Irving Fisher Committee
on Central Bank Statistics

IFC Bulletin

No 44

Big Data

Proceedings of the IFC Satellite Seminar on “Big Data” at the ISI Regional Statistics Conference 2017 in Bali, Indonesia, on 21 March 2017

September 2017

BANK FOR INTERNATIONAL SETTLEMENTS

Contributions in this volume were prepared for the IFC Satellite Seminar on "Big Data" at the ISI Regional Statistics Conference 2017, co-organised by the Irving Fisher Committee on Central Bank Statistics (IFC) and Bank Indonesia in Bali on 21 March 2017. The views expressed are those of the authors and do not necessarily reflect the views of the IFC, its members, the BIS and the institutions represented at the meeting.

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

ISSN 1991-7511 (online)

ISBN 978-92-9259-092-5 (online)

Big data

IFC Bulletin no 44
September 2017

Proceedings of the IFC Satellite Seminar at the ISI Regional Statistics Conference 2017 on big data
Bali, Indonesia, 21 March 2017

Seminar overview

Opening remarks

Keynote speech

Session 1 – Big data for central banks

Session 2 – Internet data sets

Session 3 – Financial, administrative and commercial data sets

Session 4 – Central bank communication

Panel – Issues on big data governance – big data work in central bank HR and IT issues

Big data and central banking

Overview of the IFC satellite meeting

Bruno Tissot[1]

1. Introduction – Big data issues for central banks

Background

"Big data" is a key topic in data creation, storage, retrieval, methodology and analysis. The private sector is already using data patterns from micro data sets to produce new and timely indicators. For central banks, the flexibility and real-time availability of big data open up the possibility of extracting more timely economic signals, applying new statistical methodologies, enhancing economic forecasts and financial stability assessments, and obtaining rapid feedback on policy impacts.[2] Yet the public and private sector may have different areas of concern, not least on account of data quality.[3]

As confirmed by a recent IFC survey,[4] the central banking community is indeed taking a strong interest in big data, particularly at senior policy level. A key message of the survey was that big data could prove useful in conducting central bank policy, and that it was perceived as a potentially effective tool in supporting micro- and macroeconomic as well as financial stability analyses.[5]

Yet central banks' actual use of big data is still limited, due mainly to resource and IT constraints, but also to more fundamental challenges. In particular, exploring big data is a complex, multifaceted task, and any regular production of big data-based information would take time, given the lack of transparency in methodologies and the poor quality of some data sources. From this perspective, big data may also

[1] Head of Statistics and Research Support, BIS, and Head of the IFC Secretariat (Bruno.Tissot@bis.org). The views expressed here are those of the author and do not necessarily reflect those of the Bank for International Settlements (BIS) or the Irving Fisher Committee on Central Bank Statistics (IFC). This overview benefited from comments by K Hennings, R Kirchner and J Mehrhoff.

[2] As evidence for central banks' increasing interest in using big data, see D Bholat, "Big data and central banks", Bank of England, *Quarterly Bulletin*, March 2015, https://ssrn.com/abstract=2577759.

[3] For instance, while online retailers targeting potential customers based on past web searches might find it acceptable to be "wrong" once out of five times, official statisticians would usually consider such an 80% accuracy level as inadequate.

[4] See Irving Fisher Committee on Central Bank Statistics, "Central banks' use of and interest in 'big data'", October 2015.

[5] See in particular the various techniques presented at the ECB Workshop on *Using big data for forecasting and statistics*, organised in cooperation with the International Institute of Forecasters in April 2014, www.ecb.europa.eu/pub/conferences/html/20140407_workshop_on_using_big_data.en.html.

create new information/research needs, and international cooperation could add value in this context.

One caveat is to define what big data really is.[6] In general terms, one can think of extremely large data sets that are often a by-product of commercial or social activities and provide a huge amount of granular information at the level of individual transactions. This form of data is available in, or close to, real time and can be used to identify behavioural patterns or economic trends. Yet, there is no formally agreed definition that would cover all possible cases.[7] For instance, it may not be sufficient for a data set to be large to qualify as "big data" – indeed, national statistical authorities have been dealing with large data sets covering millions of records (for instance census data) for many decades without branding them as "big data". In particular, a key factor to consider is whether the data set is structured and can be handled with "traditional" statistical techniques, or if it is unstructured and requires new tools to process the information.

In practice, it is usually referred to "big data" when (i) the data set is unstructured (and often quite large), as a by-product of a non-statistical activity – in contrast to traditional data sets, produced for statistical purposes, which are, by design, clearly structured; or (ii) large volumes of records that are relatively well structured but nevertheless difficult to handle because of their size, granularity or complexity – and which could benefit from the application of big data tools (eg IT architecture, software packages, modelling techniques) to process the information more efficiently. In any case, assessing what is big data leaves room for judgment, and depends on a number of criteria such as the following "Vs":[8] **volume** (ie number of records and attributes); **velocity** (speed of data production eg tick data); and **variety** (eg structure and format of the data set). Some observers have added other "Vs" to this list, such as **veracity** (accuracy and uncertainty of big data sets that usually comprise large individual records), valence (interconnectedness of the data) and **value** (the data collected are often a by-product of an activity and can trigger a monetary reward; hence they are usually not available as a public good, due either to commercial considerations or confidentiality constraints).[9]

[6] See P Nymand-Andersen, "Big data – the hunt for timely insights and decision certainty: Central banking reflections on the use of big data for policy purposes", *IFC Working Paper*, no 14, 2015.

[7] Following the work conducted under the aegis of the United Nations (see Meeting of the Expert Group on International Statistical Classifications, "Classification of types of big data", United Nations Department of Economic and Social Affairs, ESA/STAT/AC.289/26, May 2015), big data can be classified into three types: (1) social networks (human-sourced information, such as blogs, videos, internet searches); (2) traditional business systems (process-mediated data, such as data produced in the context of commercial transactions, e-commerce, credit cards); and (3) internet of things (machine-generated data, such as data produced by pollution or traffic sensors, mobile phone locational information, and logs registered by computer systems).

[8] For these Gartner "3Vs", see D Laney, *3D data management: controlling data volume, velocity, and variety*, META Group (now Gartner), 2001, https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[9] Not all observers agree on the precise list and definitions of the Vs one should consider; for a presentation, see P Nymand-Andersen, op cit, as well as C Hammer, D Kostroch, G Quiros and STA Internal Group, "Big data: potential, challenges, and statistical implications", *IMF Staff Discussion Note*, SDN/17/06, September 2017.

Moreover, big data raises a number of challenges for central banks, especially as regards its handling and use for policymaking. The handling of big data requires significant resources (not least on the IT side) and proper arrangements for managing the information.[10] Turning to its policymaking use, big data creates opportunities but is not without risks, such as that of generating a false sense of certainty and precision. From this perspective, the apparent benefits of big data (in terms, say, of lower production costs or speed in producing information) should be balanced against the potential large economic and social costs of misguided policy decisions that might be based on inadequate statistics.

The IFC satellite meeting

In view of the various challenges, and given the strong interest expressed by the central banking community, IFC members joined forces to monitor developments and issues related to big data – such as the methodologies for its analysis, its value compared with "traditional" statistics, and the specific structure of the data sets. This was done by focusing on a few pilot projects on which central banks have been invited to cooperate so as to share specific experiences. The pilots were intended to cover four main areas: (i) internet data; (ii) administrative data; (iii) commercial data; and (iv) financial market data. A key milestone was the presentation of this initial work at this IFC Satellite meeting on big data co-organised with Bank of Indonesia in March 2017 on the occasion of the Regional Statistics Conference of the International Statistical Institute (ISI).

Opening the meeting, Yati Kurniati, Bank Indonesia, underlined the traditional importance played by data in central banks' day-to-day work. This has been reinforced by the ongoing "big data" revolution: public authorities have access to an increasing supply of information, in terms of volume, velocity and variety; in turn, this information can be used to produce new types of indicator to support policy. Yet this raises new, sometimes acute challenges, with many of them relating to the statistical production process itself – eg collecting, cleaning and storing the new data sets, as well as extracting meaningful information with adequate technologies etc.

Aurel Schubert, European Central Bank (ECB) and Vice Chair of the IFC, also acknowledged in his opening remarks the importance of these challenges. But he felt that there was a clear opportunity for policymakers to access new and complementary information sources.[11] This puts a premium on collaborative work in the central banking community to explore the synergies and benefits of using big data.

The keynote speech by Agus Sudjianto, Executive Vice President and Head of Corporate Model Risk at Wells Fargo, was an opportunity to learn from commercial banks' experiences in dealing with large data sets. Their risk management work and the related use of data has clearly expanded since the Great Financial Crisis (GFC) of 2007–09, not least because of the need to comply with more stringent regulation. In particular, the production of stress tests requires an increasing amount of information and sophisticated quantitative tools.[12] Commercial banks now amass large data

[10] For an overview of the challenges posed by using big data for official statistics more generally, see C Hammer et al, op cit.

[11] See also J Mehrhoff, "Demystifying big data in official statistics – it is not rocket science!", presentation at the Second Statistics Conference of the Central Bank of Chile, October 2017.

[12] See Basel Committee on Banking Supervision, "Making supervisory stress tests more macroprudential: Considering liquidity and solvency interactions and systemic risk", Working Paper, no 29, November 2015.

volumes at a highly disaggregated level, for instance to measure changes in their portfolios over time, capture the drivers of their risk profiles with sufficient sensitivity, and validate their modelling tools. To do that, they have to deal with big data sets and use "big data algorithms", eg new machine learning techniques. And a key aspect was the very considerable amount of work required in terms of data cleaning and data reconciliation.

Another consequence was that financial institutions need highly skilled staff, especially graduates in mathematics, finance and statistics. As highlighted in the address by Vijay Nair, former ISI President and University of Michigan, this testifies to the emergence of data science as a key mode of scientific discovery, on a par with experimental, theoretical, and computational analysis.

The meeting was fruitful in offering various perspectives on these issues. The first session discussed the importance of big data for central banks in general. The second and third sessions focused on specific big data sets, ie internet data sets and financial, administrative and commercial data sets, respectively. The last session reviewed the implications of big data sources in central bank communication. The event ended with a panel discussion on big data governance, the related challenges and the implications in terms of resources, especially in HR and IT.

2. Big data for central banks

The first session, chaired by Robert Kirchner of the Deutsche Bundesbank, discussed the importance of big data for central banks. The initial presentation, by Eurostat, highlighted the need for central bank statisticians to carefully consider the characteristics of big data sets. In particular, the quality of an apparently large non-random sample of data is determined not by its absolute number of records but by its relative size compared to the population of interest.[13] Moreover, the statistical representativeness of a sample depends on its possible coverage bias – that is, the extent to which the structure of the sample is representative of the structure of the entire population studied given the methodology used. From this perspective, a key problem with large, non-random big data sets is their organic nature: the data are often self-reported or is the by-product of social activities (eg financial transactions, internet clicks). As a result, the coverage bias of these samples is unknown and can be significant. For instance, social media sources will yield information whose quality depends on differences in the usage intensity of these social medias; that is, the less one uses them the less one is represented. Such big data sets (eg internet-based) may thus be much lower in statistical quality than (comparatively smaller) probabilistic samples that are designed to be representative of the population of interest. In other words, using (very) large amounts of data is no guarantee of accuracy, and there is a key misperception of the intrinsic value of big data from this perspective.

Perhaps more fundamentally, it is important to distinguish between "data" and "information"; the latter depends on processing the former.[14] Traditional official

[13] See X Meng, "A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it)", in X Lin, C Genest, D Banks, G Molenberghs, D Scott and J-L Wang (eds), *Past, present, and future of statistical science*, Chapman and Hall, 2014, pp 537–62.

[14] See R Groves, "Designed data and organic data", in the Director's Blog of the US Census Bureau, 2011, www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html.

statistics can be described as "designed data", since they are collected for a specified statistical purpose through adequate statistical processes such as surveys and censuses; the collection of these designed data sets is almost by definition organised in order to extract meaningful information, a key difference to "organic" big data. With the increasing supply of organic data relative to that of designed data, the risk of confusion between "data" and "information" is clearly on the rise. The challenge is thus to complement, instead of replace, the collection of designed data with organic big data so as to maximise information-to-data ratios.

The second presentation, by the consulting firm BearingPoint, also underlined the importance of big data for central banks with a focus on the actual challenges posed by data collection and analytics.[15] One key lesson was that central banks are more and more developing their own data platforms and reporting frameworks, in particular to handle regulatory data collection. Such data are increasingly important, given central banks' greater post-crisis involvement in financial stability issues and /or supervisory functions. Thus, decisions on data have become of strategic importance to central banks. Many are now rethinking their IT system architecture and data governance framework to access big data sources or use big data techniques. Some have appointed chief data officers and have put in place coherent data strategies.

Yet the view from central banks was that existing processes should be enhanced to effectively handle the new and increasing amounts of supervisory, statistical and financial markets information. A critical factor from this perspective is the limitations faced in terms of human and IT resources. It is also essential to set up a new information value chain to replace traditional template-driven data collections; to access granular, micro-level data from various different sources and at reasonable cost; to develop process automation; and to facilitate the link between micro data points and aggregated macro indicators so as to "go beyond the aggregates". But this requires a greater harmonisation of data sets and the integration of various IT systems, both among reporters and between supervisory authorities and supervised entities. One way forward is to introduce automated secure data transfer mechanisms based on standards such as XBRL[16] and SMDX[17] and to explore innovations such as blockchain and distributed ledger technology (DLT).[18]

The third presentation, by the Business Reporting Advisory Group, highlighted the large number of big data pools generated by various regulations. This information can be very rich for central banks, with potential applications for eg financial supervision, inflation assessment, the monitoring of specific market participants.

[15] See also E Glass, "Survey analysis – Big data in central banks", Central Banking Focus Report, 2016, www.centralbanking.com/central-banking/content-hub/2474744/big-data-in-central-banks-focus-report.

[16] eXtensible Business Reporting Language.

[17] Statistical Data and Metadata eXchange; see IFC, "Central banks' use of the SDMX standard", March 2016.

[18] See Committee on Payments and Market Infrastructures, "Distributed ledger technology in payment, clearing and settlement – An analytical framework", February 2017, especially p 2. DLT refers to the processes and related technologies that enable nodes in a network (that is, a computer participating in the operation of a DLT arrangement) to securely propose, validate and record state changes (or updates) to a synchronised ledger that is distributed across the network's nodes. Financial transactions are recorded in a "block" (or batch of transactions), which is added to a chain comprising a history of transactions and is known as a "blockchain".

While individually these data pools may not fall into the category of big data analysis, together they constitute a "data lake" that can be usefully exploited via big data algorithms. To this end, it is important to develop adequate data standards, identifiers and dictionaries. Fortunately, central banks (like other financial institutions) are increasingly relying on a number of data standards (in particular SDMX, XBRL as well as ISO 20022 for payments standards) and identifiers when collecting and processing the data, and this should be strongly supported.

3. Internet data sets

The second session, chaired by Gülbin Şahinbeyoğlu, Central Bank of the Republic of Turkey, reviewed central banks' use of internet data sets. This web-based information comprises a variety of indicators, such as search queries, the recording of clicks on specific pages, the display of commercial information (see the third session of the seminar) and text posted online (see the fourth session).

The first presentation by the Central Bank of Armenia described its recent experience in collecting various kinds of web-based data for different purposes. A key message was that the increasing amount of information generated by web and electronic devices can be effectively used to complement official statistics. One example is the collection of prices posted online by supermarkets on a daily basis, which allows advanced estimates of consumer inflation to be computed. A second example is the collection of housing prices displayed by real estate agencies on their websites, which has helped to create a housing price index (there is no such index based on traditional statistics in Armenia). In addition, the fact that one can easily capture the various housing characteristics related to these web announcements has facilitated the application of hedonic price methodologies. A third example is the collection of job announcements posted by employment agencies on the web, which has led to the computing of a leading indicator of business activity. Yet these experiences highlighted a number of challenges for the central bank. One was the need to use new techniques (eg web-scraping tools) and methodologies. Another was the difficulty of automatically and easily accessing the data, as well as collecting the information consistently over time (a particular issue when collecting the prices of goods that are kept identical on the web only for a short period). Lastly, experience points to the limited quality of the data collected, especially when announcement prices captured on the web differ from actual transaction prices (an issue of particular importance for house prices). Despite these challenges, the central bank was actively seeking to expand its big data work to exploit administrative and social media sources.

These findings were echoed by Sveriges Riksbank, which also uses internet data sets for its inflation forecasts, although on a more limited scale. The central bank scrapes from the internet sales price information as regularly posted by grocery retailers with an online presence. However, this collection focuses on the specific component of the price index covering the consumption of fruit and vegetables. The reason is the high volatility of this price component and the related difficulties in forecasting it. Overall, the approach followed appears to be a useful way of enhancing short-term forecasts of inflation patterns. Moreover, the data collection process has proved to be robust and scalable, and requires little in the way of maintenance costs.

A presentation by the Central Bank of Aruba focused on using internet search queries to forecast tourism receipts, something of particular importance for this island economy. The approach tackles three key issues. One is the lag involved in producing the official tourism statistics, which can be "nowcasted" using online data sources. The second advantage is the possibility of capturing unsuspected patterns in the data: instead of inferring statistical relationships, as with "traditional" statistical modelling, big data algorithms such as machine learning models allow a wide range of effects to be incorporated (for instance, seasonal patterns, non-linearities, lagged effects) without the need to make ex ante assumptions. Third, these new techniques appear efficient in terms of predictive capacity, and can be implemented easily and in an automated way using standard packages developed by the industry. Yet one drawback is the limited interpretability of such relationships derived from "black box" calculations.

The final presentation, by the Bundesbank, described the use of web data to capture depositors' expectations. This work highlighted three important features of internet data sets for policymakers. One is that they can be used as a proxy when no available data exist: in that case, information on internet queries related to the term "*deposit insurance*" proved to be a valuable proxy for depositor concerns about funds held in banks. Second, internet-based information can be usefully complemented with other, more traditional data sources – in this case, the Bundesbank's statistics on interest rates and the balances of overnight banking deposits. Third, the (near) real time availability of web data allows trends in deposits to be anticipated and the risk of bank runs to be analysed, depending on the causality patterns found in the relevant variables. From this perspective, the Bundesbank is set to expand its approach to other big data sets, eg social media.

4. Financial, administrative and commercial data sets

The third session, chaired by Aurel Schubert, considered the broader range of data sets that qualify as "big data", in particular those comprising financial, administrative and commercial records. Certainly, the distinction between these data sets and web-based data sets can be artificial, since a significant part of the information collected on the web can be the result of commercial activities (eg the examples presented in the second session).

One example was the presentation by Bank Indonesia on collecting price lists to construct property market indicators. Previously, this information used to be posted by property agents or in newspaper ads. But most of it has now moved to the web, and in particular to a small number of property online websites – the three largest covered by Bank Indonesia's data collection represent a market share of above 50%. One advantage for the Bank is to complement its traditional statistics on property prices, which are derived from surveys available only quarterly and for a limited number of large cities. Moreover, the data are relatively easy to access, representing a significant amount of information – more than 2 million data points per month. However, significant data quality issues have arisen. One reason is the questionable accuracy of the information that individuals input to the web, which can be prone to errors and typos. Another limitation is that the information is not well structured, particularly as to property locations. Moreover, values are duplicated, since the same property can be advertised in various places. Furthermore, information accuracy varies over time, since previous advertisements can be re-posted after the expiration time, or because the announcement can continue to be posted even after the property is sold. To address these challenges, the central bank had to set up a clear and precise information process distinguishing four main phases: data acquisition (ie downloading the information); data preparation (ie detection of the characteristics such as the location of the property, removal of duplicated advertisements); data processing (ie removal of outliers and extraction of indices); and data validation.

Another presentation, by the Central Bank of Chile, was based on the collection of a fiscal database covering more than 10 years of (anonymised) tax records for roughly 25 million taxpayers. A key advantage was the richness of this data set, which allows the computation of various indicators over time and, in turn, the analysis of the factors driving business demography (eg survival rates). Yet a key challenge, apart from confidentiality constraints and the need to anonymise data, is the lack of quality control as well as the significant number of missing values. Hence cleaning the data requires significant preparatory work, for instance, to delete extreme values as well as deal with missing records.

Another example presented by the Bank of Portugal focused on credit registries, which have become the largest data sets maintained by some central banks. These data are well structured, but they qualify as "big data sets" since the information is highly granular (covering most individuals and corporations applying for a credit), contains multiple characteristics (eg on the debtor, the credit extended, the instrument used etc) and is often complex to analyse. In Europe, for instance, the AnaCredit[19] project is leading to the collection of almost 200 attributes per data point on a monthly basis (and on a daily basis for a subset). Reflecting the complexity of this information as well as its sheer importance for central bank functions, the project has triggered a full rethink of central bank information management frameworks. In particular, attention has focused on the rationalisation of data collection and management; the need to harmonise the underlying statistical concepts and ensure that consistent data can be used for multiple purposes; the set up of a single entry point for reporting and accessing the information; and the willingness to limit the associated reporting burden as possible. All in all, the project has proved instrumental in steering central banks' attention to the need to manage information in an integrated way across units.

5. Central bank communication

The fourth session, chaired by Toh Hock Chai, Central Bank of Malaysia, reviewed the implications of big data sources in central bank communication. The first presentation by the BIS recalled that that finding text similarities across a large sample of documents can be very difficult. Big data techniques, in particular text-mining technologies, can facilitate such work. One way is to build a semantic similarity database: all the words are first extracted from the textual information of interest (in the BIS study, the speeches delivered by the Federal Reserve Board members over two decades); these words are then characterised by attributes covering various dimensions in a vectoral space; and similarities between two words can be measured by the proximity of these attributes. For instance, the exercise showed that the word

[19] See also A Schubert, "AnaCredit: banking with (pretty) big data", Central Banking Focus Report, 2016.

"*forward*" appears to have close similarity with "*guidance*" and "*communicate*". Interestingly, the techniques allow these relationships to be tracked over time. For instance, and not surprisingly, the word "*systemic*" was deemed to be associated with "*macroprudential*" in the post-GFC period, but less so before 2007.

The second presentation, by the ECB, showed how such techniques could be used to assess the impact of policy communication and expectations for policy decisions. This is a relatively new topic of interest for central banks. In the past, attention focused mainly on comparing outcomes in financial markets with central bank intentions, for instance, by conducting "event studies" around the times of policy decisions. The new techniques now allow the perception of public messages by the various stakeholders to be assessed, thus providing a possible way of fine-tuning policy communication. For instance, the ECB applies computational linguistic techniques to select the words used in its statements that have the highest discriminative power and can thereby gauge the tone of its communications. Based on a global news database covering the ECB press conferences, the index allows communication phases with a "hawkish tone" to be distinguished from those "dovish" ones. The usefulness of this index can be checked by looking at its correlation with other variables (eg interest rates, to see whether actual policy changes are correctly anticipated by media reports). This experience suggests that a quantitative approach to central bank communication is possible and can provide useful insights.

The third presentation by the Bank of Italy noted that most information available on the web is textual and can therefore be exploited through ad hoc techniques. Moreover, it was particularly important to have an objective measure of the central bank's communication and of its impact on the sentiment of stakeholders. To this end, textual information is used to create a heatmap showing the usage of specific words over time. The approach can also provide insights to assess the readability and formality of central bank communication. Lastly, this allows for semantic analysis, by extracting the contextual usage of specific words and analysing similarities across documents.

6. Big data governance

The seminar ended with a panel discussion on issues related to big data governance, chaired by Katherine Hennings, Central Bank of Brazil and IFC Vice Chair. The discussions reviewed big data work in central banks and covered related resource issues, especially in HR and IT.

One view was that central banks are relatively new in exploiting big data, in contrast to the long-standing experience of national statistical offices (NSOs) in handling large and confidential data sets such as censuses and administrative records. A key reason was that central banks have traditionally been data users rather than data producers. They have been catching up rapidly, especially since the GFC, with more and more central banks being called on to collect, produce and use large data sets. Yet the lessons learned are mainly tentative, as big data sources of information are still under evaluation in most central banks.

What is unknown was whether these new data developments will lead to a change in the institutions' business models. A commonly shared view was that this would not fundamentally change what central banks do, given their role as data users

– unlike most NSOs, which are usually not obliged to use their data for policy purposes. Yet there were specific areas in which central banks' processes may have to change significantly with the advent of big data – for instance, short-term forecasting and nowcasting, IT security etc. This puts a premium on enhancing the governance of related data sets, in particular by establishing formal *Memoranda of Understanding* with the related data providers. In any case, central banks have multifaceted mandates, particularly in the post-crisis era. Dealing with the increasing supply of big data while combining old and new roles may thus prove challenging.

Another governance issue is to maximise the use of new information available to support central bank policy functions while managing the associated risks. This calls for existing data governance frameworks to be revamped. For instance, use of internet-based information raises several difficulties, as compared with the production of traditional statistics. First are the legal, financial and ethical issues posed by accessing (often private) information that is a by-product of commercial activities. Then there are the operational, legal and reputational risks entailed in dealing with transaction-level information that is potentially confidential; the responsibility for authorising such data collections, not least as regards aspects such as confidentiality protection, data ownership and privacy; the degree of statistical accuracy of these data and the level of confidence in their sources; and even the information content of data derived from self-generated activities – for instance, the information value of the number of clicks on a specific topic will vary as these clicks are influenced by the search engines and based on users' past searches. One view was that the complexity of these issues might well increase as central banks move from experimentation to the actual regular production and use of big data-based information.

From this perspective, the consensus was that cooperation (both internationally among central banks and domestically among statistical authorities) should and would indeed expand to facilitate the exploration of the big data area. This reflects the fact that there was a lot to learn from each other. Yet one view was that such cooperation may prove temporary and may well recede once big data has become a more mature area and sufficient work expertise has been developed in central banks.

Turning to resource challenges resulting from handling big data, these mainly reflect the sheer size of the data sets, their lack of structure and the often poor quality of raw data obtained from internet streaming, large administrative records or other sources. Moreover, sophisticated statistical techniques are often required to derive meaningful information from such data.

A major area is IT. The implications of big data for central banks' information systems are potentially huge. There are large IT processing costs and difficult and expensive technology choices have to be made. One risk is to spend more time and resources on cumbersome activities – cleaning the data, organising the underlying platforms etc – rather than on actually analysing and using the data. One way to address this risk is to focus on specific use cases and resist the temptation to collect various large data sets covering an excessively wide range of purposes.

Another issue was that public statisticians have a tendency to be "cloud computing-adverse", mainly because of the disclosure risks posed for confidential information. Most prefer to operate in a "secluded" data environment. This may well reduce the scope for public authorities to benefit from new big data techniques developed in the marketplace – for instance, some applications may be available only as part of a cloud-based solution. Nevertheless, it was also recognised that a number of big data sources have a public nature, implying that central banks may have sufficient opportunities to make use of private sector solutions. And, within central banks, important organisational changes were also expected to better deal with big data, including the creation of internal centres for big data statistics, data lakes, internal clouds etc. In any case, experimentation will shed more light on these aspects.

A second key area is staff. The necessary skills may not be available in-house, especially in IT, data science and methodology as well legal expertise. Given the limited supply of graduates, central banks may well face a "war for talent". This could be a key obstacle, since skilled staff are a prerequisite if central banks are to benefit from big data opportunities as well as manage the associated risks. Moreover, the skills shortage also raises questions around compensation and staff career paths, as well as management issues – one view, for instance, was that the relatively important role played by economists in central banks' managerial positions might well be called into question by these developments.

In any case, these challenges highlighted the likelihood that significant time and effort will be needed before a regular production of big data-based information can be undertaken to support central bank statistical and analytical work on a large scale.

IFC–Bank Indonesia Satellite Seminar on "Big Data", Bali, 21 March 2017

Opening remarks by Yati Kurniati, Executive Director of Statistics Department, Bank Indonesia

It is a great pleasure for me to welcome all of you to the IFC Satellite Seminar on "Big Data", here in the beautiful island of Bali. Bank Indonesia is honored to jointly organize this event with the Irving Fisher Committee on Central Bank Statistics (IFC). On behalf of Bank Indonesia, I would like to express my appreciation to all of those – IFC Executives and members, distinguished guests, speakers and participants – who are contributing and taking part in this IFC Seminar. The seminar is held in conjunction with the 2017 International Statistical Institute (ISI) Regional Statistics Conference, which will be held on 22-24 March 2017 at this same place. This second regional conference of ISI is organized by the ISI and its South East Asia Regional Network (ISI-SEA Network) together with Bank Indonesia as the co-host, and in collaboration with Badan Pusat Statistik (National Statistics Office of Indonesia), Ikatan Perstatistikan Indonesia (Association of Indonesia Statistician), and Forum Masyarakat Statistik (Indonesia Statistician Forum). We have a very stimulating program lined up for the coming days, which I am positive will prompt very rich discussions and exchanges of views. I would therefore extend my appreciation to all those who will be contributing to and taking part in the ISI regional conference.

Data and statistics are basically a cornerstone of the central bank's work. In recent years, the supply of data has increased dramatically and this trend is set to continue as an ever-greater amount of activities are stored in different ways. This data revolution, which has given rise to concepts such as Big Data, challenges traditional thinking while placing new demands on processing and analysis.

The private sector has been aware of the value of Big Data for some years now. Central banks, however, tend to be more cautious, but there is a strong interest in Big Data in the central banking community, in particular at senior policy level, as being shown from the results of a survey conducted by the Irving Fisher Committee on Central Bank Statistics in October 2015.

Many of us believe that Big Data has the potential to open up new possibilities for monetary policy-making, financial system supervision, and economic research. Data mining, text mining, image processing and other new techniques have been made possible by improvements in processing technology. Central banks have also pointed out what they see as areas of high potential for Big Data. These areas include, among others, enhancing real-time awareness by providing real-time information, strengthening macro-prudential oversight by building up new risk indicators based on households and corporation behavior, and improving business cycle analysis by enriching sentiment indicators.

As many central banks start to make the use of Big Data, challenges come to the forefront. To access, prepare, and analyze data sets presents a series of institutional and technical challenges. One of the first challenges which must be overcome in order to conduct Big Data analysis is for the data to be accessed. This is particularly important for the institutions which do not, themselves, generate the data of interest. This is also linked with the issue of personal data ownership and confidentiality. To this end, business models need to be developed to ensure that private sectors are willing to share data. Also, authorities need to design policy to help capture the value of Big Data and enable sharing across agencies. Once data is accessible, preparing the data and ensuring its readiness for further analysis requires challenges such as the time and effort to clean data. After the data is cleaned, a variety of algorithms must be deployed to extract the values from the data. Each phase of the data processing mentioned above highlights the technological capabilities necessary to work with Big Data effectively. In terms of accessing data, it is important to have the necessary infrastructures (hardware and software) to collect data. If data is dynamically fed from an online source such as Twitter, for example, then the infrastructures must allow for such real-time, continuous updated analysis. If, instead, data is being downloaded from some source and then kept for later analysis, it is important to ensure sufficient hardware capacity to store such data. Given a well-prepared dataset, a series of considerations must also be kept in mind to interpret the data. In particular, the statistical problem of selection bias, in which a large datasets collected is unrepresentative of a population, may occur in various forms.

To sum it up, utilizing Big Data poses considerable challenges for central bank. Another major concern has been the difficulty to recruit people with the skills of data scientists. In terms of infrastructures, the IT capacity of many central banks have not always lived up to the requirements laid down for processing large amounts of data. These all take new considerations in terms of strategy, organization, and budgets. Despite the challenges, the central banks are now in the process of make the best use of it. In fact, today we are going to see some of the interim results of fascinating pilot projects using various source of Big Data presented by our fellow central bankers.

The age of Big Data is upon us.

The value, volume, and variety of data are undoubtedly multiplying, the methodologies that enable us to analyze those data are maturing, and the technologies to access and processing those data are emerging. Through a concerted and collaborative effort at various levels, I believe that Big Data can be utilized to strengthen central bank to serve its goals. And I am very pleased to note that the IFC and the ISI are with us in this endeavor.

With this, ladies and gentlemen, I conclude my opening remarks. I wish all of you a productive and engaging seminar. I also hope that you are able to enjoy at least, a glimpse of Bali’s most natural and cultural wonders. Thank you.

IFC–Bank Indonesia Satellite Seminar on "Big Data", Bali, 21 March 2017

Opening remarks by Aurel Schubert, Director General Statistics, Statistics Department, European Central Bank, and Vice Chair, IFC Executive

Thank you very much Ms Yati Kurniati and the Bank Indonesia for your hospitality and your kind words. It is my pleasure and honour to supplement these welcoming remarks as one of the two Vice chairs of the IFC Executive and to guide you through today's big data session or pretty big data, as I often call it;

In these days, when in the political debate facts and figures are often ignored, altered or replaced by "alternative facts", our efforts for good data become even more important than ever. The advent and existence of Big Data might create important complementary sources for our work. To take stock in this exploration journey, this is what today's seminar is all about.

This seminar demonstrates the importance of the Irving Fisher Committee on Central Bank Statistics (IFC) as part of getting the central banking community together in exploring the synergies and benefits of using big data for central banking purposes. In this aspect, I would like to thank the coordinators, in particular Bruno Tissot (BIS) and Claudia Huber (from the IFC secretariat) and Per Nymand-Andersen, from the ECB, and to the leaders of the four pilot groups.

You may recall that the IFC launched an online survey on central banks' use of and interest in big data, whereby close to fifty central banks expressed interest to cooperate with other IFC members on "big data" and to contribute to a common roadmap on its usefulness for central banking.

This seminar is indeed an outcome of this ground work and a core milestone as part of the IFC's structured and coordinated journey;

Indeed now several central banks have joined forces working on four different pilot categories – a special thanks to all of them - relating to

1. Administrative data
2. Internet datasets
3. Commercial datasets
4. Financial market data

The concept is laid out whereby all pilots follow a similar structure and approach in managing the work, processes and expected outcome within a certain time frame.

Each group of participating central banks are describing the characteristics of the data and supplier according to a standardised set of key information for each of the five statistical production processes covering;

1. “Input”,
2. “Quality assessment”,
3. “Production”,
4. “Results” and
5. “Assessment”

The idea is to collect these pilot reports and to provide a consolidated IFC publication.

I look forward to learn from these preliminary results which will be presented at today’s Seminar and I encourage you to contribute to the discussion.

Let me just swiftly walk you through todays programme; please first be inspired by the opening keynote speaker Mr Agus Sudjianto, Executive Vice President, Head of Corporate Model Risk at Wells Fargo, who will share with us how a large commercial bank performs stress testing with lots of data to evaluate its own safety and soundness.

This will be followed by two sessions presenting the progress of the IFC pilots on using big data for central banks and internet datasets.

Please join me as well for the lunch presentation on ”Data science – Some Perspectives” by Vijay Nair, former President of ISI and a good friend of the central banking statistics community.

In the afternoon, I have the pleasure of chairing session three, on the progress with financial, administrative and commercial large datasets. This will be followed by a new subject on central banking communication and a panel discussion on big data governance, and concluding remarks by Katherine Hennings, the other Vice chair of the IFC.

This is a pretty BIG agenda for big data. So, let the seminar begin; Mr Agus Sudjianto, the floor is yours;

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Quantitative risk management and stress test to ensure safety and soundness of financial institutions[1]

Agus Sudjianto,
Executive Vice President, Head of Corporate Model Risk, Wells Fargo

[1] This presentation was prepared for the meeting. 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 meeting.

WELLS FARGO

Quantitative risk management and stress test to ensure safety and soundness of financial institutions

Agus Sudjianto, Ph.D.

Executive Vice President, Head of Corporate Model Risk

March 21, 2017

Together we'll go far

Stress test

- In the wake of the financial crisis, U.S. Congress enacted the Dodd-Frank Act
  - Requires the Federal Reserve to conduct an annual stress test
  - Seeks to ensure BHCs have sufficient capital to continue operations throughout times of economic and financial market stress
- Projects balance sheets, RWAs, net income, and resulting post-stress capital over a nine-quarter "planning horizon"
  - BHC stress scenario: internally generated scenarios (Baseline and Adverse) customized to idiosyncratic risk of BHC
  - Supervisory scenario: Baseline, Adverse, Severely Adverse

FRB guidance for quantitative methodologies/ models

- Stress test is a forward-looking quantitative evaluation of the impact of stressful economic and financial market conditions on BHC capital
- Specific expectations in terms of quantitative tools/models and their governance:
  - SR15-18: FRB Capital Planning Guidance
    - Use of Models and Other Estimation Approaches
    - Model Overlays
    - Use of Benchmark Models
    - Sensitivity Analysis and Assumptions Management
  - SR11-7: FRB Model Risk Management Guidance
    - Model Development, Implementation and Use
    - Model Validation
    - Model Governance, Policy, and Control

Applications of models

- Economic Scenario Generation
  - Firm-specific scenarios: specific vulnerabilities of the firm’s risk profile
  - Multiple stressful conditions or events can occur simultaneously or in rapid succession
- Loss Estimation
  - Credit risk losses on loans and securities
  - Fair-value losses on loans and securities
  - Market and default risks on trading and counterparty exposures
  - Operational-risk losses

Applications of models (continued)

- Pre-Provision Net Revenue (PPNR)
  - Net interest income
  - Non-interest income
  - Non-interest expense
- Risk Weighted Asset (RWA)

Model data/input and sources

- SR15-18 Guidance
  - Disaggregated levels to capture observed variations in risk characteristics and performance across sub-portfolios/segments under changing conditions
  - Internal data to estimate Losses and PPNR when possible
- Data quality and relevance
  - Downturn historical data
  - Suitability for the model and consistent with the modeling framework
    - Included/excluded data and proxies for model development population, rationale, and impact on results
    - Representative of the bank's portfolio
    - Reconciles with general reporting information (e.g., GL) as applicable

Modeling consideration

- SR15-18 Guidance
  - Separately estimate Losses and PPNR for portfolios or business lines that are sensitive to different risk drivers
  - Qualitative Approaches are allowable in limited cases
- Model requires both accuracy and sensitivity; where the later might be more important
  - Loss forecasting: performance both for short- and long-term predictions are important
  - Stress Test: sensitivity is more important than model fit
- Proper granularity and segmentations are critical to deal with changing portfolio composition

Modeling consideration (continued)

- Beware of correlation between dynamic input or “time” dummy variables which can mute the impact of macroeconomic variables
- Treatment dynamic variables which cannot be predicted
  - Time-varying behavioral variables

Modeling framework

- Credit/PPNR Models
  - Account level modeling
    - Conditional (i.e., hazard) model/panel regression
    - Credit rating migration model
  - Pool level models: vintage, segment, or behavior pool
  - Time-series regression
  - Choice consideration: granularity to capture portfolio changes, ability to capture important drivers, data availability, resource/timing, and on-going maintenance
- Market Models
  - Full revaluation using Front Office pricing model
    - Need to evaluate the model function properly during stress condition: stability, convergence, no arbitrage
  - Approximation (Greek-based) models
  - Need Risk not in Model to deal with limitation

General modeling framework

- Let *T* a random time of account closing (e.g., due to default or attrition/prepayment), the hazard function is modeled as a regression with *g*(.) link function and covariates *Z*(*s*)

$$\lambda\{t|Z(s)\} = g[\lambda_0(t), Z(s)]$$

- Where $\lambda_0(t)$ is the baseline hazard to represent the effects of unobserved factors and *s* is the observation time which can be:
  - Static such as time of origination, *s* = 0
  - Dynamics
    - Last snapshot information without future prediction
    - Including future prediction, i.e. *s* = *t* and prediction model *Z*(*t*) is available such as PPNR models (e.g., utilization or spend rate) or macro-economic factors

$$Z\{t|X(s)\} = h[Z_0(t), X(s)]$$

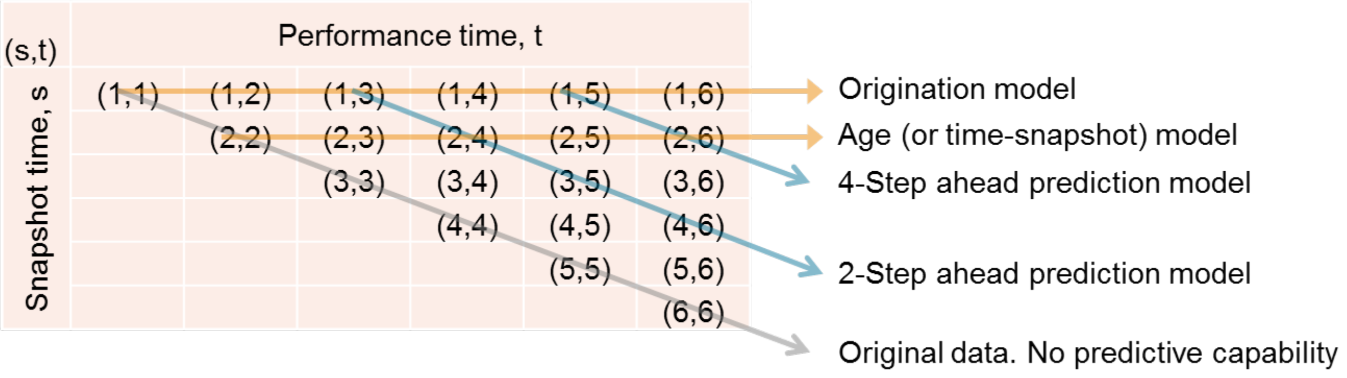
Dynamic covariates and data stacking

- Dynamic factors that no future prediction are available but they are critical such as refreshed FICO, Utilization, etc., and need to be handled through ‘data stacking’ approach

Observation Data

Snapshot Date, s	Snapshot FICO, x1	Snapshot Delinquency, x2	Performance Date, t	MOB, m	Unemployment, x3	Default	Time after snapshot, k
13-Jan(1)	675	Current	13-Jan(1)	1	7.2	0	0
13-Feb(2)	666	Current	13-Feb(2)	2	7.2	0	0
13-Mar(3)	630	30	13-Mar(3)	3	7.2	0	0
13-Apr(4)	620	60	13-Apr(4)	4	7.2	0	0
13-May(5)	620	90	13-May(5)	5	7	0	0
13-Jun(6)	620	120	13-Jun(6)	6	6.7	1	0

Dynamic Factor without future predicted values



(s,t)	Performance time, t					
Snapshot time, s	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
		(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
			(3,3)	(3,4)	(3,5)	(3,6)
				(4,4)	(4,5)	(4,6)
					(5,5)	(5,6)
						(6,6)

Origination model

Age (or time-snapshot) model

4-Step ahead prediction model

2-Step ahead prediction model

Original data. No predictive capability

Model validation depth and scope

- Soundness of modeling approach
  - Methodology, granularity, data quality, and treatment (coverage, proxy, etc.), parameter estimation/calibration
- Model stability under market shock
  - computational stability, parameter stability, reasonable outcome
- Rigor of model performance evaluation
  - Backtesting to previous stress condition
  - Out-of-sample and out-of-time testing
  - Sensitivity to risk varying risk drivers
    - Separation across different scenarios
    - Consistency with respect to scenarios
- Issues and limitations
  - Risk in model, risk not in model, parameter uncertainty
- Holistic approach
  - Not only focus on the targeted core models, but also include critical upstream and downstream models and tools
- Thorough documentation

Model validation:

Replication

- Independently rerunning/recoding models to confirm and evaluate model outputs
- In-sample backtesting
  - Multiple forecast starting points covering different parts of the economic cycle
  - Model performance for all segments and alternative segments.
- Out-of-sample/out-of-time performance
  - Out-of-development periods test
  - Model performance when “stress-time window” is excluded from parameter estimation
    - Appropriateness for future scenarios where such scenarios do not exist in the development sample
    - Out-of-time forecast performance
    - Parameter stability
- Sensitivity analysis and testing
  - Model sensitivity under distinct economic scenarios
  - Sensitivity to input changes

Model validation:

Benchmarking

- Distinct modeling alternatives
- Evaluate model performance when the true outcomes are unknown (i.e., Stress testing models)
- Diagnose appropriateness of modeling choice
  - Model structure including the simplification choice
  - Segmentation
  - Variable selection, non-linearity, interactions
- Model alternatives used by validators needs to be comprehensive and insightful and are likely to be more complicated and perform better than production models
  - Not constrained by the requirement for model maintenance and operational computation time

Evaluating the dynamics of stress testing models

Dynamics of Horizon Prediction:

$$\lambda_i(t|s) = \beta_0(k) + \boldsymbol{x}_i^T(s,t)\boldsymbol{\beta}(k)$$

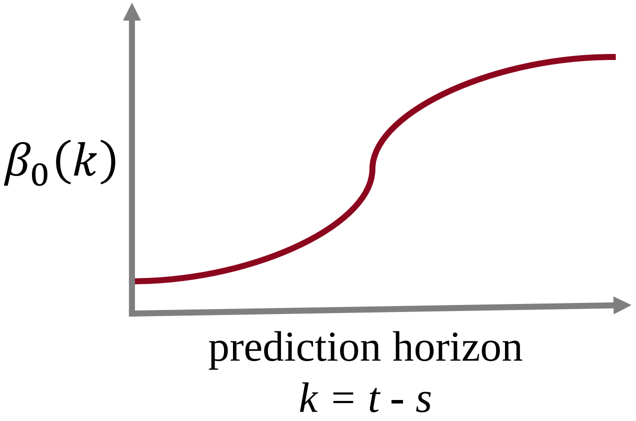
Prediction of time t given the 'snapshot' information at time s

Dynamic covariates:

- Economic factor s<–t
- Behavioral covariate t<–s

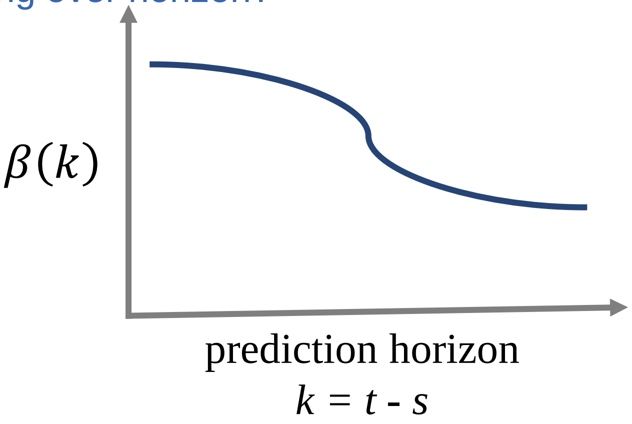
Is there effect from unobserved variables?

- e.g., baseline hazard in PD model



Is the sensitivity change over the horizon?

- e.g., is the effect of FICO at time snapshot decaying over horizon?



Machine learning for variable selections

Alternative Model: Machine Learning (ML)

Model importance ranking

- ML embedded method importance measure (e.g. gradient boosting machine(GBM), random forest)
- ML filter methods ranking(univariate and multivariate)

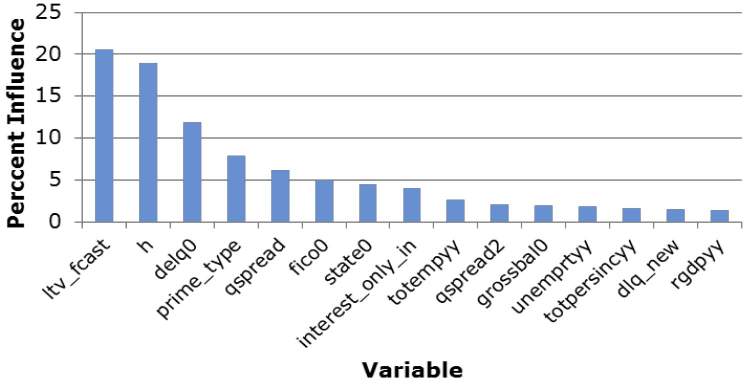
Model interaction selection

- ML H-statistics/ML 2D partial dependent plot
- GLM elastic net with regularization on interactions

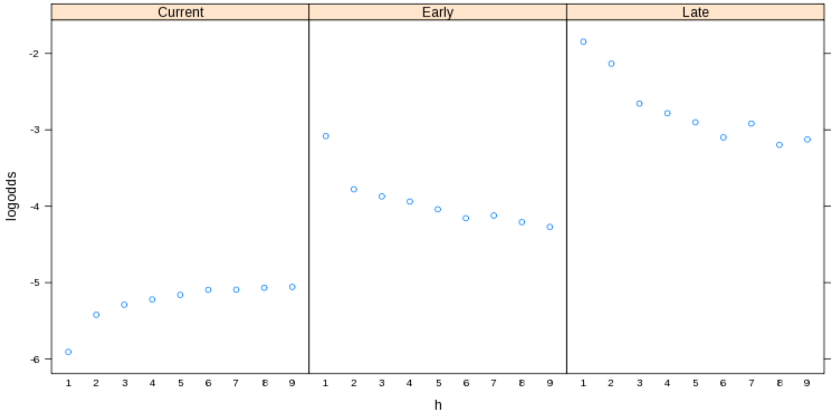
Nonlinearity detection

- ML 1D partial dependent plot

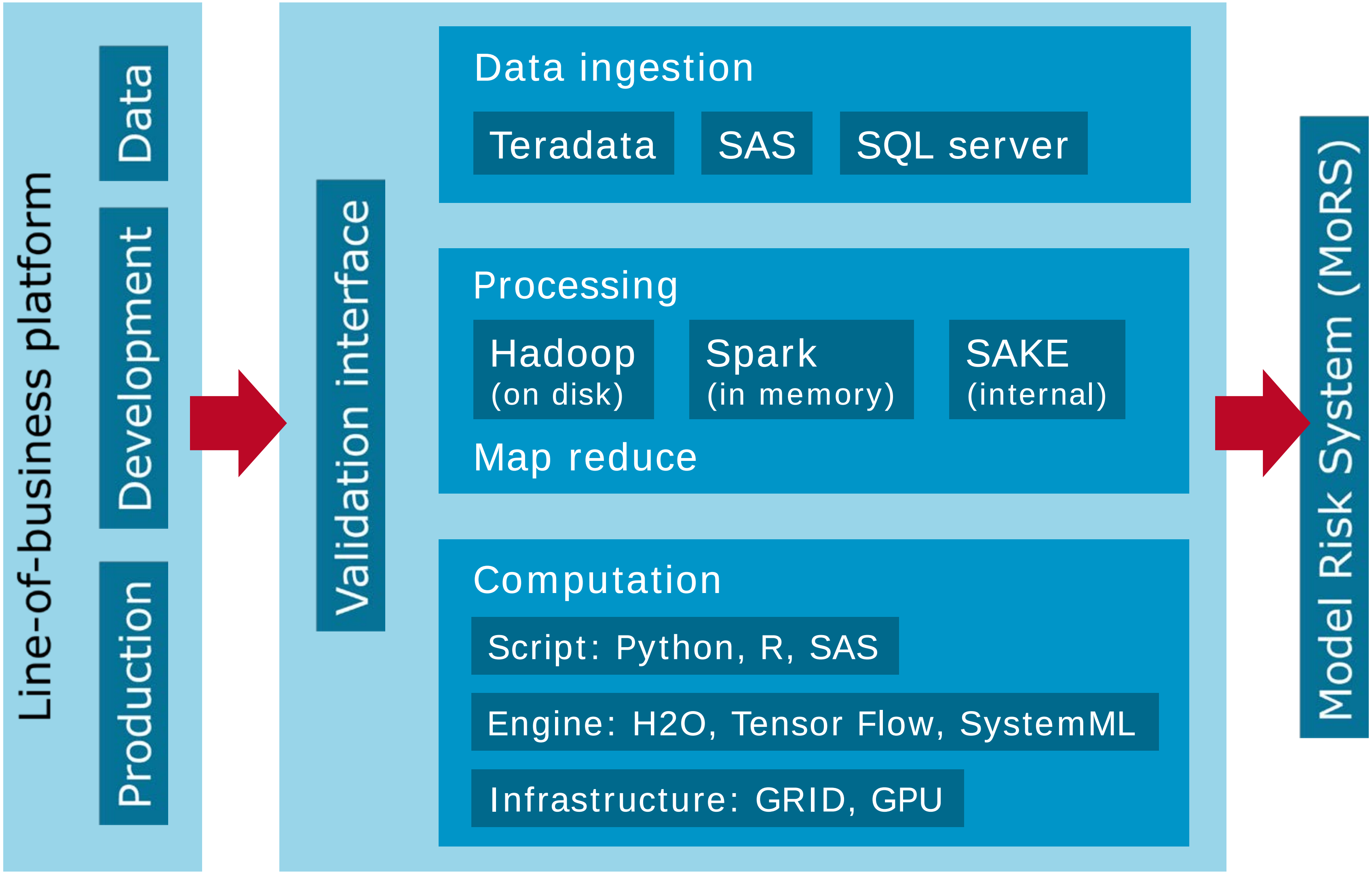
Importance ranking using GBM



Nonlinearity and Interaction



Validation platform



Line-of-business platform

Production

Development

Data

Validation interface

Data ingestion

Teradata

SAS

SQL server

Processing

Hadoop (on disk)

Spark (in memory)

SAKE (internal)

Map reduce

Computation

Script: Python, R, SAS

Engine: H2O, Tensor Flow, SystemML

Infrastructure: GRID, GPU

Model Risk System (MoRS)

Compensating model weakness during usage:

Overlays

- Models are often have weakness and limitation due to:
  - Risk in Model:
    - Outstanding issues, limitations, or restriction identified during model validations or performance monitoring
    - Model dependency
      - Weakness of upstream (feeder) models
      - Uncertainty of input assumptions
  - Risk Not in Model: model limitation to capture risk drivers listed in the stress test risk identification process
    - Factors in economic scenario that are not in the models
    - Idiosyncratic factors both external events or business drivers/strategy

Compensating model weakness during usage:

Overlays

- Compensating factors such as model overlays are typically applied for model weakness
  - Quantitative overlay: model benchmark, quantitative analysis, back testing, sensitivity analysis
  - Qualitative overlay: management judgment

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Central banks' use of and interest in "big data"[1]

Jens Mehrhoff, Eurostat

[1] This presentation was prepared for the meeting. 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 meeting.

DEUTSCHE
BUNDESBANK
EUROSYSTEM

Central banks’ use of and interest in “big data”

Session 1: “Big Data for Central Banks”

Jens Mehrhoff*, currently on secondment to the European Commission

* This presentation represents the author’s personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank, the European Commission or its staff.

Structure of the presentation

1. Introduction

2. Identifying sources

3. Joint conceptual framework and roadmap

4. Statistical paradises and paradoxes

5. One way forward for official statistics

“*But the ‘big data’ that interests many companies is what we might call ‘found data’, the digital exhaust of web searches, credit card payments and mobiles pinging the nearest phone mast.”* Tim Harford, Financial Times, 28 March 2014.

1. Introduction

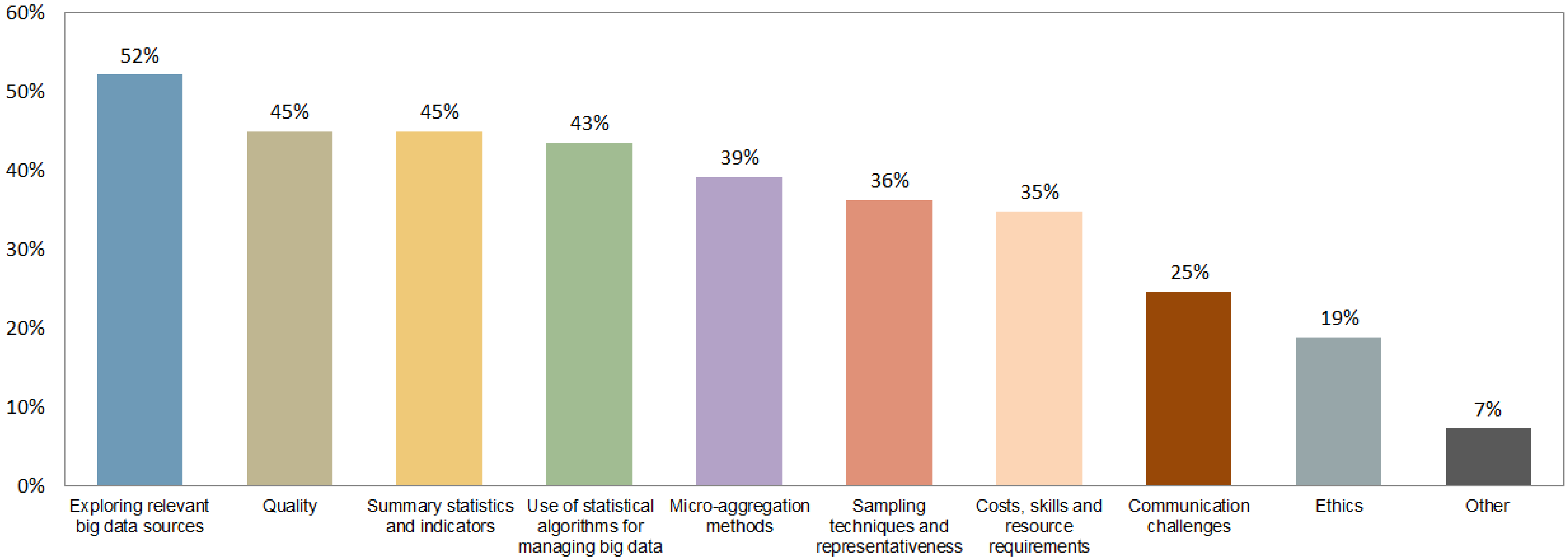
- In January 2015 the Irving Fisher Committee on Central Bank Statistics launched an **online survey on central banks' use of and interest in "big data"**.
- The aim of this survey was twofold:
  - To **take stock of central banking experience** in the use of big data; and
  - To **explore central banks' interest** in this topic with a view to defining a roadmap for further action.
- The **vast majority (69) of IFC member central banks responded**, representing a response rate of 83%.
- The **main conclusions** of the survey are the following. http://www.bis.org/ifc/publ/ifc-report-bigdata.pdf

1. Introduction

- There is **strong interest in big data** in the central banking community, in particular at **senior policy level**.
- Central banks actual involvement in the **use of big data is currently limited**.
- Big data can be **useful for conducting central bank policies**.
- Big data are perceived as a **potentially effective tool** in supporting macroeconomic and financial stability analyses.
- Big data may also **create new information/research needs**.
- **International cooperation can add value.**
- Exploring big data is a **complex, multifaceted task**.
- **Regular production** of big data-based information **will take time, especially because of resource issues**.

1. Introduction

Which statistical topics would be of interest to you as part of the big data subject?



Note: multiple responses possible.

2. Identifying sources

- Despite the **limited current experience** in the use of “big data”, there is a **strong interest** within the central banking community to **cooperate and share experiences** on the use of big data.

- The IFC Executive decided to select **few case studies for piloting the usefulness** of “big data”, as a potentially effective tool-kit, in supporting central banks’ monetary policy, financial stability and/or banking supervision policies and invited the **IFC community to cooperate in the piloting phase**.

- These **supplementary statistics** may provide further insights contributing to **guiding central bankers’ policy actions** as well as to **assessing the subsequent impact and associated risks** of these policy decisions on the financial system and real economy.

- The way forward may be to take **small steps in developing and applying** a structural approach for piloting the use of big data, or non-official sources, for central banking purposes.

2. Identifying sources

- The following **four of group of sources and tentative pilot projects** for showcasing could be envisaged:

1. **Administrative dataset:** internal central banks and statistics offices databases and other public databases
2. **Internet dataset:** patterns and behaviour of internet activities; examples could be internet search machines, social media consumer internet purchases
3. **Commercial dataset:** relate to for instance transactional data from payments, settlements or trading systems or mobile banking and operators
4. **Financial market data:** financial market data or data relating to individual financial instruments

3. Joint conceptual framework and roadmap

- One benefit of conducting joint pilot studies would be to have a **similar overall structure and approach** in managing the work, processes and expected outcome of the pilot projects within a certain time frame.

- Despite of differences of the “big data” sources, it is important that each group of participating central banks describe in details the **characteristics of the supplier** according to a **standardised set of key information** for each of the **five statistical production processes** covering

  - **“Input”**,
  - **“Quality”**,
  - **“Production”**,
  - **“Results”** and
  - **“Assessment”**

  as part of exploring its relevance for central banking tool-kits.

3. Joint conceptual framework and roadmap

- **Input:** e.g. Who are the source provider and what type of relevant information is available? Include an example of the information content of the source.

- **Quality:** e.g. How transparent is the source on its methodology? Please describe the sample and its representativeness.

- **Production:** e.g. Please describe the production process required. What types of modelling, statistical algorithms, machine learning techniques, text mining and semantic analysis is required?

- **Results:** e.g. Which types of statistics information/indicators can be made available? Please describe how the indicators could be used for central banking purposes.

- **Assessment:** e.g. Please provide a short overview/summary of the pilot study. Can the source easily be used for statistics production purposes?

4. Statistical paradises and paradoxes
Meng, X.L. (2014), in: Past, Present, and Future of Statistical Science.

- “*Is an 80% non-random sample ‘better’ than a 5% random sample in measurable terms? 90%? 95%? 99%?*” (Wu, 2012)

- Let us consider a case where we have an **administrative record** covering f_a percent of the population, and a **simple random sample (SRS)** from the **same population** which only covers f_s percent, where $f_s \ll f_a$.

- How large should f_a/f_s be before an estimator from the **administrative record dominates** the corresponding one from the **SRS, say in terms of MSE**?

- $\bar{x}_a = \frac{1}{n_a}\sum_{i=1}^{N} x_i R_i\,, R_i = \begin{cases} 1 & \text{if } x_i \text{ is recorded,} \\ 0 & \text{otherwise.} \end{cases}$

- The **administrative record has no probabilistic mechanism** imposed by the data collector.

4. Statistical paradises and paradoxes
Meng, X.L. (2014), in: Past, Present, and Future of Statistical Science.

- Expressing the **exact error**, where $f_a = n_a/N$:

$$\bar{x}_a - \bar{X}_N = \frac{\mathrm{E}[xR]}{\mathrm{E}[R]} - \mathrm{E}[x] = \frac{\mathrm{Cov}[x,R]}{\mathrm{E}[R]} = \underbrace{\rho_{x,R}}_{\text{Data Quality}} \cdot \underbrace{\sigma_x}_{\text{Problem Difficulty}} \cdot \underbrace{\sqrt{\frac{1-f_a}{f_a}}}_{\text{Data Quantity}} .$$

- The **MSE** of $\bar{x}_a$ is more complicated, mostly because R_i depends on x_i:

$$\mathrm{MSE}[\bar{x}_a] = \mathrm{E}\left[\rho_{x,R}^2\right] \cdot \sigma_x^2 \cdot \left(\frac{1-f_a}{f_a}\right).$$

- For **biased estimators** resulting from a large self-selected sample, the **MSE is dominated (and bounded below) by the squared bias term**, which is **controlled by the relative sample size f_a**.

- The **non-sampling errors** can be made arbitrarily **small only when the relative size f_a is made arbitrarily large**, that is $f_a \to 1$; just **making the absolute size n_a large will not do the trick**.

4. Statistical paradises and paradoxes
Meng, X.L. (2014), in: Past, Present, and Future of Statistical Science.

- A **key message** here is that, as far as statistical inference goes, what makes a **“big data”** set big is typically **not its absolute size**, but its **relative size to its population**.

- Therefore, the question **which data set one should trust more** is unanswerable without knowing N.

- But the general message is the same: when dealing with self-reported data sets, **do not be fooled by their apparent large sizes** or by common wisdom from studying probabilistic samples.

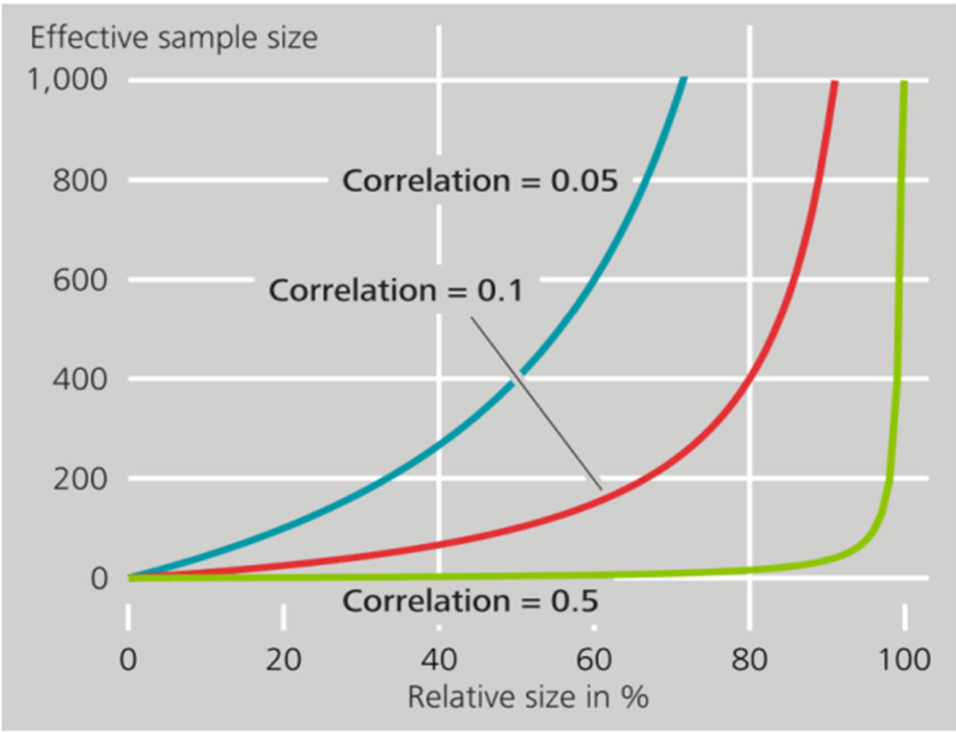
- This reconfirms the **power of probabilistic sampling** and reminds us of the **danger in blindly trusting that “big data”** must give us better answers.

- **Lesson 1:** What matters **most is the quality**, not the quantity.

4. Statistical paradises and paradoxes

Meng, X.L. (2014), in: Past, Present, and Future of Statistical Science.

The effective sample size of a “Big Data” in terms of SRS size



Deutsche Bundesbank S3IN0428.Chart

- Imagine that we are given a **SRS** with $n_s = 400$.
- If $\rho_{x,R} = 0.05$ and our intended **population is the USA**, then $N \approx 320{,}000{,}000$, and hence we will need $f_a = 50\%$ or $n_a \approx 160{,}000{,}000$ to place more trust in $\bar{x}_a$ than in $\bar{x}_s$.
- If $\rho_{x,R} = 0.1$, we will need $f_a = 80\%$ or $n_a \approx 256{,}000{,}000$ to dominate $n_s = 400$.
- If $\rho_{x,R} = 0.5$, we will need over 99% of the population to beat a SRS with $n_s = 400$.

4. Statistical paradises and paradoxes
Meng, X.L. (2014), in: Past, Present, and Future of Statistical Science.

- However, the **availability of both small random sample(s) and large non-random sample(s)** opens up many possibilities. The following (non-random) sample of questions touch on this:
  - Given **partial knowledge of the collection/response mechanism** for a (large) biased sample, what is the **optimal way to create an intentionally biased sub-sampling scheme** to counter-balance the original bias so the resulting sub-sample is guaranteed to be **less biased** than the original biased sample in terms of the sample mean, or other estimators, or predictive power?
  - What should be the **key considerations when combining small random samples with large non-random samples**, and what are the sensible **“corner-cutting” guidelines when facing resource constraints**?
- **Lesson 2:** Do not ignore seemingly tiny probabilistic datasets when combining data sources.

5. One way forward for official statistics
Groves, R.M. (2012), in: Director's Blog – Census Bureau.

- What's the **difference between "data" and "information"**?

- We're entering a world where **data will be the cheapest commodity around**, simply because the society has created **systems that automatically track transactions** of all sorts.

- Collectively, the society is **assembling data on massive amounts** of its behaviours.

- Indeed, if you think of these **processes as an ecosystem**, it is **self-measuring in increasingly broad scope**.

- Indeed, we might **label these data as "organic"**, a now-natural feature of this ecosystem.

5. One way forward for official statistics
Groves, R.M. (2012), in: Director’s Blog – Census Bureau.

- **Information is produced from data by uses.** Data streams have no meaning until they are used.
- The user finds meaning in data by **bringing questions to the data** and **finding their answers in the data**.
- An old quip notes that **a thousand monkeys at typewriters** will eventually produce the **complete works of Shakespeare**.
- The **monkeys produce “data” with every keystroke**. Only we, as **“users”, identify the Shakespearian content**.
- **Data without a user** are merely the jumbled-together **shadows of a past reality**.

5. One way forward for official statistics
Groves, R.M. (2012), in: Director’s Blog – Census Bureau.

- **What’s this got to do with official statistics?** For decades, **official statistics has created “designed data”** in contrast to “organic data.”

- The questions we ask of businesses and households **create data with a pre-specified purpose**, with a use in mind.

- Indeed, designed data through surveys and censuses are **often created by the users**.

- This means that the **ratio of information to data (for those uses) is very high**, relative to much organic data.

- **Direct estimates are made from each data item** – no need to search for a Shakespearian sonnet within the masses of data.

5. One way forward for official statistics
Groves, R.M. (2012), in: Director’s Blog – Census Bureau.

- What has changed is that the **volume of organic data produced now swamps the volume of designed data**. The **risk of confusing data with information** has grown exponentially.

- We must **collectively figure out the role of organic data** in extracting useful information about the society.

- The **challenge is to discover how to combine designed data with organic data**, to produce resources with the most efficient information-to-data ratio.

- This means we **need to learn how surveys and censuses can be designed to incorporate transaction data** continuously produced by the internet and other systems in useful ways.

- Combining data sources to **produce new information not contained in any single source is the future**. The **biggest payoff will lie in new combinations** of designed data and organic data, not in one type alone.

5. One way forward for official statistics
Groves, R.M. (2012), in: Director’s Blog – Census Bureau.

- To continue the monkey-typewriter metaphor, the **internet and other computer systems are like typewriters that have an unknown set of keys disabled**.

- Some keys are missing **but we don’t know which ones are missing**. They’re **not capturing all behaviours in the society**, just some.

- The Shakespearian library may or may not be result of the monkeys pounding on the keys. In contrast to the beauty of the bard’s words, **we may only find pedestrian jingles and conclude that’s as good as it gets**.

- We **need designed data for the missing keys**; then we **need to piece them together** with the masses of organic data from the present keys.

- **The combination of designed data with organic data is the ticket to the future.**

Contact

JENS MEHRHOFF

European Commission
Directorate-General Eurostat
Price statistics. Purchasing power parities. Housing statistics

BECH A2/038
5, Rue Alphonse Weicker
L-2721 Luxembourg
+352 4301-31405
Jens.MEHRHOFF@ec.europa.eu

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Data as a critical factor for central banks[1]

Maciej Piechocki and Anne Leslie-Bini,
BearingPoint / Central Banking

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

Data as a critical factor for central banks

Dr. Maciej Piechocki, Anne Leslie-Bini, BearingPoint

„ … it should also be clear to everyone that we are now standing only at the start of a long road in terms of data. The big challenge for statistics in the coming years is not only "many more numbers", but perhaps much more so, the reconciliation of statistical information collected in support of monetary policy and financial stability with the up-to-now rather separate world of supervisory information. It is one thing to have information, which, like blood, flows through the veins of the system, it is another to ensure that everything beats at the same rhythm and all organs in the body get all they need from the same single flow." (Mario Draghi President of the ECB, Seventh ECB Statistics Conference "Towards the banking Union. Opportunities and challenges for statistics", Frankfurt am Main, 15 October 2014")

Abstract of conclusions and survey results from the Central Banking Big Data Focus Report

Central banking statistical stability or supervisory function have been increasingly driven by (big) data, but little has changed in the methodology of supervisory data collection and management, which is still widely reliant on the document-oriented approach. This is intrinsically time-consuming, costly and complex. Data gaps still exist and so data remains a critical factor for central banks. Innovative solutions are necessary, to effectively handle “Big Data”.

The Central Banking Big Data Focus Report is a joint initiative of the Central Banking Journal and BearingPoint. The report builds upon the results of the recent IFC survey and takes a closer look at how central banks actually handle the challenge of data collection and analytics with regard to technical platforms and standards, resources and data governance.

The report investigates the concrete action plans of central banks regarding data management challenges in light of FinTech/RegTech developments and the objective of transparent and effective risk-based supervision but also plans for central banks statistics for “going beyond the aggregates” especially for the micro-granular data handling. Finally, central banks are evaluated how the BCBS 239 principle in an adapted version would apply to them today.

The focus report will draw on views from central bankers, industry experts, academics and observers to look at:

- Financial stability and supervisory applications
- Direct uses in economics and modelling
- Who should ‘own’ big data?
- Resourcing and budgets
- Future developments
- Operational challenges – gathering, structuring, storing and processing data

The Central Banking Big Data Focus Report aims at giving a clear picture of where central banks stand today with supervisory data management and defining fields of action.

Our part in the report sets out the results of a survey of how central banks view big data and data governance in their institutions. The survey was conducted by Central Banking Publications, in association with BearingPoint, in August and September 2016. The work has only been possible with the support and cooperation of the central bankers who agreed to take part. They did so on the condition that neither their names nor those of their central banks would be mentioned in the report.

Summary of key findings[1]

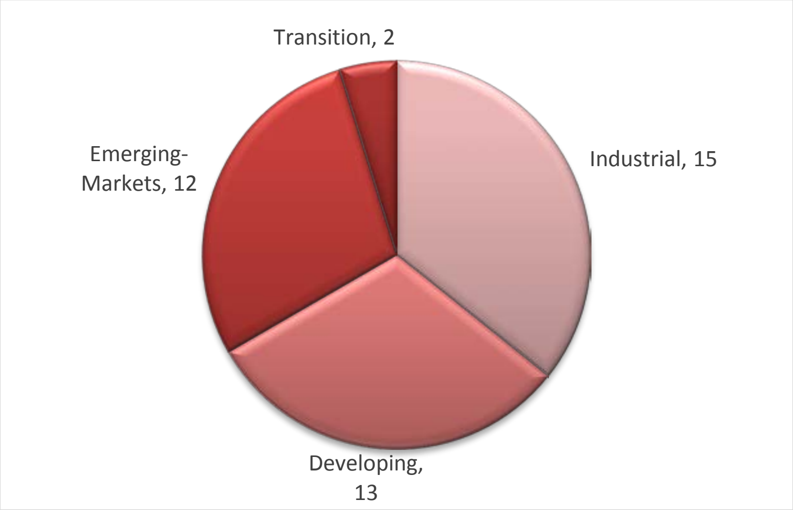
- Central banks have an active interest in big data. This is manifested in improving processing technology, adapting institutional strategies and increasing staff awareness of the area.
- Central banks typically see big data as unstructured data that is sourced externally, though this view is not universally held.
- Overwhelmingly, central banks develop their own data platforms to handle regulatory data collection, a role that has taken on greater significance since the financial crisis as central banks have expanded their involvement in financial stability.
- Big data is predominantly regarded as useful for research, but significant minorities see immediate involvement in policy-making, or scope for this.
- Lack of support from policy-makers is seen as the most significant challenge to increase use of big data.
- Central banks do not in the main have a dedicated budget for the handling of data (including big data), though many are seeking one.
- A little over 80% of respondents said they do not have any intra-departmental or divisional bodies dedicated to big data.
- More broadly, central bankers have concerns over the arrangements in place for managing data in their institutions. Many are looking to improve data governance.
- Over three-quarters of respondents indicated they had a shared internal platform, typically in the form of reporting frameworks and data warehouses.
- Central banks generally source their own big data sets, though a significant minority increasingly look elsewhere for these. Overwhelmingly, they process these themselves and there is no indication of a desire for this to change.
- Monetary policy is seen as standing to benefit most from big data, though it is expected to have a significant impact on macro-prudential policy as well.
- Support from the executive-level and policy-makers divided respondents: 35% saw it as top priority for investment within the central bank, 38% saw it as the lowest.
- Central banks have developed their own data platforms to deal with regulatory data analytics, often used in conjunction with other options. Excel remains popular.
- Central bankers broadly welcome the idea of self-assessment of data management using an adapted version of the Basel Committee’s BCBS 239 principles for supervisory data aggregation.

[1] Central Banking Publications, Big Data in Central Banks: 2016 survey results by Emma Glass

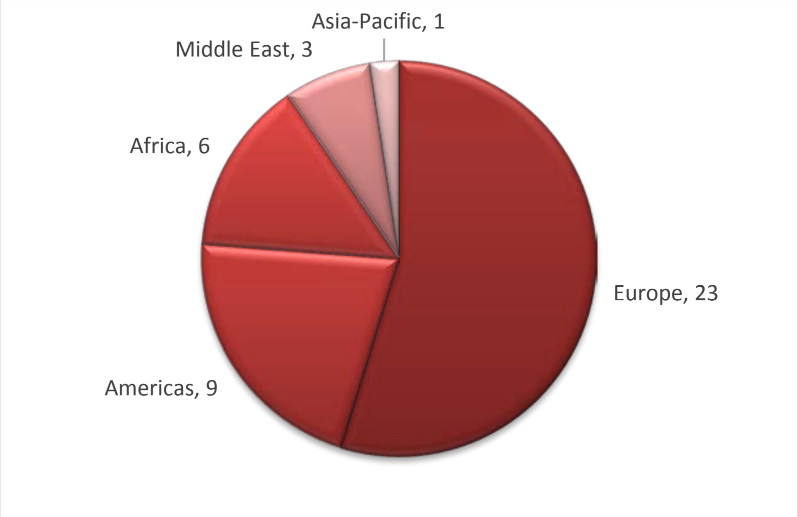
Article[2] content

Responses to the Central Banking Big Data Focus Report were received from 42 central banks al across the globe. The average staff size was 1,652 and three-quarters of respondents had less than 2,000 employees. Just over half of respondents were from central banks in Europe. Those individuals taking part in the survey were drawn in the main from the statistics function: 32, or three-quarters of respondents, were located in this area. Three were from information technology, and responses were also received from research, banking supervision, international relations, data management, infrastructure and technology, and data collection departments.

Economic classification

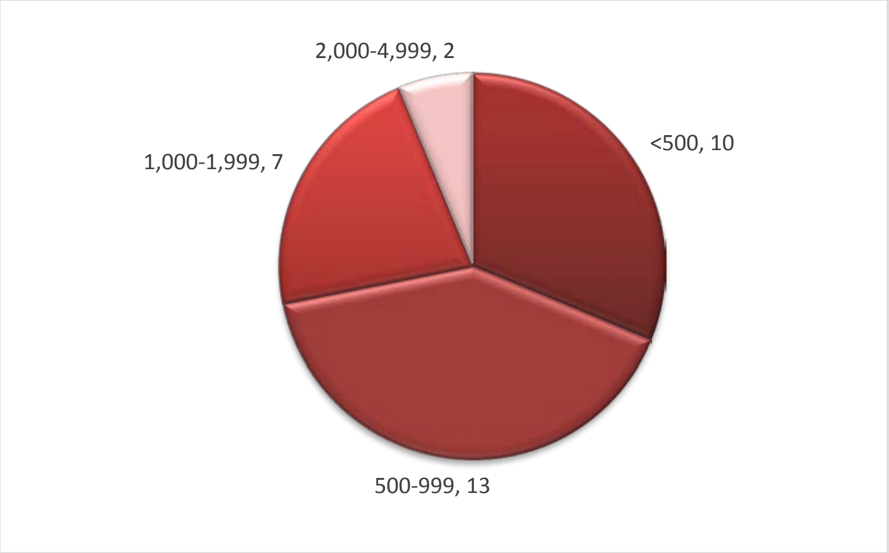


Geography

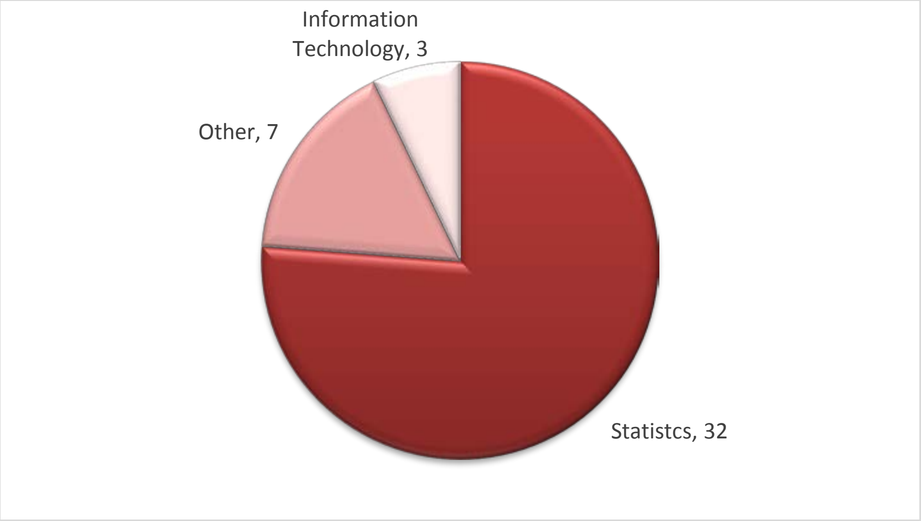


[2] This article is prepared based on the result of the survey conducted by the Central Banking in cooperation with the author and BearingPoint. Detailed results of the survey can be found in the Central Banking Big Data Focus Report 2016.

Staff size



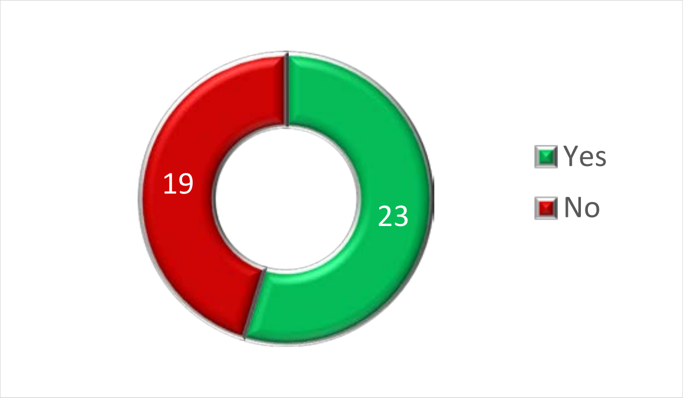
Department



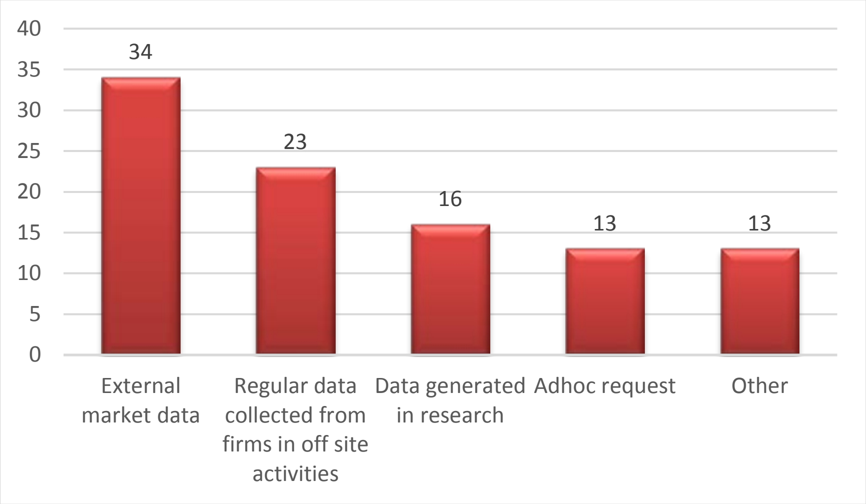
Central banks have an active interest in big data, particularly improving processing technology, adapting institutional strategies and increasing staff awareness of the area. Just over half of respondents said they had been giving the area active consideration in the past 12 months. This group of 23 central banks was drawn from around the world, though, reflecting the make-up of participants, European central banks figured prominently.

