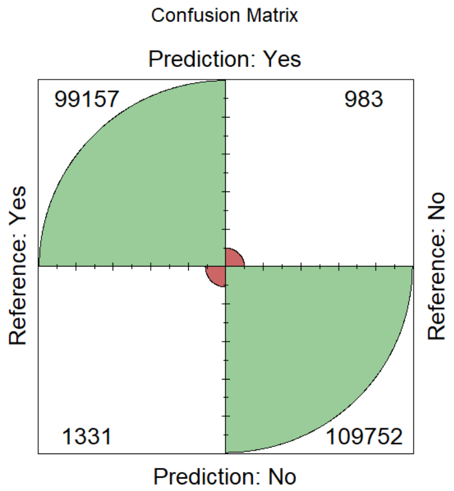
“Fast”


No
0.52
100%
yes
TYPE = ЮЛ
no
Yes
0.11
53%
LOAN_TP = 1.1,1.2,1.3,1.4,1.5,1.6,5,5.1
Yes
0.05
50%
RAZDEL_OKVED = ,A,B,C,D,E,F,G,H,I,J,L,M,N,Q,R,S
Yes
0.01
47%
No
0.78
3%
No
0.99
3%
No
1.00
47%

“Balanced”



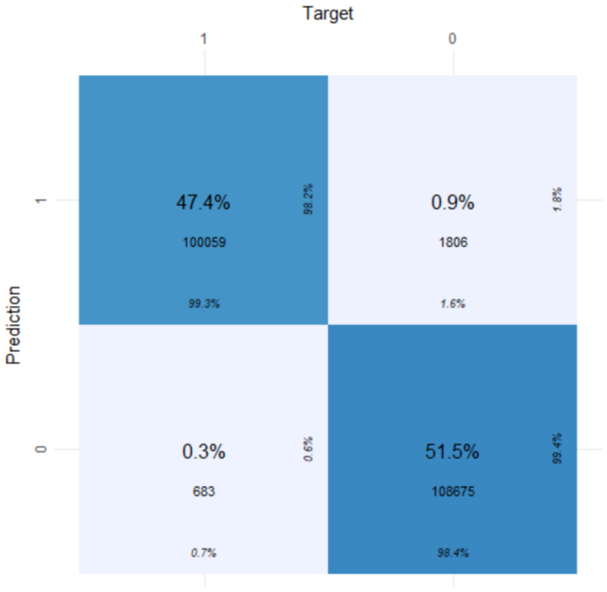


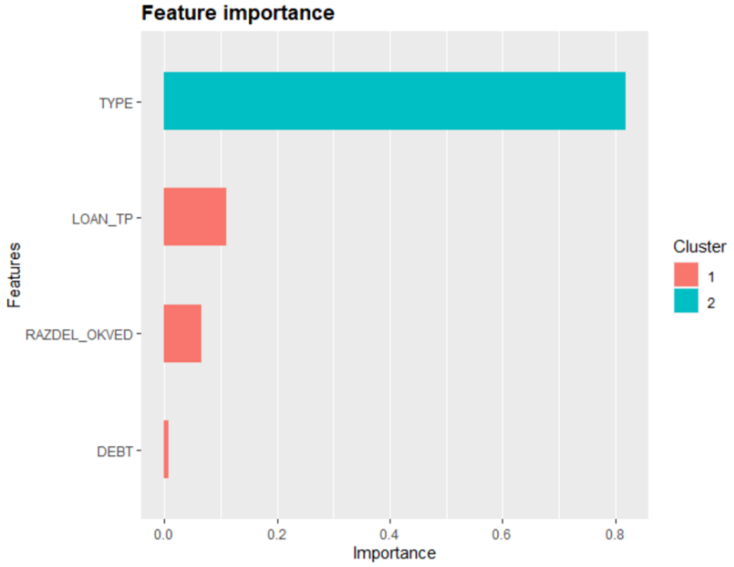


Alternative solutions with XGBoost and logistic regression

We achieved the same results in terms of efficiency but with a longer computational time.

Model type	Mean accuracy	Mean precision	Mean recall	Mean time of fit and predict (per 1kk rows), seconds	Relative efficiency score
Decision tree	0,989	0,987	0,990	17	1,939
XGBoost	0,988	0,993	0,984	45	0,731
Logistic regression	0,989	0,99	0,989	113	0,291





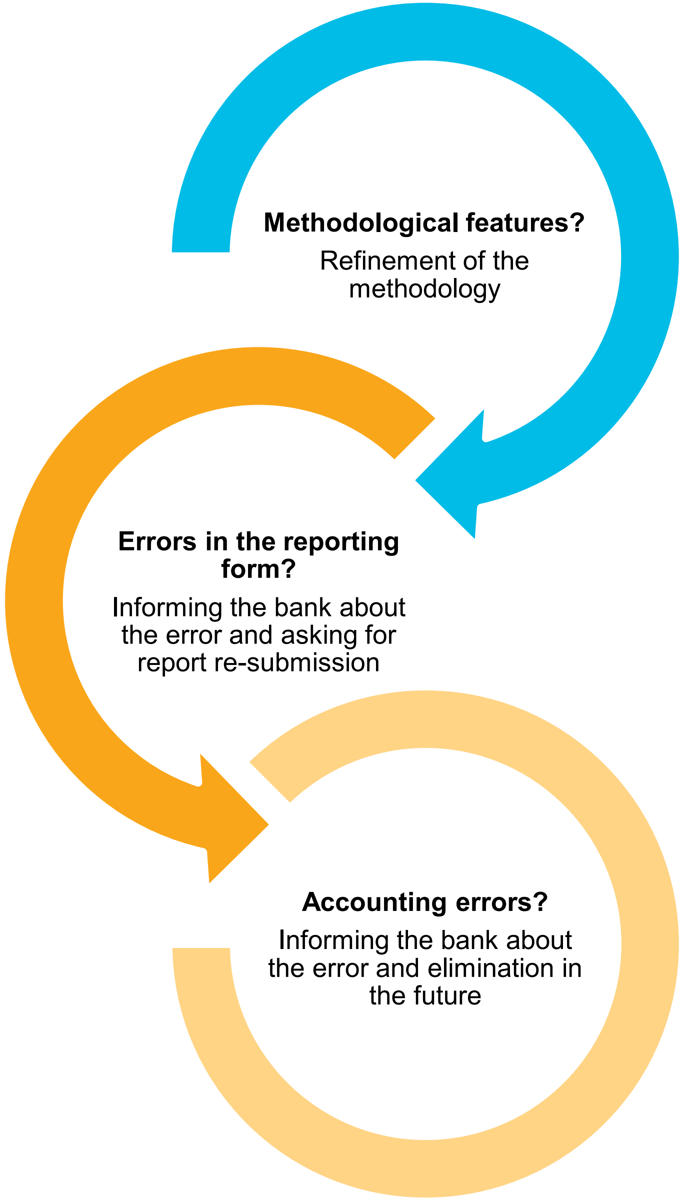
Deviation analysis for Case 1

1

0.66% or 1331 loans were classified by the decision tree as loans of non-financial companies (code 452), but in fact they are reported on other accounts (FP, for example, 453 - loans to non-residents or 451 - loans to financial companies).

2

0.44% or 983 loans were classified by the decision tree as loans that should be recorded as non-452, but in fact they are recorded in 452 accounts (FN, mainly financial companies, which should be reflected with code 451)



Methodological features?

Refinement of the methodology

Errors in the reporting form?

Informing the bank about the error and asking for report re-submission

Accounting errors?

Informing the bank about the error and elimination in the future

Case 2. Balance sheet code 451 with imbalanced distribution

TARGET:

BALANCE SHEET CODE == “451”

Train data: 845k loans;

Test data: 211k loans (20%).

The dependent variable is not evenly distributed in the training and test samples

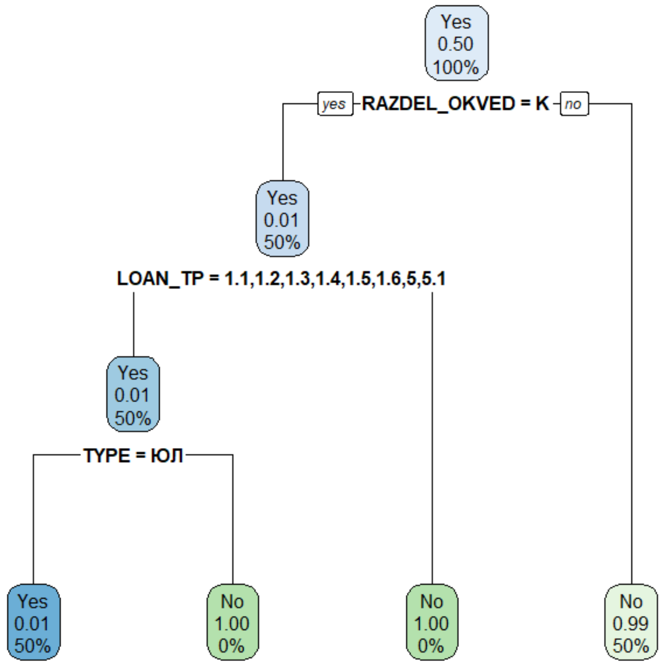
–> imbalanced samples approach

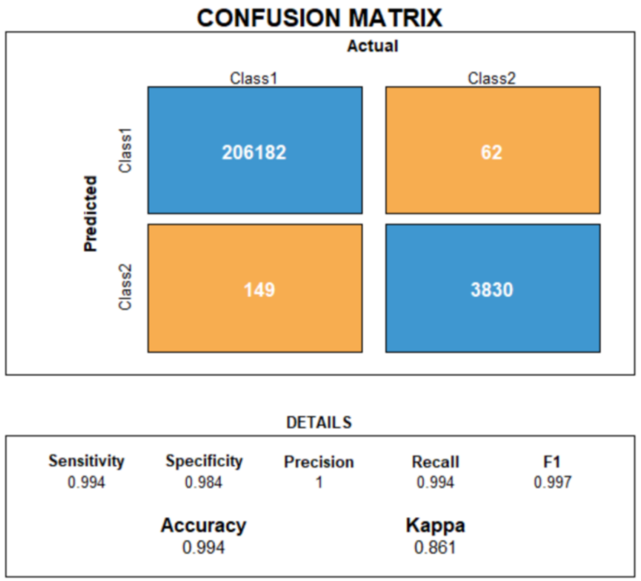
Data type	Yes	No
Train before up-sampling	1,80%	98,20%
Train after up-sampling	50%	50%
Test	1,82%	98,18%

Formula	Mean F1_score, %	Mean time (seconds per 1kk rows)	Relative Efficiency2
~ ECON_ACTIVITY + LOAN_TYPE + DEBT + BORROWER_TYPE	86,465	7,334	11,79
~ ECON_ACTIVITY + BORROWER_TYPE	85,812	6,053	14,18
~ ECON_ACTIVITY + LOAN_TYPE + DEBT	83,952	4,871	17,24
~ ECON_ACTIVITY + LOAN_TYPE	83,718	4,587	18,25
~ ECON_ACTIVITY + BORROWER_TYPE + DEBT	80,405	4,129	19,47
~ ECON_ACTIVITY + BORROWER_TYPE + LOAN_TYPE	80,298	2,916	27,54
~ ECON_ACTIVITY + DEBT	78,346	3,848	20,36
~ ECON_ACTIVITY	77,568	2,731	28,40
~ BORROWER_TYPE + LOAN_TYPE + DEBT	9,313	4,729	1,97
~ BORROWER_TYPE + LOAN_TYPE	9,057	4,157	2,18
~ BORROWER_TYPE + DEBT	6,569	3,753	1,75
~ BORROWER_TYPE	6,565	2,162	3,04
~ LOAN_TYPE + DEBT	4,944	4,052	1,22
~ LOAN_TYPE	4,73	3,662	1,29

$$Relative\ Efficiency2 = \frac{E(F1\_score)}{E(Time)} \quad (3)$$

Adaptation of the approach to the classification of financial companies





CONFUSION MATRIX

		Actual	
		Class1	Class2
Predicted	Class1	206182	62
	Class2	149	3830

DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.994	0.984	1	0.994	0.997

Accuracy	Kappa
0.994	0.861

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1\_score = \frac{2 * Precision * Recall}{Precision+Recall} \quad (6)$$

The affiliation of the legal entity to the industry K and specific loan types loan both determine the belonging to the code 451.

Despite the high accuracy (99%), the confidence in the model is significantly lower since the training and test samples are unbalanced.

Accuracy in this situation is an incorrect metric.

Precision (4) and Recall (5) metrics do not depend on the ratio of classes and therefore are applicable in conditions of unbalanced samples.

F1_score is a good balance between these two metrics.

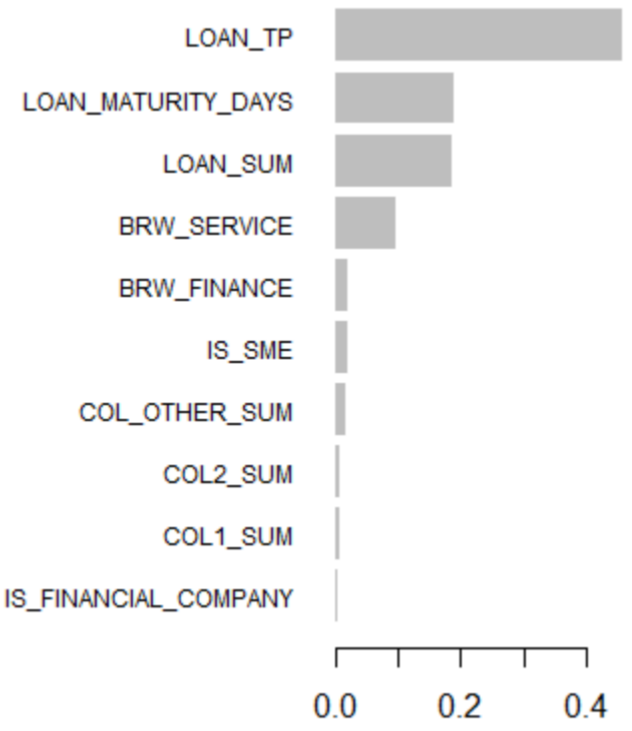
Summary of results for decision tree, XGBoost, and logistic regression approaches to classification task for account code “451”

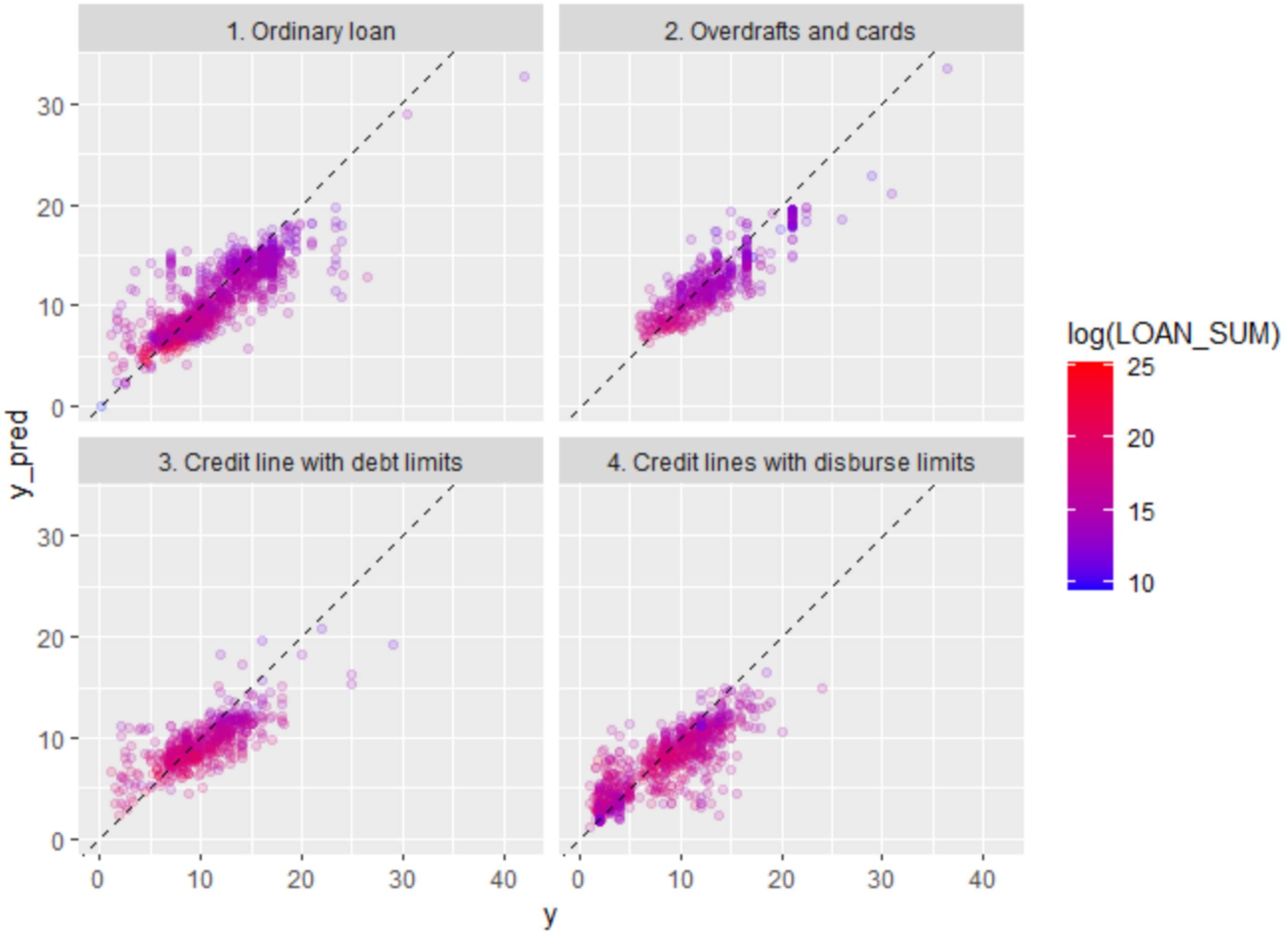
Train data type	Model type	Accuracy, %	Fit and predict time (seconds per 1kk rows)	F1_score, %	Relative efficiency score 2
Downsampled (30k rows)	Decision tree	98,98%	9,976	99,48%	0,100
	Logistic regression	99,35%	62,34	99,67%	0,016
	XGBoost	99,35%	50,606	99,67%	0,020
Original - imbalanced (845k rows)	Decision tree	99,43%	14,081	99,71%	0,071
	Logistic regression	99,44%	136,62	99,71%	0,007
	XGBoost	99,32%	38,783	99,65%	0,026
Upsampled (1658k rows)	Decision tree	98,99%	15,331	99,48%	0,065
	Logistic regression	99,39%	125,16	99,69%	0,008
	XGBoost	99,35%	70,38	99,67%	0,014

Case 3. Validation of interest rates data with eXtreme gradient boosting

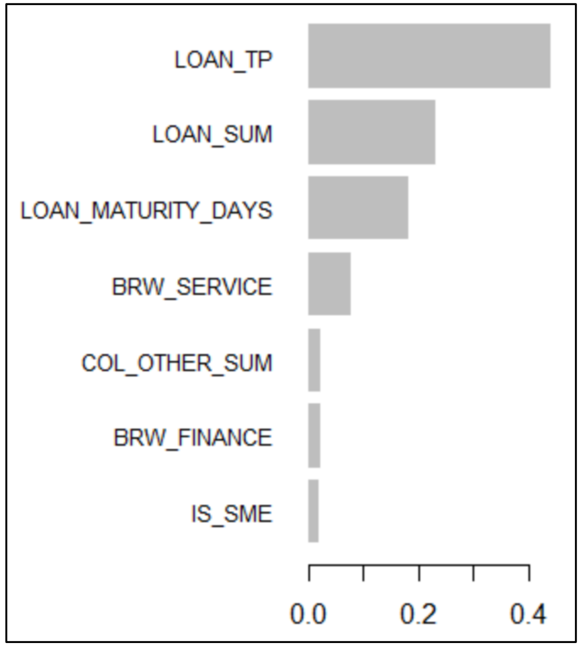
Train data: 421k loans;

Test data: 105k loans (20%).



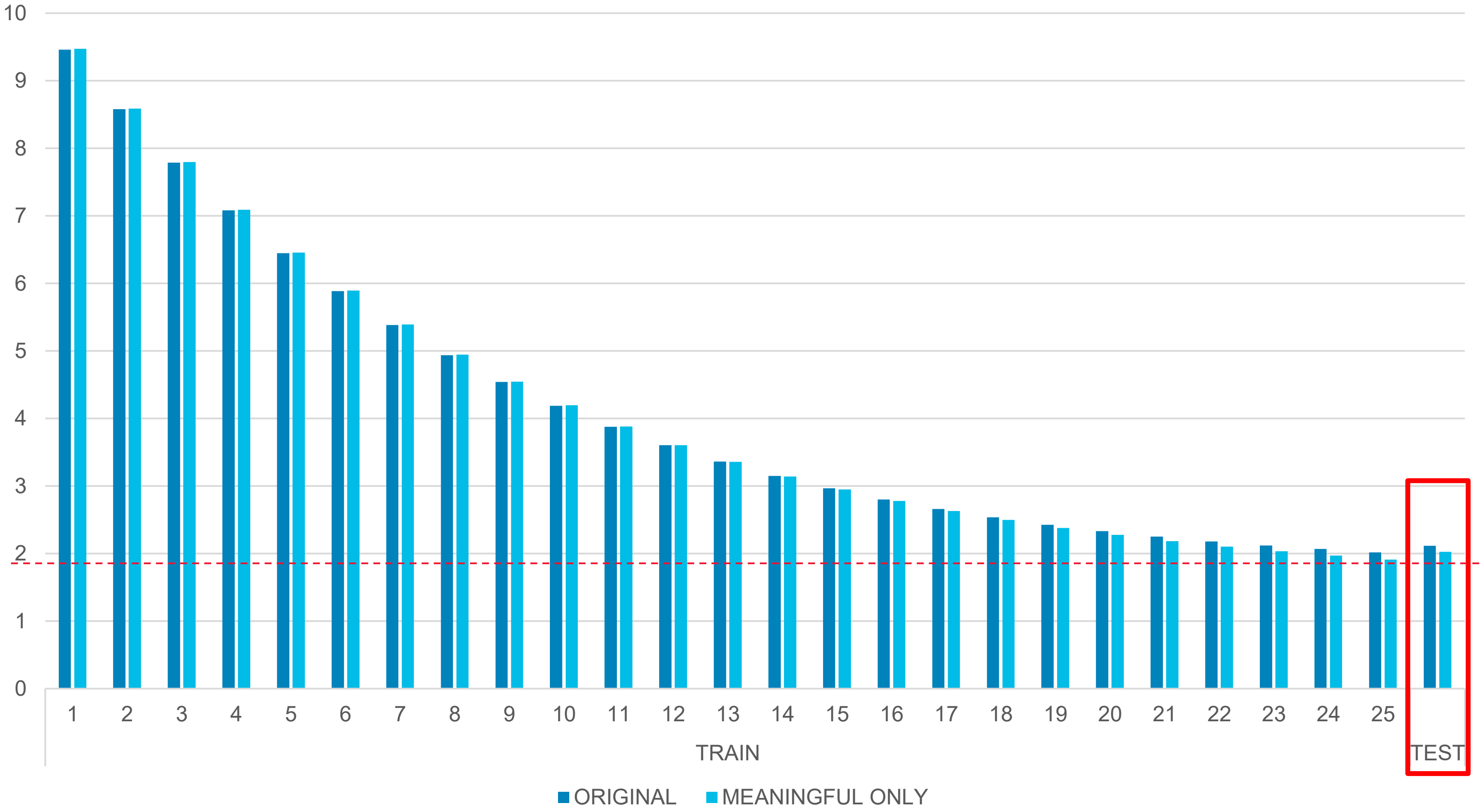


Case 3. Validation of interest rates data with eXtreme gradient boosting



Model	RMSE
XGBoost	2,04
Neural net (caret and nnet)	3,25
Linear Regression	3,85

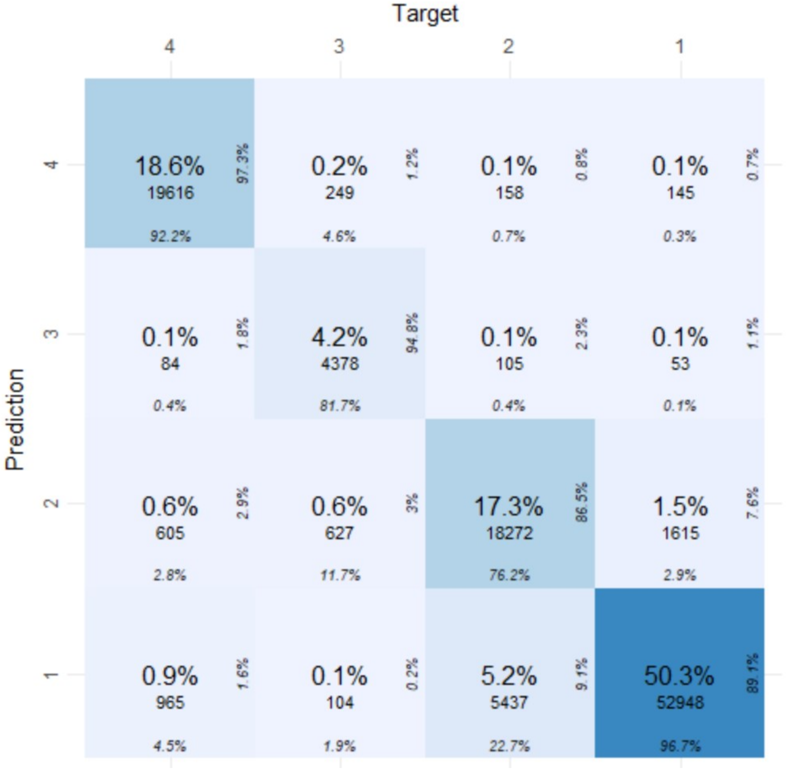
RMSE FOR XGBOOST TRAINING AND TESTING



Case 4. XGBoost multiclassification for SME size validation

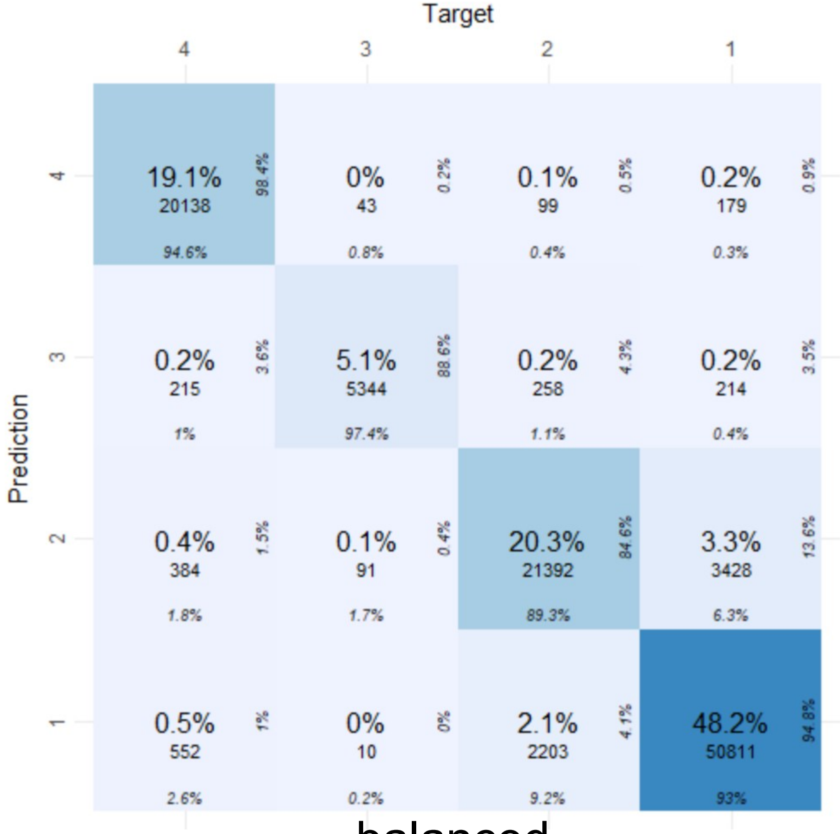
Powerful tool with high speed and very high accuracy on millions of rows