Although big data was not included in its strategic plan for 2016-2021, a respondent from the Caribbean said it would move up on the list of priorities in the future, and, as a result, they have introduced training initiatives.

Has your central bank given any active consideration to big data in the past 12 months?

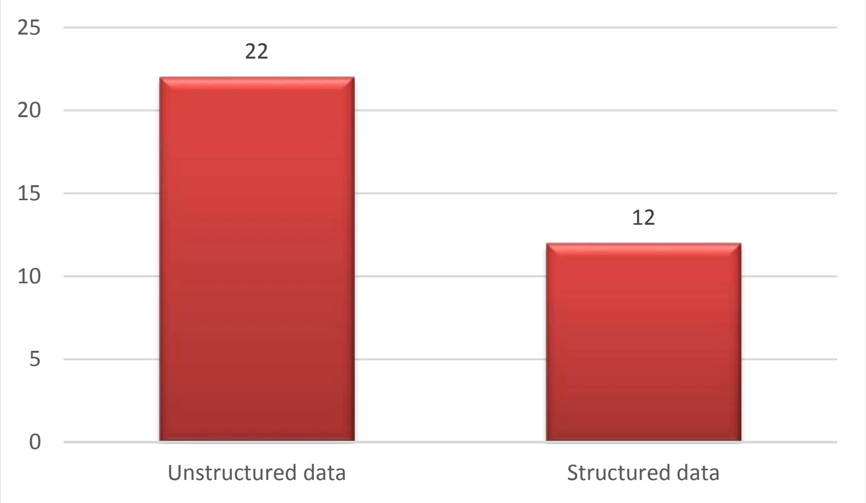


Which of the following types of data would you classify as big data?



All respondents took part in this question: most gave multiple answers.

Which statement, in your view, represents the most accurate definition of big data?

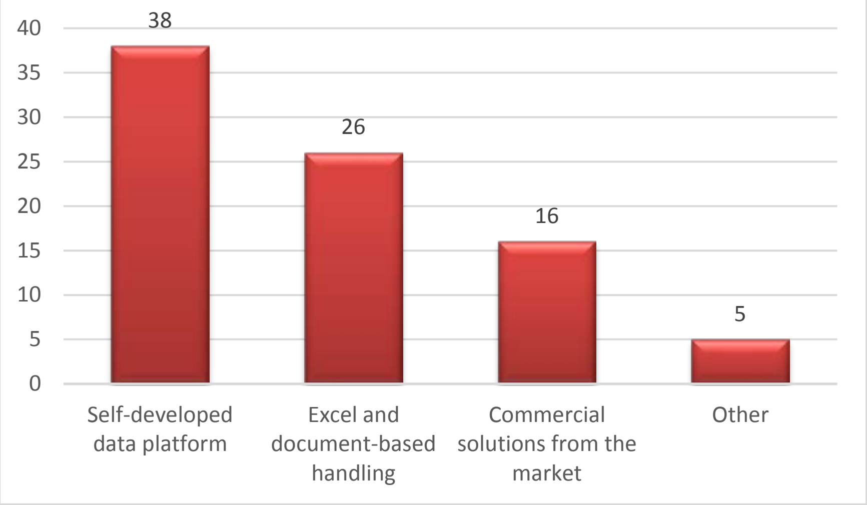


Forty-two central banks answered this question; eight of whom checked both "structured" and "unstructured".

While there is no single agreed definition of big data, industry has gravitated towards viewing it as large volumes of data, both structured and unstructured, which a standard desktop computer cannot handle.[3] Central banks typically see big data as unstructured data that is sourced externally, though this view was not universally held.

Just under two-thirds of respondents believe big data should be defined as unstructured data sets. In a comment, a European central banker said: "'Unstructured' seems to be a more accurate definition of big data because at least it requires unstructured data processing, which is more challenging."

How does your central bank deal with regulatory data collection?



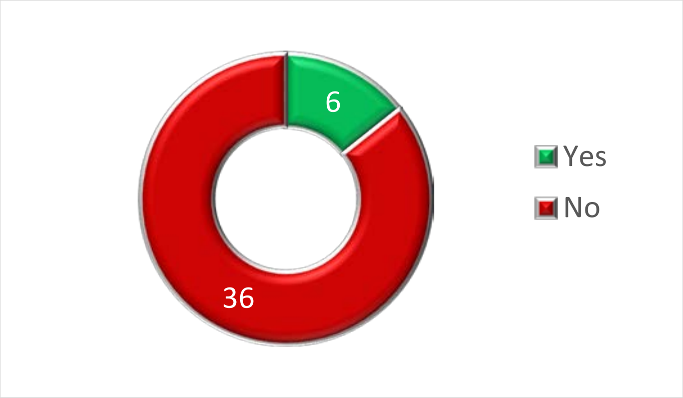
All respondents gave at least one answer.

As central banks have expanded the depth and breadth of their involvement in financial stability policy-making since the financial crisis, so the importance of collecting accurate, complete and timely data from institutions has increased. Overwhelmingly, central banks develop their own data platforms to handle regulatory data collection. This was the view of 38, or 90%, of the 42 respondents.

Commercial solutions from the market provide a viable alternative for central banks, and this option was chosen by nearly 40% of respondents. This was typically used in conjunction with self-developed platforms, as was the case for three-quarters of this group.

[3] For examples of different definitions, please see Bholat, D. (2015) "Big data and central banks" http://www.bankofengland.co.uk/publications/Documents/quarterlybulletin/2015/q108.pdf and Hinge, D. (2015) "The big data revolution and central banks" http://www.centralbanking.com/central-banking-journal/feature/2434258/the-big-data-revolution-and-central-banking.

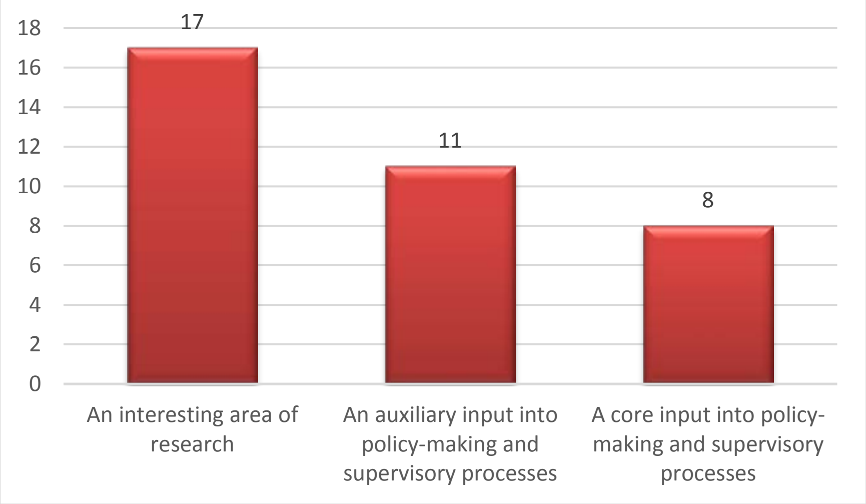
Has the approach to data collection changed in the past 12 months?



Arrangements for regulatory data collection display a high degree of continuity. Thirty-six respondents said their approach to regulatory data collection was unchanged in the past 12 months. Few respondents volunteered a comment explaining their view, though a handful indicated a transition is underway. A statistician from the Caribbean said their regulatory data collection approach is expected to change over the next six months. Similarly, a respondent from the Americas said: "We are in the process of creating better foundations for policy decisions and have therefore changed strategies for managing data and statistics." One European respondent noted they are reviewing future plans, while another said their central bank is implementing software for both the banking supervision and statistics departments, demonstrating the collaborative nature of big data projects.

Six central banks have changed how they deal with regulatory data collection in the past 12 months. This group was dominated by central banks based in Europe with one adding that it was necessary to change their data collection process "for technological and business reasons."

Which best represents your central bank's view of big data?



All respondents answered this question: six gave multiple answers.

Big data is predominantly regarded as useful for research in central banks, but significant minorities see immediate involvement in policy-making, or scope for this. Nearly 50% of respondents chose “an interesting area of research” as the best match for how their central bank views big data. A European respondent echoed several others, saying: “the central bank’s view of big data is going to change over time. Right now, it is an active area of research”. A respondent from an advanced economy implied big data would influence policy: “the optimisation of data is of key importance in driving decision-making.”

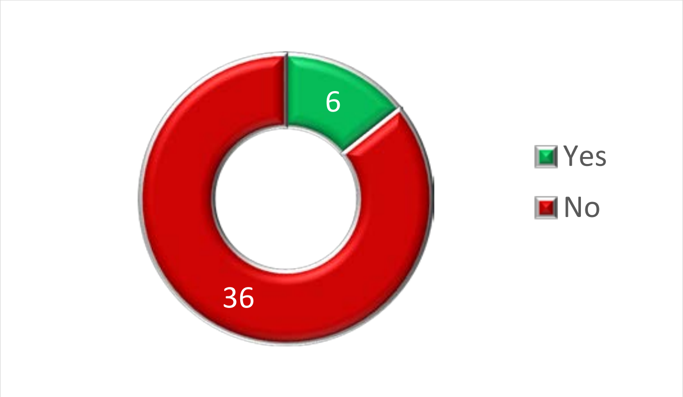
The second most popular choice was ‘an auxiliary input’. One European respondent said they could see the potential “in the future” of big data as an auxiliary input in policy and supervision. A central banker from a developing economy said its statistics department manages a handful of micro-databases that are useful for cross-checking data:

The use of microdata also brings flexibility to data management in a way that we can readily adjust and satisfy ad hoc requests, in some cases tailor-made to our customers’ needs. Finally, the establishment of protocols with other institutions gives us access to external and complementary information to our own sources, which is one of the primary keys to ensure data quality.

Eight central banks indicated big data was a core input into policy-making and supervisory processes. This included the largest central bank that participated in this survey, as well as two respondents from the Middle East. One commented:

An efficient data management is obviously needed in the course of all missions for which the central bank (including the supervisory authority) is responsible. Big data techniques are useful in that respect.

Has your central bank’s view changed in the past 12 months?



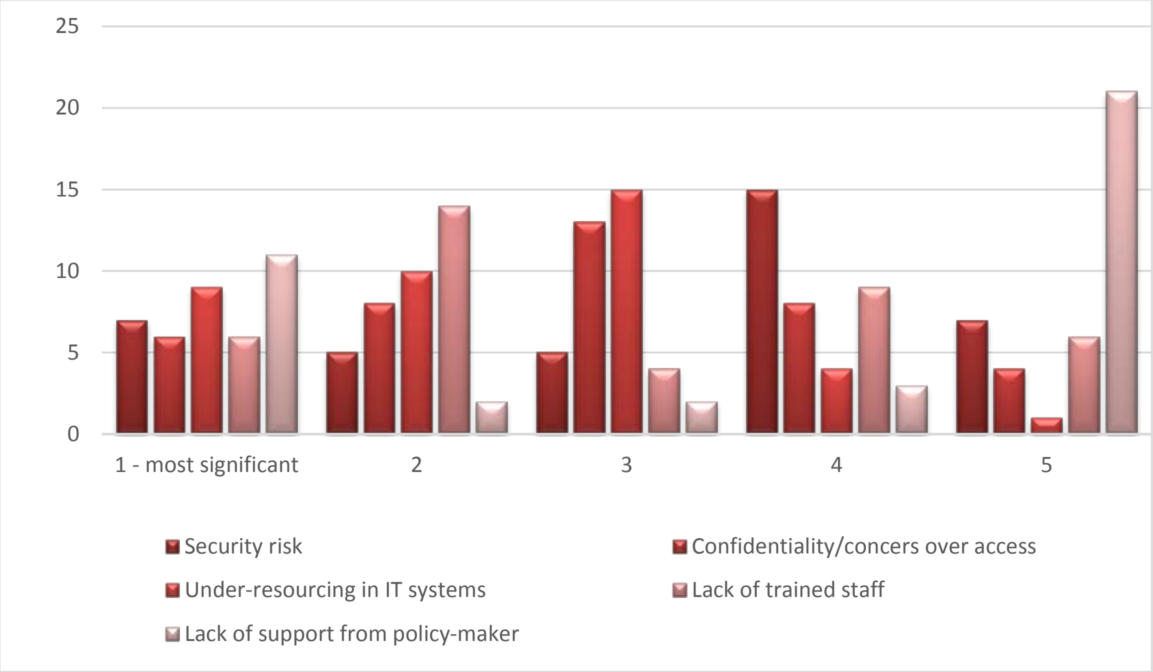
Views of big data’s role in a central bank are largely unchanged. Over 85% said their view has not changed in the past 12 months. “We expect this to change in the medium term,” observed a European central banker, who said they have dedicated a team to this work and therefore their view may change “as a consequence of our research conclusions”.

The six central banks that have made a change in the past 12 months highlighted how big data was having a catalytic effect on their institution. A statistician from Europe commented that data management is part of the central bank's new strategic plan. Another European central bank is advancing their work in this area, with dedicated teams within divisions outside of the statistics department:

There has been a change over the past 12-24 months which has seen the creation of dedicated analytics teams within supervisory divisions and a move towards creating a more data-centric organisation, which is reflected in organisational level objectives.

An officer from the Americas looked to the importance of an information strategy for the future development of big data: "To ensure that information gathered is handled in an appropriate manner, a vision for information supply is therefore required, along with an accompanying strategy which guides how data is required and processed."

Which in your view present the greatest roadblocks or challenges to increased use of data sets in your central bank? (Please rank the following 1-5 with 1 being the most significant.)

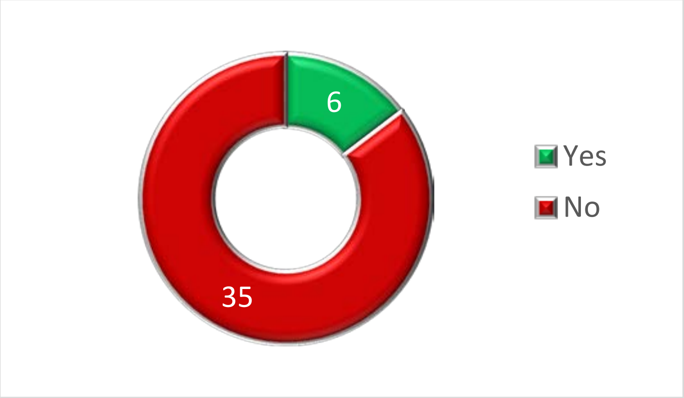


Three respondents did not reply.

Lack of support from policy-makers was most commonly identified as the greatest challenge and, intriguingly, also received the most votes as the least challenging. Twenty-eight per cent of respondents – seven Europeans, and four from the Americas – scored this as the most significant hindrance. "There are no problems with security, but there is not enough support," one European said. Conversely, 21, or 54% of, respondents saw a lack of support from policy-makers as the least significant challenge. This group was largely made up of developing and emerging market economies.

Concerns over skillsets figure prominently in central bankers' thinking. Just over half of respondents placed a lack of trained staff as the first or second most significant challenge. This group contained a significant number of central banks from advanced economies. A respondent from Europe noted: "Specific skills are needed for an efficient management of large data sets, both in IT and statistical departments." One respondent sounded a despondent note: "We expect additional human capital costs to be higher than acceptable compared to the possible benefits from using big data."

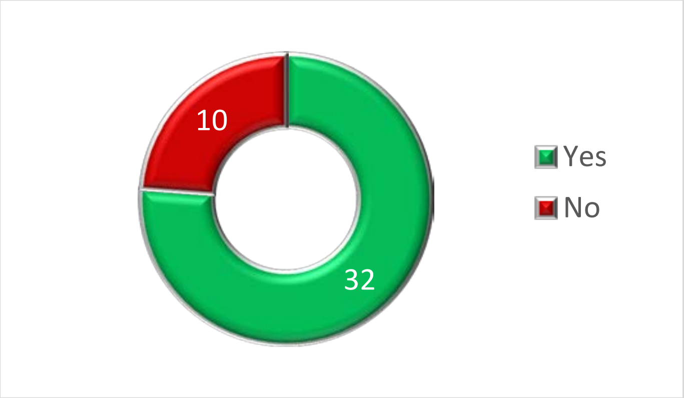
Does your central bank have a single allocated budget for the handling of data (including big data)?



One respondent did not reply.

The vast majority, 85% of respondents, said their central bank did not have a single allocated budget for data. This group was largely made up of central bankers from advanced and developing economies. In comments, respondents typically attributed this to budgeting being divided on departmental or project-specific bases rather than by resource, function or output. In this way, data was often included as part of the technology budget. One central banker commented: "several entities are involved in the handling of data."

Does your central bank have a shared internal platform to enable different areas of the central bank to access data resources?

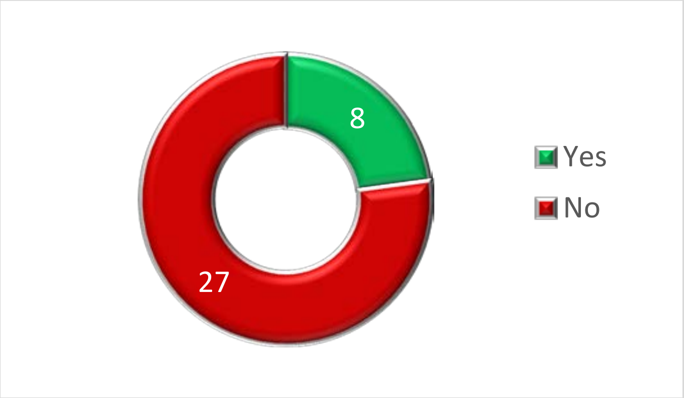


Over three-quarters of respondents indicated they had a shared internal platform, typically in the form of reporting frameworks and data warehouses. A respondent from a European central bank described the two-pronged approach used in their institution:

We have a shared internal platform for accessing most of our supervision data. For other macroeconomic data we have another platform contributed by the economic department and used by everyone inside the bank.

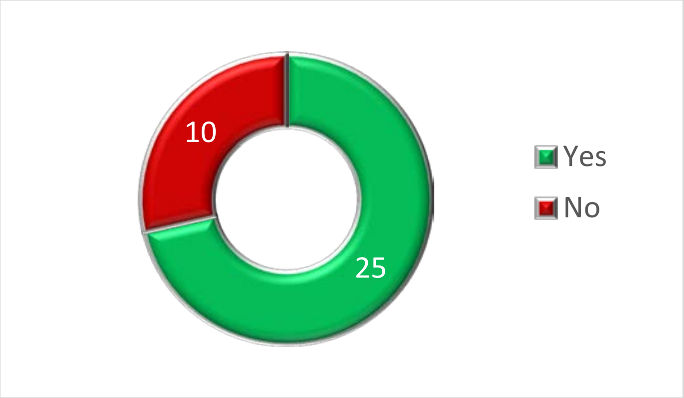
A respondent from a large central bank explained how they granted access to the data: "The first key element in the data strategy was to allow users with an appropriate business case to view all data relevant to them across the organisation. A technology solution was implemented to allow for this."

If yes, i) can this platform be accessed externally (ie for researchers or dissemination purposes?)



Seven respondents did not reply.

ii) was this shared platform built by the central bank?



Seven respondents did not reply.

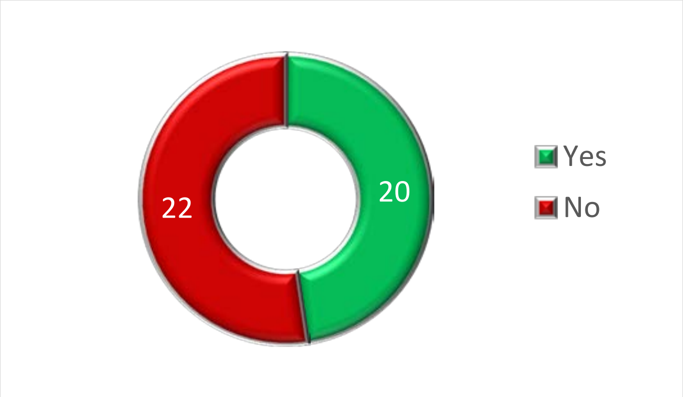
Data-sharing platforms in central banks are typically built internally by the central bank and are not available for external use.

The eight central banks that do allow external access to their self-developed platforms consisted of five European central banks, one from the Americas, one from the Middle East and one from Africa. A European central banker said: "the data software is externally accessed by banking supervisors when examining the financial institutions. This platform is restricted for researchers and dissemination purposes." Of the eight central banks that do allow external access, six built their own platform independently.

Several respondents from developed countries, however, said they were establishing platforms that would be available for external use. One central banker said "external access is still under development", while another commented that some parts of the database were being made available to registered users.

In the main, central banks build rather than buy shared platforms: this was the experience of just over 70% of respondents. One European central bank said they used "in-house developments with standard big data infrastructures". Ten central banks indicated they did not build their own platforms. Their reasons ranged from seeking assistance from another central bank in the development to subcontracting the process.

Does your central bank have clear data governance, with defined roles and responsibilities eg, chief data officer, data stewards?

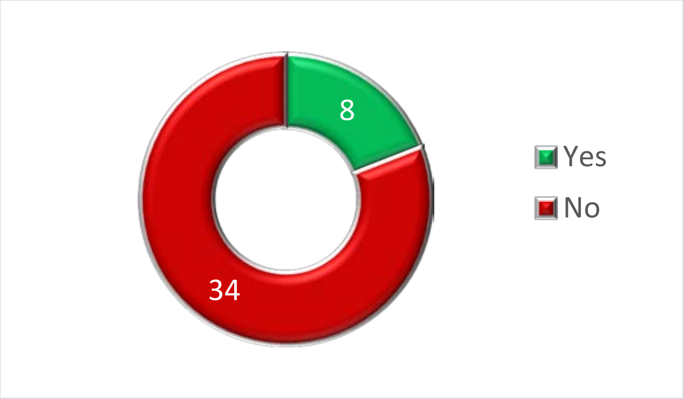


Many central bankers are concerned by data management arrangements, and just over half of respondents said their banks did not have clear data governance. European central banks featured prominently in this group, but there was also a significant number of respondents from the Americas. One said "we have no chief data officer or equivalent".

This is clearly an area of intense activity, however, as central banks are striving to improve data governance frameworks. A central bank from the Caribbean said the framework is "in its infancy" but is "expected to mature" in the coming years. An officer from a central bank in Africa commented a clear data governance structure "is now being put in place". Several central bankers from Europe said it is being discussed, however it had not been formalised, a view typified by this statistician's comment: "Certain relevant people are more aware of their roles and responsibilities as data owners or data stewards, but it has not been fully formalised and implemented yet." A developed country central banker noted: "We have a CDO and data stewards identified in place, however, more robust governance is being put in place over the coming months."

Twenty respondents were more confident in the arrangements for data governance. One European central banker commented they have "data owners" who can grant others the access to the data. A statistician from the Americas said: "for each information-specialised area, there is a work team." A central bank with over 5,000 staff has committed a whole team to the supervision of big data: "Through the project, we have built a specific organisation in charge of security policy and the high-level supervision. A dedicated committee chaired by the Director of General Statistics including high-level representatives of all business areas is responsible for the high level monitoring of the platform."

Does your central bank have any intra-departmental or divisional bodies, e.g. committees or working groups, dedicated to big data?

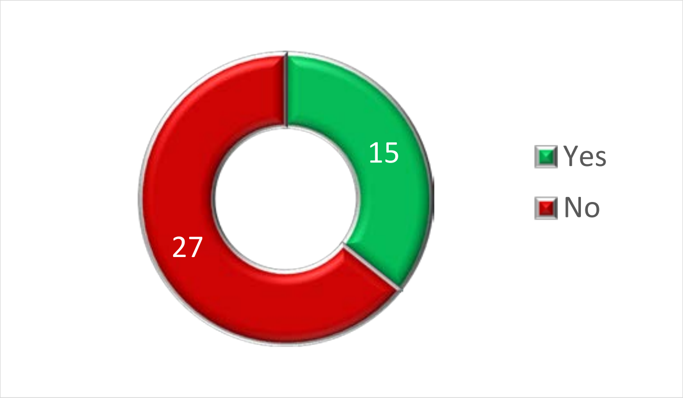


Big data is generally confined to departments or divisions in central banks. Just over 80% of respondents said they do not have any intra-departmental or divisional bodies dedicated to big data. This group featured five African central banks and all nine of the central banks from the Americas, along with two central banks from the Middle East.

Comments made by the 19% of central banks which have intra-departmental or divisional bodies centred around bank-wide data teams and committees dedicated to big data. One central bank commented: "This year we have started an inter-departmental team dedicated to big data issues", while another made reference to big data being discussed within a team of people focused on data in general. One central banker from an advanced economy said big data plays a part throughout the central bank, however it is governed by the statistics department director: "Many projects deal with big data issues: the corresponding steering committees are temporarily chaired by the Director of General Statistics."

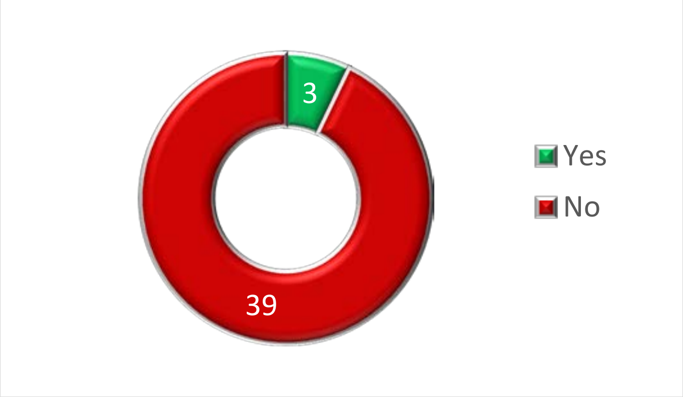
Although 34 central banks indicated they do not have an intra-departmental or divisional bodies at their central banks, three said it is a field they are looking to develop. One such central bank in the Americas commented: "We have some employees working on big data.

Does your central bank use external data providers for big data sets?



Central banks generally source their own big data sets, though a significant minority look elsewhere. Respondents that do not use external data providers were predominantly from emerging-market economies, including 17 central banks from Europe. Those that did use external providers, tended to turn to commercial banks and firms, social media and google blogs, and mobile phone operators. “We employ data from blogs, social media and private websites,” said one respondent. A statistician from Europe commented on the techniques used to collect the data from websites: “Currently, we are doing web scraping projects, collecting data from different websites.”

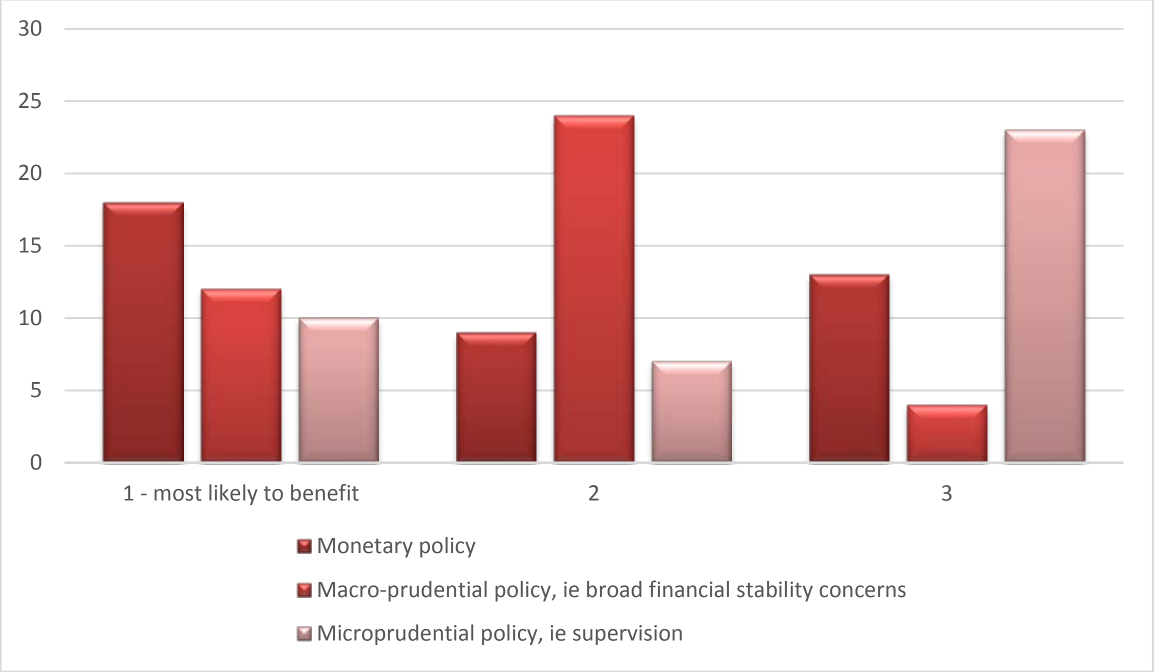
Does your central bank outsource any data processing for big data sets?



Overwhelmingly, central banks process their own big data sets and there is no indication of a desire for this to change. More than 90% of respondents answered they do not outsource any data processing. A central bank in the Americas said straightforwardly: “Data processing is run in-house”, while one central bank in Europe commented: “The data management and statistical analysis of big data sets is taken on by highly qualified staff in Director General of Statistics with the support of the dedicated team in IT department.” Three central banks outsource data processing for big data sets, however they all declined to comment.

Of the 39 central banks who indicated they did not outsource data processing, only three noted this was likely to change in the near future. One central bank in the Americas is currently hiring in this sector in order to cater for in-house data processing: “The central bank is now recruiting an expert in the field of data architecture in order to develop intra-departmental systems and working routines of data.”

In your view which of the following areas stands to benefit most from big data, in practical terms? (Please rank 1-3 with 1 being the most likely to benefit)

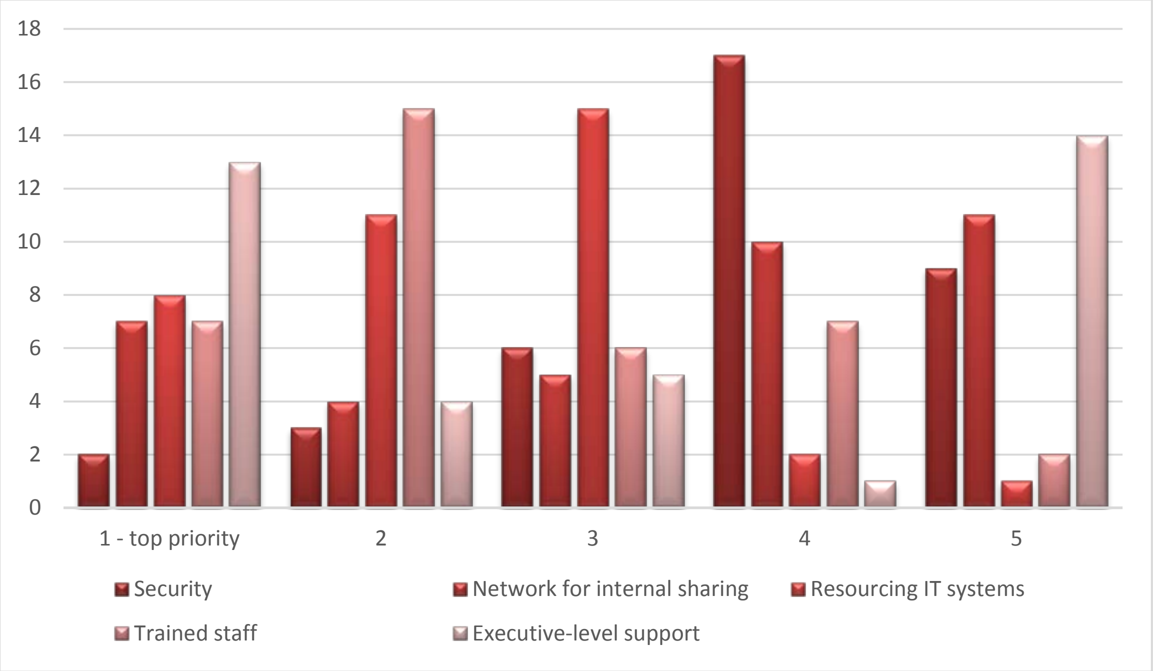


Two respondents did not reply.

Respondents saw monetary policy as likely to benefit most from big data, though they said it will have a significant impact on macro-prudential policy as well. As a central banker from an industrial economy explained: “Monetary policy, macro-prudential and micro-prudential policies can benefit from big data. Monetary policy can benefit from better and timelier nowcasts of macroeconomic variables. Macro- and micro-prudential policies might benefit as well.”

Interestingly, 90% ranked macro-prudential policy in either first or second place. Those that ranked it in first place typically chose monetary policy as their second choice. Micro-prudential policy was ranked in third place by 58% of respondents but 10 respondents did place it first place. One noted: “Big data applications shine at the most detailed level.”

Which area do you consider the priority for investment to increase big data use in your central bank? (Please rank 1-5 with 1 being top priority)



Five respondents did not reply.

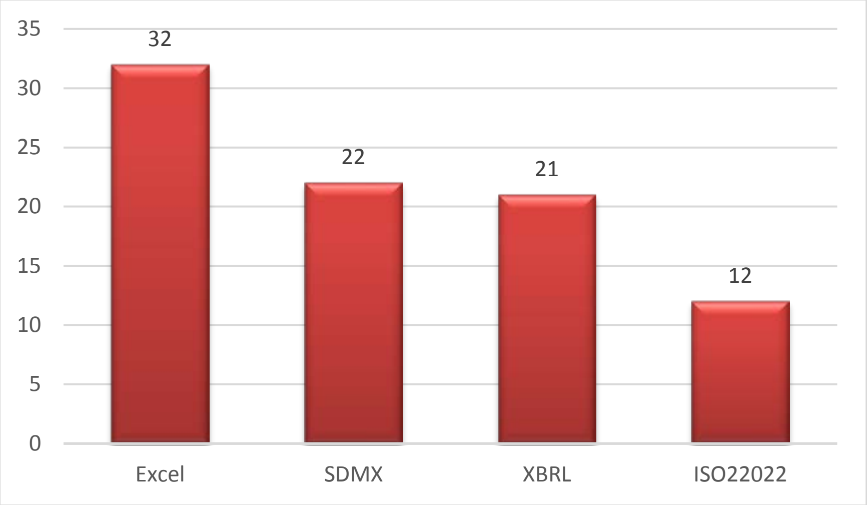
Support from the executive-level and policy-makers again divided respondents: 35% saw it as top priority, 38% saw it as the lowest. Those that saw it as a top priority were mainly central banks from Europe and the Americas. A European officer stressed the importance of having a budget: "Obtaining a budget is of fundamental importance in an initiative so therefore is a key priority."

Conversely, many saw executive-level support as a minor challenge. A statistician from a small central bank in the Americas explained their focus:

Big data also poses considerable challenges for the central bank, both technically and methodologically. In addition, new considerations need to be made in terms of strategy, organisation, skills, budgets and risks.

Sixty per cent of respondents ranked trained staff in either first or second place. This group included half a dozen emerging markets. However, these central banks declined to comment. Resourcing IT systems was most commonly ranked in third place. One European central bank commented: "Ensuring that IT infrastructure can handle additional demands is a key component of where this investment will be required." Network for internal sharing was largely ranked in fourth and fifth place by respondents, as was security.

Which of the following standards does your central bank use, or plan to use, for dealing with data exchange and collections?



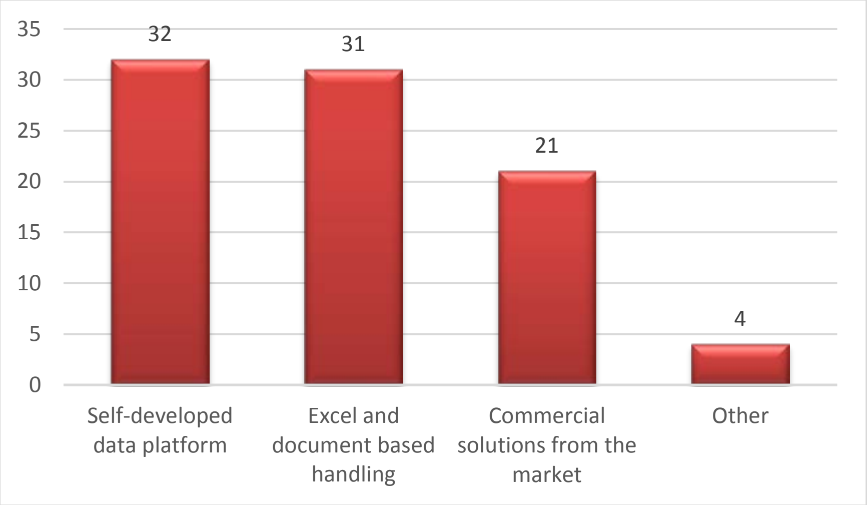
Three respondents did not reply. Respondents checked multiple answers.

Excel is the most popular standard for central banks when dealing with data exchange and collection but it is typically used in conjunction with another solution. Of the 39 respondents, 82% use Excel, and around 40% of those that use Excel use it exclusively. Of the remainder, most used either SDMX or XBRL as well, and three-quarters use both. "The central bank collects data from financial institutions using the XML format combined with Excel format", noted one statistician from a developing economy. An officer from a developed country commented: "XML with structured data prescribed by the central bank is also used for data collection and data exchange".

SDMX was the second most popular standard choice for the majority of developed, European respondents. One central banker commented: "SDMX is used for data exchange of statistical data, primarily in data exchange application with ECB." The statistics department of an African central bank is working towards implementing this standard: "SDMX is an ongoing project."

The standards of XBRL and ISO20022 are used typically for specific functions: supervision and payments, respectively. Twenty-one respondents indicated that they use SDMX at their central bank. A central bank from an advanced economy noted: "XBRL is only used in the transactional system component of RIAD application (Register of Institutions and Affiliates Database)." Although the least popular answer, ISO20022 was likewise described by respondents as useful for specific functions. One European officer noted: "Currently we are using the three standards above-mentioned: SDMX and XBRL are widely used, whereas ISO20022 is only being used in very specific exchanges". An African central bank commented it is under consideration: "ISO20022 is being considered for payments."

How does your central bank deal with regulatory data analytics?

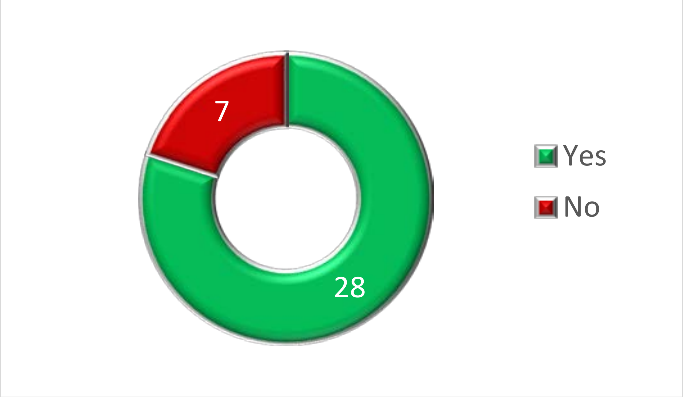


One respondent did not reply.

Central banks have developed their own data platforms to deal with regulatory data analytics, but this is often used in conjunction with Excel and document-based handling. This was the answer from over three-quarters of respondents, the majority of which were from developed countries. Only four of the 32 above use purely self-developed regulatory data analytics while the remaining 28 combine it with another standard. Nearly all combined this with Excel and document-based handling and just under half reported a commercial element to their regulatory data analytics workflow.

Commercial solutions from the market proved popular among respondents, with over half choosing this option. Interestingly, more central banks use commercial solutions from the market for regulatory data analytics than they do for regulatory data collection, with only 16 central banks.

As of January 2016 Globally Systemically Important Banks (G-SIBS) have been regulated to meet the BCBS 239 Principles, after carrying out self-assessments, in respect of data collection, data aggregation and dissemination capabilities. Do you think it would be useful for central banks to self-assess using an adapted version of these principles?



Seven respondents did not reply.

Central bankers largely welcome the idea of self-assessment using an adapted version of the BCBS 239 Principles for Globally Systemically Important Banks (G-SIBs).[4] Of the 35 respondents, 80% said such self-assessment would be useful.

As databases increase in size and complexity, there is naturally a concern that they are policed and governed properly. Furthermore, one central banker noted existing systems were under pressure: "Due to increasingly large and complex data that is now challenging traditional database systems."

Several central banks drew attention to the internal processes already in place to regulate big data processing. One European respondent from an advanced economy noted: "There is a benefit that can be accrued from this. It is important to note however that our internal audit function already undertakes full audits in this area." Another European said BCBS 239 was handled by the supervision department:

However, if the objective of these principles is to strengthen banks' risk data aggregation capabilities and internal risk reporting practices, it might be relevant to make a similar self-assessment in central banks.

Conversely, seven central banks disagreed it would be useful for central banks to self-assess based on an adapted set of principles. All were European except one from the Middle East. The BCBS 239 Principles are not applicable to data collection, a European central banker noted:

BCBS 239 addresses issues pertaining to risk data aggregation and reporting capabilities for banks, so as to achieve full compliance with regulatory expectations. The principles thereof focus on a specific use of data and do not take into account other dimensions of data valuation (e.g. statistical analysis for other needs than those underlying the collection of these data).

Summary

Increasing requirements to central banks' data management due to regulatory trends and technological innovations

Data stay a critical factor for central banks. The financial crisis showed that some of the deepest fissures were caused by gaps in data and exposed the need for high quality, comparable and timely data on the global financial network. Since then, policymakers, supervisory authorities and standard-setters across the globe have been collaborating to greater harmonize and standardize regulatory data in financial services. According to a

[4] The BCBS 239 Principles were created by the Basel Committee on Banking Supervision and put into place in January 2016. They consist of 14 principles for supervisors to follow when considering data aggregation in the banks they supervise.

recent BearingPoint Institute paper, urgent debate is still needed on how the world's financial services industry could be better and less onerously supervised via a smarter approach to regulatory reporting and data exchange. [5]

Financial supervision and central banks momentary statistics and financial stability functions are vastly driven by data. In the aftermath of the financial crisis, a "regulatory tsunami" flooded the financial services industry. Especially after the adoption of the Basel III framework, regulatory requirements have significantly increased. New regulations such as AnaCredit, BCBS 239, Solvency 2, Dodd Frank or IFRS 9 have posed new challenges to the banking and insurance sector on global, regional and local levels. Moreover, regulations like the EMIR (European Market Infrastructure Regulation), Money Market Statistical Reporting (MMSR), the Markets in Financial Instruments Regulation (MiFIR) and the Securities Financing Transaction Regulation (SFTR) oblige the major Monetary Financial Institutions (MFIs) to report derivatives or money market data on a daily basis.

"Big data" is a common buzzword in this context. Due to new media and technologies, new data sources appeared like e.g. Internet-based data, data from Social Media, but also from official sources and internal public databases such as banking supervisory data[6]. According to a BearingPoint Institute article on big data, the amount of information available in the world increased by a factor of 100 in last 20 years.

However, in the central banking area, while no single agreed definition exists, big data has already been heralded as offering a wide range of central banking applications: from nowcasting to modelling, to early warning systems and systemic risk indicators. For some it opens a new chapter in policymaking. In a recent study, the Institute of International Finance (IIF) stated that "'RegTech', defined as 'the use of new technologies to solve regulatory and compliance requirements more effectively and efficiently' has enormous potential to … improve the quality and efficiency of supervision, and reduce risk in the system."[7]

According to the 2015 IFC report on "Central banks' use of and interest in 'big data'" central banks have a strong interest in big data, but their actual involvement is still limited.[8]

BearingPoint is noticing two significant key trends worth looking at when discussing (big) data management in central banking in respect to financial services: the replacing of form-based collections with granular, micro-level data[9] and the need to go beyond reporting data validation, i.e. to integrate regulatory Key Performance Indicators (KPIs) into the overall

[5] "Reforming Regulatory Reporting. From Templates to Cubes.", Dr. Maciej Piechocki, Tim Dabringhausen

[6] IFC report „Central banks' use of and interest in „big data", October 2015, p. 19

[7] Institute of International Finance, "RegTech in Financial Services: Technology Solutions for Compliance and Reporting.", March 2016, p. 3f

[8] IFC report „Central banks' use of and interest in „big data", October 2015, p. 1

[9] IFC Working Paper No. 14, "Big data: The hunt for timely insights and decision certainty", February 2016, p. 15

operational supervisory process. However a number of further developments at central banks are observable. For example from governance perspectives central banks recently started to appoint "chief data officers" and implement harmonised "data strategies". Numbers of central banks are currently rethinking their data infrastructures which today are rather siloed and demonstrating the legacy of the past decades with no central approach to data handling.

Challenges for central banks

Notwithstanding the huge potential big data provides, decision making is now even harder than before, and business need adequate solutions to analyse this data.[10] A crucial point is how to mine all this information from the different sources exhaustively and at reasonable cost. Despite innovative tools and technologies like blockchain, cloud computing and machine-learning, even today plans often fail because the required processing power outweighs the potential returns or computing time is too long.[11]

The specific challenge for central banks in the sense of an effective 360° risk-based supervision is to rapidly access, effectively manage and timely process and analyse the increasing amounts of supervisory, statistical and markets (big) data. Especially the near or real-time access and efficient processing are regarded as critical factors due to limitations in human and IT resources.[12] According to the IIF report, some regulators still use outdated portal solutions and methods, which are inefficient and increase chances of introducing error.[13] The IIF recommends automated secure data transfer mechanisms based on standards like XBRL (eXtensible Business Reporting Language). But even with use of standard such as XBRL or SMDX (Statistical Data and Metadata eXchange) central banks must abandon "paper-" or "document-oriented" world and think of data in integrated and interlinked manner.

Current systems do not meet today's requirements when regulators have to deal with large amounts of data of various kinds - collected from supervised entities for statistical, prudential or stability purposes, provided by information providers or obtained from internal research and analysis. Such data span from granular micro information on single mortgage loans, securities traded and counterparties affected to macro-economic analysis of countries or regions to form-based collections of financial and risk data or ad-hoc supervisory exercises.

Some of this data will remain only in the perimeter of the central bank some will be remitted to other stakeholders such as the European Supervisory Authorities (ESAs), country

[10] BearingPoint Institute Issue 002, "Seeing beyond the big (data) picture, p.3-4
[11] Ibid., p. 6
[12] IFC report „Central banks' use of and interest in „big data", October 2015, p. 11
[13] Institute of International Finance, "RegTech in Financial Services: Technology Solutions for Compliance and Reporting.", March 2016, p. 22-23

governments, the International Monetary Fund (IMF) or the Bank for International Settlements (BIS), some will be disseminated to the wider public or research community.

Therefore, it is mission-critical for regulators to

- effectively handle the large amounts of increasingly granular data from various sources, i.e. rethink existing IT system architectures and landscapes
- gain transparency on the status of the reporting agents in the collection and dissemination process
- consider interlinkages between micro and macro data sets in “going beyond the aggregates” from macro and financial stability perspectives
- get a timely overview of relevant micro and macro developments in the financial markets and
- execute reliable trend analyses on KPIs and Key Risk Indicators (KRIs) based on validated collected data

Essentially, it is a question of scalability in various dimensions across the usual value chain or “lifecycle” of processing supervisory data, as investigated in detail in an article published in Banque de France’s Financial Stability Review.[14]

The expanding requirements have proved to be a great challenge and cost driver for IT departments of regulators. IT infrastructure and processes have to be optimised in order to collect, process, analyse and disseminate supervisory and statistical data from different sources and in various formats. But process automation and innovative solutions are required to increase quality and efficiency of supervision, to reduce expenditures, operational burdens and time to market for new supervisory requirements.

[14] Mark d. Flood, H.V. Jagadish, Louiqa Raschid: “Big data challenges and opportunities in financial stability monitoring”, Banque de France Financial Stability Review, No. 20, April 2016.

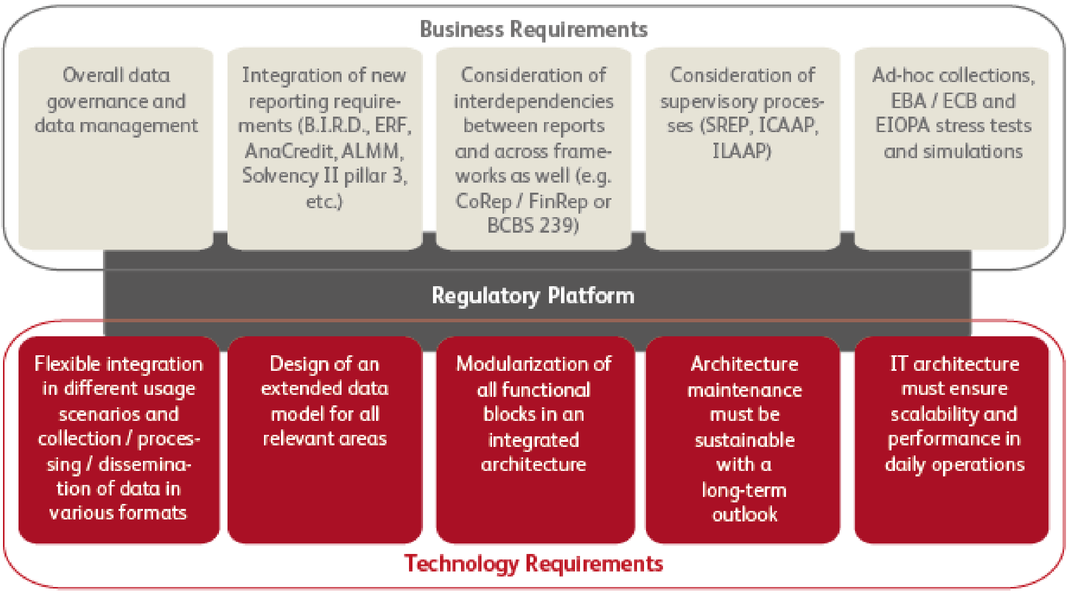

Business Requirements
Overall data governance and data management
Integration of new reporting requirements (B.I.R.D., ERF, AnaCredit, ALMM, Solvency II pillar 3, etc.)
Consideration of interdependencies between reports and across frameworks as well (e.g. CoRep / FinRep or BCBS 239)
Consideration of supervisory processes (SREP, ICAAP, ILAAP)
Ad-hoc collections, EBA / ECB and EIOPA stress tests and simulations
Regulatory Platform
Flexible integration in different usage scenarios and collection / processing / dissemination of data in various formats
Design of an extended data model for all relevant areas
Modularization of all functional blocks in an integrated architecture
Architecture maintenance must be sustainable with a long-term outlook
IT architecture must ensure scalability and performance in daily operations
Technology Requirements

FIGURE 1: REQUIREMENTS FOR FUTURE-ORIENTED REGULATORY PLATFORMS

Innovative approaches – shared utilities, integrated platforms for data management and analytics and Regulatory-as-a-Service (RaaS)

In view of the developments as described before, it is undisputable that it is mission-critical for central banks to reshape their data management and further automate industrialise processes of handling data. Automation helps to minimize risk, reduce errors, and increase transparency and thereby to deliver a better basis for decision-making.

According to a BearingPoint Institute article[15], a new information value chain is needed for reporting which helps to increase efficiency of supervisory processes, minimize risk, allocate resources effectively and improve the basis for decision-making by higher transparency and faster availability of data. We further notice a trend to shared utilities, Regulatory-as-a-Service.

A prominent example is the Austrian solution, where the national central bank, Oesterreichische Nationalbank (OeNB) and the supervised banks joined forces to stepwise replace the template-driven model and use innovative technologies to create a new regulatory value chain. The initiative is based on greater harmonization and integration of data within banks as well as greater integration of the IT systems of the supervisory authority and the supervised entities. The way it works is through a common data model (GMP) developed by central bank in cooperation with Austrian bank and a shared utility, called Austrian Reporting Services GmbH (AuRep), which is co-owned by the largest Austrian

[15] BearingPoint Institute, "Reforming regulatory reporting: are we headed toward real-time?", 2015

banking groups. This model allows cost-sharing of compliance as well as standardization of data collection.

AuRep runs on a common platform, which works as the central interface between the banks and the OeNB. Granular bank data sets are captured automatically for supervisors to interrogate in whichever way they want, whilst the banks retain control over their commercially sensitive data, maintaining only the so-called ‘passive data interface’ on the AuRep platform.

Other regulators are also aware of the limits of the template-based reporting and see the benefits of an input approach with granular datasets. While the Banca d’Italia has been providing such a shared data model named PUMA2 for some decades recently the European Central Bank (ECB) has also launched an initiative to evaluate a European “input approach”. The Expert Group on Statistical and Banking Data Dictionary was established to develop a Banks' Integrated Reporting Dictionary (BIRD), which defines a harmonized model for input data as well as rules for the transformation of input data to reporting data. BIRD should be seen as a blueprint for the banks. It forms the conceptual basis of an input approach, i.e., a data model for the organization of the regulatory reporting process within the banks. The approach is similar to the Italian and Austrian model.

Besides harmonized data definitions, new and high-performing supervisory data management platforms are necessary allowing for timely and efficient collection, analysis and sharing of the data.

These platforms could be deployed for instance as a closed solution for the regulator, as an open solution also for firms providing them advanced portal functionality as a service (RaaS or Regulatory-as-a-Service) or even as a shared services platform like in the Austrian case.

With regards to functional scope, new generation platforms should provide functionality for highly automated processing of data and regulatory business intelligence including statistical analysis, monitoring and controlling supervisory Key Risk Indicators (KRIs).

Visions

In the RegTech (Regulatory Technology) developments of recent years, one radical innovation in particular cannot go unnoticed – the Blockchain, and the Distributed Ledger Technology (DLT) that underlies it. Blockchain is rattling conventional finance by transforming business models, connecting new counterparties and eliminating friction. Blockchain offers huge potential cost and time savings in securities and derivatives transactions, especially in clearing, but the fact is that significant work needs to be done before the full benefits of the underlying technology are realised.

For central banks, Blockchain will almost certainly play a role in market infrastructure, although there is not yet a consensus on whether Blockchain is a help or a hindrance to financial market integration, due to the currently haphazard uptake of numerous different Blockchain technologies. The European Central Bank (ECB), for example, is in the process of assessing the relevance of Blockchain applications for the purposes of improving the region's securities and payments settlement system. However, the ECB has spoken out quite harshly with regard to digital currencies which are underpinned by Blockchain technology, describing them as inherently unstable and presenting a potentially negative impact on the ECB's ability to conduct monetary policy.

At the other end of the spectrum, the Bank of England (a front-runner in this area) has estimated that it could increase GDP by 3% if it introduced a Central Bank Digital Currency (CBDC), as a result of the reduced interest rates, reduced tax rates and lower transaction costs that implementing a digital currency would bring. Such a digital currency is touted as having the potential contribute to the stabilization of the wider economy, as it would give central banks another lever with which to control their currency. This would be particularly effective in times of economic shock, such as Brexit.

In addition to these already significant benefits, a Central Bank Digital Currency could also inject some much needed transparency into the financial system. During the 2008 financial crisis, it became painfully apparent how little visibility was available into the counterparty credit exposure of one major financial institution with regard to others. Eight years down the line, and thousands of pages of regulation later, this requisite level of visibility is still sorely lacking. Blockchain could remediate this opacity, as well as offering protections for privacy.

Following this line of thinking, further visionary uses of Distributed Ledger Technology (DLT) include the delivery of data to regulators, by leveraging so-called Smart Contracts that self-execute in a predefined manner once certain conditions are met. By using Smart Contracts on the Blockchain with specially designed algorithms, stakeholders in the financial system could have the ability to store regulatory information in a secure and immutable form on the shared ledger, and then share it with authorised parties when a specific event occurs. Given the growing multiplicity of data flows, with the ensuing challenges of timeliness, consistency and confidentiality, such an approach could deliver significant value in transaction-based regulatory reporting regimes such as EMIR or MIFID II.

Blockchain and distributed ledger technology (DLT) has a powerful enabling capability that may allow it to become the omnipresent 'soft infrastructure' that strengthens the efficiency, resiliency and accessibility of systems which facilitate global monetary and financial transactions. However, as a technology that is still in its infancy, there are substantial functional, operational, governance and legal aspects which need to be carefully examined before thinking about possible mass adoption. Blockchain and DLT certainly hold great potential, but these fundamental issues must be resolved before the benefits can be fully realized.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Data as a critical factor for central banks[1]

Maciej Piechocki,

BearingPoint / Central Banking

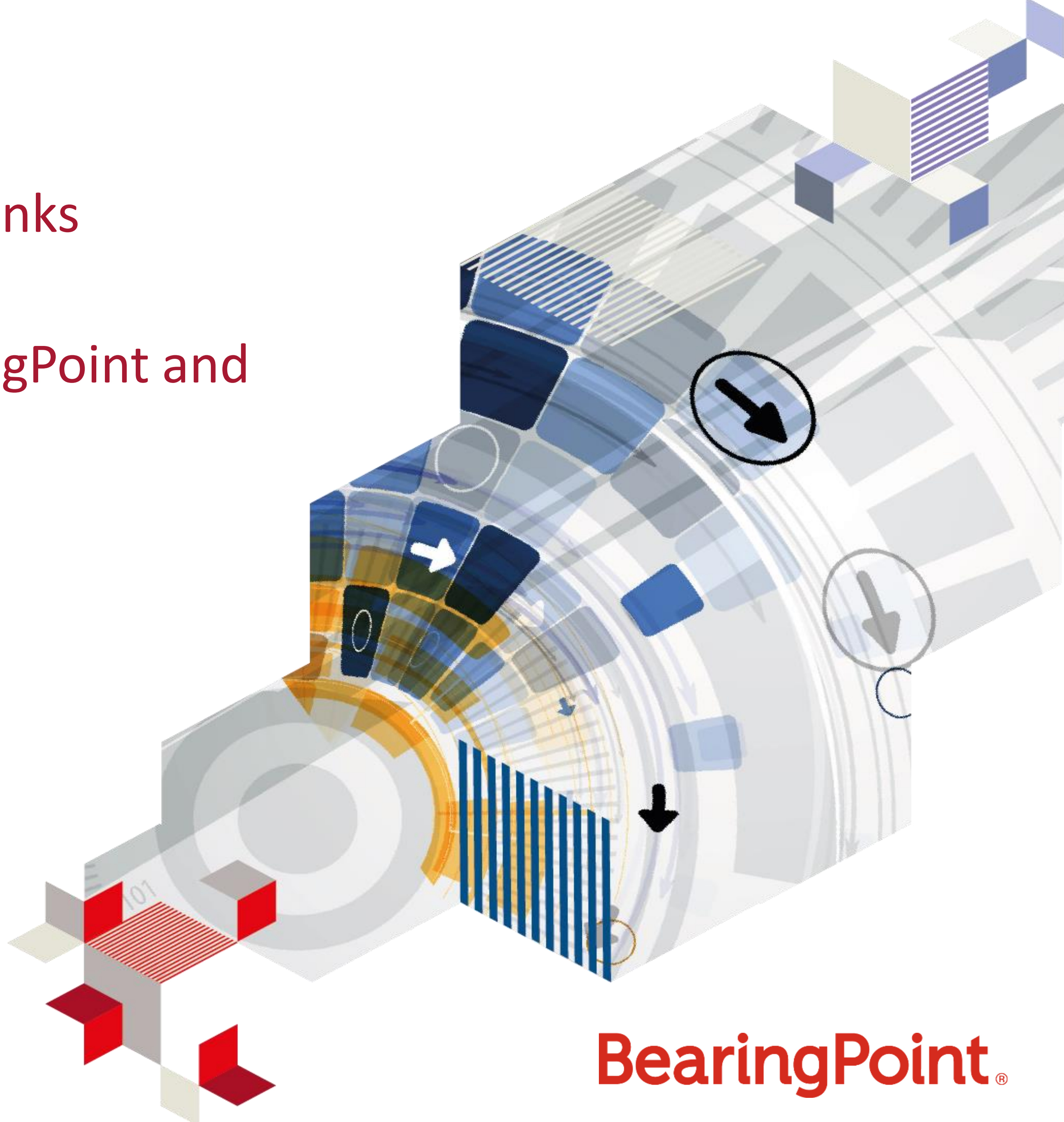
[1] This presentation was prepared for the meeting. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Big Data in Central Banks

Joint survey of BearingPoint and Central Banking

Dr. Maciej Piechocki

Bali, March 21st, 2017

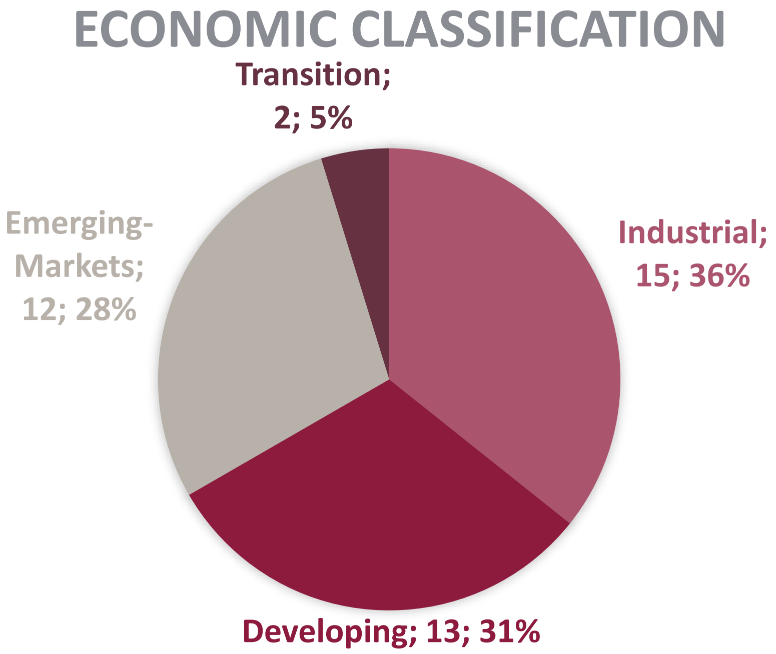


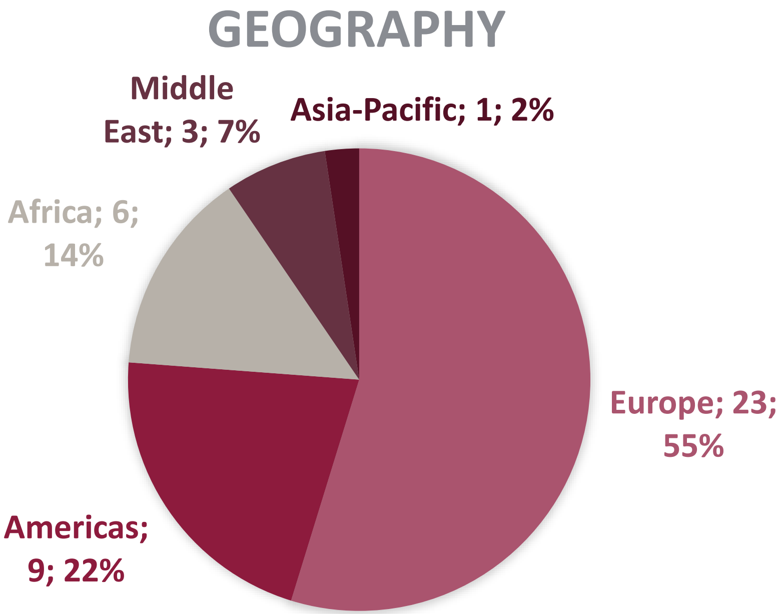
CENTRAL BANKING AWARDS 2017
CONSULTANCY AND ADVISORY SERVICES PROVIDER OF THE YEAR (CORE)

CENTRAL BANKING AWARDS 2016
TECHNOLOGY PROVIDER OF THE YEAR

BearingPoint®

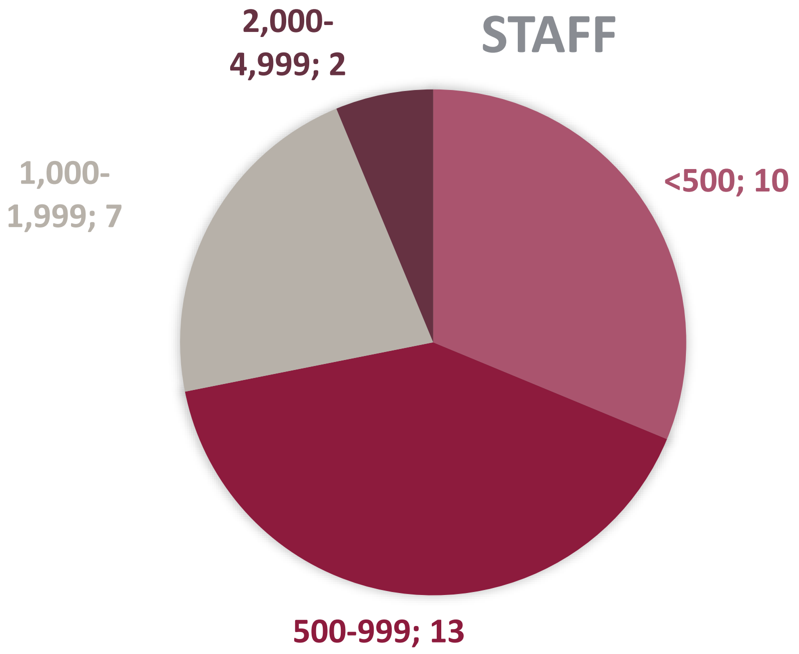
Economic classification and geographical structure of central banking respondents

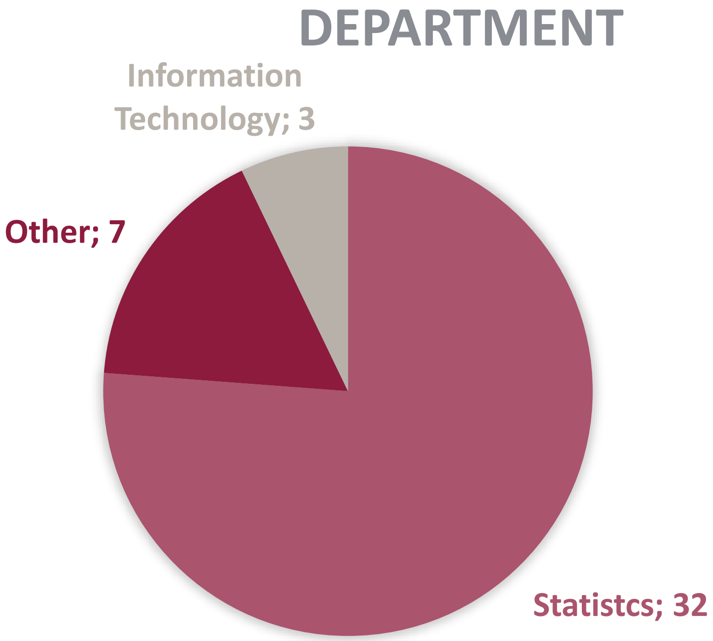




BearingPoint.

Staff size and departmental structure of central banking respondents





BearingPoint.

23 out of 42

central banks have an active interest in big data, particularly **improving processing technology, adapting institutional strategies and increasing staff awareness**

More than

80% of respondents said an example of big data was ‘external market data’

and over **55%** of respondents referred to ‘regular data collected from firms in off-site activities’

Big data is predominantly **regarded as useful for research** in central banks

Monetary policy

likely to benefit most from big data

but also significant impact on macro-prudential policy

Over 85%

said their view of big data’s role in a central bank has not changed in the past 12 months

“We expect this to change in the medium term as a consequence of our research conclusions” *

*a European central banker

Lack of **support from policy-makers** was most commonly identified as the greatest challenge followed closely by lack of **trained staff**

The vast majority,

85%

of respondents, said their central bank did not have a single allocated budget for data

Over 3/4

indicated they had a shared internal platform, typically in the form of reporting frameworks and data warehouses

“We have a shared internal platform for accessing most of our supervision data. For other macroeconomic data we have another platform contributed by the economic department and used by everyone inside the bank.” *

*a European central banker

Data-sharing platforms in central banks are typically built

internally

by the central bank and are **not** **available for**

external use

Many central bankers are concerned by **data management arrangements**, and just over half of respondents said their banks have **no clear data governance**

Just over 80%

of respondents said they do not have any intra-departmental or divisional bodies dedicated to big data

> ***"This year we have started an inter-departmental team dedicated to big data issues." ****

*a central banker

Top priority for investment to increase big data use in central bank needs support from the **executive-level and policy-makers***

> ***"Obtaining a budget is of fundamental importance in an initiative so therefore is a key priority." ****

*mainly central banks from Europe and the Americas

**European officer

38

central banks develop their own data platforms to handle regulatory data collection

For 40% commercial solutions from the market provide a viable alternative

Excel is the most popular standard for

82%

of central banks when dealing with data exchange and collection but it is typically used in conjunction with another solution

75% of the

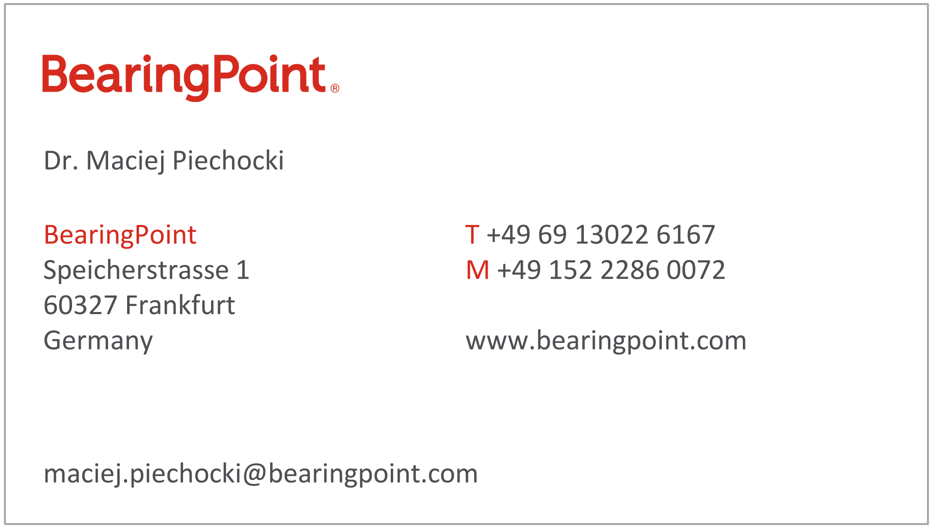
central banks have developed their **own data platforms** to deal with regulatory data analytics, but this is often used in conjunction with Excel and document-based handling

80% **of the central bankers largely welcome the idea of self-assessment using an adapted version of the BCBS 239 Principles for Globally Systemically Important Banks (G-SIBs)**

“Due to increasingly large and complex data that is now challenging traditional database systems.”*

*a central banker

Business card



BearingPoint®

Dr. Maciej Piechocki

BearingPoint
Speicherstrasse 1
60327 Frankfurt
Germany

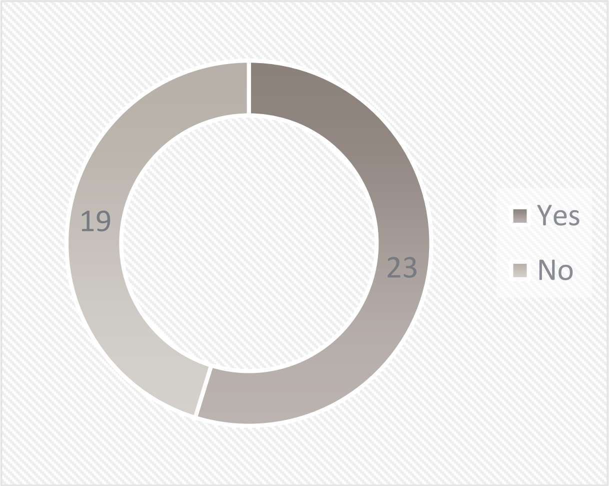
T +49 69 13022 6167
M +49 152 2286 0072

www.bearingpoint.com

maciej.piechocki@bearingpoint.com

BearingPoint®

Has your central bank given any active consideration to big data in the past 12 months?

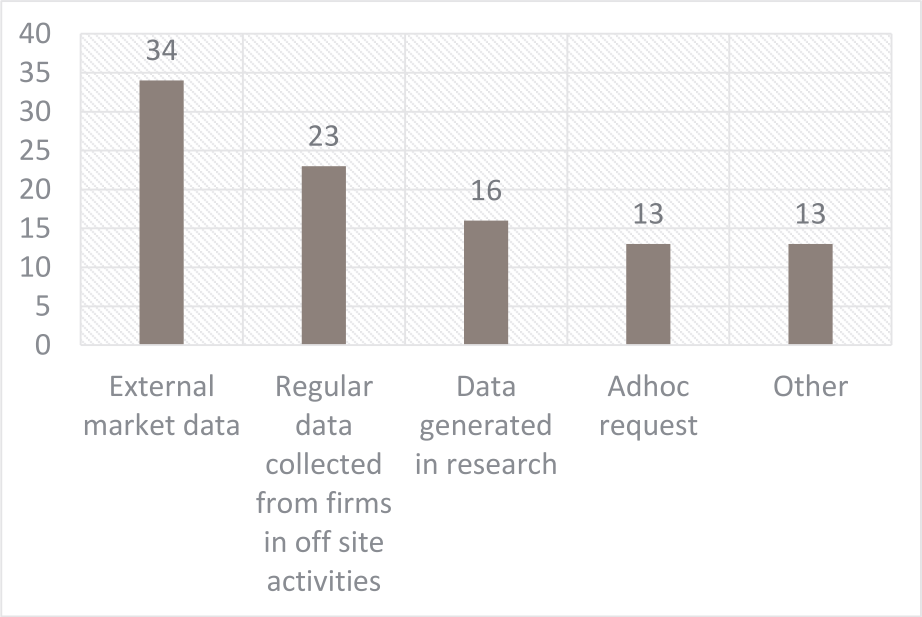


- Central banks have an **active interest in big data**, particularly **improving processing technology, adapting institutional strategies and increasing staff awareness** of the area. Just over half of respondents said they had been giving the area active consideration in the past 12 months.

- This group of 23 central banks was drawn from around the world, though, reflecting the make-up of participants, European central banks figured prominently.

- Although big data was not included in its strategic plan for 2016 -2021, a respondent from the Caribbean said it would move up on the list of priorities in the future, and, as a result, they have introduced **training initiatives**.

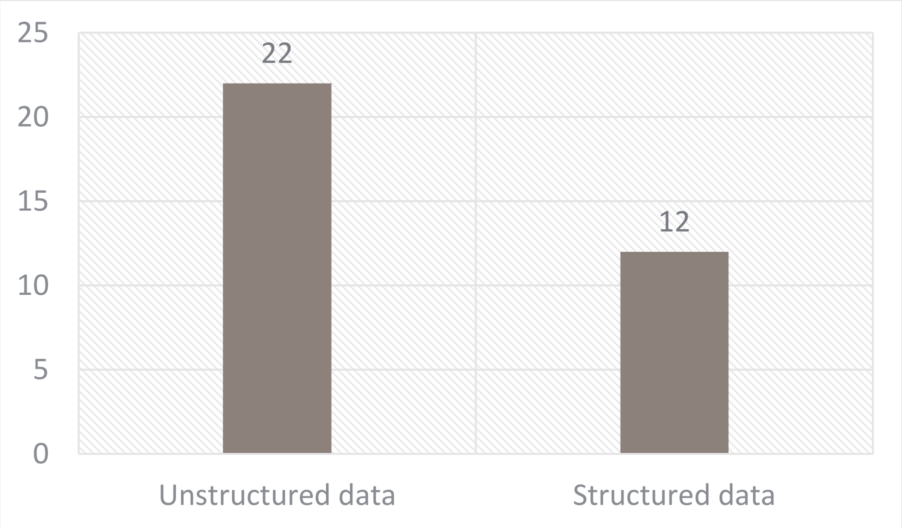
Which statement, in your view, represents the most accurate definition of big data?



- More than 80% of respondents said an example of big data was ‘**external market data**’, but only two respondents, both from Europe, checked this option exclusively. Nineteen of the 34 also chose ‘**regular data collected from firms in off-site activities’**. A statistician from a central bank in Africa included all five classifications in their answer and added “data from ministries, departments and agencies”.
- A European central bank ignored the suggested classification and offered the category of “commercial and administrative databases”

All respondents took part in this question: most gave multiple answers.

Which of the following types of data would you classify as big data?

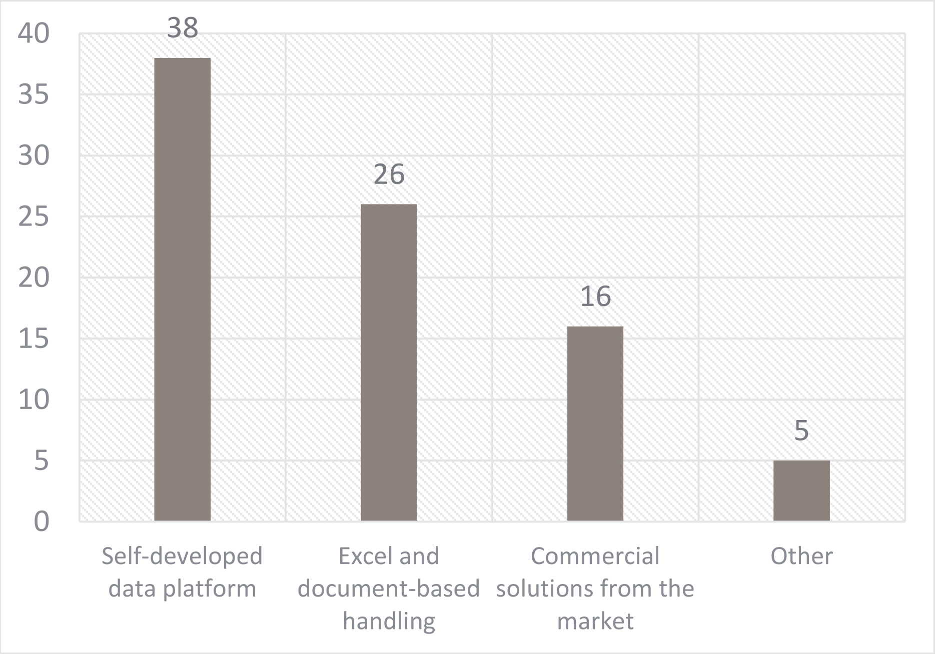


Forty-two central banks answered this question; eight of whom checked both “structured” and “unstructured”.

- While there is no single agreed definition of big data, industry has gravitated towards viewing it **as large volumes of data,** both structured and unstructured, which a standard desktop computer cannot handle. Central banks typically see **big data as unstructured data** that is sourced externally, though this view was not universally held.

- Just under two-thirds of respondents believe big data should be defined as unstructured data sets. In a comment, a European central banker said: “‘Unstructured’ seems to be a more accurate definition of big data because at least it requires unstructured data processing, which is more challenging.”

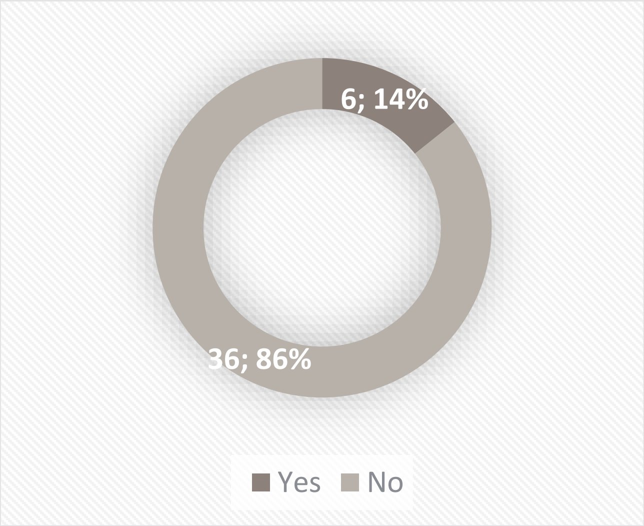
How does your central bank deal with regulatory data collection?



All respondents gave at least one answer.

- As central banks have expanded the depth and breadth of their involvement in financial stability policy-making since the financial crisis, so the importance of collecting accurate, complete and timely data from institutions has increased. Overwhelmingly, central banks **develop their own data platforms to handle regulatory data collection**. This was the view of 38, or 90%, of the 42 respondents.
- **Commercial solutions from the market provide** a viable **alternative** for central banks, and this option was chosen by nearly 40% of respondents. This was typically used in conjunction with self-developed platforms, as was the case for three-quarters of this group.

Has the approach to data collection changed in the past 12 months?

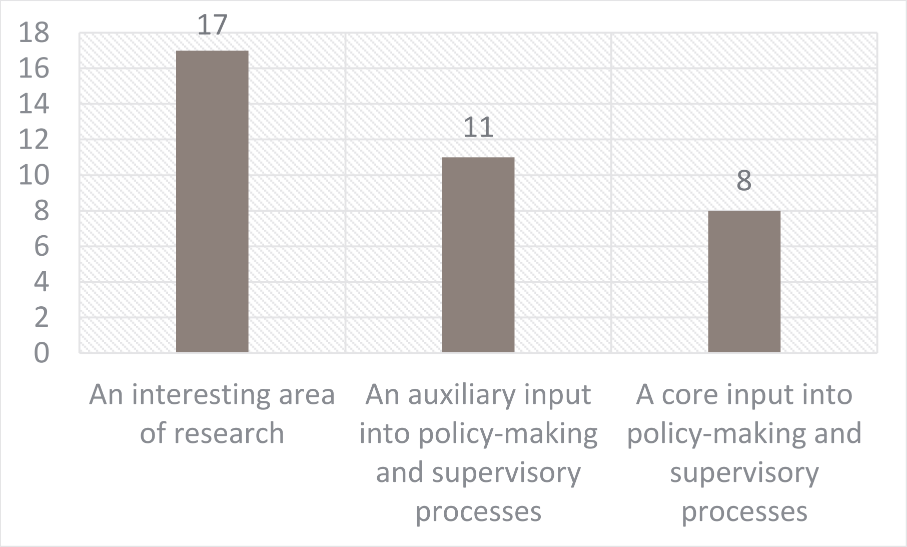


- Arrangements for regulatory data collection display a high degree of continuity. Thirty-six respondents said their approach to regulatory data collection **was unchanged** in the past 12 months. Few respondents volunteered a comment explaining their view, though a handful indicated a transition is underway.

- A statistician from the Caribbean said their regulatory data collection approach is **expected to change over the next six months**. Similarly, a respondent from the Americas said: “We are in the process of creating better foundations for policy decisions and have therefore changed strategies for managing data and statistics.” One European respondent noted they are reviewing future plans, while another said their central bank is implementing software for both the banking supervision and statistics departments, demonstrating the collaborative nature of big data projects.

- Six central banks have changed how they deal with regulatory data collection in the past 12 months. This group was dominated by central banks based in Europe with one adding that it was necessary to change their data collection process “for technological and business reasons.”

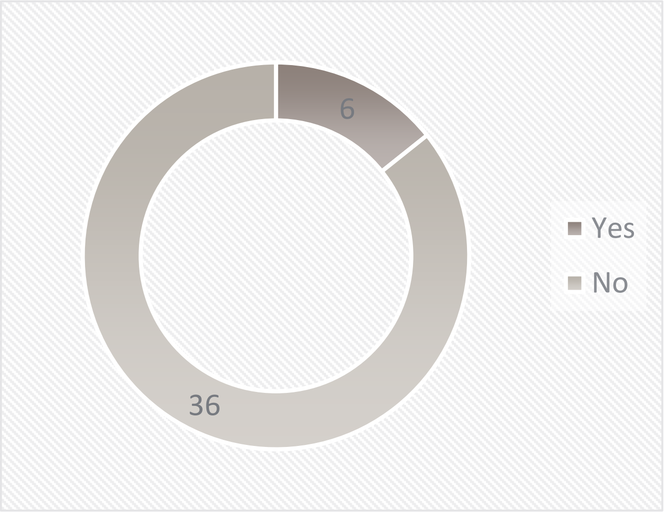
Which best represents your central bank's view of big data?



All respondents answered this question: six gave multiple answers.

- Big data is predominantly regarded as **useful for research in central banks**, but significant minorities see immediate involvement in policy-making, or scope for this. Nearly 50% of respondents chose **"an interesting area of research"** as the best match for how their central bank views big data. A European respondent echoed several others, saying: "the central bank's view of big data is going to change over time. Right now, it is an active area of research". A respondent from an advanced economy implied big data would influence policy: "the optimisation of data is of key importance in driving decision-making."
- Eight central banks indicated big data was a core input into policy-making and supervisory processes. This included the largest central bank that participated in this survey, as well as two respondents from the Middle East. One commented:
- An efficient data management is obviously needed in the course of all missions for which the central bank (including the supervisory authority) is responsible. Big data techniques are useful in that respect.

Has your central bank's view changed in the past 12 months?



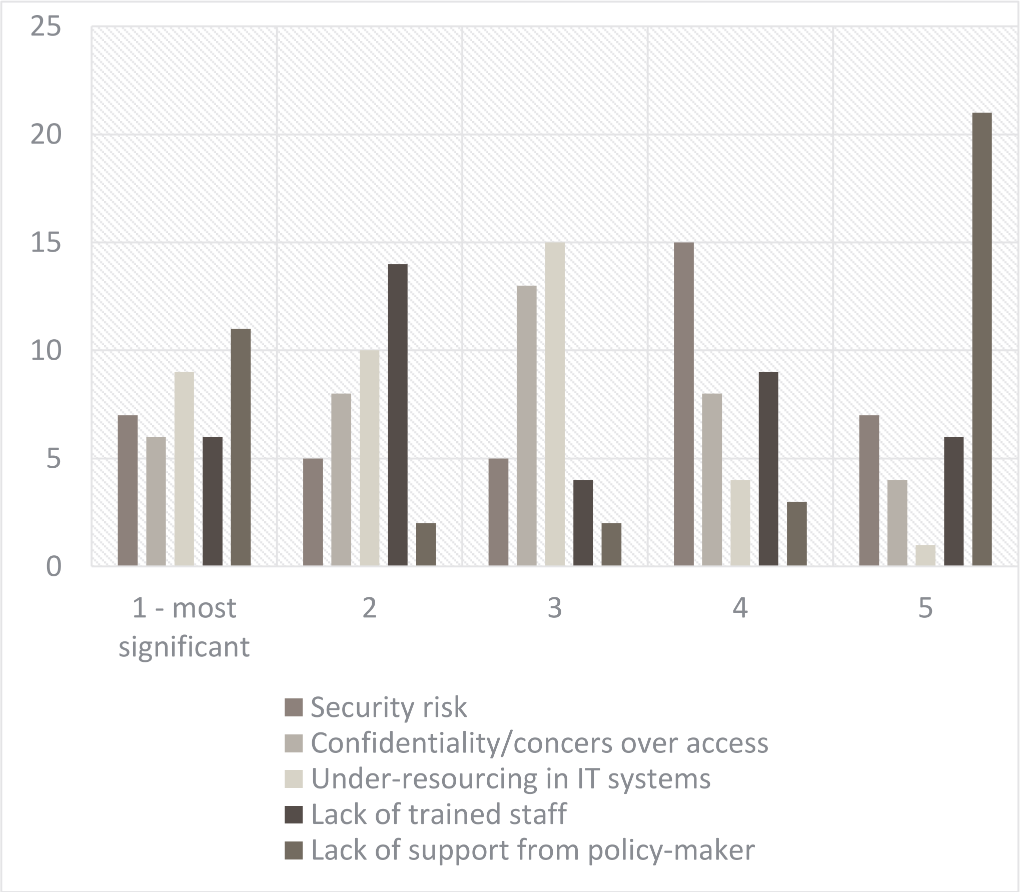
- Views of big data's role in a central bank are largely unchanged. Over 85% said their **view has not changed** in the past 12 months.

- "We expect this to change in the medium term," observed a European central banker, who said they have dedicated a team to this work and therefore their view may change "as a consequence of our research conclusions".

- The six central banks that have made a change in the past 12 months **highlighted how big data was having a catalytic effect on their institution**. A statistician from Europe commented that data management is part of the central bank's new strategic plan. Another European central bank is advancing their work in this area, with dedicated teams within divisions outside of the statistics department: "There has been a change over the past 12-24 months which has seen the creation of dedicated analytics teams within supervisory divisions and a move towards creating a more data-centric organisation, which is reflected in organisational level objectives."

- An officer from the Americas looked to the importance of an information strategy for the future development of big data: "To ensure that information gathered is handled in an appropriate manner, a **vision for information supply is therefore required, along with an accompanying strategy which guides how data is required and processed**."

Which in your view present the greatest roadblocks or challenges to increased use of data sets in your central bank?

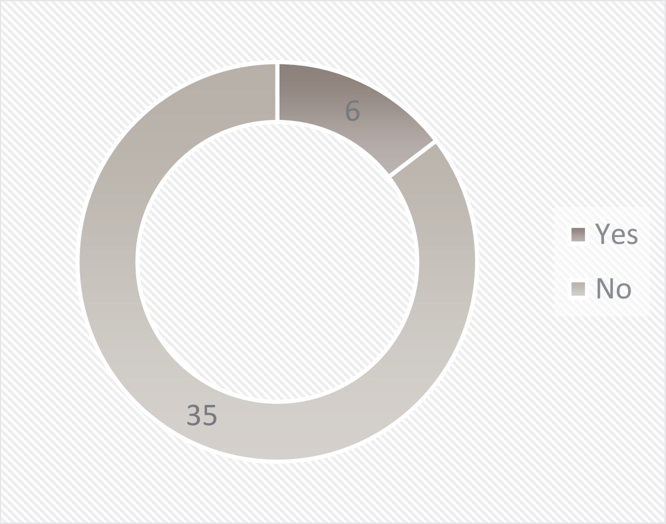


Votes were cast using a scale of 1-5, where 1 denotes the most significant roadblock, and 5 the least significant.
Three respondents did not reply.

- **Lack of support from policy-makers** was most commonly identified as the greatest challenge and, intriguingly, also received the most votes as the least challenging. Twenty-eight per cent of respondents – seven Europeans, and four from the Americas – scored this as the most significant hindrance. “There are no problems with security, but there is not enough support,” one European said. Conversely, 21, or 54% of, respondents saw a lack of support from policy-makers as the least significant challenge. This group was largely made up of developing and emerging market economies.

- Concerns over skillsets figure prominently in central bankers’ thinking. Just over half of respondents placed a lack of trained staff as the first or second most significant challenge.

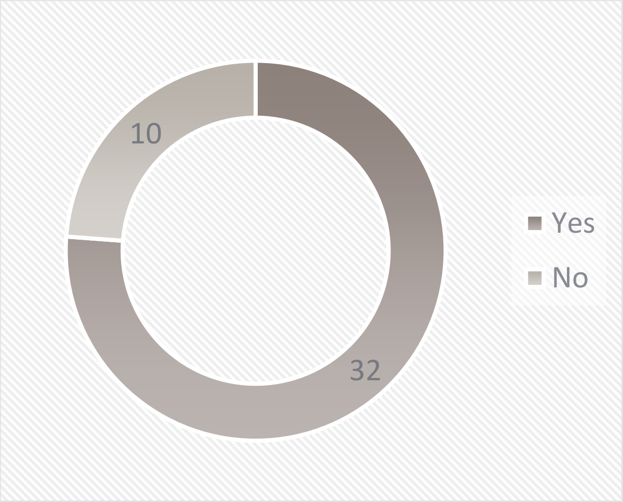
Does your central bank have a single allocated budget for the handling of data (including big data)?



- The vast majority, 85% of respondents, said their central bank **did not have a single allocated budget** for data. This group was largely made up of central bankers from advanced and developing economies.
- In comments, respondents typically attributed this to budgeting being **divided on departmental or project-specific bases r**ather than by resource, function or output. In this way, data was often included as part of the technology budget. One central banker commented: “several entities are involved in the handling of data.”

One respondent did not reply.

Does your central bank have a shared internal platform to enable different areas of the central bank to access data resources?

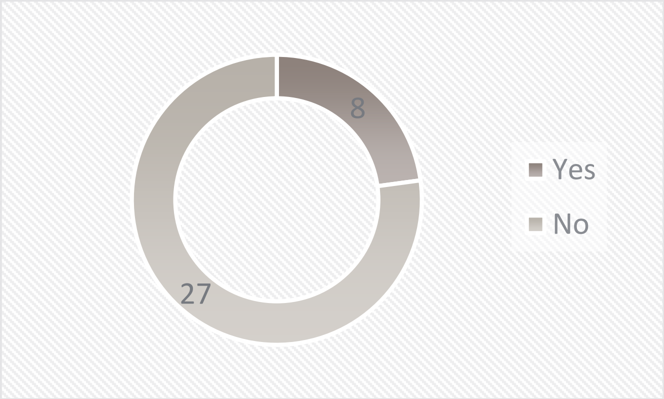


- Over three-quarters of respondents indicated they **had a shared internal platform**, typically in the form of reporting frameworks and data warehouses. A respondent from a European central bank described the two-pronged approach used in their institution: "We have a shared internal platform for accessing most of our supervision data. For other macroeconomic data we have another platform contributed by the economic department and used by everyone inside the bank."

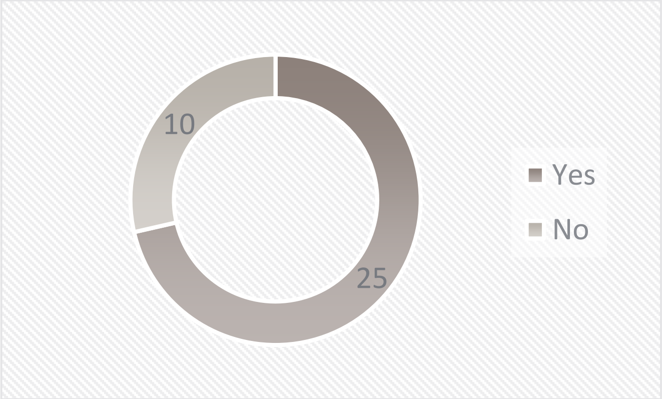
- A respondent from a large central bank explained how they granted access to the data: "The first key element in the data strategy was to allow users with an appropriate business case to view all data relevant to them across the organisation. A technology solution was implemented to allow for this."

Does your central bank have a shared internal platform to enable different areas of the central bank to access data resources?

If yes,
i) can this platform be accessed externally
(ie for researchers or dissemination purposes?)



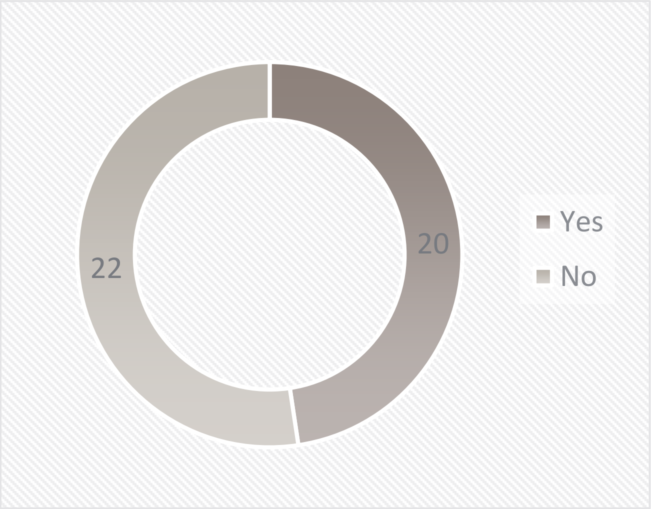
ii) was this shared platform built by the central bank?



(1) Footnotes and sources in speaker notes

- Data-sharing platforms in central banks are typically built **internally** by the central bank and are not available for external use.
- The eight central banks that do **allow external access** to their self-developed platforms consisted of five European central banks, one from the Americas, one from the Middle East and one from Africa. A European central banker said: “the data software is externally accessed by banking supervisors when examining the financial institutions. This platform is restricted for researchers and dissemination purposes.” Of the eight central banks that do allow external access, six built their own platform independently.
- Several respondents from developed countries, however, said they **were establishing platforms that would be available for external use**. One central banker said “external access is still under development”, while another commented that some parts of the database were being made available to registered users.
- In the main, central banks **build rather than buy shared platforms**: this was the experience of just over 70% of respondents. One European central bank said they used “in-house developments with standard big data infrastructures”. Ten central banks indicated they did not build their own platforms. Their reasons ranged from seeking assistance from another central bank in the development to subcontracting the process.

Does your central bank have clear data governance, with defined roles and responsibilities eg, chief data officer, data stewards?

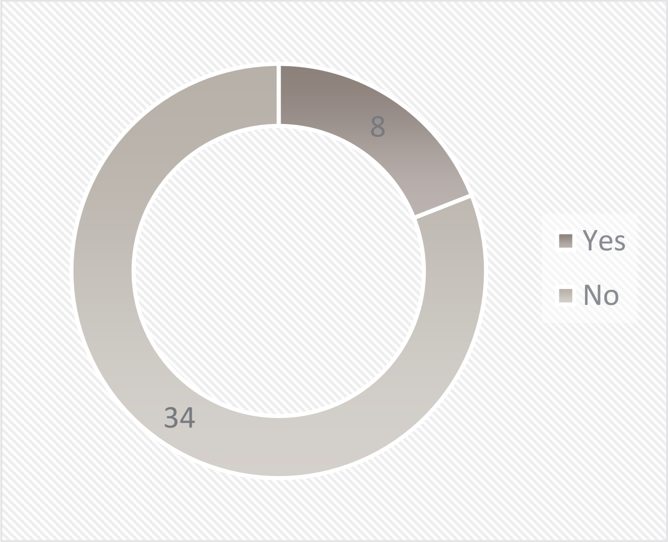


- Many central bankers are **concerned by data management arrangements**, and just over half of respondents said their banks **did not have clear data governance**. European central banks featured prominently in this group, but there was also a significant number of respondents from the Americas. One said "we have no chief data officer or equivalent".

- This is clearly an area of intense activity, however, as central banks are striving to improve data governance frameworks.

- A central bank from the Caribbean said the framework is "in its infancy" but is "expected to mature" in the coming years. An officer from a central bank in Africa commented a clear data governance structure "is now being put in place". Several central bankers from Europe said it is being discussed, however it had not been formalised, a view typified by this statistician's comment: "Certain relevant people are more aware of their roles and responsibilities as data owners or data stewards, but it has not been fully formalised and implemented yet." A developed country central banker noted: "We have a CDO and data stewards identified in place, however, more robust governance is being put in place over the coming months."

Does your central bank have any intra-departmental or divisional bodies, e.g. committees or working groups, dedicated to big data?

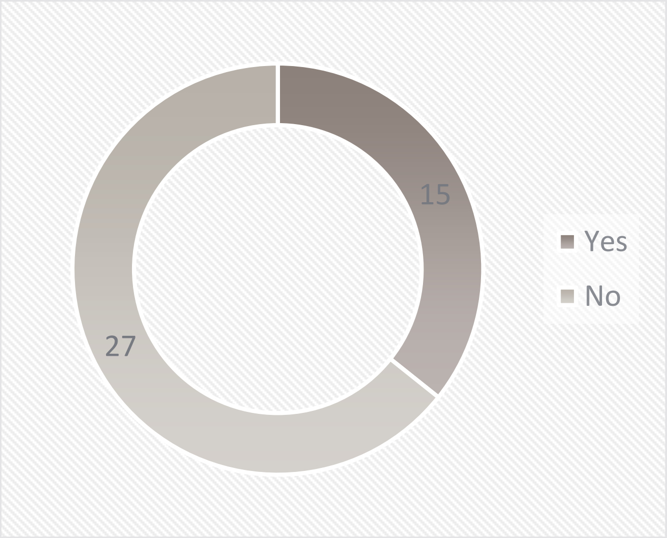


- Big data is **generally confined to departments or divisions** in central banks. Just over 80% of respondents said they **do not have** any intra-departmental or divisional bodies dedicated to big data. This group featured five African central banks and all nine of the central banks from the Americas, along with two central banks from the Middle East.

- Comments made by the 19% of central banks which have intra-departmental or divisional bodies centred around bank-wide data teams and committees dedicated to big data. One central bank commented: "This year we have started an inter-departmental team dedicated to big data issues", while another made reference to big data being discussed within a team of people focused on data in general. One central banker from an advanced economy said big data plays a part throughout the central bank, however it is governed by the statistics department director: "Many projects deal with big data issues: the corresponding steering committees are temporarily chaired by the Director of General Statistics."

- Although 34 central banks indicated they do not have an intra-departmental or divisional bodies at their central banks, three said it is a field they are looking to develop. One such central bank in the Americas commented: "We have some employees working on big data.

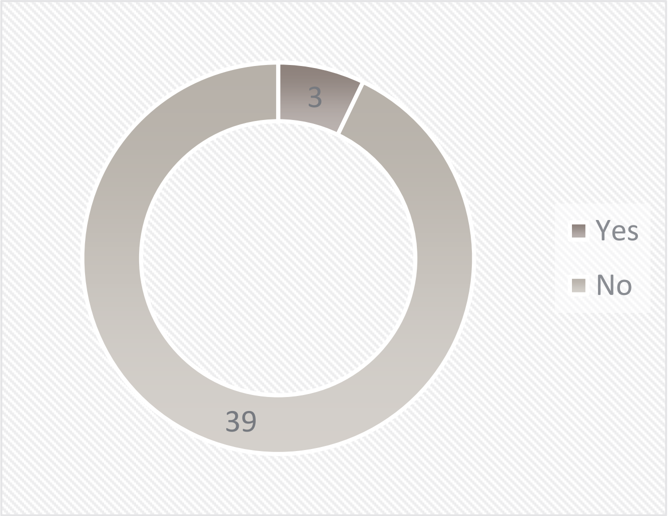
Does your central bank use external data providers for big data sets?



- Central banks **generally source their own big data sets**, though a significant minority look elsewhere. Respondents that do not use external data providers were predominantly from emerging-market economies, including 17 central banks from Europe.

- Those that did use external providers, tended to turn to commercial banks and firms, social media and google blogs, and mobile phone operators. “We employ data from blogs, social media and private websites,” said one respondent. A statistician from Europe commented on the techniques used to collect the data from websites: “Currently, we are doing web scraping projects, collecting data from different websites.”

Does your central bank outsource any data processing for big data sets?

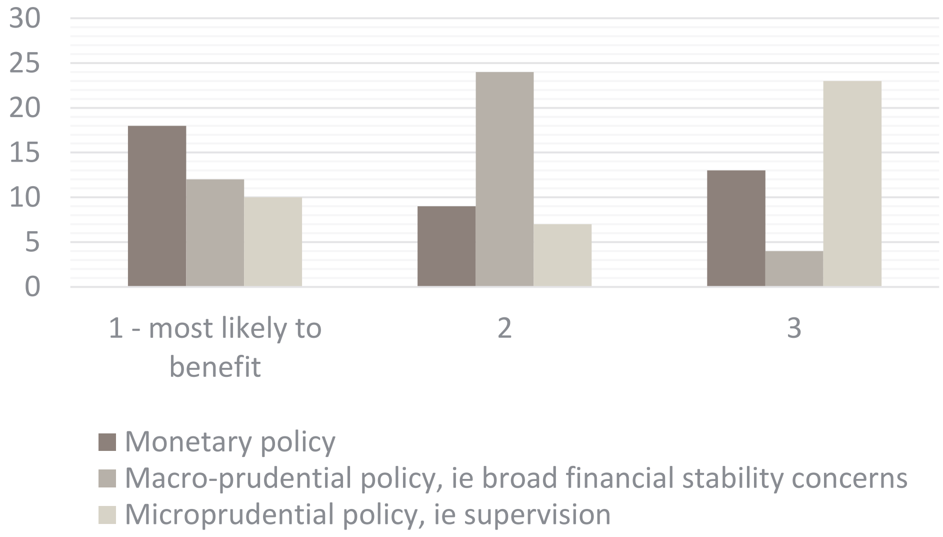


- Overwhelmingly, **central banks process their own big data sets** and there is **no indication of a desire for this to change**. More than 90% of respondents answered they do not outsource any data processing.

- A central bank in the Americas said straightforwardly: “Data processing is run in-house”, while one central bank in Europe commented: “The data management and statistical analysis of big data sets is taken on by highly qualified staff in Director General of Statistics with the support of the dedicated team in IT department.” Three central banks outsource data processing for big data sets, however they all declined to comment.

- Of the 39 central banks who indicated they did not outsource data processing, only three noted this was likely to change in the near future. One central bank in the Americas is currently hiring in this sector in order to cater for in-house data processing: “The central bank is now recruiting an expert in the field of data architecture in order to develop intra-departmental systems and working routines of data.”

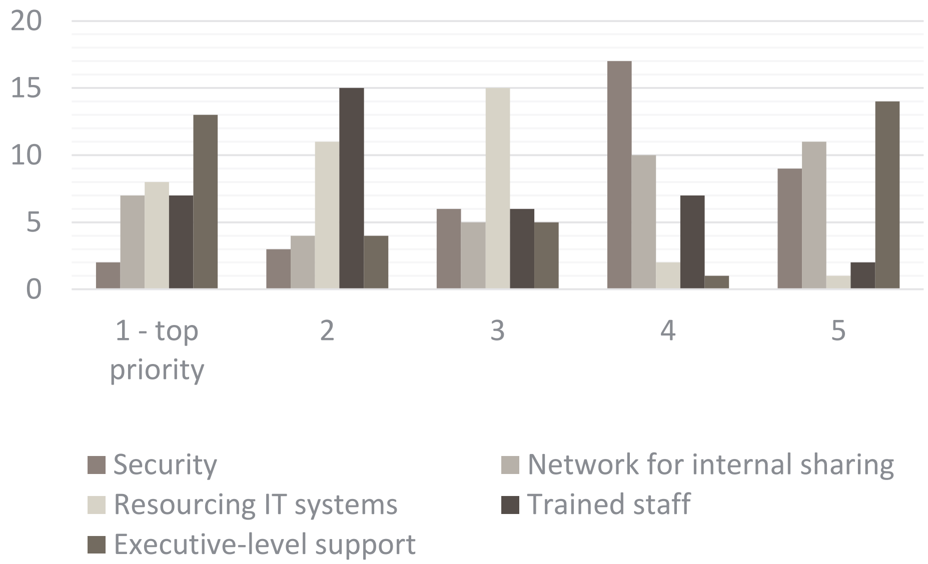
In your view which of the following areas stands to benefit most from big data, in practical terms?



Two respondents did not reply.

- Respondents saw **monetary policy as likely to benefit most from big data,** though they said it will have a significant impact on macro-prudential policy as well. As a central banker from an industrial economy explained: “Monetary policy, macro-prudential and micro-prudential policies can benefit from big data. Monetary policy can benefit from better and timelier nowcasts of macroeconomic variables. Macro- and micro-prudential policies might benefit as well.”
- Interestingly, **90% ranked macro-prudential policy in either first or second place**. Those that ranked it in first place typically chose monetary policy as their second choice. Micro-prudential policy was ranked in third place by 58% of respondents but 10 respondents did place it first place.

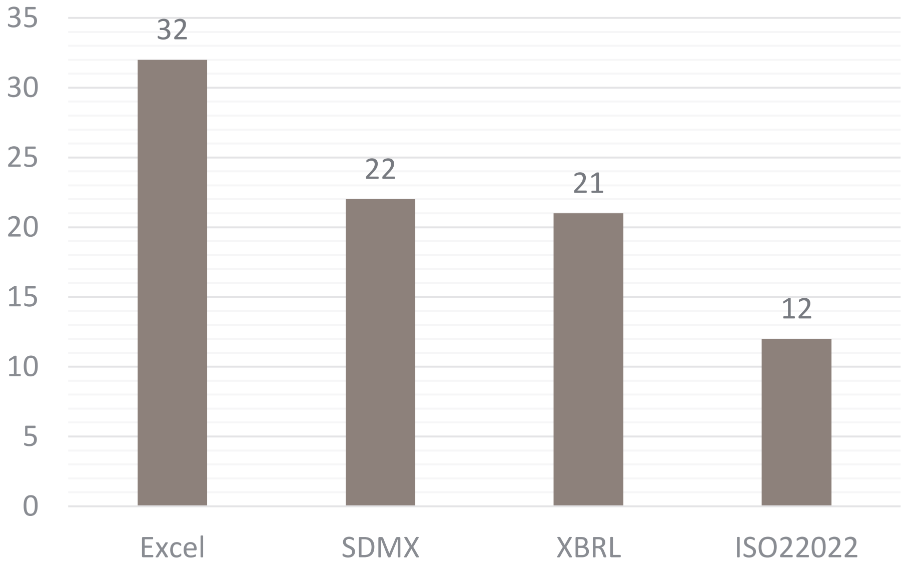
Which area do you consider the priority for investment to increase big data use in your central bank?



Five respondents did not reply.

- **Support from the executive-level and** policy-makers again divided respondents: 35% saw it as top priority, 38% saw it as the lowest. Those that saw it as a top priority were mainly central banks from Europe and the Americas. A European officer stressed the importance of having a budget: "Obtaining a budget is of fundamental importance in an initiative so therefore is a key priority."
- Sixty per cent of respondents ranked trained staff in either first or second place. This group included half a dozen emerging markets. However, these central banks declined to comment. Resourcing IT systems was most commonly ranked in third place. One European central bank commented: "Ensuring that IT infrastructure can handle additional demands is a key component of where this investment will be required." Network for internal sharing was largely ranked in fourth and fifth place by respondents, as was security.

Which of the following standards does your central bank use, or plan to use, for dealing with data exchange and collections?

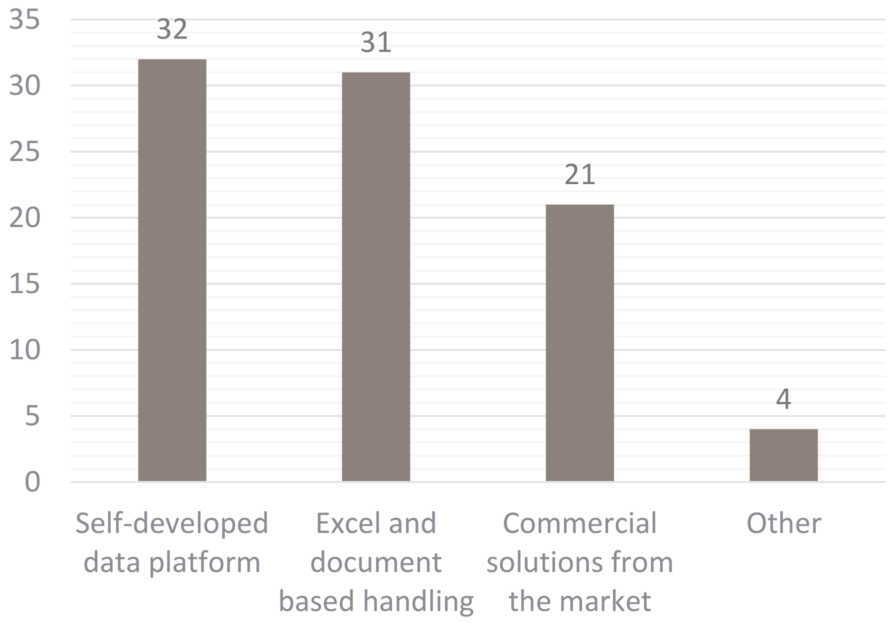


Three respondents did not reply.
Respondents checked multiple answers.

- **Excel is the most popular standard** for central banks when dealing with data exchange and collection but it is typically used in conjunction with another solution. Of the 39 respondents, 82% use Excel, and around 40% of those that use Excel use it exclusively. Of the remainder, most used either SDMX or XBRL as well, and three-quarters use both.

- The standards of XBRL and ISO20022 are used typically for specific functions: supervision and payments, respectively. Twenty-one respondents indicated that they use SDMX at their central bank. A central bank from an advanced economy noted: “XBRL is only used in the transactional system component of RIAD application (Register of Institutions and Affiliates Database).” Although the least popular answer, ISO20022 was likewise described by respondents as useful for specific functions. One European officer noted: “Currently we are using the three standards above-mentioned: SDMX and XBRL are widely used, whereas ISO20022 is only being used in very specific exchanges”.

How does your central bank deal with regulatory data analytics?

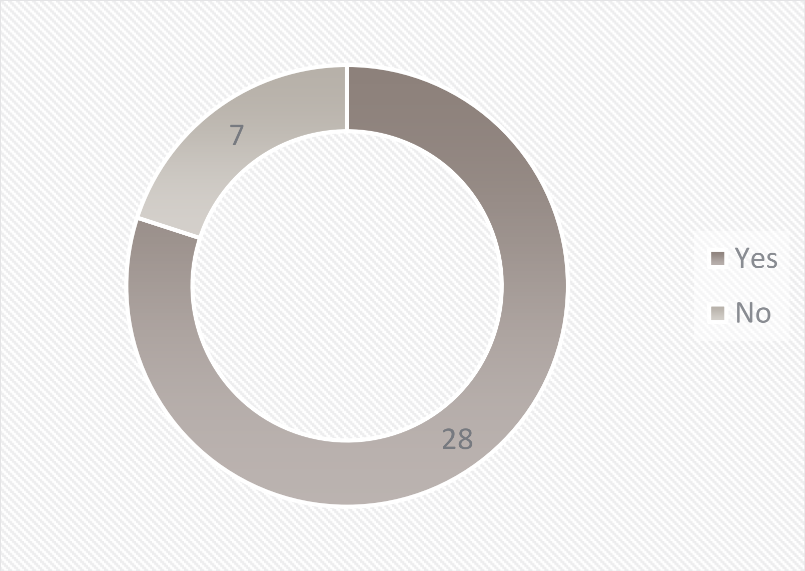


One respondents did not reply.

- Central banks have **developed their own data platforms to deal with regulatory data analytics**, but this is often used in conjunction with Excel and document-based handling. This was the answer from over three-quarters of respondents, the majority of which were from developed countries. Only four of the 32 above use purely self-developed regulatory data analytics while the remaining 28 combine it with another standard. Nearly all combined this with Excel and document-based handling and just under half reported a commercial element to their regulatory data analytics workflow.

- Commercial solutions from the market proved popular among respondents, with over half choosing this option. Interestingly, **more central banks use commercial solutions from the market** for regulatory data analytics than they do for regulatory data collection, with only 16 central banks.

As of January 2016 Globally Systemically Important Banks (G-SIBS) have been regulated to meet the BCBS 239 Principles, after carrying out self-assessments, in respect of data collection, data aggregation and dissemination capabilities. Do you think it would be useful for central banks to self-assess using an adapted version of these principles?



Seven respondents did not reply.

- Central bankers largely welcome the idea of self-assessment using an adapted version of the BCBS 239 Principles for Globally Systemically Important Banks (G-SIBs). Of the 35 respondents, 80% said such self-assessment would be useful.
- As databases increase in size and complexity, there is naturally a concern that they are policed and governed properly. Furthermore, one central banker noted existing systems were under pressure: "Due to increasingly large and complex data that is now challenging traditional database systems."
- Several central banks drew attention to the internal processes already in place to regulate big data processing. One European respondent from an advanced economy noted: "There is a benefit that can be accrued from this. It is important to note however that our internal audit function already undertakes full audits in this area." However, if the objective of these principles is to strengthen banks' risk data aggregation capabilities and internal risk reporting practices, it might be relevant to make a similar self-assessment in central banks.

Results of survey of how central banks view big data and data governance

Summary of key findings (1/2)

- Central banks have an **active interest** in big data. This is manifested in improving processing technology, adapting institutional strategies and increasing staff awareness of the area.
- Central banks typically see big data as **unstructured data** that is sourced **externally**, though this view is not universally held.
- Overwhelmingly, central banks develop their **own data platforms** to handle regulatory data collection, a role that has taken on greater significance since the financial crisis as central banks have expanded their involvement in financial stability.
- Big data is predominantly regarded as **useful for research**, but significant minorities see immediate involvement in policy-making, or scope for this.
- **Lack of support from policy-makers** is seen as the most significant challenge to increase use of big data.
- Central banks **do not in the main have a dedicated budget** for the handling of data (including big data), though many are seeking one.
- A little over 80% of respondents said they **do not have any intra-departmental or divisional bodies** dedicated to big data.
- More broadly, central bankers have **concerns** over the arrangements in place **for managing data** in their institutions. Many are looking to improve data governance.

Results of survey of how central banks view big data and data governance

Summary of key findings (2/2)

- Over three-quarters of respondents indicated they had a **shared internal platform**, typically in the form of reporting frameworks and data warehouses.
- Central banks generally **source their own big data sets**, though a significant minority increasingly look elsewhere for these. Overwhelmingly, they process these themselves and there is no indication of a desire for this to change.
- **Monetary policy** is seen as standing to benefit most from big data, though it is expected to have a significant impact on macro-prudential policy as well.
- **Support from the executive-level and policy-makers** divided respondents: 35% saw it as top priority for investment within the central bank, 38% saw it as the lowest.
- Central banks have developed their **own data platforms to deal with regulatory data analytics**, often used in conjunction with other options. Excel remains popular.
- Central bankers broadly **welcome the idea of self-assessment of data management** using an adapted version of the Basel Committee’s BCBS 239 principles for supervisory data aggregation.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Overview of international experiences with data standards and identifiers applicable for big data analysis[1]

Michal Piechocki,
Business Reporting-Advisory Group

[1] This paper was prepared for the meeting. 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 meeting.

Overview of international experiences with data standards and identifiers applicable for big data analysis

Potential of use of big data analytical methods with standardised, regulatory, financial data sets

Author: Michal Piechocki (michal.piechocki@br-ag.eu)

Abstract

Financial regulators collect and process data, that both meets and contradicts volume, velocity and variety criteria commonly accepted for big data analysis. A number of regulatory data frameworks are described using international standards and identifiers, that may aid in improving efficiency of big data and machine learning algorithms, especially applied with granular data sets, which are increasingly requested by regulators. However, even application of big data methods requires understanding of the researched data and accuracy of information, for precise, unbiased identification of correlations and causations. This paper discusses how standards and identifiers, used across regulatory frameworks, may support application of big data analysis. The paper concludes with identification of further research fields, arising from combination of standardised, regulatory data pools with public data feeds, for discovery of new regulatory insights.

Keywords: data standards, SDMX, XBRL, ISO 20022, granular, transactional, big data, analysis, algorithms, LEI, UTI, UPI

JEL classification: C55, C8, E58, G28, 032, 033

Contents

Excerpts from central banks' leaders speeches

"Big data analytics enables better quantification and pricing of risks, and helps strengthen ex-ante risk resilience measures."

Ravi Menon, Managing Director Monetary Authority of Singapore

"I can see that we stand at the start of a period where central banks, like everyone else, will make use of "big data" and we should learn how to use them to maximize their benefits. While these potential benefits are large, the effort needed is equally significant. We need to invest in information technology infrastructure, but we also need to educate our statisticians how to deal with the new larger and more complicated data sets. (...) Instead of receiving readily usable processed information, we are beginning to demand from reporting agents huge amounts of granular information that is then processed in-house by our statisticians. There is a need to streamline the process of collecting data. In particular, we should exploit to the maximum synergies between the collection of supervisory and (traditionally) statistical data, by developing common definitions to the extent possible, or simple rules to transpose the ones into the others. (...) Central banks are leaving the small safe harbor of simple, aggregate data and are opening up to the brave new world of granular big data. In order not to get lost, we need new skills, more crew, that is statisticians, and stronger vessels, that is better and more versatile models"

Yannis Stournaras, Governor of the Bank of Greece

"Systemic risk, for example, is defined as the contribution of the distress of individual financial institutions (or a group of financial institutions) to overall stress in the financial system, with adverse repercussions on the real economy. The contribution of individual financial institutions to systemic risk is higher, the greater the risk of an individual institution, the larger an institution is (too big to fail), the more connected an institution is (too connected to fail), or the more financial institutions are exposed to common risk factors (too many to fail). This definition shows that systemic risk cannot be analyzed without making use of detailed and granular data on financial institutions. (...)Technological progress has contributed to improved access to micro data and to improved handling of large, granular datasets. (...)It will be possible to reap the full benefits of micro data in terms of efficiency and effectiveness of reporting only if there is close coordination between the scope of existing statistics and newly collected micro data. This may, in some instances, require the scope of existing statistics to be adjusted, and it requires detailed planning when designing new data requirements."

Prof Claudia Buch, Deputy President of the Deutsche Bundesbank

"We are also exploring how we - and others - could use the data the Bank collects more effectively. Big Data has the potential to help the Bank's policy committees identify trends in systemic risk and the economy."

Mark Carney, Governor of the Bank of England

Overview of data standards and identifiers used in the financial industry

Introduction

The concept of application of analytical methods over voluminous, high-frequency and diversified data sets has settled well within regulatory environments. As highlighted by the MITSloan Management Review[1], over the past years a number of implementations indicated value of such analysis, for example for estimation of inflation, examination of housing and employment market conditions or studying the impact of high-frequency trading on stock markets by looking at equity transactions.[2] The most commonly used big data analysis methods including: association rule learning; classification tree analysis; genetic algorithms; machine learning; regression analysis; sentiment analysis; social network analysis and other, , coupled with granular data sets, promise, and in some cases already deliver, insightful results. In the process of applying big data methods over public and regulatory data sets researchers[3] and regulators[4] observed however a number of challenges related primarily to:

1. data governance, architecture and understanding;
2. data fragmentation in siloes;
3. performance of IT infrastructure;
4. statistical and analytical methods to limit false positives, data bias or noise accumulation and
5. knowledge and researches availability.

While some obstacles, such as 3 and 4, are gradually being overcome through technological advancements, other, like 1, 2 and 5 remain a key burden in realising the big data potential.

In this article, we will focus on understanding the conditions of and potential solutions to data governance, architecture, understanding and fragmentation hurdles, affecting efficiency of big data analysis.

[1] https://sloanreview.mit.edu/case-study/better-data-brings-a-renewal-at-the-bank-of-england/

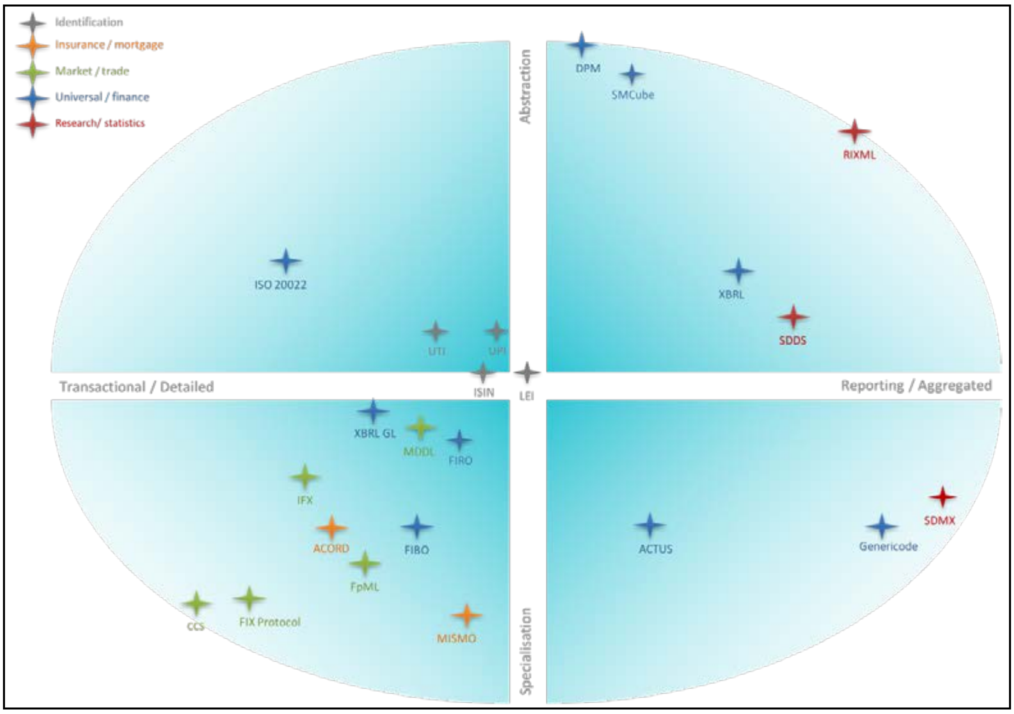
[2] N. McLaren and R. Shanbhogue, "Using Internet Search Data As Economic Indicators," Bank of England Quarterly Bulletin 51, no. 2 (2011): 134-140; D. Pimlott and T. Bradshaw, "Bank of England Googles to Track Latest Trends," Financial Times, June 13, 2011; E. Benos and S. Sagade, "High-Frequency Trading Behavior and Its Impact On Market Quality: Evidence From the UK Trading Market," working paper no. 469, Bank of England, London, December 2012, www.bankofengland.co.uk; and E. Benos, A. Wetherilt, and F. Zikes, "The Structure and Dynamics of the UK Credit Default Swap Market," Financial Stability Paper no. 25, Bank of England, London, November 2013, www.bankofengland.co.uk.

[3] Xu z, Shi Y., Exploring Big Data Analysis: Fundamental Scientific Problems https://link.springer.com/article/10.1007/s40745-015-0063-7

[4] Nagel J. How the Banking Union has transformed banks' IT requirements http://www.bis.org/review/r141209a.pdf

Financial industry and regulatory data standards

Among the main tools utilised by financial regulators, in order to gain understanding of data they process, data standards and identifiers form an important subset, due to their role of enabling data collection, validation and organisation. Contrary to the popular expectation, there exists a large variety of financial data standards. Regulatory and industry experts forming the Frankfurt Group and its Technical Workshop[5] have analysed most common standards applied within the banking and insurance industries, and classified them according to the granular-aggregated axis and generic-specialised axis as presented on the Standards Map[6] below.



Picture 1: Financial data Standards Map

While classification of data standards and initiatives may be subject to experts' perceptions, the standards map demonstrates the heterogeneity of standardisation efforts, often competing across a variety of financial instruments, counterparties or other fields of interest. In summary, the map provides a classification of:

- 2 data description methodologies (DPM, SMCube) applicable to both granular and aggregated data sets;
- 4 granular data identifiers (ISIN, LEI, UTI, UPI);

[5] The Frankfurt Group Technical Workshop (FGTW) on Data Standards Interoperability is a discussion forum, organised under auspices of the European Central Bank, gathering regulatory standards experts and conveying quarterly workshops on the topics of data standards, identifiers, methodologies and technologies. The author serves as a chairman of the FGTW.

[6] The Standards Map was first published in the internal document of the FGTW: Piechocki M., McKenna K, Dill J. Note on Technical Vision of Standards Interoperability, 2014-06-23

- 16 data standards:
  - 11 granular standards (FixProtocol, FIBO, FIRO, CCS, FPML, MDDL, ISO20022, ACORD, IFX, MISMO, XBRL GL)
  - 5 aggregated standards ((XBRL, SDMX, RIXML, SDDS, Genericode)
- 1 data initiative (ACTUS) describing extremely granular-level data.

The table presents a brief explanation of each component positioned on the map:

Table 1: List of data standards used by financial regulators

Abbreviation	Full name	Purpose
ACORD	ACORD Data Standards and Framework	Data standards for life and annuity property and casualty and for Global Reinsurance & Large Commercial. Claims and settlements messages.
ACTUS	ACTUS Financial Research Foundation	Data and algorithmic standard aiming to break down the diversity in financial instruments into a manageable number of cash flow patterns
CCS	Clearing and connectivity standard	Clearing of OTS transactions
DPM	Data Point Model	Multidimensional data modelling
FIBO	Financial Industry Business Ontology	Define financial industry terms, definitions and synonyms using RDF/OWL and UML
FIRO	Financial Industry Regulatory Ontology	Ontology for description of financial services regulatory domain
FIXProtocol	FIX Protocol	Protocol for international real-time exchange of information related to the securities transactions and markets
FPML	Financial Product Markup Language	Business information exchange standard for electronic dealing and processing of financial derivatives instruments
Genericode	Generic Code	Generic code list representation
IFX	Interactive Financial eXchange	Interoperability of systems seeking to exchange financial information internally and externally
ISIN	International Securities Identification Number	Unique international identification of securities
ISO 20022	Universal financial industry message scheme	Universal financial industry message scheme
LEI	Legal Entity Identifier	Standard for identification of business entities
MDDL	Market Data Definition Language	Standard to describe financial instruments, corporate events and market related indicators
MISMO	Mortgage Industry Standards Maintenance Organization	Data standards that cover the entire mortgage life cycle
RIXML	Research Information Exchange Markup Language	Language for description of investment research documents and other research

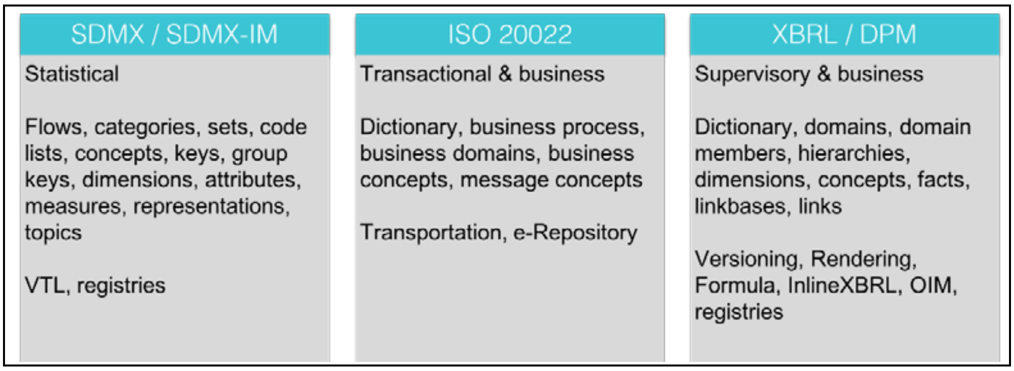
SDDS	Special Data Dissemination Standard	Standard for dissemination of statistical information
SDMX	Statistical Data Metadata Exchange	Statistical time series
SMCube	Single Multidimensional Metadata Model	Model used to define the structure of a group of datasets that have been compiled following different modelling methodologies (e.g. SDMX, DPM/XBRL).
UPI	Universal Product Identifier	Unique identification of the OTC derivatives data elements
UTI	Universal Transaction Identifier	Unique identification of individual OTC derivatives transactions required by authorities to be reported to trade repositories
XBRL	Extensible Business Reporting Language	Electronic business reporting
XBRL GL	Extensible Business Reporting Language Global Ledger	Open standard for transactional reporting

It is noteworthy to mention that a number of other initiatives is under way, such as schema.org[7] approach to describe details of financial instruments and transactions, through advanced blend of ontological descriptions with elements of existing standards and identifiers.

Furthermore, it is necessary to mention that the above classification should not, by any means, be understood as canonical. Rather, it represents an early approach to use the key purpose or origin of the specific data standard for initial categorisation. Nevertheless, the authors of the Standards Map recognise that real-world application of various standards crosses boundaries indicated by original intents. For instance, SDMX and XBRL are used to collect highly-granular data in a number of regulatory projects, as will be discussed further in this article. Similarly, granular data standards are increasingly coupled with aggregation mechanisms, in order to reflect aggregated indicators, cubes or groups of data.

Three data standards have been identified as key for the financial industry, and most commonly applied across multitude of regulations: ISO 20022, SDMX and XBRL. It is important to note, that FIX Protocol has also been widely adopted, however, due to strong cooperation between FIXProtocol and ISO 20022, the latter was taken into account. The diagram presents a brief summary of the standards.

[7] http://schema.org



SDMX / SDMX-IM	ISO 20022	XBRL / DPM
Statistical	Transactional & business	Supervisory & business
Flows, categories, sets, code lists, concepts, keys, group keys, dimensions, attributes, measures, representations, topics	Dictionary, business process, business domains, business concepts, message concepts	Dictionary, domains, domain members, hierarchies, dimensions, concepts, facts, linkbases, links
VTL, registries	Transportation, e-Repository	Versioning, Rendering, Formula, InlineXBRL, OIM, registries

Picture 2: Three most popular financial data standards

The ISO 20022 standard is a comprehensive, XML-based standardisation approach that includes a methodology, process and repository to be used by financial standards initiatives. As of date of publication of this paper the ISO 20022 describes processes, data repositories and messages for five domains: payments, securities, trade services, cards and FX.

The SDMX is an initiative and an XML-based standard led by major global regulatory and statistical bodies such as the IMF, the World Bank, the ECB or Eurostat. It describes time-series of data captured through variables according to a defined information model and exchanged through one of technical syntaxes supported by the standard.

The XBRL is an open, XML-based standard for exchange of multidimensional business information described in dictionaries called taxonomies, jointly with mathematical and logical business rules and exchanged as XBRL instance documents.

Application of data standards across regulatory data pools

Based on the European Union example it is possible to analyse which data pools, commonly collected and processed by financial regulators such as central banks, utilise data standards.

We have analysed 15 European Union regulations, initiatives or projects that include data standardisation within banking, insurance or capital market segments. The list of regulations follows:

1. Capital Requirements Directive IV / Capital Requirements Regulation (CRD / CRR)
2. Money Market Statistical Reporting (MMSR)
3. AnaCredit (AnaCredit)
4. Balance Sheet Items – Monetary Interest Rates (BSI-MIR)
5. Securities Holding Statistics (SHS)
6. European Markets Infrastructure Regulation (EMIR)
7. Markets in Financial Instruments Directive II / Markets in Financial Instruments Regulation (MiFID / MiFIR)
8. Securities Financing Transactions (SFT)
9. Undertakings for Collective Investment in Transferable Securities (UCITS)

10. Alternative Investment Funds Markets Directive (AIFMD)

11. Solvency II (Solvency II)

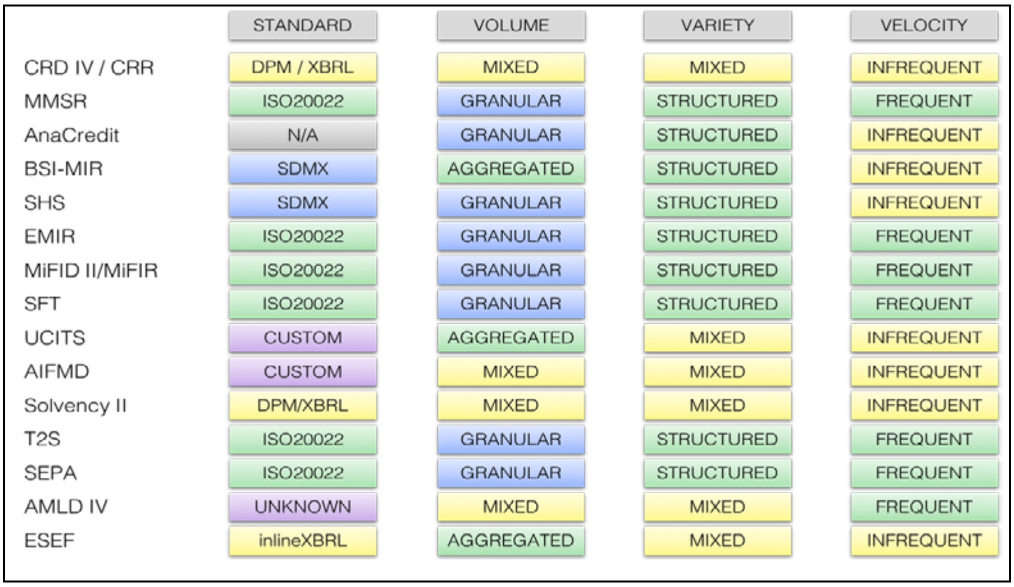
12. Target 2 Securities (T2S)

13. Single European Payments Area (SEPA)

14. Anti Money Laundering Directive IV (AMLD IV)

15. European Single Electronic Format (ESEF)

Each regulation was assessed for factors potentially applicable to big data analysis that is: volume, variety and velocity. For each regulation, a key data standard was identified.



	STANDARD	VOLUME	VARIETY	VELOCITY
CRD IV / CRR	DPM / XBRL	MIXED	MIXED	INFREQUENT
MMSR	ISO20022	GRANULAR	STRUCTURED	FREQUENT
AnaCredit	N/A	GRANULAR	STRUCTURED	INFREQUENT
BSI-MIR	SDMX	AGGREGATED	STRUCTURED	INFREQUENT
SHS	SDMX	GRANULAR	STRUCTURED	INFREQUENT
EMIR	ISO20022	GRANULAR	STRUCTURED	FREQUENT
MiFID II/MiFIR	ISO20022	GRANULAR	STRUCTURED	FREQUENT
SFT	ISO20022	GRANULAR	STRUCTURED	FREQUENT
UCITS	CUSTOM	AGGREGATED	MIXED	INFREQUENT
AIFMD	CUSTOM	MIXED	MIXED	INFREQUENT
Solvency II	DPM/XBRL	MIXED	MIXED	INFREQUENT
T2S	ISO20022	GRANULAR	STRUCTURED	FREQUENT
SEPA	ISO20022	GRANULAR	STRUCTURED	FREQUENT
AMLD IV	UNKNOWN	MIXED	MIXED	FREQUENT
ESEF	inlineXBRL	AGGREGATED	MIXED	INFREQUENT

Picture 3: Financial regulations and data standards (EU)

Despite relative subjectivity introduced in quantifiers, it is possible to observe, that data sets already collected and processed by financial regulators, while independently not meeting criteria for big data analysis, treated jointly constitute a voluminous, diversified and high-frequency data pool, that may benefit from application of big data algorithms.

Furthermore, the table demonstrates that financial regulators are slowly, yet steadily, harmonising their data requirements and applying common standards, rather than developing custom approaches. Importantly, most regulatory initiatives rely heavily on data dictionaries, and regulators, such as the ECB, are implementing standardised data dictionaries both, within the organisation (in the ECB: Statistical Data Dictionary based on SMCube), and for communication with supervised parties (for banks: Banking Integrated Reporting Dictionary).

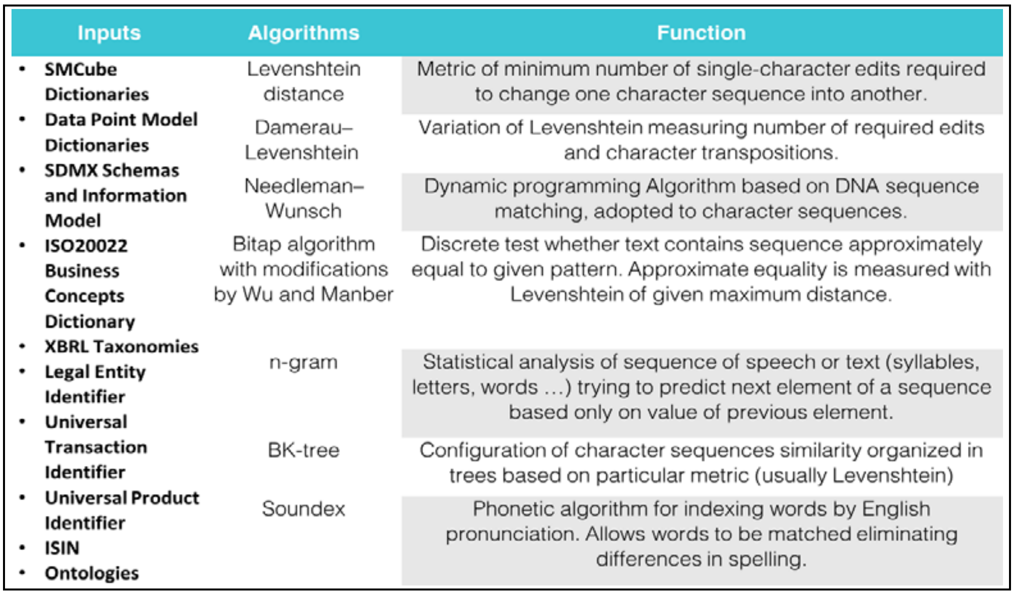
Potential of data standards for big data analysis

As demonstrated on the example of regulations from the European Union, a typical central bank may process standardised data sets that, jointly, may be subject of big data analysis. For example, a central bank in Asia or Latin America may collect and process similar to CRD/CRR, Basel Acord-driven data sets, such as information on own funds, market, operational and credit risk, leverage, liquidity, large exposures etc. In addition, many central banks already collect detailed granular data on loans or securities, such as AnaCredit or SHS. Collection of granular data related to payments is, by nature, a common experience among central banks. Increasing number of these data sets are collected and processed using standards such as ISO 20022, XBRL or SDMX.

Standardisation of regulatory data brings about at least two advantages for application of big data algorithms:

1. potentially removes the burdens indicated in introduction section related to data governance, architecture and understanding and data fragmentation in siloes;
2. standardised dictionaries, schemas and identifiers provide for valuable inputs for big data algorithms such as: keywords, keys, links and relations.

The picture below presents potential inputs from common regulatory data standards, for a variety of big data algorithms.

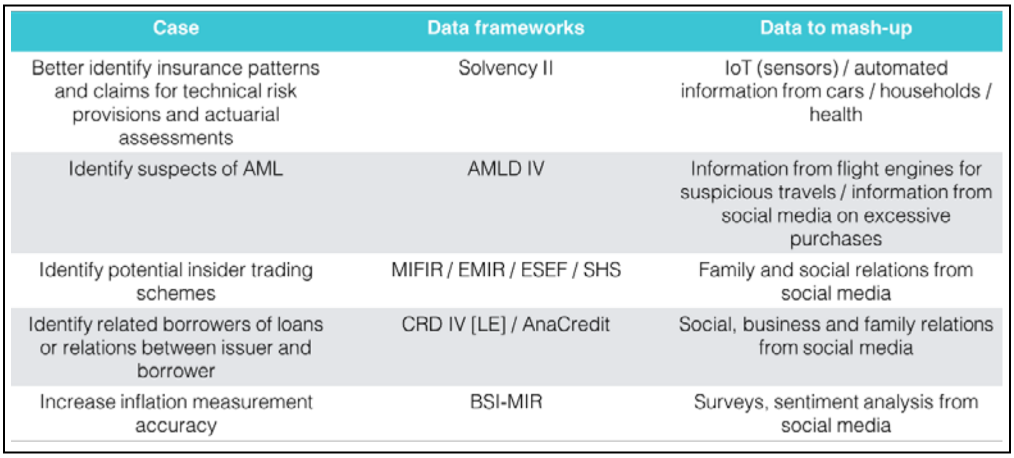


Inputs	Algorithms	Function
• **SMCube Dictionaries** • **Data Point Model Dictionaries** • **SDMX Schemas and Information Model** • **ISO20022 Business Concepts Dictionary** • **XBRL Taxonomies** • **Legal Entity Identifier** • **Universal Transaction Identifier** • **Universal Product Identifier** • **ISIN** • **Ontologies**	Levenshtein distance	Metric of minimum number of single-character edits required to change one character sequence into another.
	Damerau–Levenshtein	Variation of Levenshtein measuring number of required edits and character transpositions.
	Needleman–Wunsch	Dynamic programming Algorithm based on DNA sequence matching, adopted to character sequences.
	Bitap algorithm with modifications by Wu and Manber	Discrete test whether text contains sequence approximately equal to given pattern. Approximate equality is measured with Levenshtein of given maximum distance.
	n-gram	Statistical analysis of sequence of speech or text (syllables, letters, words …) trying to predict next element of a sequence based only on value of previous element.
	BK-tree	Configuration of character sequences similarity organized in trees based on particular metric (usually Levenshtein)
	Soundex	Phonetic algorithm for indexing words by English pronunciation. Allows words to be matched eliminating differences in spelling.

Picture 4: Inputs for big data algorithms

Since many big data algorithms rely on character-based analysis, structured dictionaries, classifications, ontologies and categorisations provide significant input for training of networks and machine learning algorithms in analysis of regulatory data pools. Particularly methodologies such as the Data Point Model, SMCube or SDMX-IM (SDMX Information Model) may provide for important inputs for big data analytical approaches.

The potential expands, if financial regulators consider combination of regulatory data sets, with public and commercial data pools, through a variety of mash-up techniques. The picture below presents a sample of potential application cases, where regulatory, standardised data, mashed-up with public sets, may provide new insights.



Case	Data frameworks	Data to mash-up
Better identify insurance patterns and claims for technical risk provisions and actuarial assessments	Solvency II	IoT (sensors) / automated information from cars / households / health
Identify suspects of AML	AMLD IV	Information from flight engines for suspicious travels / information from social media on excessive purchases
Identify potential insider trading schemes	MIFIR / EMIR / ESEF / SHS	Family and social relations from social media
Identify related borrowers of loans or relations between issuer and borrower	CRD IV [LE] / AnaCredit	Social, business and family relations from social media
Increase inflation measurement accuracy	BSI-MIR	Surveys, sentiment analysis from social media

Picture 5: Cases for potential application of big data analysis

The aggregated financial information collected in Solvency II tables, together with detailed technical risk provisions information, and detailed assets identification, combined with Internet-of-Things (IoT) sensors, delivering automated information from cars, households or inhabitants, may provide for better identification of insurance patterns, claims for technical risk provisions and actuarial assessments.

Information defined under proposed AMLD IV, combined with information from flight engines for suspicious travels and information from social media on excessive purchases, may support identification of suspects of money laundering.

Mashing-up of transactional data from trade repositories and securities information databases such as SHS, with family and social relations from social media, may aid in identification of potential insider trading schemes.

Similarly, family and social relations information mashed-up with loans data from CRD IV and AnaCredit, may provide for better identification of related borrowers of loans or identify relations between issuers and borrowers.

Last but not least, datasets like BSI-MIR, coupled with surveys and sentiment analysis from social media like Twitter or Facebook, may increase accuracy of inflation measurements.

Conclusions

Big data analysis promises valuable insights and discovery of new correlations and causations across voluminous, diversified and high-frequency data sets. While data sets typically collected by financial regulators like central banks may, individually, not meet criteria commonly accepted for big data analysis, the combination of regulatory data, coupled with public or commercial data, should open a new field of regulatory, supervisory and statistical financial analysis.

In order to realise the benefits of application of big data algorithms regulators should understand and standardise data sets they collect and process according to variety of regulations. Use of global, standardised identifiers should reduce potential bias and enable comparison of analytical results across industries, geographies or instruments.

Numerous standardisation efforts are under way and most advanced financial regulators have embarked on creation of data dictionaries, in order to introduce common understanding of data across organisation. These dictionaries are gradually being extended into the industry, to streamline data sourcing and mapping and therefore increase data quality and accuracy.

Application of big data algorithms over data already collected by regulators, combined with openly available sets, such as social feeds available through API (Application Programming Interface), provide central banks with unprecedented opportunity, to examine practical potential and value of big data analysis.

As highlighted by the speakers cited at the beginning of this paper, analysis of granular data is the new reality for central banks, however its efficient implementation requires understanding and mitigation of challenges identified by researchers and regulators. We, hopefully, demonstrated, that financial data standards, dictionaries and identifiers are not only a commonly applied foundation for building this efficiency, but they may also bring about unexpected value, through discovery and identification of new financial trends and phenomena.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Overview of international experiences with data standards and identifiers applicable for big data analysis[1]

Michal Piechocki,
Business Reporting-Advisory Group

[1] This presentation was prepared for the meeting. 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 meeting.

BRAG

Overview of international experiences with data standards and identifiers applicable for big data analysis

Michal Piechocki
Chairman | Frankfurt Group Technical Workshop
Director | XBRL International Board of Directors
CEO | BR-AG

Bali, March 2017

Agenda

❑ Overview of data standards and identifiers used in the financial industry

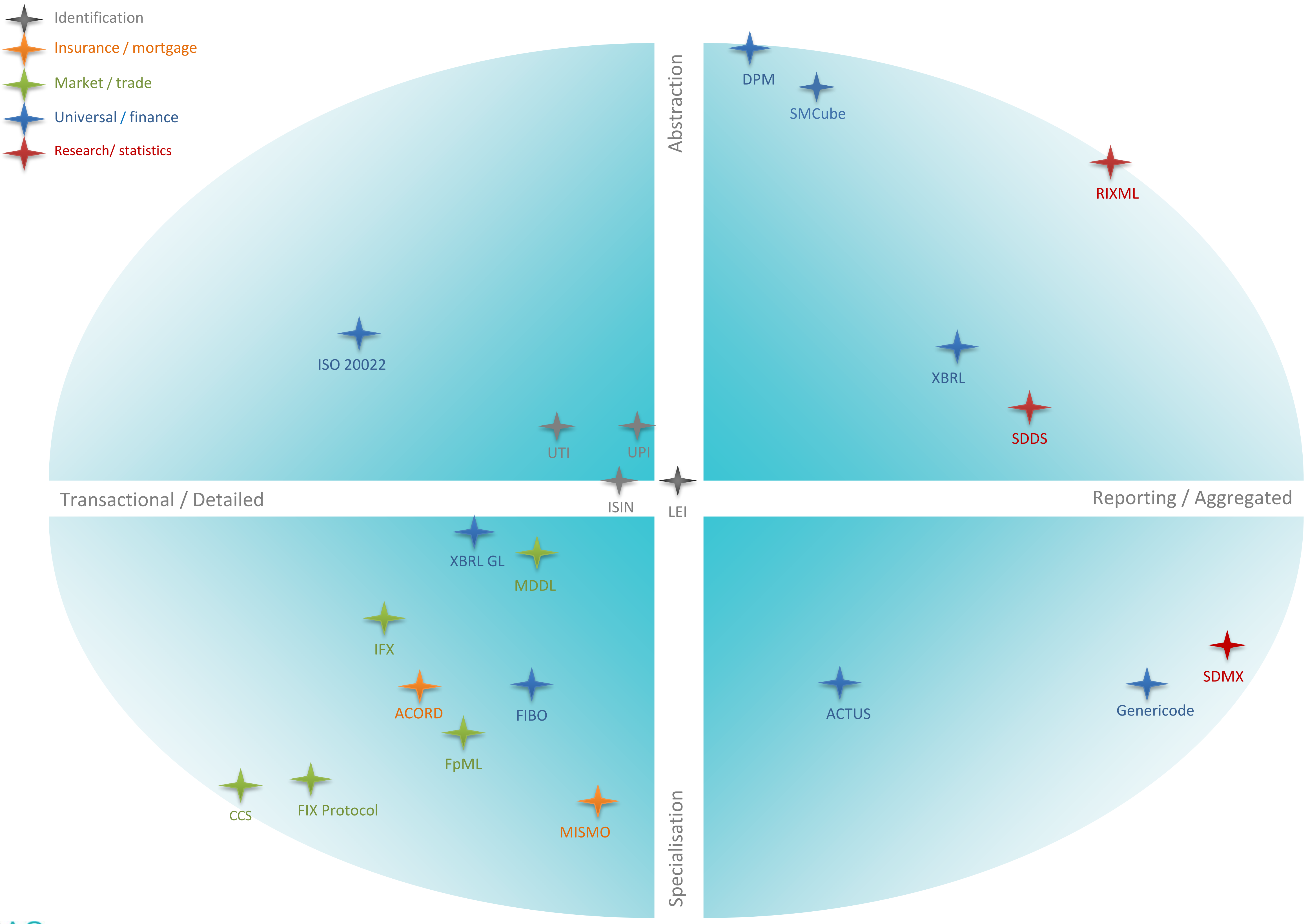
❑ Analysis of data frameworks applicable to financial institutions

❑ Verification of big data requirements

❑ Forward-thinking considerations

Central banks, financial supervisors and financial institutions operate at least several data standards and a few identifiers

Data standards and identifiers: map



BRAG

Key data standards in the financial sector include SDMX, XBRL/DPM and ISO20022 and are applicable across multitude of regulations

Key data standards: highlights

SDMX / SDMX-IM	ISO 20022	XBRL / DPM
Statistical	Transactional & business	Supervisory & business
Flows, categories, sets, code lists, concepts, keys, group keys, dimensions, attributes, measures, representations, topics	Dictionary, business process, business domains, business concepts, message concepts	Dictionary, domains, domain members, hierarchies, dimensions, concepts, facts, linkbases, links
VTL, registries	Transportation, e-Repository	Versioning, Rendering, Formula, InlineXBRL, OIM, registries

Based on the European example a typical financial regulator uses a large number of data pools stemming from variety of regulations

Financial data frameworks: overview

1. Capital Requirements Directive IV / Capital Requirements Regulation
2. Money Market Statistical Reporting
3. AnaCredit
4. Balance Sheet Items – Monetary Interest Rates
5. Securities Holding Statistics
6. European Markets Infrastructure Regulation
7. Markets in Financial Instruments Directive II / Markets in Financial Instruments Regulation
8. Securities Financing Transactions
9. Undertakings for Collective Investment in Transferable Securities
10. Alternative Investment Funds Markets Directive
11. Solvency II
12. Target 2 Securities
13. Single European Payments Area
14. Anti Money Laundering Directive IV
15. European Single Electronic Format

Individually none of these data pools falls into the category of big data analysis, but together they may constitute a data lake applicable for big data algorithms

Financial data frameworks: mix

	STANDARD	VOLUME	VARIETY	VELOCITY
CRD IV / CRR	DPM / XBRL	MIXED	MIXED	INFREQUENT
MMSR	ISO20022	GRANULAR	STRUCTURED	FREQUENT
AnaCredit	N/A	GRANULAR	STRUCTURED	INFREQUENT
BSI-MIR	SDMX	AGGREGATED	STRUCTURED	INFREQUENT
SHS	SDMX	GRANULAR	STRUCTURED	INFREQUENT
EMIR	ISO20022	GRANULAR	STRUCTURED	FREQUENT
MiFID II/MiFIR	ISO20022	GRANULAR	STRUCTURED	FREQUENT
SFT	ISO20022	GRANULAR	STRUCTURED	FREQUENT
UCITS	CUSTOM	AGGREGATED	MIXED	INFREQUENT
AIFMD	CUSTOM	MIXED	MIXED	INFREQUENT
Solvency II	DPM/XBRL	MIXED	MIXED	INFREQUENT
T2S	ISO20022	GRANULAR	STRUCTURED	FREQUENT
SEPA	ISO20022	GRANULAR	STRUCTURED	FREQUENT
AMLD IV	UNKNOWN	MIXED	MIXED	FREQUENT
ESEF	inlineXBRL	AGGREGATED	MIXED	INFREQUENT

Importantly data standards, identifiers and dictionaries provide for valuable inputs for big data algorithms: keywords, keys, links and relations

Inputs for big data algorithms

Inputs	Algorithms	Function
• **SMCube Dictionaries** • **Data Point Model Dictionaries** • **SDMX Schemas and Information Model** • **ISO20022 Business Concepts Dictionary** • **XBRL Taxonomies** • **Legal Entity Identifier** • **Universal Transaction Identifier** • **Universal Product Identifier** • **ISIN** • **Ontologies** • **…**	Levenshtein distance	Metric of minimum number of single-character edits required to change one character sequence into another.
	Damerau–Levenshtein	Variation of Levenshtein measuring number of required edits and character transpositions.
	Needleman–Wunsch	Dynamic programming Algorithm based on DNA sequence matching, adopted to character sequences.
	Bitap algorithm with modifications by Wu and Manber	Discrete test whether text contains sequence approximately equal to given pattern. Approximate equality is measured with Levenshtein of given maximum distance.
	n-gram	Statistical analysis of sequence of speech or text (syllables, letters, words …) trying to predict next element of a sequence based only on value of previous element.
	BK-tree	Configuration of character sequences similarity organized in trees based on particular metric (usually Levenshtein)
	Soundex	Phonetic algorithm for indexing words by English pronunciation. Allows words to be matched eliminating differences in spelling.

If we consider these pools jointly with variety of identifiers and potential of mash-up with other data sets the big data algorithms become even more useful

Potential applications

Case	Data frameworks	Data to mash-up
Better identify insurance patterns and claims for technical risk provisions and actuarial assessments	Solvency II	IoT (sensors) / automated information from cars / households / health
Identify suspects of AML	AMLD IV	Information from flight engines for suspicious travels / information from social media on excessive purchases
Identify potential insider trading schemes	MIFIR / EMIR / ESEF / SHS	Family and social relations from social media
Identify related borrowers of loans or relations between issuer and borrower	CRD IV [LE] / AnaCredit	Social, business and family relations from social media
Increase inflation measurement accuracy	BSI-MIR	Surveys, sentiment analysis from social media

THANK YOU

Michal Piechocki

e: michal.piechocki@br-ag.eu
m: +48505558628

Acknowledgments: Michal Skopowski

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

The use of Big Data in Central Bank of Armenia[1]

Gagik Aghajanyan, Tigran Baghdasaryan and Gor Lazyan,
Central Bank of Armenia

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

Central Bank of Armenia

Statistics department

The use of Big Data in Central Bank of Armenia

G.Lazyan, T.Baghdasaryan, G.Aghajanyan

Abstract

In this modern world, where the technology develops in a very high speed, more and more data are generated, especially by web and electronic devices. The concept of "Big data" arose in order to describe the huge amount of generated data in the world, which could be characterized with the "3 V's": volume, velocity and variety. The use of big data by policy makers and researchers became more popular and acceptable in recent years. The Central Bank of Armenia does not fall behind the other policy makers in the world and have started the collection of Big data since 2016. The main source is Internet data sets. The sources of the data are online markets, supermarkets, service providers, realty agencies, employment agencies, etc. At this moment the main goals of using this information is to follow real-time price dynamics of goods and services in the market (food, non-food, services) and get flash estimates of the CPI, track housing prices in order to explore real estate market and compute housing price index and finally estimate the demand in labor market by industries of economy and type of occupation. The technology used to access the data is web scraping.

Keywords: Big Data, price statistics, real estate market, labor market, central banks

Contents

Introduction

In this pace of technological progress, more and more data are generated from web and various devices and sensors. This vast amount of generated data is known as “Big Data”. It could be described by “3V”s: volume, variety, velocity. Every single second and almost everything around us is generating data: mobile phones, computers, smart watches, cameras, Facebook and twitter posts and so on. These vast amounts of data are very different from each other and generally unstructured. Some authors may add another “V”s, like veracity which refers to the accuracy of the data, valence – the connectedness of the data, variability, value, etc.

Nowadays big data is widely used in both business and public policy decisions. Many governments and central banks around the world already use big data statistics in their policies.

The Central Bank of Armenia (CBA) have started the collection of big data since 2016. The main objectives are to generate flash estimate of consumer price index (CPI), examine real estate market and use the outcome in the estimation of Housing price index in the future, and to get estimates of labor demand. The main sources of data are supermarkets’, realty agencies’ and labor organizations’ websites.

We will first discuss the main challenges of CBA, and then will turn to the sources of data, methodology and the results. Lastly, we will talk about main limitations that the CBA met and future contributions to this paper.

The main challenges for Central Bank of Armenia

As we have mentioned in the introduction one of the main objectives of big data use is to get a flash estimate of CPI. The National Statistical Service of the Republic of Armenia (NSSRA) publishes preliminary CPI in the 3rd working day after the reference month.[1] As an authority, responsible for price stability, CBA needs the estimates of CPI as soon as possible in order to implement an effective forward-looking policy. Understanding this and taking into account the wide coverage of modern statistics and data availability, having resources and an opportunity as well the Statistics department took up the matter to collect price data from supermarkets’ websites and compute flash estimates of CPI.

The next challenge is important from the point of view of both monetary policy and financial stability. Housing price index (HPI) is widely used in the world, although it is not available yet in Armenia. Thus, we decided to try to analyze real estate market in order to get some trends and characteristics of housing market that will also help in the more precise estimation of the HPI.

[1] National statistical service of the Republic of Armenia

Finally, the last, but not least, challenge is the estimation of labor market demand. It could be used as a leading indicator for business sectors. NSSRA collects a good statistics on labor market, and do have indicators describing the demand of labor market, but it takes one month to get the data and it is grouped by industries, not by profession.

These are only the challenges the Statistics department set in 2016 for the first time, and in case we succeed, the scope of the challenges will get larger in the future.

Data and methodology

In this section, we will talk about data and the main results of our analysis based on this data. We have used only data from internet datasets, using web-scraping tools with different intervals (generally daily data) depending the data availability and update interval.

Consumer price index

In order to get flash estimates of CPI, we started the collection of price statistics in 2016 using supermarkets data, which have an online web store. After the research both on internet and business register, we found only two supermarkets, which have listed all its products in its website. One of them has about 4000 listed items and another one about 9000.[2] For our analysis, we took only one supermarket data, which has more listed items. The interval in which we scrape the data is 10 days, so that we have price data for every 10th, 20th and 30th day of a month. We tried to get close to the methodology used by NSSRA for the computation of CPI as much as possible. Based on this methodology, 470 goods and services are included in the basket to compute the CPI, which divided into 3 main groups: food, non-food and services, with corresponding weights. Seasonal products (food) have a large share in Armenian consumer basket and usually play a "key role" in CPI fluctuations. Therefore, we took only food products and created indices of each subgroup of products to the previous period, as we only had a one-year data. In the first step, we have created indices to the December of previous year for each product, based on the daily price data. Then weights (set by the NSSRA) were applied in order to get the price index of each group of products and divided by the corresponding index in order to get price indices over the previous month. The graphs in Appendix A show the comparison of the official (published by the NSSRA) and the computed indices based on our analysis.

As we can see from the graphs, although errors from the official index are sometimes larger, the overall trends are pretty much the same.

Further contribution to this challenge would be a computation of CPI using different weighting methods like principal components analysis or factor analysis, data envelopment analysis etc.[3]

[2] We scrape the information about the name (including the item name, type and weight) and price.

[3] "Handbook on constructing composite indicators: methodology and user guide", OECD, 2008

Real estate market

Monitoring the trends in real estate market plays a vital role from central banks perspectives. Changes in real estate market may have an important impact on the monetary policy transmission mechanism and may have influence on aggregate demand and inflation. However, there is no reliable and representative indicator of housing price changes in Armenia. Developing such a measure is very challenging and one of the main problems is the availability of data.

Starting from January 2017, we scrape the web page of one of the largest realty agencies in the market, which includes more than 8000 announcements. The main advantage of the collected data is that it is very detailed; there are more than 40 properties of apartments:

- Full address, including city, region and district
- Number of Rooms and bathrooms
- Area in sq.m.
- Ceiling height
- Floor/total floors
- Building type (monolit, panel, stone, other)
- Condition (newly repaired, good, zero condition)
- Available facilities (gas, water, heating, view, close to a bus station, etc.)
- Price

This will enable us to construct a hedonic price index. However, indices based on advertised prices have a major drawback. Houses can be withdrawn from market and the agreed selling price may not equal the seller's asking price.[4] Taking into account the mentioned drawbacks the data collected might have some use at least being considered as an indicator of housing supply price.

The main idea of hedonic regression method is that heterogeneous goods can be described by their attributes or characteristics. That is, a good is essentially a bundle of (performance) characteristics. In the housing context, this bundle may contain attributes of both the structure and the location of the properties. There is no market for characteristics, since they cannot be sold separately, so the prices of the characteristics are not independently observed. Price index is calculated using dummy variable hedonic model.

$$ln\, p_n^t = \beta_0 + \sum_{t=1}^{T} \delta^t D_n^t + \sum_{k=1}^{K} \beta_k z_{nk}^t + \varepsilon_n^t$$

Where β_0 and β_k are the intercept term and the characteristics parameters to be estimated.

The time dummy variable D_n^t has the value 1 if the observation comes from period t and 0 otherwise.

[4] "Handbook on Residential Property Prices Indices (RPPIs)" European Union, International Labor Organization, International Monetary Fund, Organization for Economic Co-operation and Development, United Nations Economic Commission for Europe, The World Bank, 2013

According to the results obtained, the price index of February 2017 compared to the previous month is equal 100.15%.[5] In comparison, the preliminary estimations of price index computed based on the State Committee of Real Estate Cadastre of RA data is 100.35%.

Labor market

Labor statistics is very developed in Armenia, and NSSRA collects really a wide range of information and has many indicators describing labor market. However, sometimes for policy makers it is necessary to get the data much earlier than it is published. That is why we started the collection of job announcements in a daily basis. We use web-scraping technics again to get the data of online job announcements. The information include the position to which a person is applying, the industry the company is in, the starting date and deadline for application, and approximate salary (not always shown). This data could be useful in creating a leading indicator of each business sector.

Limitations

On the way of achieving our goals, we faced some limitations, which delayed or created obstacles, which need to be solved or avoided. Here are some of them:

- Not too much online supermarkets. The price sources are not too much, and this fact has a negative impact on reliability of the data. It is very difficult to follow the price dynamics of a particular good or a group, since it always changing: you may not find the item in next two months or six months later, and it is very difficult to find a substitute one. The problem may also be behind the website or web scraping tool.

- No Application-programming interface (API) available. During this time, no any supermarket or agency provided an API in order to access the data easily. This is one of the main technical issues in our way.

- Sometimes data is not much reliable. This mainly refers to online announcements. Anyone could create an announcement on the website and moderators not always concerned about it and just move forward into the website. Sometimes one could see an apartment for only 10$. These types of data should be considered during the analysis.