Performance on imbalanced sample is just 2% less accurate, than on upsampled and balanced (90% VS 92%)



	4	3	2	1
4	18.6% 19616 92.2% 97.3%	0.2% 249 4.6% 1.2%	0.1% 158 0.7% 0.8%	0.1% 145 0.3% 0.7%
3	0.1% 84 0.4% 1.8%	4.2% 4378 81.7% 94.8%	0.1% 105 0.4% 2.3%	0.1% 53 0.1% 1.1%
2	0.6% 605 2.8% 2.9%	0.6% 627 11.7% 3%	17.3% 18272 76.2% 86.5%	1.5% 1615 2.9% 7.6%
1	0.9% 965 4.5% 1.6%	0.1% 104 1.9% 0.2%	5.2% 5437 22.7% 9.1%	50.3% 52948 96.7% 89.1%

imbalanced

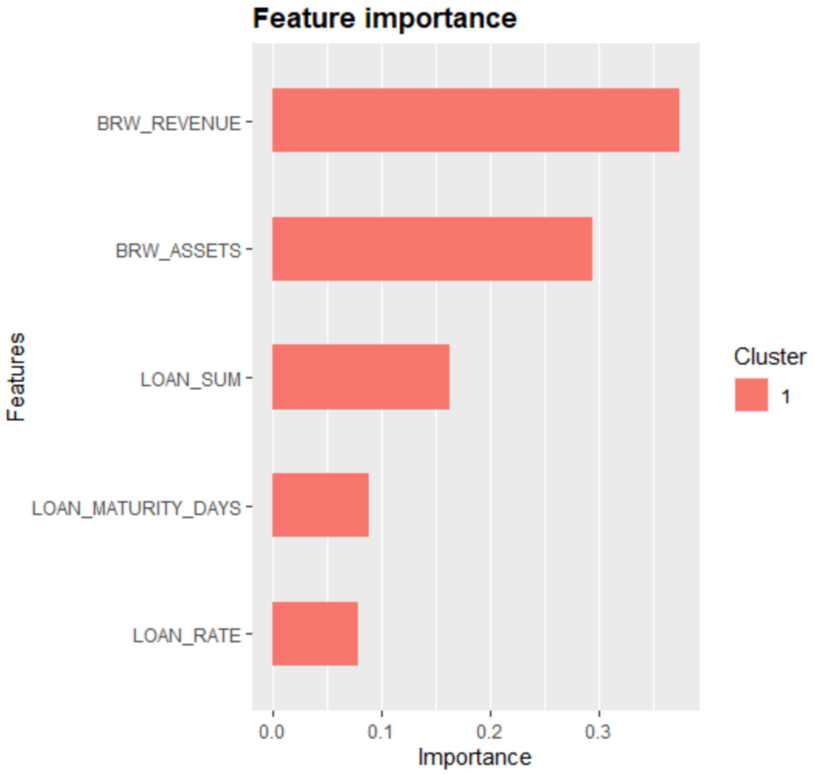


	4	3	2	1
4	19.1% 20138 94.6% 98.4%	0% 43 0.8% 0.2%	0.1% 99 0.4% 0.5%	0.2% 179 0.3% 0.9%
3	0.2% 215 1% 3.6%	5.1% 5344 97.4% 88.6%	0.2% 258 1.1% 4.3%	0.2% 214 0.4% 3.5%
2	0.4% 384 1.8% 1.5%	0.1% 91 1.7% 0.4%	20.3% 21392 89.3% 84.6%	3.3% 3428 6.3% 13.6%
1	0.5% 552 2.6% 1%	0% 10 0.2% 0%	2.1% 2203 9.2% 4.1%	48.2% 50811 93% 94.8%

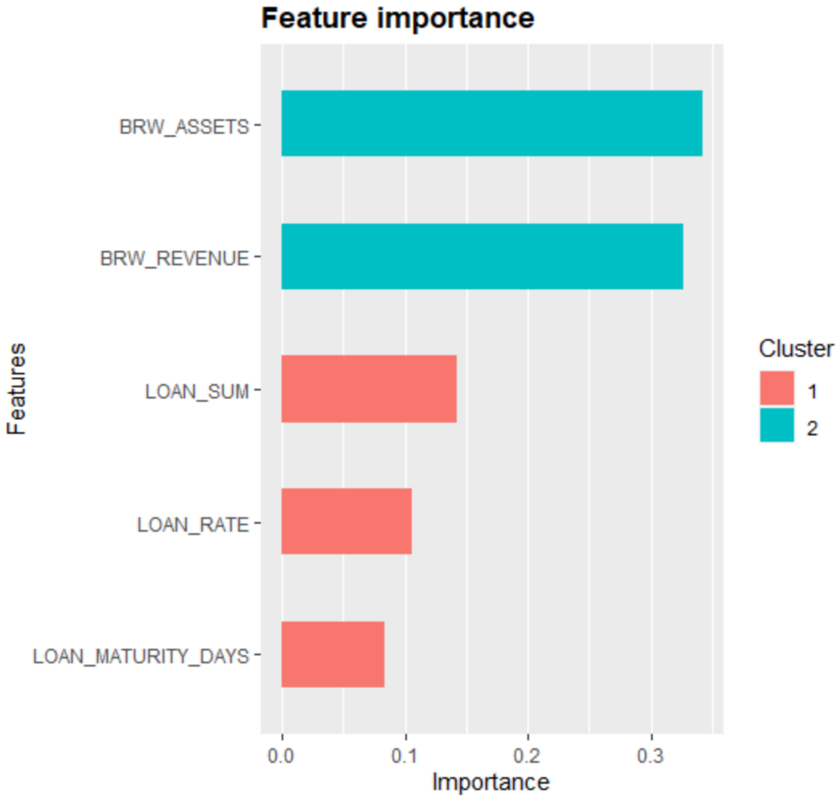
balanced

Case 4. XGBoost multiclassification for SME size validation

But the balanced sample provided more insight on the importance of features, so computational power may be saved



imbalanced



balanced

Case 5. Validation of SME status with neural networks

Problem

Belonging to the SME Registry (provided by Federal Tax Service) defines a borrower’s business size. This attribute is extremely important for the analysis of the economic situation. However, really small and insignificant companies may be excluded from the Registry for various reasons. We need to be able to check the validity of any given status.

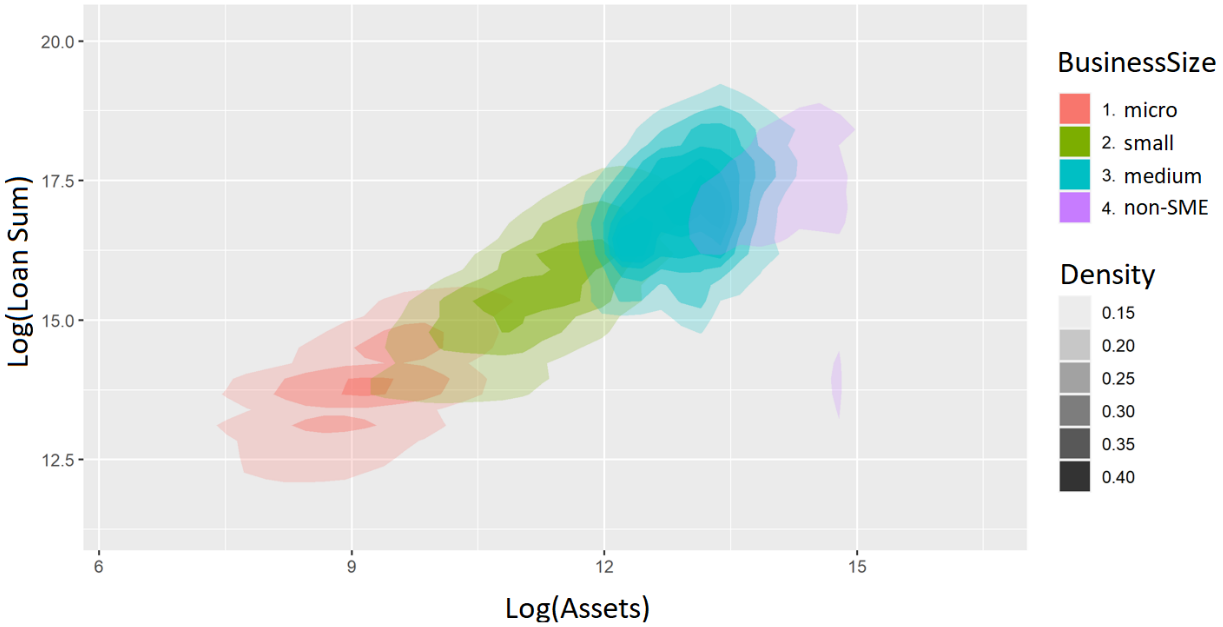
Intuition

Large borrowers have more assets and, accordingly, apply for more significant amounts of loans.

SME borrowers are usually smaller in business size and balance sheet.

Implementation

Build a neural network that can classify companies as SMEs and non-SMEs based on various quantitative indicators characterizing the loan and the borrower.



Finding balance between complexity and accuracy

TARGET:
COMPANY IS SME == “TRUE”

The dependent variable is not evenly distributed in the training and test samples

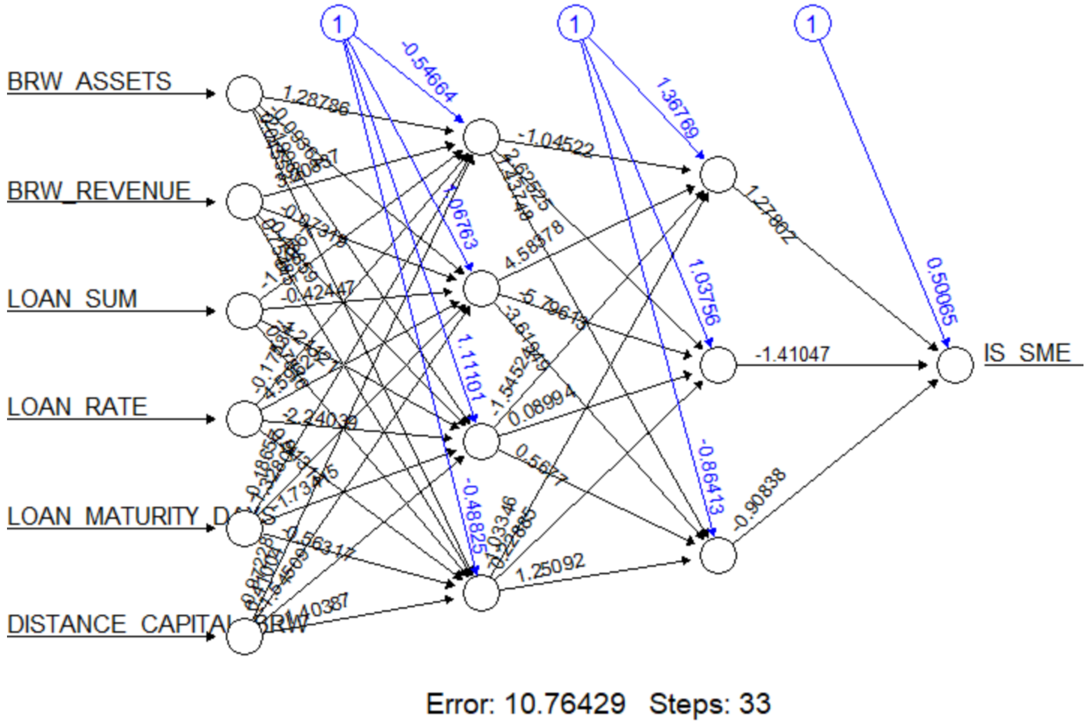
1. Imbalanced samples approach

Data type	Yes	No
Train before up-sampling	75,48%	24,54%
Train after up-sampling	50%	50%
Test	75,09%	24,91%

2. Normalizing the input information for the model

+ scaling or normalization to the range [-1; 1]

+ sigmoid activation function



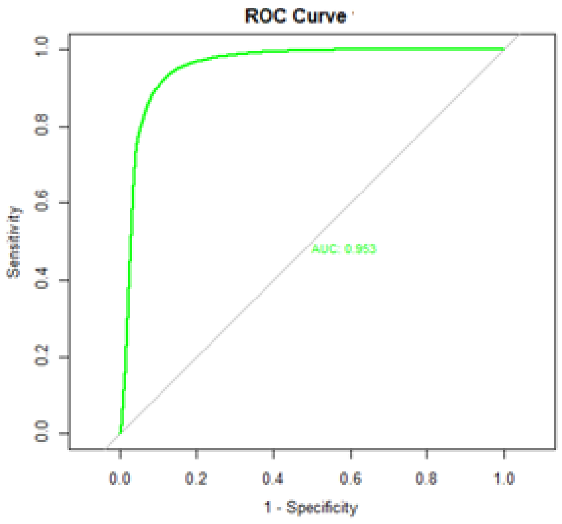
2 hidden layers, 4 and 3 neurons
Accuracy - 90.4%

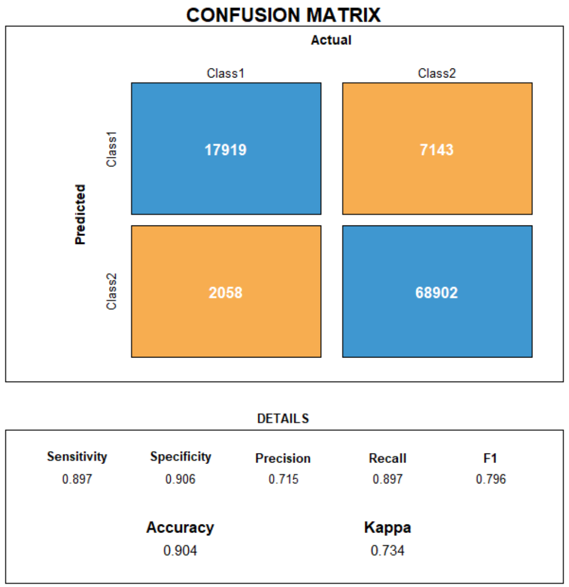
Training time ~ 619 sec. for 1kk objects

Choice of optimal neural network

Hidden layers	Accuracy	Fitting time	Prediction time	Weighted score
4,3	90,4%	619,78	138,86	10,77211413
2,2	67,2%	102,54	334,39	10,33535241
3,2	73,2%	232,73	299,22	10,07282431
4	79,8%	518,30	245,60	8,336222782
5,3	71,0%	339,18	305,19	7,823119106
4,2	71,2%	264,82	420,88	7,393118623
4,1	80,6%	607,07	346,24	6,814520418
5,3	58,2%	7,68	493,28	6,761507372
3	78,8%	484,72	444,25	6,684266136
4,1	47,4%	5,19	337,11	6,563785262
4,3	55,8%	11,47	539,54	5,650794018
4,2	70,8%	619,94	354,98	5,141586833
2	77,6%	1063,63	156,36	4,935904051
3,2	88,4%	1398,23	281,46	4,652387824
5	69,4%	321,55	760,02	4,453117059
4,3	43,8%	6,85	426,56	4,426365879
4,2	70,6%	796,28	380,96	4,233930891
5,4	76,8%	1431,68	377,27	3,260586787
6,2	77,8%	1619,66	292,87	3,16482278

A simpler neural network sometimes may turn out as more accurate and efficient in terms of consumption of computational resources (when it comes to big data, this issue becomes very significant)




CONFUSION MATRIX

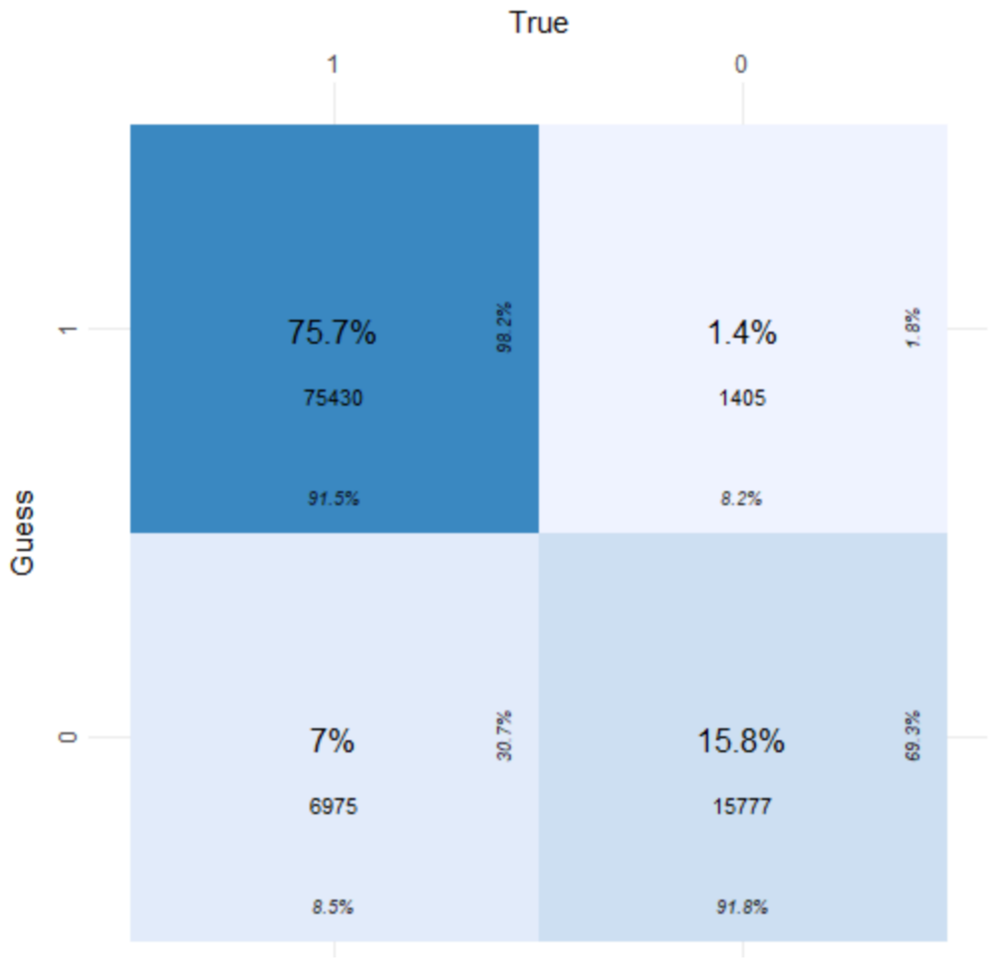
Predicted \ Actual	Class1	Class2
Class1	17919	7143
Class2	2058	68902

DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.897	0.906	0.715	0.897	0.796

Accuracy	Kappa
0.904	0.734

Alternative approach: random forest and logistic regression



Model	Accuracy	F1_score	Computation time (per 1kk rows), seconds
Neural net	90,4%	80,1%	757
Random forest	92,8%	81,4%	318
Log regression	93,3%	83,1%	108

We approached the same task with random forest and logistic regression, and achieved the same accuracy and F1_score results, but faster.

In terms of computational speed, the efficiency of random forest is two times higher, while results are even slightly better.

Traditional classifier logistic regression has beaten neural networks and random forests in terms of result/speed ratio.

Main conclusions

Interpretability, controllability, the possibility of automatic selection of informative features of decision trees, and regressions were the reason for their use as the primary tool for efficient classification of processing large amounts of data, searching for atypical values for subsequent filtering, and identifying erroneous values in categorical variables.

Due to human disabilities, to analyze the reliability of a large and diverse set of data, expanding the field of applied machine learning methods will increase the quality of data and decisions made on their basis.

When solving classification problems, metrics should be monitored carefully and problems solved under business logic.

With unequal classes, metrics should be selected carefully, and up-sampling or down-sampling applied when necessary.

Simpler models often give more balanced and correct results during cross-validation on test data.

Bank of Russia

THANK YOU

Work email: dyachkovdv@cbr.ru
Personal email: d.djachkov@gmail.com

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Classifying payment patterns with artificial neural networks: an autoencoder approach[1]

Luis Gerardo Gage and Raúl Morales-Resendiz,
Centre for Latin American Monetary Studies (CEMLA)

John Arroyo and Jeniffer Rubio, Banco Central del Ecuador;

Paolo Barucca, University College London

[1] This presentation 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 central banks and other institutions represented at the event.

An extended version is published in the Latin American Journal of Central Banking 1 (2020) 100013.

Classifying payment patterns with artificial neural networks: an autoencoder approach[1]

Jeniffer Rubio[2], Paolo Barucca[3], Gerardo Gage[4], John Arroyo[5] and Raúl Morales-Resendiz[6]

Abstract

Payments and market infrastructures are the backbone of modern financial systems and play a key role in the economy. One of their main goals is to manage systemic risk, especially in the case of systemically important payment systems (SIPS) serving interbank funds transfers. We develop an autoencoder for the *Sistema de Pagos Interbancarios* (SPI) of Ecuador, which is the largest SIPS, to detect potential anomalies stemming from payment patterns. Our work is similar to Triepels-Daniels-Heijmans (2018) and Sabetti-Heijmans (2020). We train four different autoencoder models using intraday data structured in three time-intervals for the SPI settlement activity to reconstruct its related payments network. We introduce bank run simulations to feature a baseline scenario and identify relevant autoencoder parametrizations for anomaly detection.

The main contribution of our work is training an autoencoder to detect a wide range of anomalies in a payment system, ranging from the unusual behavior of individual banks to systemic changes in the overall structure of the payments network. We also found that these novel techniques are robust enough to support the monitoring of payments' and market infrastructures' functioning, but need to be accompanied by the expert judgement of payments overseers.

Keywords: Market Infrastructure, Neural Network, Anomaly Detection, Autoencoder, Artificial intelligence, Retail Payments, Machine Learning.

JEL classifications: C45, E42, E58.