Further contribution

As a further contribution for big data usage in CBA, we can mention the followings:

- Further and deep analysis of price data in order to get an indicator comparable with CPI

- Implementation of the methodology of Housing price index in Armenia based on the data from realty agencies

[5] See full Stata output table in Appendix B

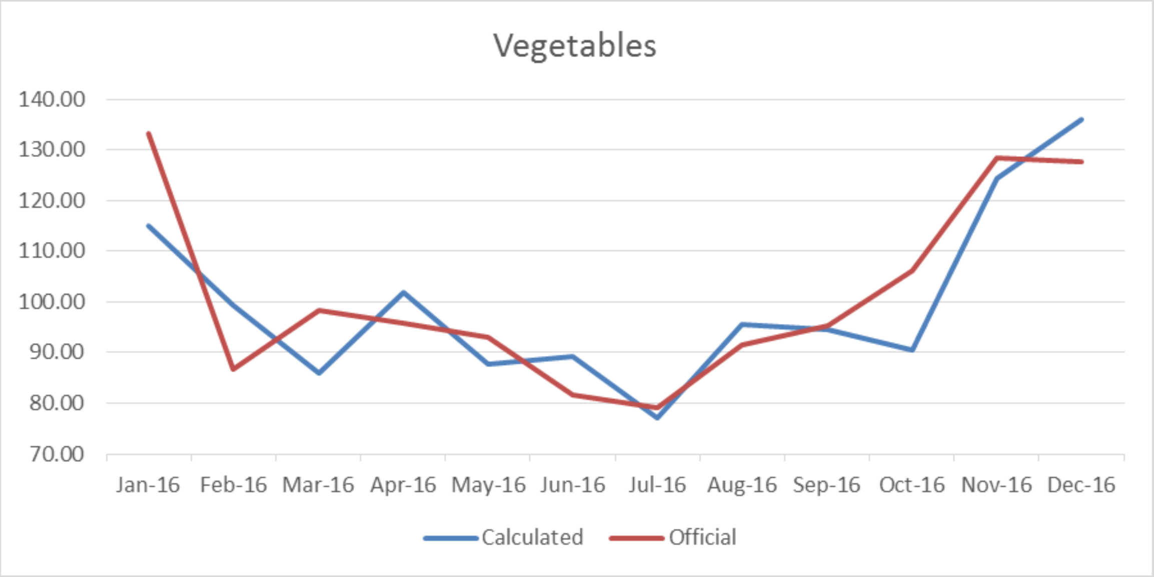
- Widen the range of the usage of big data in Armenia in terms of administrative and social media sources

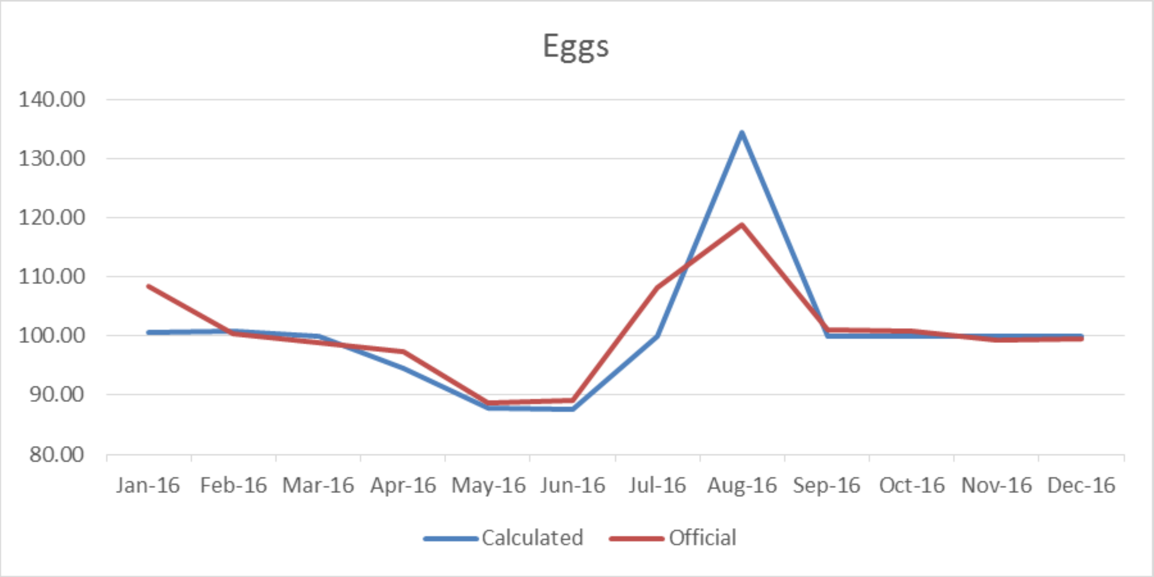
Conclusion

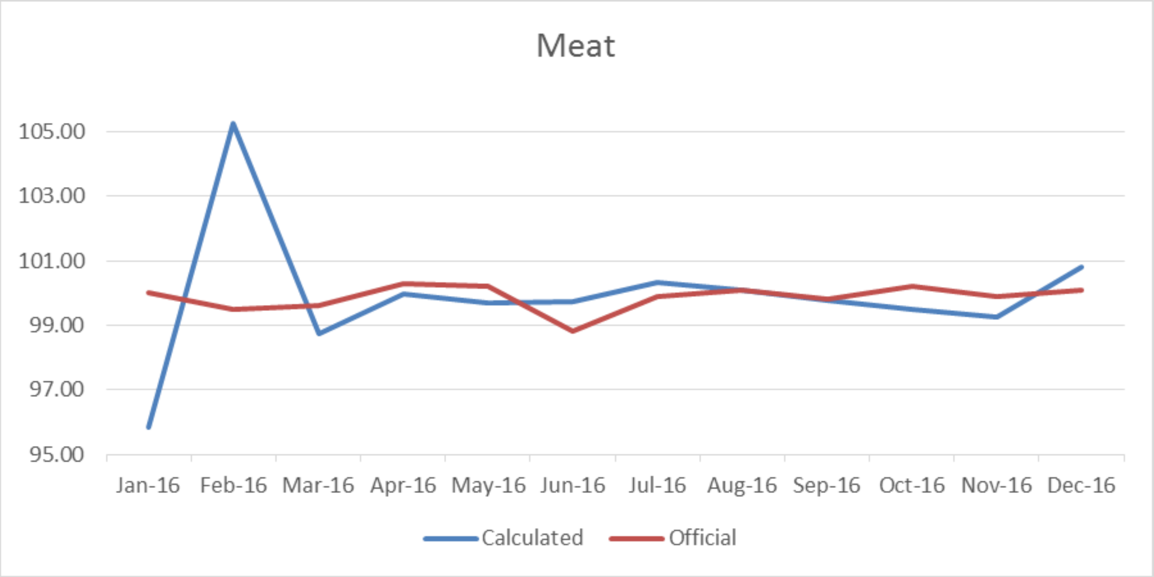
Concluding we can say that nowadays the use of big data is not only useful but it is already necessary for central banks, governments and other authorities and national statistical services, of course parallel with the official statistics. Big data is not yet substitute but complementary for the official statistics and it requires new technics and methodology other than traditional. It could be used in different aspects and for different purposes: in tourism statistics, traffic and transport statistics, price statistics and also in consumer behavior via social media.

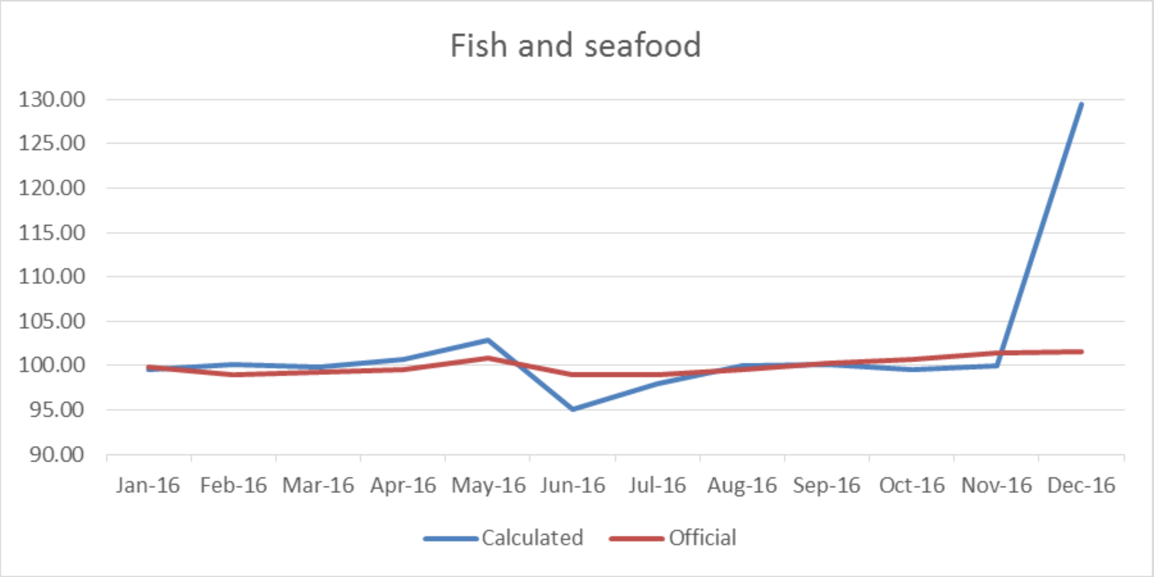
Appendix

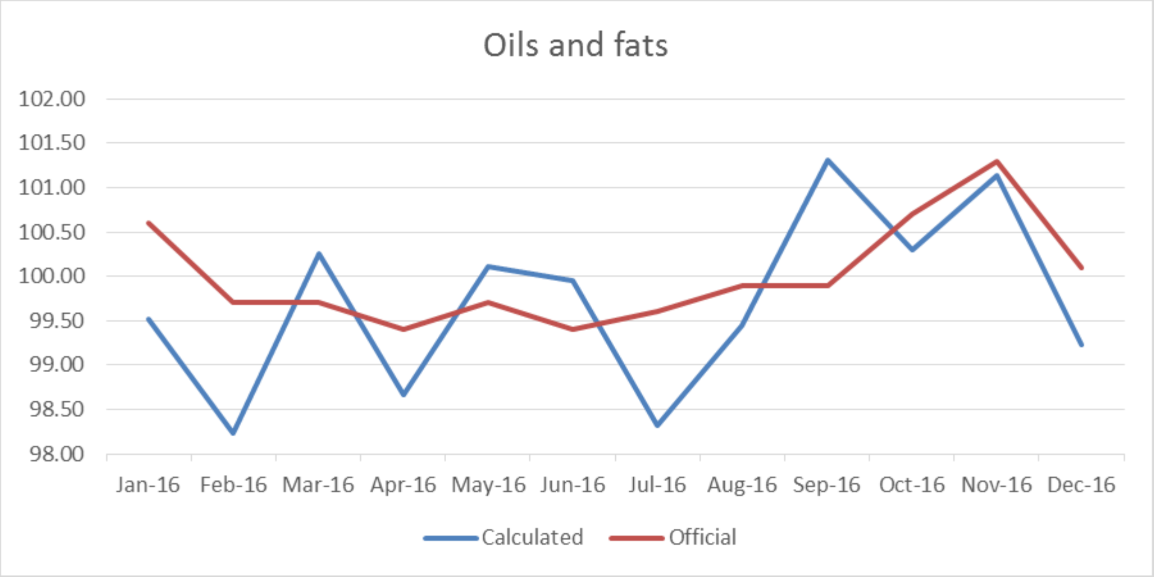
Appendix A

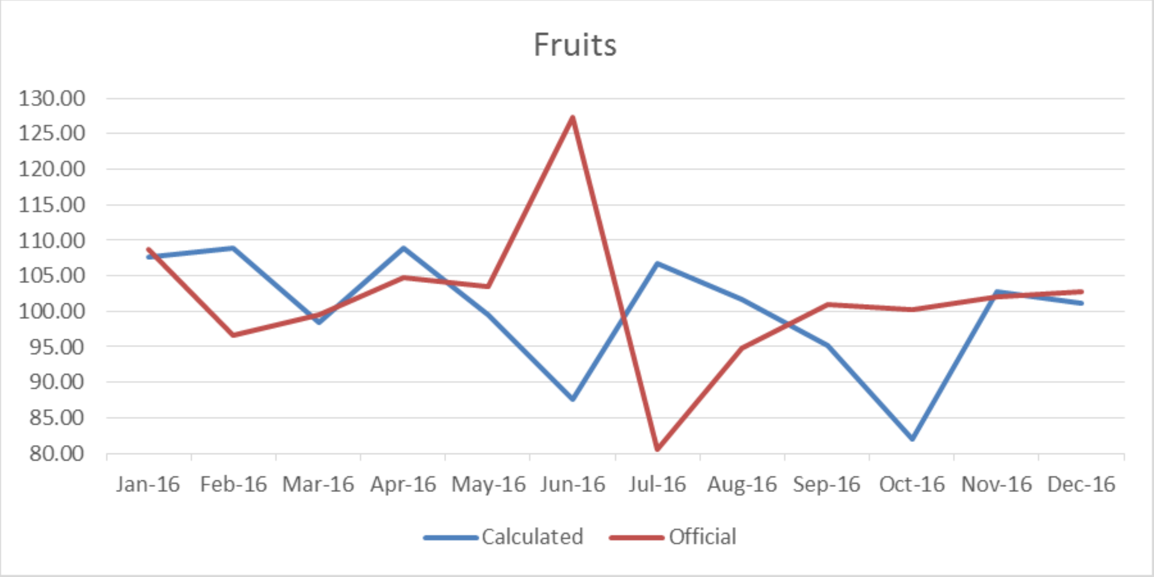


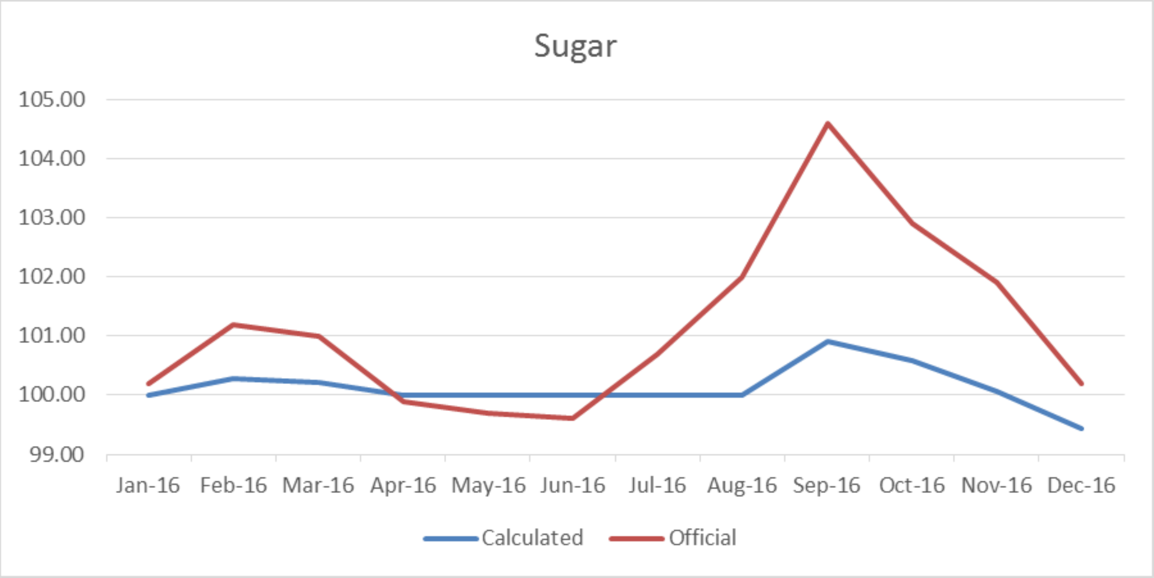

Eggs
140.00
130.00
120.00
110.00
100.00
90.00
80.00
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Oct-16
Nov-16
Dec-16
Calculated
Official

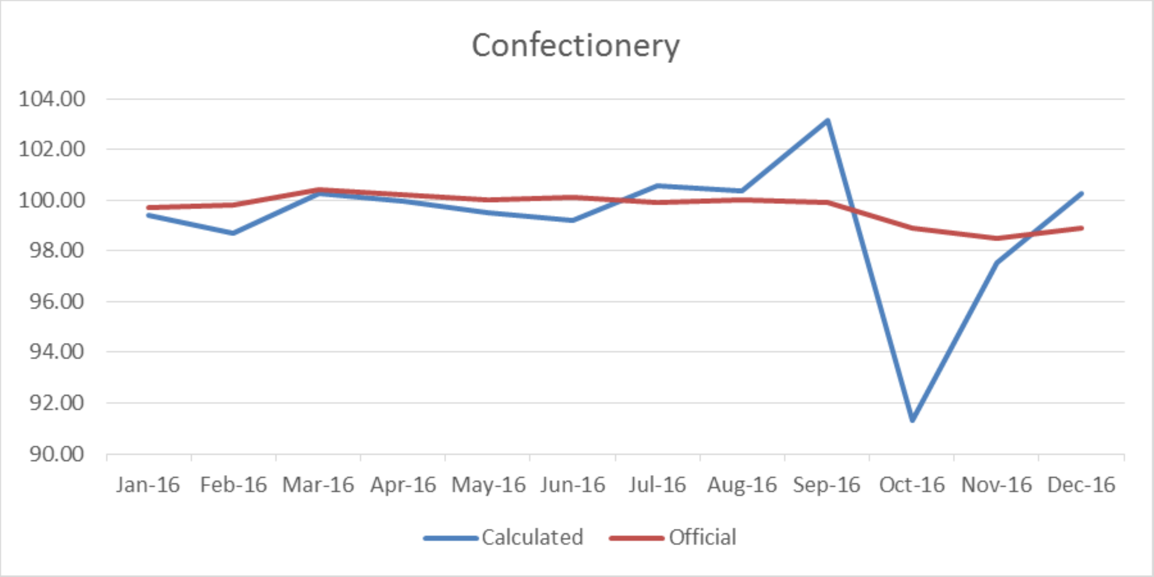

Meat
105.00
103.00
101.00
99.00
97.00
95.00
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Oct-16
Nov-16
Dec-16
Calculated
Official


Fish and seafood
130.00
125.00
120.00
115.00
110.00
105.00
100.00
95.00
90.00
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Oct-16
Nov-16
Dec-16
Calculated
Official

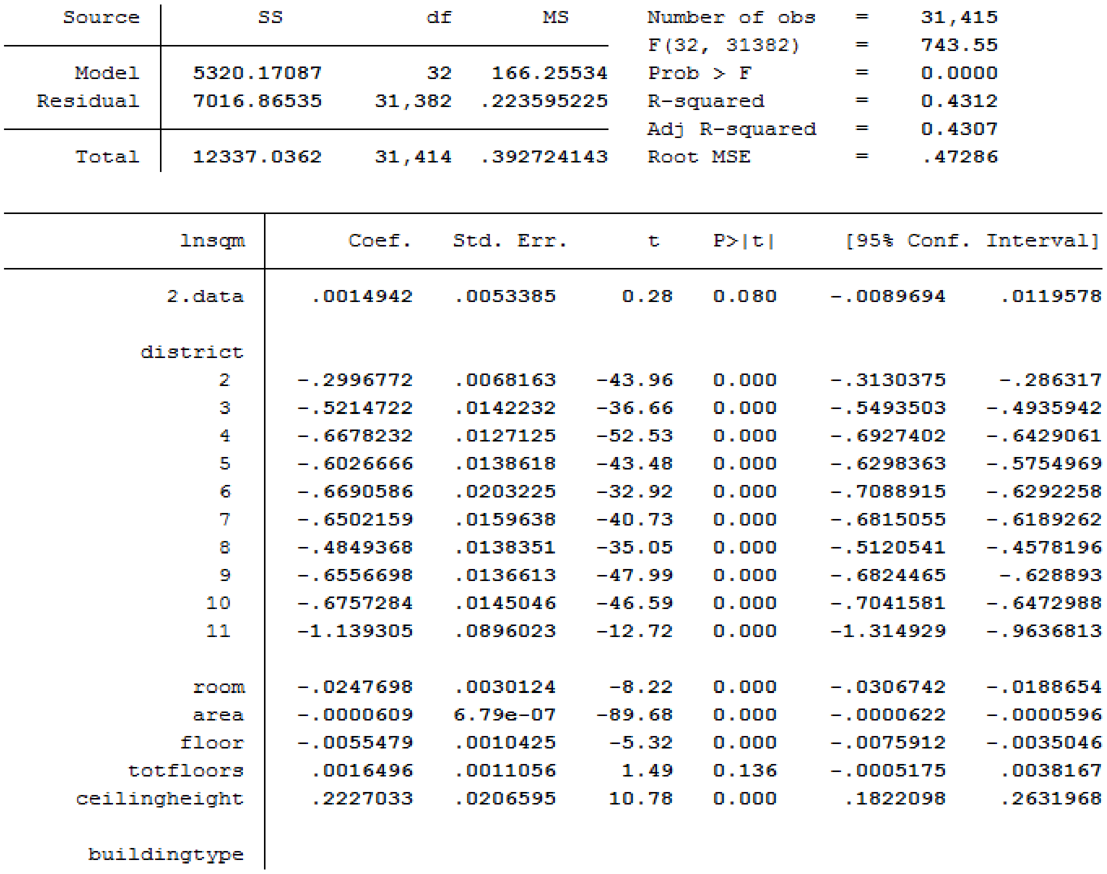

Oils and fats
102.00
101.50
101.00
100.50
100.00
99.50
99.00
98.50
98.00
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Oct-16
Nov-16
Dec-16
Calculated
Official


Fruits
130.00
125.00
120.00
115.00
110.00
105.00
100.00
95.00
90.00
85.00
80.00
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Oct-16
Nov-16
Dec-16
Calculated
Official


Sugar
105.00
104.00
103.00
102.00
101.00
100.00
99.00
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Oct-16
Nov-16
Dec-16
Calculated
Official

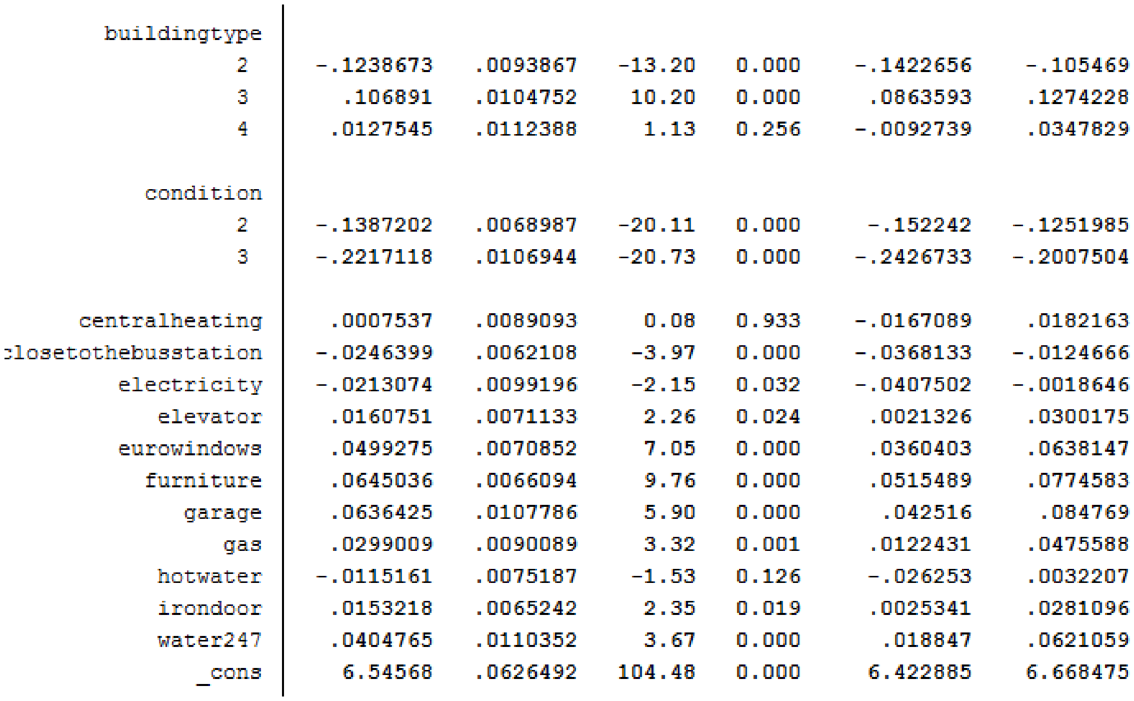

Confectionery
104.00
102.00
100.00
98.00
96.00
94.00
92.00
90.00
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Oct-16
Nov-16
Dec-16
Calculated
Official

Appendix B



```
      Source |       SS           df       MS      Number of obs   =    31,415
-------------+----------------------------------   F(32, 31382)    =    743.55
       Model |  5320.17087        32  166.25534   Prob > F        =    0.0000
    Residual |  7016.86535    31,382  .223595225   R-squared       =    0.4312
-------------+----------------------------------   Adj R-squared   =    0.4307
       Total |  12337.0362    31,414  .392724143   Root MSE        =    .47286

------------------------------------------------------------------------------
       lnsqm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      2.data |   .0014942   .0053385     0.28   0.080    -.0089694    .0119578
             |
    district |
          2  |  -.2996772   .0068163   -43.96   0.000    -.3130375    -.286317
          3  |  -.5214722   .0142232   -36.66   0.000    -.5493503   -.4935942
          4  |  -.6678232   .0127125   -52.53   0.000    -.6927402   -.6429061
          5  |  -.6026666   .0138618   -43.48   0.000    -.6298363   -.5754969
          6  |  -.6690586   .0203225   -32.92   0.000    -.7088915   -.6292258
          7  |  -.6502159   .0159638   -40.73   0.000    -.6815055   -.6189262
          8  |  -.4849368   .0138351   -35.05   0.000    -.5120541   -.4578196
          9  |  -.6556698   .0136613   -47.99   0.000    -.6824465    -.628893
         10  |  -.6757284   .0145046   -46.59   0.000    -.7041581   -.6472988
         11  |  -1.139305   .0896023   -12.72   0.000    -1.314929   -.9636813
             |
        room |  -.0247698   .0030124    -8.22   0.000    -.0306742   -.0188654
        area |  -.0000609   6.79e-07   -89.68   0.000    -.0000622   -.0000596
       floor |  -.0055479   .0010425    -5.32   0.000    -.0075912   -.0035046
   totfloors |   .0016496   .0011056     1.49   0.136    -.0005175    .0038167
ceilingheight |   .2227033   .0206595    10.78   0.000     .1822098    .2631968
             |
buildingtype |
```



```
 buildingtype |
           2  |  -.1238673   .0093867   -13.20   0.000    -.1422656    -.105469
           3  |    .106891   .0104752    10.20   0.000     .0863593    .1274228
           4  |   .0127545   .0112388     1.13   0.256    -.0092739    .0347829
              |
    condition |
           2  |  -.1387202   .0068987   -20.11   0.000     -.152242   -.1251985
           3  |  -.2217118   .0106944   -20.73   0.000    -.2426733   -.2007504
              |
centralheating |   .0007537   .0089093     0.08   0.933    -.0167089    .0182163
closetothebusstation |  -.0246399   .0062108    -3.97   0.000    -.0368133   -.0124666
  electricity |  -.0213074   .0099196    -2.15   0.032    -.0407502   -.0018646
     elevator |   .0160751   .0071133     2.26   0.024     .0021326    .0300175
 eurowindows |   .0499275   .0070852     7.05   0.000     .0360403    .0638147
    furniture |   .0645036   .0066094     9.76   0.000     .0515489    .0774583
       garage |   .0636425   .0107786     5.90   0.000      .042516     .084769
          gas |   .0299009   .0090089     3.32   0.001     .0122431    .0475588
     hotwater |  -.0115161   .0075187    -1.53   0.126     -.026253    .0032207
     irondoor |   .0153218   .0065242     2.35   0.019     .0025341    .0281096
     water247 |   .0404765   .0110352     3.67   0.000      .018847    .0621059
        _cons |    6.54568   .0626492   104.48   0.000     6.422885    6.668475
```

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

The use of Big Data in Central Bank of Armenia[1]

Gagik Aghajanyan, Tigran Baghdasaryan and Gor Lazyan,
Central Bank of Armenia

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

The use of Big Data in Central Bank of Armenia

Gor Lazyan
Gagik Aghajanyan

Statistics department
Central Bank of Armenia

IFC-BI Satellite Seminar on Big Data
Bali, Indonesia
21 March 2017

Outline

1 Introduction

2 The main challenges of CBA
- Consumer price index (CPI)
- Housing price index (HPI)
- Labor market

3 Limitations

4 Further contributions

Introduction

Characteristics of Big data

- Volume
- Variety
- Velocity
- Veracity
- Valence
- Value

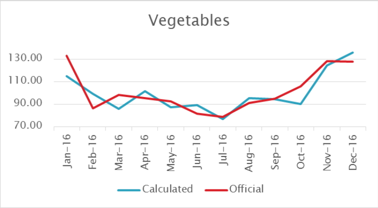
The main challenges of CBA

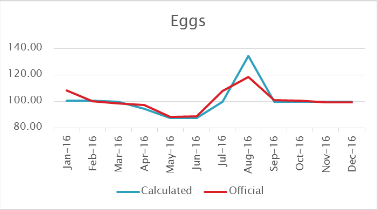
- Estimation of CPI based on supermarkets price database
- Estimation of Housing price index based on online announcements
- Labor demand analysis

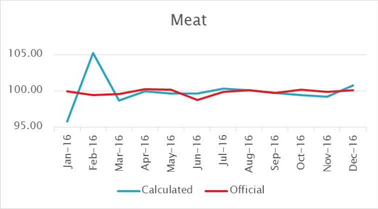
Consumer price index (CPI)

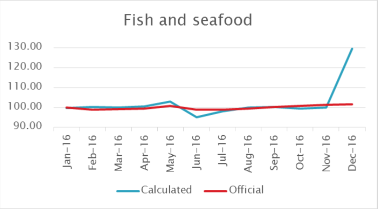
- On-line supermarkets
- 10 days interval, starting from 2016
- More than 9000 products
- Food products
- Over previous period

Consumer price index (CPI)









Housing price index (HPI)

Daily data of on-line announcements (supply), from 2016

- Full address, including city, region and district
- Number of Rooms and bathrooms
- Area in sq.m.
- Ceiling height
- Floor/total floors
- Building type (monolit, panel, stone, other)
- Condition (newly repaired, good, zero condition)
- Available facilities (gas, water, heating, view, close to a bus station, etc.)
- Price

Housing price index (HPI)

Dummy variable hedonic model

$$\ln p_n^t = \beta_0 + \sum_{t=1}^{T} \delta^t D_n^t + \sum_{k=1}^{K} \beta_k z_{nk}^t + \varepsilon_n^t$$

The price index in February 2017 compared to the previous month is **100.15%***

*preliminary estimations of price index computed based on the State Committee of Real Estate Cadastre of RA data is 100.35%

Labor market

Daily data of on-line announcements (demand), from 2016

- Position
- Industry
- Starting date
- Deadline
- Salary (not always shown)

Limitations

- Data range
- Number of available on-line supermarkets and on-line announcements websites
- Data accuracy

Further contributions

- Deep analysis of price data and applying this technique to the other product groups of CPI
- Computation of CPI using different weighting methods like factor analysis
- Implementation of the methodology of Housing price index in Armenia based on the data from realty agencies
- Creation of labor demand index which could be used as leading indicator for each industry
- Widen the range of the usage of big data in Armenia in terms of administrative and social media sources

Thank you

Questions?

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Price information collected online and short-term inflation forecasts[1]

Isaiah Hull, Marten Löf and Markus Tibblin,
Sveriges Riksbank

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

Price information collected online and short-term inflation forecasts

Isaiah Hull, Mårten Löf and Markus Tibblin[1]

Abstract

Forecasting short-term inflation developments (e.g. inflation over the coming months) is important for a central bank. There are certain elements within the published inflation figures that are volatile and inherently hard to forecast even in the short-run. Fruit and vegetable prices, energy prices and air travel prices are examples of product groups within the inflation measure that historically have held a high degree of volatility in Sweden. An automatic internet data collection process was developed to collect sales prices daily for selected fruits and vegetables from a number of Swedish online retailers. The results indicate that the information from the daily data could increase the precision in short-term inflation forecasts in Sweden.

Keywords: inflation, forecasting, big data

JEL classification: L11, L81, E30, C55

1. Short-term inflation forecasts matter to central banks

The objective for monetary policy at the Riksbank, like many other central banks, is to maintain price stability. This is interpreted as keeping inflation, i.e. consumer prices measured through an index based on a basket of goods and services households tends to consume, low and stable.[2] The consumer price index (CPI) consists of a range of sub-categories. In Sweden, fruit and vegetables, air travel and fuel prices are examples of sub-categories, which historically have shown a relatively high level of price volatility. This volatility may at times create challenges in forecasting CPI in the short-term (e.g. the coming months) although these sub-categories sum to only a small share of total CPI. Short-term CPI volatility and related short-term forecast errors may seem like a minor issue, given the fact that monetary policy decisions are based on inflation development over a longer time horizon. However, accurate forecasts of future inflation are dependent on precise information regarding current inflation, as well as good forecasts of short-term inflation developments. It is important from a forecasting, and also a monetary policy, perspective to be able to decide to what extent a large deviation between forecast and outcome is due to temporary or more permanent factors. A large forecast error that is due to permanent

[1] The authors work at the Monetary Policy Department of the Riksbank. The opinions here are the sole responsibility of the authors and should not be viewed as reflecting the views of the Riksbank. The authors would like to thank colleagues at the Riksbank for their valuable comments on previous drafts.

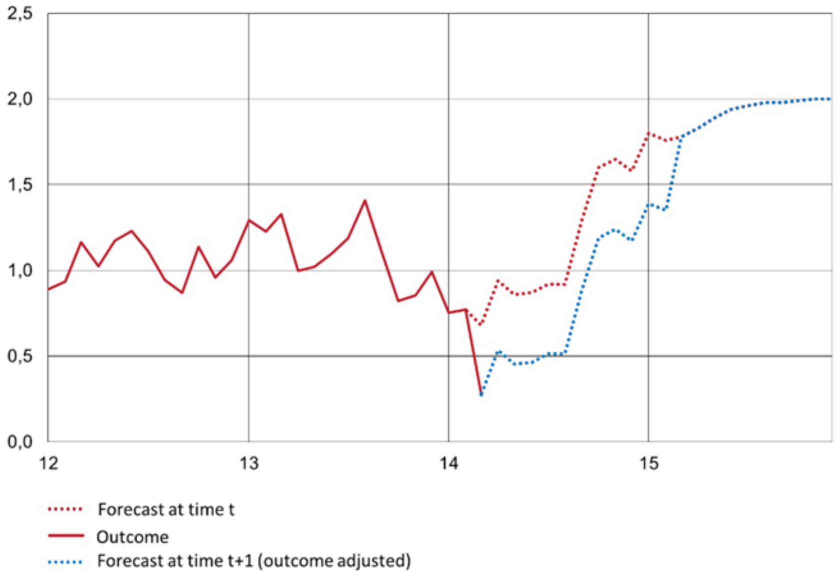
[2] http://www.riksbank.se/en/Monetary-policy/

factors can affect the path of the new inflation forecast over the next twelve months, assuming that no other changes will occur, see Figure 1.

Inflation outcome and forecast revision due to short-term forecast error

Yearly percentage change

Figure 1



Note: The figure illustrates how a forecast error at time t (difference between the solid and the dashed red line) may affect the forecast path for inflation in the coming year (blue dashed line).

2. New data and new analytical methods

The information published on the Internet is growing very fast. It has been widely discussed how, for example, central banks can make use of new methods to collect and analyse this type of data. The Riksbank arranged a Big Data workshop in September 2015, at which a number of central banks, researchers and private firms laid out examples where new data and methods arising from Big Data may support analysis and decision-making at central banks.[3] One area of growing interest among central banks is the increasing amount of sales price information available online. Leading research in this area has been conducted in the "Billion prices project". This project was initiated by Cavallo and Rigobon, who have in a number of studies shown that a price index produced using online price data follows the official measures of consumer prices such as CPI fairly well (Cavallo and Rigobon, 2014). Also, through discussions with retailers in relation to the Riksbank's Business survey it is clear that there are in general small differences in consumer prices between offline and online prices in Sweden,[4] which is also in line with a large international comparison study conducted by Cavallo (2016). Furthermore, it has been shown that online pricing data

[3] http://www.riksbank.se/en/Press-and-published/Notices/2015/The-Riksbank-organises-a-workshop-on-big-data/

[4] http://www.riksbank.se/Documents/Rapporter/Foretagsintervjuer/2016/rap_foretagsundersokning_160615_eng.pdf

performs well in forecasting CPI produced by national statistical offices and for some countries outperform models that include offline data (Aparicio and Bertolotto, 2016).

3. The pilot project –collecting online fruit and vegetable prices from the internet

Utilising online price data may potentially overcome a number of current issues in short-term inflation forecasts. First, there are no time lags in data collection, as online prices can be collected in real-time i.e. price data are available before official inflation figures are published. Moreover, collecting price data online generates information at a very granular level. This enable a more detailed analysis, which for example could give information on whether an unusual price development is temporary or not. Also, as online data collection can be automated, daily price indices can be produced and included in models at a low cost and with very limited resources.

A small pilot study was initiated to investigate whether prices of fruit and vegetables that are available online could improve the accuracy of short-term inflation forecasts in Sweden. Prices of fruit and vegetables make up 3 percent of the CPI basket. Although these prices constitute a small part of the basket, strong price movements in this sub-index can have a clear impact on the aggregated figures. The pilot project was set up with the aim to:

1. Create a process for automatic online data collection collecting online price data daily for a few selected fruit and vegetables from Swedish retailers with e-commerce. Online prices have been collected for oranges, bananas, peppers, apples, cucumbers, cabbage, grapes, cauliflowers, pears, leeks and tomatoes. The correlation between the subset of selected fruit and vegetable prices and the total index of fruit and vegetables in the CPI is fairly high, see Figure 2 below. Hence, the subset of prices captures the variation in the total index for fruits and vegetables in the CPI quite well. It was therefore assumed that internet data for selected products would be useful when forecasting the overall CPI-index for fruit and vegetables.

2. Create a weighted monthly price index based on observed price changes online for the selected fruit and vegetables mentioned in 1.

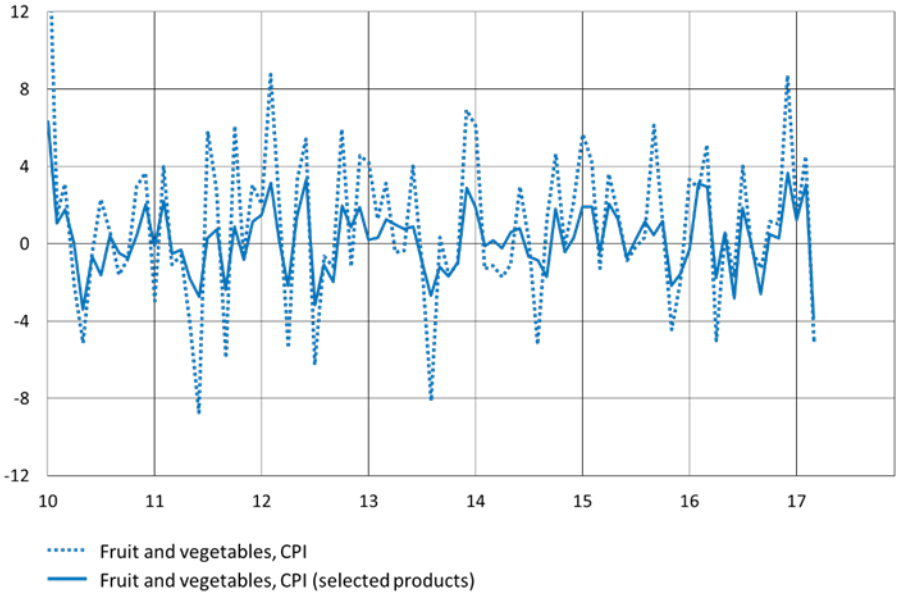
3. Test whether the constructed online price index adds value to the existing short-term forecasting models for fruit and vegetables currently used by the Riksbank.

The rest of the paper is outlined as follows: First a brief description of the data collection method is given, then the index construction and data transformations are described. Finally, the results are presented, including a discussion and overall conclusions drawn so far from the pilot project.

Prices of selected products in CPI and the aggregate of fruit and vegetables in the CPI

Monthly percentage change

Figure 2



4. Scraping

We use a technique called scraping to collect the data for this project. Scraping involves sending programmatic requests to a website's server. The server responds by returning the underlying code that would be executed in a browser if you were to visit a given page on its website. The code's structure can then be employed to identify targeted items, such as product names and prices. We use additional tools to schedule the scraping tasks so that they are automatically performed on the targeted sites at the same time each day. We also limit the speed at which we send requests to each website to ensure that no strain is placed on their servers.

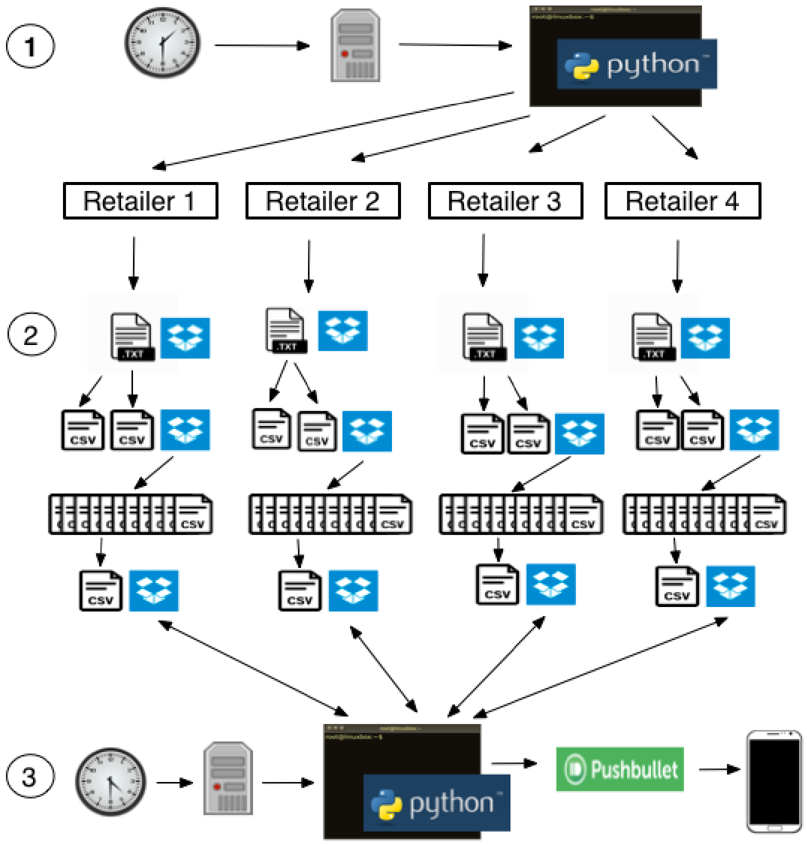
All data collection tasks are performed on a Linux virtual private server (VPS). The server executes three scripts in sequence at the same time each day, as illustrated by Figure 3. The first visits the websites of four large grocery retailers. It extracts the code from all pages related to fruit and vegetables. A list of all product prices and names is then identified in the code and saved in a .csv file for each location and day. The raw code is also saved in .txt format for 90 days, allowing us to correct errors discovered later. The script then uses regular expressions, which identify patterns in text to filter the data, creating a second .csv file for each location and day that consists only of targeted fruits and vegetables.

After the scrape is finished, the server executes the second script, which merges the filtered location and day files with past data. It first identifies matches between products in the new data and products identified in previous scrapes. Items that were not identified in previous scrapes are assigned new location-product IDs and products that already have IDs are merged with existing location-product time series.

Finally, the server executes the third script, which checks for errors. This script identifies the number of files produced, the size of those files, and the data types used in those files. The server then delivers an error report via SMS, allowing us to quickly identify and correct errors.

The daily scraping process

Figure 3



5. Forming indices

In a first step the collected online prices are sorted into different groups, orange prices in one group, and apple prices in another group and so on. The dataset is then truncated so that it exactly matches the measurement weeks in the Swedish CPI survey.[5] In the next step a geometric mean is calculated for each product and month, i.e. an average price of oranges in January, one for orange prices in February and so on.

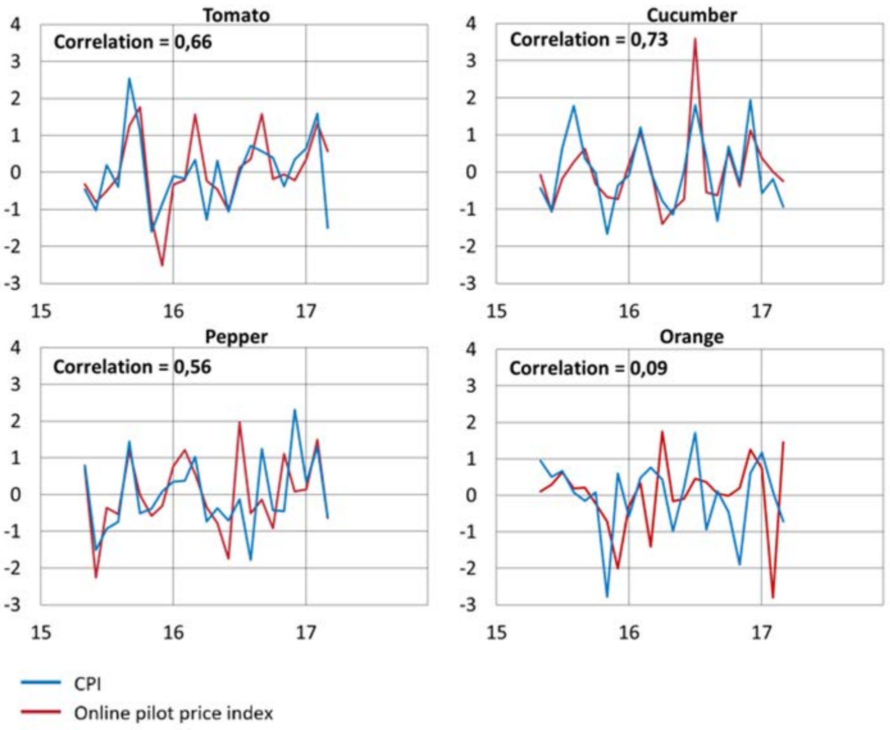
[5] Price data on fruit and vegetables in the CPI were collected for three weeks in the middle of each month (a total of 21 days) until December 2016. Thereafter, Statistics Sweden has made some changes.

Figure 4 shows the monthly percentage changes in prices for oranges, cucumbers, peppers, and tomatoes from May 2015 to March 2017. The blue lines show monthly percentage price changes according to the CPI, while the red lines show the corresponding changes based on information from the internet. The correlation between the CPI prices and prices available online are relatively high. The correlation is highest for prices on cucumbers and tomatoes. The lowest correlations are measured for prices on oranges.

Price changes according to internet data and corresponding price changes according to the CPI

Monthly percentage change May 2015 – March 2017

Figure 4



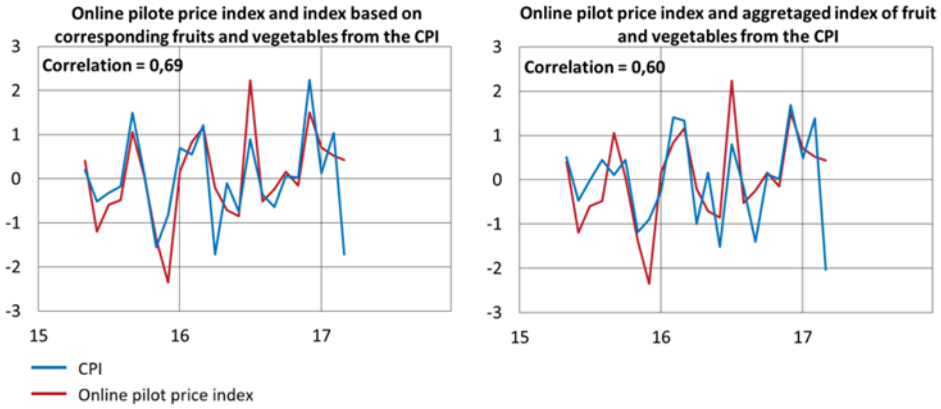
Note: The red lines indicate online data while the blue lines indicate data from official CPI

These price changes are then merged in to an overall index (henceforth denoted the online pilot price index) using the CPI weights according to CPI for the fruit and vegetable prices collected online. The left panel of Figure 5 compares the online pilot price index (red line) with a weighted CPI-index for the corresponding fruit and vegetable prices (blue line). The right panel in Figure 5 shows the online pilot price index together with the total price index for fruit and vegetables in the CPI (i.e. not only the fruit and vegetables collected in the pilot study). Naturally, the correlation decreases when comparing the online pilot price index with the total index for fruit and vegetables (left panel compared to right panel in figure 5). However, it is still fairly high and the online pilot price index captures most of the volatility in the official fruit and vegetables index.

Comparison between an aggregated index based on internet data and two different indexes with CPI data

Monthly percentage change May 2015 – March 2017

Figure 5



Note: The figure show aggregated price changes according to the data from the internet (red lines). Blue lines show price changes according to CPI. The blue line in the left hand panel show price changes if one use the same products as in the internet collection. The blue line in the right panel shows price changes according to the index for fruit and vegetables in CPI.

6. Forecast evaluation

In this section we want to test more formally whether the internet prices can be used when forecasting the index for fruit and vegetables in the CPI.

Here we compare the forecasts from the Riksbank's current models for fruit and vegetables with the forecasts generated using the internet prices. A variety of indicators are included in the models now used by the Riksbank for short-term forecasting. The models for prices of fruit and vegetables include one indicator at a time, together with a moving average of an exchange rate index (KIX). In addition, lags of the dependent variable, dummies for outliers and in some cases moving average terms are included in the models. The results are summarized as the mean forecast from all of these specifications. This approach is denoted MEAN in the evaluation below. Principal component analysis is also used to summarise the information from the indicators in the first step. The resulting Principal component indices can be seen as a weighted averages of all the indicators. In that case the models include one or more of these summary indices instead of the individual indicators. This approach is denoted PC in the evaluation below. These model-based forecasts are compared with an approach using the online price index. Here we simply use the online price index as a forecast for the aggregate of fruit and vegetables in the CPI. We denote this second approach (OP).

The evaluation period is May 2015 to March 2017. The root mean square error (RMSE) is used to compare the forecasting ability between the approaches. The RMSE summarises the standard deviation in the forecast errors and their systematic deviation. The lower the estimated RMSE, the better the forecasting ability. A forecast that is always correct has a zero RMSE.

Table 1 below summarizes the results. During this short evaluation period the best approach has been OP, where the online price index are used directly. Hence, it seems like online price information could add value in forecasting price changes for fruit and vegetables in the coming month. However, the evaluation period is very short, and longer time series are required to draw clearer conclusions.

RMSE for different approaches, May 2015-March 2017. Table 1

	Nowcasting procedures		Online prices
Horizon	Mean	PC	OP
1 month	2.0	2.0	1.6

7. Discussion

The work on developing scrapers for automatically collecting and analysing online price data has been a bit of a trial-and-error exercise. The codes developed have had to be tailored to handle a range of different, changing website layouts. However, it has been possible to build scrapers coping with changing layouts, thus making it possible to maintain a collection process with very limited resources.

The journey of analysing online micro data has just started during this pilot project. A fair amount of time has been devoted to investigating whether further transformation of the online data could increase forecasting ability. For example, an index only based on online prices showing high correlation with corresponding official CPI prices was constructed. The online price information has also been summarized using Principal Component Analysis. These types of indices have at times performed very well, but also shown large variability and overall not been as good as just the simple online index described above. Thus, no transformation is currently performed on the data. This could potentially be an area of further investigation.

Additionally, the collected daily data could also be analysed from the perspective of firms' pricing behaviour. Detailed micro data may be a source of understanding when, why and how firms change their prices. This could also be a potential area of further investigation.

8. Conclusions

The pilot project has revealed a number of insights regarding online price collection and the analysis of such data in relation to inflation forecasting.

First of all, it has been proved possible to consistently scrape online retail prices from e-commerce websites held by retailers in Sweden. On the one hand, creating scripts and IT-processes ensuring a stable data collection takes time and requires programming competence not traditionally available at central banks. On the other hand, once scrapers have been put in place, only minimum maintenance and development have been required.

Furthermore, the result so far indicates that online pricing data add some value when forecasting short-term developments of consumer prices for fruit and vegetables.

Given this outcome, there may be scope for further expanding the collection of prices available online and using them as input in short-term forecasting models.

The aggregated time series of online prices is still short and further analysis is required to ensure that the online price index have forecasting ability also in the future. The result also indicates that there may be room for increasing the collection of prices available on the internet and using them in short-term forecast models. However, there are still many questions that should be investigated, such as how to use the information in the best possible way.

References

Aparicio, Diego, & Manuel Bertolotto. (2016). “Forecasting Inflation with Online Prices.” Working Paper - MIT.

Bernanke, B. S. & Boivin, J. (2003), “Monetary policy in a data-rich environment,” Journal of Monetary Economics 50 (3), 525-546.

Bertoloto, M. & Cavallo, A. & Rigobon, R. (2014), "Using Online Prices to Anticipate Official CPI Inflation," UTokyo Price Project Working Paper Series 031, University of Tokyo, Graduate School of Economics.

Cavallo, F. Alberto, (2016), "Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers," NBER Working Papers 22142, National Bureau of Economic Research, Inc.

Stock, J. H. & Watson, M.W. (2002), “Forecasting using principal components from a large number of predictors,” Journal of the American Statistical Association 97:460, 1167-1179.

Stock, J. H. & M. W. Watson, (2004), “Combination forecasts of output growth in a seven- country data set,” Journal of Forecasting 23 (Issue 6), 405–430.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Scraped sales price information and short-term CPI forecasts[1]

Isaiah Hull, Marten Löf and Markus Tibblin,
Sveriges Riksbank

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

Scraped Sales Price Information and Short-Term CPI Forecasts

Isaiah Hull, Mårten Löf, and Markus Tibblin

Sveriges Riksbank

March 3, 2017

Background

Does Sales Price Information Scraped from the Internet Increase the Precision of Short-Term Inflation Forecasts?

Background

- Pilot project started in December 2014
- Collected price data from Swedish grocery retailers with online presence
  - Brick-and-mortar stores
  - Internet-only retailers
  - Multiple store-locations for largest retailer
- Constructed indices that are used as input for short-term inflation forecast

Background

- Focused on fruits and vegetables
  - High price variation over time
  - Difficult to forecast
- Limited scope to subset with known CPI weights
  - oranges, apples, bananas, cucumbers, peppers, tomatoes, pears, cabbage

Methods

- A server runs three scripts in sequence daily
- The first script visits a number of grocery retailers
- It identifies and collects code associated with all fruits and vegetables on the website

Methods

- The script parses the code to extract product names and prices
- It also applies a filter to generate a second file that contains only targeted fruits and vegetables
- The second script merges all daily scrape files and updates time series

Methods

- Finally the server executes third script that scans all new files to determine whether there were any errors

- The first script also maintains a 90-day rolling archive of the raw code extracted from the website, so that revisions can be made if any errors are discovered at a later date

Results

Filtered dataset

- Data stored using Dropbox
- 3000 price series
- Series means first computed for selected time window
- Means computed across items of same type

Results

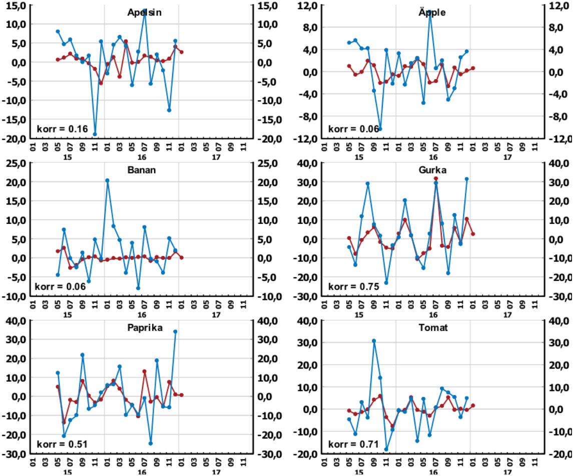
Calculate averages for price series using CPI weights:

$$index = w^{cpi}_{orange} d(p_{orange}) + ... + w^{cpi}_{tomato} d(p_{tomato}) \tag{1}$$

Use index as input to short term CPI forecast

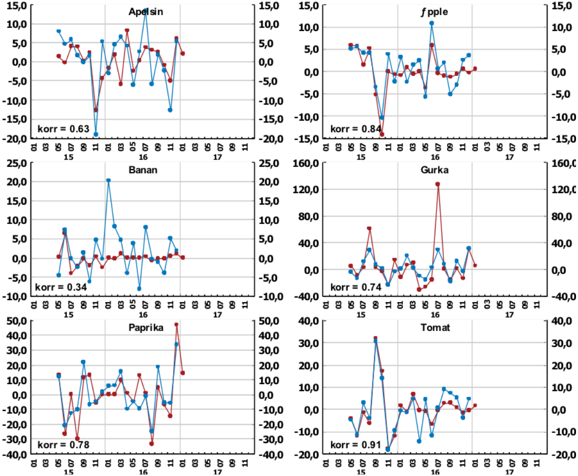
Results

CPI vs. Scraped Data



Results

CPI vs. (Adjusted) Scraped Data



Summary

- Online price collection adds value to current CPI forecasts
  - Forecast error reduced
  - Must perform additional cross validation
- Constructed robust, low-maintenance, scalable system for price collection
  - No major code changes required since completion
  - Maintenance infrequent

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Forecasting tourism demand through search queries and machine learning[1]

Rendell E. de Kort,
Central Bank of Aruba

[1] This paper was prepared for the meeting. 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 meeting.

Forecasting tourism demand through search queries and machine learning[1]

Rendell E. de Kort[2]

Abstract

This paper utilizes different machine learning techniques for tourism demand forecasting. Considering the magnitude of tourism in terms of economic contribution to Small Island Developing States (SIDS), policy making could benefit greatly from accurate tourism demand forecasting. This paper pursues a novel approach of identifying relevant search query features through google correlate and applying machine learning techniques to estimate individual source market series prior to aggregation. The prediction performance of several machine learning methods is assessed when applied to monthly tourist arrivals from individual source countries to Aruba from 1994 to 2016. The results indicate that machine learning techniques in combination with novel internet datasets sets pose great potential for achieving accurate tourism demand forecasts.

Keywords: Forecast combination, machine learning, feature selection, tourism demand forecasting, random forest, search data.

JEL classification: C22, C40, C52, C63

[1] The views expressed here are those of the author and do not necessarily reflect those of the Centrale Bank van Aruba.

[2] (R.e.dekort@cbaruba.org) Economist, Research Department, Centrale Bank van Aruba.

1. Introduction

A key challenge in many tourism destinations is the accurate forecasting of inbound tourism to support destination management decisions and to guide macroeconomic policy. The importance of the tourism industry is particularly evident in the case of Aruba, as it ranked second among tourism destinations in terms of relative contribution of travel and tourism to GDP in 2016 and where jobs in the tourism industry accounted for an estimated 89.3 percent of total employment (World Travel and Tourism Council, 2017).

However, by its very nature, tourism forecasting remains a very tricky endeavour. The sector is so unpredictable that even a small disturbance in the environment of the host country may bring down the level of demand significantly. Be it predictions about changes in the economic scenario leading to sudden inflation or deflation, any expected occurrences of hostile activities like war or terrorism, any warned natural disasters like earthquakes or floods, any likely incidences of cultural hostility or any kind of threat to public health owing to environmental imbalance or spread of some contagious diseases; all such factors have massive impact on the demand in tourism, making it almost impossible to forecast demand (O'Mahony et al., 2008).

Yet, given the significant impact of tourism on the wider Aruban economy, accurate forecasting of tourism demand is a fundamental input for key decisions on investments, as well as for gauging the conjectural situation and planning for demand flow.

Unfortunately, despite the consensus on the need to develop more accurate forecasts and the recognition of their corresponding benefits, there is no one model that stands out in terms of forecasting accuracy (Claveria et al, 2013). With recent advancements in Internet search technology, a new field has emerged (Google Econometrics), which utilizes time series data on Internet activity by obtaining correlations between keyword searches and macro-economic variables, including unemployment, tourism and consumer demand. Spurred by recent computational advancements, including refinements to the capacity to both efficiently process large volumes of data and run computationally intensive algorithms, there has been an increasing interest in machine learning techniques, including Artificial Neural Networks (ANN) and Random Forests (RF).

In the case of Aruba, tourism demand analysis faces several challenges in terms of, e.g., data availability, erratic factors and a dynamic economy that is inherently vulnerable. Destinations may be inherently vulnerable because they are open to both internal and external human and natural factors, and may have different capabilities to cope with the changes and disturbances originating from these factors (Ridderstaat, 2015). In an effort to counter some of these challenges, this paper conducts a forecasting exercise for tourism demand to Aruba by leveraging the availability of internet search data in combination with recent advances machine learning techniques.

The remainder of the paper is structured as follows. In section 2 the relevant literature is discussed while in section 3 I describe the main methodological frameworks utilized. The results are presented in section 4 and to finalize section 5 presents some concluding remarks.

2. Literature review

Search queries reflect how people show interest and attention on specific topics on the internet and has caught the attention of researchers as a potential useful source of information to model real world phenomenon (Mohebbi et al, 2011). Google provides two data sources that are useful in this context, namely Google correlate and Google Trends. While economic data is often reported with a lag of months or quarters, Google query data is available in real time. This means that queries are contemporaneously correlated with an economic time series, which may be helpful for economic 'nowcasting' (Stephens-Davidowitz and Varian, 2015). Furthermore, existing studies have demonstrated that these data can predict future trends (Choi and Varian, 2012). In the field of tourism, this development has not gone unnoticed, as the predictive power of internet searches has been explored to predict the number of visitors (Saidi et al, 2010; Li, 2016; Yang et al, 2014).

Given a temporal pattern of interest, Google Correlate provides an online, automated method for query selection which determines which queries best mimic the data (Mohebi et al., 2011). More specifically, when time series are uploaded, Google Correlate computes the Pearson Correlation Coefficient (r) between the time series of interest and the frequency time series for every query in the google database. Correlation coefficients range from r=-1.0 to r=+1.0. The queries that Google Correlate shows are the ones with the highest correlation coefficient (i.e. nearest to r=1.0) (Mohebbi, M. et al, 2011). Tourism demand modelling and forecasting studies have focused predominantly on tourist arrivals as proxy for tourism demand (Song and Li, 2008). However, the literature has presented at least three classes of tourism models, namely, those explaining the tourist expenditure, tourist arrivals and length of stay. The most accepted measure of tourism demand is tourism expenditure (Ahmed, 2013). For this study, tourism receipts are utilized since it is available and provides a closer proxy to what tourists contribute to the economy in monetary terms. The literature suggest that tourism demand very often exhibit patterns in term of seasonal, cyclic and trend components (Cankurt and Subasi, 2015). This is a challenge to traditional forecasting techniques to which machine learning could potentially aid. Also, real-time macroeconomic data are typically incomplete for today and the immediate past ('ragged edge') and subject to revision. To enable more timely forecasts, the 'ragged edge' issue can be framed as a standard "nowcasting" problem and addressed in similar fashion to the nowcasting framework of the Centrale Bank van Aruba, as outlined in Zult and Schreuder (2011).

In terms of techniques, compared to econometric models, machine learning based approaches count on several significant advantages, particularly when modelling large data sets. Machine-learning techniques are gaining ground among econometricians, and are particularly well suited to the nowcasting problem. Traditionally, econometrics and machine learning have focused on different types of problems, and have developed separately. Econometrics has generally focused on explanation, with particular attention to issues of causality, and a premium placed on models that are easy to interpret. A "good" model in this framework is mostly assessed on the basis of statistical significance and in-sample goodness-of-fit. Machine learning, on the other hand, has focused more on prediction, with emphasis instead on a model's accuracy rather than its interpretability. A "good" machine-learning model, then, is often determined by looking at its likely out-of-sample success, based on bootstrap-style simulation techniques (Tiffen, 2016).

Another interesting insight that has emerged from the machine learning literature is that averaging over many small models tends to give better out-of-sample prediction than choosing a single model (Varian, 2013). Furthermore, there has been an increasing interest in Artificial Neural Networks (ANN) due to controversial issues related to how to model the seasonal and trend components in time series and the limitations of linear methods (Claveria et al, 2013). In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyse data used for classification and regression (Li, 2016). Machine learning models are also deemed superior in recognizing and learning the seasonal patterns without removing them from the raw data (Cankurt and Subasi, 2015).

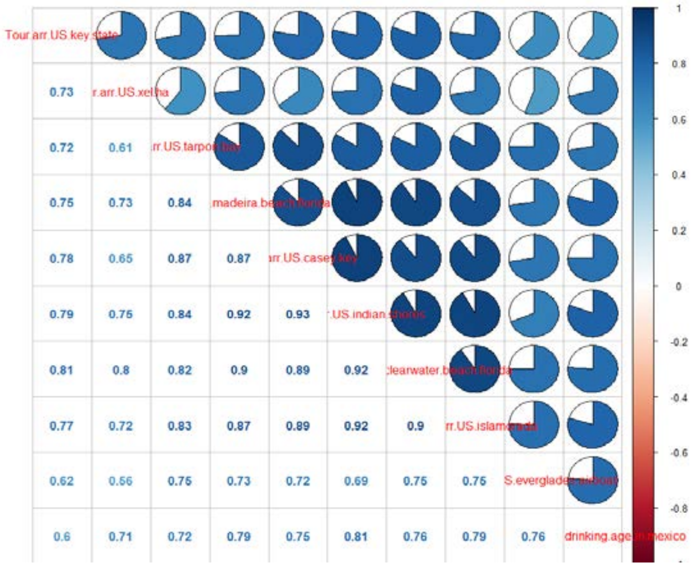
3. Methodology

As a proxy for the dependent variable representing tourism demand, quarterly tourism receipts were collected from the Centrale Bank van Aruba from 2004 to 2016. Tourism arrivals and nights for the 5 largest source markets are collected on a monthly basis and passed through google correlate to identify search terms that have a similar pattern of activity as our dependent variables ("features"). Google correlate surfaces search queries whose temporal patterns are most highly correlated (R2) with our target pattern. Google correlate employs a novel approximate nearest neighbour (ANN) algorithm over millions of candidate queries in an online search trees to produce results. The top 5 source countries combined are found to account for about 90 percent of arrivals/nights.

In total 100 features are collected (see Table 1). The fact that most of the features identified by google correlate are related to tourism provides initial face validity of their inclusion.

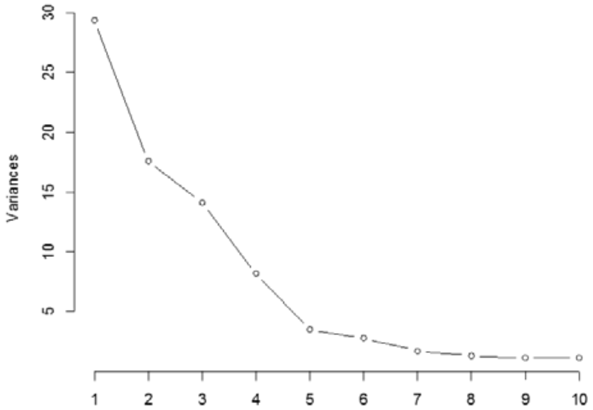
We utilize Google trends to download the features. In practice two ways to achieve this were considered: by (tediously) downloading CSV files from the Google Trends website or by scripting a connection to Google trends data through packages like "gtrendsR" in the R statistical software. Once the data was obtained, unsurprisingly, many of the collected variables illustrated strong co-movement. Figure 1 provides a correlogram of the first 10 features.

Figure 1: Correlogram (First ten features)



As Figure 1 illustrates, the 100 collected Google search term series exhibit a high degree of co-movement. The bulk of their dynamics can therefore be captured by relatively few common factors, effectively reducing the dimensions of the full dataset to a more manageable set (5) of key drivers (see Figure 2). The assumption being that the 5 principle components represent a concise and sufficient summary of underlying processes that drive tourism demand. As is evident by Figure 2, the marginal improvements in captured variance diminishes greatly after the 5th principal component. In terms of approach, the reduction through PCA closely resembles the methodology adopted by Zult and Schreuder (2011).

Figure 2: Correlogram Principle Components



The 'ragged edge' issue of incomplete real-time macroeconomic data is particularly apparent in Aruba, were the dependent variable of interest (tourism receipts) can have a lag of up to 6 months in comparison to real-time Google data. Therefore, to fully take advantage of the monthly frequency and timely availability of the collected features, the dependent variable (tourism receipts) is disaggregated using the "Chow Lin" with 'sum' disaggregation method which converts the series from a quarterly to monthly frequency (see: Sax and Steiner, 2013) using the following equation:

$$REC_t = \propto + \beta_1 Arrivals_t + \beta_2 Nights_t + \beta_3 Time_t + \beta_4 D1 + \beta_5 D3 + e_t,$$

Where the dependent variable tourism receipt is a function of total tourist arrivals, total nights, a time variable, and 2 seasonal dummies.

Temporal disaggregation

Table 1

Variables	Coefficient	Std. Error	T value	Prob
(Intercept)	1.27.0e+02	0.2423	5.243	<0.001
Arrivals	-2.543e-03	7.817e-04	-3.253	0.002
Nights	3.928e-04	9.608e-05	4.088	<0.001
Time	5.503e+01	6.386e-02	8.658	<0.001
D1	3.299e+01	3.805e+00	8.670	<0.001
D3	-1.325e+01	3.085e+00	-4.294	<0.001

Chow-Lin Min RSS Ecotrim disaggregation with ‘sum’ conversion.

In general, learning algorithms benefit from standardization of the data set. The intention is to counteract the effects of different features having different scales (which then causes models to assign incorrect weights). The data is therefore normalized between 0 and 1 by:

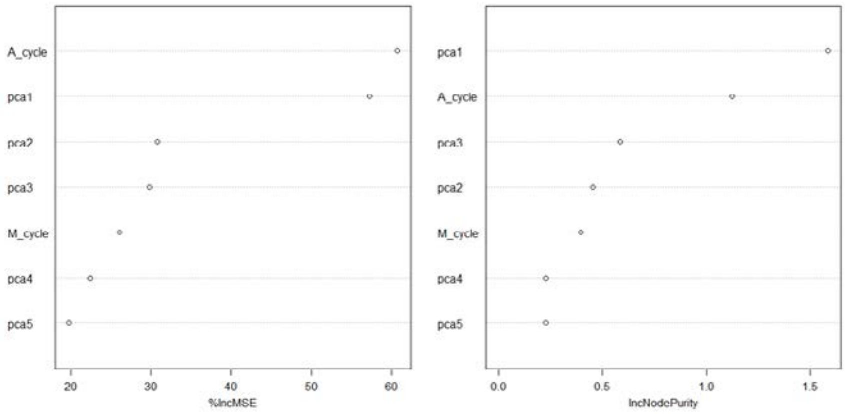
$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

To implement machine learning algorithms, the prediction is framed as a supervised learning problem where we have to infer from historical data the possibly nonlinear dependence between the input and the output (future value). To run the machine learning algorithms, the dataset is split between a training set (January 2004 – December 2014) and a test set (January 2015 – December 2016). The forecast period is defined to cover 12 month beyond the test set (January 2017 – December 2017). More specifically, 3 machine learning techniques are implemented, namely: random forest including Google data, neural network auto regression, and a neural network including Google data.

Random Forest (RF)

At core, these methods are based on the notion of a decision tree, which aims to deliver a structured set of yes/no questions that can quickly sort through a wide set of features, and produce an accurate prediction of a particular outcome. Decision trees are computationally efficient, and work well for problems where there are important nonlinearities. The RF algorithm seeks to improve the model’s predictive ability by growing numerous (unpruned) trees and combining the result. This method produces surprisingly good out-of-sample results, particularly with highly nonlinear data. In fact, Random Forests have been accredited as the most successful general-purpose algorithm in modern times (Varian, 2013). A more detailed methodological discussion on how RF works in the context of time series forecasting is provided by Tiffen (2016). In constructing the RF, the 5 Google based principle components are utilized along with two additional time variables to account for annual and monthly cyclical behaviour (Figure 3).

Figure 3: Random Forest



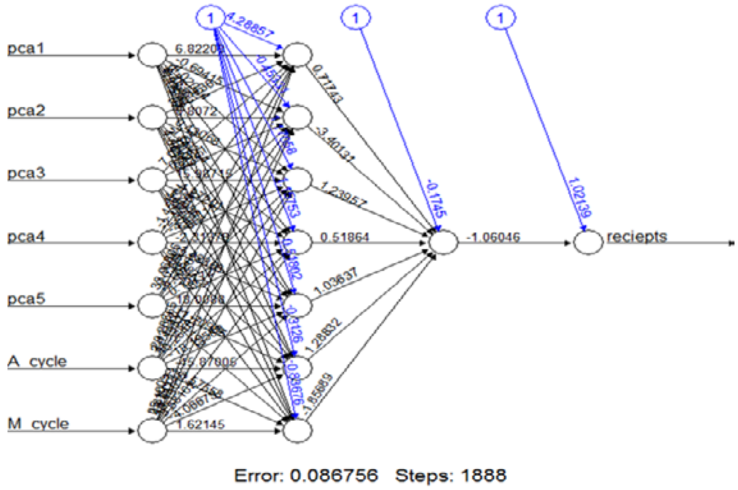
Neural Network Autoregression (NNA)

Artificial neural networks are forecasting methods that are based on simple mathematical models of the brain. They allow complex nonlinear relationships between the response variable and its predictors. A neural network consists of an input layer, an output layer, and usually one or more hidden layers. Each of these layers contains nodes, and these nodes are connected to nodes at adjacent layer(s). In the neural network autoregression, lagged values of the time series are used as inputs to a neural network (similar to a linear autoregressive model). We consider a feed-forward network with one hidden layer. The forecasts are obtained by a linear combination of the inputs. The weights are selected in the neural network framework using a "learning algorithm" that minimises a "cost function" such as MSE (Hyndman and Athanaspoloulos, 2013).

Neural Network (using Google data)

Consistent with the input in the previously mentioned RF calculation, the 5 Google based principle components are supplemented with two time variables to account for annual and monthly cyclical behaviour. We consider a feed-forward network with one hidden layer. Figure 4 provides a visual representation of the neural network and the inter-relationship between the different layers.

Figure 4: Neural Network (using Google data)



4. Results

In this section we evaluate the forecasting accuracy of the three machine learning techniques (Random Forest, Neural Network Auto Regression and Neural Network including Google data) by examining out-of-sample predictions of tourism receipts in Aruba. The collected data was divided in training, validation and test sets to assess the performance of the algorithms on unseen data. The forecasting performances are compared in terms of their relative performance for the test set (January 2015 – December 2016). The results of our forecasting competition are shown in Table 2.

Forecast model accuracy

Table 2

	ME	RMSE	MAE	MPE	MAPE	ACF1	Thiel's U
Random Forest	0.093	0.137	0.097	11.571	12.404	0.562	1.018
Neural Network AR	0.037	0.051	0.042	5.023	5.972	0.352	0.405
Neural Network (Google)	-0.037	0.072	0.063	-7.334	10.164	0.493	0.675

When comparing forecasting performance, the various measures are consistent in contending that the prediction error is substantially less for the Neural Network AR model, followed by the Neural Network using the Google variables.

Annex 2 provides a visual example of a Neural network model fitted based on the training dataset, tested for accuracy using the test set and forecasted for 12 months ahead.

5. Conclusion

In terms of interpretability, it should be noted that both neural networks and random forest resemble black boxes: explaining their outcome is much more difficult than explaining the outcome of simpler models (such as a linear models) due to their complexity. Nevertheless, these models have the advantage of providing fairly accurate estimates and despite their computational complexity, improvements in computing technology enable relatively quick execution of machine learning algorithms. This paper provided an example where the combination of near real-time Google search information along with machine learning techniques provides forecasters with a new set of tools to model complex relationships such as tourism demand, but which could easily be transferred to similar macro-economic variables within other domains.

6. References

Ahmed, Y. (2013). Analytical review of tourism demand studies from 1960 to 2014. International Journal of Science and Research (IJSR).

Breiman, L. (2001). Statistical Modeling: The two cultures. Statistical Science 2001, Vol. 16, No. 3, 199-231

Cankurt, S. and Subasi, A. (2015). Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components. Balkan Journal of Electrical & Computer Engineering, 2015, Vol.3, No.1.

Claveria, O. et al (2013). Tourism demand forecasting with different neural network models. Research Institute of Applied Economics. Working paper 2013/21.

Croes, R. and Vanegas, M. (2005). An econometric study of tourist arrivals in Aruba and its implications. Tourism Management. December 2005.

Hassink, W., de Kort, R. and Ridderstaat, J. (2015) "De economische consequenties van de verdwijning van Natalee Holloway", Me Judice, 30 mei 2015.

Hyndman, R.J. and Athanasopoulos, G. (2013) Forecasting: principles and practice. OTexts: Melbourne, Australia. http://otexts.org/fpp/. Accessed on 9/14/2017.

Law, R. and Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. Tourism Management 20 (1999) 89-97.

Mohebbi, M. et al (2011). Google Correlate Whitepaper. Draft date: June 9, 2011.

O'Mahony, B., Lee, C., Bergin-Seers, S., Galloway, G. & McMurray, A. (2008). Seasonality in the Tourism Industry: Impacts and Strategies.

Ridderstaat, J.R. (2015). Studies on Determinants of Tourism Demand: Dynamics in a Small Island Destination. The Case of Aruba

Sax, C. and Steiner, P. (2013). Temporal Disaggregation of Time Series. The R journal Vol. 5/2, December 2013.

Shmueli, G. and Lichtendahl, K. (2016). Practical Time Sries Forecasting with R: A hands-On Guide [2nd Edition].

Song, H. and Li, G. (2008). Tourism Demand Modelling and Forecasting. A review or Recent Research.

Stephens-Davidowitz, S. and Varian, H. (2015). A hands-on guide to Google data. Draft date: March 7, 2015.

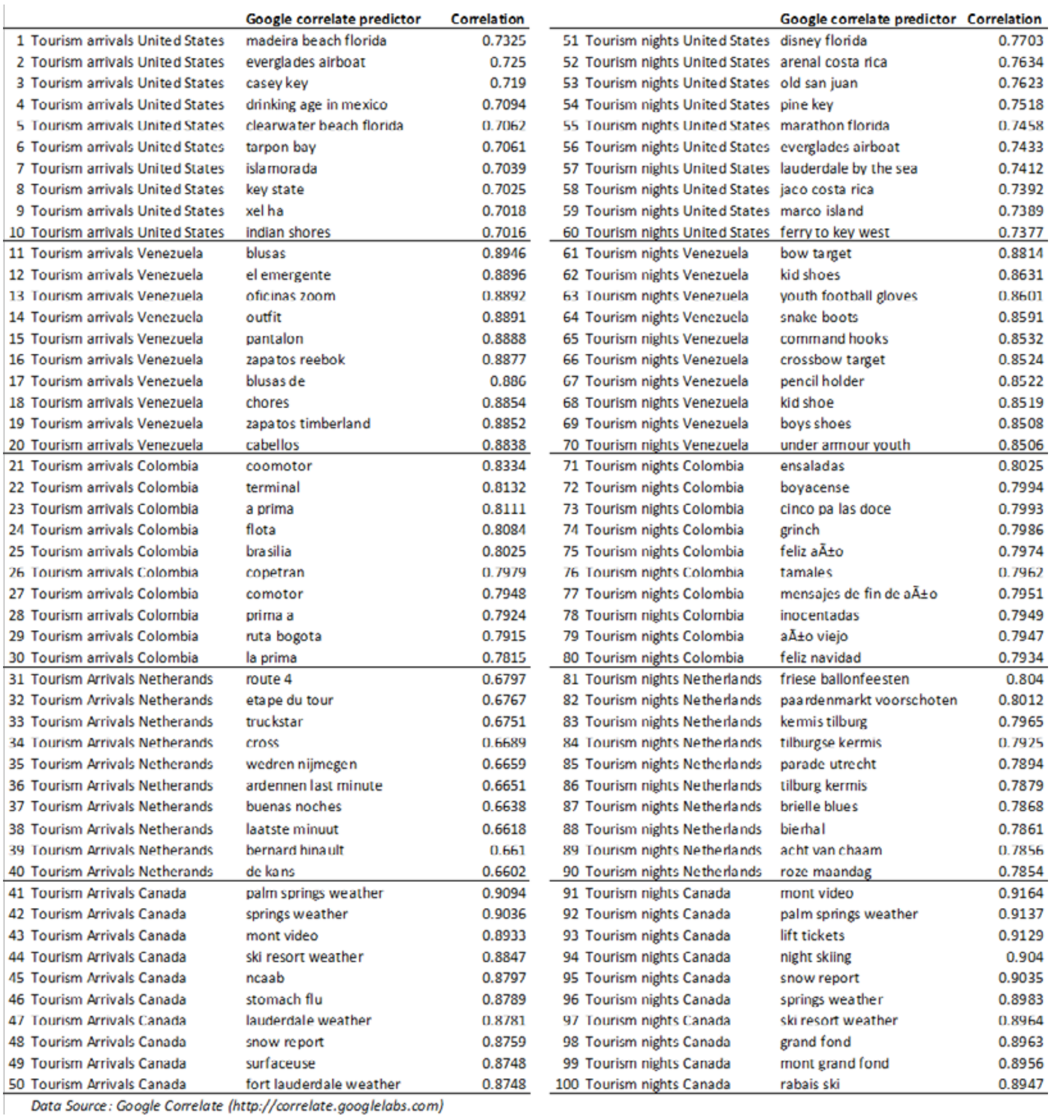
Tiffen, A. (2016). Seeing in the Dark: A Machine-Learning Approach to Nowcasting in Lebanon. IMF Working Paper. WP/16/56

World Travel and Tourism Council (2017). Economic Impact 2017 Aruba.

Varian, H. (2013). Big Data: New Tricks for Econometrics.

Zult, D. and Schreuder, G. (2011). Monthly Nowcast of Aruban Year-on-Year Growth in GDP. Statistics Netherlands. February 2011.

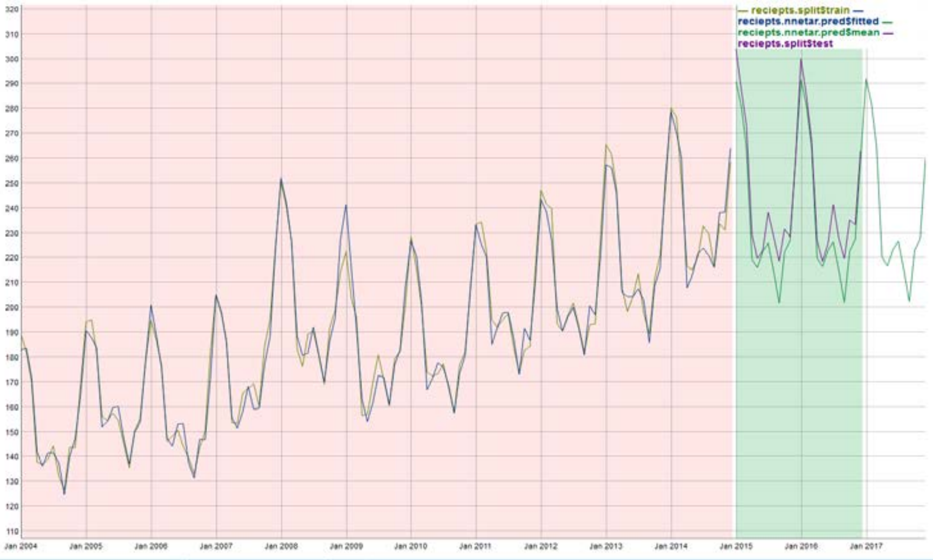
Appendix 1: google variable selection



		Google correlate predictor	Correlation			Google correlate predictor	Correlation
1	Tourism arrivals United States	madeira beach florida	0.7325	51	Tourism nights United States	disney florida	0.7703
2	Tourism arrivals United States	everglades airboat	0.725	52	Tourism nights United States	arenal costa rica	0.7634
3	Tourism arrivals United States	casey key	0.719	53	Tourism nights United States	old san juan	0.7623
4	Tourism arrivals United States	drinking age in mexico	0.7094	54	Tourism nights United States	pine key	0.7518
5	Tourism arrivals United States	clearwater beach florida	0.7062	55	Tourism nights United States	marathon florida	0.7458
6	Tourism arrivals United States	tarpon bay	0.7061	56	Tourism nights United States	everglades airboat	0.7433
7	Tourism arrivals United States	islamorada	0.7039	57	Tourism nights United States	lauderdale by the sea	0.7412
8	Tourism arrivals United States	key state	0.7025	58	Tourism nights United States	jaco costa rica	0.7392
9	Tourism arrivals United States	xel ha	0.7018	59	Tourism nights United States	marco island	0.7389
10	Tourism arrivals United States	indian shores	0.7016	60	Tourism nights United States	ferry to key west	0.7377
11	Tourism arrivals Venezuela	blusas	0.8946	61	Tourism nights Venezuela	bow target	0.8814
12	Tourism arrivals Venezuela	el emergente	0.8896	62	Tourism nights Venezuela	kid shoes	0.8631
13	Tourism arrivals Venezuela	oficinas zoom	0.8892	63	Tourism nights Venezuela	youth football gloves	0.8601
14	Tourism arrivals Venezuela	outfit	0.8891	64	Tourism nights Venezuela	snake boots	0.8591
15	Tourism arrivals Venezuela	pantalon	0.8888	65	Tourism nights Venezuela	command hooks	0.8532
16	Tourism arrivals Venezuela	zapatos reebok	0.8877	66	Tourism nights Venezuela	crossbow target	0.8524
17	Tourism arrivals Venezuela	blusas de	0.886	67	Tourism nights Venezuela	pencil holder	0.8522
18	Tourism arrivals Venezuela	chores	0.8854	68	Tourism nights Venezuela	kid shoe	0.8519
19	Tourism arrivals Venezuela	zapatos timberland	0.8852	69	Tourism nights Venezuela	boys shoes	0.8508
20	Tourism arrivals Venezuela	cabellos	0.8838	70	Tourism nights Venezuela	under armour youth	0.8506
21	Tourism arrivals Colombia	coomotor	0.8334	71	Tourism nights Colombia	ensaladas	0.8025
22	Tourism arrivals Colombia	terminal	0.8132	72	Tourism nights Colombia	boyacense	0.7994
23	Tourism arrivals Colombia	a prima	0.8111	73	Tourism nights Colombia	cinco pa las doce	0.7993
24	Tourism arrivals Colombia	flota	0.8084	74	Tourism nights Colombia	grinch	0.7986
25	Tourism arrivals Colombia	brasilia	0.8025	75	Tourism nights Colombia	feliz aÃ±o	0.7974
26	Tourism arrivals Colombia	copetran	0.7979	76	Tourism nights Colombia	tamales	0.7962
27	Tourism arrivals Colombia	comotor	0.7948	77	Tourism nights Colombia	mensajes de fin de aÃ±o	0.7951
28	Tourism arrivals Colombia	prima a	0.7924	78	Tourism nights Colombia	inocentadas	0.7949
29	Tourism arrivals Colombia	ruta bogota	0.7915	79	Tourism nights Colombia	aÃ±o viejo	0.7947
30	Tourism arrivals Colombia	la prima	0.7815	80	Tourism nights Colombia	feliz navidad	0.7934
31	Tourism Arrivals Netherands	route 4	0.6797	81	Tourism nights Netherlands	friese ballonfeesten	0.804
32	Tourism Arrivals Netherands	etape du tour	0.6767	82	Tourism nights Netherlands	paardenmarkt voorschoten	0.8012
33	Tourism Arrivals Netherands	truckstar	0.6751	83	Tourism nights Netherlands	kermis tilburg	0.7965
34	Tourism Arrivals Netherands	cross	0.6689	84	Tourism nights Netherlands	tilburgse kermis	0.7925
35	Tourism Arrivals Netherands	wedren nijmegen	0.6659	85	Tourism nights Netherlands	parade utrecht	0.7894
36	Tourism Arrivals Netherands	ardennen last minute	0.6651	86	Tourism nights Netherlands	tilburg kermis	0.7879
37	Tourism Arrivals Netherands	buenas noches	0.6638	87	Tourism nights Netherlands	brielle blues	0.7868
38	Tourism Arrivals Netherands	laatste minuut	0.6618	88	Tourism nights Netherlands	bierhal	0.7861
39	Tourism Arrivals Netherands	bernard hinault	0.661	89	Tourism nights Netherlands	acht van chaam	0.7856
40	Tourism Arrivals Netherands	de kans	0.6602	90	Tourism nights Netherlands	roze maandag	0.7854
41	Tourism Arrivals Canada	palm springs weather	0.9094	91	Tourism nights Canada	mont video	0.9164
42	Tourism Arrivals Canada	springs weather	0.9036	92	Tourism nights Canada	palm springs weather	0.9137
43	Tourism Arrivals Canada	mont video	0.8933	93	Tourism nights Canada	lift tickets	0.9129
44	Tourism Arrivals Canada	ski resort weather	0.8847	94	Tourism nights Canada	night skiing	0.904
45	Tourism Arrivals Canada	ncaab	0.8797	95	Tourism nights Canada	snow report	0.9035
46	Tourism Arrivals Canada	stomach flu	0.8789	96	Tourism nights Canada	springs weather	0.8983
47	Tourism Arrivals Canada	lauderdale weather	0.8781	97	Tourism nights Canada	ski resort weather	0.8964
48	Tourism Arrivals Canada	snow report	0.8759	98	Tourism nights Canada	grand fond	0.8963
49	Tourism Arrivals Canada	surfaceuse	0.8748	99	Tourism nights Canada	mont grand fond	0.8956
50	Tourism Arrivals Canada	fort lauderdale weather	0.8748	100	Tourism nights Canada	rabais ski	0.8947

Data Source: Google Correlate (http://correlate.googlelabs.com)

Appendix 2: Forecast example



Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Forecasting tourism demand through search queries and machine learning[1]

Rendell E. de Kort,
Central Bank of Aruba

[1] This presentation was prepared for the meeting. 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 meeting.

CENTRALE BANK VAN ARUBA

Forecasting Tourism demand through search queries and machine learning

Rendell E. de Kort

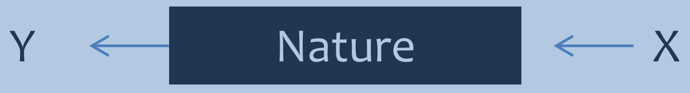
IFC – Bank Indonesia Satellite Seminar on “Big Data”, Bali, Indonesia, 21 March 2017

The content of this presentation represents the author’s personal opinions and do note necessarily reflect the views of the Centrale Bank van Aruba (CBA). The CBA does not assume any liability for any loss that may result from the reliance upon this information.

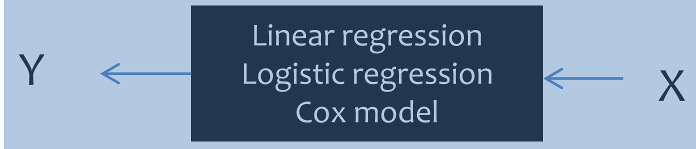
1. Background

Statistical modeling: The two cultures

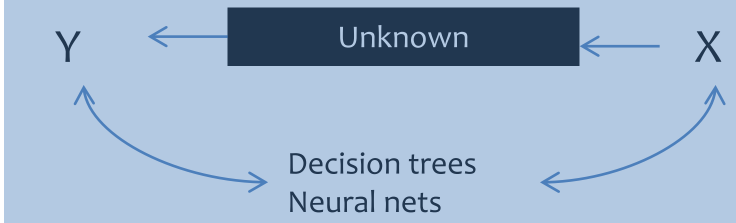
- Prediction
- Information



The data modeling culture



The algorithmic modeling culture



Source: Breiman, L. (2001). Statistical Modeling: The two cultures. Statistical Science 2001, Vol. 16, No. 3, 199-231

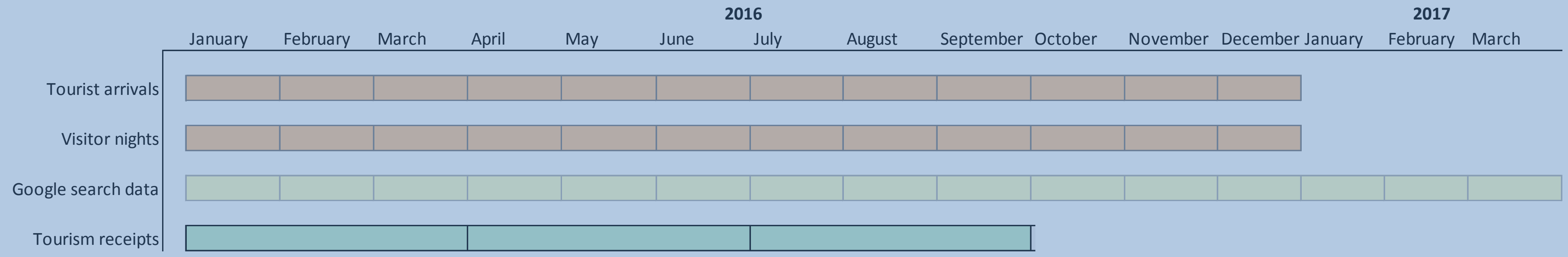
2. Considerations

- Traditional statistical models require eliminating the effects of seasonality prior to forecasting.
- Literature points to the ability of machine learning models to recognize and learn seasonal patterns without removing them from the raw data.
- Traditional statistical forecasting methods are mostly linear models while the literature indicates machine learning techniques cope well with possible nonlinearities.

2. Considerations

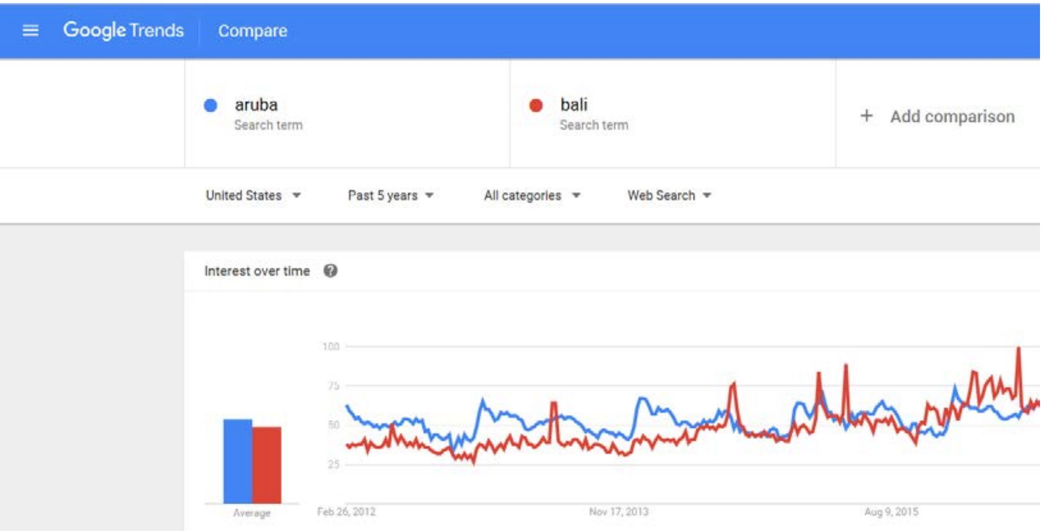
- Real-time macroeconomic data are typically incomplete for today and the immediate past (‘ragged edge’) and subject to revision.
- To enable more timely forecasts the issue is initially framed as a standard “*nowcasting*” problem.

Figure 1: Ragged edge



3. Preprocess

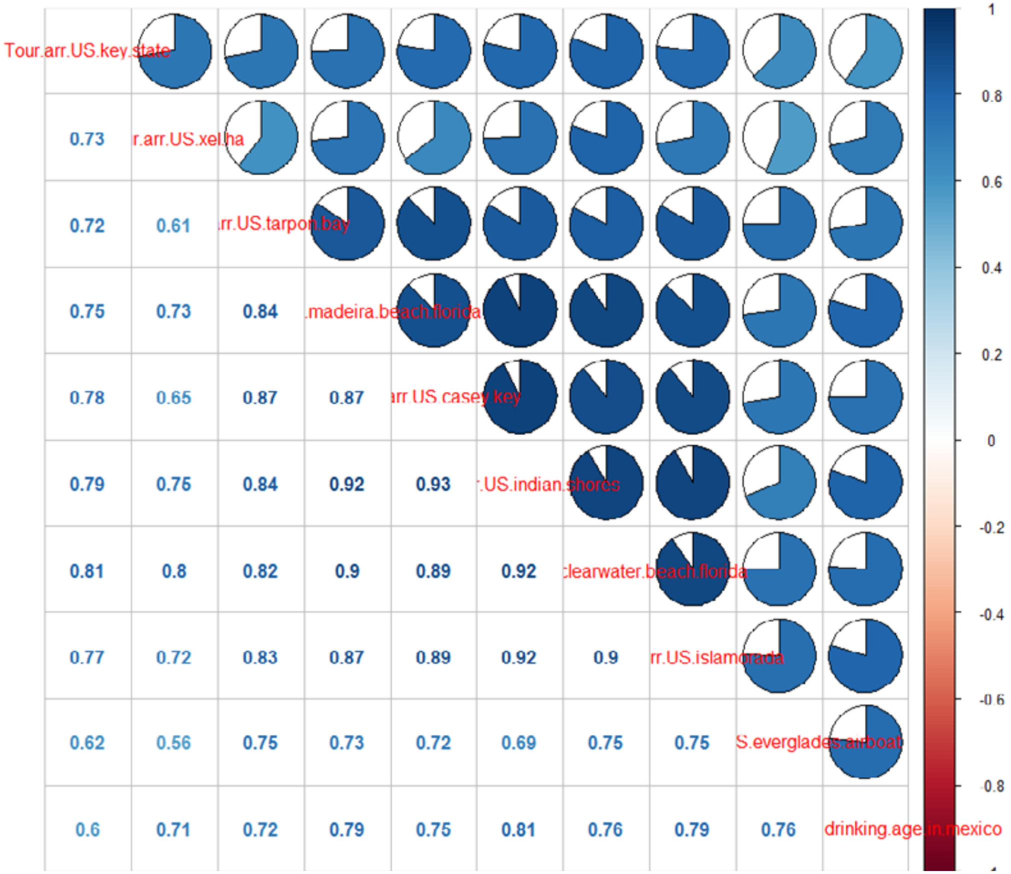
- Google trends enables easy download of Google search query time series.
- Download can take place by:
  - (tediously) downloading CSV files from the Google Trends website.
  - Using packages like “gtrendsR” to connect with your Google account and downloading Trends data directly into R with a simple script.

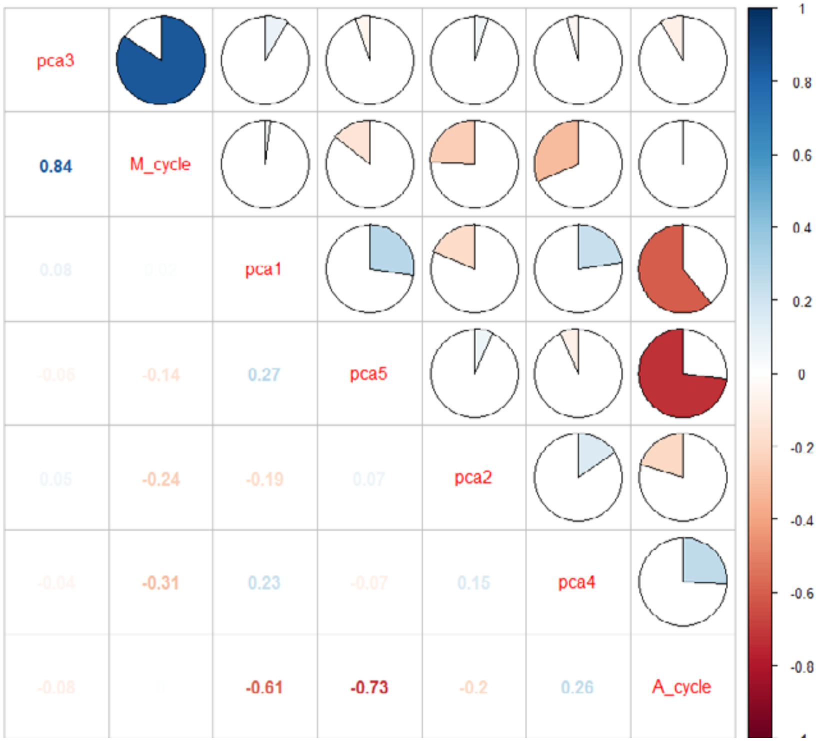


Google correlate predictor	Correlation
madeira beach florida	0.7325
everglades airboat	0.725
casey key	0.719
drinking age in mexico	0.7094
clearwater beach florida	0.7062
tarpon bay	0.7061
islamorada	0.7039
key state	0.7025
xel ha	0.7018
indian shores	0.7016

Data Source: Google Correlate (http://www.google.com/trends/correlate)

3. Preprocess

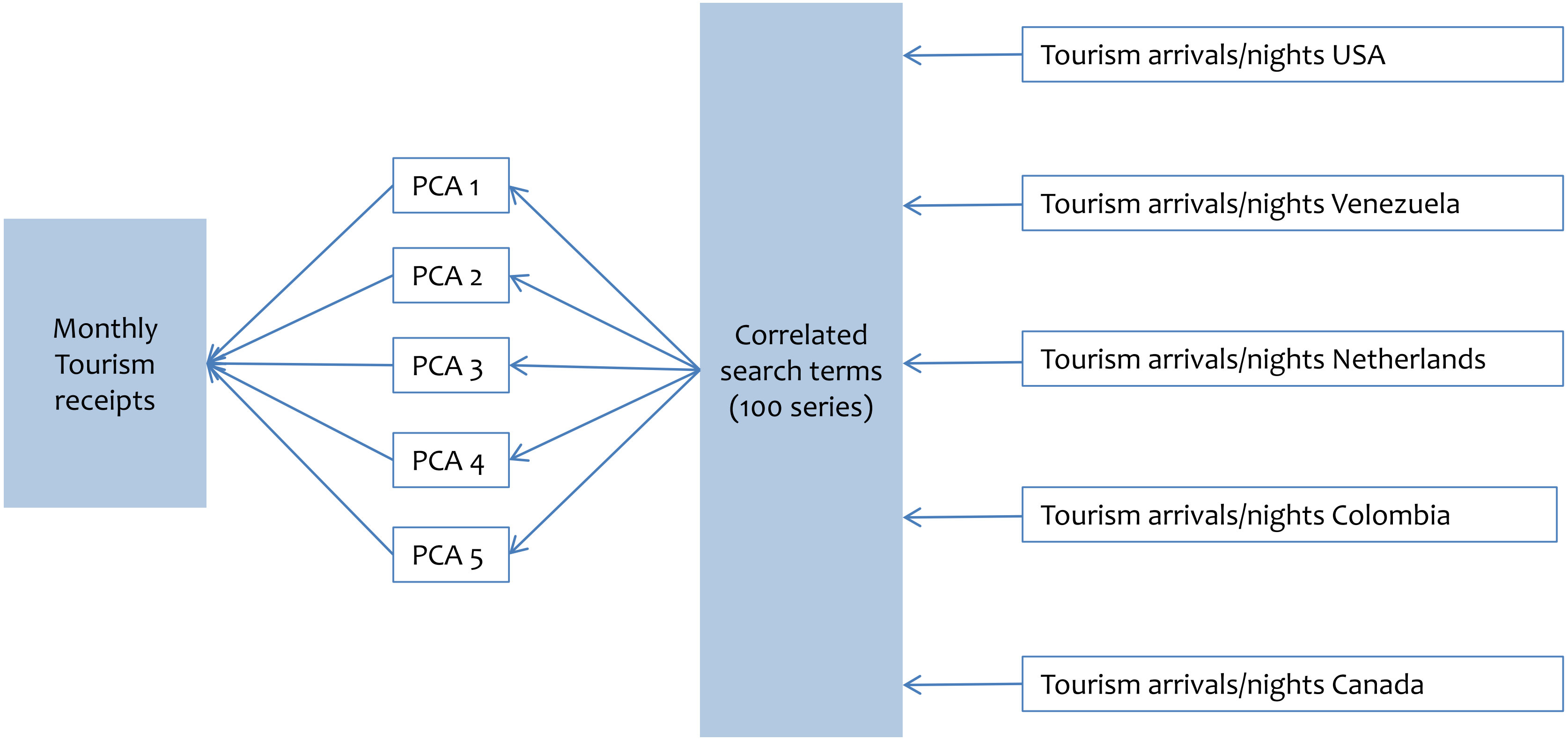




* Important to be mindful of extracting information from a large number of correlated proxies (100 in our example).

3. Preprocess

- The dependent variable "tourism receipts" is measured on a quarterly basis. To take full advantage of features collected on a monthly basis, we're disaggregating the series.


Monthly Tourism receipts
PCA 1
PCA 2
PCA 3
PCA 4
PCA 5
Correlated search terms (100 series)
Tourism arrivals/nights USA
Tourism arrivals/nights Venezuela
Tourism arrivals/nights Netherlands
Tourism arrivals/nights Colombia
Tourism arrivals/nights Canada

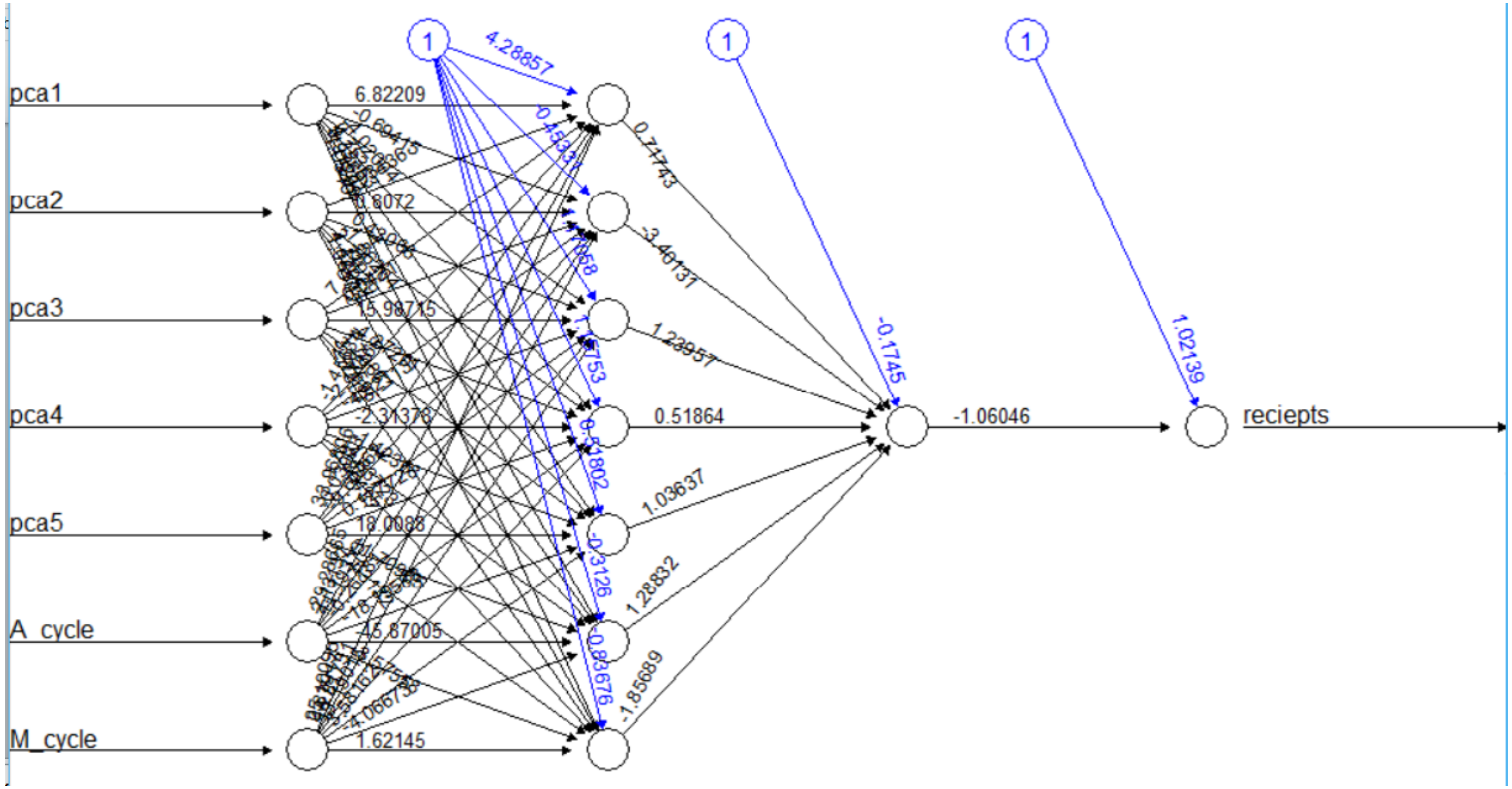
4. Machine learning estimations

Random Forest

- At core, these methods are based on the notion of a decision tree, which aims to deliver a structured set of yes/no questions that can quickly sort through a wide set of features, and produce an accurate prediction of a particular outcome.
- Decision trees are computationally efficient, and work well for problems where there are important nonlinearities.
- The RF algorithm seeks to improve the model's predictive ability by growing numerous (unpruned) trees and combining the result.

4. Machine learning estimations

Neural Network (NN)



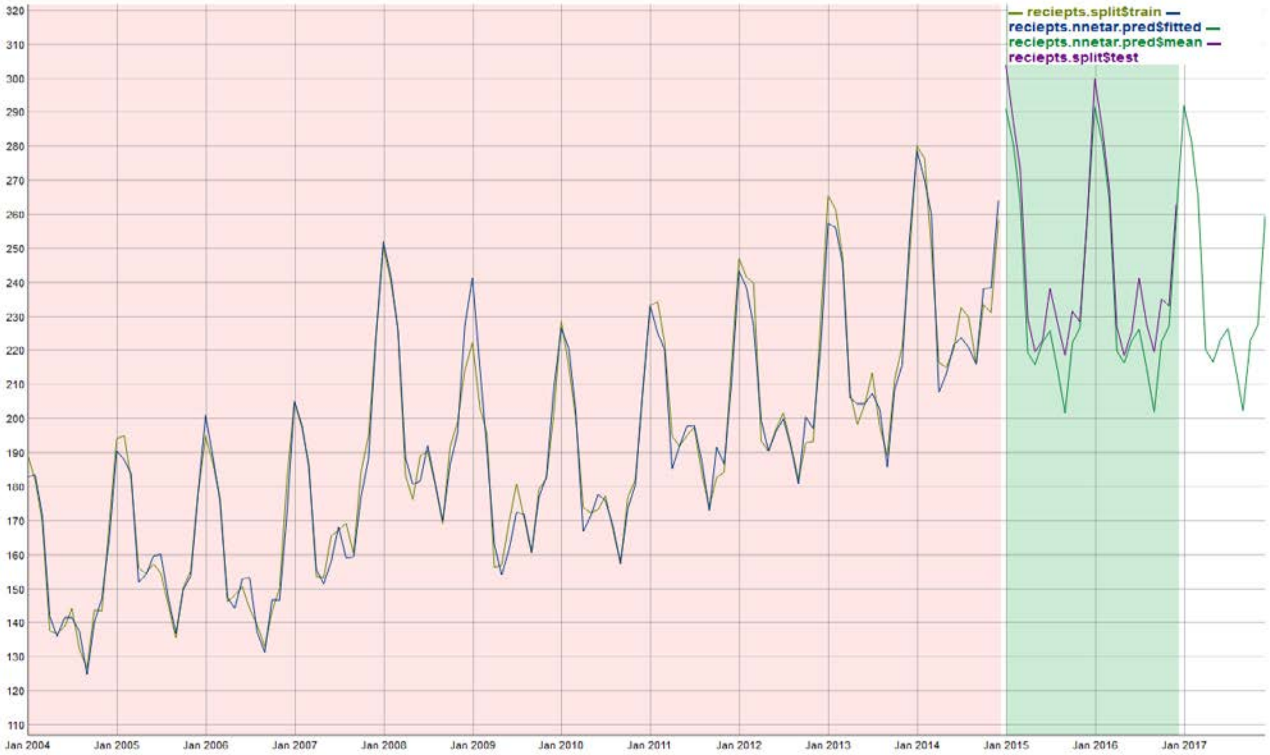
4. Machine learning estimations

Neural network autoregression

- With time series data, lagged values of the time series can be used as inputs to a neural network (similar to a linear autoregressive model).
- we consider a feed-forward network with one hidden layer
- Using the “nnetar” function in R

4. Machine learning estimations

Example



	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Random Forest	0.093	0.137	0.097	11.571	12.404	0.562	1.018
Nueral Network AR	0.037	0.051	0.042	5.023	5.972	0.352	0.405
Nueral Network (Google)	-0.037	0.072	0.063	-7.334	10.164	0.493	0.675

5. Concluding remarks

- Machine learning models provide “good” out-of-sample success.
- Takes advantage of additional search information.
- Tradeoff between interpretability of the model and forecasting performance (predictive not descriptive).
- Benefit: ease of processing once the script is in place.
- Opportunities exist to include the google data in a expanded framework to forecast economic growth.
- The R script available on GitHub: https://github.com/rendell

THANK YOU

CENTRALE BANK VAN ARUBA

TERIMA KASIH

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Capturing depositors' expectations with Google data[1]

Patrick Weber, Falko Fecht and Stefan Thum,
Deutsche Bundesbank

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

DEUTSCHE
BUNDESBANK
EUROSYSTEM

Capturing depositors' expectations with Google Data

Falko Fecht, Stefan Thum and Patrick Weber

IFC-BI Satellite Seminar on Big Data, 21 March 2017

DISCLAIMER
The views expressed in this presentation do not necessarily reflect the views of the Deutsche Bundesbank.

1. Motivation

Strategic complementarities and financial crises

Motivation

Can Google searches be used as a predictor for a deposit run on banks (time series and cross section)?

Many financial institutions exposed to self-fulfilling liquidity crises

- Financial institutions performing liquidity transformation are exposed to runs by depositors
- Worries that others excessively withdraw induce investors to withdraw
- Some empirical evidence…
  - ➢ Mutual funds: Chen, Goldstein, & Jiang (JFE 2010)
  - ➢ Open end real estate funds: Fecht & Wedow (JFI 2014)

2. Contribution
Capturing depositor‘s expectations

How to measure investors’ expectations?

- Google searches might serve as a proxy for investors' worries
- Searches serve as an early warning indicator for liquidity crises

Exploit particularities of German banking system

- Savings banks de facto government-guaranteed
- Suitable reference group: Credit cooperative banks

Natural experiment: Blanket guarantee for German banks’ liabilities

- Public announcement on 5 October 2008 by Chancellor Merkel
- All retail deposits are safe: deposit insurance scheme
- Intention was to avoid possible bank run

3. Data
Google data are obtained via Google Trends

We obtain Google Trends Data via *www.google.com/trends*

- **Relevant data:**
  Relative search interest in search terms related to **deposit insurance** in Germany at the local level

- **Breakdowns:**
  Web searches by federal state

3. Data
Augment the Google data set with central bank data sources

We **augment Google** data with…

1. Bundesbank Balance Sheet Items statistics

- **Outstanding Euro amounts of overnight deposits** at a monthly frequency at bank level (census approach)

2. Bundesbank MFI Interest Rate statistic

- **Interest rates on outstanding amounts of overnight deposits** at a monthly frequency at bank level for roughly 230 German banks (sample approach)

4. Key variables
Measuring deposit flows

Deposit shift variable

$$\text{Deposit Shift}_{j,t} = \frac{\text{Volume Savings Banks}_{j,t}}{\text{Volume Cooperative Banks}_{j,t}}$$

$$\Delta\text{Deposit Shift}_{j,t} = \text{Deposit Shift}_{j,t} - \text{Deposit Shift}_{j,t-6}$$

j	Federal State
t	Month

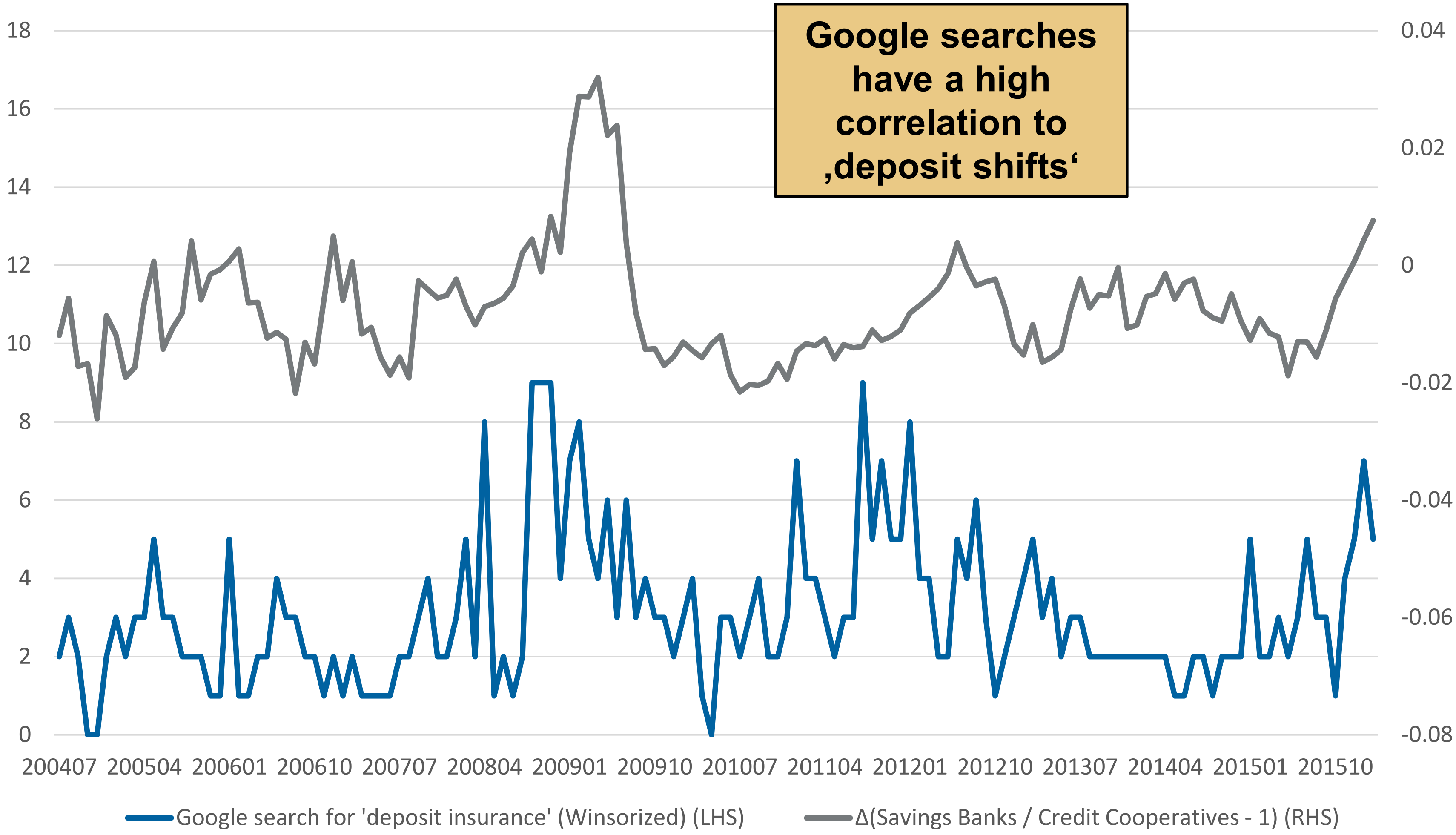
Control variable: Interest rate margin

$$\text{Interest Margin}_{j,t} = \text{Interest Rate Savings Banks}_{j,t} - \text{Interest Rate Coop. Banks}_{j,t}$$

$$\Delta\text{Interest Margin}_{j,t} = \text{Interest Margin}_{j,t-1} - \text{Interest Margin}_{j,t-7}$$

5. Descriptive Statistics

Google search interest versus deposit shifts



Google searches have a high correlation to ‚deposit shifts‘

6. Methodology

Model setup

Inter-temporal Analysis

1. VAR analysis and Granger causality (Perspective: Germany only)
2. Standard OLS Regression (Perspective: Germany only)

Panel Perspective

3. Analysis of impact of government guarantee (Perspective: State level)

$$\begin{aligned}&\Delta\left(\text{Deposit Shift}_{j,t}\right)\\&= \alpha_j + \alpha_t + \beta_1 \text{Guarantee}_t + \beta_2 \text{Google}_{f,j,t} + \beta_3 \text{Google}_{f,j,t} * \text{Guarantee}_t\\&+ \beta_4\left(\Delta\text{Interest Margin}_j\right) * \text{Guarantee}_t + \beta_5\left(\Delta\text{Interest Margin}_j\right) * \text{NoGuarantee}_t + u_{j,t}\end{aligned}$$

α_j	Fixed effect of state j
α_t	Monthly time fixed effect

4. Bank-level panel analysis (Perspective: Individual banks)

7. Results

Granger Causality

Ordering: *Google, Interest Spread, Deposit shift*

Equation		*Factor*	*Chi2*	*df (lags)*	*p-value*
Google	=	Interest Spread	1.9922	2	0.369
Google	=	Deposit shift	0.40469	2	0.817
Interest Spread	=	Google	10.967	2	**0.004**
Interest Spread	=	Deposit shift	2.9827	2	0.225
Deposit shift	=	Google	7.5309	2	**0.023**
Deposit shift	=	Interest Spread	13.643	2	**0.001**

Google searches Granger cause the ***Interest Spread*** *and* ***Deposit Shifts***

Alternative Google searches also Granger cause the ***Interest Spread*** *and, for most of the time,* ***Deposit Shifts***

The results hold ***independent of the ordering***

English translation of search term	Correlation
banks deposit insurance	90%
deposit insurance banks	90%
how safe is my money	83%
secure banks	78%
state guarantee	63%
bank bankruptcy	60%
deposit insurance savings banks	55%
dexia communal bank	49%
money market saving	49%
statutory deposit insurance	37%

7. Results
Standard Panel Regression (Perspective: State level)

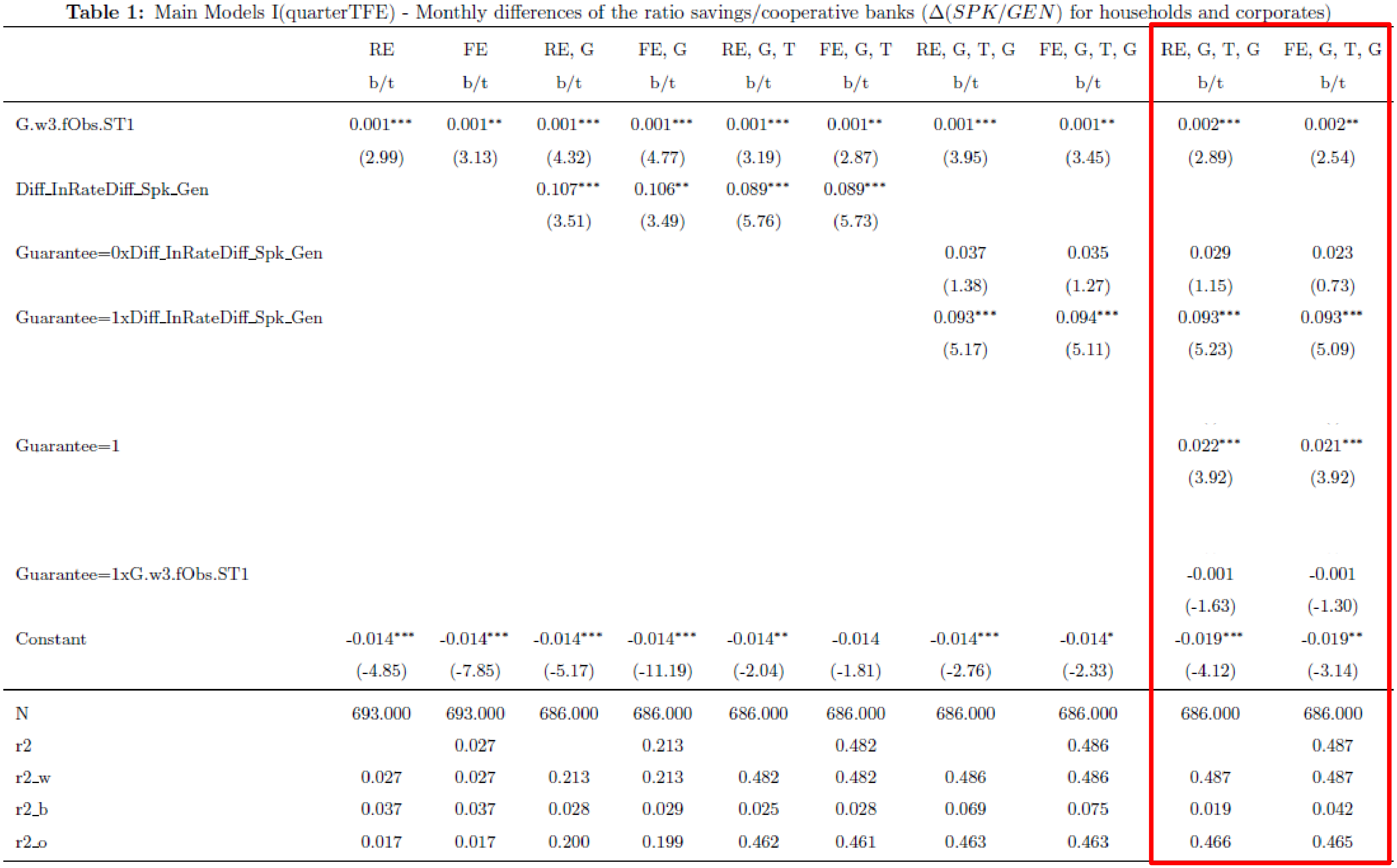


Table 1: Main Models I(quarterTFE) - Monthly differences of the ratio savings/cooperative banks ($\Delta(SPK/GEN)$ for households and corporates)

	RE	FE	RE, G	FE, G	RE, G, T	FE, G, T	RE, G, T, G	FE, G, T, G	RE, G, T, G	FE, G, T, G
	b/t	b/t	b/t	b/t	b/t	b/t	b/t	b/t	b/t	b/t
G.w3.fObs.ST1	0.001^{***}	0.001^{**}	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{**}	0.001^{***}	0.001^{**}	0.002^{***}	0.002^{**}
	(2.99)	(3.13)	(4.32)	(4.77)	(3.19)	(2.87)	(3.95)	(3.45)	(2.89)	(2.54)
Diff_lnRateDiff_Spk_Gen			0.107^{***}	0.106^{**}	0.089^{***}	0.089^{***}				
			(3.51)	(3.49)	(5.76)	(5.73)				
Guarantee=0xDiff_lnRateDiff_Spk_Gen							0.037	0.035	0.029	0.023
							(1.38)	(1.27)	(1.15)	(0.73)
Guarantee=1xDiff_lnRateDiff_Spk_Gen							0.093^{***}	0.094^{***}	0.093^{***}	0.093^{***}
							(5.17)	(5.11)	(5.23)	(5.09)
Guarantee=1									0.022^{***}	0.021^{***}
									(3.92)	(3.92)
Guarantee=1xG.w3.fObs.ST1									-0.001	-0.001
									(-1.63)	(-1.30)
Constant	-0.014^{***}	-0.014^{***}	-0.014^{***}	-0.014^{***}	-0.014^{**}	-0.014	-0.014^{***}	-0.014^{*}	-0.019^{***}	-0.019^{**}
	(-4.85)	(-7.85)	(-5.17)	(-11.19)	(-2.04)	(-1.81)	(-2.76)	(-2.33)	(-4.12)	(-3.14)
N	693.000	693.000	686.000	686.000	686.000	686.000	686.000	686.000	686.000	686.000
r2		0.027		0.213		0.482		0.486		0.487
r2_w	0.027	0.027	0.213	0.213	0.482	0.482	0.486	0.486	0.487	0.487
r2_b	0.037	0.037	0.028	0.029	0.025	0.028	0.069	0.075	0.019	0.042
r2_o	0.017	0.017	0.200	0.199	0.462	0.461	0.463	0.463	0.466	0.465

Includes monthly time fixed effects

8. Conclusions

Google searches are a valuable measure for depositors’ expectations

Google searches can indeed be used as a measure for the concern of depositors

- Indication of run-type phenomena in local deposit markets
- Effect is more pronounced for private households than for non-financial corporations

Blanket guarantee during the crisis led to a level playing field between private and public banks

- Deposit flows became more sensitive to interest rate spreads
- Fiercer competition in the deposit market
- Potentially more excessive risk-taking

Next step: Augment our analysis by Twitter data

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Determinants on firm survival in Chile: Evidence from cohort 2010 for the period 2011-2015[1]

Diana López, Daymler O´Farrill, Josué Pérez and Beatriz Velasquez,
Central Bank of Chile

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

Determinants on firm survival in Chile: Evidence from cohort 2010 for the period 2011-2015 (*)

Diana López Statistics Division

Daymler O´Farrill Statistics Division

Josué Pérez Statistics Division

Beatriz Velasquez Statistics Division

Abstract

In this paper we present evidence on the probability of firm survival in Chile for the period 2011-2015. Using information from the Chilean IRS we investigate the impact of financial variables on firms' survival. We first calculate the number of firms that "were born" in 2010 and follow them throughout the entire period of time and then we estimate the determinants of firm survival using the proportional hazard model. We find a positive and strong relationship between efficiency, leverage and profitability and the firms´ survival. Furthermore, we find that probability of survival has an inverted-U shape, which is in line with current literature.

Keywords: firms´ survival, proportional hazard model, economic industries

(*) Disclaimer: This document was first presented at the IFC *Satellite Seminar on Big data* that took place in Bali, Indonesia in March, 2017. The views, thoughts, and opinions expressed in the text belong solely to the authors, and do not reflect the view of the Central Bank of Chile. Any errors or omissions are our own. E-mail**:** dlopez@bcentral.cl, dofarril@bcentral.cl, jnperezt@bcentral.cl, bvelasqu@bcentral.cl.

Contents

1. INTRODUCTION

The process of entry and exit of firms is determinant in the corporate structure in a country and it must be understood as a core element for describing and comprehending the economic growth. This process allows us to go beyond the traditional view of economy since is based on the measure of the added value from those industries participant in the productive process.

In this context the policies to promote the creation of firms are focused on establishing the conditions to improve innovation, competitiveness, the use of technology and employment generation. In this sense, it is relevant to examine the performance of firms in terms of markets, size, leverage, age, assets, liabilities, etc.

The studies on business demography offer statistical information about the population of firms in national borders. In general, these studies rely on a set of indicators that represent the transformations suffered by firms over a time span. The performance of these indicators over time might be indicative of the evolution of more aggregated variables such as employment, capital stock and economic growth.

In developing this kind of studies a dataset that covers a period of at least five years is necessary. Also, it is required the information available to be as much detailed as possible at the firm level in both qualitatively and quantitatively terms. In the case of Chile the Central Bank receives information from the Chilean Internal Revenue Service (IRS, Servicio de Impuestos Internos) on a regularly basis. In this paper we work with the IRS' anonymized administrative records data base that gathers information on annual income taxes.

The main objective of this paper is to characterize the survival of firms created in 2010 in Chile for the period 2011-2015; in order to do the latter we estimate the proportional hazard model using partial information from the firm´s financial statements. In general, we want to add new insights to the comprehension of firms´ dynamics in terms of their survival and "destruction".

We find a positive and strong relationship between profitability, leverage and efficiency and the firms´ survival. In particular, we find a statistical significant difference between "survivor" and "non-survivor" firms regarding these indicators. These results remain stables even after being controlled by other variables. Additionally, we find that the probability of survival has an inverted-U shape, which is in line with most of the literature on this subject. This study complement the results from Pérez and Suazo (2014) in two directions: a) we offer new insights on the dynamic of firms´ survival in Chile and b) we estimate a survival model to predict the viability of a business initiative based on indicators of economic activity, leverage and profitability.

The rest of the paper is organized as follows: Section two describes the literature review regarding the determinants of firms´ survival. Section three briefly explains the empirical methodology we use to address the problem of this research. In Section four the data is described and the relationship between "survivors" and "non-survivors" regarding our main variables is addressed. Section six reports the main results. Section seven concludes.

2. LITERATURE REVIEW

There is a wide-range of literature analyzing the demography of firms and its determinants as well. For instance, OCDE (2014) finds that the probability that firms in specific markets survive beyond two years goes from 60% to 80%. In other words, the probability of "fail" (firm exit) in *t+2* for firms created in *t* is in the range of 20% to 40%. Even more, only 40% to 50% of firms that were born in the same year survive beyond the seventh year.

López-García and Puente (2006) study the determinants of firm exit in Spain. They conclude that the bankruptcy rate for Spanish firms is lower than that of firms from similar countries like Italy, Germany and the United Kingdom. This finding is confirmed over time, even after being controlled by industries, with the main exception of financial services, insurance, real states and wholesale and retail where firms showed entry and exit patterns similar to those found in neighbor countries. These authors also find that the bankruptcy risk-function has an inverted-U shape, reaching a maximum in the fourth year of activity, which is confirmed for all industries in the economy.

Other studies (see below) reach the same conclusions using a different set of information; thus it is possible to safely rule out the effects of sample bias on these results. Both the survival rate and the inverted-U risk-function might be relevant for explaining the employment evolution and productivity growth, not only in Spain but also in other countries, which is why it is important to consider them in future analysis.

Audretsch, Santarelli and Vivarelli (1999) estimate an unconditional risk-function for manufacturing firms in Italy. They conclude this function increases up to two years and then falls down thereafter. Bhattacharjee (2005) estimates both a conditional and an unconditional risk-function in modelling the probability of survival of those firms traded on the London Stock Exchange. He gets to the conclusion that these firms survive up to three years after their stock exchange opening and exit the market afterwards. Wagner (1994) analyzes the demography of firms in Germany. He finds firms survive up to a maximum of three years before exiting the market. Bartelsman et al. (2003) also find an inverted-U risk-function for the United Kingdom, Italy and the United States using a different information set.

Ericson and Pakes (1998) and Bhattacharjee (2005) argue that an inverted-U risk-function is consistent with the theoretical models of active and passive learning since firms need time to learn about their efficiency. Brüderl and Schüssler (1990) and Fichman y Levinthal (1991) explain that new firms often possess a stock of initial resources that help them "to survive" for a while, a period in which firms can establish new operational structures. Those initial operations might explain why firms take time to comprehend they are not as efficient as they supposed to and, as a result, they must exit the market. The latter is even more evident when firms face high fixed costs in order to initiate their operations. In those cases, firms try to stay in business as much as their initial resources allow them to before taking the decision of shutting down.

Finally, there are studies on firms' demography and survival probabilities that use specific variables for each firm, including industry-specific characteristics. The choice of these variables depends on the economic theory and previous empirical analysis.[1] First, the initial investment of firms and their financial conditions at the time of "birth", or even one year later, are taken into account. For instance, the main findings point out to the existence of a non-lineal relationship between indebtedness and the survival of firms. The sign of such a relationship changes accordingly to the firms´ initial level of indebtedness: if the firm is not highly indebted increasing the level of indebtedness is favorable for the firm survival; on the contrary, increasing the level of debt when the firm is already highly indebted increases the probability of fail.

In what follows, we describe the empirical methodology we use in this paper to address the determinants of firm survival in Chile.

3. EMPIRICAL METHODOLOGY

Our empirical study is based on survival analysis using binary choice models and the Cox´s proportional hazard model for Chilean firms during the period 2011-2015. We first document the determinants of firm survival per each cohort using a probit model of all firms and not only the ones that were born in 2010 and then estimate the proportional hazard model over the entire period using only the sample of firms that were created in 2010. In doing the latter we document the firm survival on a yearly basis and show a more robust method in estimating the survival of firms when there are censored observations in the data.

3.1. Discrete choice models

In the literature is common to answer the dichotomy questions using the probability linear model.[2] However, it is well known that this model poses two important challenges to researchers: a) since the dependent variable is not bounded between 0 and 1, predictions in terms of probability are not useful and

[1] See, for instance, Mata, Portugal and Guimaraes (1995), Geroski, Mata and Portugal (2003), Stiglitz and Weiss (1981), Evans and Jovanovic (1989), Blanchflower and Oswald (1998), Holtz-Eakin, Joulfaian and Rosen (1994), Robert Cressy (1996), Audretsch, Houwelling and Thurik (1997), Weiss (1976).

[2] Wooldridge (2010) shows a complete guide of studies using the probability linear model to answer this kind of questions.

b) linearity does not make much sense conceptually. To walk around those challenges, non-lineal type models are considered. The Probit and Logit setups are two well-known examples.

Consider the following model:

$$\Pr(y=1|x) = G(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad (1)$$

$$\Pr(y=1|x) = G(\boldsymbol{x\beta})$$

where *G* is a function taking values strictly between 0 and 1: 0<*G(z)*<1, for all real numbers z. *G* is the cumulative density function and is monotonous increasing in index z with

$\Pr(y=1|x) \to 1$ **when** $\boldsymbol{x\beta} \to \infty$

$\Pr(y=1|x) \to 0$ **when** $\boldsymbol{x\beta} \to -\infty$

G can be approximated using the logistic distribution, which supports the Logit model, or the normal standard distribution, which supports the probit model. For the Logit model

$G(\boldsymbol{x\beta}) = \frac{\boldsymbol{exp(x\beta)}}{\boldsymbol{1+exp(x\beta)}} = \boldsymbol{\Lambda(x\beta)},$

which is between 0 y 1 for all values of $\boldsymbol{x\beta}$**.** This constitutes the cumulative function for the logistic variable. For Probit, *G* is the normal standard *cdf* expressed as an integral

$G(\boldsymbol{x\beta}) = \Phi(\boldsymbol{x\beta}) = \int_{-\infty}^{x\beta} \phi(v)dv$ **,**

where

$\phi(v) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{v^2}{2}),$

is the standard normal density. Writing *G* in this way makes sure that the probability of success be strictly between 0 y 1 for all values of parameters and regressors.

3.2. The proportional hazard model

Modelling firm exit (survival) using OLS provokes a sampling bias since in this case some firms are more likely to stay in business than others. It is possible to perform a logit or probit analysis on firm survival, but one would need to observe all firms from entry to exit. This generates an addition problem since the sample period ends before most of the firms leave the market. As a result, a censored data problem emerges and we need other methods to tackle it.

The issue when performing survival analysis is the use of the information on survivor firms. A widespread approach uses the proportional hazard model to perform event history analysis. This analysis allows us to study what happens over a time span before some event takes place; in this study, the event is the firm exit. A key process of event history analysis is the specification of the survival function which describes the probability of firms´ survival until a certain time has elapsed.

The survival function is presented as follows

$$S(t) = \Pr(T \geq t) \qquad (2)$$

where *T* is the duration of survival of a firm and *t* is a certain time point. In particular, the function shows the probability of survival at time *t* as a function of *t*. Other important concept is that of the hazard function: it describes the probability of the risk of some event happening. If we denote the probability density function of event occurrence as $f(t)$, then the hazard function can be written in this way

$$\lambda(t) = \frac{f(t)}{s(t)} \tag{3}$$

The hazard function calculates the probability that some event occur (exit) in a lapse of time, conditional on no occurrence of the event until time *t*, which in our case means, conditional on firm survival until time *t*. An important issue to take into account is the specification of the probability distribution of firms' exit. It is not possible to know, *ex ante*, such a distribution, which make problematic to empirically specify the functional form for the hazard function.

Cox (1972, 1975) uses the hazard function to investigate the relationship between the probability that some event happens and several regressors. Under the condition of "hazard proportionality", which defines that the proportion of two kinds of hazard keeps constant over time, the analysis of regressors is developed without specifying a hazard function.

In the proportional hazard model, each sample´s rate $\lambda_i(t)$ is a function of a group of regressors. Conceptually, a) there is a baseline hazard $\lambda_0(t)$ that does not depend on any regressors and b) the proportion of $\lambda_i(t)$ and $\lambda_0(t)$ is constant. The latter is based on the assumption of hazard proportionality. As a result, the proportion of hazards is analyzed as a function of regressors.

Consider the vector of regressors as $\boldsymbol{x_k}$. We can write the proportional hazard model as follows

$$\frac{\lambda_i(t)}{\lambda_0(t)} = \exp(\boldsymbol{\beta x_k}) \tag{4}$$

or

$$\lambda_i(t) = \lambda_0(t)\exp(\boldsymbol{\beta x_k}) \tag{5}$$

Taking logarithm of both sides in (5) we obtain

$$\log\lambda_i(t) = \log\lambda_0(t) + \boldsymbol{\beta x_k} \tag{6}$$

In this set up we analyze the factors that influence the height of hazard rates. A negative regressor coefficient is correlated with a higher probability of survival. On the contrary, a positive regressor coefficient is correlated with a lower probability of survival.

Since we do not know the distribution of the hazard, the baseline hazard is estimated after a regression with all samples. With this information one can estimate the baseline survival function $S_0(t)$ using the following

$$S_0(t) = \exp\{-\Gamma_0(t)\} \tag{7}$$

where $\Gamma_0(t)$ is the cumulative function of the baseline hazard $\lambda_o(t)$. The relationship between $S_0(t)$ and $\lambda_o(t)$ is calculated from equation (3) as $\lambda_o(t) = -\frac{d(\log(S(t))}{dt}$. In this paper, the baseline survival function shows the survival pattern of firms when regressors do not impact their survival. According to (7), the probability of exit is higher in early stages before regressors are considered. Thus, regressors explain the deviations of actual hazard from baseline hazard $\lambda_o(t)$.

4. CONCEPTS AND DATA

Our main source of information is the Internal Revenue Service's anonymized administrative records data base (Annual Income Tax Return) from the Chilean IRS (Servicios de Impuestos Internos). It groups a wide range of information on legal entities (corporations, partnerships, sole proprietors and among others) from all industries in Chile that pay the capital gain tax.

The Central Bank of Chile has been receiving this information since the end of 90s. From 2007 to 2015 we have gathered roughly 25 million of administrative records. In this paper we focus our analysis on the information that covers the period 2010 to 2015. This data base has many potential uses being one of them the computation of a set of indicators related to business dynamics, firm entry and exit and survival patterns. The definitions below are the main concepts we use throughout this paper:

- Firm stock: Number of entities that have remained in business during part or whole year. Firms remaining active at the end of the period and those who shut their operations down during that period are also considered as part of the firm stock.
- Firm birth: Set of entities that have created a combination of new production factors within a year. There is not relationship between these set of new entities and previous existent ones or whatsoever. We say that the firm was born in period *t* if and only if we do not observe it in *t-j*.
- Firm survival: It is the set of entities that keep in business five years after the time of birth.

Furthermore, we can build some useful definitions from the concept above. In this regard we have:

$$\textbf{Birth rate:}\qquad \frac{\sum N^{ti}}{\sum T^{t}_{i}} \qquad \begin{array}{l}\text{Enterprise births in the sector i, at the year t}\\ \text{Stock of enterprises in the sector i, at the year t}\end{array}$$

$$\textbf{Survival rate:}\qquad \frac{\sum S^{t+k}}{\sum T^{t}_{i}} \qquad \begin{array}{l}\text{Enterprise survivals in the sector i, year t+ k= 1,…, 5}\\ \text{Stock of enterprises in the sector i, at the year t}\end{array}$$

Also, we consider total assets, total liabilities, turnover and total costs to build a set of economic activity (Gross margin), leverage (Indebtedness), which also can be understood as a measure of access to credit, and profitability (Return on Assets) ratios we later use in our estimation. Gross margin is a firm's total sales revenue minus its cost of goods sold divided by total sales revenue, expressed as a percentage. Indebtedness is the difference between total assets and firm equity divided by total assets. Finally, the Return on Assets is the ratio of profits to total assets. We also use the Inventory turnover ratio. It is an efficiency calculation used to control and manage turns by comparing cost of goods sold and average inventory.

Just like any other administrative register data, the information we use is affected by different types of errors. Most frequently are records with very high or low values and presence of missing information. As a first step we use a statistical procedure to detect the outliers in our data and then we run an imputation method to fill the missing records.[3]

In order to have a better comprehension of survival of firms we grouped them in our sample using both an industry and a size classification. At the industry level we create five groups: Agriculture-fishing, Mining-manufacturing-utilities[4] (MMU), Construction, Wholesale and retail trade-repair of motor vehicles and motorcycles-accommodation and food service activities (Trade) and Services[5]; at the size level we use two groups of firms: micro-small and medium-large.[6] We later control for the effects of both classification groups in the probability of firm survival.

Table 1 below summarizes the information on the number of firms from 2010 to 2015 by industry and size. First, notice we work with more than 600 thousands firms per year on average. Most of them are "Trade" and "Services" firms, which together represent almost an 80% of the entire sample. Also, it is important to notice how "Services" has been increasing its importance overall whereas "Trade" has been reducing it: in 2015 the proportion of "Trade" and "Services" was the inverse of that of 2010. Furthermore, around 20% of firms are producers of good, with almost 50% of them being part of MMU.

With respect to firm demography, 91 thousand firms were created in Chile in 2010, which accounts for a 14.1% of total stock of firms in that year. By industries, "Services" show the higher creation rate with a 15.7% whereas "Trade" shows the lower creation rate, accounting for a 12.6%. The left side of Table 1 shows the survival rates over the time span. The first year (2011), the survival rate reached 80.4% with slightly differences between industries. In the second year, 62,567 from a total of 91,067 firms "survived", which represents a 68.7%. Again, the differences between industries are not relevant; however there is a marginal difference between those firms that produce goods and those producers of services.

Overall, five years after the entry of a firm, the survival rates are almost 46%. This number is slightly larger for "MMU" (48.4%) and slightly smaller for "Construction" and "Trade" (44.9% and 44.7% respectively). These rates are located at the bottom of the statistical distribution showed by Eurostat (2007) and are consistent with the statistical fact that producers of services are more dynamic than producers of goods.

[3] We cluster the data using a defined classification by industry and size and detect the outliers using the Tukey´s (1972) method and impute the missing information using a median within each group of firms.

[4] By "Utilities" we mean electricity, gas, steam and air conditioning supply and water supply, sewerage, waste management and remediation activities.

[5] "Services" include transportation and storage, information and communication, financial and insurance activities, real estate activities, professional, scientific and technical activities and administrative and support service activities.

[6] Our definition of "size" is based on firm's level of turnover, in contrast to other studies that use firms' number of workers instead. This is mainly due to the lack of information on numbers of workers at the firm level. In this sense we define micro-small as those firms with level of turnover<1000.000 of USD and medium-large as those firms with level of turnover>=1000.000 USD in 2010 figures.

Table 1 Number of firms, Cohort 2010 and Survival by Industry

Number of Firms by Broad Economic Sectors: 2010-2015						Birth Rate					
						(% total)	% of survival				
	2011	2012	2013	2014	2015	2010	2011	2012	2013	2014	2015
Agriculture & Fishin	30.087	30.603	30.782	30.976	32.862	13,8	78,1	68,7	60,1	52,3	45,5
MMU	60.366	62.298	62.951	65.201	69.170	14,7	82,2	71,4	62,6	55,0	48,4
Construction	39.059	40.708	43.162	45.814	48.603	14,7	82,1	70,3	60,4	52,3	44,9
Trade	272.565	272.945	274.075	272.866	289.478	12,6	81,4	68,4	59,1	51,4	44,7
Services	255.155	277.098	294.376	304.134	322.650	15,7	79,1	68,1	59,6	52,4	45,6
Total	657.232	683.652	705.346	718.991	762.763	14,1	80,4	68,7	59,8	52,2	45,5

Source: own calculations

The analysis by firms' size shows interesting results too. Table 2 below encompasses this information. For instance, Micro-Small firms represent a 95% of total stock of firms for the entire period. In 2010, 14.4% of total "births" corresponded to Micro-Small firms and only a 7.6% fraction to Medium-Large ones. Nevertheless, the latter group shows the higher rates of survival; certainly, after five years of creation, 84.6% of medium-large firms were still in the market, which represents almost twice the percentage showed by micro-small firms. Thus the size is relevant when explaining the patterns of firms´ survival.

Table 2 Number of firms, Cohort 2010 and Survival by Size

Number of Firms by Size: 2010-2015						Birth Rate					
						(% total)	% of survival				
	2011	2012	2013	2014	2015	2010	2011	2012	2013	2014	2015
Micro & Small	625.166	197.436	669.837	682.441	725.175	14,4	79,9	52,6	46,1	51,3	44,5
Medium & Large	32.066	34.164	35.509	36.550	37.588	7,6	98,5	95,5	9,2	88,5	84,6
Total	657.232	231.600	705.346	718.991	762.763	14,1	80,4	53,6	45,2	52,2	45,5

Source: own calculations

4.1. DETERMINANTS OF FIRM SURVIVAL IN CHILE

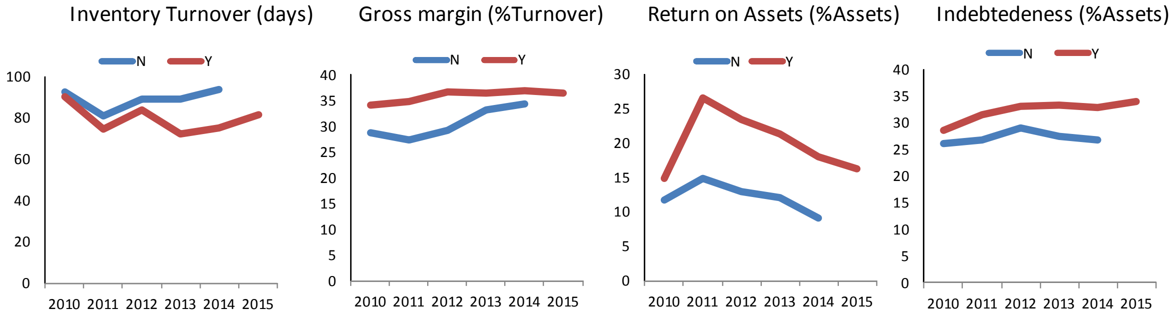
Here we show the pattern of firm survival by broader economic industries and size. Just to clear the things up: in all graphs below the "N (blue line)" refers to "non-survivor" firms whereas the "Y (red line)" refers to "survivor" firms.

4.1.1. Descriptive

In general those firms that survive show a lower Inventory turnover, higher leverage rates, higher Gross margin and higher profitability. The latter is highly significant. Notice that the difference between "survivors" and "non-survivors" increases over the years for all variables considered with the only exception of Gross margin (See Graph 1 below).

Graph 1 Economic activity, financing and profitability 2010-2015: All

(Cohort 2010, median)



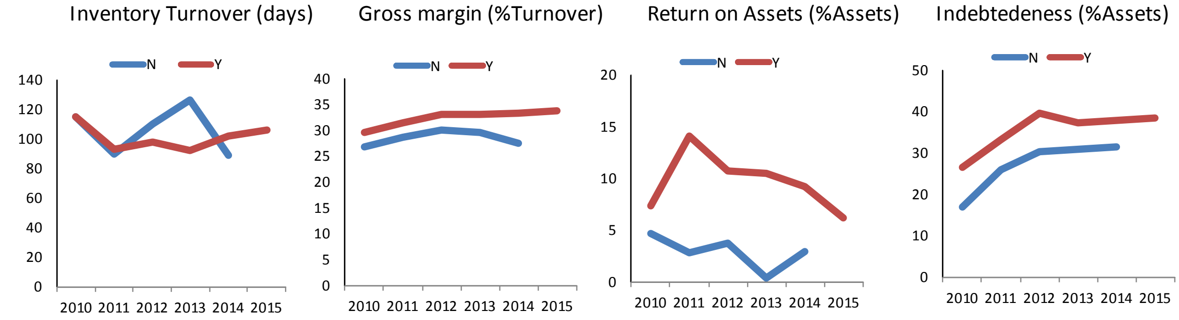
The results above are general in that they consider the whole group of firms. A more interesting issue is to disentangle the dynamics of survival by industry and size. After all, we expect relevant differences to emerge between firms of different size or pertaining to different industries.

4.1.1.1. By industry

In Agriculture-fishing we observe that "non-survivor" firms have lower leverage rates compared to that of "survivors". They also enjoy a relatively lower Gross margin, lower rate of profitability and higher Inventory turnover rates. The latter is less conclusive; in 2010, 2011 and 2014 the Inventory turnover is the same for "survivor" and "non-survivor" firms.

Graph 2 Economic activity, financing and profitability 2010-2015: Agriculture-fishing

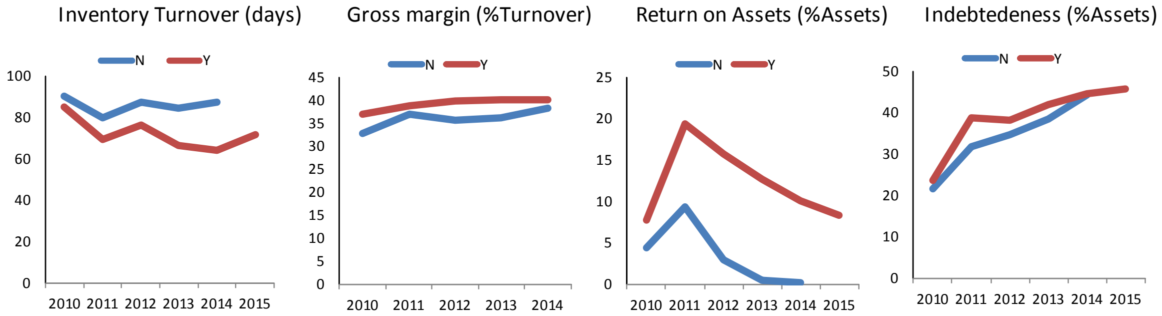
(Cohort 2010, median)



Within MMU (Mining, Manufacturing-Utilities) the firms that survive are characterized by lower rates of Inventory turnover, significant differences in profitability, higher Gross margin and higher leverage. Yet, the differences in the degree of indebtedness seem to be inconclusive, or at least not highly significant, for the entire period of analysis (See Graph 3).

Graph 3 Economic activity, financing and profitability 2010-2015: MMU

(Cohort 2010, median)

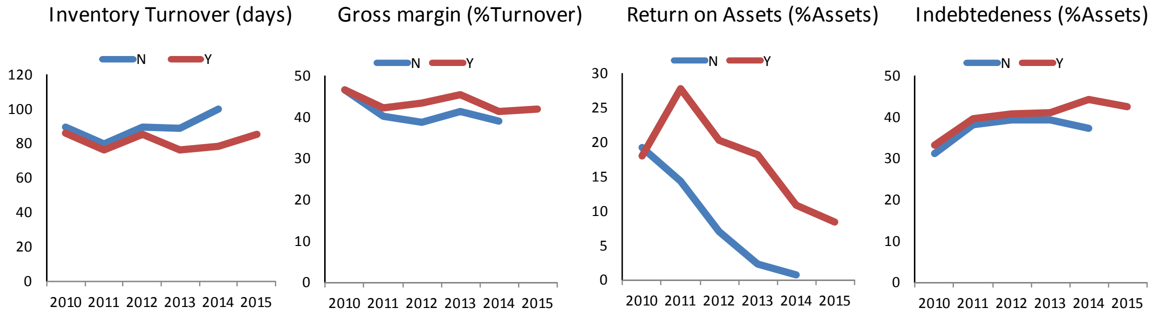


Inventory Turnover (days) Gross margin (%Turnover) Return on Assets (%Assets) Indebtedeness (%Assets)

For Construction it is important to notice how the level of profitability of those firms that do not survive rapidly falls after the first year. Actually it falls for both groups of firms but in the case "survivors" the falling starts after the second year. Notice also, the huge differences in levels for this variable. As for the rest of the indicators, Gross margin is higher for "survivors" and so it is the leverage rate, whereas the Inventory turnover remains the same for both groups until 2012 and then a gap emerges afterwards (Graph 4 below).

Graph 4 Economic activity, financing and profitability 2010-2015: Construction

(Cohort 2010, median)

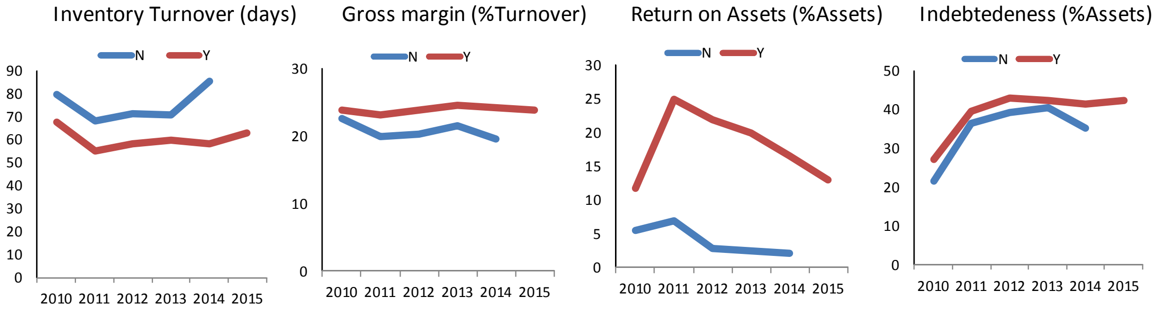


Inventory Turnover (days) Gross margin (%Turnover) Return on Assets (%Assets) Indebtedeness (%Assets)

The "survivor" firms within Trade show lower rates of Inventory turnover rate, higher Gross margin and profitability. There is, however, a close relationship between "survivors" and "non-survivors" regarding their leverage until 2013; although, the firms that survive clearly show higher rates of indebtedness.

Graph 5 Economic activity, financing and profitability 2010-2015: Trade

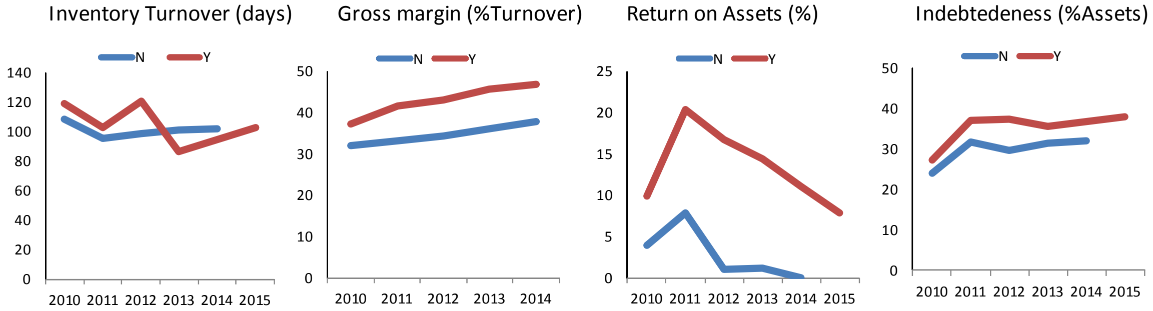
(Cohort 2010, median)



Finally, in Services "survivors" enjoy higher leverage rates than "non-survivors" and a higher Gross margin rate, particularly toward the final periods. The result from Inventory turnover is not informative. Also the level of profitability looks rare in this case. All in all, profitability is higher for those firms that survive.

Graph 6 Economic activity, financing and profitability 2010-2015: Services

(Cohort 2010, median)



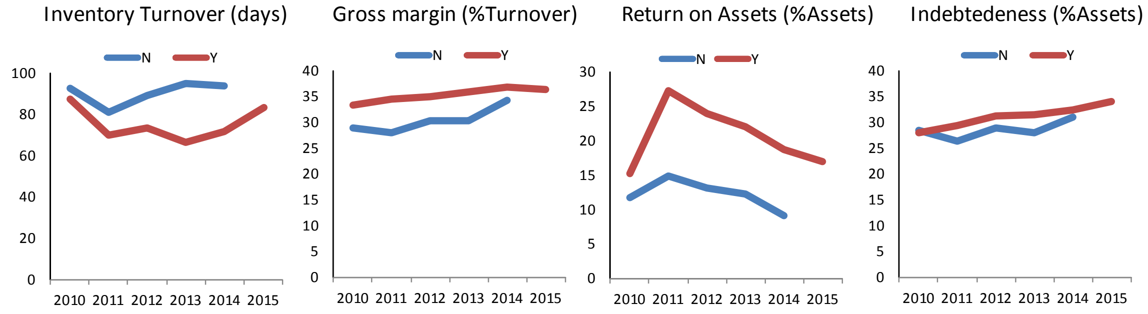
To sum up, there are noticeable differences between the set of "survivors" and "non-survivors" in all variables considered. In general, those firms that survive show higher rates of Gross margin, a higher access to credit markets and profitability is also higher. The Inventory ratio is not statically significant in some cases and in others is not informative at all.

4.1.1.2. By size

The analysis above is also interesting when looking at the differences between "survivors" and "non-survivors" by size. As it can see in Graph 7 micro-small "survivors" show higher results in terms of economic activity, leverage and profitability when compared to "non-survivors". The latter is highly significant for the entire period of analysis.

Graph 7 Economic activity, financing and profitability 2010-2015: Micro-Small

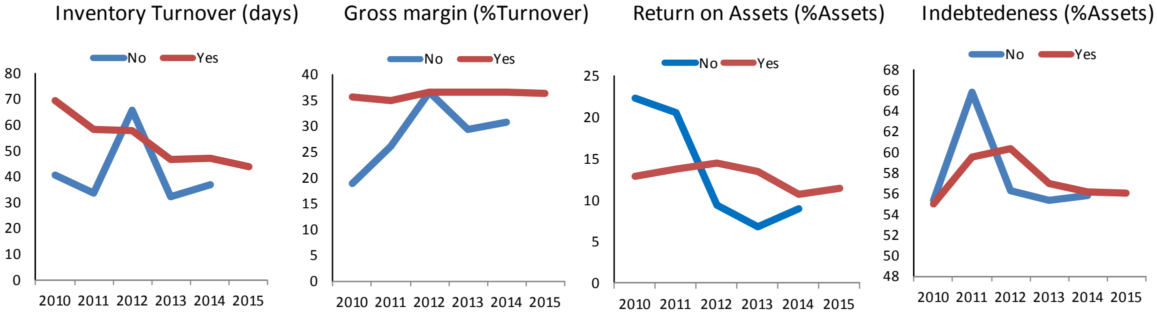
(Cohort 2010, median)



On the other hand, the medium-large firms that remain in business are characterized by higher rates of Gross margin; notice the those firms show a lower level of indebtedness during the first two years and then the relation with "non-survivors" changes afterwards. This is interesting: it seems like those firms that have a low level of indebtedness at first are more likely to survive than those who were highly indebted at the same moment. The profitability ratio starts lower for "survivors" but after 2011 it remains higher. The result for Inventory turnover is odd. Probably, the low numbers of firms in this group is provoking this kind of result.

Graph 8 Economic activity, financing and profitability 2010-2015: Medium-Large

(Cohort 2010, median)



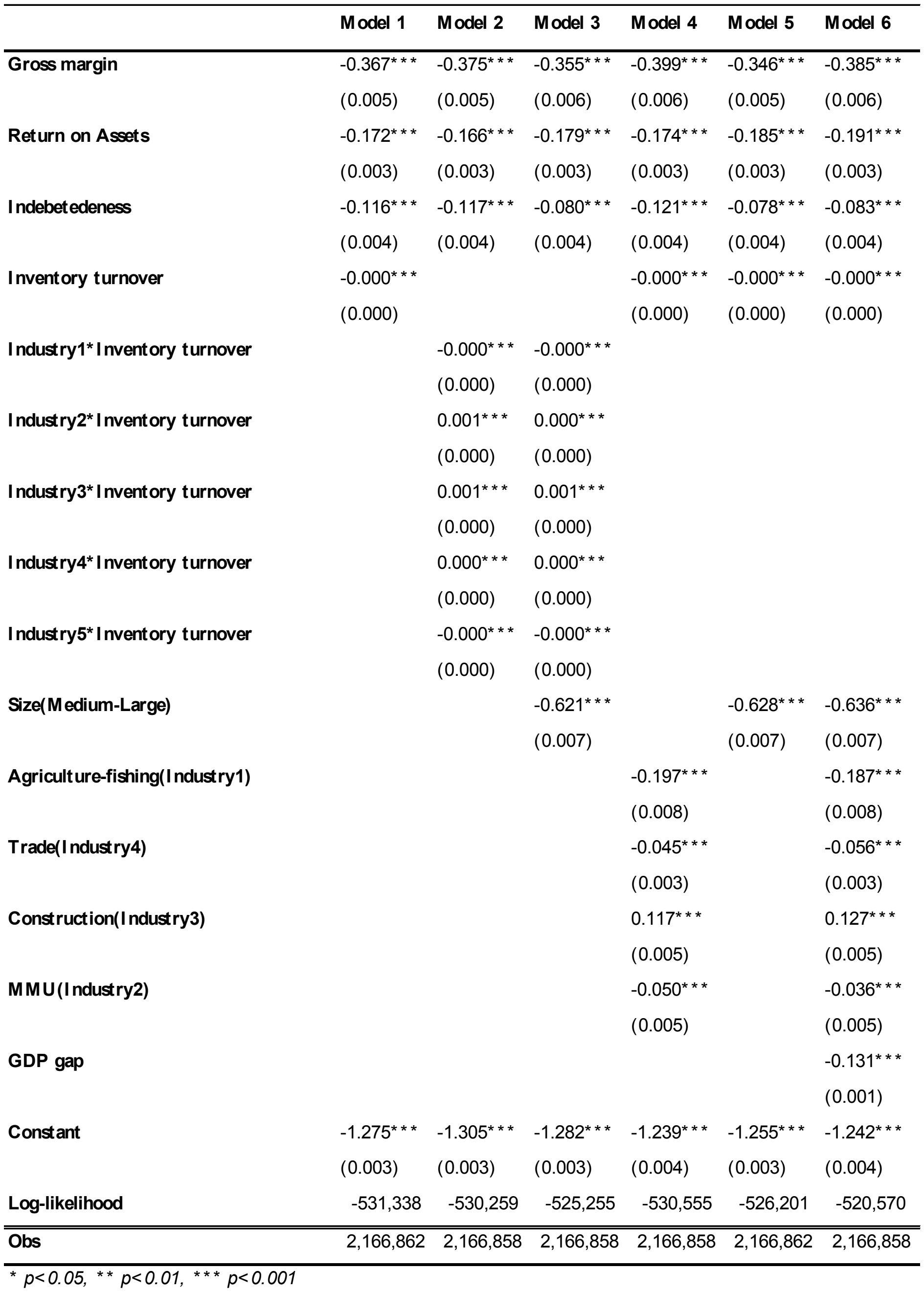
5. MAIN RESULTS

In this section we present the main results obtained from the estimation of our proportional hazard model and discuss their implications. For comparative reasons, we estimate the probability of firm survival for the entire stock of firms using a probit model.