[1] Authors are responsible for the statements found in this research article, and do not represent those of the institutions to which the authors are affiliated. Any errors are their own responsibility. A special acknowledgement to Dr. Serafín Martínez-Jaramillo, CEMLA, and Dr. Ronald Heijmans, De Nederlandsche Bank, for their valuable comments and suggestions

[2] Banco Central del Ecuador (BCE)

[3] University College London (UCL)

[4] Centro de Estudios Monetarios Latinoamericanos (CEMLA), México.

[5] Banco Central del Ecuador (BCE)

[6] Centro de Estudios Monetarios Latinoamericanos (CEMLA), México.

1. Introduction

Financial market infrastructures (FMIs) underpin the financial system and the economy by enabling multilateral transactions under certain rules and common platforms. They entail by design financial and operational risks related to interbank funds transfers. Given their systemic importance, central banks need to be able to monitor their activity and to identify anomalous events, and for these purposes artificial neural networks result purposeful.

According to the CPMI-IOSCO Principles for Financial Market Infrastructures (PFMI), payment systems and FMIs should be properly designed to support their participants to manage and mitigate risks more efficiently, and have better liquidity management. Their performance is key to support the financial system's health. Indeed, when payment processing rules or arrangements are not clear or comprehensive, payments and FMIs participants could take unnecessary. Likewise, if platforms are not resilient, the entire network could be endangered by cyber threats. In fact, poorly designed and operated payment systems and FMIs can contribute to exacerbate systemic crises. Contagion risk could impact the overall stability of the financial system. Therefore, the payment systems and FMIs must be robust and reliable, available even in times of stress (CPMI-IOSCO, 2012).

Monitoring payments and financial market infrastructures is one of the primary objectives for central banking to ensure that the above events take place. By overseeing the functioning of FMI and SIPS they can identify and address systemic risk events. Central banks have long worked in establishing an appropriate monitoring and risk management framework for SIPS, and other prominent payment systems and FMI. Oversight is also thought as a relevant task that fosters FMI good performance. This task is supported by several quantitative and qualitative tools, including international standards such as the PFMI, risk policies, Business Intelligence and other software tools for liquidity and collateral monitoring, business continuity plans, among others. Yet understanding the complexity of FMI and SIPS requires a powerful and well-designed toolkit. (CPMI-IOSCO, 2012)

In light of this challenging task, we present an application of artificial neural networks for outlier detection, tailored to payment systems and FMIs. In our work, we focus on the major FMI in Ecuador, the *Sistema de Pagos Interbancarios* (SPI), managing both wholesale and retail payment transactions. The SPI is a hybrid payments system performing Real-Time Gross Settlement and Deferred Net Settlement features.

For the purpose of our work, payment systems can be efficiently represented as directed networks, where institutions are nodes and payment flows from an institution to another one constitute edges. There are payment flows or patterns of payment flows, i.e. network patterns, that can pose a systemic risk in case a particular participant, or a set of participants, is unable to settle their transactions in a specific time interval. From an oversight point of view, it is crucial to be able to determine a normal pattern in a payment network as well as to identify risky events arising from anomalous patterns.

In our work, we introduce and test a pattern recognition tool for anomaly detection based on an autoencoder architecture. An autoencoder is an unsupervised feed-forward neural network that aims to reconstruct the input data at the output layer by passing through a lossy compression process that creates a lower-dimensional representation. This makes the model learn the most important features inherent to the data. Once the autoencoder is trained and has learned the usual patterns, the anomaly detection is done through flagging those instances that have high reconstruction errors, which perhaps is an indicative of abnormal patterns.

We trained four autoencoder models to identify common and anomalous payment patterns of financial entities (participants) in the SPI payments network. We present the results of the models that consist of anomalies between normal and unknown patterns of payment flows. These results can significantly contribute to oversight experts, to establish an alerting system and to anticipate potential risks in the SPI. The detailed set up, training and testing of our models is further explained in the methodology section where we also provide information on the different architectures, dataset partitions' and data preprocessing taken.

Our work is related to Triepels-Daniels-Heijmans (2018), Triepels-Heuver (2019), and Sabetti-Heijmans (2020). They also developed autoencoder approaches for both wholesale and retail payment systems, by training different models using daily payment flows, in some cases to identify financial stress in entities that have gone bankrupt.

The main contribution of our work relates to training autoencoders able to detect a wide range of anomalies in the SPI, ranging from spotting the anomalous behavior of individual banks to detecting changes in the overall activity of the payments network. Our work highlights that these novel techniques are robust enough to support payments' and market infrastructures' oversight, and ultimately to monitor financial stability, but need to be always in tandem with the expert judgement of central banking overseers.

The remainder of the work is structured as follows. Section 2 surveys relevant literature that is closely related to our work. Section 3 describes the *Sistema de Pagos Interbancarios* and also provides a statistical analysis of the data. Section 4 introduces the methodology and autoencoder setup. Section 5 presents key results on the autoencoder performance and the bank run simulations. In Section 6 we discuss how our work can be further advanced.

2. A brief survey of literature

Detecting outliers in a dataset is an old challenge in statistics (Edgeworth, 1887). Although the definition of anomaly can vary across different disciplines, the underlying statistical definition of anomaly is the same, i.e. a subset of data that behaves according to different patterns with respect to those identified as the normal ones. This general definition suits perfectly the logic of many known machine learning algorithms. Not surprisingly, these algorithms have been developed and applied within many domains, e.g. intrusion detection, fraud detection, fault detection, as well as medical anomaly detection (Chandola, 2008).

Hodge and Austin (2004) described three fundamental approaches for outlier detection. *Type 1* where the outliers are determined without any prior knowledge, this approach is analogous to unsupervised learning, where the learning algorithm is provided with an unlabeled dataset and it aims to find hidden patterns within the data. *Type II* requires pre-labeled data targeted as normal or abnormal, this approach is analogous to supervised learning, where the learning algorithm objective is to fit a function that reproduces the behavior of pre-labeled data in order to make predictions on unseen data. And, *Type III* that requires also pre-labeled but it only models normality - and in few cases abnormality - helping to define a boundary for normality, this approach is analogous to semi-supervised learning, which is in the middle ground between supervised and unsupervised learning and can use both labeled and unlabeled data during the learning process. Hodge and Austin described different techniques from three fields: statistics, neural networks and machine learning. One can also find hybrid systems that combine techniques from these fields. As our work analyzes a dataset that *a priori* do not contain labels on what is and what is not an anomaly, our methodology can be classified as *Type I* at first glance, i.e. analogous to unsupervised learning. But given that after the fitting process we create bank run simulations, which can be considered in a way as anomalies, to test the capacities of the autoencoder to identify them, it is rather appropriate to state that our methodology falls within *Type III*, i.e. analogous to semi-supervised learning.

A challenge for outlier detection is associated with the presence of high dimensionality in the data. There are different approaches (Barnett and Lewis, 1994; Arning et al., 1995) to dimensionality reduction, including clustering methods (Knorr and Ng., 1998). One way to tackle high dimensionality is to make parsimonious projections in lower spaces and then proceed with the anomaly detection as proposed in Aggarwal (2001). One of the main properties of the data in this paper is that the number of features is close to the number of observations, thus high dimensionality needs to be taken under consideration; remarkably, an advantage of the autoencoder is that - by design - the encoding creates a lower representation of the data that learns the most relevant features.

Autoencoders have been previously used for outlier detection. Hawkins et al., (2002) developed a methodology for an autoencoder with two different datasets, one dataset for network intrusion detection and the other for breast cancer identification. In the first case all outliers were identified and, in the second case over 75% out of the total, showing the robustness and transferability of the methodology.

Another case in which an autoencoder is implemented to detect anomalies is found in Williams et al. (2002). Their results were compared with three techniques: i) the Donoho-Stahel estimator, ii) an outlyingness estimator proposed in Hadi (1994) based on both the means and covariances of the variables and the Mahalanobis distance, iii) and a model of mixture-models clustering. The comparison was done fitting and testing the techniques on many datasets, where each dataset contained labels that identified the abnormal instances; the datasets relates to information from different areas such as breast cancer, internet intrusions and other topics. The results show that for small datasets, the compared techniques show a good level of performance, with clustering being the one that presented the most difficulties in detection, but in the case of longer datasets, the autoencoder showed better performance identifying anomalies.

In the context of finance, applications of anomaly detection are found in Aleskerov et al. (1997), Ghosh and Reilly, (1994), Dorronsoro et.al. (1997), and Baruse et al. (1999), and are mainly concerned with credit card fraud detection. Nevertheless, the application of unsupervised machine learning methodologies for payment systems' oversight is relatively new among central banks and relevant authorities. In a recent series of papers (Triepels-Daniels-Heijmans, 2018; Sabetti-Hejmans, 2020), the autoencoder architecture has been shown to be effective for learning patterns of normal transaction data and to detect anomalous payments, using the autoencoder reconstruction error. In Triepels-Daniels-Heijmans (2018) two autoencoders with one hidden layer were trained - one used a linear activation and the other one used a sigmoid activation in the hidden layer. They used data from TARGET2 (the RTGS for the Eurosystem) settlement system to reconstruct liquidity-related information and also introduced a bank run simulation. The paper reported that the data presented relevant features of the payments network enabling the autoencoders to detect changes in the payments flow behavior. Sabetti-Heijmans (2020) compared the performance between one hidden layer and two hidden layers autoencoders, using data from the Canadian ACSS (a Canadian retail payment system). They found that the one hidden layer autoencoder had lower validation error compared with the two hidden layers autoencoder, but the two hidden layers autoencoder displayed a lower variance that can lead to better results when using testing data.

Our work follows the general approach of Triepels-Daniels-Heijmans (2018) and Sabetti-Heijmans (2020), and it contributes to the literature by analysing a new dataset for the Ecuador SPI and presenting a detailed review of alerts, illustrating the ample range of anomalies that can be detected by the autoencoder.

3. The *Sistema de Pagos Interbancarios*

3.1 SPI main features

The Central Bank of Ecuador (BCE) must provide the physical and electronic means of payment necessary for the proper functioning of the country's economy. In this respect, the BCE is the owner and operator of several payment systems, which as a whole are known as the Central Payment System (SCP). The SCP entails the interbank funds transferring system for large value payments and it also supports settlement of private retail payment systems and securities clearing and settlement systems. Thus, the SCP represents the most relevant payment infrastructure for Ecuador.

The underlying system that makes up the SCP is the *Sistema de Pagos Interbancarios* (SPI). The SPI settles 60% of the total payments in the SCP, for which reason, this system is deemed as the major SIPS in Ecuador. In light of its importance, our work focuses on the SPI activity. The relevance of the SPI is paramount. It provides an infrastructure for different types of participants. Within this universe, there are banks with a higher activity and make payments with the rest of the SPI participants. Some SPI participants only make transactions with few entities in the payments network. It also channels all Government payments as well as 98% of wholesale and retail payments from the private sector. On average, the SPI processes 300,000 transactions per day totaling USD 450 million.

In terms of the clearing and settlement mechanism, the SPI makes the settlement of payments in three daily time-intervals. Each time interval represents an intraday settlement period for the interbank payments ordered by the financial entities in the SPI. These time-intervals are carried out at three different hours (08:30, 11:00 and, 16:30). Each time-interval is exclusive of the other, the net amounts between financial entities are cleared and settled at the end of each time interval. The SPI only settles the operations of financial institutions that have liquidity in their accounts at the BCE to cover their net debit position at the time of the settlement of each of these time intervals, otherwise the financial institution is excluded from the process. The latter in order to avoid liquidity risks for the entire SPI participants.

The SPI as any other payments and market infrastructure is subject to operational and financial risks. Technological advances such as malicious intruders or operational events experienced by a single participant, can both represent a major risk for the SPI and its participants, with undesirable negative effects in the financial system and, ultimately, the economy.

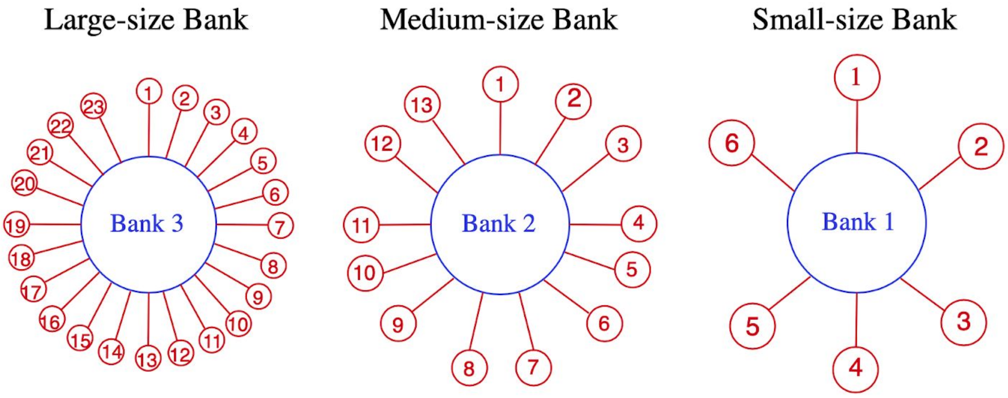
Developing an automated oversight tool to detect atypical payments or payment behavior is a significant contribution to better identify malicious activity in SPI and other prominent payment infrastructures. Such an alert system should be also able to allow the monitoring authorities and the own operator to understand normal behavior of financial institutions as they participate in the SPI. For such purposes, we work on the autoencoder feedforward neural network to anticipate and identify potential risks in this systemically important market infrastructure of Ecuador.

3.2 The dataset

Since the SPI implementation in 2002, the total amount and number of transactions settled in this system have grown year after year. In 2018, the SPI settled over USD 100 billion with a corresponding number of transactions of almost 70,000, representing a daily average of USD 400 million and 300 thousand transactions. For the purpose of our work, we used transaction information from 24 financial institutions in 2018, which represent around 90% of the amount channeled by SPI by the private sector.

As seen in Figure 1, the typical flows for a large, medium, and small bank can significantly vary. In our work, we consider a subnetwork of payments, i.e. the maximum number of flows that a financial institution can have is 24, reflecting the fact that it can send payments to the rest of the 23 banks and to itself.

Figure 1. Payment connections in the SPI for large-, medium- and small- banks

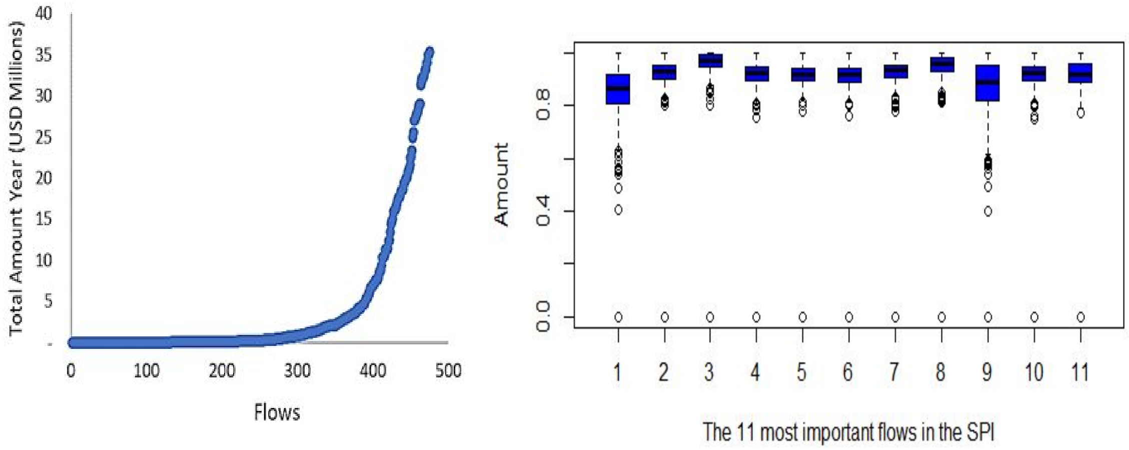


For the purposes of this investigation, 741 time intervals corresponding to 247 working days of the year 2018 were used. In our analysis, each payment flow (i.e. connection) represents the exchange of interbank payments between bank A and bank B. If we analyze the frequency of participation of a bank in the SPI, on average, the analyzed payments flows, i.e. interbank payments by SPI participants, take place in nearly half of the 741 time intervals, that is 376 time intervals. A fraction of 25% of the analyzed flows appeared more than 731 times, in effect these are particularly recurrent transactions for 2018. Conversely, 25% of payment flows only took place in 56 times intervals. A small set of unique payment flows occurred only once or up to 15 times.

The average volume of payments per time interval amounts to USD 76 million. The maximum amount for 2018 time-interval was USD 212 million and the minimum, USD 31 million. Nearly 75% of the intervals registered payments for over USD 60 million, each.

Considering that there are interbank payment flows (i.e. payments between Bank A and Bank B) among all 24 banks, there can be a total of 576 (24 banks x 24 banks) possible connections. However, the SPI dataset shows that for 101 connections there was no single exchange. Therefore, we only analyzed a total of 475 connections for 2018. We can observe that there is an average of 241 connections for each time interval along the year. The lowest number of connections for a time interval was 47 payment flows, while the maximum, 298 flows. It is worth mentioning that in 95% of the intervals there were more than 207 flows. The average payment per flow was USD 2 million over 2018, while the maximum value for a payment connection reached USD 4 billion.

Figure 2. Payment flows in the SPI



A. Average value

B. Largest flows (USD millions)

Figure 2A represents the average amount of all the flows of the SPI in 2018. The 75% of the connections amounted to over USD 130 billion, 25% of the flows amounted to more than USD 20 million. Figure 2B is a boxplot of the 11 most important flows in the SPI for payments made in the year. These flows represent around 50% of the total amount of payments made by the SPI in 2018.

The SPI dataset includes large and small, as well as more and less frequent, payment flows for 2018. However, for anomaly detection, each connection can be important regardless of the magnitude or frequency of interbank payments. There are payment flows that can pose a systemic risk in case a particular participant is unable to settle them in a specific time interval. From a payments system oversight point of view, it is crucial to be able to determine a normal behaviour and to identify risky events arising from anomalous behaviours. With this goal in mind, in the following sections, we introduced and tested a pattern recognition tool for anomaly detection based on an autoencoder architecture.

4. Methodology

In this section we begin by stating the general anomaly detection framework in the context of a payment system. First we define the basic structures that will be used, such as the set of participants, the time intervals, and the matrix that represents the interactions between participants. Once the above has been defined, the next step consists of the setup of the anomaly detection task, which will be based on the measurement of the reconstructions' quality made by a compression model.

The section continues with a detailed description of the autoencoder, as the compression selected model, providing details of its operation and why it can be used to detect abnormal patterns in the data. The section concludes with a discussion on the preprocessing of the data that is made prior to the training of the models, to then continue with a review of the adjustment and testing process of the models to finally introduce the bank run simulations.

4.1. Definition of the general framework for the anomaly detection task

Following Triepels (2018), let $B = \{b_1, b_2, ..., b_n\}$ be the set of SPI participants that settles transactions between them. Now let us consider $T = \{t_1, t_2, ..., t_m\}$ an ordered set of m time intervals where each $t_i = [\tau_{i-1}, \tau_i)$ having that i ranges from 1 to m, and where τ_i are specific timestamps delimiting time intervals. In our case each time interval represents one of the three intervals taking place in a SPI working day.

Then, we define the structure for liquidity transmission among institutions within different time intervals. Let $a_{ij}^{(k)}$ be total amount of liquidity transferred from institution b_i to institution b_j within the time interval t_k . The liquidity matrix $A^{(k)}$ accounts for the liquidity transferred between all the institutions within the interval t_k:

$$A^{(k)} = \begin{bmatrix} a_{11}^{(k)} & \cdots & a_{1n}^{(k)} \\ \vdots & \ddots & \vdots \\ a_{n1}^{(k)} & \cdots & a_{nn}^{(k)} \end{bmatrix}$$

The diagonal elements $a_{ii}^{(k)}$ indicate the total amount of liquidity that b_i transfers between its own accounts. The elements of $A^{(k)}$ can be interpreted as the weights of a network where nodes are institutions and edges are liquidity flows (payments). In order to feed the model with a simple data structure, every $A^{(k)}$ is mapped to a liquidity vector $\overline{a}^{(k)}$ with the form:

$$\overline{a}^{(k)} = [a_{11}^{(k)}, ..., a_{n1}^{(k)}, ..., a_{1n}^{(k)}, ..., a_{nn}^{(k)}]^T,$$

where $\overline{a}^{(k)}$ is a n^2 column vector that consists of $A^{(k)}$ appended columns.

The data we study contains information on the liquidity vectors for a series of periods at a payments system. We aim to reconstruct the vectors by compressing and decompressing the dataset under use. For this purpose, we implemented a lossy compression, which generates a particular type of representation that allows the data to not be exactly learnt and some of the information to be lost. When this type of compression is implemented, the relevant patterns present in the data are learned. Once the compression model learns the common patterns, that is, the most observed, and a new liquidity vector is fed for its reconstruction; if the reconstruction is bad, this is explained by the fact that vector information differs from the normal patterns that the model learned, indicating the possibility of a potential anomaly. The quality of the reconstructions will be measured through the reconstruction error, in other words, the differences between values yielded by the lossy compression and the real vector values.

More formally, given a lossy compression model, let RE be the non-negative function that measures the reconstruction error of liquidity vector $\overline{a}^{(k)}$; $RE : D \rightarrow [0, \infty)$ where D is the set of liquidity vectors for all time intervals. Our main objective is to find all the liquidity flows (i.e. payment connections) in a particular time interval corresponding to reconstruction errors greater than a given threshold $\varepsilon > 0$, i.e., given a set of liquidity vectors D we aim to find the set $F = \{\overline{a}^{(k)} \in D \mid RE(\overline{a}^{(k)}) \geq \varepsilon\}$. It is noteworthy that there is no rule or methodology to follow in order to set the value for ε, instead it has to be set according to the particular characteristics of the data and prior knowledge on the respective payments system.

4.2. Autoencoder modeling

For our work, we select the autoencoder as the lossy compression model. The autoencoder falls in the category of artificial neural networks techniques. The basic unit of a neural network is the neuron or node, which can be both fed directly with the data or fed through other connected neurons, and depending on the type of neuron it will be defined how the information will be processed to generate an output. As can be seen in Figure 3, the neurons of one layer connect with those of the next but never between them. All neural networks have an input layer, an output layer, and at least one hidden layer (the case of having more than two hidden layers it's considered as deep learning). The types of neurons are:

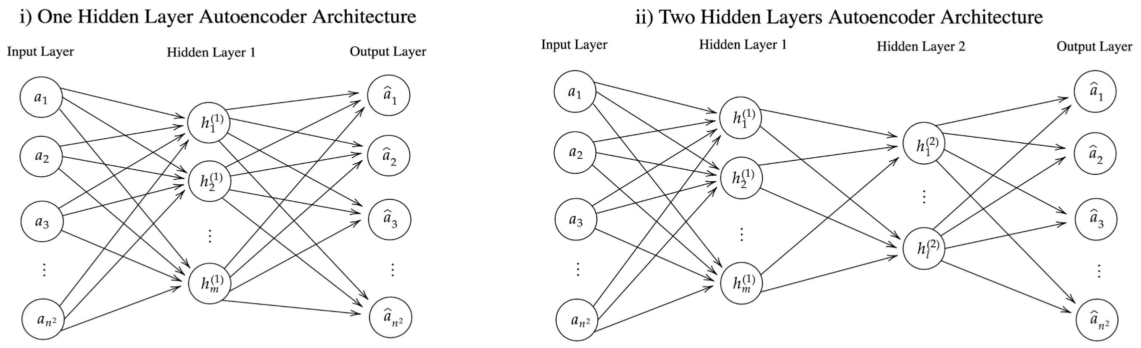
- *Input neuron*: They are fed with the data directly and this conforms the output that feeds the next layer.
- *Hidden neurons*: Each of them is fed by all the neurons of the previous layer ($x_i's$) and multiplied by a set of weights ($w_i's$). The output of these neurons is generated by first computing a weighted sum of the weights and the outputs of the previous layer and then applying a function f to it called activation function, i.e., the output is given by $f(\sum_i w_i x_i + b)$. The term b refers to the bias that data poses.
- *Output neurons*: It follows the same process as hidden neurons, but the output it generates rather than feeding other neurons is the resultant final prediction.

The weights $w_i's$ decide how much of the information in each neuron should be transmitted to the next layer; these are the parameters that will be learnt during the training process. The *activation function* f, has the purpose to learn non-linear relationships between the components of the data.

We utilized two different activation functions, Rectified Linear Unit (ReLU), namely $ReLU(x) = max(0,x)$, and the hyperbolic tangent (Tanh). The latter maps the value to the interval (-1, 1) and falls within the category of sigmoidal functions (s-shape) which provides a simple model for the firing of a real neuron. An issue regarding sigmoidal functions is that derivatives can become very small far from zero, affecting the learning process which is based on gradient methods; ReLU overcomes this issue - being linear for positive values - and also its computation is simpler, but if during the learning process the weighted sums gets below zero, then most of the neurons in the neural network will go to zero, potentially leading to non sensitivity and poor fitting[7][8].