5.1. Probit estimates

In general, the results provided are stable and deliver evidence of our intuition. Table 3, 4 and 5 below shows the estimates from the probit model. We estimate the probability of survival overall and for each year in the sample. The Model 1 is the "basic model" since we use only four regressands in the estimation. The signs of the coefficients are as expected, with the only exception of the Inventory turnover ratio. The latter might be explained, first, by the fact we do not control for firm entry in our model and second, by the fact that Inventory turnover is not an efficiency proxy for some activities, like agriculture and fishing, where business are affected by seasonality factors.

Table 3 Probit estimates: overall



	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Gross margin	-0.367***	-0.375***	-0.355***	-0.399***	-0.346***	-0.385***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)	(0.006)
Return on Assets	-0.172***	-0.166***	-0.179***	-0.174***	-0.185***	-0.191***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Indebetedeness	-0.116***	-0.117***	-0.080***	-0.121***	-0.078***	-0.083***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Inventory turnover	-0.000***			-0.000***	-0.000***	-0.000***
	(0.000)			(0.000)	(0.000)	(0.000)
Industry1*Inventory turnover		-0.000***	-0.000***			
		(0.000)	(0.000)			
Industry2*Inventory turnover		0.001***	0.000***			
		(0.000)	(0.000)			
Industry3*Inventory turnover		0.001***	0.001***			
		(0.000)	(0.000)			
Industry4*Inventory turnover		0.000***	0.000***			
		(0.000)	(0.000)			
Industry5*Inventory turnover		-0.000***	-0.000***			
		(0.000)	(0.000)			
Size(Medium-Large)			-0.621***		-0.628***	-0.636***
			(0.007)		(0.007)	(0.007)
Agriculture-fishing(Industry1)				-0.197***		-0.187***
				(0.008)		(0.008)
Trade(Industry4)				-0.045***		-0.056***
				(0.003)		(0.003)
Construction(Industry3)				0.117***		0.127***
				(0.005)		(0.005)
MMU(Industry2)				-0.050***		-0.036***
				(0.005)		(0.005)
GDP gap						-0.131***
						(0.001)
Constant	-1.275***	-1.305***	-1.282***	-1.239***	-1.255***	-1.242***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)
Log-likelihood	-531,338	-530,259	-525,255	-530,555	-526,201	-520,570
Obs	2,166,862	2,166,858	2,166,858	2,166,858	2,166,862	2,166,858

** p<0.05, ** p<0.01, *** p<0.001*

Source: own calculations

In the Model 2 we add interactions of the Inventory turnover ratio and each industry dummy. Notice that in the case of Agriculture-fishing the sign of the coefficient is negative, this is also true for Services. For

MMU, Construction and Trade the sign is positive. In general, in Agriculture-fishing and Services the more time firms spend in selling their inventories the higher their probability of survival overall. Nevertheless, the value of the coefficients is rather low for all interactions and even for Inventory turnover itself. As a result, we drop this variable from the rest of our estimations.

Additionally, we include a variable of size and the GDP gap for the period under analysis (Model 3 to Model 5) together with dummies for each industry. The coefficients of Gross margin, indebtedness and profitability remain stable. Additionally a larger size of firms is associated to a higher probability of survival. Other things being equal, firms are more likely to survive in Agriculture-fishing, MMU and Trade industries compared to Services. The contrary is true for Construction. These results are strongly significant at a 95% level of statistical significance.

Table 4 Probit estimates: 2011-2014

	Model 7	Model 8	Model 9	Model 10
Gross margin	-0.324***	-0.368***	-0.333***	-0.319***
	(-0.012)	(-0.011)	(-0.011)	(-0.01)
Return on Assets	-0.195***	-0.197***	-0.203***	-0.188***
	(-0.006)	(-0.006)	(-0.006)	(-0.006)
Indebetedeness	-0.078***	-0.089***	-0.112***	-0.066***
	(-0.008)	(-0.008)	(-0.008)	(-0.008)
Size(Medium-Large)	-0.721***	-0.709***	-0.631***	-0.649***
	(-0.018)	(-0.017)	(-0.015)	(-0.014)
Agriculture-fishing(Industry1)	-0.196***	-0.199***	-0.188***	-0.159***
	(-0.018)	(-0.017)	(-0.017)	(-0.016)
Trade(Industry4)	0.005	-0.026***	-0.025***	-0.019**
	(-0.007)	(-0.007)	(-0.007)	(-0.006)
Construction(Industry3)	0.153***	0.154***	0.135***	0.175***
	(-0.012)	(-0.012)	(-0.011)	(-0.01)
MMU(Industry2)	-0.012	0.015	-0.025*	0.020*
	(-0.011)	(-0.01)	(-0.01)	(-0.01)
Constant	-1.215***	-1.137***	-1.098***	-1.065***
	(-0.008)	(-0.008)	(-0.007)	(-0.007)
Log-likelihood	-105,998	-114,919	-123,321	-133,091
Obs	405,037	416,729	425,293	422,862

** p<0.05, ** p<0.01, *** p<0.001*

Source: own calculations

On a yearly basis (See Model 7 to Model 14) the results above remain stable. We remove the GDP gap from the estimations since the effect is irrelevant in this particular set up. Again, the coefficient signs from Gross margin, indebtedness and profitability are in line with our intuition and are highly significant. Finally, Construction is the only industry where firms have a lower probability of survival compare to Services.

Table 5 Probit estimates: 2011-2014

	Model 11	Model 12	Model 13	Model 14
Gross margin	-0.325***	-0.370***	-0.336***	-0.320***
	(-0.012)	(-0.011)	(-0.011)	(-0.01)
Return on Assets	-0.188***	-0.187***	-0.194***	-0.180***
	(-0.006)	(-0.006)	(-0.006)	(-0.006)
Indebetedeness	-0.115***	-0.127***	-0.148***	-0.105***
	(-0.008)	(-0.008)	(-0.008)	(-0.008)
Agriculture-fishing(Industry1)	-0.205***	-0.210***	-0.202***	-0.175***
	(-0.018)	(-0.017)	(-0.017)	(-0.016)
Trade(Industry4)	0.014*	-0.018**	-0.018**	-0.013*
	(-0.007)	(-0.007)	(-0.007)	(-0.006)
Construction(Industry3)	0.139***	0.136***	0.122***	0.165***
	(-0.012)	(-0.012)	(-0.011)	(-0.01)
MMU(Industry2)	-0.032**	-0.005	-0.043***	0.003
	(-0.011)	(-0.01)	(-0.01)	(-0.01)
Constant	-1.238***	-1.161***	-1.121***	-1.089***
	(-0.008)	(-0.008)	(-0.007)	(-0.007)
Log-likehood	-107,110	-116,155	-124,468	-134,442
Obs	405,037	416,729	425,293	422,862

** p<0.05, ** p<0.01, *** p<0.001*

Source: own calculations

5.2. Proportional hazard estimates

The results related in Table 6 are consistent with the ones obtained in Table 3, 4 and 5. The main difference here is we address the probability of survival using the proportional hazard model and focus on the set of firms that were born in 2010.

First, the signs of our main variables remain strong. A higher Gross margin, for instance, is related with a higher probability of survival. This is intuitive; a Gross margin can be used as a measure of firm efficiency; as a result, the more efficient is the firm the higher its probability of remaining in business, *ceteris paribus*. Notice that the sign of the Gross margin coefficient remains strong in each of the five models considered; which means that, regardless the industry and the size of the firm, a higher Gross margin is always a good new for enhancing the probability of survival.

Table 6 Proportional hazard model estimates

	Model 1	Model 2	Model 3	Model 4	Model 5
Gross margin	-0.708***	-0.715***	-0.685***	-0.710***	-0.681***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
Return on Assets	-0.347***	-0.340***	-0.346***	-0.341***	-0.346***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Indebtedeness	-0.528***	-0.524***	-0.528***	-0.472***	-0.476***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
GDP gap		0.093***	0.092***	0.097***	0.096***
		(0.007)	(0.007)	(0.007)	(0.007)
Agriculture-fishing			-0.420***		-0.433***
			(0.046)		(0.046)
Trade			-0.095***		-0.117***
			(0.027)		(0.027)
MMU			-0.203***		-0.204***
			(0.033)		(0.033)
Construction			-0.190***		-0.214***
			(0.027)		(0.027)
Size(Medium-Large)				-1.478***	-1.481***
				(0.061)	(0.061)
Obs	206192	206192	206192	206192	206192

**** p< 0.05, ** p< 0.01, *** p< 0.001***

Source: own calculations

Second, firms with a higher degree of profitability are more likely to remain in business than firms with a lower degree of profitability. Truly, the Return on Assets coefficient is negative and highly significant at a 95% level of significance. This result remains stable after being controlled by other variables.

Third, the probability of survival is positive related to a higher degree of indebtedness. Say it differently, higher indebted firms are more likely to remain in business than lower indebted ones. According to the financial theory a highly leveraged firm is the one who enjoys tax deduction benefits and also is able to send a good signal into financial markets; putting all these together results in a higher value firm which means the probability of survival is higher.

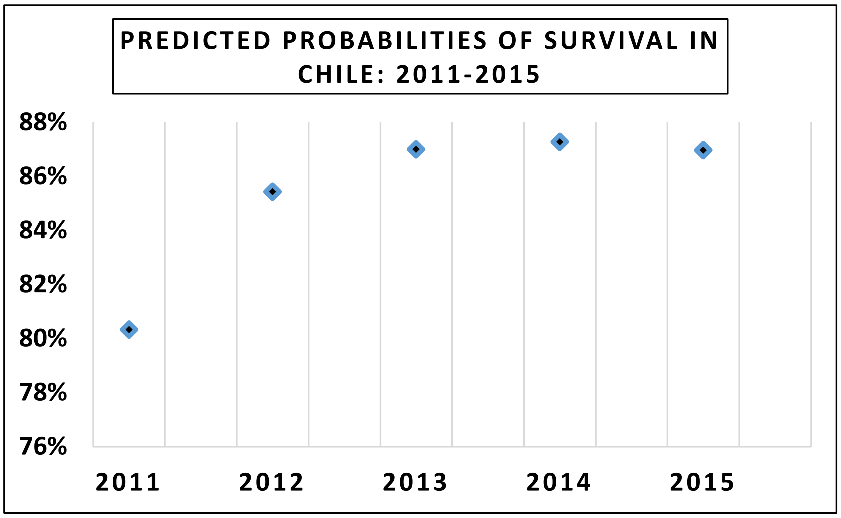
Fourth, larger firms are more likely to survive than smaller ones. Notice, also, that the coefficient on Gross margin is reduced when controlling by size. This is not surprising since we build the size of firms using turnover, which is also the same variable we employed to calculate the Gross margin.

In general a higher GDP gap is correlated with a lower probability of survival. The coefficient remains the same in all of the five models presented. A higher GDP gap (negative or positive) suggest that the economy is working inefficiently. Accordingly, there might be a negative effect on firm survival since economics conditions are adverse.

Trade is the more dynamic industry and Agriculture-fishing is the less dynamic one when comparing to Services. The probabilities of survival for both industries are 0.11% and 0.43% respectively. Construction and MMU, on the other hand, show probabilities of survival of 0.2% approximately.

Finally, we can see in Graph 9 and 10 the shape of our predicted probabilities. Overall, the predicted probability of survival is almost 80% in the first year; it reaches its peak in 2014, four years after the "birth" of the firms, and then falls afterwards (See Graph 9).

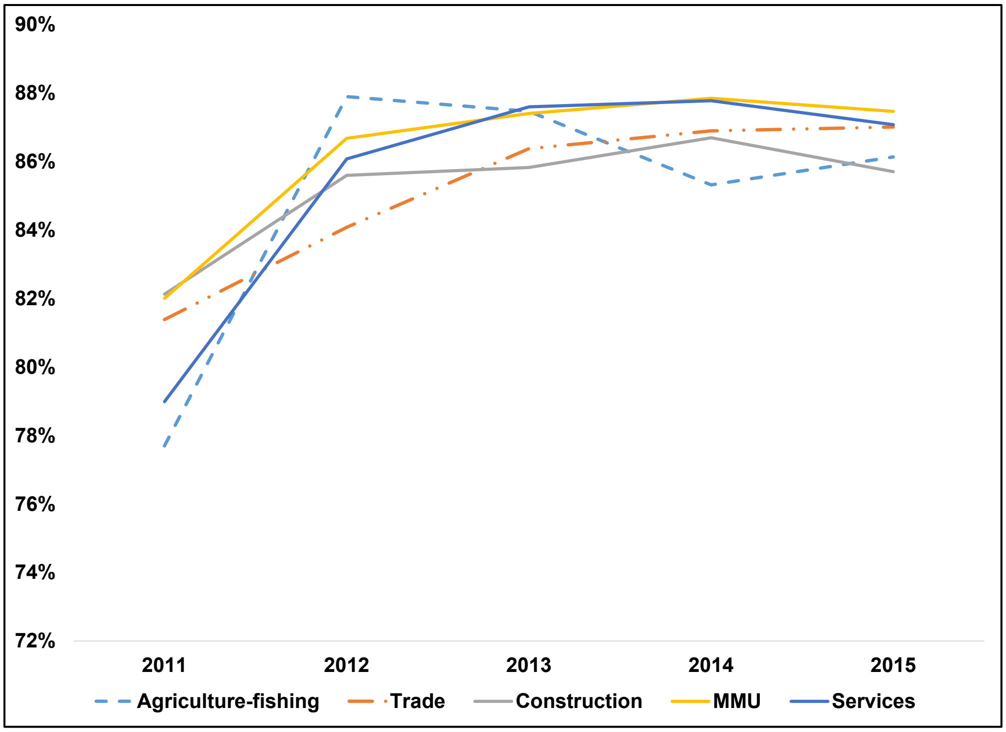
Graph 9 Inverted-U shape survivals



Source: own calculations

Looking at the probability of survival at the industry level the results are diverse. For Agriculture-fishing the peak is reached two years after the "birth", falling immediately afterwards. The same seems to be true for Construction, although in this case the tendency is not clear. In Trade and Services the survival rate reaches its peak in 2014, four years after the firms´ "birth"; the latter is also true for MMU. In conclusion Trade, Service and MMU seems to be leading the dynamic in terms of firm survival overall (See Graph 10).

Graph 10 Inverted-U shape survivals by industry



Source: own calculations

6. CONCLUSION

In this paper we analyze the survival of Chilean firms using information from the Chilean IRS for the period 2010-2015. The importance of this analysis relies on the fact that we take a sample of firms that were born in the same year, share similar characteristics and yet show a different pattern of survival over time.

The results confirm our intuition regarding the impact of firms´ economic and financial variables in their probability of survival. First, firms are more likely to stay in business the higher their Gross margin, the higher their level of indebtedness and the higher their profitability level. This is confirmed in both the probit and the proportional hazard model. All results are highly significant at 95% level of statistical significance.

In the case of Return on Assets and Gross margin, the correlation is very intuitive; after all, both variables are indicators of firm´s efficiency and profitability respectively. For Indebtedness the result seems to be odd at a first sight, however, this goes in line with the financial theory related to the effect of leverage on economic performance of firms. All results remain stable after being controlled by other variables.

Second, the probability of survival increases with the firm size: larger firms are more likely to survive than smaller ones. This finding is similar to others using a different set of information. Notice also this result remains stable for the entire period 2011-2015 and on a yearly basis as well.

At the industry level firms face higher probability of survival in Agriculture-fishing than that in Trade. Furthermore, comparative speaking, there is not difference in being part of Services or MMU regarding the survival of firms. Both industries show similar probabilities of survival.

Finally, the pattern of survival of Chilean firms is similar to that in other countries. Overall, the survival rate is 80% in the first year after the firm "birth" and reaches its peak in the fourth year. At the industry level the results are mainly different; however some similarities between industries emerge. The survival in Trade, Services and MMU reaches its maximum point in the fourth year, whereas in Agriculture-fishing the maximum point is in the second year.

7. REFERENCES

Audretsch, Houwelling and Thurik (1997)." New Firm Survival: Industry versus Firm Effects". No 97-063/3, Tinbergen Institute Discussion Papers from Tinbergen Institute.

Audretsch, Santarelli and Vivarelli (1999). "Start-up size and industrial dynamics: some evidence from Italian manufacturing", International Journal of Industrial Organization, 1999, vol. 17, issue 7, 965-983.

Bartelsman et al. (2003), "Comparative Analysis of Firm Demographics and Survival: Micro-Level Evidence for the OECD Countries", OECD Economics Department Working Paper No. 348, OECD, Paris.

Benavente, J.M. (2008). La dinámica empresarial en Chile (1999-2006). Ministerio de Economía, FUNDES e INTELIS.

Bhattacharjee (2005)." The Effect of P2P File Sharing on Music Markets: A Survival Analysis of Albums on Ranking Charts"

Blanchflower and Oswald (1998). "What Makes an Entrepreneur?", Journal of Labor Economics, 1998, 16(1), pp. 26-60.

Brüderl and Schüssler (1990)." Organizational Mortality: The Liabilities of Newness and Adolescence". Administrative Science Quarterly Vol. 35, No. 3 (Sep., 1990), pp. 530-547

Crespi, G. (2003). "PYME en Chile: nace, crece y… muere: Análisis de su desarrollo en los últimos siete años". FUNDES Chile.

Cox, D. R. (1972)." The Cox Proportional Hazards model", Journal of the Royal Statistical Society 34:187-220.

Cox, D. R. (1975). "Partial likelihood", Biometrika, Vol. 62, No. 2 (Aug., 1975), pp. 269-276.

Correa, C. y Echavarría, G. (2013), "Estimación del aporte de las Pyme a la actividad en Chile, 2008-2011". Estudio Económico Estadístico N°101, 2013.

EUROSTAT (2012) "*Structural business statistics overview*".

Ericson and Pakes (1998)." Empirical Implications of Alternative Models of Firm Dynamics". Journal of Economic Theory, 1998, vol. 79, issue 1, 1-45

Evans and Jovanovic (1989). "An Estimated Model of Entrepreneurial Choice under Liquidity Constraints", Journal of Political Economy Vol. 97, No. 4 (Aug., 1989), pp. 808-827.

Fichman y Levinthal (1991). "Honeymoons and the Liability of Adolescence: A New Perspective on Duration Dependence in Social and Organisational Relationships". Academy of Management Review 16:442-468

Geroski, Mata and Portugal (2003). "Founding Conditions and the Survival of New Firms".

Hermann Consultores (2013), "Dinámica Empresarial 2006-2012".

Holtz-Eakin, Joulfaian and Rosen (1994). "Entrepreneurial Decisions and Liquidity Constraints", The RAND Journal of Economics Vol. 25, No. 2 (Summer, 1994), pp. 334-347

Informa (2012), "Evolución de la Demografía Empresarial en España 2007-2012. Comparativa con Alemania, Francia, Italia y Portugal".

Joachim Wagner (2012)." Exports, imports and firm survival: first evidence for manufacturing enterprises in Germany"

Kimura Fukunari and Takamune Fujii (2003), "Globalizing activities and the rate of survival: Panel data analysis on japenese firms", 2003.

López-García, P., and S. Puente (2006)."Business Demography in Spain: Determinants of firm survival".

Mata, Portugal and Guimaraes (1995)." The survival of new plants: Start-up conditions and post-entry evolution". International Journal of Industrial Organization", 1995, vol. 13, issue 4, 459-481.

Nunes, A. and E. Sarmento (2010), "Business Demography Dynamics in Portugal: A Non-parametric Survival Analysis"

Nuñez, S. (2004), "Salida, entrada y tamaño de las empresas españolas", Boletín económico de Marzo, 2004.

OCDE (1998), "Industrial structure statistics", 1998.

OCDE (2000), "Structural Statistics for Industry and Services Edition, 2000".

OCDE (2006), "Structural Statistics for Industry and Services Edition, 2006".

OCDE, EUROSTAT (2007). "Eurostat-OECD Manual on Business Demography Statistics"

OCDE (2009), "Business Demography: employment and survival".

OCDE (2009), "Structural and Demographic Business Statistics".

Pérez, J. (2010), "Una caracterización de las empresas privadas no financieras de Chile", Estudio Económico Estadístico N°83, 2010.

Robert, C. (1996). "Are Business Startups Debt-Rationed?" Economic Journal, 1996, vol. 106, issue 438, 1253-70.

Stiglitz and Weiss (1981). "Credit Rationing in Markets with Imperfect Information", American Economic Review, 1981, vol. 71, issue 3, 393-410.

Tukey, J. (1972). "Data analysis, computation and mathematics", Special issue, symposium on "The future of applied mathematics".

Wagner, J. (1994), "The Post-Entry Performance of New Small Firms in German Manufacturing Industries", Journal of Industrial Economics, Vol. 42, No. 2, pp. 141-154.Bartelsman et al. (2003).

Wooldridge (2010). "Econometric Analysis of Cross Section and Panel Data", Second Edition.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Determinants of business demography in Chile over the 2010-15 period[1]

Beatriz Velasquez,
Central Bank of Chile

[1] This presentation was prepared for the meeting. 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 meeting.

Determinants of business demography in Chile over the 2010 – 2015 period

Central Bank of Chile – Statistics Division

Diana López
Daymler O'Farril
Josué Pérez
Beatriz Velasquez

Outline of the presentation

- Motivation
- Background
- Data
- Preliminary Findings

Motivation

- Using Big Data to exploit the potential of the Internal Revenue Service’s anonymized administrative records data base.

- Using Big Data to find evidence on the probability of survival of non-financial enterprises in Chile by analyzing the reasons behind their definitive closure.

Background

- The Central Bank has been receiving the tax records data base from the Chilean IRS by way of special orders since 2010 and only since 2015 through a formal agreement.

- This formalization has allowed the creation of an historical data base of roughly 25 million administrative registers from the year 2007 to 2015. The software that managed this DB was also acquired in 2015, and it is SAS.

- This data base has many potential uses, being one of them, computing a set of indicators related to business dynamics, describing births, deaths and survival patterns of companies.

- In 2014 Suazo and Pérez presented the first paper of Chilean business demography, based on this data.

Data

- Annual income tax returns (anonymized) filed by individuals and corporations from the calendars years 2007 to 2015. Source: Chilean Internal Revenue Service (IRS).

- Around 25 million of registers including individuals and firms, of which, approximately 7 million correspond to enterprises.

- Corporations fill a balance sheet reduced version too. Decision rule: Corporate tax $\neq 0$.

- Issues in use of this administrative data: lack of quality control over the data and missing items or missing records

Data: Imputation Process

1. Append to the corporate tax record data base the economic sector according to National Accounts & size stratum.
2. Calculus of 15 financial ratios with two pivots: Total assets, sales income. These ratios are standardized and clustered by economic sector (80), business structure (10) and size stratum (4).
3. Extreme values deletion by cluster. Threshold +/-2.5 standard deviation.
4. Compute the M-Estimator for the 15 financial ratios within each cluster.
5. Ratio imputation of missing values in levels.

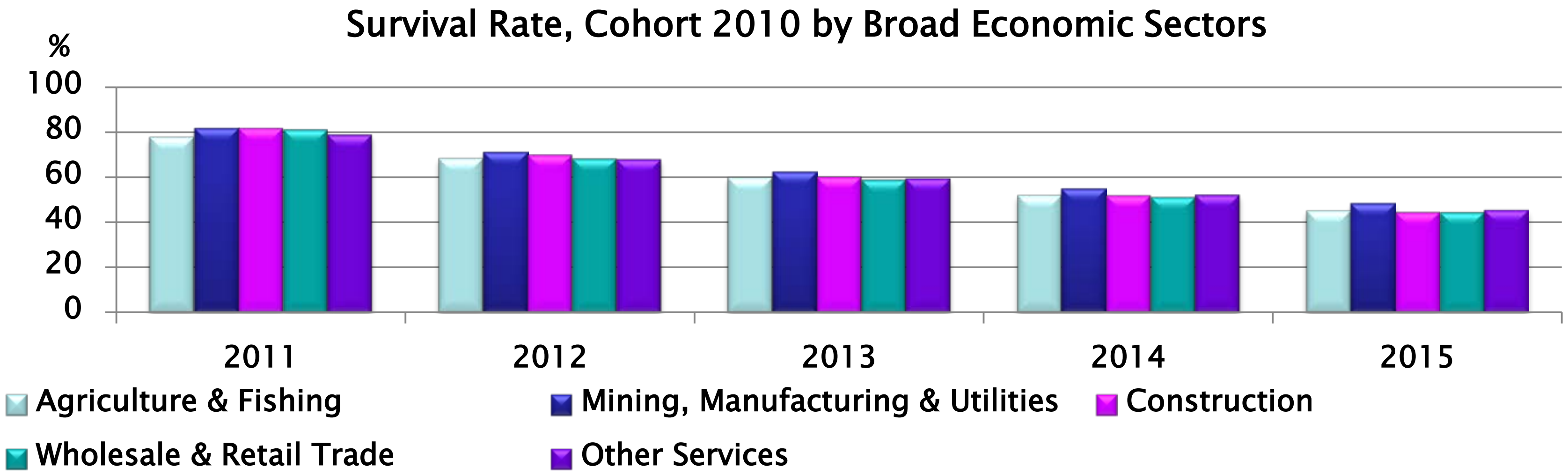
Definitions

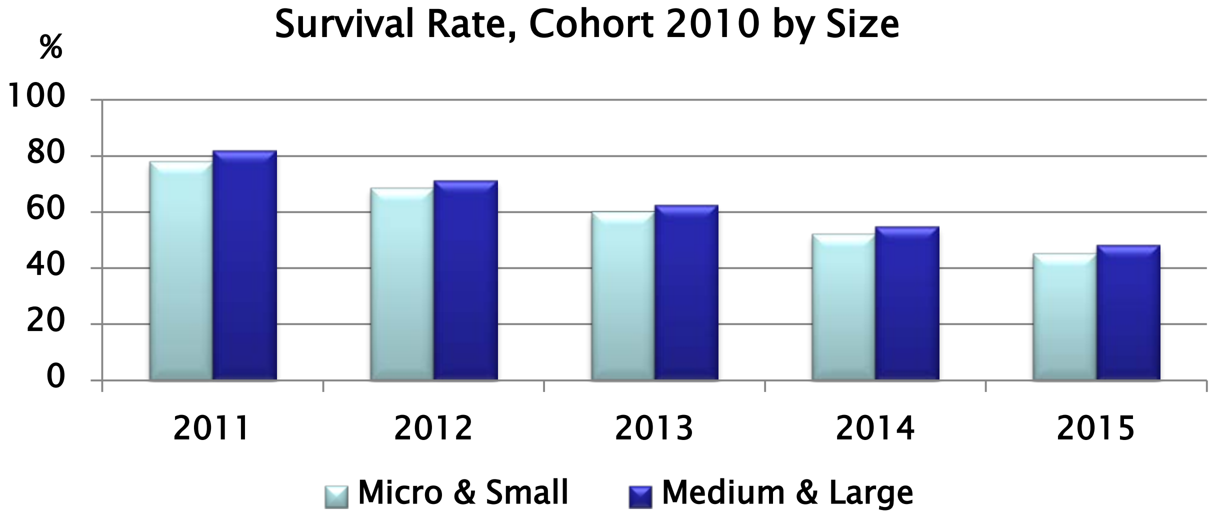
- Birth rate: $\dfrac{\Sigma N^{ti}}{\Sigma T^{t}_{i}}$
  Enterprise births in the sector i, at the year t
  Stock of enterprises in the sector i, at the year t

- Survival rate: $\dfrac{\Sigma S^{t+k}}{\Sigma T^{t}_{i}}$
  Enterprise survivals in the sector i, year t+ k=1,…, 5
  Stock of enterprises in the sector i, at the year t

- Death rate: $\dfrac{\Sigma M^{ti}}{\Sigma T^{t}_{i}}$
  Enterprise deaths in the sector i, at the year t
  Stock of enterprises in the sector i, at the year t

Preliminary Findings: Survival Rates

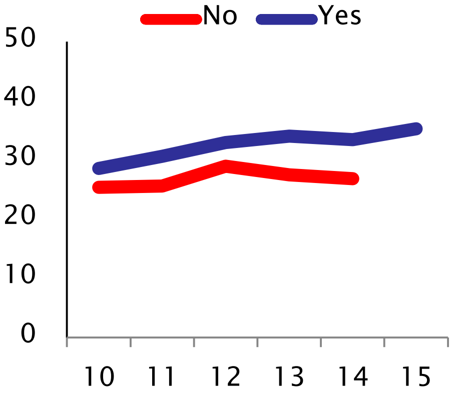


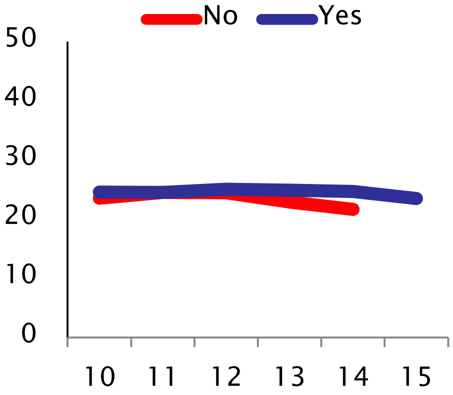


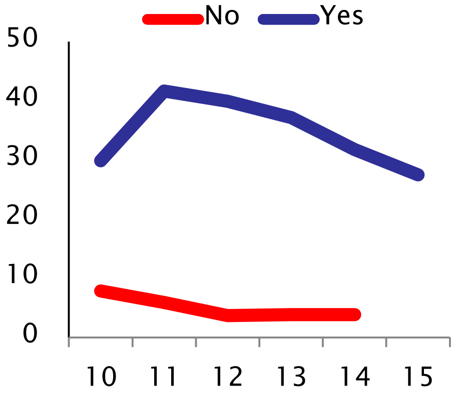
Preliminary Findings: Financial Indicators, cohort 2010 over the 2010 – 2015 period

$$\frac{\text{Total Debt}}{\text{Total Assets}} \qquad \frac{(\text{Revenue} - \text{Cost of goods Sold})}{\text{Revenue}} \qquad \frac{\text{Net Income}}{\text{Shareholder's Equity}}$$

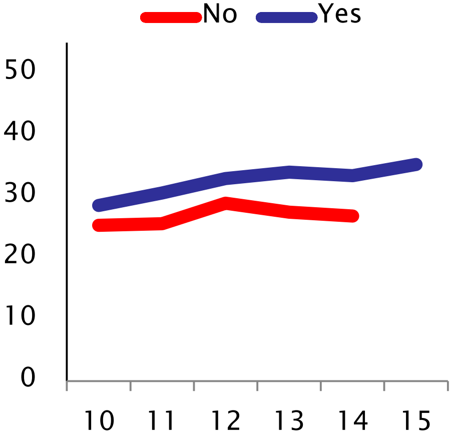
Agriculture & Fishing

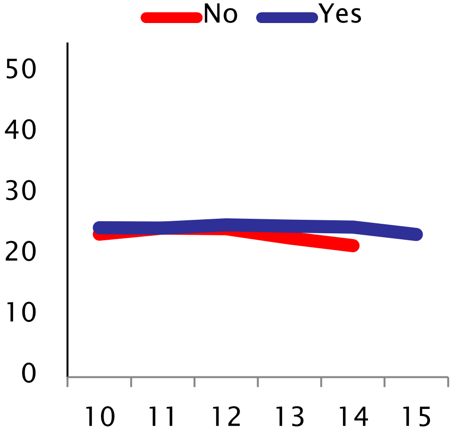


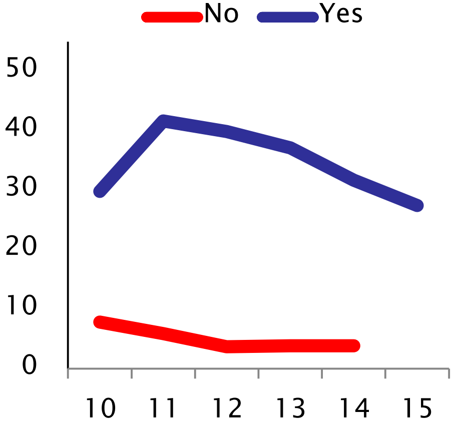




Construction



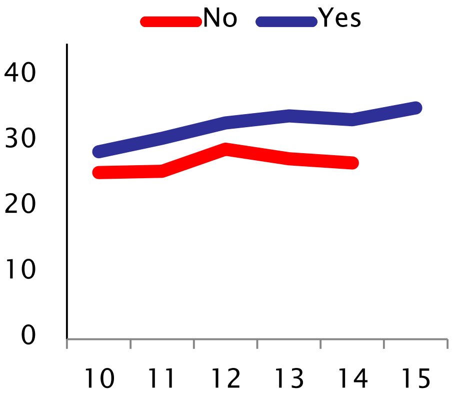


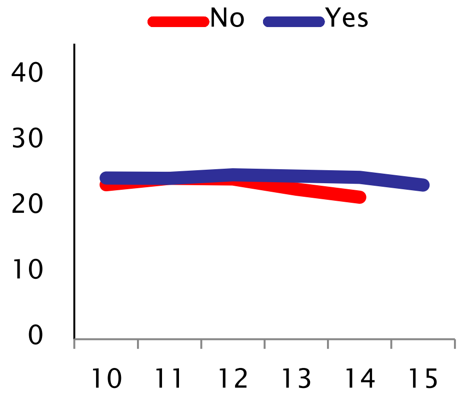


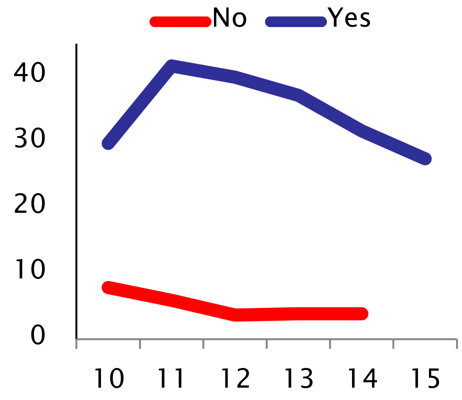
Preliminary Findings: Financial Indicators, cohort 2010 over the 2010 – 2015 period

$$\frac{\text{Total Debt}}{\text{Total Assets}} \qquad \frac{(\text{Revenue} - \text{Cost of goods Sold})}{\text{Revenue}} \qquad \frac{\text{Net Income}}{\text{Shareholder's Equity}}$$

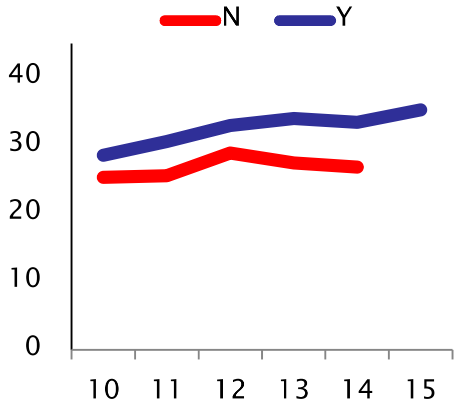
Mining, Manufacturing & Utilities

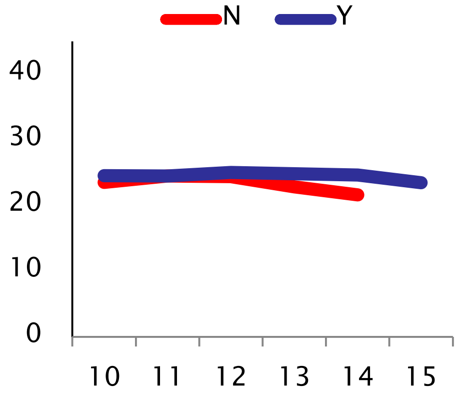


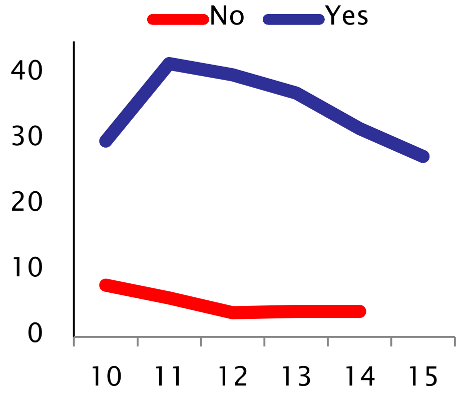




Wholesale & Retail Trade



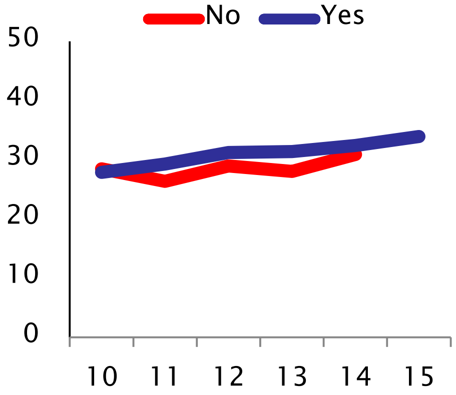


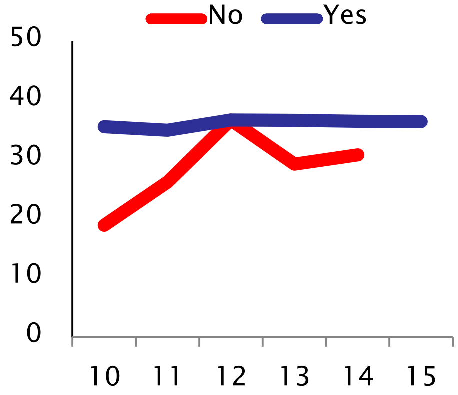


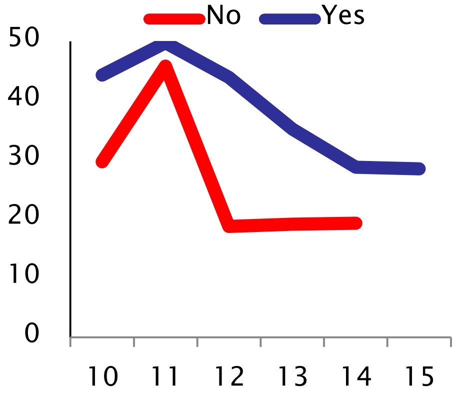
Preliminary Findings: Financial Indicators, cohort 2010 over the 2010 – 2015 period

$$\frac{\textbf{Total Debt}}{\textbf{Total Assets}} \qquad \frac{\textbf{(Revenue − Cost of goods Sold)}}{\textbf{Revenue}} \qquad \frac{\textbf{Net Income}}{\textbf{Shareholder's Equity}}$$

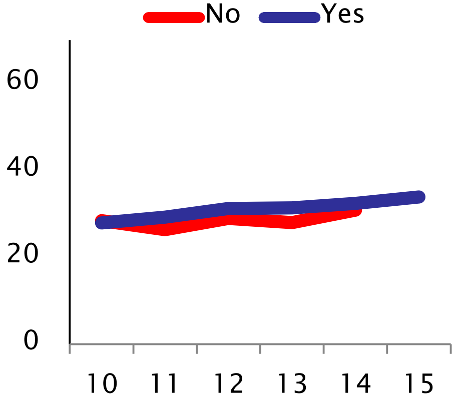
Micro & Small

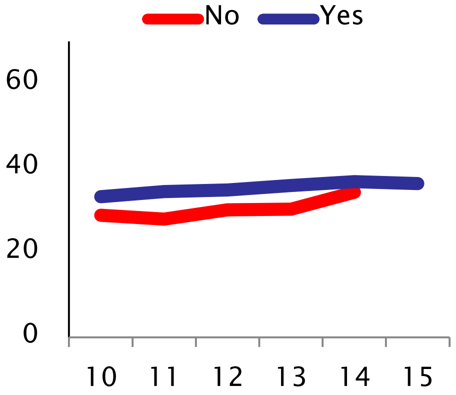


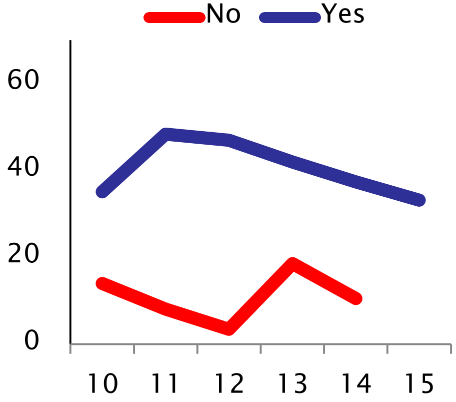




Medium & Large







Preliminary Findings: A Probit model (work in progress…)

- Consider the model below:

$$\Pr(\text{Survival}_{it}|x_{it-1}) = G(\beta_1 + \beta_2 \text{Leverage}_{it-1} + \beta_3 \text{MP}_{it-1} + \beta_4 \text{ROE}_{it-1} \ + \beta_5 \text{IT}_{it-1})$$

- where:
  - $\Pr(\text{Survival}_{it}|x_{it-1})$: Survival probability for the firm “i” in the period “t” conditional on its financial performance at the period “t-1”.
  - Leverage_{it-1}: Debt to assets ratio for the firm “i” in the period “t-1”.
  - MP_{it-1} : Margin price ratio for the firm “i” in the period “t-1”.
  - ROE_{it-1} : Return on equity ratio for the firm “i” in the period “t-1”.
  - IT_{it-1} : Inventory turnover ratio for the firm “i” in the period “t-1”.

Preliminary Findings: A Probit model (work in progress…)

Table 1 Probit Estimates

	2010	2011	2012	2013	2014
Intercept	0.6141***	0.7191***	0.8100***	0.7765***	0.7795***
	(0.0141)	(0.0178)	(0.0204)	(0.0221)	(0.0236)
Leverage	1.1720***	0.7903***	0.6539***	0.7316***	0.7857***
	(0.0399)	(0.0363)	(0.0388)	(0.0418)	(0.0443)
Margin Price	0.1900***	0.4138***	0.3595***	0.3664***	0.2425***
	(0.0325)	(0.0380)	(0.0421)	(0.0446)	(0.0476)
Return on Equity	0.3754***	0.1748***	0.2910***	0.4190***	0.5504***
	(0.0191)	(0.0167)	(0.0222)	(0.0279)	(0.0342)
Inventory Turnover	−0.0000	−0.0002	−0.0006***	−0.0004**	−0.0002
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)

Note: The dependent variable is a dummy equal to one if the firm "i" survives in year t, and zero otherwise. Robust White Standard Errors are presented in the parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

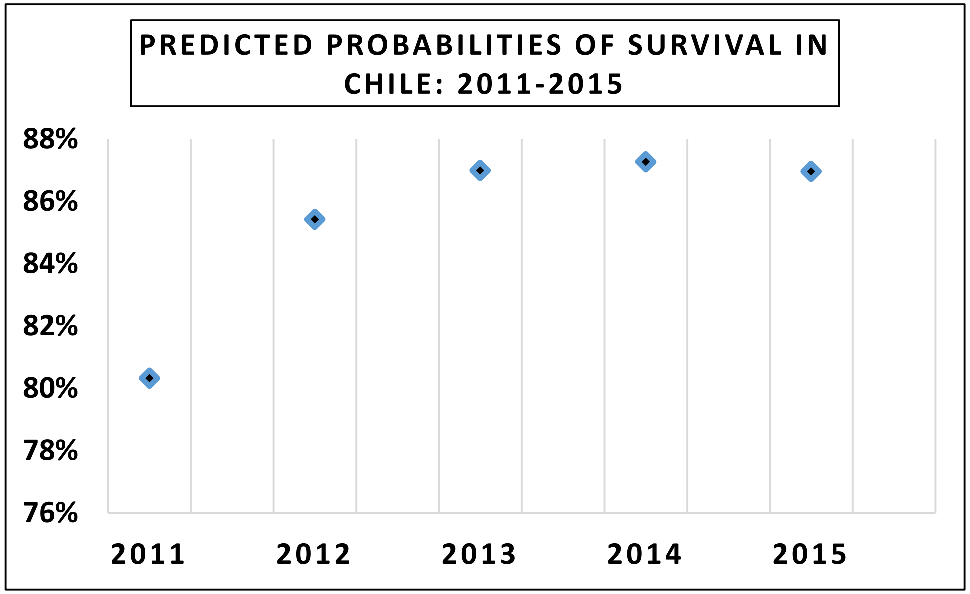
Table 2 Linear Hypothesis Tests

	Wald Chi-square				
	2010	2011	2012	2013	2014
Leverage=0	861.94	472.77	284.55	306.39	315.22
Margin Price=0	341.17	118.87	73.04	67.44	25.96
Return on Equity=0	387.98	108.93	171.73	226.18	259.72
Inventory Turnover=0	0.16	1.53	8.94	2.98	0.99
All four covariates are equal	1717.63	974.76	775.63	886.40	908.45

Preliminary Findings: A Probit model (work in progress…)

Table 3 Clasification table

	2011		2012		2013		2014		2015	
Y/N	N	Y	N	Y	N	Y	N	Y	N	Y
Correct	4	73147	1	62550	0	54435	9	47517	15	41355
Total		**91057**		**73221**		**62567**		**54452**		**47567**
Survival Probabilities	***80.34%***		***85.43%***		***87.00%***		***87.28%***		***86.97%***	



Thank you!

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Using online property advertisements data as a proxy for property market indicators[1]

Kumala Kristiawardani and Irfan Sampe,
Bank Indonesia

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

Using Online Property Advertisements Data as a Proxy for Property Market Indicators

Bank Indonesia:
Kumala Kristiawardani
Irfan Sampe

IFC – Bank Indonesia Satellite Seminar on Big Data
Bali, 21 March 2017

Topics Covered

- Background
- Data Sources
- Methodology
  - Data Acquisition
  - Data Issues
  - Data Preparation
  - Data Processing
- Results
- Conclusions

Background

- A boom and bust in residential property prices is perhaps the most widely discussed topic in recent financial crises
  - ❑ Residential property prices were fell in the 1990s, following the US recession in 1990-1991;
  - ❑ In Japan, residential property prices fell continuously as the economy collapsed in Japan around 1990;
  - ❑ In 2007, the housing market crash was the cause of the financial crisis in US.
- Bank Indonesia has an important task to not only to safeguard monetary stability, but also financial system stability
- Hence, monitoring residential property prices (with other asset prices) is crucial for Bank Indonesia to achieve its main task.

Background

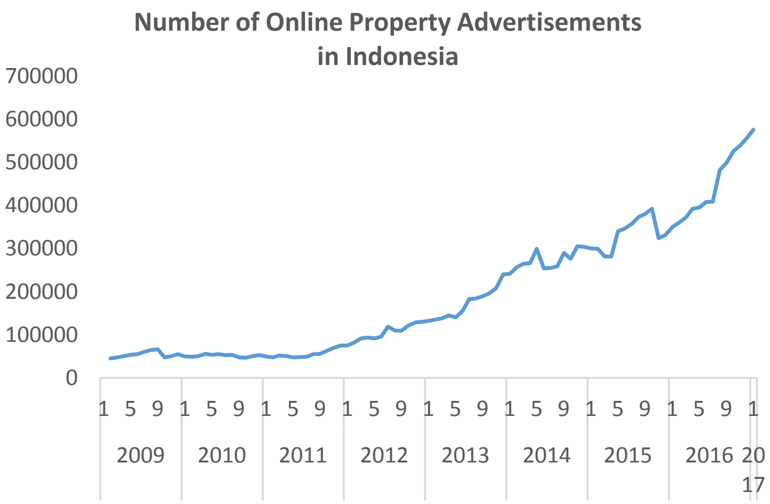
- Currently, the Bank Indonesia’s primary data sources for monitoring Residential Property price are:
  - ❑ Residential Property Price Survey for primary house, conducted quarterly in **16 big cities**.
  - ❑ Residential Property Price Survey for secondary market, conducted quarterly only in **9 big cities**.

  The data published at **six weeks** after the end of the survey period

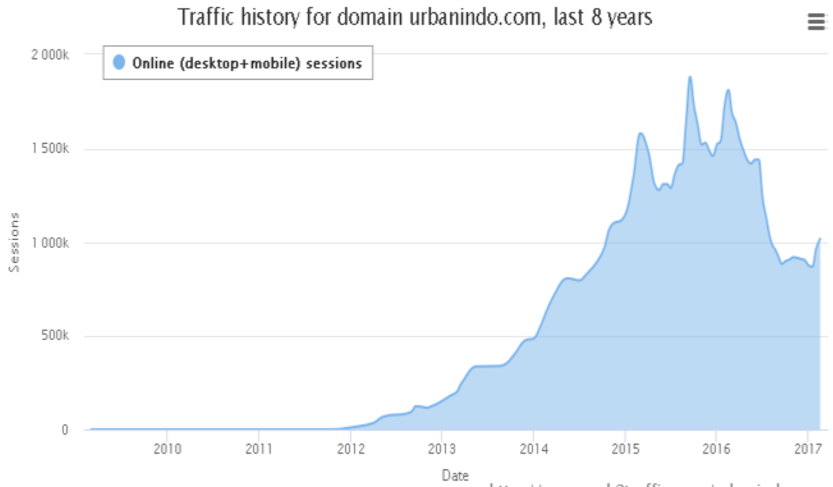
- How do “big data” give the added value for Bank Indonesia in monitoring residential property market?

Background

- The people's behaviour change in finding and selling the house (especially for secondary market)
  - ❑Traditional: property agent, advertisement in newspaper
  - ❑Now: search through internet (google, property online website, mobile apps)



Number of Online Property Advertisements in Indonesia



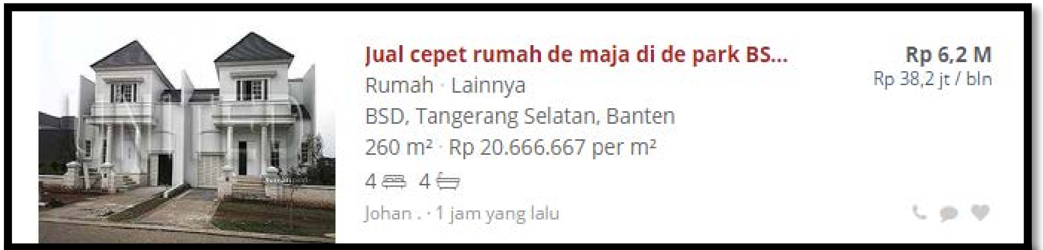
Traffic history for domain urbanindo.com, last 8 years

http://www.rank2traffic.com/urbanindo.com

Data Sources

3 biggest property online website in Indonesia (share 56 %)

- Title
- Status of property : sell/rent
- Type of property (house/apartment/villa/condotel/condominium)
- Advertising time : Starting & end date
- Property price
- Land & building size
- Number of bedroom & bathroom
- Address



Jual cepet rumah de maja di de park BS...

Rumah · Lainnya

BSD, Tangerang Selatan, Banten

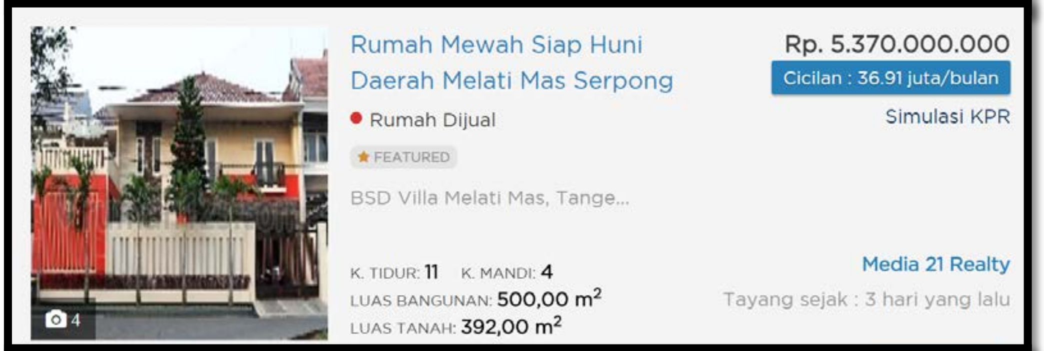
260 m² · Rp 20.666.667 per m²

4 🛏 4 🛁

Johan . · 1 jam yang lalu

Rp 6,2 M

Rp 38,2 jt / bln



Rumah Mewah Siap Huni Daerah Melati Mas Serpong

● Rumah Dijual

★ FEATURED

BSD Villa Melati Mas, Tange...

K. TIDUR: 11 K. MANDI: 4

LUAS BANGUNAN: 500,00 m^2

LUAS TANAH: 392,00 m^2

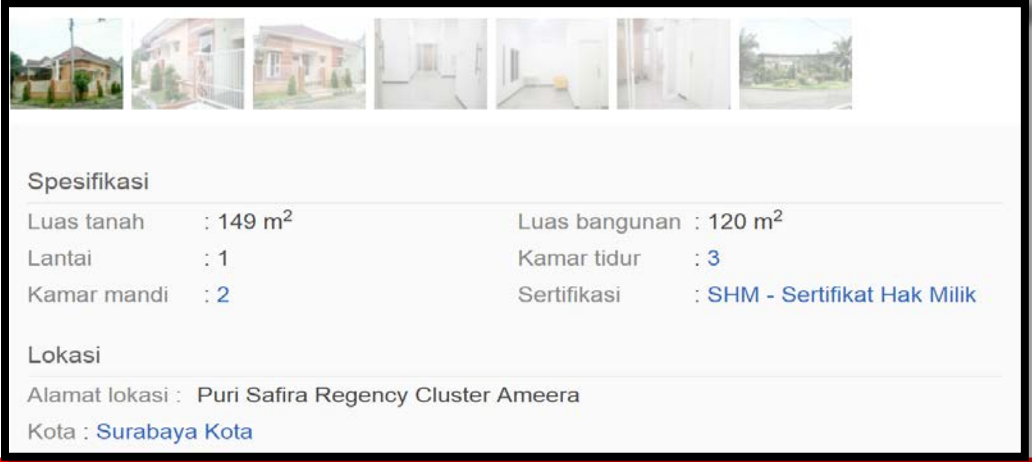
Rp. 5.370.000.000

Cicilan : 36.91 juta/bulan

Simulasi KPR

Media 21 Realty

Tayang sejak : 3 hari yang lalu



Spesifikasi

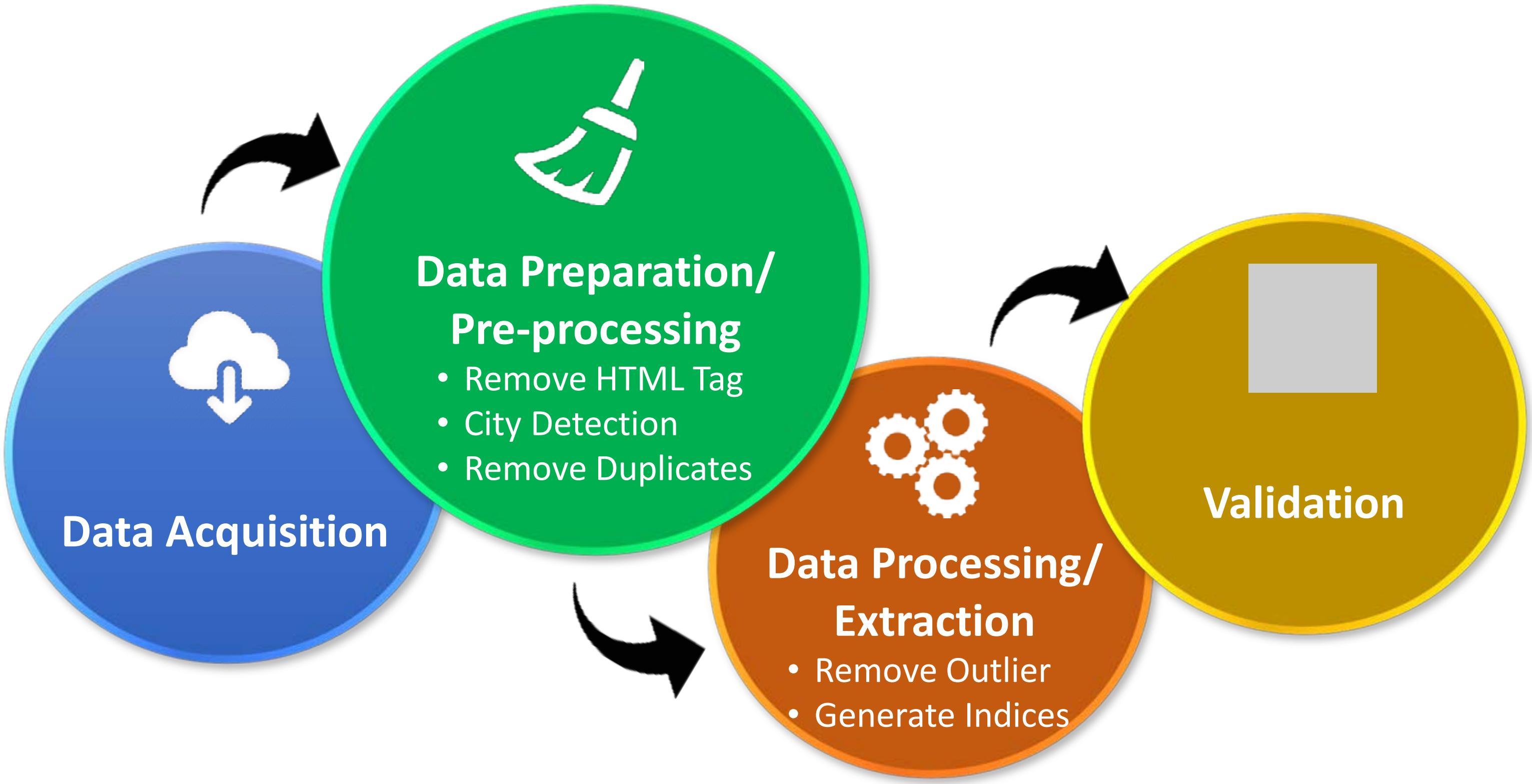
Luas tanah	: 149 m^2	Luas bangunan	: 120 m^2
Lantai	: 1	Kamar tidur	: 3
Kamar mandi	: 2	Sertifikasi	: SHM - Sertifikat Hak Milik

Lokasi

Alamat lokasi : Puri Safira Regency Cluster Ameera

Kota : Surabaya Kota

Methodology



Data Acquisition

Data Preparation/ Pre-processing

- Remove HTML Tag
- City Detection
- Remove Duplicates

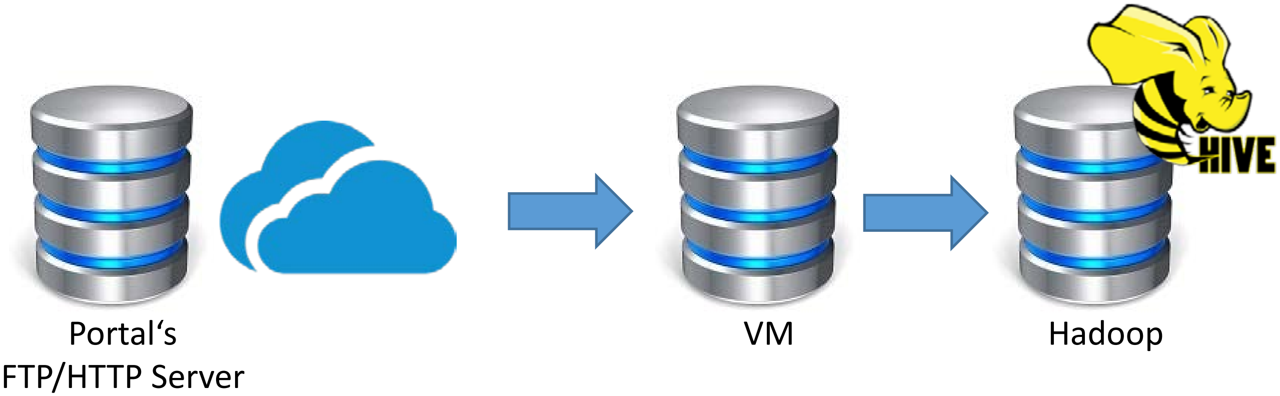
Data Processing/ Extraction

- Remove Outlier
- Generate Indices

Validation

Data Acquisition

- Property portal shared the data using FTPS/HTTPS. The files are password protected
- Available in the 1st week every month
- Loaded into Hadoop
- ≈ 2.2 million ads/month



Data Issues

- Human error in data entry, i.e:
  - ❑ Price = Rp. 0, Price = Rp. 16 trillion ($ 1.2 billion) on small size property
  - ❑ Land Size = 0 sqm, Land Size = 1 sqm
  - ❑ Typo on city/regency name
- Not standardized address data (freetext field)
  - ❑ District/sub district, e.g: Bogor, Bgr
  - ❑ Street name without district name, e.g: Jl. Kesadaran Sukmajaya
- Duplicate ads that are caused by:
  - ❑ One property can be advertised by more than one seller in a single portal
  - ❑ One property can be advertised by one seller across portals
  - ❑ Ads re-post after expiration date

Data Preparation/Pre-Processing

City Detection

- Map district/sub-district into city/regency using BPS's* Master Kabupaten,
- Map address into city/regency using Google Maps Geocoding API

Kampung Rambutan → Jakarta Selatan
Jl. Kesadaran Sukmajaya → Depok

Remove Duplicates

Advertisements are identic if:

- The same attributes values on city/regency, land size, building size, number of bathrooms, and number of bedroom
- Price difference ≤ 5%
- String similarity score for address and ads title ≥ 0.8 (scale of 1) → using Levenshtein Distance

*Indonesian Central Bureau of Statistics (BPS)

Data Processing/Extraction

Remove Outlier

- Removing properties with:
  - ❑ Land size and bulding size is empty (NULL)
  - ❑ Land size < 21 sqm and
    > 10.000 sqm
  - ❑ building size < 21 sqm
    > 10.000 sqm
- Applying price/sqm threshold
- Applying Median Absolute Deviation (MAD)

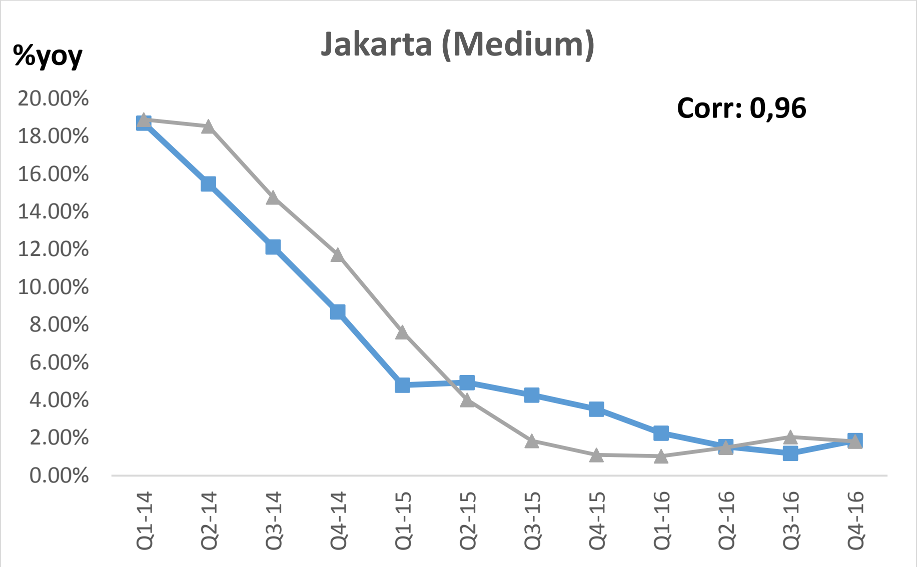
Generate Indices

- Landed house only
- Properties are divided into 3 types (based on building size):
  - ❑ Small: < 80 sqm
  - ❑ Medium: 80 – 150 sqm
  - ❑ Large: > 150 sqm
- Indices are generated per city/regency
  - ❑ Price (AVG: average of property price)
  - ❑ Supply (COUNT: number of active property ads)

Results Obtained

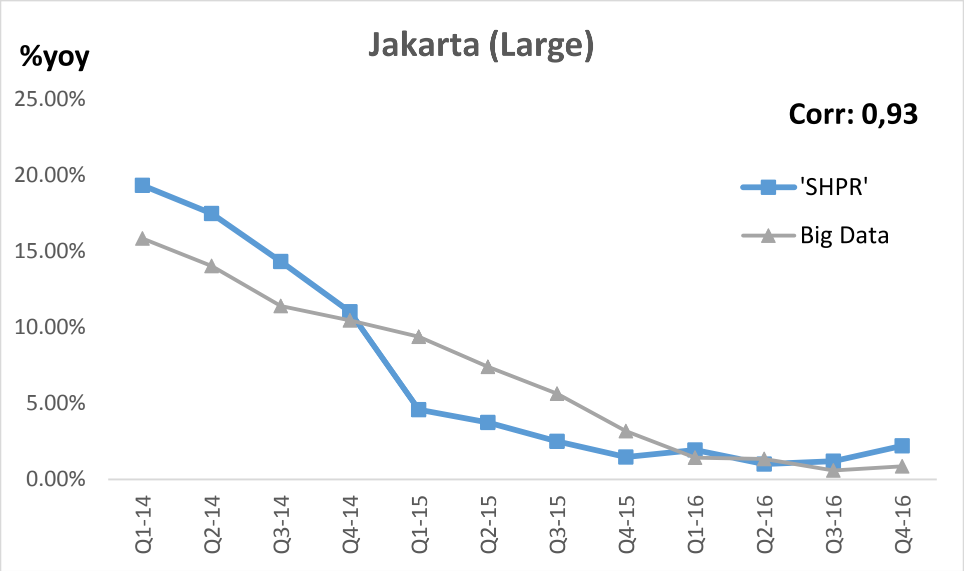
Price Index

Base period: Q2 2015



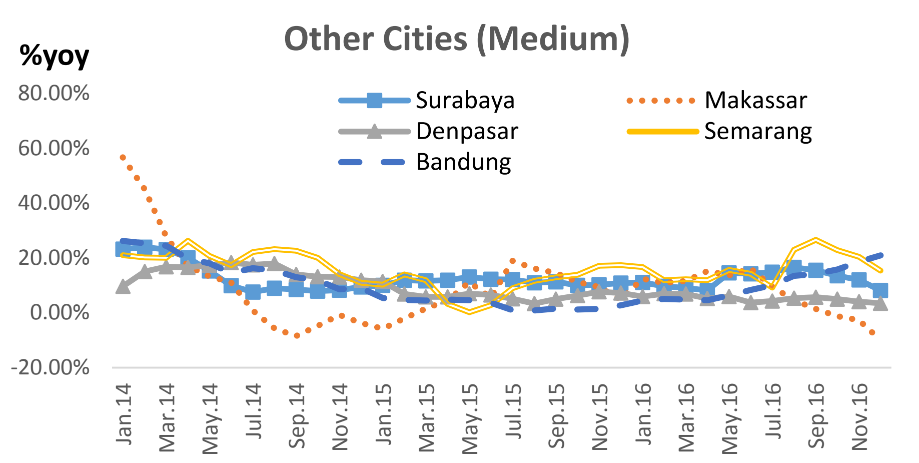
Jakarta (Medium)

Corr: 0,96

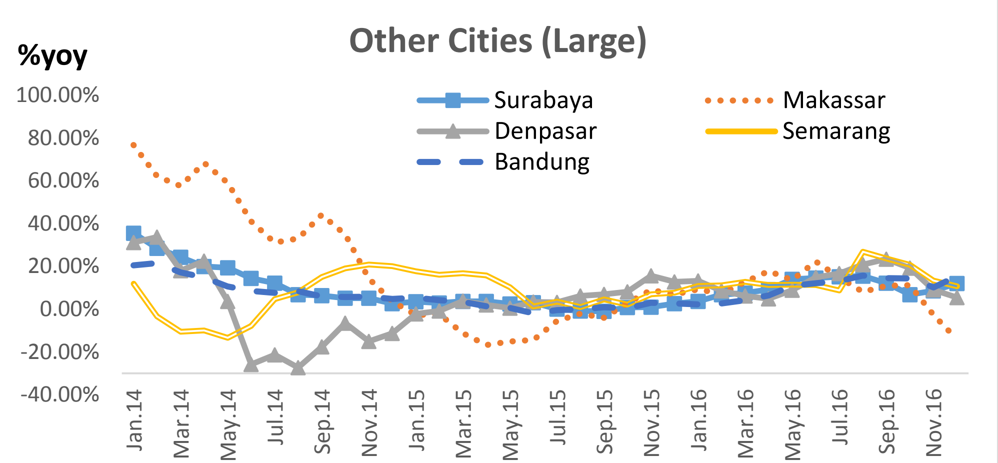


Jakarta (Large)

Corr: 0,93



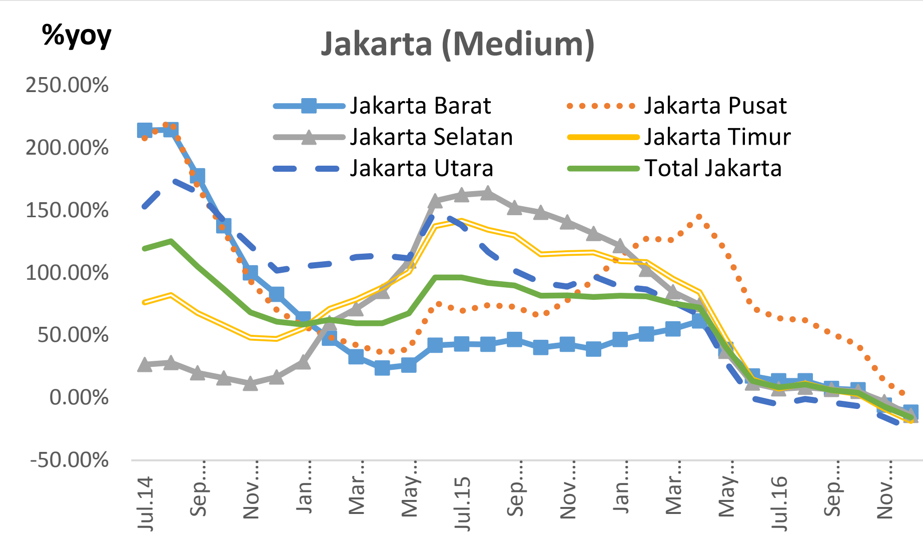
Other Cities (Medium)

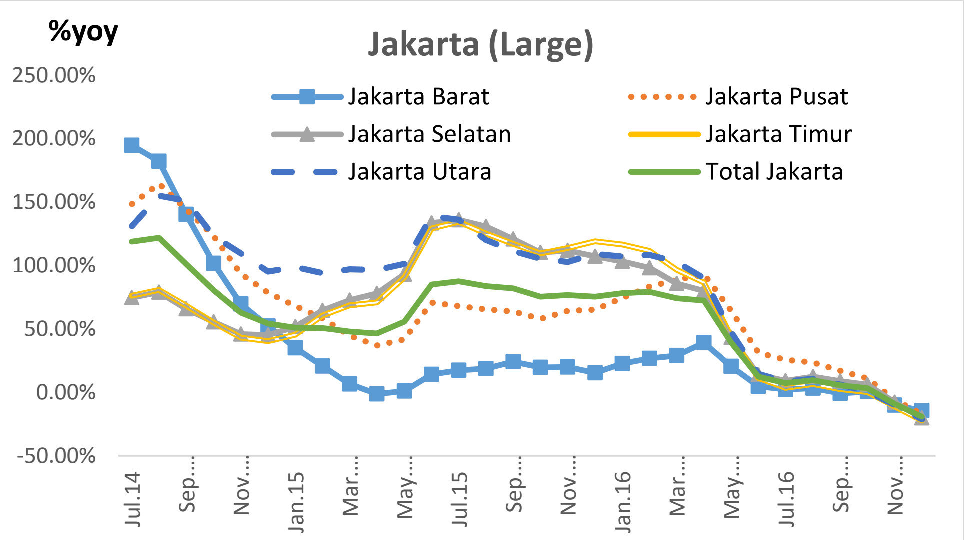


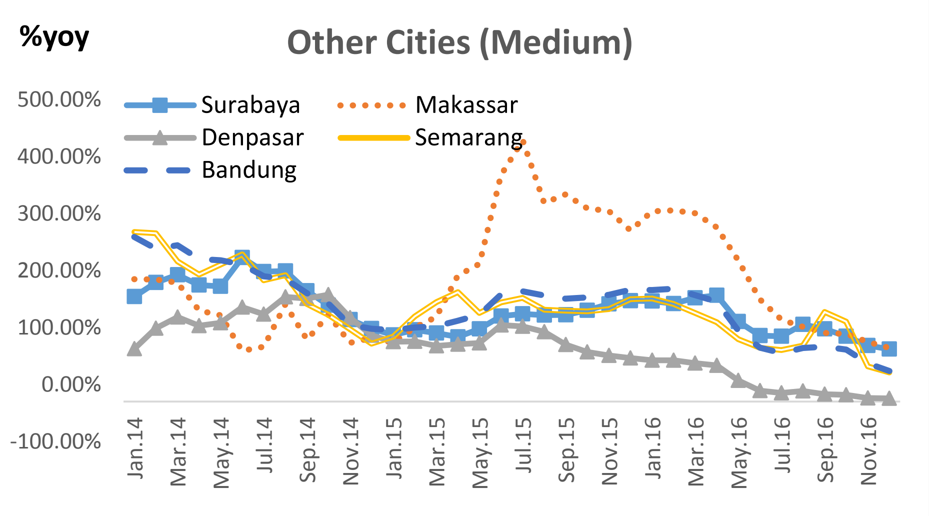
Other Cities (Large)

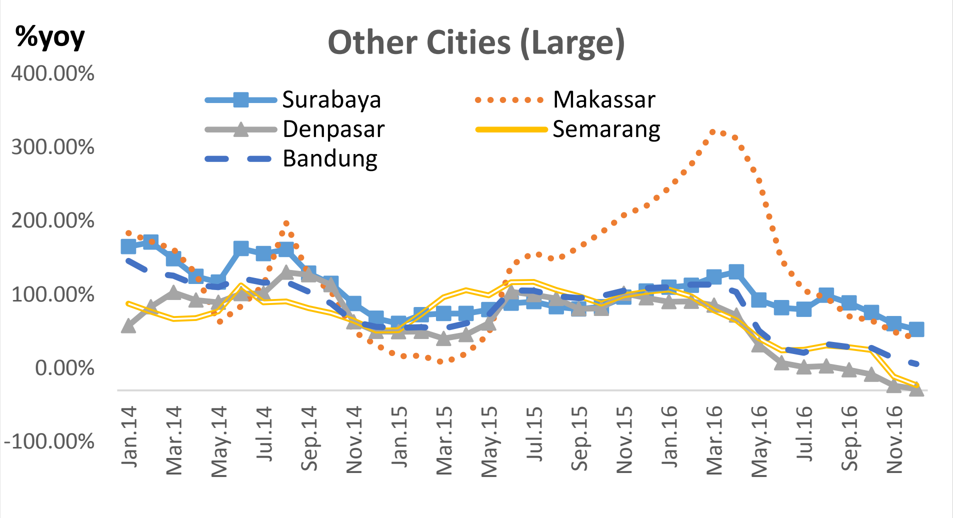
Results Obtained

Supply Index









Conclusions

- Online property ads data are potentially used as a proxy of price and supply indicators in Indonesia’s residential property market.
- However, there are some limitations in conducting the research due to data availability and quality, i.e:
  - ❑ Short periode of data (only available since 2013)
  - ❑ The sold status is rarely updated by the seller

Terima Kasih

(Thank you)

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Integrated management of credit data - Turning threats into opportunities[1]

Luis Teles Dias,
Bank of Portugal

[1] This presentation was prepared for the meeting. 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 meeting.

Integrated management of credit data

Turning threats into opportunities

BANCO DE PORTUGAL
EUROSYSTEM

Luís Teles Dias

Deputy-Director ● Statistics Department

IFC – Bank Indonesia Satellite Seminar on "Big Data"

Bali, Indonesia | 21 March 2017

Central banks “love” credit data

- Central banks (CBs) have a vested interest in credit data
- Usefulness for a number of CBs’ functions:
  - Monetary policy
  - Supervision
  - Financial Stability
  - Statistics
  - Research

Credit Registers

- **Credit registers** (CRs) are the **main source of credit data** for many central banks
- Depending on their scope, number of data attributes, threshold, frequency of reporting or level of granularity, CRs are quite often one of the **largest databases managed by central banks**
- Although not complying with the five Vs commonly used to describe Big Data, CRs are frequently referred by central banks as *Big Data projects*

Credit Registers and the five Vs

Volume	High, at least for CBs' standards
Velocity	Low, although daily reporting can introduce extra challenges
Variety	None; data are typically very structured and uniformly formatted
Veracity	Data refers to individuals (natural and legal persons) who are normally entitled to access their own credit information, thus the trustworthiness of the data is crucial
Value	High, either for the CB and the credit institutions (when the CR is also a service provided by the CB)

The AnaCredit project

- The financial crisis increased the need to **explore indebtedness** across the European Union
- In order to address this desideratum, the ECB (together with the euro area and some non-euro area NCBs) launched the **AnaCredit** project in 2011
- **Harmonized and highly granular credit data** was considered crucial to support several CB functions, such as decision-making in monetary policy and macroprudential supervision

Challenging EU central banks

A big challenge ahead for EU central banks !

Not only because of the volumes of data but also due to very different backgrounds

The "long and winding road" to comply with AnaCredit

- The impact of the AnaCredit project was very diverse among the euro area national central banks (NCBs)

STARTING POINT		IMPACT
NCBs running a CR	Loan-by-loan CR	Low-Medium
	Debtor-by-debtor CR	High
NCBs not running a CR		High

BdP case

- Furthermore, the wide range of data attributes originates a high degree of complexity

Strategic approach to the AnaCredit challenge

- Managing a CR is a long-time legal responsibility of *Banco de Portugal* (since 1978)
- Triggered by AnaCredit, it was **inevitable to develop a new CR** (and not simply revamp the existing one)
- Before embarking in the development of the new CR we have decided to **look at the challenge strategically**:

  How could we profit from this threat to became more efficient and, at the same time, to convey this efficiency to the reporting obligations of credit institutions?

The new CR – an opportunity for BdP

The new CR is a golden opportunity to rationalize all the reporting of credit data to the central bank, irrespectively of the purpose behind the different existing reports

The reporting to the new CR will integrate all credit data needs of the different departments of the Bank and will operate as a "single-entry point" for all those data

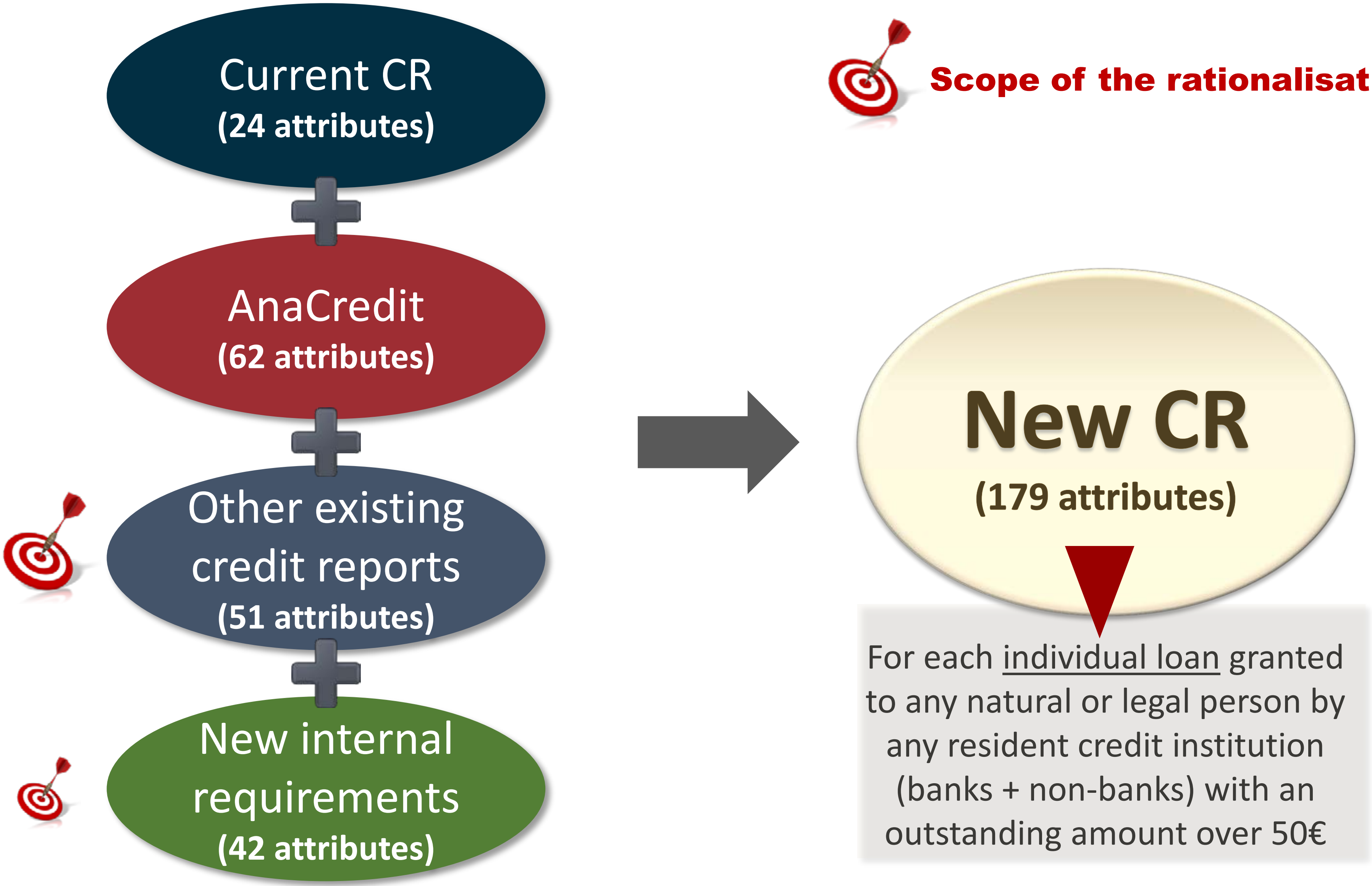
Expected benefits of the chosen approach

- Rationalization of reports
- Harmonization of concepts, granularity, frequency and timeliness related to credit data
- Minimization of efforts required to ensure the coherence of credit data used for different purposes
- Efficiency in the management of quantitative data
- Existence of a single moment and a single entry-point for communicating credit data
- Contribution for reducing context costs of the financial system regarding their reporting obligations

Full support of this approach by credit institutions

The new Credit Register in a nutshell

Scope of the rationalisation



Current CR
(24 attributes)

AnaCredit
(62 attributes)

Other existing credit reports
(51 attributes)

New internal requirements
(42 attributes)

New CR
(179 attributes)

For each individual loan granted to any natural or legal person by any resident credit institution (banks + non-banks) with an outstanding amount over 50€

The new Credit Register in a nutshell

Monthly **(all attributes):**

- Credit data
- Credit risk data

Daily **(small subset of attributes):**

- New loans above 15K €
- Full anticipated repayments
- New overdue loans
- Repayment of overdue loans

Managing information in a strategic way

- The new CR will be an important cornerstone of the ***Integrated Management of Information*** – a strategic approach adopted by BdP two years ago
- Information, a major asset of a CB, should be managed in an integrated way – like other resources such as HR, IT or Budget
- The new CR will contribute to reinforce the role of the Statistics Department as the operational manager (not “owner”) of quantitative information of the Bank

The new Credit Register project

Major milestones

SEPTEMBER
2015

Start of the Investigation Phase

OCTOBER
2016

Start of the Realisation Phase

… TODAY …

JUNE
2018

Go-live!