The autoencoder is made up of two components, the encoder and the decoder. The encoder is the initial part of the autoencoder and it has the task to create an accurate lower-dimensional representation of the data. The second part of the autoencoder, the decoder, is in charge to carry out the reconstruction of the data. The encoder goes from the input layer to the layer with the lowest number of neurons, which is commonly called bottleneck given that is the layer of the network where data is the most compressed. The encoder can be represented as a function $h = f(X)$, where X represents the input data[9]. On the other hand, the decoder goes from the *bottleneck* to the output layer, it can be represented by the function $r = g(h)$. The autoencoder final objective is to find f and g such that $X \approx g(f(X))$.

Figure 3. Autoencoders Architectures


i) One Hidden Layer Autoencoder Architecture

ii) Two Hidden Layers Autoencoder Architecture

In Figure 3 we can observe the architecture for two autoencoders, one hidden layer (left panel) and two hidden layers (right panel), where $m < n^2$ and $l < m$; this tells us that the input data

[7] An alternative to overcome the issues that arise from the use of ReLU as activation function is to use the Leaky ReLU that has the same value for the non-negative values but for a negative variable (x) it assigns the correspondent value 0.01x. The use of this ReLU variation is left for future work.

[8] A further description on feed-forward neural network, components and learning process can be found in (Goodfellow et. al. 2016).

[9] Input data is represented by the liquidity flows.

will be compressed through a projection from a n^2-dimensional space to a m-dimensional space, for the one hidden layer, having an extra compression from m dimensions to l dimensions in the case of having two hidden layers. Adding one more layer to the autoencoder compresses the bottleneck, forcing the neural network to learn a lower dimensional representation. Yet adding many layers increases the number of parameters and can cause the neural network to overfit the training data, thus failing to generalize. In our study we trained both one and two hidden layers autoencoders.

The learning of the autoencoder, and in general for neural networks, is achieved by the minimization of a cost or loss function with respect to the weights and biases (mentioned above). For our work, this function will correspond to the reconstruction error (RE) of our lossy compression. More precisely, the reconstruction error will be given by the mean of the squared differences between the liquidity vectors and its reconstruction, in other words the the Mean Squared Error (MSE), for all time intervals, i.e. the loss function will depend on the set D:

$$RE_D = \frac{1}{m} \sum_{k=1}^{m} (\overline{a}^{(k)} - g(h(\overline{a}^{(k)})))^2$$

The autoencoder's weights are learnt through mini-batch backpropagation[10]. We strived to fine-tune the autoencoder's hyper-parameters to improve its performance. This is accomplished through cross-validation, which performs an exhaustive search within a predefined set of hyper-parameters for multiple data partitions, where the performance of each hyper-parameter setting is evaluated. We also carry out pre-processing of the data, which can lead to a significant improvement in the performance. This is discussed in more detail in the next subsection.

4.3. Model fitting, selection and testing

In this subsection we describe the procedure to fit the autoencoder, which involves the preprocessing and partition of the data, and the training and validation steps.

Before fitting the model, we pre-processed the data, this corresponds to a log-transformation that was followed by a min-max standardization. The former have the purpose to reduce the skewness in the payments flows, while the latter maps the values to the interval $[0, 1]$, to give the same degree of importance to all the bilateral transactions and avoid the autoencoder to be unbalanced toward the transactions with highest value. In such a pre-processing, let V be a feature (in our case we have 576 features, each one corresponding to the flow of liquidity between one institution to another) the log-transformation of V is done by applying the natural logarithm to it, this is computed for all the features.

After this step, we continued with the min-max standardization, for this end, we first found the maximum and minimum values for feature V, to then transform each feature element by:

[10] Backpropagation is a learning mechanism, based on gradient descent, which is widely used for the training of neural networks. The mini-batch indicates that the updating of the parameters learned is done after passing not the complete dataset or a single instance (that currently are another types of backpropagation), but a portion of the whole dataset, to the network; a further insight can be found in (Goodfellow et. al. 2016).

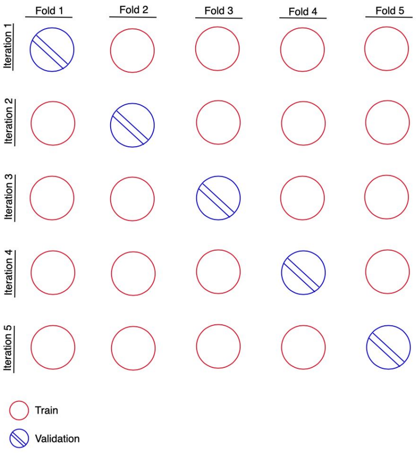
$$minmax(x_i) = \frac{x_i - min(V)}{max(V) - min(V)}$$

Where x_i is an element of V. The aforementioned steps led to smaller training and validation reconstruction errors than when the data was fed in its original form.

Once the data is preprocessed, we make a partition of it. It's a common practice in machine learning to separate the whole data into different subsets. In our work, we first randomly divide the data in two parts, the first, which will be used to perform cross-validation, is the larger and contains 80% of the original data. The remainder 20% of the data, commonly called the test set, is used in order to evaluate the performance of the model with data not observed during training or validation.

The cross-validation is performed once the data is divided in subsets, to assess the performance of the different hyper-parameters, with the ultimate goal of setting the best model configuration. This is mainly guided by using as selection criterion the validation MSEs. More specifically, cross-validation as a re-sampling technique is useful to evaluate the effectiveness of machine learning models. For our work, we implemented K-fold cross-validation with 5 folds. As mentioned above, the implementation is performed on the 80% of the whole dataset. In sum, the cross-validation process can be divided into the following: first we randomly divide the data into 5 disjoint groups, or folds; then, a hyper-parameters configuration is chosen (e.g. an autoencoder with one hidden layer with 100 neurons); in the next step which corresponds to the first iteration, the model is trained in four groups and the validation is done in the remaining group; and, the iteration ends with the computation of training and validation MSEs. The following iterations redo the same process, but as can be seen in Figure 4, the difference lies in the training and validation sets that are used.

Figure 4. 5-fold Cross-validation



Once all the iterations are completed, the mean of both training and validation MSEs is computed and this will be the performance of the model with the previously hyper-parameters chosen. The above process is carried out for each element of the set of hyper-parameters to be tested.

The determination of the best configuration of a model is made based on the validation MSEs. Generally, the model that yields the lowest MSE is selected, but it may be the case that the best model is too complex, ie, that the number of neurons in the hidden layers is very large, and one will prefer to choose a model with less complexity where the difference between their MSEs is not significant.

After performing cross-validation and the best hyper-parameters configuration is selected, the model is re-trained in the 80% of the data. In the last step the model is fed with unobserved data, which corresponds to the test set, and the instances that show highest reconstruction errors are identified.

4.4. Bank run simulations

Given that the SPI dataset presents a small amount of anomalies that could pose a level of uncertainty about its abnormality, we perform a series of bank run simulations similar to Triepels-Daniels-Heijmans (2018) and test whether the autoencoder was able or not to flag them as anomalies. The simulations were done by randomly choosing an institution b_i, then all its outgoing flows for a given period are modified according to:

$$a_{ij}^{(k)} \to a_{ij}^{(k)} + (B(k) \cdot E(k))$$

Where k is the time period, $B(k) \in \{0, 1\}$ was sampled from a $Bernoulli(p)$ random variable and decides whether extra liquidity will be added or not to the current period, $E(k) \in [0, \infty)$ was sampled from a $Exponential(\lambda)$, which decides how much extra liquidity will be added to the payment flow. The parameters p and λ determines the intensity of the bank run, the greater, the more intense the bank run will be. The autoencoder is expected to be less able to reconstruct the payment networks arising from these simulations, indicating that they are displaying anomalous behavior, in this case a bank run.

5. Results

In this section we first present the models we trained, highlighting the changes for each model related to the network architecture and the activation function. This is followed by a summary of the performance and results of every model. Next, we show the analysis of the anomalies that were detected in all or the majority of the models and that were found relevant for oversight purposes.

5.1. Results from autoencoder's fitting and performance

We analyze different setups for the autoencoder by varying the number of hidden layers, the number of neurons in each layer and the activation functions, this led us to the definition of four different models:

- ➔ Model 1: One hidden layer with TanH as the activation function
- ➔ Model 2: Two hidden layers with TanH as the activation function
- ➔ Model 3 One hidden layer with ReLU as the activation function

- ➔ Model 4: Two hidden layers with ReLU as the activation function

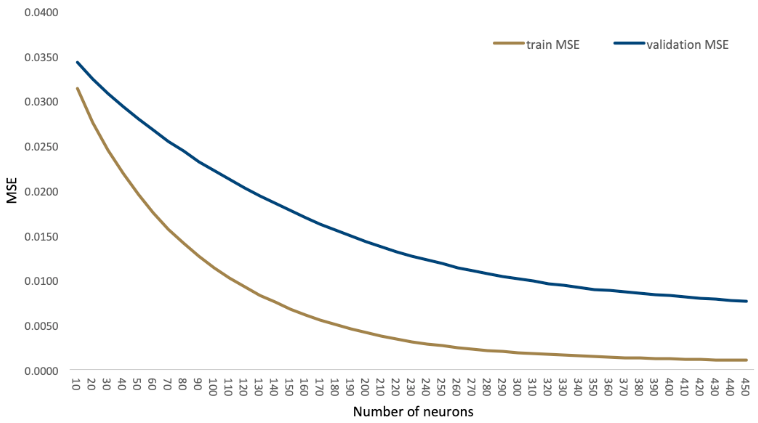
As described in subsection 4.3, each of the models were evaluated using 5-fold cross-validation to determine the optimal number of neurons for the hidden layers, the values proven for each model are summarised in Table 1. It is noteworthy that the autoencoder performs a compression, this can be observed in the reduction of number of neurons between the input layer and the hidden layers. In the case of one hidden layer we have gone from a 576-dimensional space[11] to a new space whose dimensions can range from 10 to 450. In the case of two hidden layers we have a double reduction, first a reduction similar to the case of one hidden layer takes place (now dimensions ranges from 10 to 400), then the autoencoder is forced to generate a smaller space with 8,16 or 32 dimensions. This compression enables the model to learn only the most important characteristics inherent in the data.

Table 1. Summary of the configurations evaluated for each model

	Activation Function	Neurons in input layer	Neurons in first hidden layer	Neurons in second hidden layer	Neurons in output layer
Model 1	TanH	576	(10, 20, 30,..., 450)	----------------	576
Model 2	TanH	576	(10, 20, 30,..., 400)	(8, 16, 32)	576
Model 3	ReLU	576	(10, 20, 30,..., 450)	----------------	576
Model 4	ReLU	576	(10, 20, 30,..., 400)	(8, 16, 32)	576

Figure 5 shows the Model 1 (with one hidden layer) validation and training errors. We set the final configuration at 300 neurons, which is the point where the error has no substantial reduction.

Figure 5. One Hidden Layer with TanH performance for different number of neurons

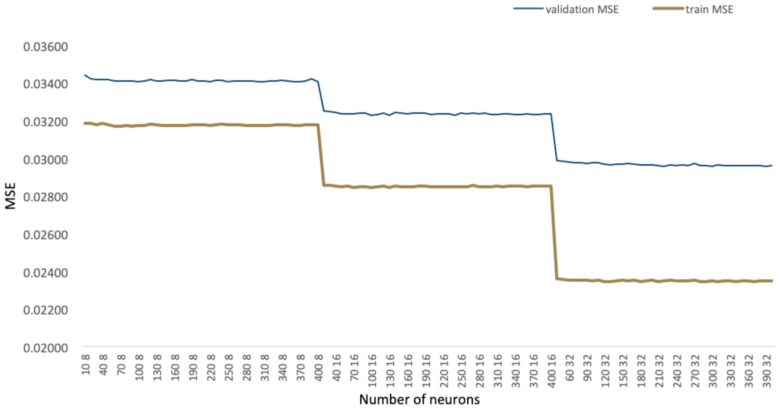


For the case of Model 2, Figure 6 shows a stepwise decrement of the errors that relates to the number of neurons used in the second hidden layer, where it can be said that having 32 neurons in the second hidden layer is the best decision. For the first hidden layer, here it is observed a

[11] The 576 dimensional space corresponds to all our features which are the 24x24 participants' settlement interactions.

continuous decreasing behavior, where the inflection point corresponds to 220 neurons. In light of this training results, the final configuration of Model 2 is set to have 220 neurons in the first layer and 32 neurons in the second layer.

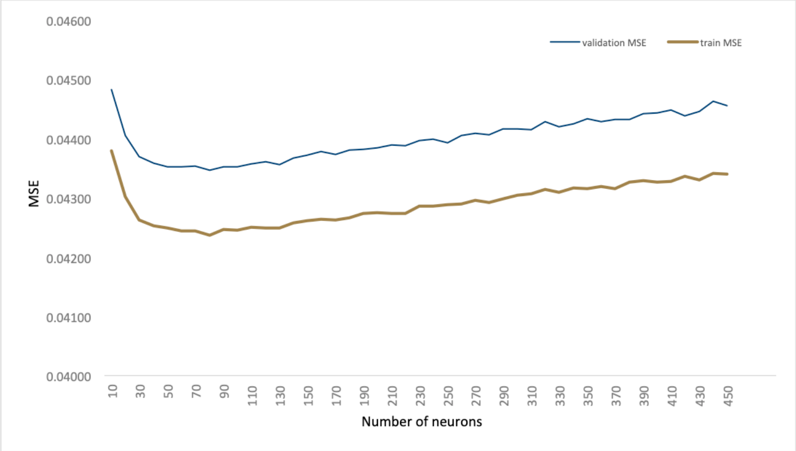
Figure 6. Two Hidden Layers with TanH performance for different number of neurons



In Figure 7 is shown the performance of Model 3, which corresponds to one hidden layer with ReLU activations. It can be observed that from 10 to around 80 neurons, the training and validation errors present a decreasing trend, but for the rest of the neurons the errors only increase, for that reason we decided to set the final configuration of the model using 80 neurons in the hidden layer.

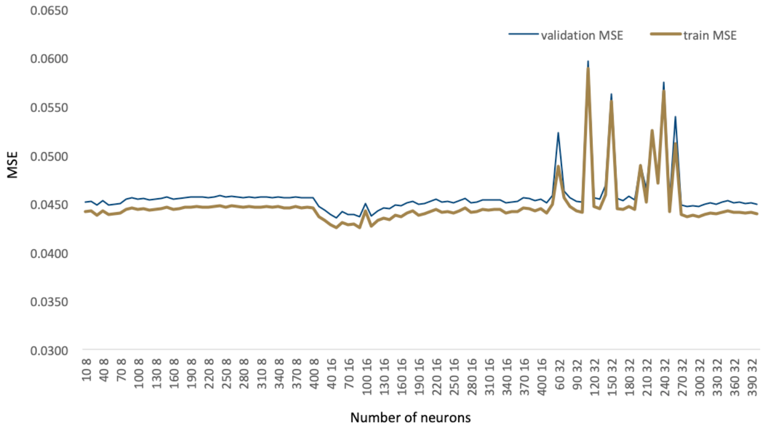
For Model 4, Figure 8 shows its performance. It can be noted that the behavior of the errors is stable for the configurations where the second hidden layer has 8 neurons, presenting only small variations. Then, for the configurations with 16 neurons in the second layer, a decrease in errors is observed, reaching the minimum with 110 neurons in the first layer; after this point a slight increase in errors is observed. In the case of 32 neurons in the second layer, the performance begins to be unstable, it is possible to observe peaks where the errors have large increases with respect to the rest of the configurations and where even the training and validation errors are the same[12].

Figure 7. One Hidden Layer with ReLU performance for different number of neurons



[12] This is not desirable given that the training data is the one used to adjust the model, thus it should show smaller errors than the validation data which is not used for the training.

Figure 8. Two Hidden Layers with ReLU performance for different number of neurons



Once the optimal configuration for each model was selected we proceed to re-train the models on the whole set used for cross-validation. It follows the feeding of the models with unseen data. The behavior of the reconstruction errors for the liquidity vectors belonging to the test set is expected to be stable and low and that only a few of them are above the average. This is given by our assumption that most of the data has a normal behavior and that only a few instances will perform abnormally; if the above does not happen, it means that our model did not perform correctly the reconstruction. This will imply that more data need to be used for training or that the chosen configuration is not adequate.

Figure 9 shows the results of the test set reconstruction for Model 1 and Model 2. It can be observed that both models identified six time intervals where the reconstruction error is higher than the average, but the rest of observations present a stable reconstruction with the difference that Model 1 shows higher variance and that Model 2 has bigger overall reconstruction errors.

Figure 9. Errors corresponding to the reconstruction of test set for Model 1(left) and Model 2 (right)

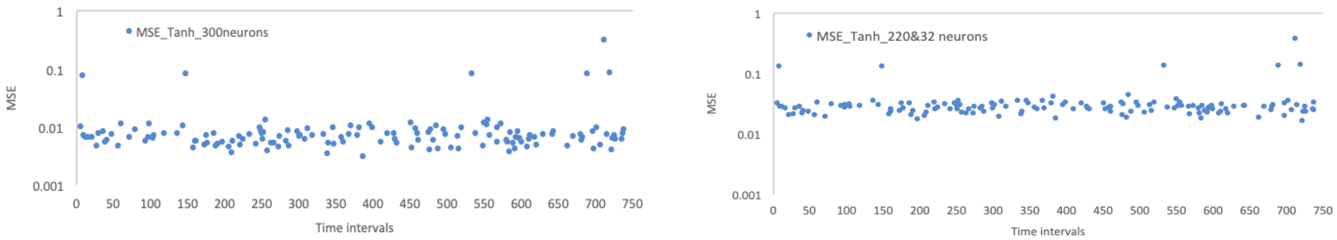
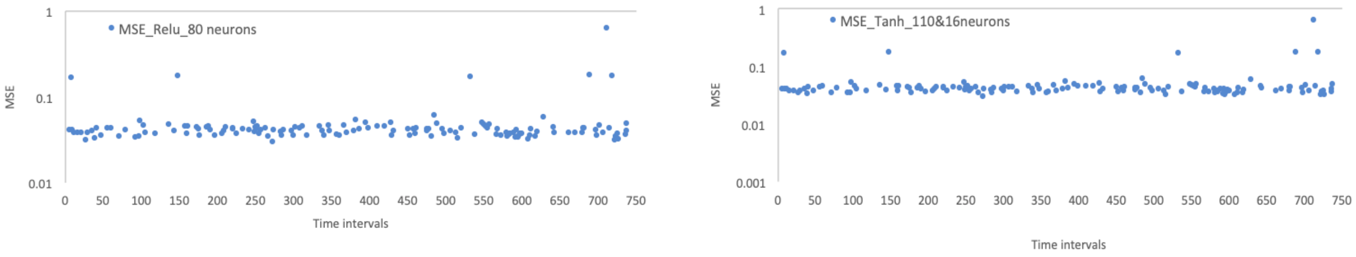


Figure 10 presents the reconstruction errors for the test set related to Model 3 and Model 4, it can be observed that both models have a similar behavior and that six time intervals present larger reconstruction errors as in the case of the previous models.

Figure 10. Errors corresponding to the reconstruction of test set for Model 3 (left) and Model 4 (right)



Figures 9 and 10 show that there are six time intervals where all the models had difficulties in carrying out the reconstruction of the corresponding liquidity vectors. This result indicates, on one hand, that our methodology is robust to changes in the architectures of the autoencoders, and on the other, that these intervals are potential anomalies. We will delve into the latter at the end of this section.

5.2. Bank runs simulations testing

After the fitting and testing of the four models, we performed bank run simulations to see if the autoencoder was able to flag them as anomalies. The bank run was done in the last 90 time intervals (from 651 to 741) for 2018, on one SPI participant. The simulation consisted of stressing institution b_i outflows toward the rest of participants. More specifically, we introduce the bank run as a random value with exponential distribution and a probability of occurrence for the selected b_i outflows, sampled from a Bernoulli distribution. Following Triepels-Daniels-Heijmans (2018), the corresponding parameters p for the *Bernoulli* and λ for the *Exponential* sampled variables are defined as follows:

$$p(x) = p_s + (p_e - p_s)(\tfrac{x}{d})^{\,r}$$

$$\lambda(x) = \lambda_s + (\lambda_e - \lambda_s)(\tfrac{x}{d})^{r}$$

Where the subscripts s and e represents starting and ending values for each parameter, $x \in \{1, 2, \ldots, 90\}$ indicates the time interval, while d is the total number of time intervals, 90, finally r is a rate that controls the increase of $\frac{x}{d}$. This parametrization increases the value of p and λ as time passes, leading to a more intense liquidity adding to the end of the period under the simulation.

Moreover, in Figure 11 and Figure 12, the MSE highlights the simulated bank run in the final time intervals. In these time intervals, the MSE rapidly changed as the payment network unexpectedly began to change as well. Below, we present the MSE of the final liquidity matrices, emphasizing that the high outgoing liquidity flows of the stressed b_i could not be accurately reconstructed, resulting in a high reconstruction error during bank runs. This is observed in all proposed architectures, having that for the TanH models when two layers are used, the autoencoder makes stricter penalizations. It can be explained because the error related to the simulations is larger than the case of one layer. On the other hand, the ReLU autoencoders show a very similar behavior, showing larger errors for both the simulations and the rest of the time intervals.

Figure 11. Bank run simulations results for Model 1 and Model 2

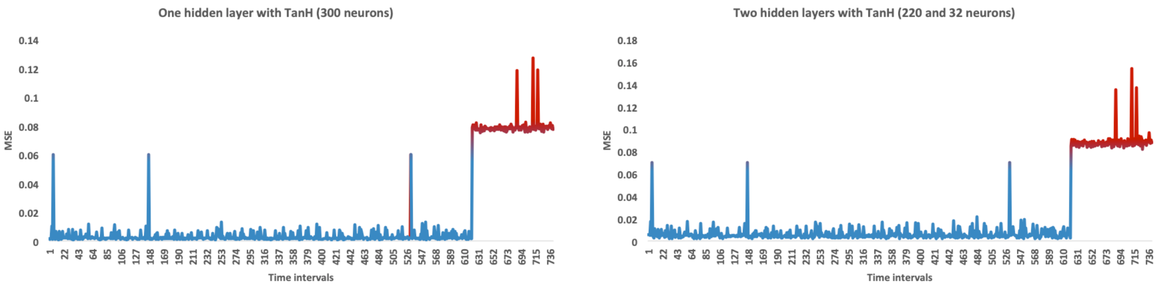
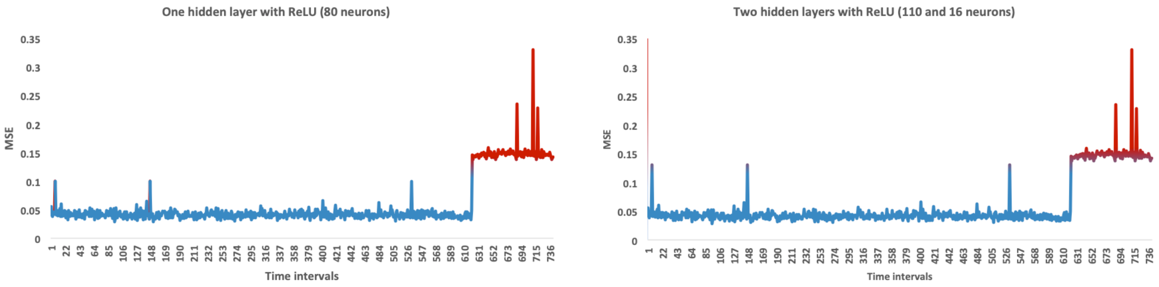


Figure 12. Bank run simulations results for Model 3 and Model 4



With the above it can be confirmed that we have achieved a compression model that is capable of learning the common patterns inherent to the data, and thus is able to recognize anomalous behavior. The next step, once the robustness of the autoencoder has been tested, is to deepen on the analysis of the anomalies detected within the test set in order to have a deeper understanding on autoencoder capacities as an oversight tool.

5.3. Alert analysis

Besides identifying intervals with anomalous patterns of payments with the autoencoder, we investigated which of the flows (one or several) caused these anomalies. In this subsection, we present the main results obtained from the models presented in the previous section, analyzed from the point of view of the system overseer to confirm that the alerts could be considered real anomalies. It is worth noting that there is a concordance in the alert results for all the trained models that indicate unusual payment patterns of systemic importance within the SPI.

The table 2 shows the top ten time-intervals that can be classified as anomalies in the SPI considering our four autoencoder models, the classification criterion is based on the set F defined in subsection 4.1, that is, the anomalous instances will be those that are greater than ε, which in our case is equal to the 90th percentile of the validation MSE of each model. The first 6 alerts (711,718,147,688,532,7) coincide in all models. The time intervals 484, 381, 628, 549, 97 are repeating alerts on some models. While the time intervals 254, 549, 394 are alerts that only Model 1 indicated.

Table 2. Top 10 of time intervals with highest errors for each model

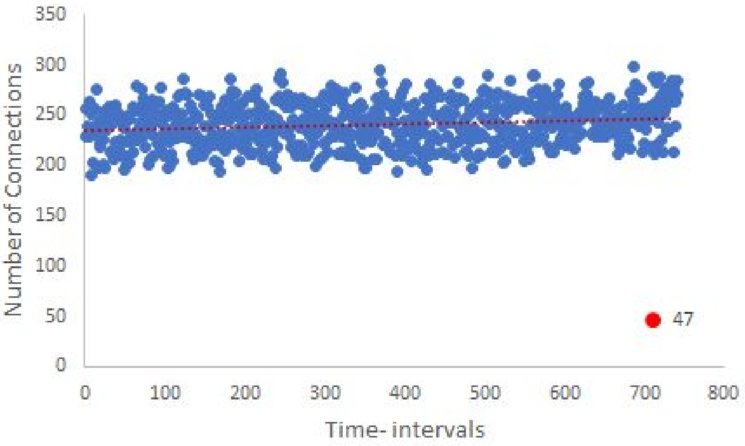
Rank	**Model 1**		**Model 2**		**Model 3**		**Model 4**	
	Time interval	MSE	Time interval	MSE	Time interval	MSE	Time interval	MSE
1	**711**	0.31	**711**	0.39	**711**	0.64	**711**	0.64
2	**718**	0.08	**718**	0.14	**688**	0.18	**688**	0.18
3	**147**	0.08	**532**	0.14	**147**	0.18	**147**	0.18
4	**688**	0.08	**688**	0.14	**718**	0.18	**718**	0.18
5	**532**	0.08	**147**	0.14	**532**	0.17	**532**	0.17
6	**7**	0.08	**7**	0.13	**7**	0.17	**7**	0.17
7	**554**	0.01	**484**	0.05	**484**	0.06	**484**	0.06
8	**254**	0.01	**381**	0.04	**628**	0.06	**628**	0.06
9	**549**	0.01	**549**	0.04	**381**	0.05	**381**	0.05
10	**394**	0.01	**345**	0.04	**97**	0.05	**97**	0.05
Median		0.007		0.028		0.041		0.041
75th percentile		0.008		0.032		0.045		0.045
90th percentile		0.011		0.034		0.049		0.050

It can be underscored that the resulting alerts show the time intervals in which there are fewer than normal payment connections. In particular, we find both intervals for which major banks do not operate and intervals for which medium or small banks do not channel payments to large banks, among other possible patterns. Results show that the autoencoder was able to alert on individual or systemic unusual patterns.

Below, we present the most significant alerts that the autoencoder identified. They are useful examples of how this analysis can support the oversight of the *Sistema de Pagos Interbancarios.*

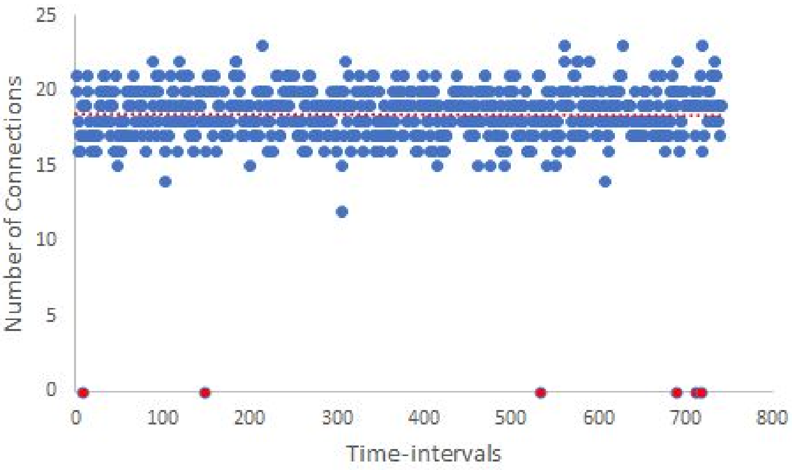
Systemic alert: The most relevant alert was given in time interval 711 (December 2018). This can be explained by an unusual low participation of banks in the SPI. In this case, only 47 payment flows (3 ordering banks) occurred, much lower than the average flows in each time interval, that is 240 payment connections. In Figure 13, we present the flows and amounts for this time-interval.

Figure 13. Alert for unusual low payment connections



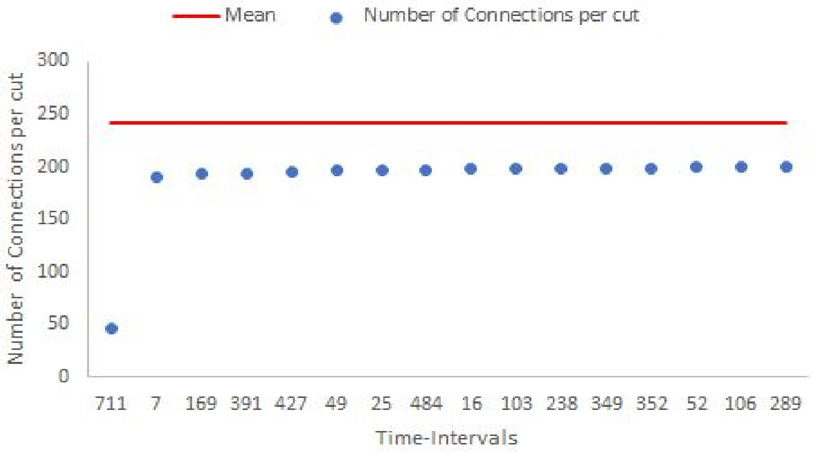
Individual alert for a SPI large participant: This large bank channelized 22% of the total in the SPI and sent and received 11% out of the total operations in 2018. The activity of this large bank is stable along a year time period; it participated in 735 of the 741 time intervals of the year. All the models indicate that there are 6 time-intervals for which the bank did not participate in the SPI at all. As a relevant fact, this bank had 17 payment connections on average over 2018. For example, it can be seen that in intervals 717 and 719, this large bank registered 23 and 24 connections - above the average - but in intervals 7, 147, 532, 688,711, 718, this SPI participant does not have a single connection. Figure 14 shows the activity of this "big bank" in all time intervals of the year.

Figure 14. Alerts for a large bank at times of no payment connections



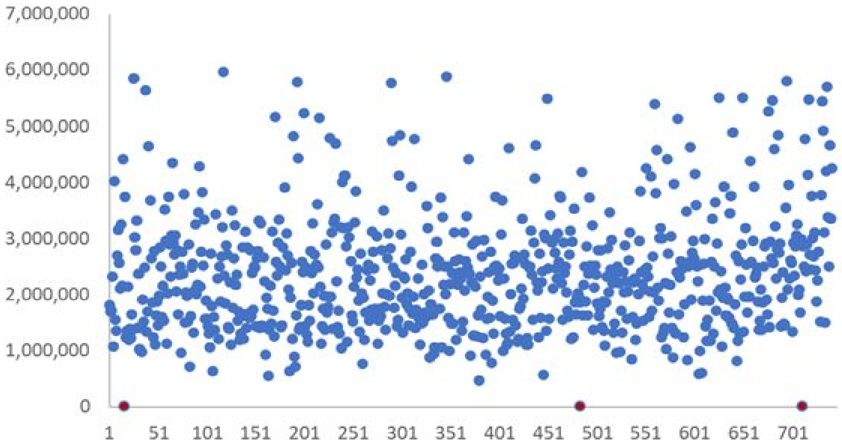
Alerts of low number of participants (and payment connections) in the SPI. In 5% of the time intervals, the number of payment flows is considerably low in comparison with the average 241 connections. The most significant alert is detected in interval 711 with only 47 payment connections. It is worth mentioning that this time interval was associated in the model with a higher alert. Figure 15 depicts the anomalous behavior for such intervals as well as others below the average, all ranked according to their number of connections. These alerts also involve time-interval 484 and 394.

Figure 15. Alerts of low number of participants (and payment connections)



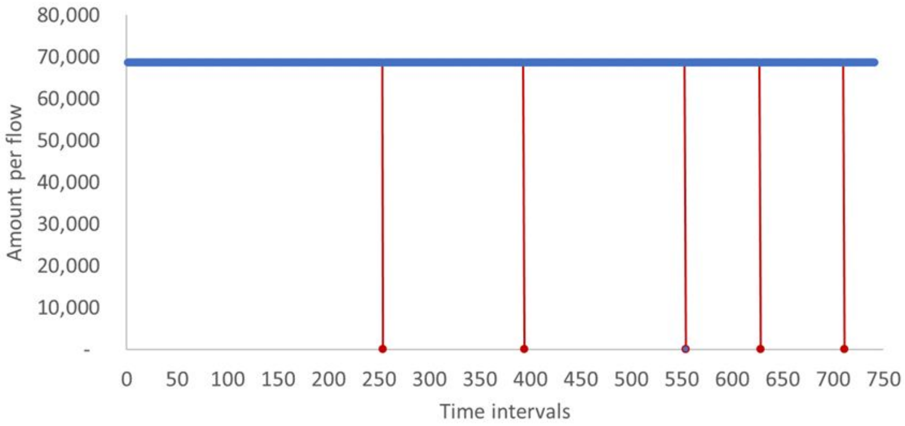
Problem with communication provider and impact on a large bank: The 484 time interval identified by the autoencoder is mainly explained by the non-participation in the SPI of a large bank as well as six other banks (1 medium and 6 little ones) that usually operate. According to the records of operation of the SPI, at this time-interval (see Figure 16) there was a problem of intermittent connection with certain banks with a provider of communication channels, which prevented them from sending operations to the Central Bank of Ecuador in this time interval.

Figure 16. Alert for problem with communication provider



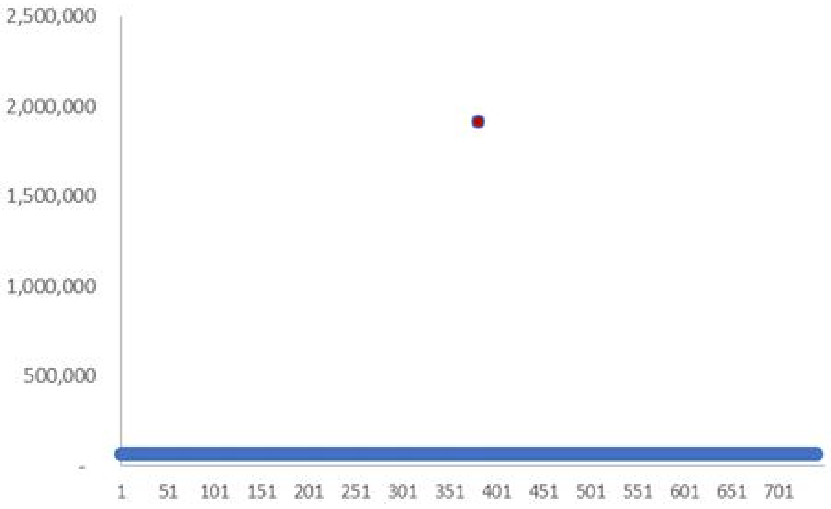
Individual alert for a SPI "average Joe" participant: An alert of a medium bank is found in three intervals (254, 554, 394, 628, 711). The autoencoder detected that in 99.5% of the 2018 intervals, this "average Joe" bank has payment connections with the 5 largest banks participating in the SPI. The alert refers to three intervals for which this bank does not participate in the system, including its regular connections with the 5 largest banks. Figure 17 demonstrates that the "average-Joe" SPI participant was absent in such time-intervals.

Figure 17. Alert for a medium size bank with no payments connections



Unusual payment amounts: In the cases the autoencoder detects events in which the entities send unusual amounts in most cases it indicates the maximum or very low amounts with respect to its normal behavior. These behaviors are evident in the time interval analyzed 381, 549, 345. For example, an entity that on average sends payments for an amount of USD 64,000, the autoencoder alerts when this entity sends USD 1.9 million, the latter being the maximum payment made by the SPI. As shown in Figure 18, it may constitute a significant alert to be analyzed further.

Figure 18. Alert for amounts of payments different from the usual ones



The above alerts provided by the autoencoder models are useful for the oversight and monitoring of the SPI because they allow us to detect payment patterns. The whole SPI dataset cannot be analyzed manually due to the quantity of information that is processed daily, so this methodology provides a powerful tool to equip oversight experts with the ability to identify both normal and anomalous payment patterns.

All in all, interpreting the autoencoder alerts requires the expert judgement of payments oversight teams who have the best understanding of how the system operates, to determine whether the alerts are relevant and if they constitute evidence for pattern changes that could represent a risk.

Finally, the autoencoder can be used as a tool to detect possible operational problems that can cause uncertainty within the system. This is in line with Klee (2010) who used an algorithm to identify outliers in the payment patterns of financial institutions in Fedwire Funds that stem from operational outages. Operational problems cause uncertainty regarding end-of-day Fed account positions that can impact rates in the Federal Funds market. The magnitude of the effects depend on the severity of the difficulty, the time it occurs, and the volume of payments made by the affected participant.

6. Discussion of results

The application of machine learning techniques to support monitoring of financial transactions in SIPS does not replace human decision making but rather it provides new tools to test both simple and complex hypotheses on a large scale over big datasets. This is a critical development for SIPS and FMIs oversight. Autoencoder models can process long time-series of payment networks and based on volume distribution and network structure, they suggest arbitrarily small subsets of transaction periods for further evaluation. In effect, the generality of the autoencoder representation -given by the non-linear decomposition into adaptive hidden units- allows us to extract common patterns in the data and single out uncommon patterns of transactions. The detailed analysis of the properties of this subset of flagged transactions demonstrated the complexity of this automated monitoring procedure, as alerts were generated for a number of different reasons, related to volume, number of connections, or absence of specific institutions.

The autoencoder for anomaly detection is a methodology originally proposed for a real-time payment system (RTGS) by Triepels-Daniels-Heijmans (2018), making the model to identify patterns in real time transactions. In our model, given the SPI does not fully perform as an RTGS but rather as a hybrid, our analysis is more similar to Sabetti-Heijmans (2020) that performs an autoencoder for a DNS system. This confirms the versatility and validity of applying an autoencoder to detect anomalies in payment systems. As we underlined in previous sections, the autoencoder needs to be applied in tandem with the thoughtful review of a payment systems oversight team, to verify the real causes of the alert.

The construction and training of the model is a careful process involving numerous validations and tests that must be carried out, but once the model has been trained its daily application in detecting anomalies in SIPS and FMI operations can take only minutes. This is accompanied by the important and opportune access to data, we were able to use the BCE dataset while fine-tuning the model, as the Central Bank is responsible for operating and overseeing the SPI. This is important as noted by León (2020) who indicates that the application of new methods in payment systems should consider low computational costs and ease in data collection.

This document contributes to identify evidence of incidents in the payments flow of a respective system, and thereby provides new tools for payments oversight, and it ultimately sets the basis for early warning tools. We were able to detect anomalies in the payments flows processed by banking entities in the major SIPS Ecuador, the *Sistema de Pagos Interbancarios*. We proposed four models to test autoencoder robustness, where we selected the best architectures by performing cross-validation. These models were trained on real banks behavior stemming from the payments network of the SPI for 2018, which makes our work novel and relevant. All the autoencoders we tested identified relevant anomalous patterns, finding problems in the reconstruction of the data and flagging specific payment networks as anomalous. In order to evaluate the models, a bank run simulation was performed, altering a major SPI participant's outgoing payment flows exponentially over a period of time.

In our approach, we were able to identify alerts that could affect the system, such as: 1) non-participation of systemically important participants, 2) a low number of connections (payment flows), 3) medium-size banks not sending payments to systemically important participants, among others. Additionally, the bank run simulation shows the ability of the autoencoder to detect risk events that may be generated in the SPI in the absence of common patterns. Importantly, we deepened the analysis of the alerts that the autoencoder signaled as potential anomalies by relying on the expert judgement of overseers of the SPI.

Further studies in machine learning for anomaly detection in payment systems can improve the accuracy, reliability, and speed of the methodology leading to financial alerts that are more spot-on, consistent, and that can be performed in real-time. Notwithstanding, these techniques should not be intended to replace the knowledge in payment systems oversight teams but rather become part of their toolkit.

References

Aggarwal, Charu C., and Philip S. Yu. "Outlier detection for high dimensional data." Proceedings of the 2001 ACM SIGMOD international conference on Management of data. 2001.

Aleskerov, Emin, Bernd Freisleben, and Bharat Rao. "Cardwatch: A neural network based database mining system for credit card fraud detection." Proceedings of the IEEE/IAFE 1997 computational intelligence for financial engineering (CIFEr). IEEE, 1997.

A. Arning, R. Agrawal, P. Raghavan. A Linear Method for Deviation Detection in Large Databases. KDD Conference Proceedings, 1995.

V. Barnett, T. Lewis. Outliers in Statistical Data. John Wiley and Sons, NY 1994.

Ben-Gal, Irad. "Outlier detection." Data mining and knowledge discovery handbook. Springer, Boston, MA, 2005. 131-146.

Brause, Rüdiger, T. Langsdorf, and Michael Hepp. "Neural data mining for credit card fraud detection." Proceedings 11th international conference on tools with artificial intelligence. IEEE, 1999.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3 (2009): 1-58.

Committee on Payment and Market Infrastructures and Technical Committee of the International Organization of Securities Commissions (CPMI-IOSCO). 2012a. “Principles for Financial Market Infrastructures.” (April).

Dorronsoro, Jose R., et al. "Neural fraud detection in credit card operations." IEEE transactions on neural networks 8.4 (1997): 827-834.

Ghosh, Sushmito, and Douglas L. Reilly. "Credit card fraud detection with a neural-network." System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on. Vol. 3. IEEE, 1994.

Goodfellow I., Bengio Y., and Courville A. "Deep Learning" MIT Press, 2016.

Hadi, A.S. A modification of a method for the detection of outliers in multivariate samples. Journal of the Royal Statistical Society, B, 56(2), 1994.

Hawkins, Simon, et al. "Outlier detection using replicator neural networks." International Conference on Data Warehousing and Knowledge Discovery. Springer, Berlin, Heidelberg, 2002.

Hodge, Victoria, and Jim Austin. "A survey of outlier detection methodologies." Artificial intelligence review 22.2 (2004): 85-126.

Klee, E. (2010). Operational outages and aggregate uncertainty in the federal funds market. *Journal of Banking & Finance*, *34*(10), 2386-2402.

León-Rincón, C. E., (2020). Detecting anomalous payments networks: A dimensionality reduction approach. *Latin American Journal of Central Banking* 1 (1-4).

E. Knorr, R. Ng. Algorithms for Mining Distance-based Outliers in Large Data Sets. VLDB Conference Proceedings, September 1998.

Sabetti, Leonard, and Ronald Heijmans. "Shallow or deep? Detecting anomalous flows in the Canadian Automated Clearing and Settlement System using an autoencoder." (2020).

Triepels, Ron, Hennie Daniels, and Ronald Heijmans. "Anomaly Detection in Real-Time Gross Settlement Systems." ICEIS (1). 2018.

Williams, G. J., Baxter, R. A., He H. X., Hawkins S. and Gu L. (2002) "A Comparative Study of RNN for Outlier Detection in Data Mining" IEEE International Conference on Data-mining (ICDM'02), Maebashi City, Japan, CSIRO Technical Report CMIS-02/102.

Classifying payment patterns with artificial neural networks: An autoencoder approach

Jeniffer Rubio, Paolo Barucca, Gerardo Gage, John Arroyo, Raúl Morales-Resendiz

Introduction

Payments and market infrastructures are the backbone of modern financial systems and play a key role in the economy by enabling multilateral transactions under certain rules and platforms. One of their main goals is to manage systemic risk, especially in the case of systemically important payment systems serving interbank funds transfers. Thus central banks need to be able to monitor their activity and to identify anomalous events.

Thanks to the availability of high data volumes and to the increment in computation capabilities it is posible to develop tools that automatically performs tasks as the identification of anomalous patterns.

In this vein, we developed a methodology based on a unsupervised neural network, the autoencoder, to detect a diverse set of anomalies arising within the *Sistemas de Pagos Interbancarios* (SPI) from Ecuador. It was found that the methodology is robust enough to support the monitoring of payment systems, but need to be acompained by the expert judgement of payments overseers.

Methodology

The data we study contains information on the liquidity vectors, we aimed to reconstruct the vectors by compressing and decompressing the dataset under use. For this purpose, we implemented a lossy compression, which generates a particular type of representation that allows the data to not be exactly learnt and some of the information to be lost.

When this type of compression is implemented, the relevant patterns present in the data are learned. Once the compression model learns the common patterns and a new liquidity vector is fed for its reconstruction; if the reconstruction is bad, this is explained by the fact that vector information differs from the normal patterns that the model learned, indicating the possibility of a potential anomaly.

The quality of the reconstructions will be measured through the reconstruction error which is the difference between the values yielded by the lossy compression and the real vector values.

Methodology

Anomaly detection framework

- $B = \{b_1, \dots, b_n\}$ set of SPI participants
- $T = \{t_1, \dots, t_m\}$ ordered set of m time intervals
- $a_{i,j}$ total amount of liquidity transferred from institution b_i to institution b_j
- $A^{(k)}$ liquidity matrix for the k-th interval

$$A^{(k)} = \begin{pmatrix} a_{1,1}^{(k)} & \cdots & a_{1,n}^{(k)} \\ \vdots & \ddots & \vdots \\ a_{n,1}^{(k)} & \cdots & a_{n,n}^{(k)} \end{pmatrix}$$

- $\bar{a}^{(k)}$ liquidity vector ($A^{(k)}$ rearrangement)
- $RE: D \to [0, \infty)$, non-negative function that measures the reconstruction error of liquidity vector, where *D* is the set of all liquidity vectors

Our goal is to find the set $F = \{\bar{a}^{(k)} \in D \,|\, RE(\bar{a}^{(k)}) \geq \epsilon\}$, where $\epsilon > 0$

CEMLA
CENTER FOR LATIN AMERICAN MONETARY STUDIES

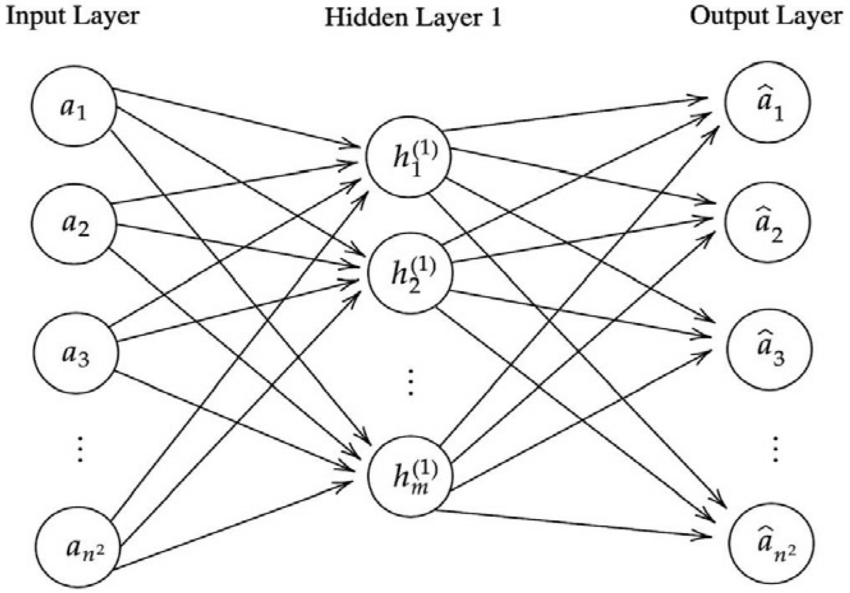
Methodology

For our work, we selected the autoencoder as the lossy compression model, which is an unsupervised neural network.

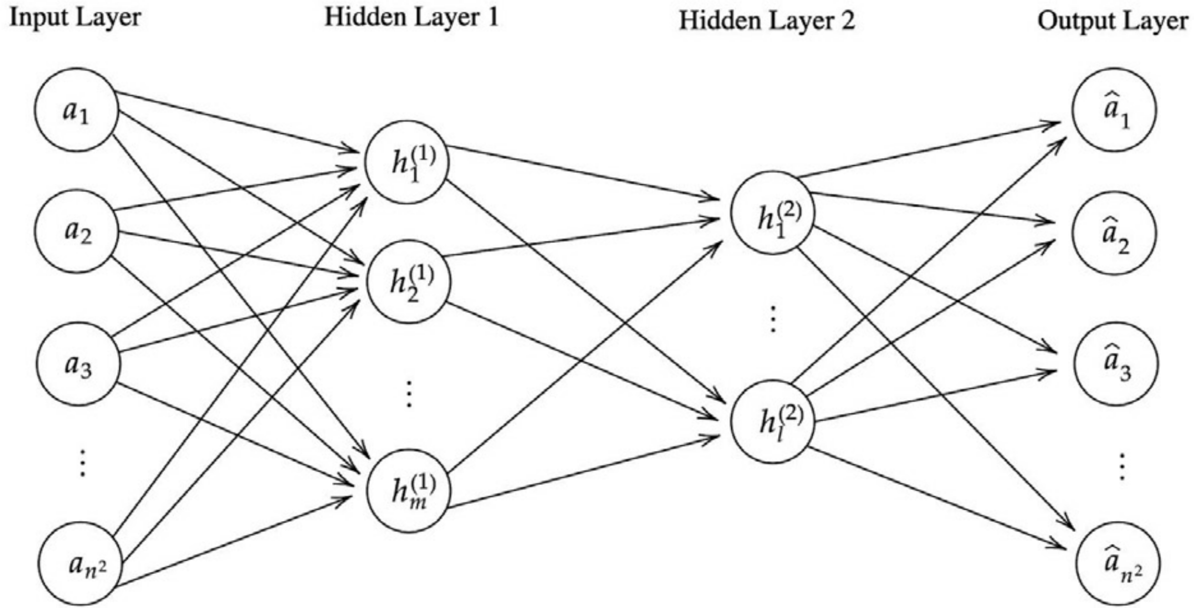
The autoencoder is made up of two components, the encoder and the decoder. The encoder is the initial part of the autoencoder and it has the task to create an accurate lower-dimensional representation of the data. The second part of the autoencoder, the decoder, is in charge to carry out the reconstruction of the data. The encoder goes from the input layer to the layer with the lowest number of neurons, which is commonly called bottleneck given that is the layer of the network where data is the most compressed. The encoder can be represented as a function $h = f(X)$, where X represents the input data. On the other hand, the decoder goes from the *bottleneck* to the output layer, it can be represented by the function $r = g(h)$. The autoencoder final objective is to find f and g such that $\mathrm{X} \approx g(f(X))$.