Thank you for your attention!

ldias@bportugal.pt

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Finding similar words in Big Data - Text mining approach of semantic similar words in the Federal Reserve Board members' speeches[1]

Christian Dembiermont and Byeungchun Kwon,
Bank for International Settlements

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

BANK FOR INTERNATIONAL SETTLEMENTS

Finding similar words in Big Data

Text mining approach of semantic similar words in the Federal Reserve Board members' speeches

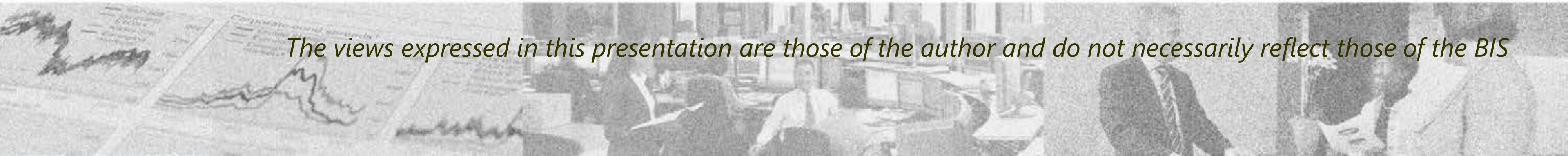
Christian Dembiermont and Byeungchun Kwon

Data Bank Services, Monetary and Economic Department,

Bank for International Settlements

Irving Fisher Committee - Bank Indonesia Satellite Seminar on "Big Data"

Bali, 21 March 2017



The views expressed in this presentation are those of the author and do not necessarily reflect those of the BIS

Overview

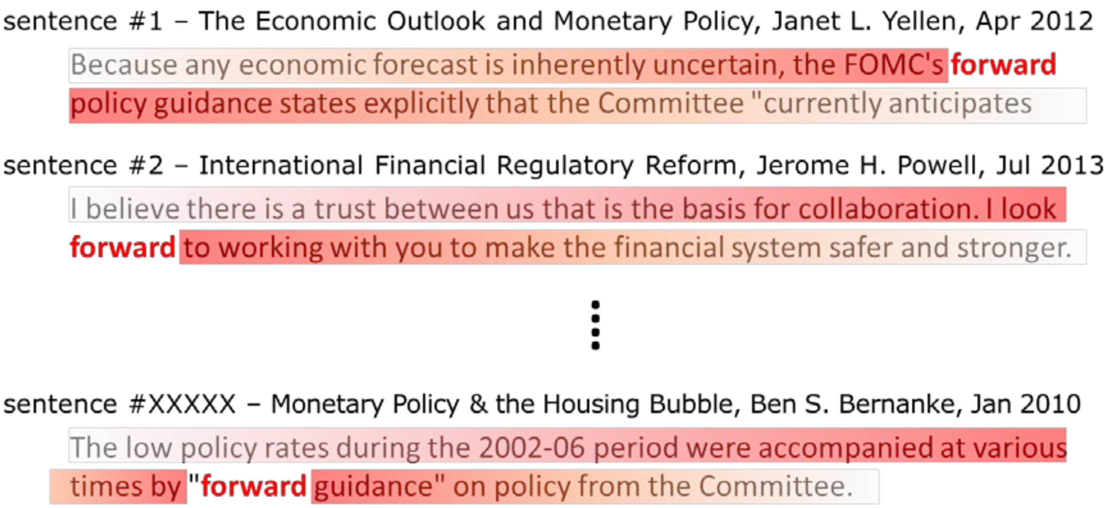
- Finding words in a corpus of thousands of documents is a difficult task
- Finding similar words in this corpus is a daunting task

- Business case: finding similar words to "forward"
- Solution: new text mining technology "Word2Vec"

Detection of words with similar meaning

Central hypothesis:

Linguistic items with similar distributions have similar meaning



sentence #1 – The Economic Outlook and Monetary Policy, Janet L. Yellen, Apr 2012

Because any economic forecast is inherently uncertain, the FOMC's **forward** policy guidance states explicitly that the Committee "currently anticipates

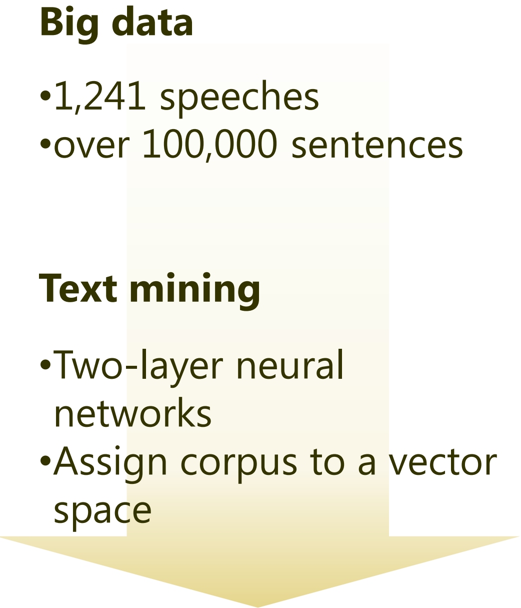
sentence #2 – International Financial Regulatory Reform, Jerome H. Powell, Jul 2013

I believe there is a trust between us that is the basis for collaboration. I look **forward** to working with you to make the financial system safer and stronger.

⋮

sentence #XXXXX – Monetary Policy & the Housing Bubble, Ben S. Bernanke, Jan 2010

The low policy rates during the 2002-06 period were accompanied at various times by "**forward** guidance" on policy from the Committee.



Big data

- 1,241 speeches
- over 100,000 sentences

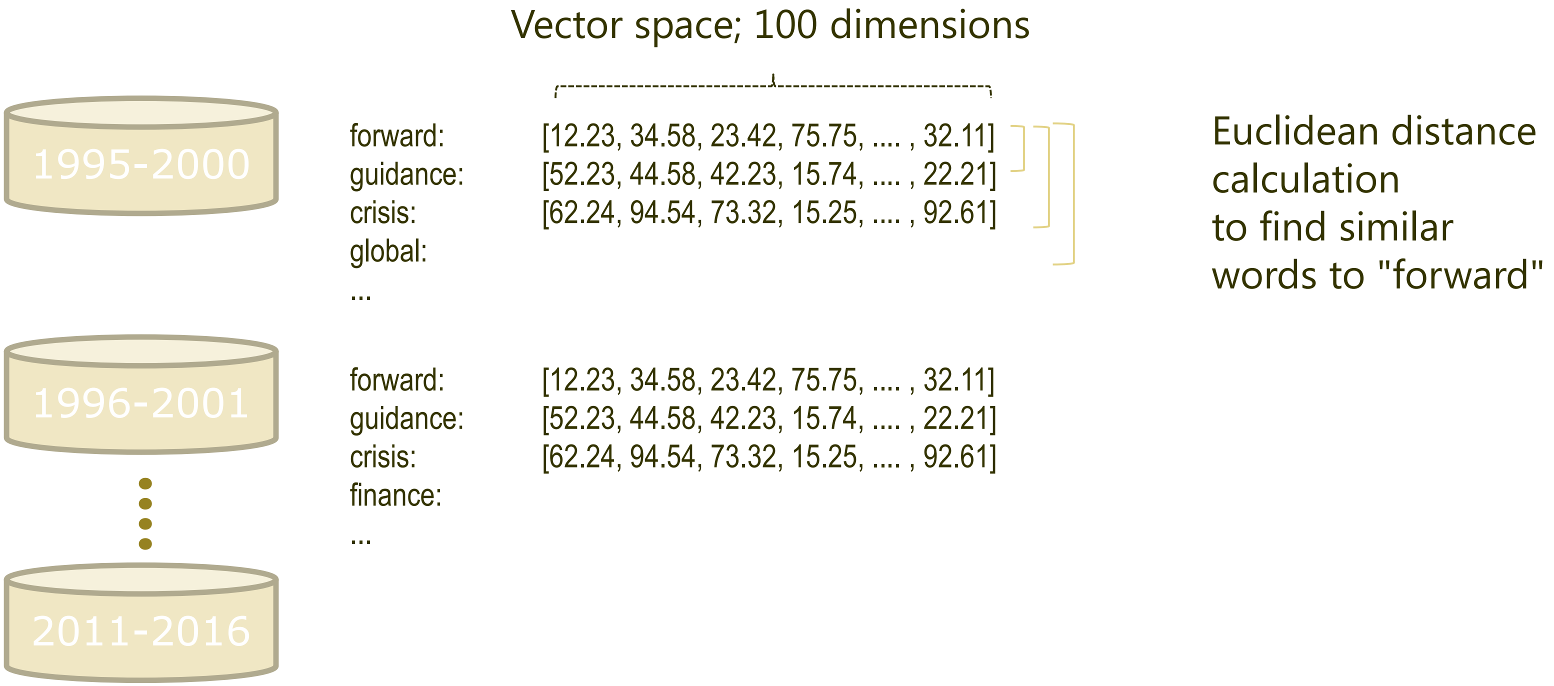
Text mining

- Two-layer neural networks
- Assign corpus to a vector space

Semantic Similarity Database

Semantic Similarity Database

Calculation of the Euclidean distance between two Word vectors



Vector space; 100 dimensions

1995-2000

forward: [12.23, 34.58, 23.42, 75.75, ... , 32.11]
guidance: [52.23, 44.58, 42.23, 15.74, ... , 22.21]
crisis: [62.24, 94.54, 73.32, 15.25, ... , 92.61]
global:
...

Euclidean distance calculation
to find similar
words to "forward"

1996-2001

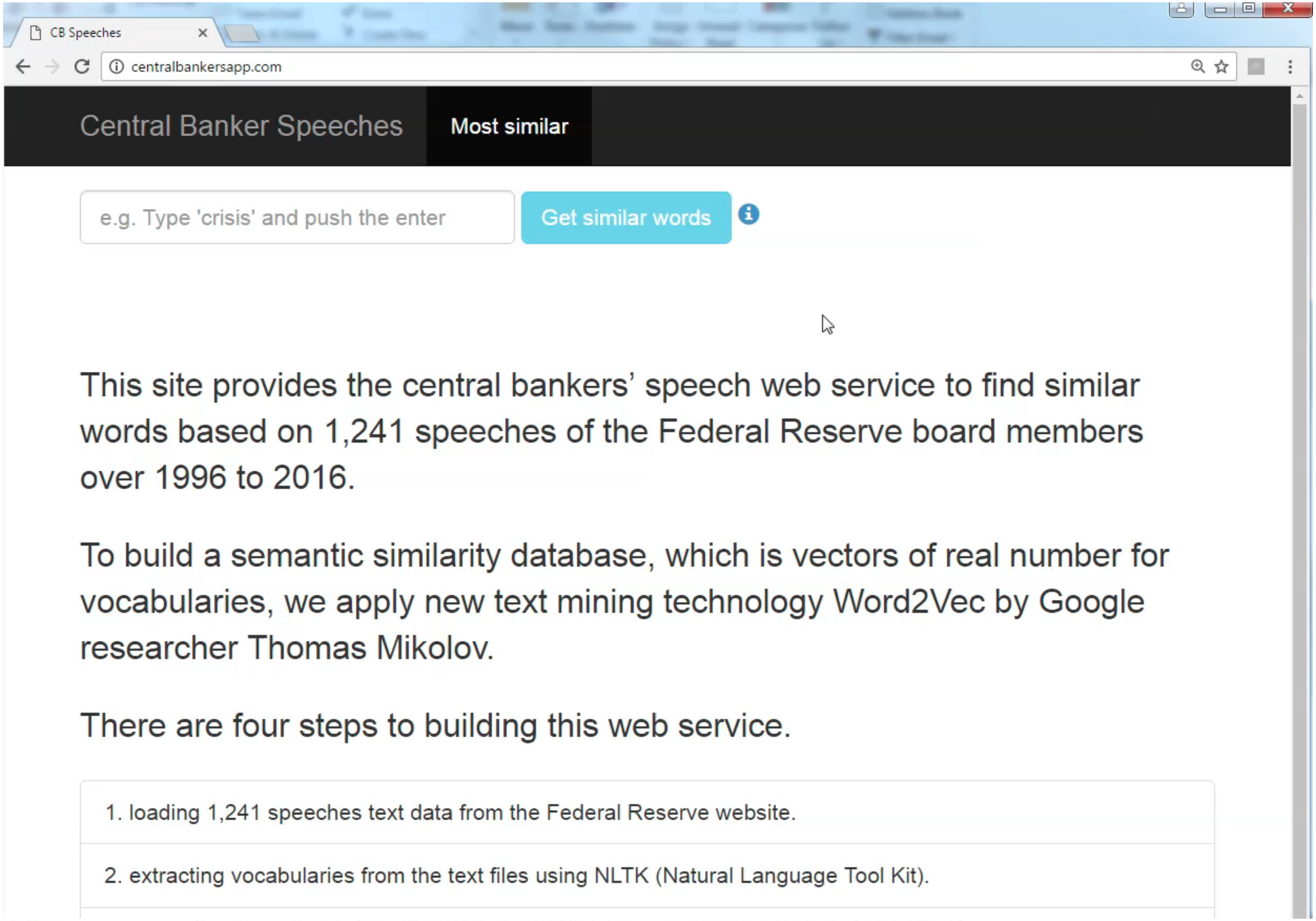
forward: [12.23, 34.58, 23.42, 75.75, ... , 32.11]
guidance: [52.23, 44.58, 42.23, 15.74, ... , 22.21]
crisis: [62.24, 94.54, 73.32, 15.25, ... , 92.61]
finance:
...

2011-2016

Behind the Semantic Similarity Database: Word2Vec

- Word2vec
  - created by a team of researchers led by Tomas Mikolov (Google)
  - input: a large corpus of text
  - output:
    - a vector space, typically of several hundred dimensions
    - each unique word in the corpus being assigned a corresponding vector in the space
    - word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. ICLR.

Live demo *(available on http://centralbankersapp.com/)*



This site provides the central bankers' speech web service to find similar words based on 1,241 speeches of the Federal Reserve board members over 1996 to 2016.

To build a semantic similarity database, which is vectors of real number for vocabularies, we apply new text mining technology Word2Vec by Google researcher Thomas Mikolov.

There are four steps to building this web service.

1. loading 1,241 speeches text data from the Federal Reserve website.
2. extracting vocabularies from the text files using NLTK (Natural Language Tool Kit).

Results of the findings

Similar words to "forward" are:

Federal Reserve Board members’ speeches: 1995-2007

forecasts, incoming, ahead, carefully

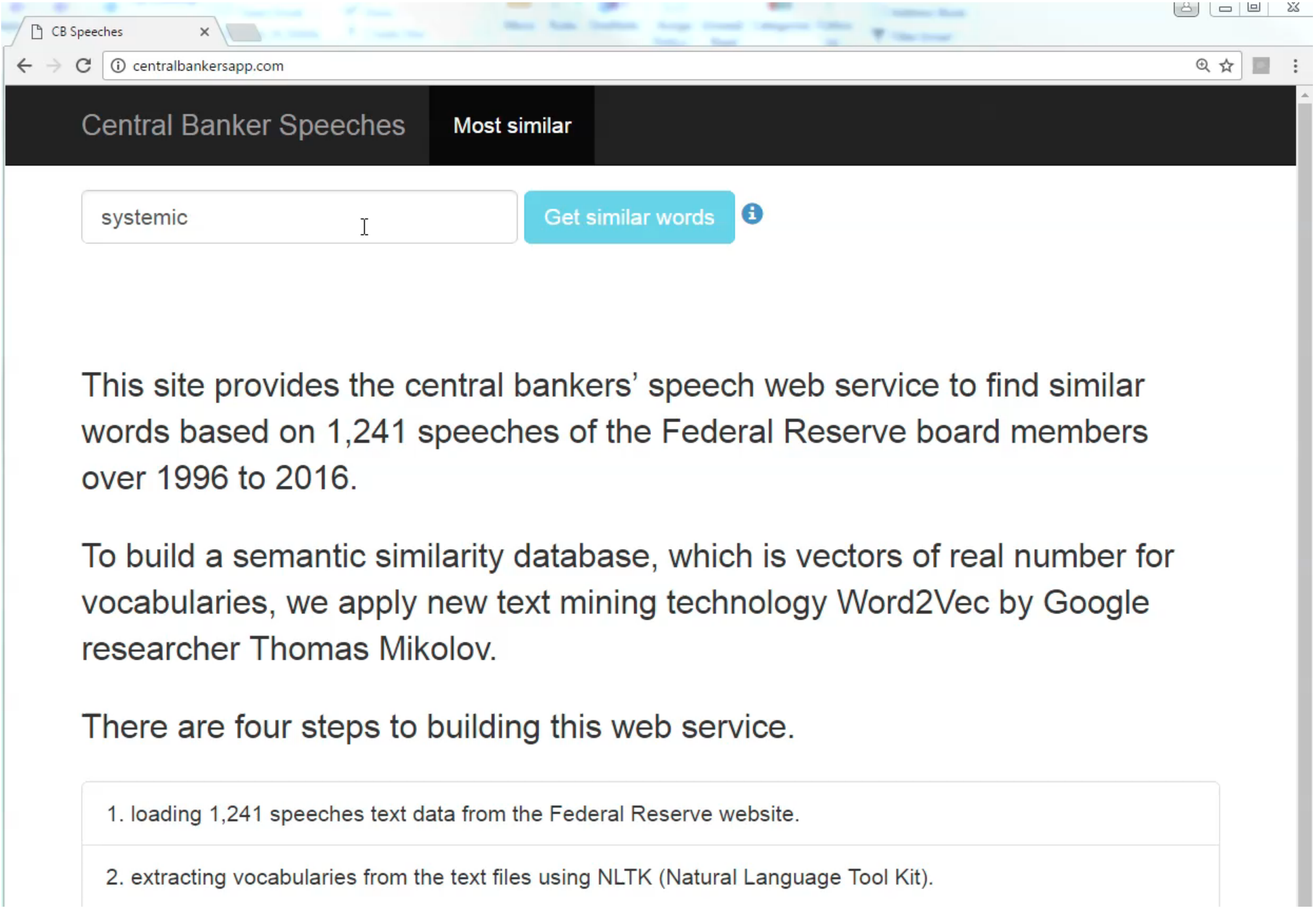
Federal Reserve Board members’ speeches: 2008-2016

guidance, communicate, intention, path

p.m.: Standard dictionary:

ahead, leading, onward, forth

Live demo *(available on http://centralbankersapp.com/)*



This site provides the central bankers' speech web service to find similar words based on 1,241 speeches of the Federal Reserve board members over 1996 to 2016.

To build a semantic similarity database, which is vectors of real number for vocabularies, we apply new text mining technology Word2Vec by Google researcher Thomas Mikolov.

There are four steps to building this web service.

1. loading 1,241 speeches text data from the Federal Reserve website.
2. extracting vocabularies from the text files using NLTK (Natural Language Tool Kit).

Results of the findings

Similar words to "systemic" are:

Federal Reserve Board members’ speeches: 1995-2007

hazard, moral, soundness, operations, sensitivity, taking

Federal Reserve Board members’ speeches: 2008-2016

macroprudential, interconnectedness, failure, structure

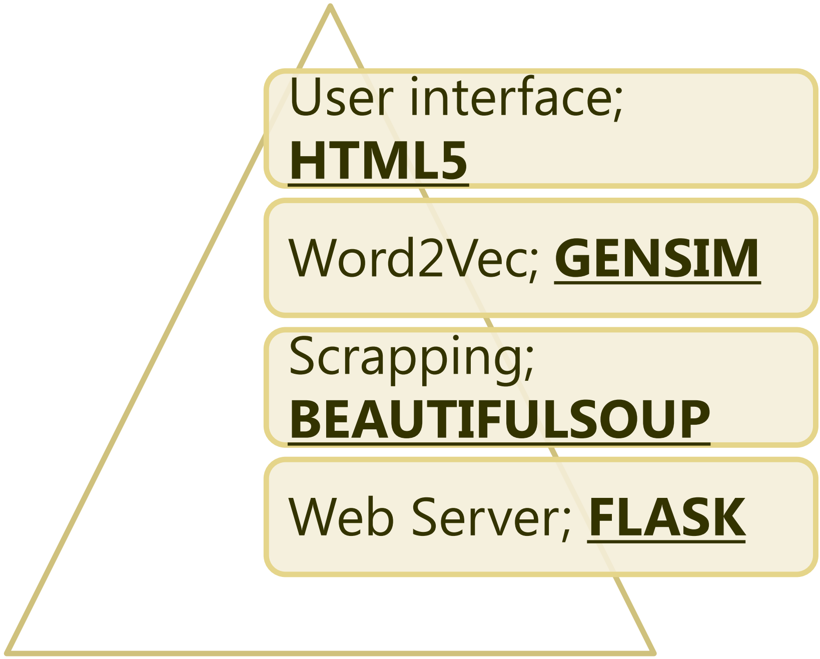
Conclusion

Characteristics of the Word2Vec text mining technique

- Applied for the first time to a central bank domain
- Applied for the first time to analyze similarity between words
- Improved the text mining beyond the word cloud (used in CBs)
- Able to trace similarity evolution over time
- Does not provide any economic forecasts or causality analysis

Source code

- All codes are written in Python language and are available at http://github.com/Byeungchun/centralbankersword2vec



User interface; **HTML5**

Word2Vec; **GENSIM**

Scrapping; **BEAUTIFULSOUP**

Web Server; **FLASK**

Thank you!

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Between hawks and doves: measuring Central Bank Communication[1]

Stefano Nardelli, European Central Bank,
David Martens and Ellen Tobback, University of Antwerp

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

Between Hawks and Doves: Measuring Central Bank Communication

Stefano Nardelli[1], David Martens[2] and Ellen Tobback[3]

Abstract

Media scrutinise the ECB's communication very attentively to extract information on likely future moves of monetary policy rates, in particular after each press conference following monetary policy meetings. Assessing media's perception requires the translation of words into a quantitative indicator. In this paper, we propose the Hawkish-Dovish (HD) indicator which is computed out of a bulk of above 9,000 media reports on the ECB press conference (i.e. newspaper articles, newswires, etc.) and translates on a numerical scale the degree of “hawkishness” or “dovishness” of the ECB press conference tone perceived by media. We compare two different methods to calculate the indicator: one is based on the semantic orientation, while the other is constructed on a text classification with a Support Vector Machines classification model. We show that the latter method tends to provide more stable and accurate measurements of the perception on a labelled test set. Furthermore, we demonstrate the potential use of this indicator with some applications. We analyse the correlations with a set of interest rates and demonstrate the indicator's ability to anticipate monetary policy moves. Additionally, we use the Latent Dirichlet Allocation (LDA) algorithm to detect the dominant topics in the articles on ECB monetary policy decisions and conclude that the media's focus has shifted from the classic interest rate movements towards the non-standard monetary policy instruments. These findings provide decisive evidence in favour of using an advanced text mining classification model to measure more accurately how media perceive ECB communication.

Keywords: Central bank policies, monetary policy, communication, big data, data mining, quantitative methods.

JEL classification: C02, C63, E52, E58

[1] This paper was prepared for the IFC-Bank Indonesia Satellite Seminar on “Big Data” at the ISI Regional Statistics Conference 2017 in Bali on 21 March 2017. The authors are grateful to participants to the conference and to all preliminary seminars for their useful comments and fruitful discussions on the paper. The views expressed here do not represent those of the European Central Bank.

Correspondence: Stefano Nardelli; European Central Bank; E-mail: stefano.nardelli@ecb.europa.eu

[2] University of Antwerp, Belgium; Email: david.martens@uantwerpen.be

[3] University of Antwerp, Belgium; Email: ellen.tobback@uantwerpen.be

Contents

1. Introduction

During the financial crisis, monetary policy interest rates reached levels close to zero in most major central banks. As a result, communication has increasingly qualified as a powerful instrument able to influence financial market developments and drive expectations. Forward guidance, i.e. a form of verbal commitment to keep interest rates at a certain level for a period of time, possibly conditional to certain economic developments, became a genuine monetary policy instrument for many major central banks in addition to the standard toolkit of interest rates and refinancing instruments for the banking sector. Mirroring the increased role of communication in central banks, economic research has started to focus on how central bank communication adds information to that which is already contained in macroeconomic variables and how it may reveal policy makers individual preference functions. The overall objective is to enhance predictability of monetary policy decisions. The traditional approach focuses mainly on how information events affect both financial market developments and expectations of future policy moves.

Media usually analyse central banks' communication with a view to extract information about likely future moves of monetary policy rates (Hayo and Neuenkirch, 2015). Unlike market indicators, extracting relevant information on monetary policy from media reports is not straightforward and requires the translation of words into a numerical representation. Ideally, the resulting quantitative indicator should represent the media's perception of the central banks message and should be able to indicate whether expectations are more on the tightening side (hawkish perception) or rather on the loosening side (dovish perception).

In recent years, a number of studies have developed quantitative tools to measure or, at least, to express on a numerical scale the information contained in official central banks' statements (e.g. KOF Swiss Economic Institute, 2007; Lucca and Trebbi, 2009; Hansen and McMahon, 2016; and Nechio and Regan, 2016). In other words, this approach can be characterised as an attempt to quantify information in communication, which is qualitative by nature. In this paper, we present an index measuring how the official monetary policy communications of the European Central Bank (ECB) are interpreted by the press, using data mining techniques. Concretely, we want to distinguish articles that perceive the ECB's statement as predominantly hawkish, i.e. pointing towards likely future policy rate increases, from those perceiving statements as dovish, i.e. hinting at declines or no increases in policy rates. We use two different techniques normally used in text analysis to measure the tone expressed in an article published after the press conference: Semantic Orientation (SO) and text classification using Support Vector Machines (SVM). The former measures how often the ECB is mentioned in a news article together with respectively hawkish and dovish words, the latter uses a classification model to predict the tone of a news article. We apply both methods to a data set of approximately 9,000 articles from the Dow Jones's Factiva global news database (which includes several million of items from nearly 33,000 sources), published between January 1999 and March 2016, in order to create the HD index (after the initials of Hawkish and Dovish). This indicator represents a sort of average tone between two hawkish and dovish extremes of the ECB communication as perceived by the media after each press conference.

This paper aims at offering on original contribution on various dimensions. First, it proposes a quantitative indicator computed from text sources, showing that data or text mining techniques can bring value in the process of monetary policy decision making (Section 2 and 3). Secondly, it compares two alternative methodologies and concludes in favour of the use of SVM (Section 4). Finally, it shows that one of the advantages of the SVM classification model there is its flexibility and the possibility to analyse in details terms most frequently used by the written media to describe monetary policy decisions (Section 4.3).

2. The HD index

The HD index was computed using two different methodologies: the first based on semantic orientation and the other on a support vector machine (SVM).

The first approach followed the methodology originally proposed by Lucca and Trebbi (2009), who analysed how Federal Open Market Committee (FOMC) monetary policy statements were reflected in press reports. Semantic orientation is a concept from computational linguistics and defines the position of a word or string of words between two opposite concepts. Turney (2002) has used Semantic Orientation (SO) to classify reviews as positive or negative. In practice, the SO score is defined as the difference between the strength of its association with a set of words associate to two opposite concepts such as "hawkish" and "dovish" in a subset R of a corpus of texts, i.e.:

$$\begin{aligned} SO &= \log\left(\frac{\Pr(Hawkish\ \&\ R)}{\Pr(Hawkish) \times \Pr(R)}\right) - \log\left(\frac{\Pr(Dovish\ \&\ R)}{\Pr(Dovish) \times \Pr(R)}\right) \\ &= \log\left(\frac{\Pr(Dovish)}{\Pr(Hawkish)}\right) + \log\left(\frac{\Pr(Hawkish\ \&\ R)}{\Pr(Dovish\ \&R)}\right) \end{aligned}$$

In this way, the values taken on a numeric scale by the SO reflects the relative frequency of two opposite concepts, in this case "hawkish"and "dovish".

The HD index based on this method is computed by counting the co-occurrences of strings with words and expressions that are normally associated with these extreme concepts on a set of texts reporting on ECB monetary policy decisions. Texts are extracted from a corpus of media reports (Dow Jones's Factiva). The formula used is the following:

$$HD_t = \frac{\sum_{s(t)}(-1) \times I[s(t), R, D] + \sum_{s(t)}(+1) \times I[s(t), R, H]}{\sum_{s(t)} I[s(t), R, D] + \sum_{s(t)} I[s(t), R, H]}$$

with $I(\ )$ being the indicator function that counts the co-occurrences of a word *s(t)* pre-classified as dovish (D) or hawkish (H) in a given set of articles referring to a time t extracted from a corpus through a filter R = {"European Central Bank" or "ECB" or "Mario Draghi" etc.}. By construction, the HD index takes values in the interval [−1; +1], where −1 indicate a maximum degree of dovishness and +1 the opposite.

The alternative methodology applies text mining techniques, i.e. a Support Vector Machines (SVM) classification model following Provost and Fawcett (2013). This technique automatically looks for patterns in text documents to select the words with the highest discriminative power. The output is a linear model in which each word is assigned a weight in favour of either class +1 (i.e. hawkish) or –1 (i.e. dovish). A clear advantage is that this algorithm looks at every document as a whole and therefore tends to overcome the limitations of a predefined set of keywords as in the previous method.

To initialise the algorithm, a so-called *training set* was formed by selecting approximately 550 articles and pre-classifying them as "hawkish" or "dovish". These articles were randomly selected from the available corpus. However, articles of uncertain classification were excluded not to introduce any bias and to enhance prediction accuracy. All articles were transformed into a 'bag-of-words' vector, i.e. $[t_0\ t_1 \dots t_j \dots t_n]$ containing all n unique words present in the training set, with t_j being the occurrences of word j in the article. From such vector a term-frequency matrix $tf(m,n)$ is built – with m being the number of articles and n the total number of words and – in which each cell (i,j) indicates the number of times a word j occurs in article i.

In order to reduce the relative weight of words occurring very frequently in the training set of articles, each term count is multiplied by the inverse document frequency (*idf*), which measures the frequency of a term across all documents (Weiss et al., 2010), i.e.:

$$idf(t,m) = \log \frac{\text{Number of articles in the training set } (m)}{\text{Number of articles in the training set where term } \boldsymbol{t} \text{ occurs}}$$

The resulting *tf-idf* matrix is used as input to the SVM algorithm searching for the "decision boundary" maximising the margin between the two classes, i.e. "dovish" and "hawkish" in this case.

Linear SVM tries to solve the following optimization problem (Fan et al., 2008):

$$\min_{w} \frac{1}{2} w^T w + C \sum_{i} \max(1 - y_i w^T x_i, 0)^2$$

with w being the vector of the weights in the model, x_i and y_i representing respectively the input vector and the label of the i-th observation, while C is a cost parameter defined exogenously.

Articles are classified based on the following linear model:

$$f(x_i) = w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_j x_{ij} + \cdots + w_n x_{in}$$

in which the weights w_j are estimated from the optimisation problem and x_{ij} is the occurrence (frequency) of the j-th unique terms of the training set in the i-th article. The sign of the resulting decision value $f(x_i)$ is the predicted class the article belongs to (i.e. hawkish or dovish), whereas the value of $f(x_i)$ approximates the article's degree of hawkishness or dovishness. In this way, the larger the decision value is, the more certain the classifier is about the chosen class. Document of uncertain classification would therefore tend to have a decision value around 0, which can be interpreted as neutral tone in this context.

3. Data and results

To compute our HD-index, we used articles extracted from Dow Jones Factiva. The selection was restricted to articles mentioning "ECB" or "European Central Bank" or the name of its President and limited to categories 'Major News and Business

Publications: Europe' and 'Major News and Business Publications: US'. Only articles in English were included in the sample. Because the focus was the ECB press conference, a three-day window around the press conference (i.e. the day before, the very day and the day after) was used to extract articles. Using these criteria, we formed a corpus of slightly less than 9,000 articles, published between January 1999 and the latest press conference, covering therefore the whole history of the euro (in this paper, however, we included data until January 2016).

Figure 1 shows the news sources included in our data set and the respective proportion of articles in the corpus.

Sources and respective proportion of articles in the corpus

Figure 1

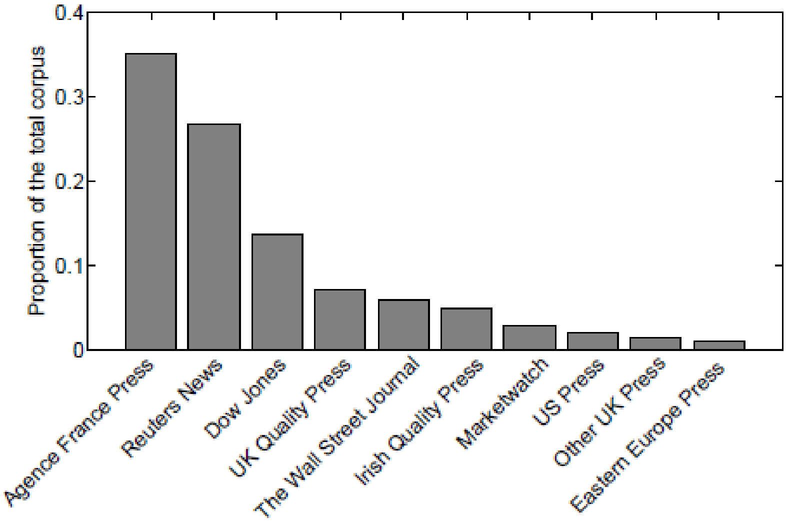
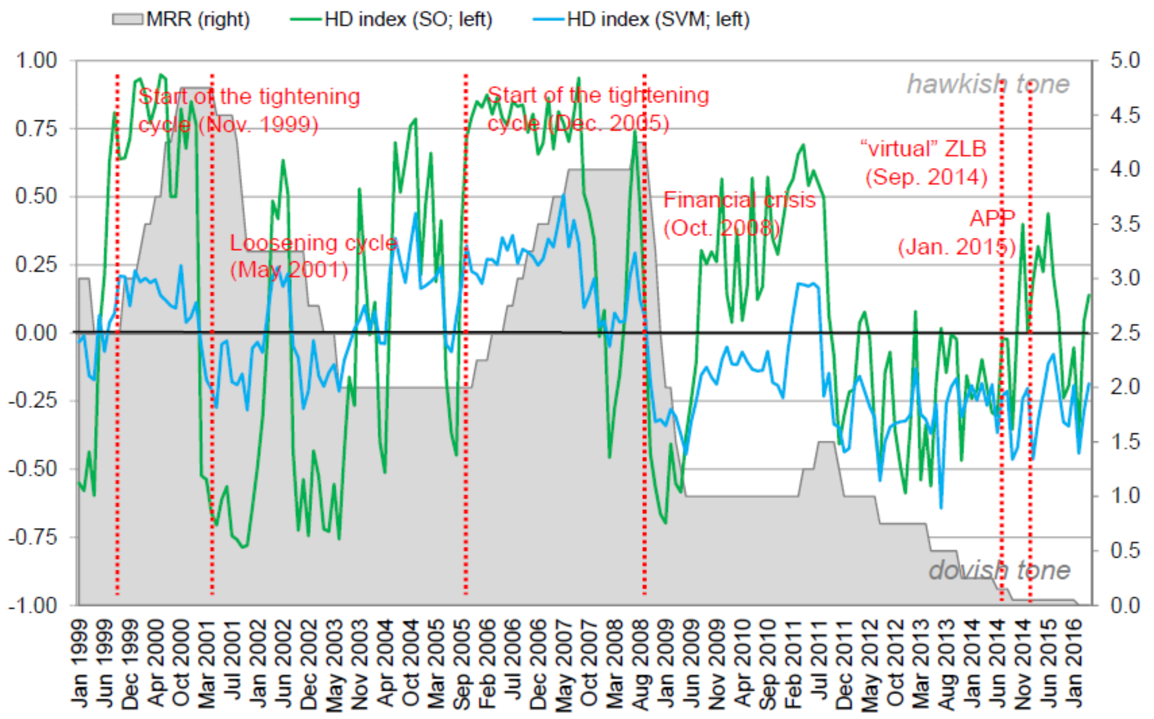


Figure 2 shows the index computed using the two methodologies presented in Section 2 for the whole sample. The two series are contrasted with the evolution of the ECB monetary policy rate (namely, the interest rate on main refinancing operations) to contextualize them. Some relevant episodes are also indicated. Overall, the HD index generally appears to anticipate turning points in the ECB monetary policy stance and to be consistent with the different monetary policy cycles with a few exceptions. In other words, changes in the tone of communication that anticipate official interest rate movements generally appear to be understood appropriately. Specifically, the HD index rose prior to the two tightening monetary policy cycles started at the end of 1999 and at the end of 2005, but also HD declined ahead of the loosening cycles started in mid-2001 and the one caused by the deterioration of the financial crisis following Lehman's collapse in autumn 2008 (in particular the SVM index moved from a very hawkish to a neutral/dovish stance already in late 2007). During the crisis period the SVM HD index has consistently hovered around negative values thus indicating a prevailing dovish tone perceived in the ECB's communication. The decline to relatively low values in 2011 anticipated the marked loosening in monetary policy conditions taking the form of unprecedented standard and non-standard measures in order to bring inflation in line with its objective. While relatively low values were reached in the OMT phase, interestingly the HD index has become relatively less dovish in the period following the announcement of the extended APP, probably reflecting the belief that the ECB loosening policy reached a bottom or, at least, that what was done was sufficient.

As regards differences in results between the two methodologies, although they broadly show a similar dynamics (the correlation between the two series is 0.78), the SO method tends to produce more marked volatile results and occasionally inconsistent results (for instance, towards the end of 2014).

HD index computed with SO and SVM and ECB official interest rate Figure 2



4. Validation criteria

Whereas the SVM methodology overcomes some obvious limitations of the SO methodology (the preselection of a fixed set of words in the first place), there is no obvious validation method to compare the results obtained under the two methods. In this section we propose on three different criteria:

1. we measure how well both methods can predict whether the tone of an article is predominantly hawkish or dovish;
2. we analyse in a deeper fashion the link of the HD index with some interest rates through correlations; and
3. we evaluate the indicators' performance qualitatively by extracting the most important topics mentioned in the news articles to investigate if they can be linked with various phases of monetary policy as identified by the HD index.

Performance analysis

The classifications (scores) obtained from the two methodologies are compared with our own manual classifications using two performance metrics: Area under the receiver operating curve (AUC) and accuracy. AUC is a standard evaluation metric for classification models that represents a model's discriminative power by measuring to what extent positively labelled observations (hawkish) are ranked higher than negatively (dovish) labelled observations (Fawcett 2006). Unlike accuracy, which represents the percentage of correctly classified observations, AUC is able to deal with unbalanced distributions. Results are shown in Table 1.

Performance results of the SVM and SO to classify out-of-sample articles

Table 1

Method	AUC	Accuracy
SO	69.24%	62.00%
SVM	98.55%	92.00%

The SVM methodology outperforms SO when classifying the test articles as hawkish or dovish. We find a difference in accuracy of 30 percentage points in favour of SVM. The more advanced SVM-classification model is able to look at the broader context of an article, which results in a dovish classification for the example article.

Correlation analysis

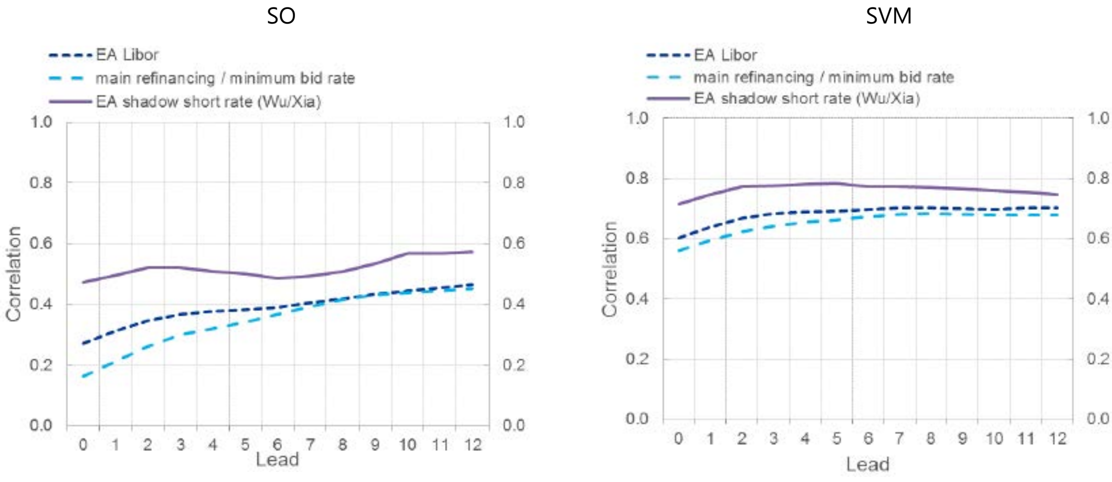
The visual inspection in Section 3 is substantiated here with the analysis of the correlation of the HD index and the ECB main refinancing rate. In addition, correlations were also computed with the Wu-Xia shadow rates to take into account the actual stance when official interest rates reached the zero lower bound and with the 12-month euro LIBOR rate, a money market interest rate reflecting expectations on official rates over the one-year horizon.

Figure 3 shows correlograms from zero to twelve meetings ahead and can be interpreted as an indication on the lead properties of the HD index on actual monetary policy moves. It is important to stress that evidence of correlation should be interpreted here as a positive indication about the ability of the press to interpret correctly communication on future monetary policy stance (and of the index to represent significantly such interpretation).

Positive correlation actually exists for the HD index and tends to be higher for the SVM methodology. The HD index tends to lead actual interest rates moves after 6 to 7 Governing Council meetings and then levels off for the SVM (8 to 9 meetings for SO). Furthermore, a relatively higher correlation (reaching a 0.8 peak) exists with the Wu-Xia shadow rates with a lead of 4 meetings in the case of SVM; also in this case SO tends to have a have a weaker correlation of only 0.6 and a lead of 11 meetings. A similar picture emerges from the correlation with the 12-month euro LIBOR rate with a lead of approximately 7 meetings and a peak correlation of about 0.7 for HD SVM and 0.45 for HD SO.

Correlograms of the HD index with various interest rates

Figure 3


SO

SVM

Topic classification

In this case, a topic classification model was applied to the dataset divided into 13 periods selected on the basis of official interest rate developments (i.e. rising, stable, declining) and consistently with the values of the HD index. The idea behind is to check whether an unsupervised algorithm is able to identify topics that can be reasonably associate with the various phases of ECB monetary policy. If that is the case, one can conclude that communication successfully conveyed (and reporters were able to relate) the most relevant elements of the monetary policy discourse.

Topics are derived using a Latent Dirichlet Allocation (LDA) as in Blei et al. (2003). This method assumes that each document can be represented by a mixture of topics, where the topic distribution is assumed to follow a Dirichlet prior. Every topic has a probability of generating a set of words, which makes it possible to characterise the topic on the base of the related words. In other words, a topic is defined by a cluster of words, which tend to appear together in a document or corpus and are identified by the LDA algorithm.

The topic subjects and a selection of the top 10 words for these topics can be found in Table 2, whereas Figure 4 shows the three most dominant topics in each period and contrasted with the HD index.

Topics identified by the LDA algorithm tend to be consistent with the tightening cycles (and the hawkish peaks of the HD index). Between May and July 2002 and between June 2004 and April 2005 both topic A – characterised by words as "rise", "hike", "increase" as well as "price" associate to a tightening monetary policy – and topic C – characterised by words "growth", "inflat[ion]", "price" all pointing to a strengthening of economic cycle – tend to dominate. In tightening and easing cycles, dominant topics are those related to the interest rate movements, i.e. topic A or B. However, topic B (pointing to rate cut) tends also to appear in the same phases, possibly indicating that a level of uncertainty or disagreement not only in periods of stable monetary policy rates but also in those with interest rates increasing. Interestingly, Figure 4 confirms also the evolution in topics from those predominantly related to interest rate decisions (A and B) or price/growth expectations (C) to crisis (E) and asset/bond purchases (F) after the onset of the financial crisis until recently. Between 2013 and 2014 the dominant topic is indeed the one containing words such as "bond", "purchase" and "debt" as well as "crisis" and "cut". Incidentally, these topics are better captured by the SVM methodology than the SO, which would require the inclusion of *ad hoc* expressions to characterise monetary policy also in periods in which non-standard instruments (and vocabulary) are predominantly used.

Top 10 words for the most frequent topic identified by the LDA

Table 2

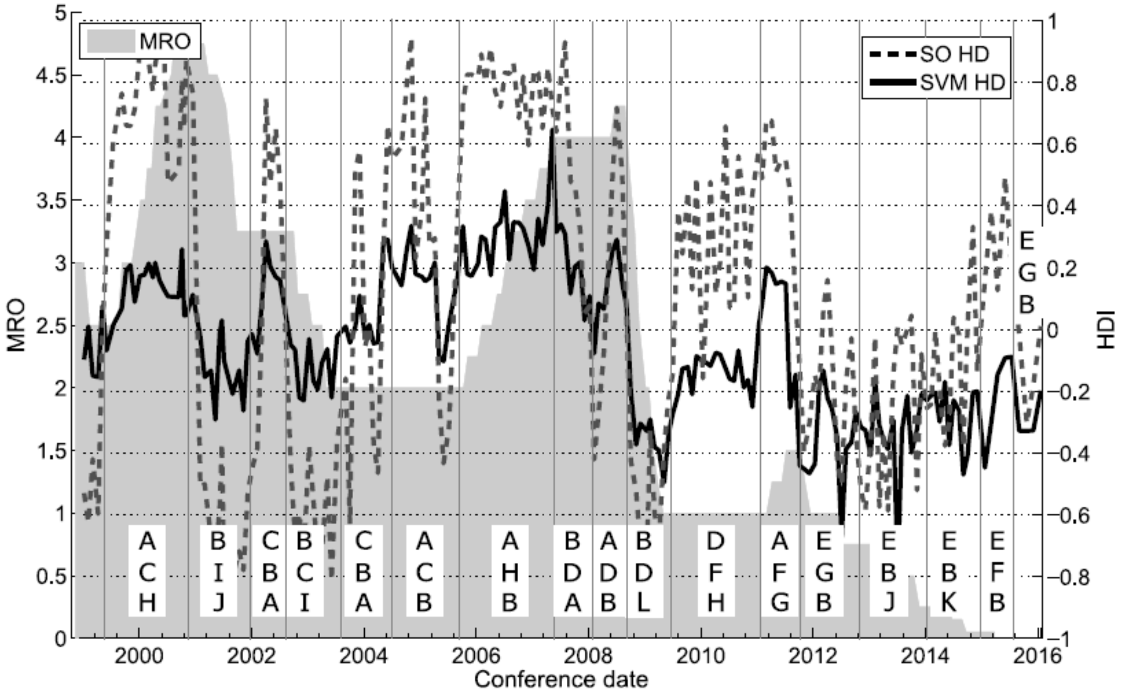
Topics

A	B	C	D	E	F	G	H	I	J	K
rate	rate	growth	Trichet	ECB	Greece	bond	price	cut	stock	rate
ECB	ECB	price	economy	Draghi	debit	market	growth	ECB	market	month
Trichet	cut	euro	market	bond	Greek	yield	remain	rate	share	Euribor
inflate	euro	economy	lend	crisis	ECB	debt	medium	point	index	market
price	bank	rate	record	cut	crisis	govern	risk	decision	bank	ECB
rise	zone	ECB	low	purchase	bond	Spain	inflate	move	expect	lend
bank	inflate	inflate	credit	monetary	default	purchase	stability	announce	rose	fixed
hike	expect	policy	recession	interest	bailout	crisis	expect	Thursday	trade	expect
interest	year	stability	financial	buy	plan	spread	develop	market	fell	overnight
increase	economy	recovery	crisis	debt	fund	Italy	continue	dollar	gain	EONIA

Note: Topics as identified by the LDA algorithm are indicated by the capital letter on top of each column

Figure 4. The HD index and most dominant topics in 13 periods

Figure 4



5. Conclusions

In this paper, we present the development of a numerical indicator that represents the perceived degree of hawkishness or dovishness based on news articles reporting on ECB monetary policy decisions. The evidence provided indicates that the HD index can bring value to central bank communication. On a methodological ground, data mining techniques such as SVM prove to be superior to develop an index measuring the tone of communication as they are more flexible and may offer more insightful information. Specifically, we show that it is possible to extract and analyse the most frequently used topics in the text data, that for the HD index confirms a shift in media focus from the standard interest rate setting to non-standard monetary policy instruments.

Apart from the results presented in this paper, the methodology can be easily applied to other communication outlets or extended to include other languages consistent with the multilingual nature of the Economic and Monetary Union to make it a fully-fledged tool to effectively monitor media reports and assessing effectiveness of ECB communication.

References

Blei, D. M., A. Y. Ng and M. I. Jordan (2003), "Latent Dirichlet Allocation", The Journal of Machine Learning Research, 3, 993–1022.

Fawcett, T. (2006), "An introduction to ROC analysis", Pattern Recognition Letters, 27, 861–874.

Hansen, S. and M. McMahon (2016), "Shocking language: Understanding the macroeconomic effects of central bank communication", Journal of International Economics, 38th NBER International Seminar on Macroeconomics: S114–S133.

Hayo, B., and M. Neuenkirch (2015), "Self-monitoring or reliance on media reporting: How do financial market participants process central bank news?", Journal of Banking and Finance, 59, 27–37.

KOF Swiss Economic Institute (2007), "KOF Monetary Policy Communicator for the Euro Area". (http://www.kof.ethz.ch)

Lucca, D.O. and F. Trebbi (2009), "Measuring central bank communication: an automated approach with application to FOMC statements". Technical Report National Bureau of Economic Research.

Nechio, F. and R. Regan (2016), "Fed Communication: Words and Numbers", FRBSF Economic Letters, 2016–26.

Provost, F. and T. Fawcett (2013), Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc.

Turney, P. D. (2002), "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 417–424.

Weiss, S. M., N. Indurkhya and T. Zhang (2010), Fundamentals of predictive text mining, Springer Science & Business Media.

Wu, J. C., and F. D. Xia (2016), "Measuring the macroeconomic impact of monetary policy at the zero lower bound", Journal of Money, Credit and Banking, 48, 253–291.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Between hawks and doves: measuring Central Bank Communication[1]

Stefano Nardelli,
European Central Bank

[1] This presentation was prepared for the meeting. 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 meeting.

EUROPEAN CENTRAL BANK

EUROSYSTEM

Stefano Nardelli
European Central Bank
Monetary Policy Strategy Division

Between Hawks and Doves
Measuring Central Bank Communication

IFC – Bank Indonesia Satellite Seminar on “Big Data”
Bali, 21 March 2017

Motivation

- Usually analysis is to evaluate **financial market developments** to check whether they are in line with what intended and avoid unintended consequences

  «*The policy stance may also be affected by a continued appreciation of the exchange rate*» (M. Draghi, 24 April 2014)

- or to extract what markets and analysts expect about **future monetary policy moves**

  «*While the governing council did not take any concrete action, the ECB president raised the rhetoric and said his central bankers would be "unanimous" in backing more radical measures, including quantitative easing, to cope with a "too prolonged a period of low inflation." Analysts viewed the shift in tone as substantial …* » (Financial Times, 4 April 2014)

- However, the message delivered by the central bank **may not be perceived in the intended way** by stakeholders (financial markets, banks, public opinion)

  «*The markets got it wrong in forming their expectations. They did indeed have higher expectations than were there and that's why they reacted like they reacted but that was not our intention*» (V. Constâncio, 3 December 2015)

  «*It is not the task of a central bank to correct the erroneous opinions of individuals ... the central bank has to do what it considers right* » (E. Nowotny, 9 December 2015)

Motivation

To assess to measure effectiveness of communication is crucial to develop **quantitative measures**

- *traditional approach:* “event study”, i.e. measure the impact on some market variables around some crucial communication events (e.g. press conference, speeches, press releases, etc.)

- *text analysis approach*: derive indicators directly on texts relevant for communication, issued either by CB or external watchers/stakeholders, using computational linguistic techniques

The HD index Two alternative methodologies

- The **HD index** is computed on articles and newswires reporting on ECB press conference
- First version computed using **Sematic Orientation** technique based on a fixed set or pre-determined word/expressions exogenously classified as either dovish of hawkish to determine the tone of a document
- Alternative version makes use of **Support Vector Machine** (**supervised machine learning algorithm**) to further reduce subjective interpretation
- This algorithm automatically looks for **patterns in text documents** to select the words with the highest discriminative power and determines the tone of a document based on them

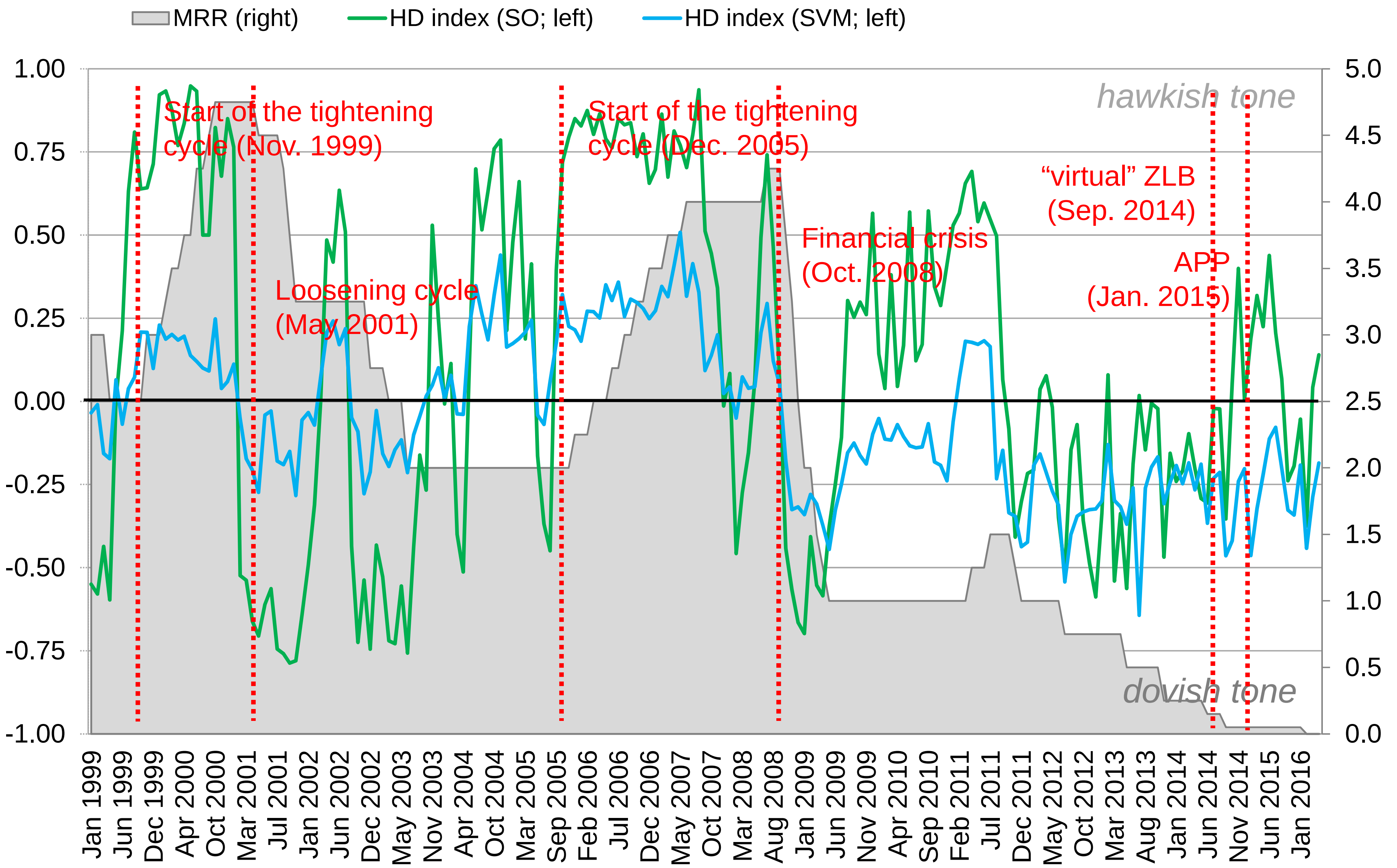
The HD index Computational details

- Main **data source** is **Factiva**, i.e. Dow Jones's global news database featuring nearly 33,000 sources (e.g. Dow Jones Newswires, The Wall Street Journal and Barron's).But *Financial Times* also included

- The **HD index** is computed on a subset of articles about the ECB press conference for a 3-day time window "around" the event (i.e. day before *t-1*, very day *t* and day after the pc *t+1*)

- The **time series** starts in January 1999 and covers the whole history of the euro (more than 9,000 articles in total)

The HD index


MRR (right)
HD index (SO; left)
HD index (SVM; left)
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
hawkish tone
dovish tone
Start of the tightening cycle (Nov. 1999)
Loosening cycle (May 2001)
Start of the tightening cycle (Dec. 2005)
Financial crisis (Oct. 2008)
“virtual” ZLB (Sep. 2014)
APP (Jan. 2015)
Jan 1999
Jun 1999
Dec 1999
Apr 2000
Oct 2000
Mar 2001
Jul 2001
Jan 2002
Jun 2002
Dec 2002
May 2003
Nov 2003
Apr 2004
Oct 2004
Mar 2005
Sep 2005
Feb 2006
Jul 2006
Dec 2006
May 2007
Oct 2007
Mar 2008
Aug 2008
Jan 2009
Jun 2009
Nov 2009
Apr 2010
Sep 2010
Feb 2011
Jul 2011
Dec 2011
May 2012
Oct 2012
Mar 2013
Aug 2013
Jan 2014
Jun 2014
Nov 2014
Jun 2015
Jan 2016

The HD index Validation of results

Question: **how to validate results?**

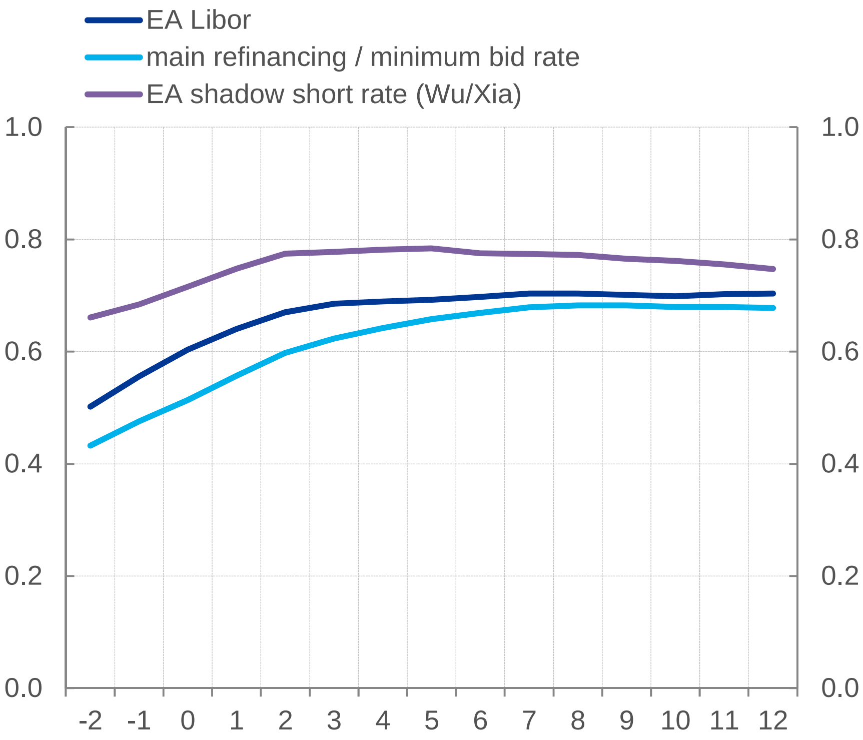
Three indicators are proposed:

- **Correlations** between HD and actual interest rates ("*are actual ECB monetary policy moves anticipated correctly by the media reports?*")
- Identify topics on news dataset using a **topic classification model** (*unsupervised machine learning algorithm*) to assess whether media reports focus on most relevant topics
- To assess superiority of one method **performance metrics** (used increasingly in machine learning and data mining research)

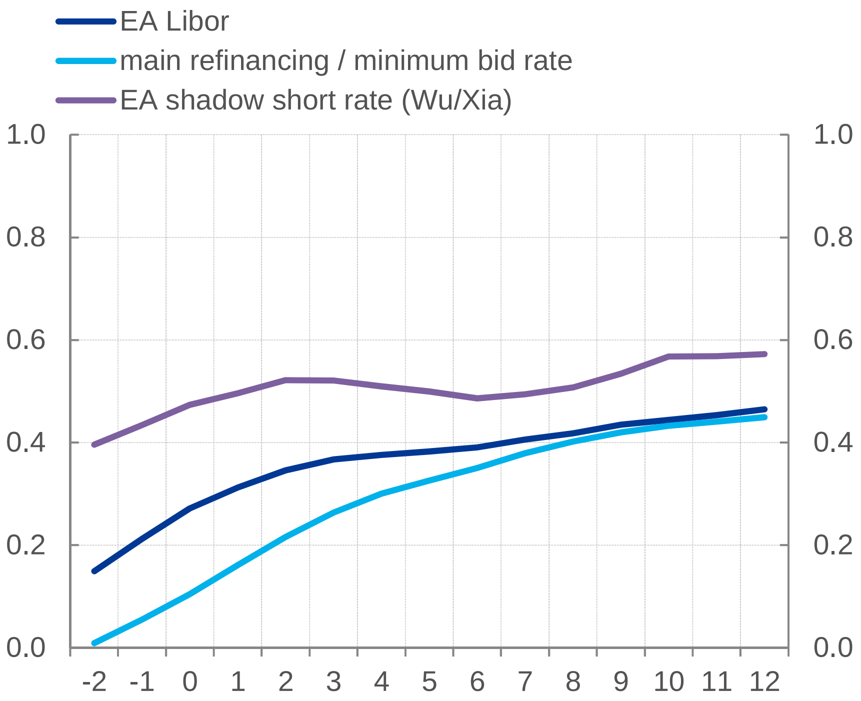
The HD index Leading properties (correlograms)

a. interest rate levels

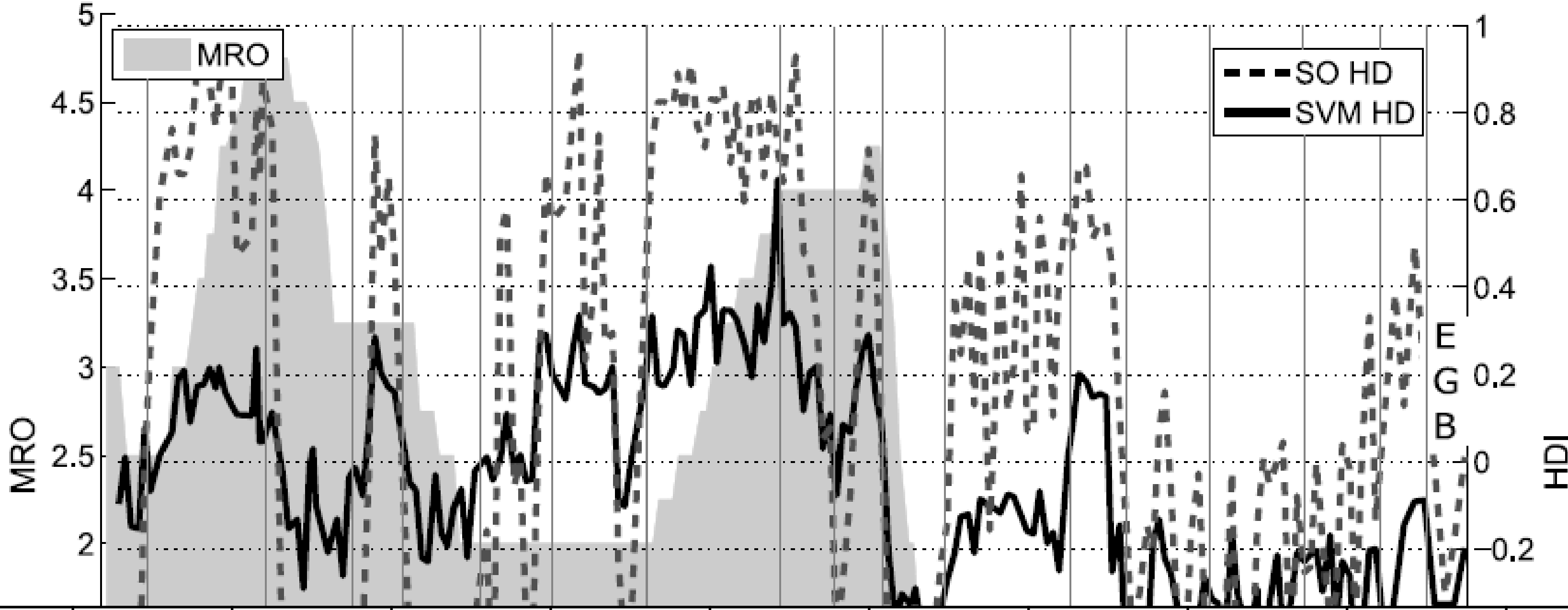
SVM methodology



SO methodology



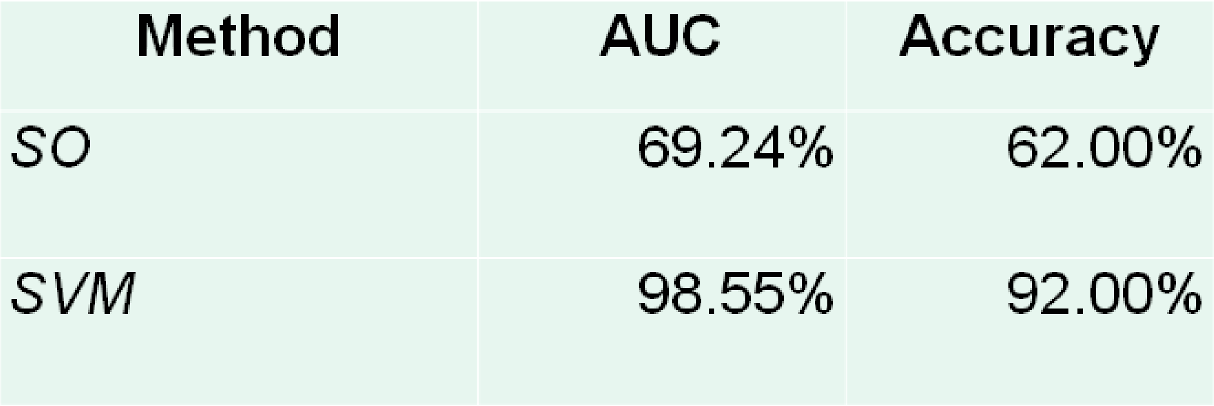
The HD index and most dominant topics in 13 periods



A	B	C	D	E	F	G	H	I	J	K
rate	*rate*	*growth*	*Trichet*	*ECB*	*Greece*	*bond*	*price*	*cut*	*stock*	*rate*
ECB	*ECB*	*price*	*economy*	*Draghi*	*debit*	*market*	*growth*	*ECB*	*market*	*month*
Trichet	*cut*	*euro*	*market*	*bond*	*Greek*	*yield*	*remain*	*rate*	*share*	*Euribor*
inflate	*euro*	*economy*	*lend*	*crisis*	*ECB*	*debt*	*medium*	*point*	*index*	*market*
price	*bank*	*rate*	*record*	*cut*	*crisis*	*govern*	*risk*	*decision*	*bank*	*ECB*
rise	*zone*	*ECB*	*low*	*purchase*	*bond*	*Spain*	*inflate*	*move*	*expect*	*lend*
bank	*inflate*	*inflate*	*credit*	*monetary*	*default*	*purchase*	*stability*	*announce*	*rose*	*fixed*
hike	*expect*	*policy*	*recession*	*interest*	*bailout*	*crisis*	*expect*	*Thursday*	*trade*	*expect*
interest	*year*	*stability*	*financial*	*buy*	*plan*	*spread*	*develop*	*market*	*fell*	*overnight*
increase	*economy*	*recovery*	*crisis*	*debt*	*fund*	*Italy*	*continue*	*dollar*	*gain*	*EONIA*

The HD index Performance analysis

- **Accuracy,** i.e. percentage of correctly classified observations
- **Area u**nder the **Receiver Operating Curve** (AUC), i.e. measure to what extent positively labelled observations (hawkish) are ranked higher than negatively (dovish) labelled observations; AUC can to deal with unbalanced distributions (Fawcett 2006)



Method	AUC	Accuracy
SO	69.24%	62.00%
SVM	98.55%	92.00%

Conclusions

- A quantitative approach to communication is possible
- Tools originally developed from computational linguistics may offer useful insights on CB communication and its perception by stakeholders (media)
- Specifically HD index is a sensible summary indicator about how communication around the ECB press conference is reproduced by media
- The SVM methodology proves superior to the "traditional" SO methodology

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Central Bank Communications: information extraction and semantic analysis[1]

Giuseppe Bruno,
Bank of Italy

[1] This paper was prepared for the meeting. 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 meeting.

Central Bank Communications: Information extraction and Semantic Analysis.

Giuseppe Bruno[*]

Bank of Italy, Economics and Statistics Directorate.

Abstract

Central Banks, among other tasks, provide a relevant amount of information for Institutions and market operators. Indeed central banks employ a multiplicity of communication channels to drive market expectations. In this paper we present some methodologies aimed to quantify the information content of official communications and we present their application to the semi-annual publication of the Financial stability report. While these methodologies are quite developed for the English and other highly spoken languages in the world, they are still in their experimental phase for the Italian language. Here the goal is twofold: on one hand we provide a transparent numerical framework to consider sub-unit of an official Central Bank report written in Italian. Moreover it is proposed an analytical tool to gauge the impact of an official document on the public. In the context of reports released by the Bank of Italy, we show how this framework can be employed to numerically characterize and extract their information content.

We deem quite relevant a quantitative evaluation of the impact of these reports in increasing the central bank transparency with the goal of enhancing the effectiveness of its institutional action.

JEL classification: C83, E58, E61.
Keywords: Text Mining, Semantic Analysis, Pointwise Mutual Information, Web search.

[1] Bank of Italy, Research Area.
[*] giuseppe.bruno@bancaditalia.it
This version February 2017
The author wishes to thank Fabio Di Bernardini for his helpful suggestions and unceasing cooperation.
The views expressed are the author's only and do not imply those of the Bank of Italy.

1 Introduction and motivation

> *For a large class of cases – though not for all – in which we employ the word meaning it can be defined thus: the meaning of a word is its use in the language game.*
>
> – Wittgenstein, Philosophical Investigations A. 43

In the last two decades, the amount of textual information available at everybody fingertips has soared. According to rough estimates 80% of the total amount of web pages on the internet is given by textual or unstructured data. In this paper we take the challenge of adopting and experimenting a methodology for quantifying the linguistic content of some official documents of the Bank of Italy. In particular we take into consideration the Italian version of the Financial Stability report (henceforth *FSR*) . This is a young publication whose first issue came out on 2010. We build a small corpus of all the available issues of the *FSR* and we show how to convert it into a vector space model by employing familiar concept of linear algebra such as eigenvalues and eigenvectors. Among these vector space models, we consider here the *Latent Semantic Analysis* (henceforth *LSA*) model. This is one of the most successful models which is emerged in computational linguistic around 20 years ago (Landauer and Dumais [9] and Landauer et al. [10]). *LSA* was patented in 1998 in the US and it is a widely used technique in Information Retrieval (*IR* and natural language processing for analyzing relationships between a set of documents and the words they contain. Within the framework of the conceptual model of lexical semantics words, statements, chapters and whole documents are represented as high dimensional vectors in the same space. The main advantage of this closed representation is the natural metric induced in the vector space. Therefore in the *LSA* model we can:

1. compute semantic similarity measures between words/documents by exploiting the statistical redundancies in text;
2. compute word neighborhood which are set of words/documents sharing semantic concepts (synonimity);
3. compute text coherence and summary of given documents;
4. answer to multiple choice questions to topics dealt with in the corpus.

We make one step further by taking inspiration from previous works on similar topics such as Lucca and Trebbi [11], Carvahlo et al. [2] and Kawamura et al. [8] and using Web search queries in Italian, we evaluate the public perceptions associated with some keywords associated with the financial stability issues. Although in recent years we have seen a great progress in sentiment classification, it is still challenging to develop a practical sentiment classifier for open applications. Lexical semantic orientation can be measured by complementing the *LSA* with the Pointwise Mutual information (*PMI*) which has proved to be a dominating indicator for sentiment classification (Turney [18]).

The web-derived computation of the *PMI* allow us to measure the people orientation about some relevant theme for the financial stability.

Aside form the previous two economic goals of the paper, here we compare the effectiveness of the most relevant web search engines[1] and the simplicity of two very popular open source software frameworks: **Python** and **R**[2]. The two software packages have been employed for the statistical analysis and for producing the code interacting with the search engines.

In particular, with reference to our corpus composed of the whole set of the Italian version of the Bank of Italy Financial stability reports, this work attempts to answer the following questions:

1) what are the most relevant concepts considered in the documents of the corpus between 2010 and 2016?

2) what is the readability and formality level of each document?

3) how can we measure the impact of the *FSR* on the readers by web searching?

The buoyant literature on these themes (see for example D. Bholat and Schonhardt-Bailey [3] and R. Nyman and Tuckett [14]) confirms the growing interest shown by public institutions and Central Banks on these issues.

The development of web tools for monitoring and extracting sentiment orientation, provides further analytical support fostering a wider adoption of text mining and semantic techniques for improving the statistical accuracy required for assuming well informed decisions.

The paper is arranged in the following way. After this introduction in section 2 we present the basic theoretical concepts behind the text-mining techniques. Section 3 introduces the algebraic concepts behind the Latent Semantic Analysis. Section 4 describes the corpus of the *FSR* for our empirical application. Section 5 presents the results of our semantic analysis carried out on a corpus containing all the 11 editions, from 2010 to 2016, of the Bank of Italy *FSR*. Section 6 presents the methodology employed for evaluating the people orientation with respect to some economic issues through web search. Finally section 7 provides some concluding remarks.

2 Building a Corpus of Text documents

Quantitative analysis of human languages allows to discover common features of spoken or written text. In order to simplify the analysis of written text we have adopted the *Bag-of-words* model (see Salton and McGill [16]). This set-up considers each document as a sequence of different words where the semantic meaning of any statement is conveyed by the word-document co-occurrence while neither grammar nor word order play any determinant role[3]. Therefore this is an orderless representation of the document where we forgo any grammatical relationship among the words. Within the *Bag-of-words* framework each document

[1] Here we have considered Bing, Google and Yahoo which reached about 90% of the market share in 2015.

[2] `www.python.org` and `cran.r-project.org` respectively

[3] A bag or *multiset* is a set-like object where only element multiplicity is accounted for, while the order of its elements is irrelevant.

is represented simply by a vector whose length is the size of the corpus vocabulary and each vector entry is a weighted count of the word occurrences.

Our analysis is based on the text mining methodology which can generally be defined, see for example Hearst [5], as the task of harnessing *a large online text collections to discover new facts and trends about the world itself.* The development of a corpus of textual homogeneous documents is the first step to address any kind of text mining and *LSA* applications. The only determinant factor for the semantic value of a text is the frequency of occurrences of each word. At first glance this seems a pretty unplausible assumption, but for the purposes of information retrieval, comparison and measures of similarity among documents will prove very effective. Our corpus is defined as a set of documents written in the same language. The whole set of different terms constitutes our vocabulary $V = \{k_1, k_2, \ldots, k_N\}$ whose size N is the cardinality of the set. In our examples the terms are words and each one of them is an independent dimension in our vector space [4]. Therefore any word/document is represented by a vector in the space $\mathbb{R}^N$. Since each one of our documents contain a subset of the Vocabulary V their vector representation will be very sparse.

It is evident from this description that the atomic building block of a corpus is a word. Therefore, for processing reasons, all the documents must be converted in plain textual format so that it is straightforward to tally the occurrences of the different words in each document belonging to the considered set. On the other hand, the choice to take into account just the main text imposes the need to give up any consideration about tables, figures and other external elements such as notes, bibliography etc. To this end we proceed by extracting the plain text from the original PDF© or MS-Word© format. Once we convert our original documents into a set of text files we can start the preproprocessing steps, which, depending on the final goals, consist in some of the following tasks:

- lower case conversion and white space removal,
- stopword and number removal,
- stemming or lemmatization,
- special characters conversion or filtering.

The realization of some of these tasks aims to reduce the size of the considered vocabulary allowing to focus on the most relevant topic-determinant words. While tasks like lowercase conversion, number and white space removal are independent of the considered language, for other tasks the language employed in the documents plays a relevant role. Although the analytical instruments for the English language are well developed, software tools for the italian language are yet in their growing phase. In this work we have tested the software capabilities of some of the **R** packages addressing their suitability for carrying out these language dependent tasks on text written in Italian.

The stopword list available in the **tm** package is composed of around 300 tokens which belong to grammatical categories such as determiners, pronouns, conjugated forms of auxiliary verbs. In our empirical analysis we had to complement this list with other italian words with poor information relevance. The stemmer provided by the **SnowballC** package has seemed suitable for the Italian language employed in our official documents[5]. Other preliminary text processing tasks such as synonymous replacement and part of speech tagging are not considered here because out of the scope in our analysis. Once the required preliminary tasks are completed our corpus of text documents is translated into a term by document matrix (henceforth *TDM*) where each entry $a_{i,j}$ gives a weighted frequency of the word i in the document j. This normed vector space representation for the set of our documents constitute the starting point for the semantic analysis described in section 3 .

[4] A single word will be a vector with all zeros but a one in the position of the word in the vocabulary.

[5] SnowballC is a R interface to the C stemmer which implements the Porter's algorithm (see Porter [12] and Porter [13]).

3 Latent Semantic Analysis: algebraic background

When we consider the learning speed with which people stockpile knowledge, we are faced with an apparent puzzle, where it is hardly possible to explain the amount of people's word knowledge by considering the shallowness of their information set. In the past, for example, Landauer and Dumais [9] or Landauer et al. [10] went even way back to Plato's hypothesis that *people must come equipped with most of their knowledge, needing only hints and contemplation to complete it.* Latent Semantic Analysis (*LSA*) tries to set up on more solid foundations the puzzle of excessive learning speed by assuming a multiplicative inference generated by the relationships available with the present stock of knowledge.

The *LSA* methodology was originally suggested by S. Deerwester and Harshman [15] in the framework of automatic indexing and information retrieval (IR). The suitability of *LSA* for different text analytics purposes has been already established in different fields among which we have the just mentioned IR. Adoption of this technique is already significant among central banks, see, for example, Boukus and Rosenberg [1] who analyze the information content in the *FOMC* of the Federal Reserve Board, Hendry and Madeley [6] who carry a similar analysis for the Central Bank of Canada. These two works perform text mining tasks on a corpus of the English language. Carvahlo et al. [2] employs the *LSA* to quantify the informational content in the portuguese version of the statements of the "Comitê de Política Monetária" [6] of the Central Bank of Brazil while Kawamura et al. [8] analyzes the Japanese version of the Monthly Report of the Bank of Japan to check whether the central banks communicate strategically being selective about the type of information they disclose.

Application of the *LSA* methodology rests on the *Bag-of-words* model, where word occurrences and co-occurrences build up the basic information set. The most interesting feature of the *LSA* is its use of the same vector space for both words and documents. In this closed framework a word meaning is not an absolute concept but it is an average of the meaning of all the statements of the corpus including this word. These averages are numerically computable and this quantitative feature is one of the main advantage of the model.

The input element for the *LSA* analysis is a *TDM* matrix as produced with the processing steps described in the chapter 2. Starting with a *TDM*, the *LSA* algorithm can be broken up in three separate steps. The first one consists in rebalancing the relative impact of low- and high-frequency word by applying a weighting scheme to the *TDM*. Among the available weighting schemes we have considered the Tf-Idf (Term frequency-Inverse document frequency) which is very popular in the IR domain[7]. This scheme consists in weighting each nonzero element of the *TDM* in the following way:

$$\omega_{i,j} = wf_{i,j} \cdot \left(\log \left(\frac{m}{df_i} \right) \right) \tag{1}$$

where: $wf_{i,j}$ is the frequency of word i in document j, df_i is the number of documents containing the word i, the total number of documents in our corpus is m, $\log\left(\frac{m}{df_i}\right)$ is the log of the inverse document frequency and $\omega_{i,j}$ is the final value given to word i in document j in the *TDM* matrix. This means we multiply the raw frequency by the logarithmically scaled inverse of document fraction containing the chosen word.

Tf-Idf is a very intuitive weighting scheme which provides a higher weight to words occurring frequently in very few documents (topic determinants words) while giving lower weight to words uniformly present in all the documents of the corpus [8].