Methodology

i) One Hidden Layer Autoencoder Architecture



ii) Two Hidden Layers Autoencoder Architecture



Model fitting, selection and testing

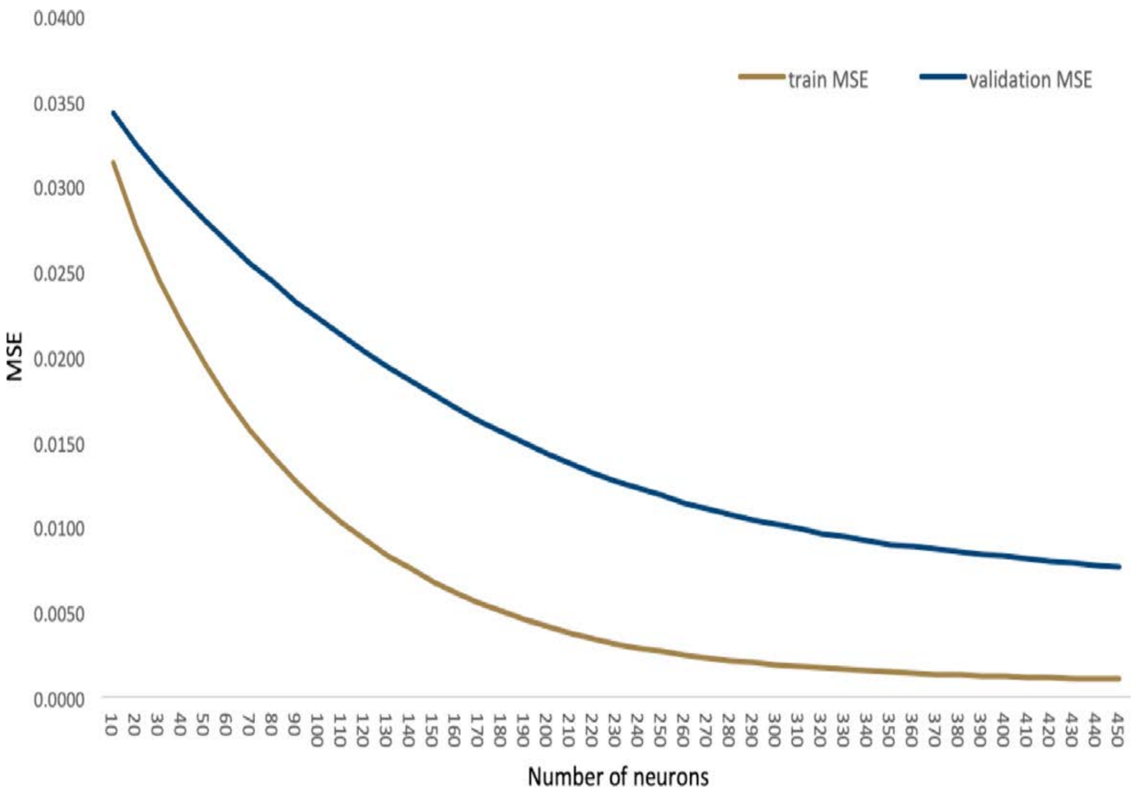
- Before fitting the model, data was pre-processed, this corresponded to a log-transformation and standardization.

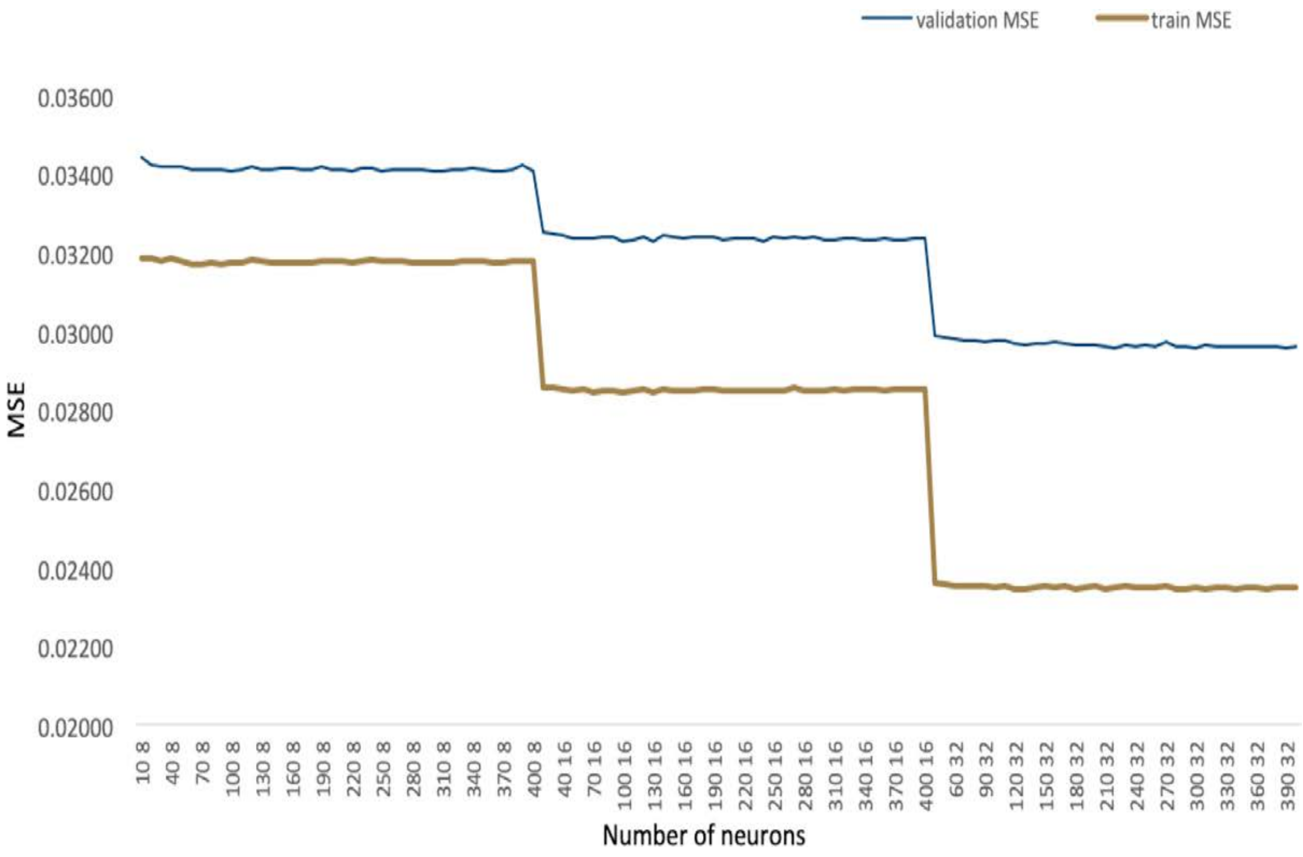
- Then the data was partitioned into cross-validation and test sets in a proportion of 80% and 20%, respectively.

- We analyzed different setups for the autoencoder by varying the number of hidden layers, the number of neurons in each layer and the activation functions (TanH and ReLU), this led us to the definition of four different models
  - Model 1: Model 1: One hidden layer with TanH as the activation function.
  - Model 2: Two hidden layers with TanH as the activation function.
  - Model 3 One hidden layer with ReLU as the activation function.
  - Model 4: Two hidden layers with ReLU as the activation function.

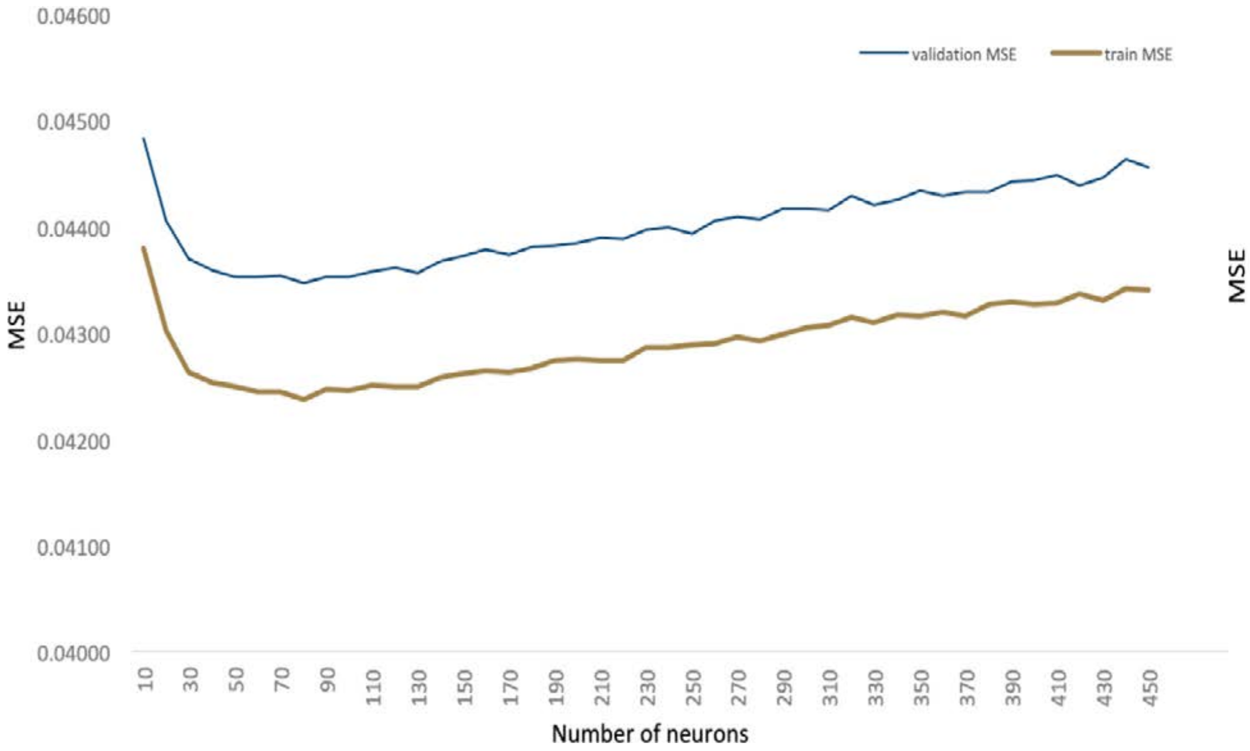
The number of neurons in each hidden layer (hyperparameters) were selected through a 5-fold crossvalidation.

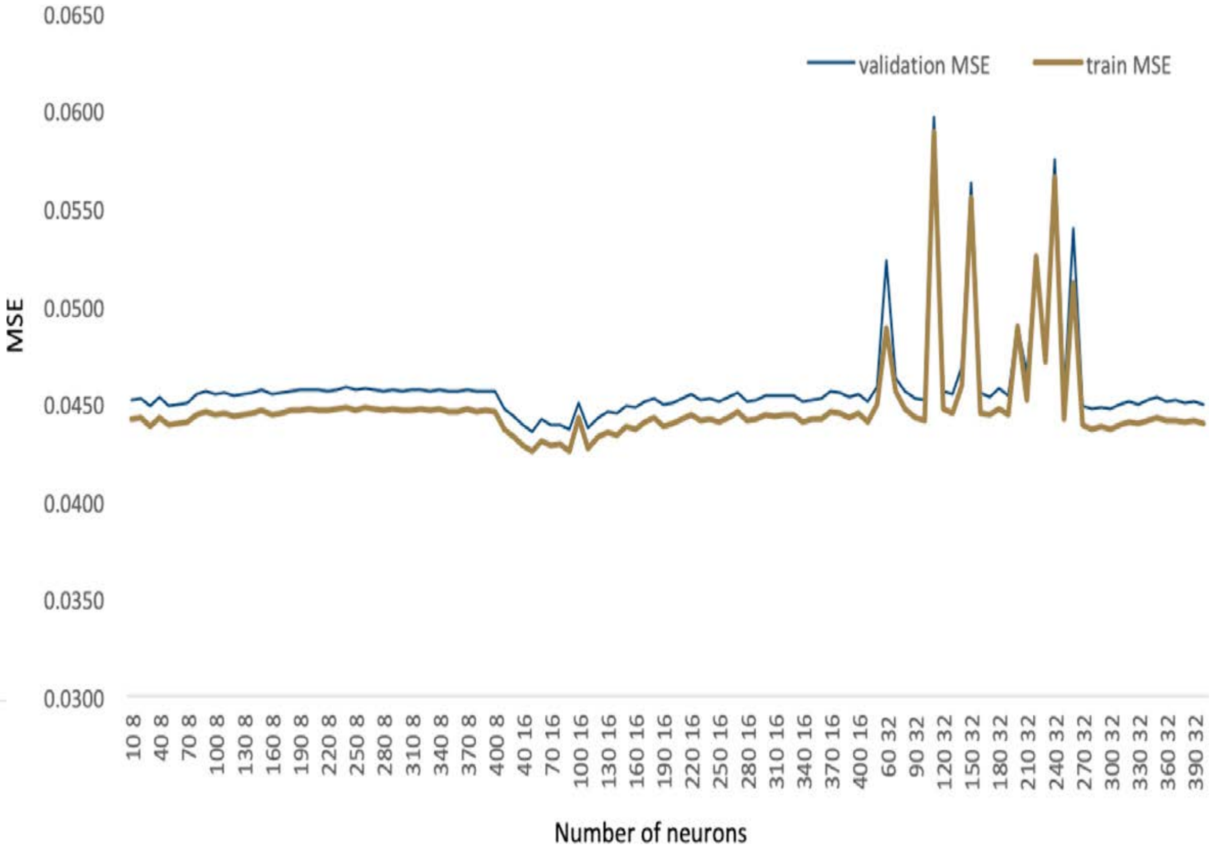
TanH models results





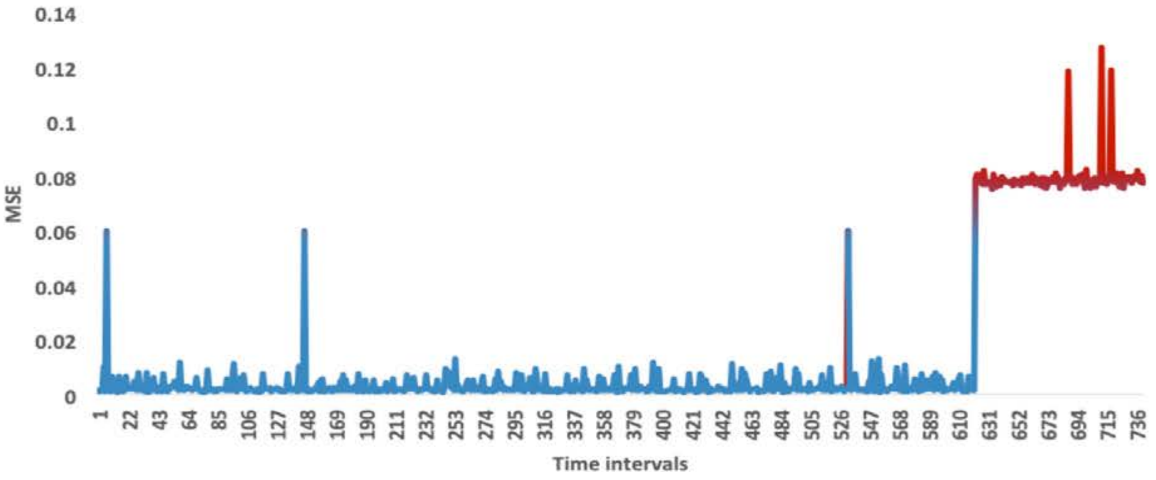
ReLU models results



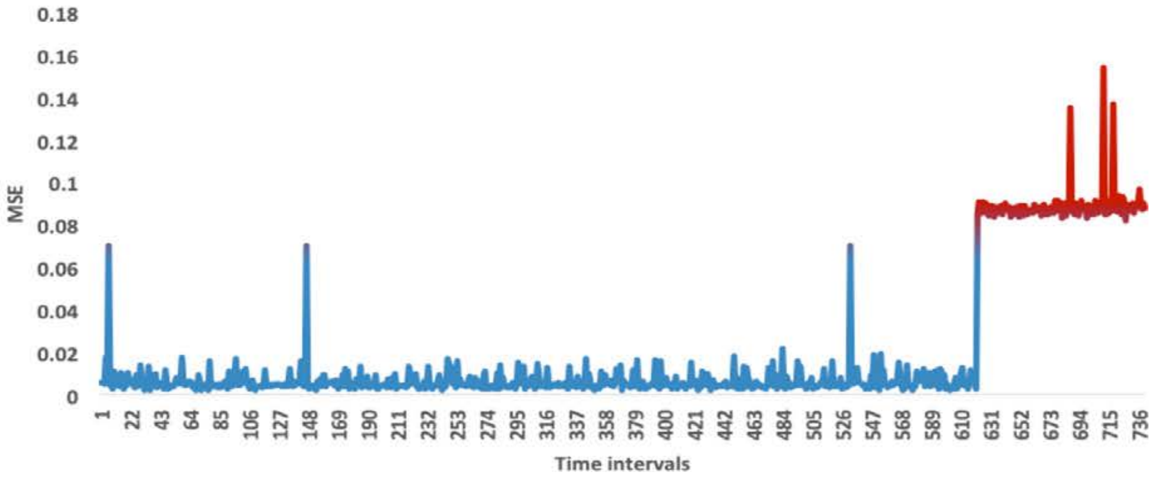


Bank run simulations

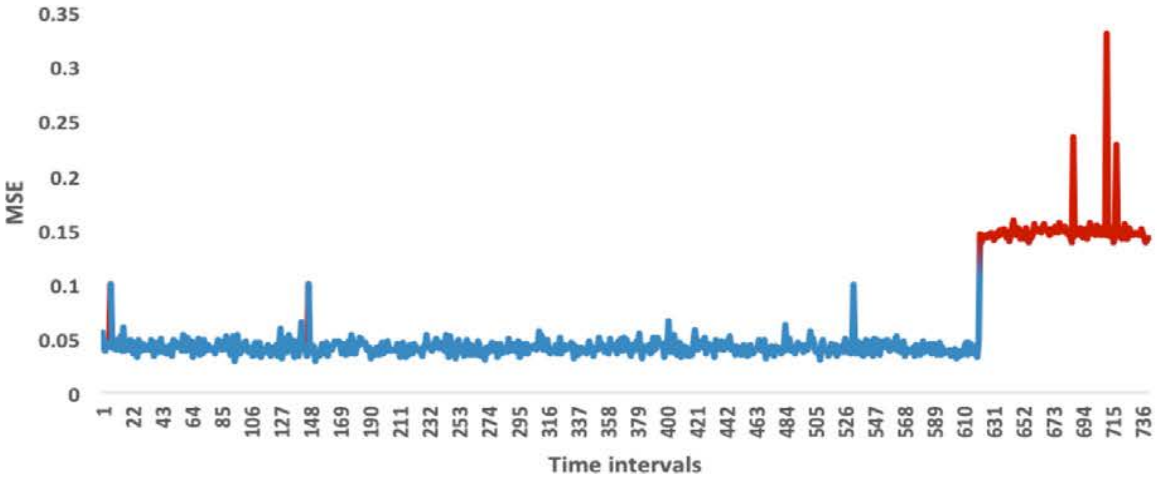
One hidden layer with TanH (300 neurons)



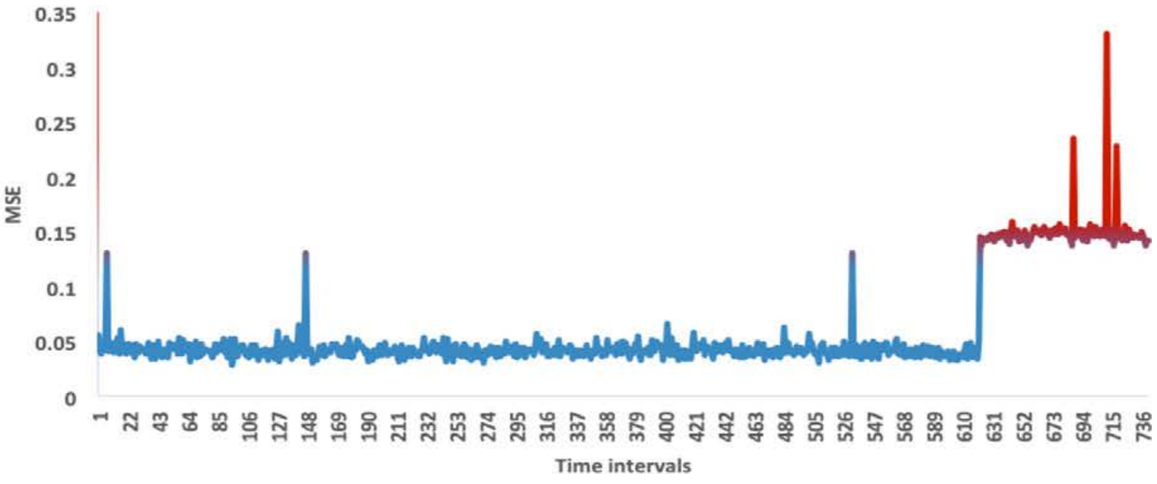
Two hidden layers with TanH (220 and 32 neurons)



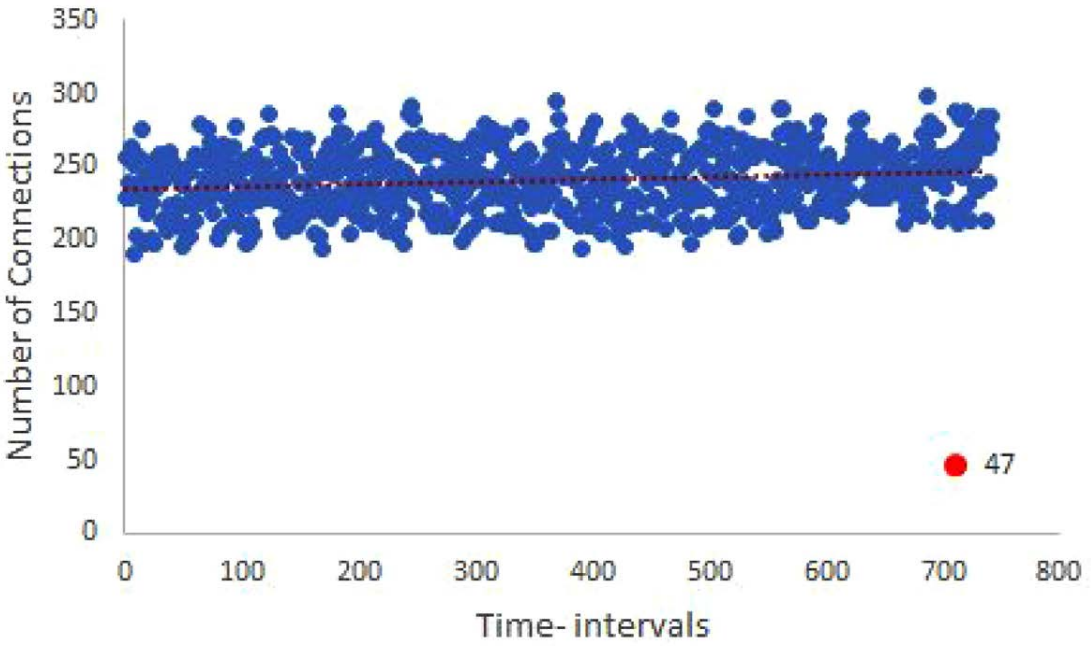
One hidden layer with ReLU (80 neurons)

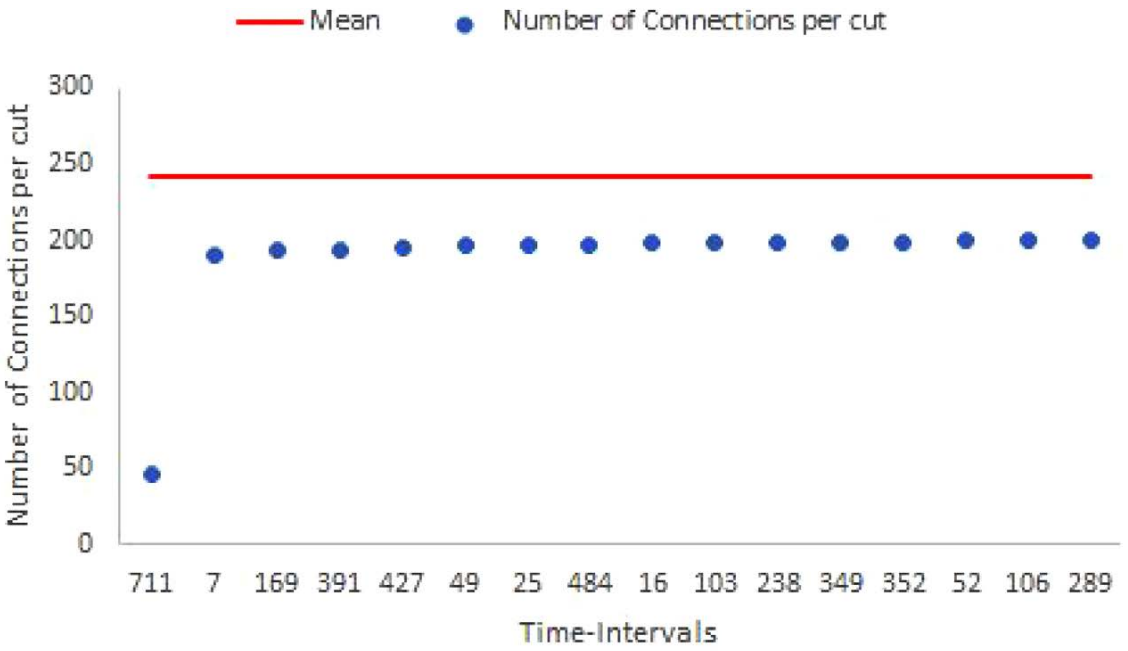


Two hidden layers with ReLU (110 and 16 neurons)

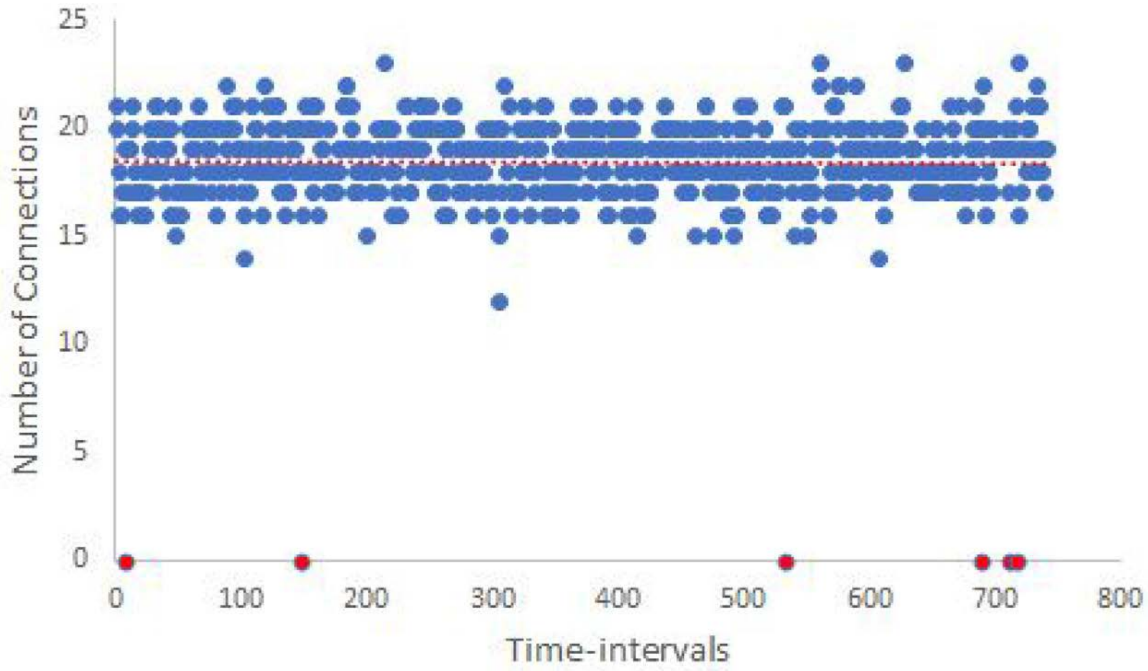


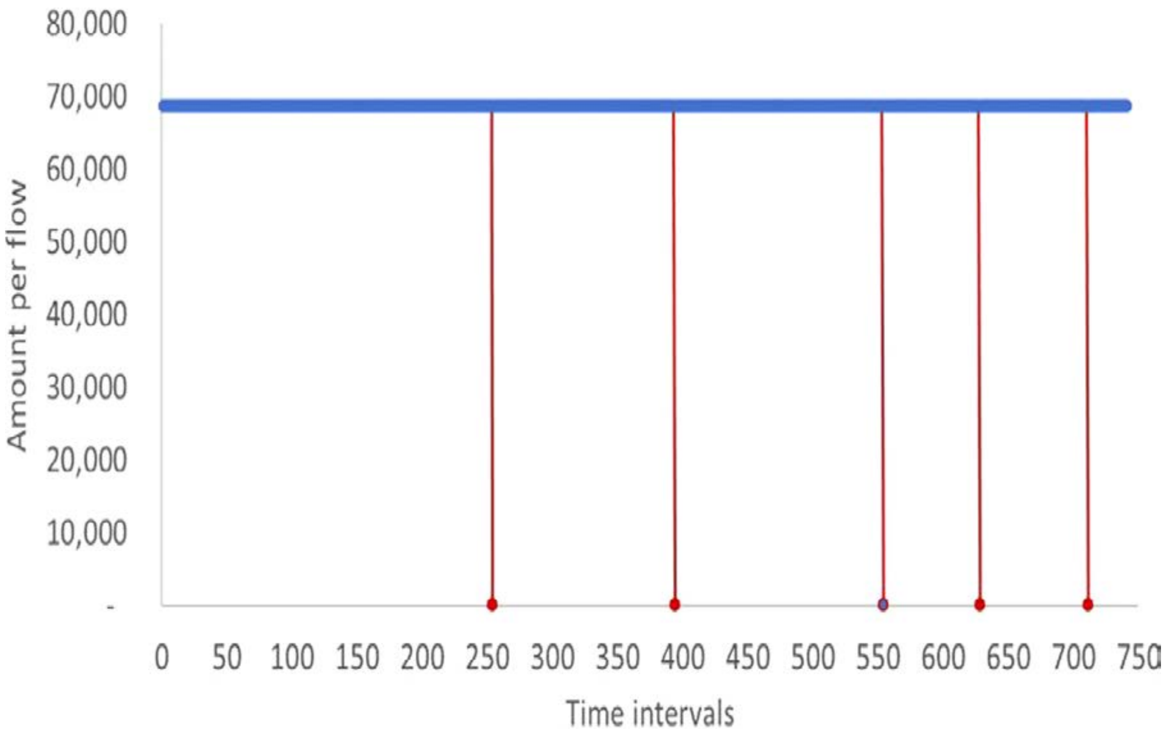
Alerts analysis



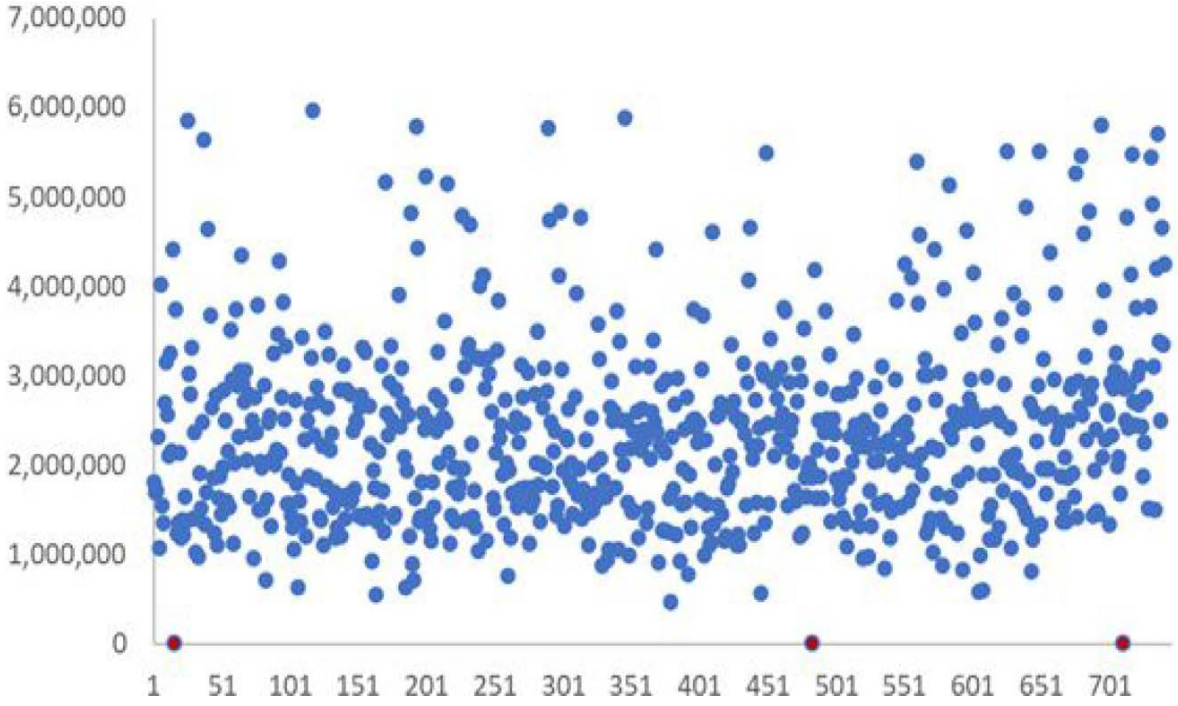


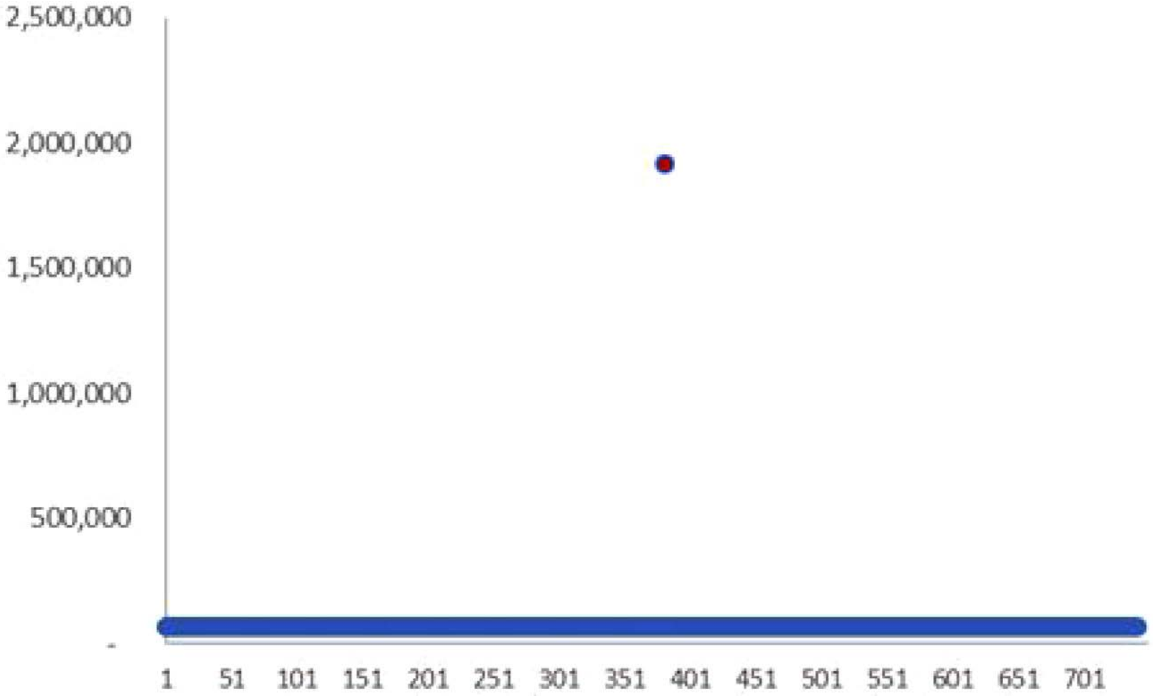
Alerts Analysis





Alerts Analysis





CEMLA
CENTER FOR LATIN AMERICAN MONETARY STUDIES

Conclusions

- The application of machine learning techniques to support monitoring of financial transactions in SIPS does not replace human decision making but rather it provides new tools to test both simple and complex hypotheses on a large scale over big datasets .
- The generality of the autoencoder representation -given by the non-linear decomposition into adaptive hidden units- allows us to extract common patterns in the data and single out uncommon patterns of transactions.
- Further studies in machine learning for anomaly detection in payment systems can improve the accuracy, reliability, and speed of the methodology leading to financial alerts that are more spot-on, consistent, and that can be performed in real-time
- These techniques should not be intended to replace the knowledge in payment systems oversight teams but rather become part of their toolkit

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Using deep learning technique to automate banknote defect classification[1]

Jiradett Kerdsri and Pucktada Treeratpituk,
Bank of Thailand

[1] This presentation 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 central banks and other institutions represented at the event.

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Using Deep Learning Technique to Automate Banknote Defect Classification

Jiradett Kerdsri, Pucktada Treeratpituk

Department of Data Management and Data Analytics
Bank of Thailand

20 October 2021

Motivation

To increase the efficiency of our banknote printing operation

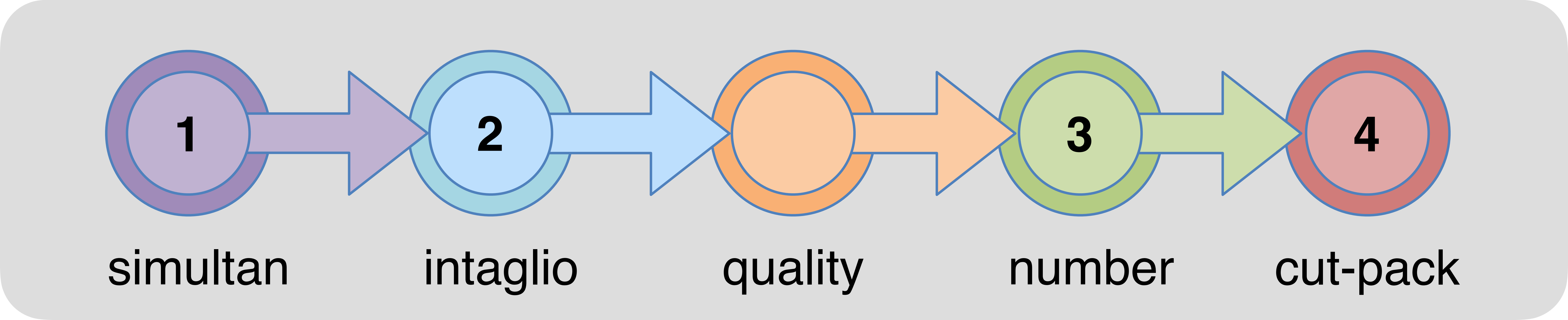
- By reducing defect rate with real-time banknote printing quality inspection
- By reducing human effort involved in the inspection process

To expand ML applications into central bank operations (BAU) especially for more automation & lean processes

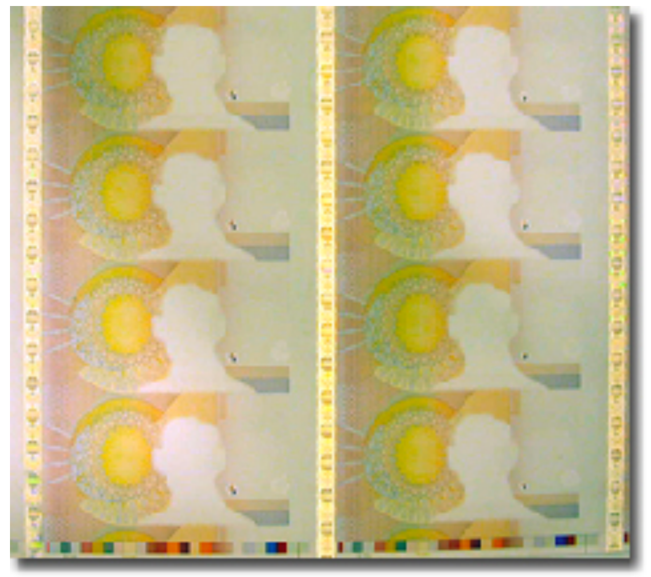
- *Statistical compilations, e.g. government spending, entity-resolution/disambiguation*
- *Human Resources*
- *Communications*
- ***Banknote printing works***

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

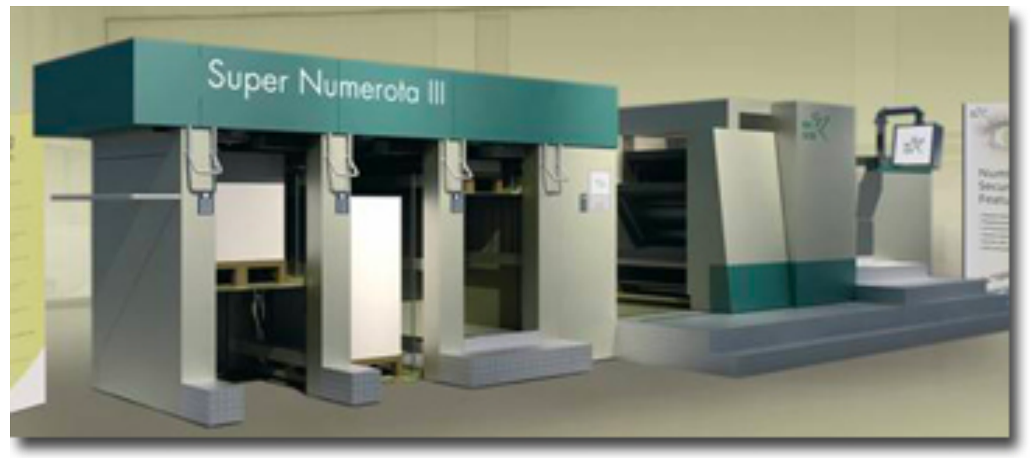
Banknote Printing Process



simultan → intaglio → quality → number → cut-pack

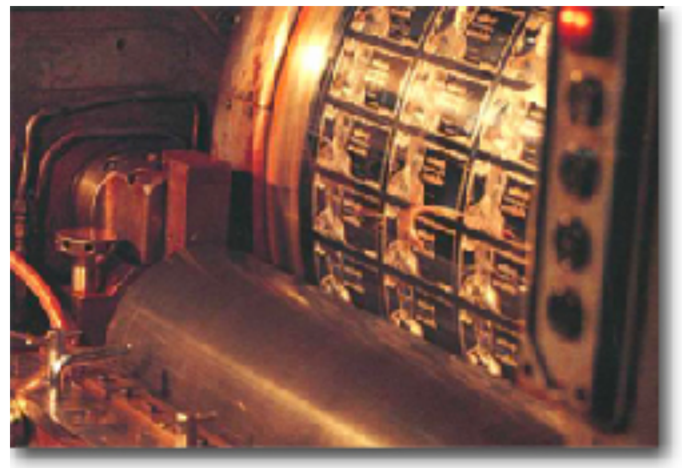


offset printing machine



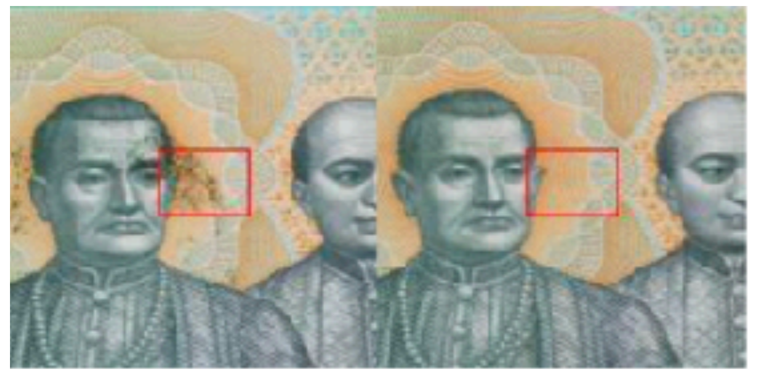
number printing machine

intaglio machine



ธนาคารแห่งประเทศไทย
BANK OF THAILAND

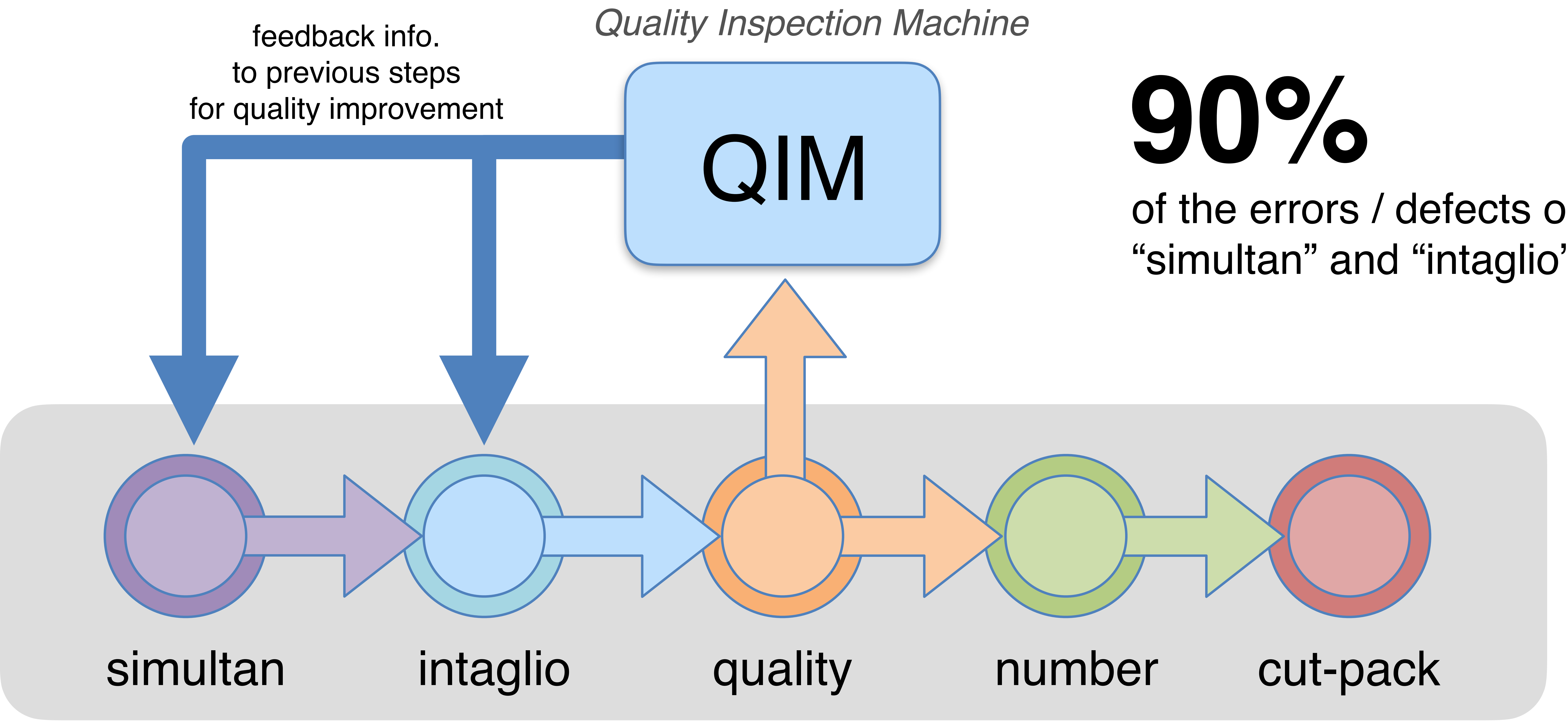
Quality Inspection



number of defects from the note printing process contributes to the cost of cash

1%
defect target

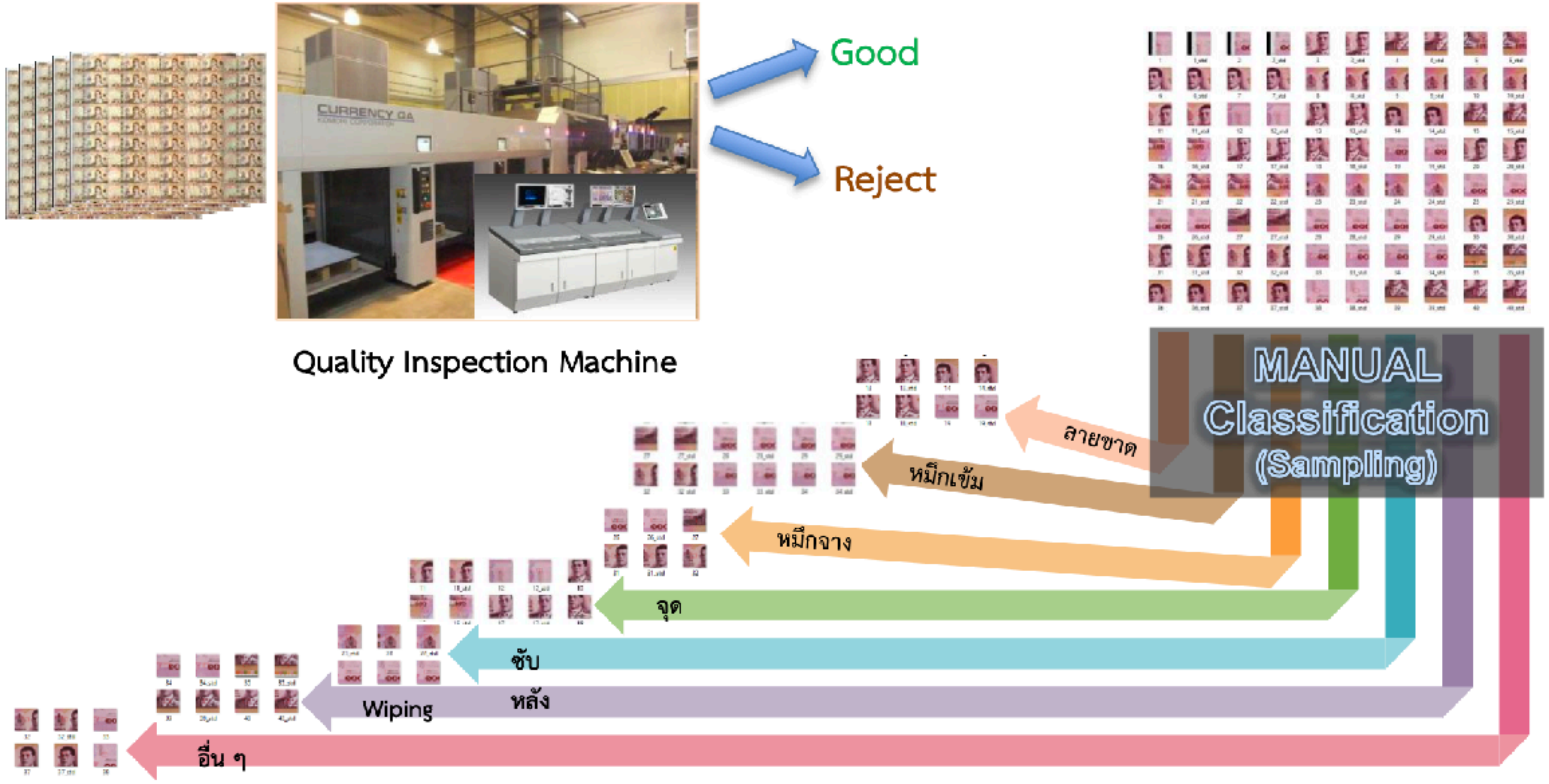
0.95%
~ actual defect percentage



90%

of the errors / defects occurs in the "simultan" and "intaglio" steps

Previous Process

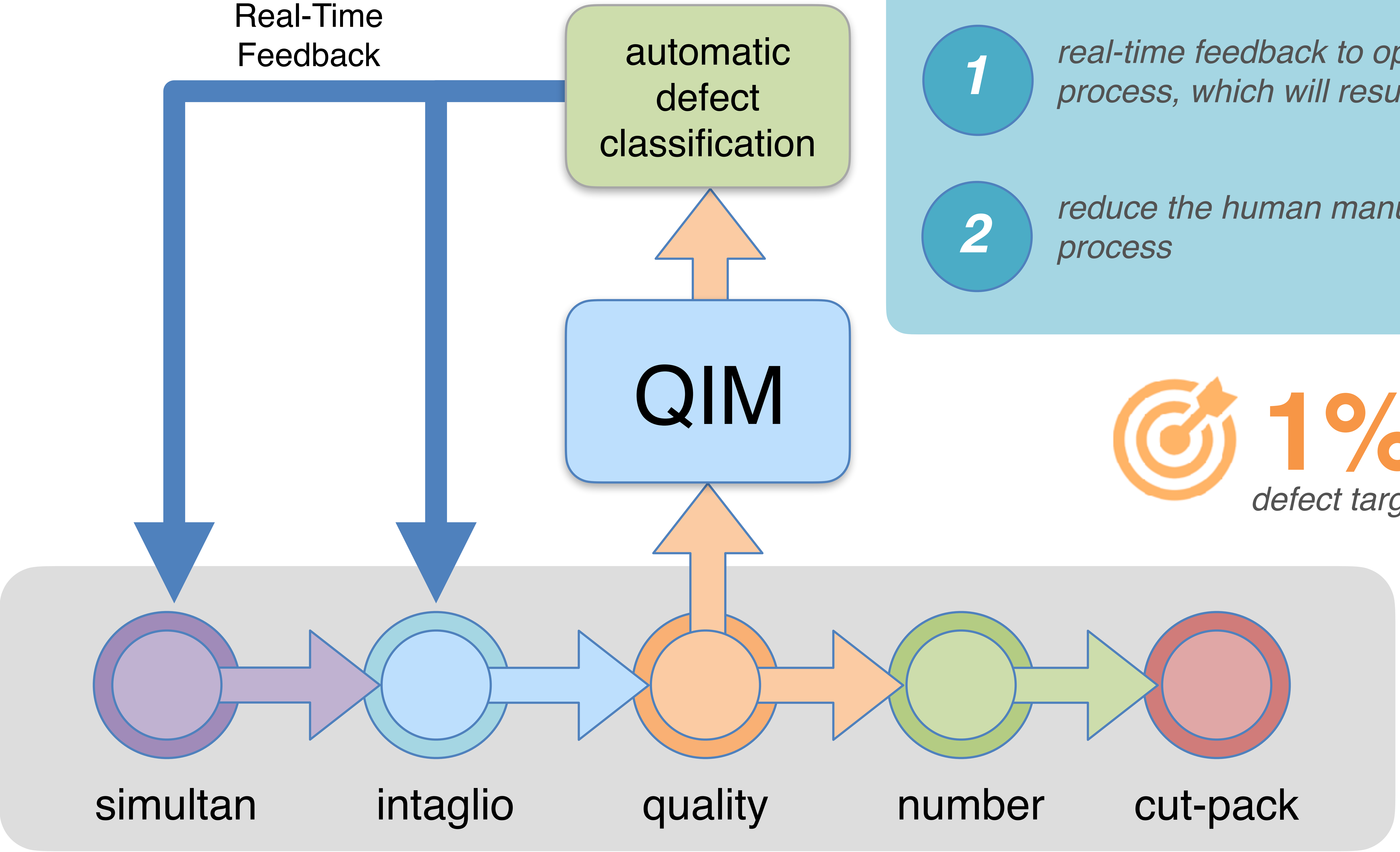


1 *require human inspector to manually classify and collect statistics of defect types to inform operators in previous steps*

2 *Inspections are carried out at the end-of-day, thus there are gaps between when the problems was found and when the adjustment will be made*

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Automatic Defect Classification



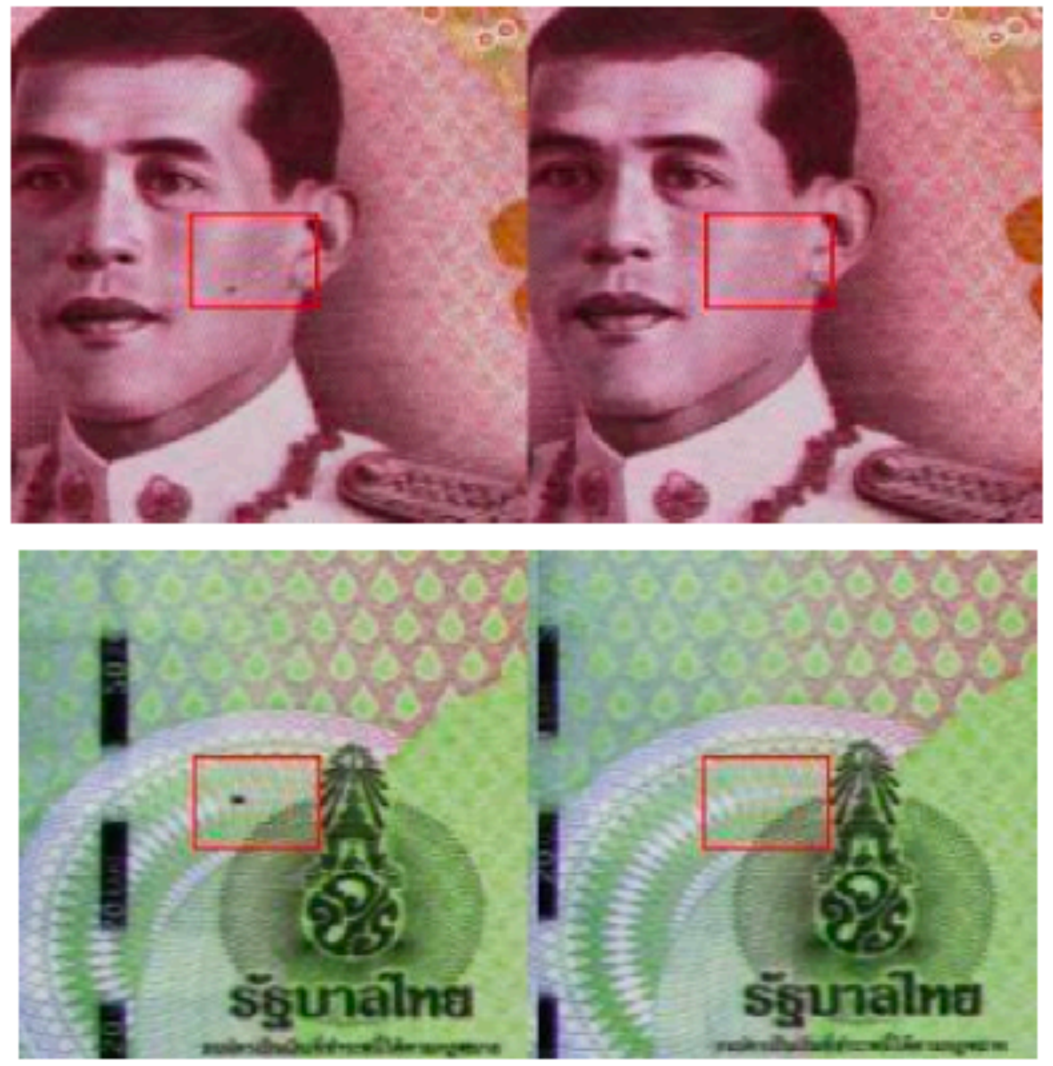
1. *real-time feedback to operators of the printing process, which will result in faster fixes*
2. *reduce the human manual work-load in the process*

1% *defect target*

20% *reduction of defect*

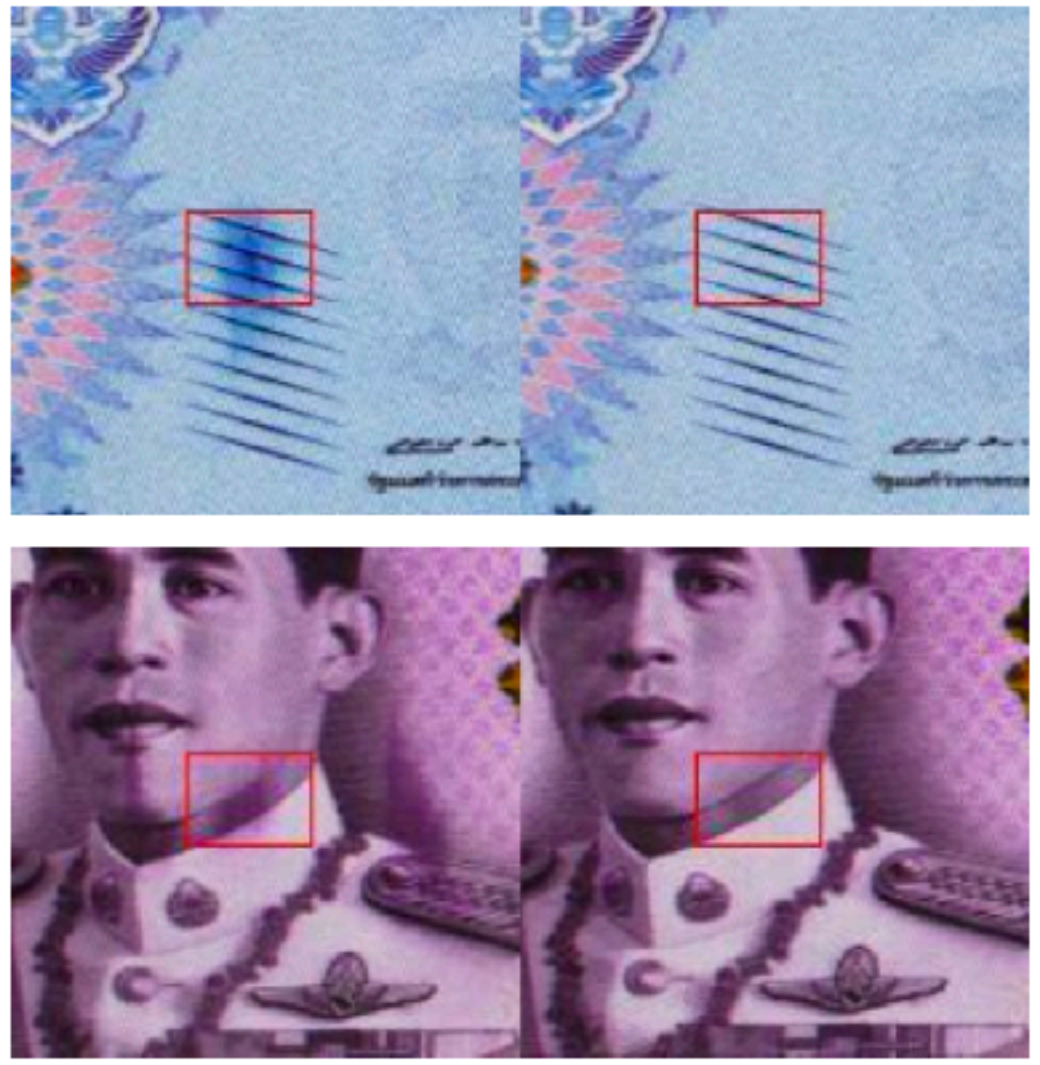
Examples: Banknote Defects

Dot/Ink Spot



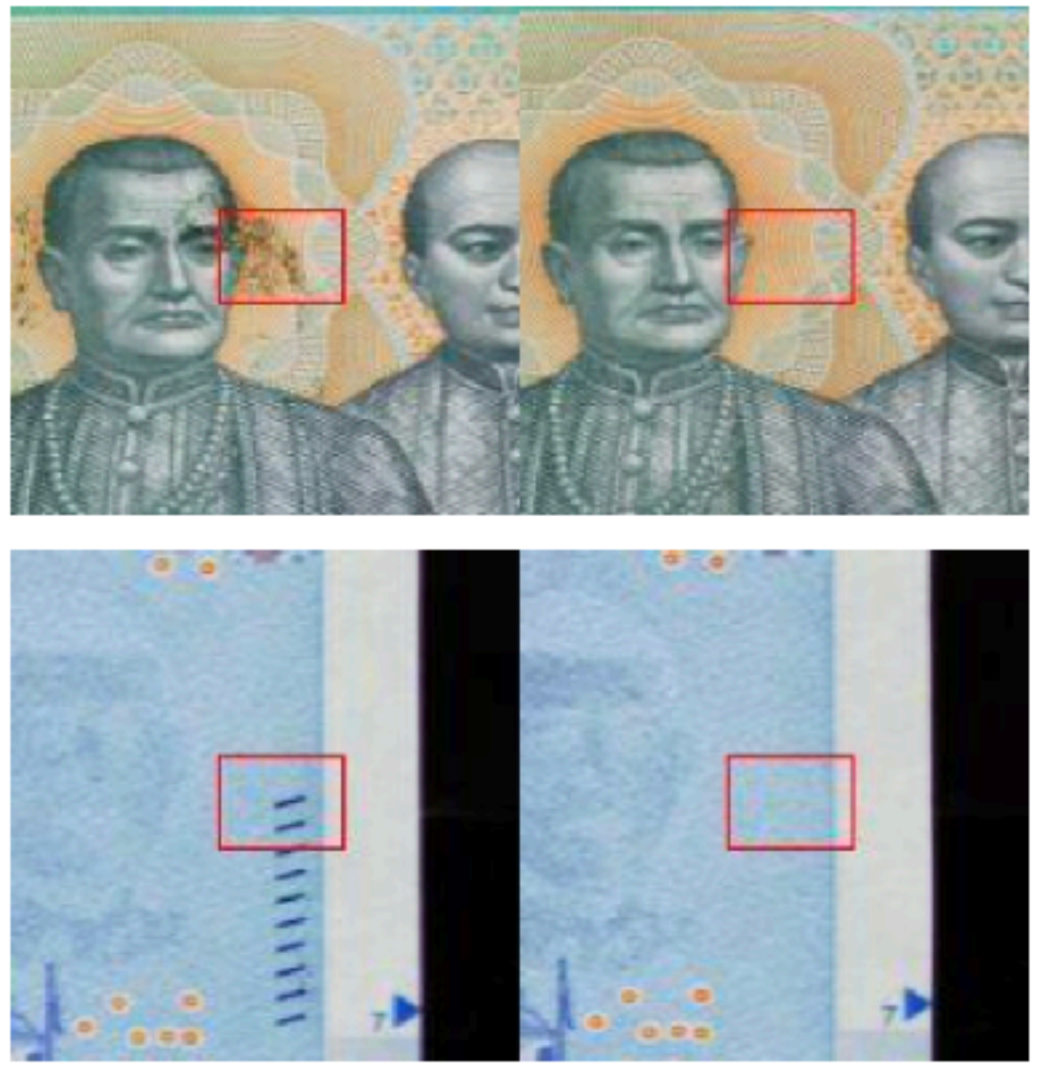
Defect *Normal*

Wiping



Defect *Normal*

Set-Off

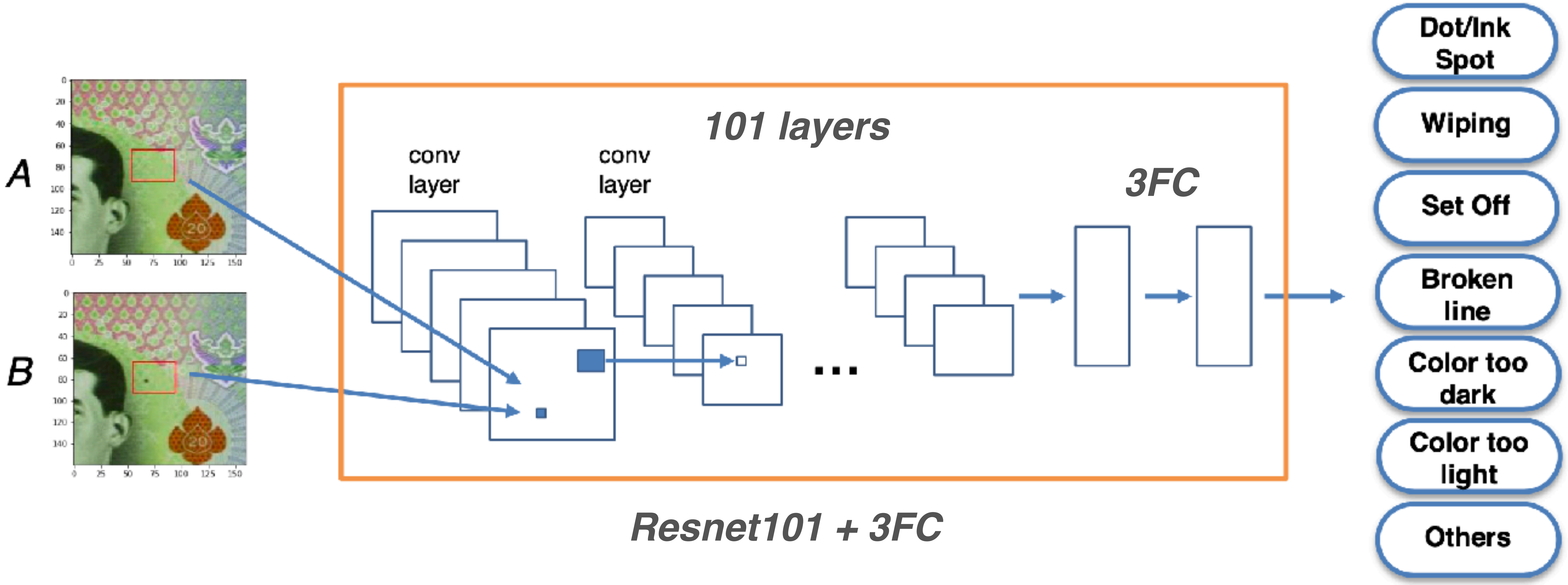


Defect *Normal*

Methodology
Automatic Defect Classification

***Model**: ResNet-101* + 3FC*
[one shared model for 5 banknote denominations]

***Input**: image pair (defect banknote + standard banknote)*
***Output**: 7 defect classes*


A

B

101 layers

conv layer

conv layer

...

3FC

Dot/Ink Spot

Wiping

Set Off

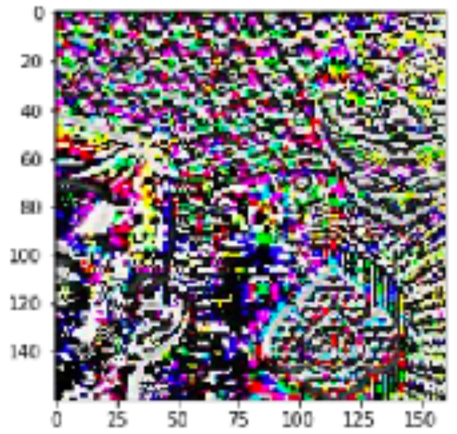
Broken line

Color too dark

Color too light

Others

Resnet101 + 3FC


A - B

photos are not properly aligned, so simply subtract two images won't work

** K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, CVPR, 2016*

Accuracy

Training (1,659 banknote defects)

97% **Test Accuracy**
(Out-Of-Sample)

Operational

74% **Accuracy**
Operational

	Dot (Front)			Dot (Back)			Wiping (Front)			Set Off (Back)		
	Model	*Manual*	*Acc*	*Model*	*Manual*	*Acc*	*Model*	*Manual*	*Acc*	*Model*	*Manual*	*Acc*
20 Baht	467	461	99%	366	330	90%	1066	1023	96%	327	306	94%
50 Baht	438	392	89%	184	102	55%	1416	692	49%	252	186	74%
100 Baht	947	946	100%	231	229	99%	575	543	94%	612	564	92%
500 Baht	247	229	93%	184	145	79%	1319	177	13%	627	525	84%
1000 Baht	123	123	100%	59	54	92%	433	163	38%	654	592	91%
Sum	2222	2151	97%	1024	860	84%	4809	2598	54%	2472	2173	88%

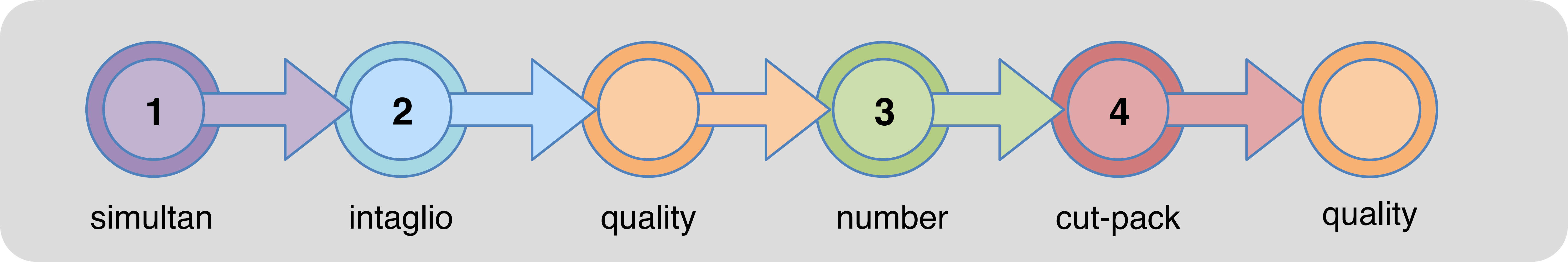
ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Future Work

? Improving performance on all 7 types of defects (90% coverage), currently work best only on the 3 biggest defect types

- Increase accuracy to 90%+, via more training data
- Expanding to front & bank variations (certain error types occur mostly only on one side)

? Expanding to the automatic quality inspection to the cut-pack step


1
simultan
2
intaglio
quality
3
number
4
cut-pack
quality