In the second step of *LSA* linear algebra kicks in. At this stage we employ the Singular Value Decomposition (SVD) to factorise our weighted rectangular *TDM* matrix in three factors. This decomposition is a generalization to the eigenvalue/eigenvector decomposition by representing each term and document in an orthonormal base. Its importance derives from the circumstance that it can be applied without restrictions to

[6] Monetary Policy Committee.

[7] This scheme provides an higher retrieval accuracy with respect to the simple Tf (Term frequency).

[8] The Idf of a word appearing in every document of the corpus will be zero.

any rectangular matrix. Formally by assuming the weighted *TDM* A_w of size $m \times n$, the SVD decomposition corresponds in determining the three matrices in the right hand side of the following equation:

$$A_w = U \cdot \Sigma \cdot V^T \tag{2}$$

where: U is a $m \times m$ orthonormal matrix containing the left eigenvector, V is a $n \times n$ orthonormal matrix containing the right eigenvector, and Σ is the $m \times n$ rectangular matrix with elements $\sigma_{i,j} = 0$ if $i \neq j$ while the elements $\sigma_{i,i}$ are the singular values. The rows of U contains the word vectors, while the rows of V contains the document vectors. The third and last step of *LSA* corresponds to the dimensionality reduction that is implemented by setting to zero all the singular values below a certain threshold. In our empirical application we considered a very small number of documents [9], therefore we did not carry out this reduction which is required when the number of documents is in the range of hundred of thousands.

The representation of words and documents as vectors in $\mathbb{R}^N$ allows for a straightforward evaluation of a numerical value for the similarity two elements in the vector space. A commonly used measure is the cosine similarity between the two document vectors $\mathbf{x}, \mathbf{y}$ which is given by:

$$\cos(\theta) = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{\sqrt{\left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} y_i^2\right)}} \tag{3}$$

where: θ is the angle between the two document vectors. In absence of weighting all the vector components are positive and $cos(\theta)$ will range from 0, for completely different documents, to 1, for very similar documents. In the more general case of logarithmic weights, therefore $cos(\theta)$ might span its whole range from -1 to 1. Here the similarity concept pertain the use of the same or co-occuring words regardless of their order[10]. In the following sections different applications of this similarity measure will be presented and employed in different circumstances.

4 The Bank of Italy Financial Stability Corpus

For our empirical application we have built a small corpus composed of all the issues of the *FSR*. Bank of Italy rolled out the publication of the FSR in 2010[11]. It started as a yearly publication but in 2012 the report became biannual. The whole report consists of 4 chapters for a total of about 40 pages for issues. Each publication includes an average of 50 graphs, 5 tables and around 10 in-depth information boxes.

To the purpose of our analysis, the documents have been converted in plain text after discarding tables, graphs and other auxiliary elements. Each Report has been split in its component chapters. At the completion of this step we had a corpus of 58 text documents. In the table 1 we show some descriptive statistics about the corpus considered in our experimental set-up. Here we can see a decreasing number of sentences over time with a stable number of word per sentence, characters per word and characters per sentence. These shallow text statistics play a key role for the estimation of the readability and formality of the considered documents. Beside from these general figures, the first statistical analysis carried out on these documents is based on word cloud and heatmaps.

The word cloud is a synthetic picture showing the principal words in the document by resizing their fonts proportionally to their relative frequency. As an example we provide the wordmap for the whole corpus of the 11 issue of the *FSR*. Some of the most relevant words are: *rischio*, *credito*, *liquidità* and *banche* [12]

[9] By splitting in their chapters the 11 *FSR* issues we ended up with a corpus of 58 documents

[10] *LSA* allow to pin down synonyms by harnessing semantic similarity.

[11] Electronic version of the documents are available online at http://www.bancaditalia.it/pubblicazioni/rapporto-stabilita/index.html.

[12] The English translation is risk, credit, liquidity and banks

Table 1

issue	n statem	#word per statem	sd #word	#char per statem	#char per word
2010_1	518	31.30	14.69	182.41	5.83
2011_1	428	32.40	15.29	190.00	5.86
2012_1	295	32.97	16.27	191.99	5.82
2012_2	364	33.18	16.06	192.01	5.78
2013_1	288	32.21	15.56	187.26	5.81
2013_2	317	31.85	15.46	185.60	5.83
2014_1	271	31.52	15.10	181.26	5.75
2014_2	379	34.21	16.64	195.40	5.71
2015_1	266	34.32	14.98	195.94	5.71
2015_2	267	32.21	14.92	183.88	5.71
2016_1	297	32.87	14.94	187.57	5.71

The heatmap is another qualitative summary representation of the *TDM* matrix where the frequency of each word in each document is coded through the color intensity. In the following we provide a heatmap for all the 11 issues where we can see the more frequently used words in these 6 years of the *FSR* publication. A normalized version of the heatmap is shown in fig. 4.2.

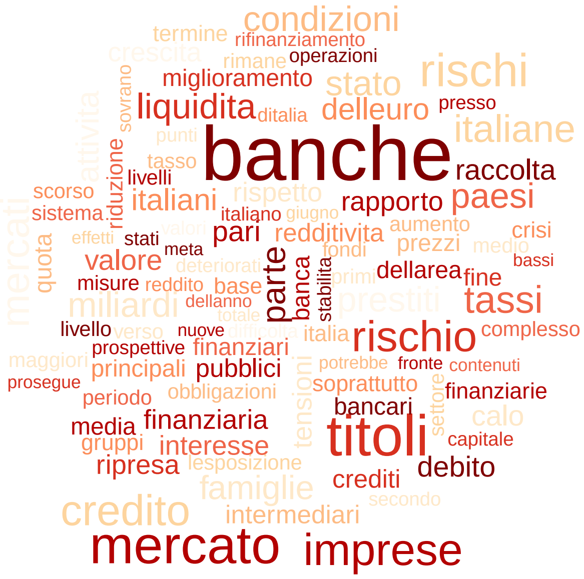
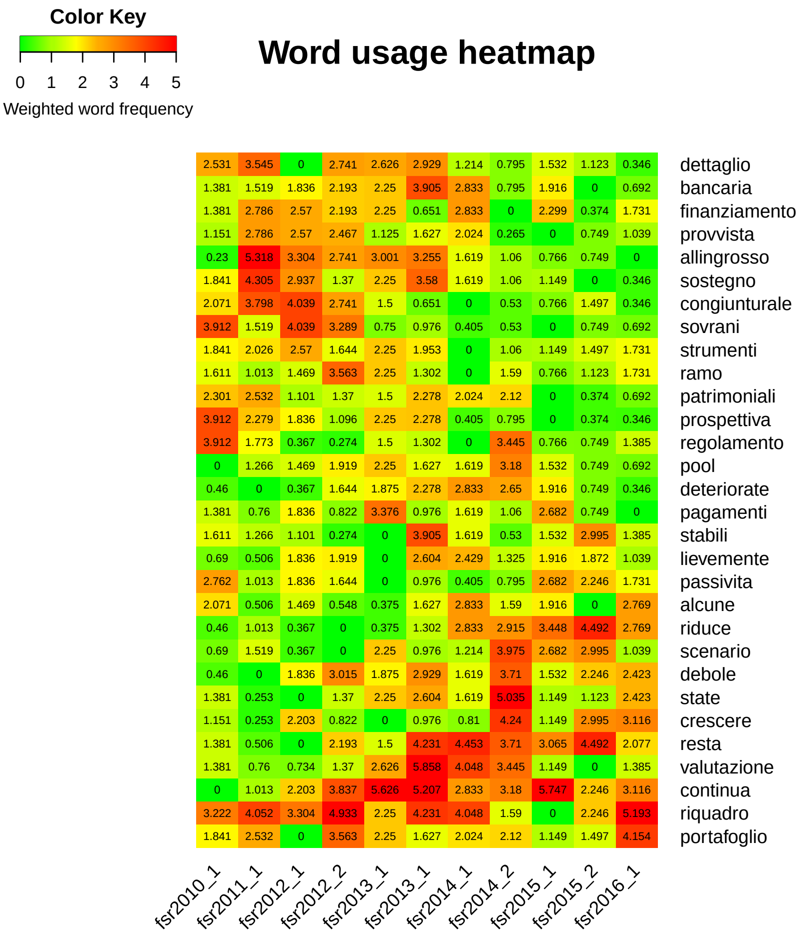


Figure 4.1: wordmap for the whole corpus of the *FSR*



	fsr2010_1	fsr2011_1	fsr2012_1	fsr2012_2	fsr2013_1	fsr2013_1	fsr2014_1	fsr2014_2	fsr2015_1	fsr2015_2	fsr2016_1
dettaglio	2.531	3.545	0	2.741	2.626	2.929	1.214	0.795	1.532	1.123	0.346
bancaria	1.381	1.519	1.836	2.193	2.25	3.905	2.833	0.795	1.916	0	0.692
finanziamento	1.381	2.786	2.57	2.193	2.25	0.651	2.833	0	2.299	0.374	1.731
provvista	1.151	2.786	2.57	2.467	1.125	1.627	2.024	0.265	0	0.749	1.039
allingrosso	0.23	5.318	3.304	2.741	3.001	3.255	1.619	1.06	0.766	0.749	0
sostegno	1.841	4.305	2.937	1.37	2.25	3.58	1.619	1.06	1.149	0	0.346
congiunturale	2.071	3.798	4.039	2.741	1.5	0.651	0	0.53	0.766	1.497	0.346
sovrani	3.912	1.519	4.039	3.289	0.75	0.976	0.405	0.53	0	0.749	0.692
strumenti	1.841	2.026	2.57	1.644	2.25	1.953	0	1.06	1.149	1.497	1.731
ramo	1.611	1.013	1.469	3.563	2.25	1.302	0	1.59	0.766	1.123	1.731
patrimoniali	2.301	2.532	1.101	1.37	1.5	2.278	2.024	2.12	0	0.374	0.692
prospettiva	3.912	2.279	1.836	1.096	2.25	2.278	0.405	0.795	0	0.374	0.346
regolamento	3.912	1.773	0.367	0.274	1.5	1.302	0	3.445	0.766	0.749	1.385
pool	0	1.266	1.469	1.919	2.25	1.627	1.619	3.18	1.532	0.749	0.692
deteriorate	0.46	0	0.367	1.644	1.875	2.278	2.833	2.65	1.916	0.749	0.346
pagamenti	1.381	0.76	1.836	0.822	3.376	0.976	1.619	1.06	2.682	0.749	0
stabili	1.611	1.266	1.101	0.274	0	3.905	1.619	0.53	1.532	2.995	1.385
lievemente	0.69	0.506	1.836	1.919	0	2.604	2.429	1.325	1.916	1.872	1.039
passivita	2.762	1.013	1.836	1.644	0	0.976	0.405	0.795	2.682	2.246	1.731
alcune	2.071	0.506	1.469	0.548	0.375	1.627	2.833	1.59	1.916	0	2.769
riduce	0.46	1.013	0.367	0	0.375	1.302	2.833	2.915	3.448	4.492	2.769
scenario	0.69	1.519	0.367	0	2.25	0.976	1.214	3.975	2.682	2.995	1.039
debole	0.46	0	1.836	3.015	1.875	2.929	1.619	3.71	1.532	2.246	2.423
state	1.381	0.253	0	1.37	2.25	2.604	1.619	5.035	1.149	1.123	2.423
crescere	1.151	0.253	2.203	0.822	0	0.976	0.81	4.24	1.149	2.995	3.116
resta	1.381	0.506	0	2.193	1.5	4.231	4.453	3.71	3.065	4.492	2.077
valutazione	1.381	0.76	0.734	1.37	2.626	5.858	4.048	3.445	1.149	0	1.385
continua	0	1.013	2.203	3.837	5.626	5.207	2.833	3.18	5.747	2.246	3.116
riquadro	3.222	4.052	3.304	4.933	2.25	4.231	4.048	1.59	0	2.246	5.193
portafoglio	1.841	2.532	0	3.563	2.25	1.627	2.024	2.12	1.149	1.497	4.154

Figure 4.2: Heatmap

From this heatmap we can easily track down the evolution of the more frequently used words over the past six years. It can be seen that in 2011 the words *sostegno* and *congiunturale* [13] were hot topics. In the second half of 2013 we have again *supporto* along with *bancaria* and *valutazione*. In the last issue *portafoglio* [14] becomes a relevant topic. These words constitute a clue in representing some topics. Further investigation with *LSA* might confirm the true role of the frequency of these words in signalling interest in given semantic concepts.

[13] In Italian they are respectively support and short-term

[14] Portfolio in Italian.

5 LSA application with Financial stability reports

In order to check the actual behavior of the available text mining procedures we have taken into account the corpus composed of the 11 issues of the *FSR*. These documents are available in PDF© format at the Bank of Italy web site.

A relevant consideration to take into account here is the language of the corpus. As a matter of fact many computational tools are already quite developed for the English, German and Chinese languages while they are still at a rather infancy stage for the Italian.

The first *LSA* analysis run on the *FSR* corpus has consisted in deriving a coherence measure on each chapter of the different 11 issues. The coherence index is a measure contained in the range $(0 \div 1)$, values above 0.5 signal a semantic similarity among the sentences composing the chapters. The **R** software provides the package **LSAfun** with a *coherence* function computing both a local (statement to statement) and a global coherence. Here we present just the global coherence which is an average value among all the statements in a chapter. In fig 5.1 we show the the evolution of the coherence over the different 58 chapters and the coherence of a 10 statements automatic summary achieved by applying once again the *LSA* methodology as proposed in Gong and Liu [4].

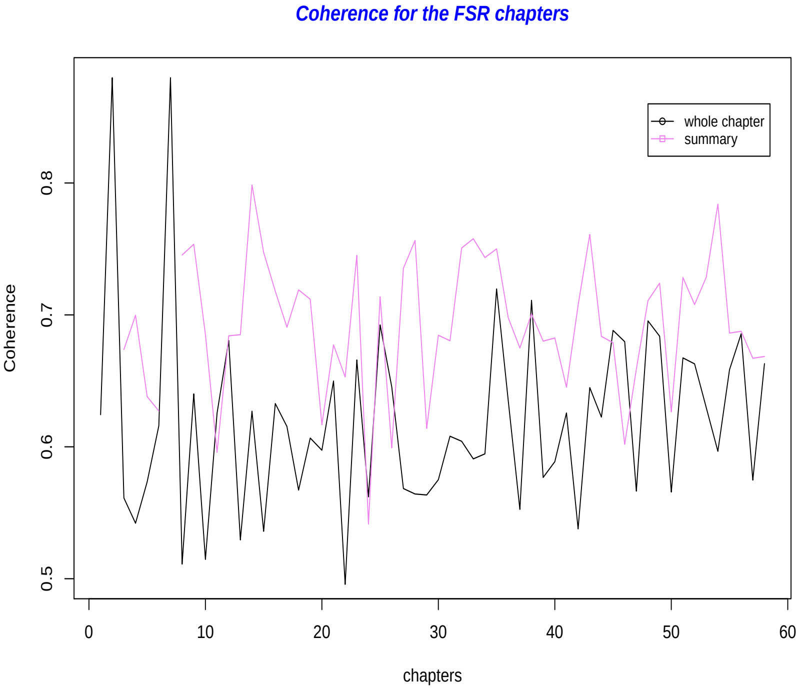


Figure 5.1: Coherence evolution for the *FSR*

In the table 2 we show the average coherence statistics for the whole set of chapters and for a summary of each issue of the *FSR*:

Table 2

variable	mean	std dev
global	0.80	0.030
summary	0.85	0.032

The first row of the table show the average coherence of the whole reports. The second row lists the average coherence of the automatic summary generated as a by product of the *LSA* methodology. These values provide an objective measure of the high coherence level shown by the chapters of the corpus.

As a second application of the *LSA* we have computed the nearest neighbors for the italian translation of the words crisis and stability[15]. For each given word a nearest neighbors is a word with semantically similar meaning. These neighbors are computed by ranking in decreasing order the similarity between the each couple of words composed of the reference and another word of the vocabulary. In the figure 5.2 we show the top five nearest neighbors to the word *crisi* using the Multidimensional Scaling on the similarity matrix of all the available couple of words. The concept of crisis emerges as strictly linked with debt, consolidation and countries. These words having semantically closeness with *crisi* could play the role of either causes or consequences.

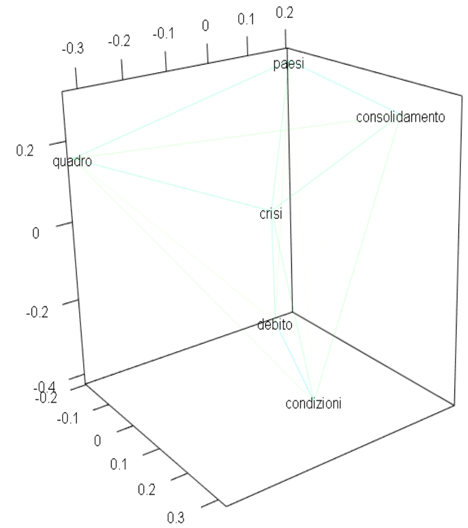


Figure 5.2: Top 5 neighbors for crisi

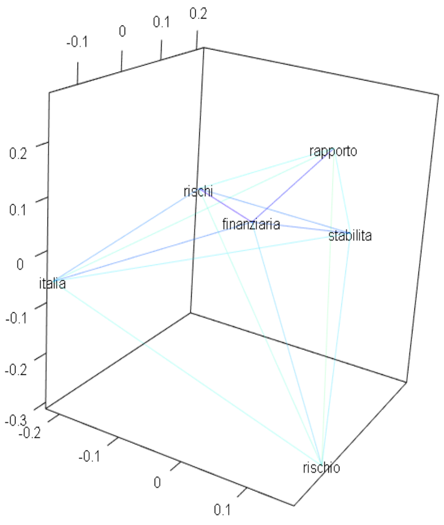


Figure 5.3: Top 5 neighbors for stabilità

The second example is presented in figure 5.3. Here we show the nearest neighbors to the word *stabilità*. This time the *stabilità* concept is obviously closely related with the word *finanziaria*[16] and with the words *rischio/rischi* and *Italia*. The close connection among some neighbor words might sporadically look mysterious, and sometimes words that should be close are not. One possible explanation for this phenomenon could come from a bias due to the too thinly sampled words (small sampling bias).

[15] In Italian they are respectively crisi and stabilità

[16] The two words stabilità and finanziaria constitute the title of the report

5.1 Gauging the Readability for the *FSR*

Evaluating in an automatic way the understandability of a text is a relevant factor for estimating its general public acceptance. Most of the classical readability metrics are linear model of few superficial features of words and sentences such as those shown in chapter 4. These readability indexes are generally given by a linear combination two proxies: a) the word difficulty measured by the number of letters per word, b) the sentence difficulty given by the number of words per sentence. In our empirical application we have chosen to employ the following functions available in the **qdap** R package. It is the following:

1. Automate Readability Index, $ARI = 4.71 \cdot \left(\frac{N_{char}}{N_{words}}\right) + .5 \cdot \left(\frac{N_{words}}{N_{sentences}}\right) - 21.43$
2. Fleisch Vacca test, $FV = 206.0 - 1.0\left(\frac{N_{words}}{N_{sentences}}\right) - 0.65\left(\frac{N_{syllables}}{N_{words}}\right)$

In the two formulae we have:

N_{char} is the character count for every word in the text,

N_{words} is the word count for every sentence and

$N_{sentences}$ is the total number of sentences in the text,

$N_{syllables}$ is the total number of syllables in the text.

These two indexes tend to reward the employment of short words and short sentences. ARI is a simple empirical derivation for the English language which provides a useful comparison tool over time and among different documents belonging to the same class. In our experiment we have used the test for a corpus written in Italian. In this case we extend to the Italian language the assumptions made about difficulty of words and sentences in English. The Fleisch-Vacca (FV) index is a modification of the Fleisch-Kincaid measure proposed for the Italian language[17]. This index approximate the readability ease and is generally comprised between 0 and 100. Values around 100 indicate a very simple reading while values below 30 are judged as texts requiring a degree for their understanding. In the pictures 5.4, 5.5, 5.6 and 5.7 we show the readability evolution for four of the past issues of the *FSR*.

In the pictures 5.8, 5.9, 5.10 and 5.11 we show the readability evolution computed with the Fleisch Vacca index for the same issues of the *FSR*.

Values of the ARI index in the pictures are around 20. In this case the measure is upward biased because the ARI Index was originally designed for the English language where words and sentences are on average shorter than their Italian translation by respectively 10% and 40%. Therefore this value represents an upper bound for the true readability. The readability values provided by the Fleisch-Vacca index are all between 30 and 35 meaning that a degree is required for their understanding. The results from the two measures appear slightly different: while the ARI index seems to indicate a readability at the college level degree, the FV measure judges the *FSR* as a text requiring a higher level of education. A possible avenue for further analysis could be the implementation of the GULPEASE index which was devised directly for the Italian language[18]. The final result can be interpreted as a general understandability of the *FSR* by people between an average to higher specialization.

5.2 Measuring the Formality for the *FSR*

Another relevant feature allowing to numerically gauge the degree of the context-dependence of a document is the formality of a document. There are some different definitions for the formality. Here we have chosen the definition suggested in Heylighen and Dewaele [7] where the Formality score is calculated according

[17] This test has been obtained modifying the coefficients provided in the Fleisch Kincaid test of the **qdap** package

[18] GULP is the Gruppo Universitario Linguistico Pedagogico (Linguistic Pedagogical University Group).

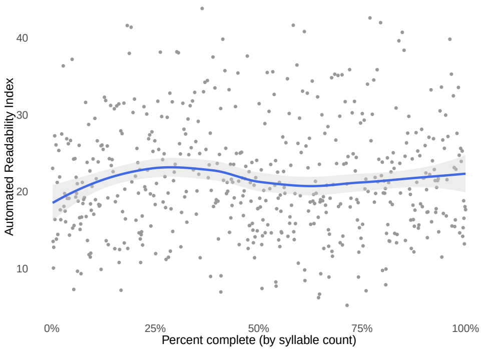


Figure 5.4: Readability of Financial Stability 2010

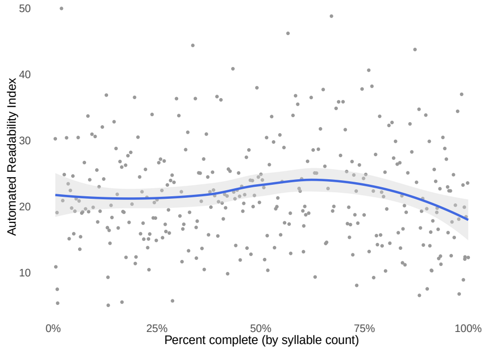


Figure 5.5: Readability of Financial Stability 2013-2

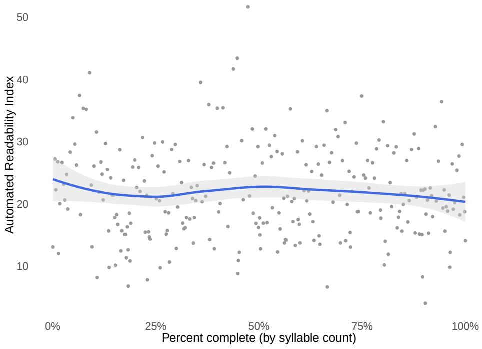


Figure 5.6: Readability of Financial Stability 2015-2

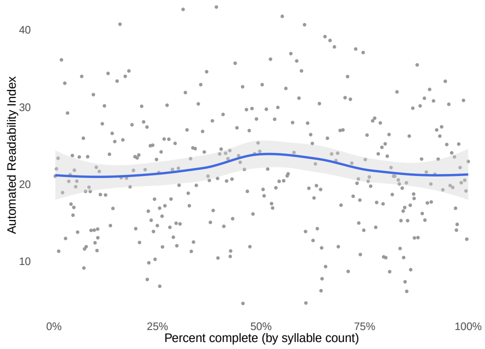


Figure 5.7: Readability of Financial Stability 2016-1

the equation:

$$F = 50 \cdot (\frac{n_f - n_c}{N} + 1) \tag{4}$$

where: $f = \{nouns, adjective, preposition, article\} \qquad n_f = |f|,$

$c = \{pronoun, adverb, verb, interjection\} \qquad n_c = |c|$

$N = \sum(f + c + conjunctions)$

This Formality score gets higher when statements make more use of nouns and adjectives rather than pronouns and adverbs.

In the picture 5.12, 5.13, 5.14 and 5.15 we show the formality evolution for four of the past issues of the *FSR*,

Reference values for the Formality measure taken from Heylighen and Dewaele [7] indicates values of 70 as highly formal. The last issues of the *FSR* feature a systematically higher values for the Formality index which is in the range $[75 \div 80]$. This result confirms our intuition of highly formal documents leaving few room to individual interpretation. The generally high formality value for the *FSR* signals a substantial absence of semantic ambiguity.

6 The web impact of the *FSR* through Google search

This section tries to answer the third question put forth in the introduction. It describes the methodology employed for evaluating the impact produced by the *FSR* on the web by counting the hits returning a web search suitably defined[19]. This technique, based on the measurement of the similarity of a pairs of phrases, has been already put forward and employed by different authors. Originally proposed by Turney [18] for the

[19] This impact is defined as semantic orientation in Lucca and Trebbi [11]

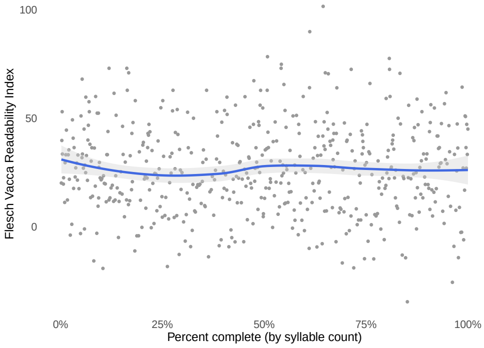


Figure 5.8: Fleisch-Vacca Readability of Financial Stability 2010

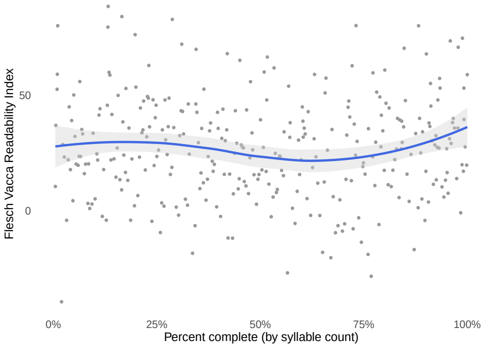


Figure 5.9: Fleisch-Vacca Readability of Financial Stability 2013-2

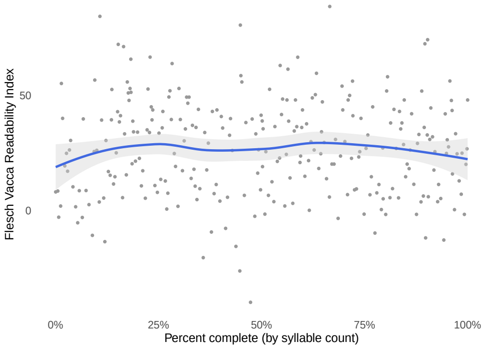


Figure 5.10: Fleisch-Vacca Readability of Financial Stability 2015-2

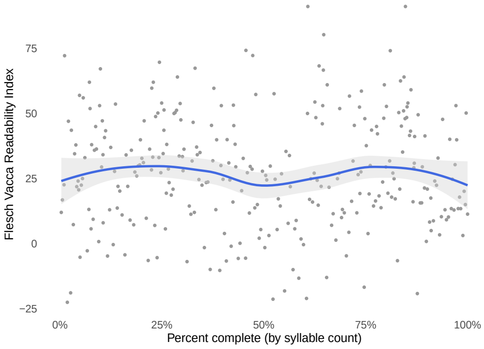


Figure 5.11: Fleisch-Vacca Readability of Financial Stability 2016-1

unsupervised classification of reviews. In the framework of Central Banks communications, to the best of our knowledge, the first paper making use Google hits count to gauge the web reaction about an economic issues is Lucca and Trebbi [11] which derive the relevant web score by applying an information-theoretic based tool. Similar technique are employed by Carvahlo et al. [2] for the evaluation of the information content of the interest rate setting statements of the Central Bank of Brazil and by Kawamura et al. [8] in analyzing the hypothesis of strategic disclosure by the Bank of Japan. The concept of orientation or value judgment towards a statement is based on the comparison of two distances. The first one is the distance between the considered phrase and a positively polarized word (e.g. ”stability”) and the second is the distance between the same phrase and a negatively polarized word (”instability”). When the phrase appears closer to the positively polarized word we assign a positive orientation to the phrase. On the other hand if the statement appears closer to the negatively polarized word we attribute a negative orientation to the statement. The distance between two words or sentences is assumed to be given by the Pointwise Mutual Information (PMI) which measures the likelihood that the first word/sentence will appear along with the second one. The formal definition of the PMI is given by

$$PMI(\phi_1, \phi_2) = \log \left[\frac{p(\phi_1, \phi_2)}{p(\phi_1) \cdot p(\phi_2)} \right] \tag{5}$$

It can take positive or negative values. It is zero when the two words/sentences are independent. Because of the practical unfeasibility to compute the probabilities considered in equation (5), we approximate them by evaluating the number of *Web search hits* in the web search of the statement of our document associated with the two opposing adjective (antonym). In their work regarding monetary policy, Lucca and Trebbi [11] provide the example of the antonymy ”hawkish” versus ”dovish”. In practice, after having extracted the more sensible statements relating to monetary policy actions, the semantic orientation *SO* is evaluated by estimating the difference between the two *PMI* of the extracted statement and each term of the antonymy. In this paper we extend the Lucca and Trebbi [11] procedure by taking into account the whole set of statements contained in each issue of the *RSF*. In this way we attempt to average the polarity with respect to a given antynomy over all the statements.

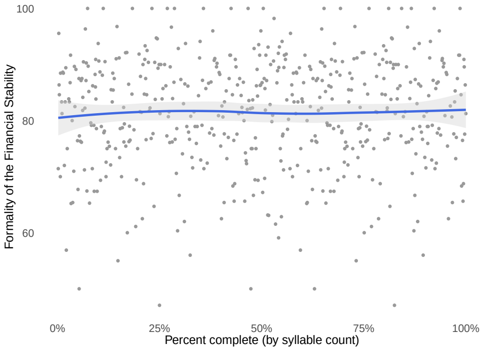


Figure 5.12: Formality of Financial Stability 2010

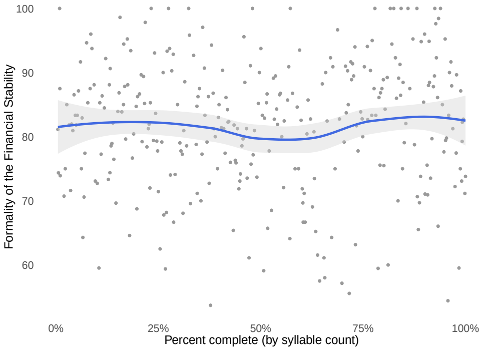


Figure 5.13: Formality of Financial Stability 2013-2

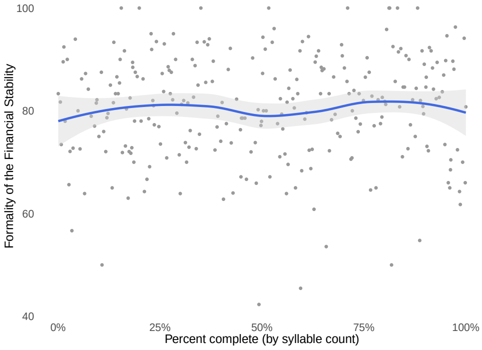


Figure 5.14: Formality of Financial Stability 2015-2

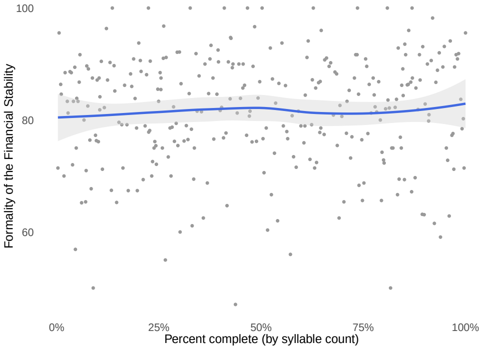


Figure 5.15: Formality of Financial Stability 2016-1

Formally, called $\mathbb{X}$ the statement under scrutiny [20], the *SO* of the statement will be estimated as:

$$SO(\mathbb{X}) = PMI(\mathbb{X}, hawkish) - PMI(\mathbb{X}, dovish) \tag{6}$$

While in the given references the web search has been carried out only on Google, in our investigation we went further on examining the stability of web search against three of the major web search engine: Bing, Google and Yahoo [21]. The details about code employed are in appendix. Here we have found that the number of web search hits are not very stable among the considered search engines.

In our empirical exercise we have considered the semantic orientation relative to each statement of the different *FSR* with respect to the following three antonyms:

1. stabilità/instabilità (stability/instability)
2. crisi/espansione (stability/instability)
3. vulnerabilità/solidità (stability/instability)

As an example we provide here the three pictures for the *PMI* referring to the three antonyms taken into account. In the figure 6.1 we show the *PMI* evolution for three different editions. The initial issue of 2010, the first issue of 2014 and finally the same variables are shown for the first issue of 2016.

[20] As an example, Lucca and Trebbi [11] consider the statement $\mathbb{X} \equiv$ ”pressures on inflation have picked up in recent months”.

[21] On the 25th of July 2016 Yahoo has been acquired by Verizon

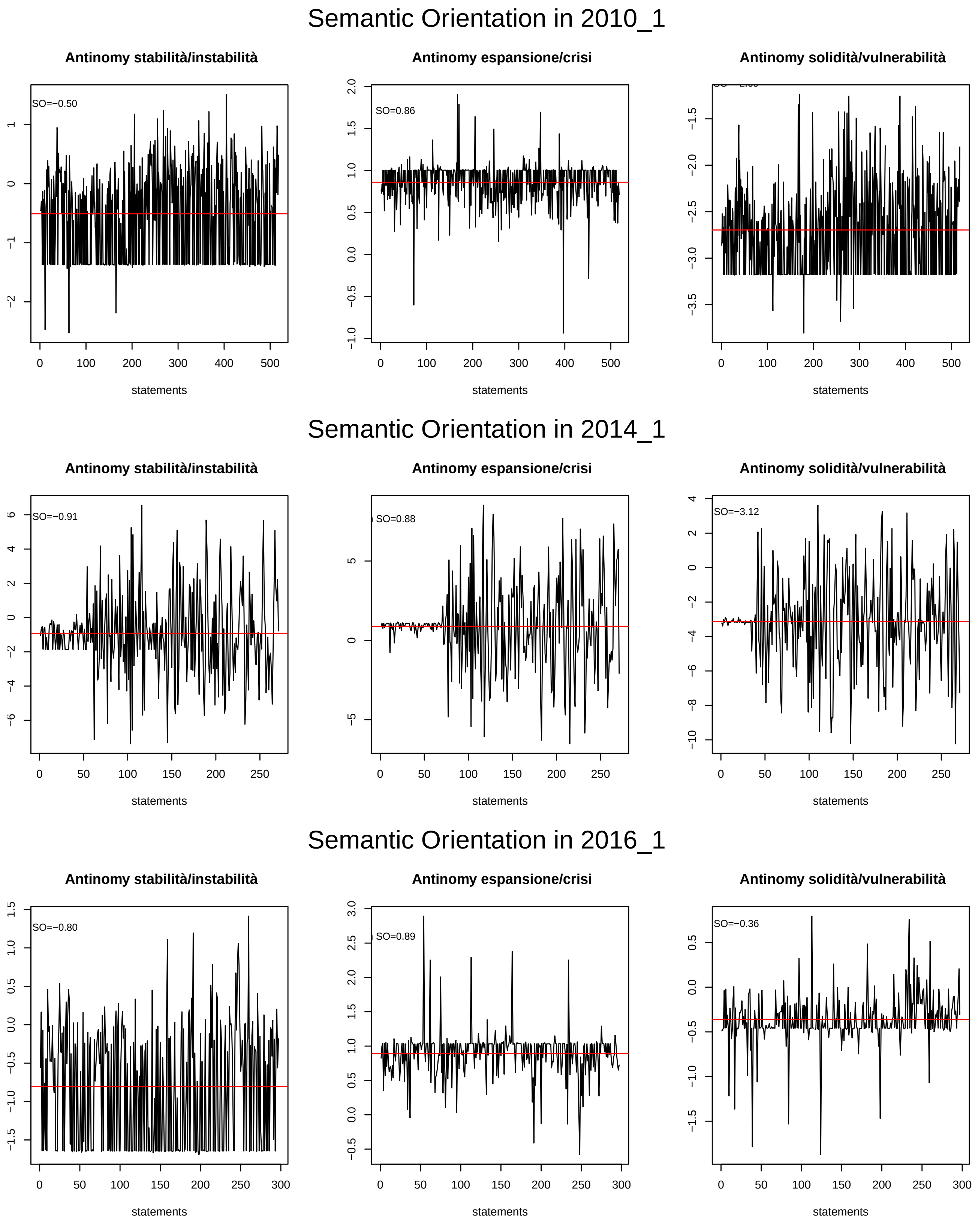

Semantic Orientation in 2010_1
Antinomy stabilità/instabilità
SO=−0.50
−2
−1
0
1
0
100
200
300
400
500
statements
Antinomy espansione/crisi
SO=0.86
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
0
100
200
300
400
500
statements
Antinomy solidità/vulnerabilità
−3.5
−3.0
−2.5
−2.0
−1.5
0
100
200
300
400
500
statements
Semantic Orientation in 2014_1
Antinomy stabilità/instabilità
SO=−0.91
−6
−4
−2
0
2
4
6
0
50
100
150
200
250
statements
Antinomy espansione/crisi
SO=0.88
−5
0
5
0
50
100
150
200
250
statements
Antinomy solidità/vulnerabilità
SO=−3.12
−10
−8
−6
−4
−2
0
2
4
0
50
100
150
200
250
statements
Semantic Orientation in 2016_1
Antinomy stabilità/instabilità
SO=−0.80
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
0
50
100
150
200
250
300
statements
Antinomy espansione/crisi
SO=0.89
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
50
100
150
200
250
300
statements
Antinomy solidità/vulnerabilità
SO=−0.36
−1.5
−1.0
−0.5
0.0
0.5
0
50
100
150
200
250
300
statements

Figure 6.1

These three pictures can be read along two dimensions. By considering the three antonyms we see that in 2010 and in 2014 there is a strong feeling of vulnerability while in 2016 we see more neutral semantic orientation. By considering time evolution, we see a more confident/positive perception on all the three antonyms. Here we cannot take any definitive position but we believe these analyses could spawn many complementary tools improving the effectiveness of the standard communication means. The same web search has been tested on different web search engines. The results are similar between Bing and Yahoo, while they are significantly different with Google. We interpret this result with the difficulty of tracking down the internal behavior of the different search engines. It seems necessary to further our analysis by cooperating with the Companies providing these web query tools.

7 Concluding Remarks

In this paper we have presented the main computational linguistic methodologies for mining the relevant information from a corpus of documents and evaluating some summary statistics. These methodologies employ a tool-set taken from the IR technology.

We have shown that these technologies can be quite helpful in quickly analyzing huge amount of documents and automatically extracting sentiment or opinion orientation and gauging polarity of these sentiments. By taking advantage of some packages available on the CRAN[22] repository, we have written some *R* procedures implementing different algorithms for text mining and sentiment analysis. These *R* procedures have been applied for the analysis of an homogeneous corpus of documents based of the Bank of Italy *Financial stability report* which started in 2010.

The main conclusions are the following:

1) the *FSR* has shown a high level of local and global coherence;

2) the ARI and the FV readability indexes show a readability which implies approximately a college to higher degree for understanding the texts;

3) the examined corpus of documents, as it could be expected, has shown a quite high average level of the Formality score. This result witnesses the scarcity of room left to ambiguities.

The present strand of research looks quite promising especially for the possibility to quickly provide institutional answers more closely connected to the social emotions and preferences of the different economic agents.

References

[1] Boukus, E. and J. V. Rosenberg (2006). The information content of fomc minutes. *Federal Reserve Bank of New York* (NA), 1 – 53.

[2] Carvahlo, C., C. Cordeiro, and J. Vargas (2013). Just words? a quantitative analysis of the communication of the central bank of brazil. *Revista Breasileira de Economia 67*(4), 443 – 455.

[3] D. Bholat, S. Hansen, P. S. and C. Schonhardt-Bailey (2015). Text Mining for Central Banks. *Centre for Central Banking Studies 33*, 1–19.

[4] Gong, Y. and X. Liu (2015). Text mining for central banks. *Centre for Central Banking Studies- Bank of England 33*, 1 – 19.

[5] Hearst, M. A. (1999). Untangling Text Data Mining. *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 3–10.

[22] Comprehensive R Archive Network, https://cran-r-project.org

[6] Hendry, S. and A. Madeley (2010). Text Mining and the information content of Bank of Canada Communications. *Bank of Canada Working Paper*, 2–53.

[7] Heylighen, F. and J. Dewaele (2002). Variation in the Contextuality of Language: an Empirical Measure. *Foundation of Science*, 293–340.

[8] Kawamura, K., Y. Kobashi, M. Shizume, and K. Ueda (2016). Strategic central bank communication: Discourse and game-theoretic analyses of the bank of japan's monthly report. *JSPS Working Paper Series* (80), 1 – 34.

[9] Landauer, T. K. and S. Dumais (1997). A solution to platos problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211 – 240.

[10] Landauer, T. K., P. Foltz, and D. Laham (1998). Introduction to latent semantic analysis. *Discourse Processes 25*, 259 – 284.

[11] Lucca, D. O. and F. Trebbi (2011). Measuring central bank communication: an automated approach with applications to fomc statements. *NBER working paper* (15367), 1 – 37.

[12] Porter, M. F. (1980). An Algorithm for suffix stripping. *Program 14*(3), 130 – 137.

[13] Porter, M. F. (2006). Stemming Algorithms for various European languages.

[14] R. Nyman, P. O. and D. Tuckett (2015). Measuring financial sentiment to predict financial instability: A new approach based on text analysis. *University College London*.

[15] S. Deerwester, S. Dumais, G. F. and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science 6*(41), 391 – 407.

[16] Salton, G. and M. J. McGill (1983). *Introduction to Modern Information retrieval*. New York: McGraw Hill Book Co.

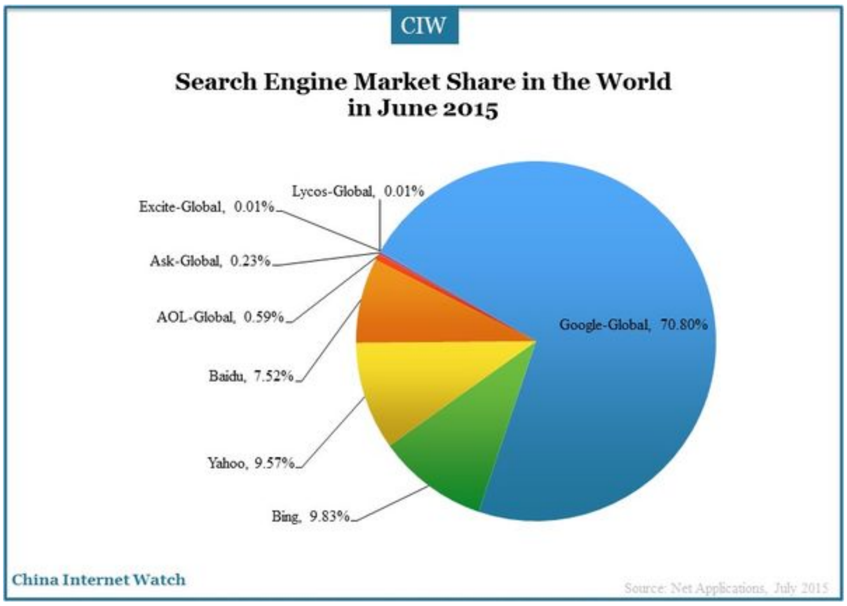
[17] Senter, R. J. and E. A. Smith (1967). Automated readability index. *Aerospace Medical Research Laboratories*.

[18] Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of review. *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, 417–424.

Appendices

A Market share distribution for Web search engines.

The following figure shows the market share for the main web search engine in 2015: from this picture we realize that the search engines we consider cover over 90% of market share.



B Code snippet for computing number of web-hits.

B.1 Google web search

For computing the the web-hits by using the Google search engine we have used the **R** programming environment. the kernel code adopted is the following:

```
require("XML")
require("RCurl")
# Function to compute the number of hits on a given Google search
# Adapted from theBioBucket at http://goo.gl/TXvTxP
GoogleHits <- function(query){
        url <- paste0("https://www.google.com/search?q=", gsub(" ", "+", query))
        CAINFO = paste0(system.file(package="RCurl"), "/CurlSSL/ca-bundle.crt")
        script <- getURL(url, followlocation=T, cainfo=CAINFO)
        doc    <- htmlParse(script)
# Results look like this:
# <div class="sd" id="resultStats">About 10,300,000 results</div>
        res <- xpathSApply(doc, ’//*/div[@id="resultStats"]’, xmlValue)
        return(as.numeric(gsub("[ˆ0-9]", "", res)))}
```

B.2 Bing and Yahoo web search

For Bing and Yahoo we have employed the Python language [23]. This choice has been taken simply for a lack of R packages for using these two search engines.

[23] Python was first released in 1991 by the dutch programmer Guido van Rossum.

```
# Bing Cognitive Search API
def bingcs(**kwargs):
    """
                Bing query language:  https://msdn.microsoft.com/en-us/library/ff795620.aspx
                Bing CS Search API:   https://msdn.microsoft.com/en-us/library/ff795657.aspx
    """
    KEY="XXXXXXXX"

    import requests
    url = 'https://api.cognitive.microsoft.com/bing/v5.0/search'
    payload = kwargs.copy()
    if 'fields' in payload.keys(): payload.pop('fields')
    headers = {'Ocp-Apim-Subscription-Key': KEY}
    r = requests.get(url, params=payload, headers=headers)
    j = r.json()
    if 'fields' in kwargs.keys():
        try:
            return _pathGet(j, kwargs['fields'])
        except KeyError:
            return 0
    else:
        return j

# Yahoo public search
def yahoo(**kwargs):
    """
    Documentazione parametri Yahoo: https://search.yahoo.com/web/advanced
    research suggestions:  https://help.yahoo.com/kb/search/improve-yahoo-search-results-sln2242.ht
    """
    import requests
    import string
    from bs4 import BeautifulSoup
    url = 'https://search.yahoo.com/search'
    payload = kwargs.copy()
    req = requests.get(url, params=payload)
    soup = BeautifulSoup(req.content, 'html.parser')
    try:
        text = soup.find("div", class_="compPagination").find("span").text
        return ''.join(ch for ch in text if ch not in string.punctuation).split()[0]
    except AttributeError:
        return 0
```

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Central Bank Communications: information extraction and semantic analysis[1]

Giuseppe Bruno,
Bank of Italy

[1] This presentation was prepared for the meeting. 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 meeting.

Central Bank Communications: Information extraction and Semantic Analysis.

Giuseppe Bruno[1]

[1] Economics and statistics Directorate
Bank of Italy

IFC - Bank Indonesia Satellite Seminar on Big Data March 21st 2017

Outline

1. Motivation
2. Shallow and Syntactic features of documents
   - Readability & Formality
3. Latent Semantic Analysis
4. Pointwise Mutual Information and Semantic Orientation
   - Web hit computed Pointwise Mutual Information.
5. Concluding Remarks

The Questions we are going to address

Extracting information from textual data

- The web provides an enormous warehouse of information. Around $4/5$ of this info is of textual nature.
- What is the impact of the Bank's communications? Can we devise an objective measurement mechanism?
- What is the measure of the sentiment caused by these communications?

[I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind. Lord Kelvin - 1883]

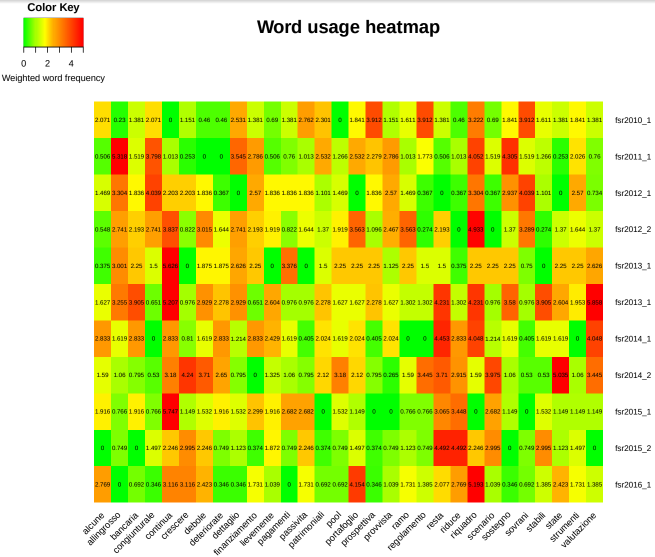
Statistics for the Financial stability report

Some linguistic statistics.

issue	#sentence	#word per sentence	sd #word	#char per sentence	#char per word
2010_1	518	31.30	14.69	182.41	5.83
2011_1	428	32.40	15.29	190.00	5.86
2012_1	295	32.97	16.27	191.99	5.82
2012_2	364	33.18	16.06	192.01	5.78
2013_1	288	32.21	15.56	187.26	5.81
2013_2	317	31.85	15.46	185.60	5.83
2014_1	271	31.52	15.10	181.26	5.75
2014_2	379	34.21	16.64	195.40	5.71
2015_1	266	34.32	14.98	195.94	5.71
2015_2	267	32.21	14.92	183.88	5.71
2016_1	297	32.87	14.94	187.57	5.71

The Financial stability report appeared in 2010. It started as a yearly publication. In 2012 the report became biannual.

It has about 40 *pages, with* 50 *graphs,* 5 *tables and around* 10 *in-depth information boxes.*

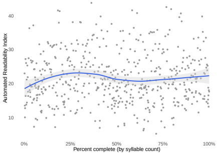


	alcune	allingrosso	bancaria	congiunturale	continua	crescere	debole	deteriorate	dettaglio	finanziamento	lievemente	pagamenti	passivita	patrimoniali
fsr2010_1	2.071	0.23	1.381	2.071	0	1.151	0.46	0.46	2.531	1.381	0.69	1.381	2.762	2.301
fsr2011_1	0.506	5.318	1.519	3.798	1.013	0.253	0	0	3.545	2.786	0.506	0.76	1.013	2.532
fsr2012_1	1.469	3.304	1.836	4.039	2.203	2.203	1.836	0.367	0	2.57	1.836	1.836	1.836	1.101
fsr2012_2	0.548	2.741	2.193	2.741	3.837	0.822	3.015	1.644	2.741	2.193	1.919	0.822	1.644	1.37
fsr2013_1	0.375	3.001	2.25	1.5	5.626	0	1.875	1.875	2.626	2.25	0	3.376	0	1.5
fsr2013_1	1.627	3.255	3.905	0.651	5.207	0.976	2.929	2.278	2.929	0.651	2.604	0.976	0.976	2.278
fsr2014_1	2.833	1.619	2.833	0	2.833	0.81	1.619	2.833	1.214	2.833	2.429	1.619	0.405	2.024
fsr2014_2	1.59	1.06	0.795	0.53	3.18	4.24	3.71	2.65	0.795	0	1.325	1.06	0.795	2.12
fsr2015_1	1.916	0.766	1.916	0.766	5.747	1.149	1.532	1.916	1.532	2.299	1.916	2.682	2.682	0
fsr2015_2	0	0.749	0	1.497	2.246	2.995	2.246	0.749	1.123	0.374	1.872	0.749	2.246	0.374
fsr2016_1	2.769	0	0.692	0.346	3.116	3.116	2.423	0.346	0.346	1.731	1.039	0	1.731	0.692

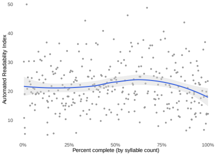
	pool	portafoglio	prospettiva	provvista	ramo	regolamento	resta	riduce	riquadro	scenario	sostegno	sovrani	stabili	state	strumenti	valutazione
fsr2010_1	0	1.841	3.912	1.151	1.611	3.912	1.381	0.46	3.222	0.69	1.841	3.912	1.611	1.381	1.841	1.381
fsr2011_1	1.266	2.532	2.279	2.786	1.013	1.773	0.506	1.013	4.052	1.519	4.305	1.519	1.266	0.253	2.026	0.76
fsr2012_1	1.469	0	1.836	2.57	1.469	0.367	0	0.367	3.304	0.367	2.937	4.039	1.101	0	2.57	0.734
fsr2012_2	1.919	3.563	1.096	2.467	3.563	0.274	2.193	0	4.933	0	1.37	3.289	0.274	1.37	1.644	1.37
fsr2013_1	2.25	2.25	2.25	1.125	2.25	1.5	1.5	0.375	2.25	2.25	2.25	0.75	0	2.25	2.25	2.626
fsr2013_1	1.627	1.627	2.278	1.627	1.302	1.302	4.231	1.302	4.231	0.976	3.58	0.976	3.905	2.604	1.953	5.858
fsr2014_1	1.619	2.024	0.405	2.024	0	0	4.453	2.833	4.048	1.214	1.619	0.405	1.619	1.619	0	4.048
fsr2014_2	3.18	2.12	0.795	0.265	1.59	3.445	3.71	2.915	1.59	3.975	1.06	0.53	0.53	5.035	1.06	3.445
fsr2015_1	1.532	1.149	0	0	0.766	0.766	3.065	3.448	0	2.682	1.149	0	1.532	1.149	1.149	1.149
fsr2015_2	0.749	1.497	0.374	0.749	1.123	0.749	4.492	4.492	2.246	2.995	0	0.749	2.995	1.123	1.497	0
fsr2016_1	0.692	4.154	0.346	1.039	1.731	1.385	2.077	2.769	5.193	1.039	0.346	0.692	1.385	2.423	1.731	1.385

Readability

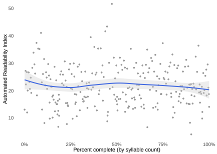
$$ARI = 4.71 \cdot \left(\frac{N_{char}}{N_{words}}\right) + .5 \cdot \left(\frac{N_{words}}{N_{sentences}}\right) - 21.43$$



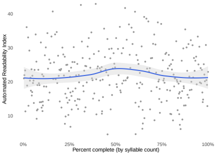
Readability FSR 2010



Readability FSR 2013-2



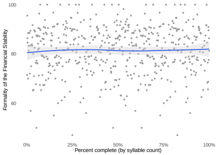
Readability FSR 2015-2



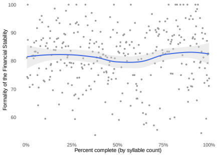
Readability FSR 2016-1

The Formality measure.

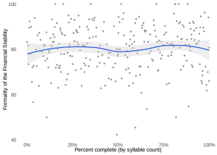
Formality of a statement is defined as the amount of expression that is immutable irrespective to changes of context.



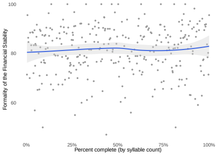
Formality of FSR 2010



Formality of FSR 2013-2



Formality of FSR 2015-2

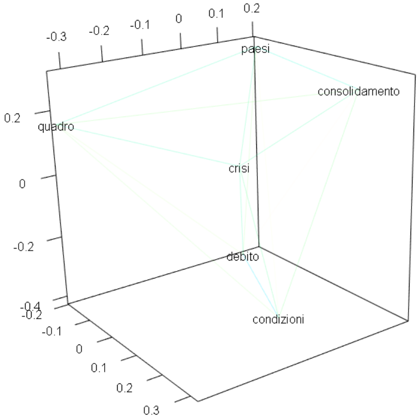


Formality of FS 2016-1

Measuring correlation between words and documents.

After completing the task of building a corpus of documents, it is possible to start the semantic analysis.
Latent Semantic Analysis (LSA) is a methodology for extracting and representing the contextual-usage of words (co-occurrence) for determining the similarity of meaning of sentences by analysis of large text corpora.
LSA methodology is well established and available in software such as Python, R and SAS.

LSA app: words most highly similar with 'crisi'



Semantic Orientation from PMI

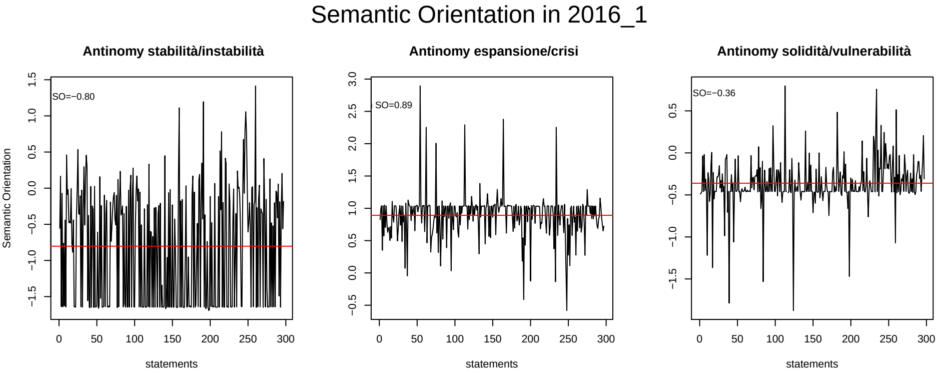
Given two events x and y, we have:

$$PMI(x; y) \equiv \log \frac{p(x, y)}{p(x) \cdot p(y)}$$

PMI measures the degree of statistical independence between x and y. The semantic orientation can be made more robust by employing an array of N antonyms:

$$SO(sent_i) \equiv \sum_{ant_j=1}^{N} \left(PMI(sent_i, ant_j[pos]) - PMI(sent_i, ant_j[neg])\right)$$

Semantic Orientation in 2016



Concluding Remarks

- We have built a Latent Semantic space to measure similarity among words and sentences in the*FSR*;
- we have evaluated some general characteristics of the *FSR* (readability and formality);
- we have extended the Semantic Orientation in Lucca(2011) by employing all the sentences as search units;
- we have shown a technique for evaluating the sentiment and polarity orientation caused by the text on the Web.

For Further Reading

F. Heylighen and J. Dewaele.
Variation on the Contextuality of Language: an Empirical Measure.
Foundation of Science, 2002.

R. Senter and E.A. Smith.
Automated Readability Index.
Aerospace Medical Research Laboratory, 2010.

D. Lucca and F. Trebbi.
Measuring Central Bank Communication: an Automated Approach with Applications to FOMC Statements.
NBER working paper, 2011.

Thank you for your attention.

Any questions?

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Issues on Big Data Governance – Big Data work in Central Banks, HR and IT issues[1]

Pedro Luis do Nascimento Silva,

President of ISI - International Statistical Institute

[1] This presentation was prepared for the meeting. 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 meeting.

International
Statistical
Institute

Issues on Big Data Governance – Big Data Work in Central Banks and HR and IT Issues

Pedro Luis do Nascimento Silva
President of ISI
Principal Researcher, ENCE, Brazil

A Perspective on Big Data

My own experience and environment is within a National Statistics Office (NSO), and not a Central Bank.

Nevertheless, I believe the two communities have much to learn from each other.

NSOs have a long history of handling governance of large and confidential datasets (such as census and administrative records).

The same is true with Banks in general.

A Perspective on Big Data

The main difference: NSOs have an obligation to disseminate statistics (and also microdata) to the public to the maximum extent possible without breaking the confidentiality ‘contract / promise’.

There’s no such pressure on most Central Banks.

How did NSOs cope: developing and adopting sophisticated methodology for confidentiality protection based on ‘risk – utility framework’.

A Perspective on Big Data

Big data is like **genetic medicine**: it is here to stay.

There are huge **promises of utility**, but also large **risks**.

However we will not give it up because of the potential utility.

The challenge is indeed how to manage the risks while maximizing the utility!

1 – Does the existing Data Governance in your CB/ institution cope with the Big Data needs?

No, if we consider some big data sources which are being created.

We are ‘cloud averse’ – few, if any, NSOs operate with ‘cloud computing’.

Reasons are simple: disclosure risks are seen as too high.

1 – Does the existing Data Governance in your CB/ institution cope with the Big Data needs?

13/03/2017 – Reuters

“Canadian statistics agency hacked, security flaw patched – officials said”

http://www.reuters.com/article/canada-cyber-idUSL2N1GQ1KH

Attack also affected Canada Revenue Agency!

Most NSOs will operate ‘secluded’ data environments for holding their data.

1 – Does the existing Data Governance in your CB/ institution cope with the Big Data needs?

Some big data sources are `public` and `cloud based`.

Thus we will need to use cloud computing at some stage.

For the confidential big data sources, some solution is needed that mitigates the risks while preserving utility ➔ methodological research!

2 – Which are the challenges posed by Big Data in terms of organization, human and IT resources, and data management responsibilities?

Need to mobilize skills which may not be currently available `in house`: IT, methodology, legal, etc.

Most likely need to upgrade IT resources & consider `cloud computing`.

But essential also to bring in the methodological expertise to operate the risk – utility approach.

3 – Will the use of Big Data encourage cooperation/partnerships between and within your institutions?

Yes! It seems unlikely that any organization can ‘do it alone’.

We will be forced to partner / cooperate if we are to fully benefit from the new opportunities created by the **Data Revolution**.

Many opportunities exist for us to learn from each other and cooperate.

4 – Is Big Data simply a new data source or does it call for changing the business model of your institutions?

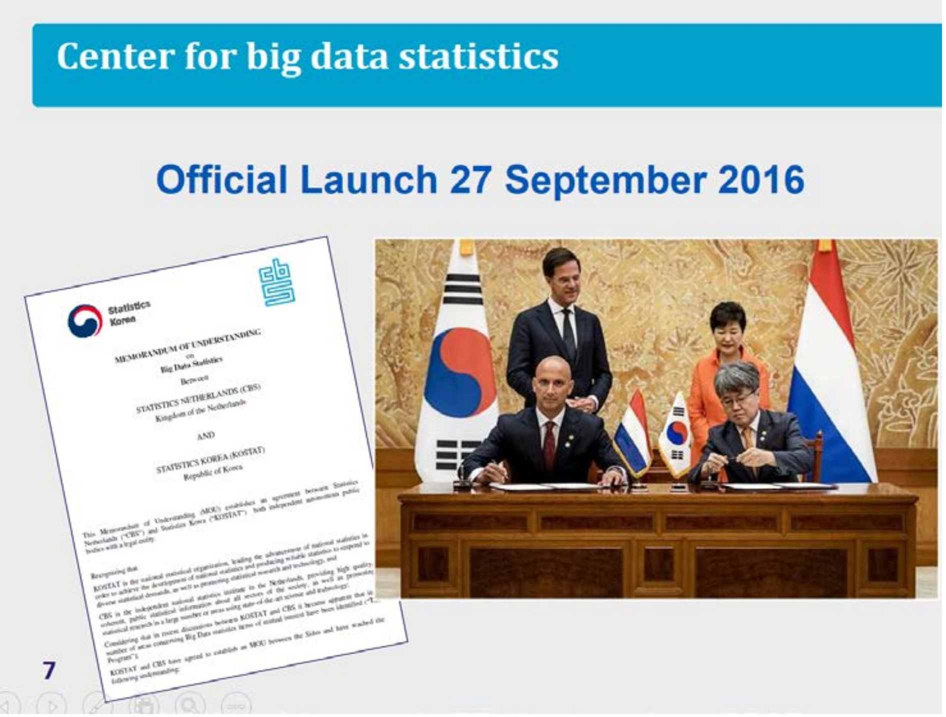
Not simple to answer – depends on the NSO or Central Bank current position.

ISI organized a side event to the 48th Session off the UNSC on the 8th March.

https://unstats.un.org/unsd/statcom/48th-session/side-events/20170308-1M-are-the-current-nso-business-models-still-relevant/

4 – Is Big Data simply a new data source or does it call for changing the business model of your institutions?

Netherlands CBS reported having created a ‘Center for big data statistics’ jointly with Korean NSO.



Center for big data statistics

Official Launch 27 September 2016

Statistics Korea

MEMORANDUM OF UNDERSTANDING
on
Big Data Statistics
Between
STATISTICS NETHERLANDS (CBS)
Kingdom of the Netherlands

AND

STATISTICS KOREA (KOSTAT)
Republic of Korea

This Memorandum of Understanding (MOU) establishes an agreement between Statistics Netherlands ("CBS") and Statistics Korea ("KOSTAT"), both independent autonomous public bodies with a legal entity.

Recognizing that:

KOSTAT is the national statistical organization, leading the advancement of national statistics in order to achieve the development of national statistics and producing reliable statistics to respond to diverse statistical demands, as well as promoting statistical research and technology, and

CBS is the independent national statistics institute in the Netherlands, providing high quality, coherent, public statistical information about all sectors of the society, as well as promoting statistical research in a large number of areas using state-of-the-art science and technology;

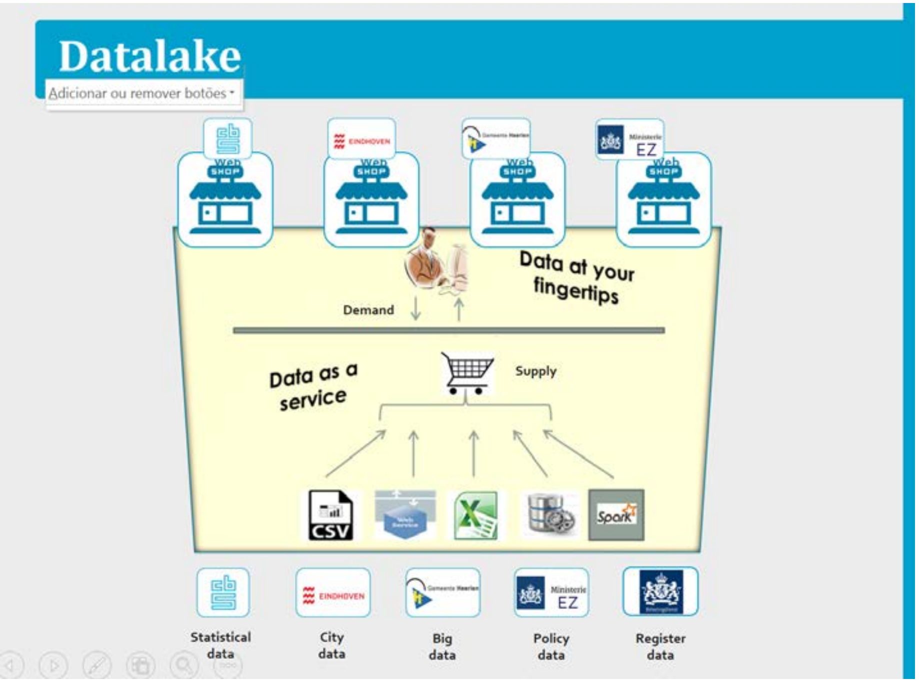
Considering that in recent discussions between KOSTAT and CBS it became apparent that in a number of areas concerning Big Data statistics items of mutual interest have been identified ("T Program");

KOSTAT and CBS have agreed to establish an MOU between the Sides and have reached the following understanding:

7

4 – Is Big Data simply a new data source or does it call for changing the business model of your institutions?

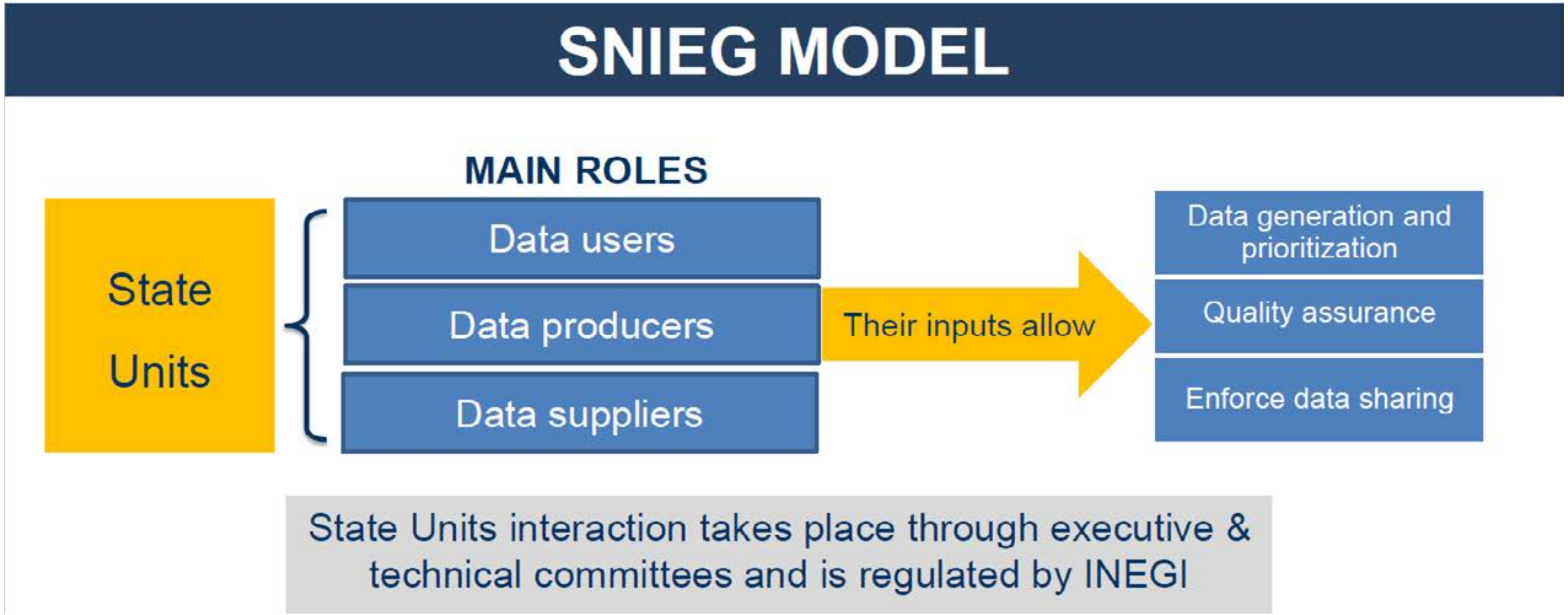
Netherlands CBS also reported having created a 'Data lake', where Big Data can be sourced and used.



4 – Is Big Data simply a new data source or does it call for changing the business model of your institutions?

Slovenian NSO called for ‘continuous adaptation’.

Mexican NSO relies on coordination role.



SNIEG MODEL

MAIN ROLES

State Units

- Data users
- Data producers
- Data suppliers

Their inputs allow

- Data generation and prioritization
- Quality assurance
- Enforce data sharing

State Units interaction takes place through executive & technical committees and is regulated by INEGI

Summarizing

Leadership.

Adaptation.

Cooperation.

Innovation.

Persistence.

ISI
Statistical Science for a Better World

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Issues on Big Data Governance – Big Data work in Central Banks, HR and IT issues[1]

Anne Leslie-Bini,

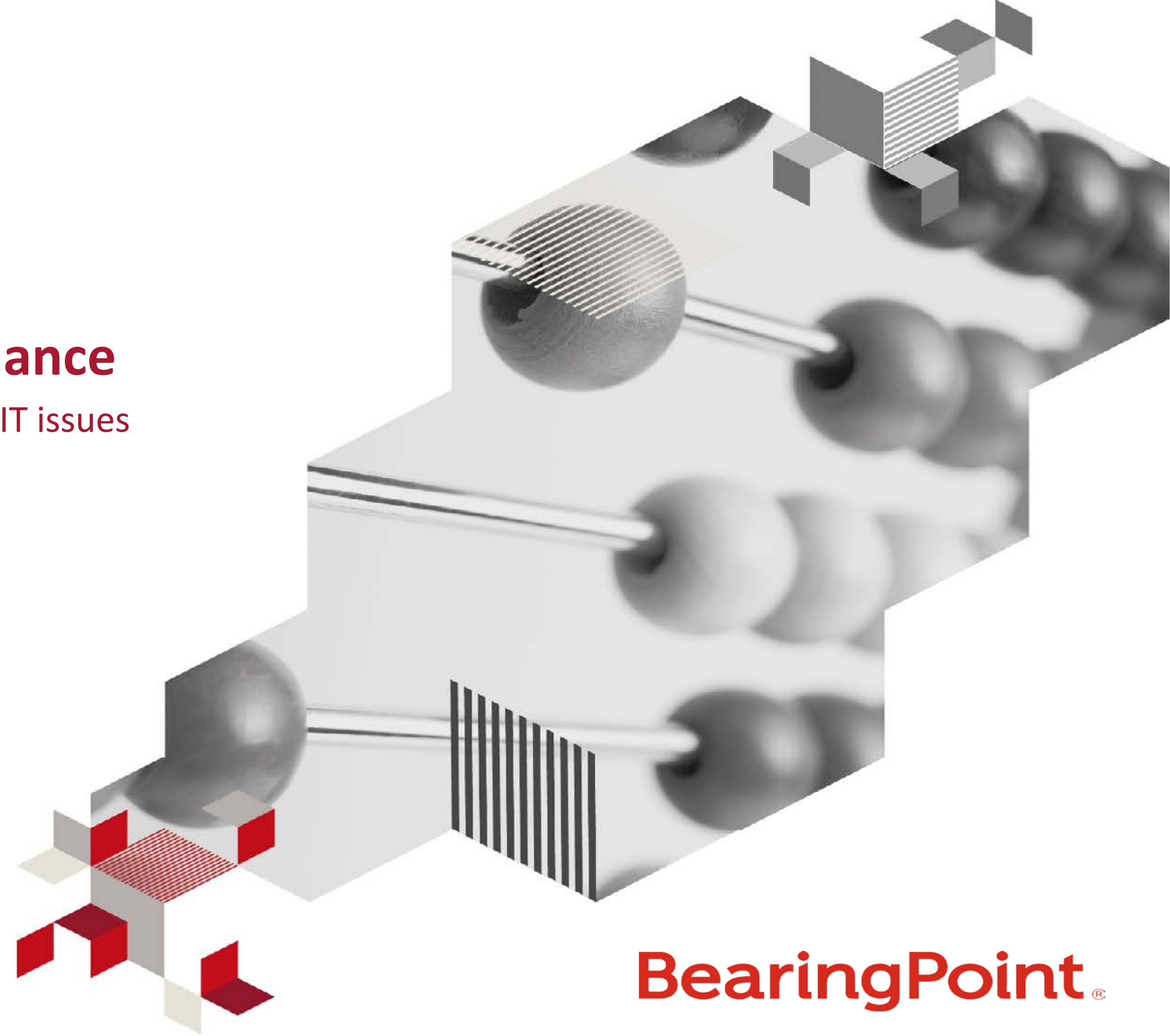
Director, Financial Services, BearingPoint / Central Banking

[1] This presentation was prepared for the meeting. 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 meeting.

Issues on Big Data Governance

Big Data work in Central Banks, HR and IT issues

Anne Leslie-Bini



A venerable institution with centuries of history…

A multi-faceted mission…

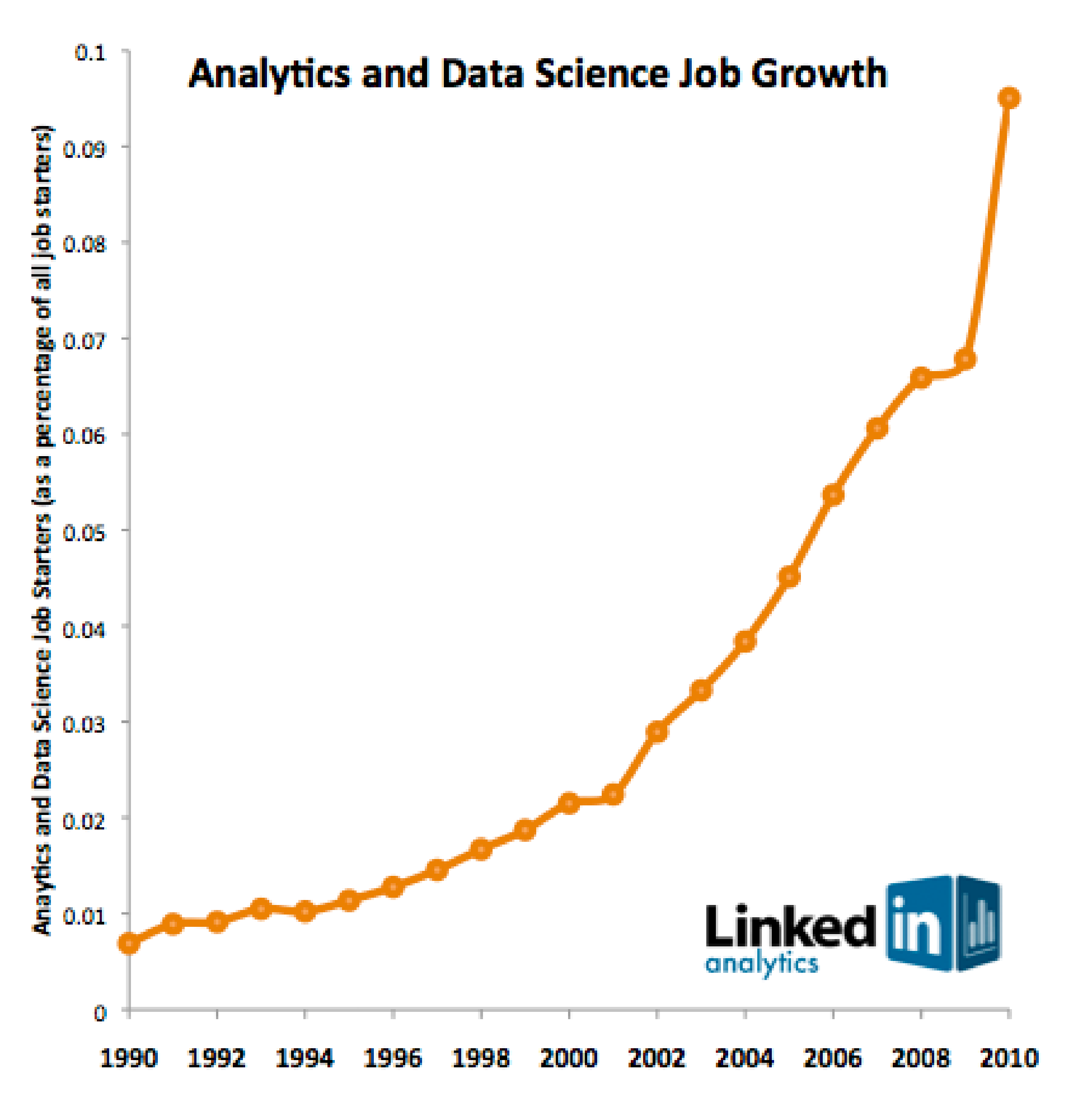
Living with legacy architecture…

And the challenges of reconciling old and new…

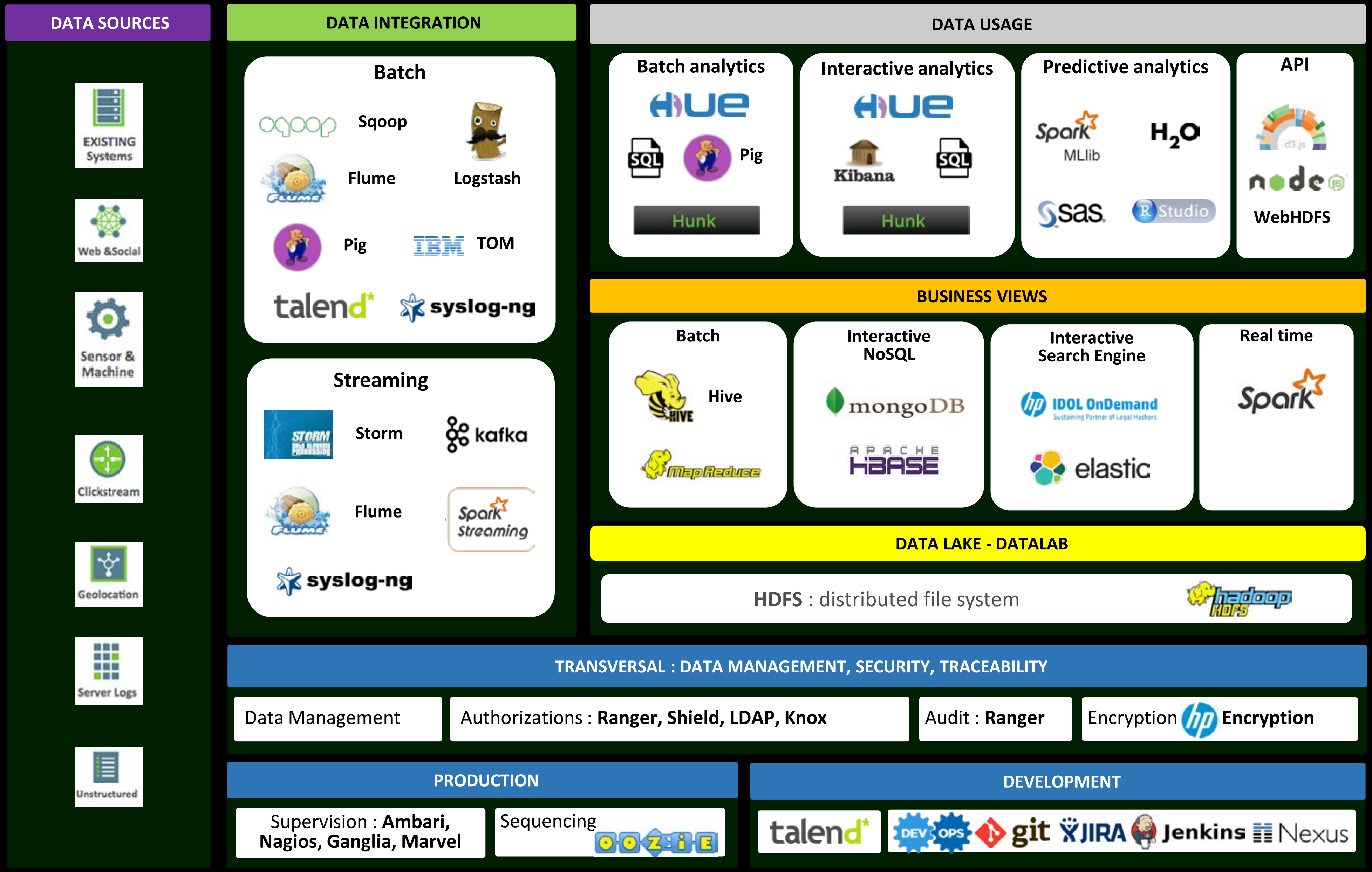
Striving for the best of both worlds…

LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE
© marketoonist.com

The war for talent…


Analytics and Data Science Job Growth
Anaytics and Data Science Job Starters (as a percentage of all job starters)
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
Linked in
analytics

USING IT
IS THE HARDEST PART.
BIG
DATA



DATA SOURCES

EXISTING Systems

Web &Social

Sensor & Machine

Clickstream

Geolocation

Server Logs

Unstructured

DATA INTEGRATION

Batch

Sqoop

Flume

Logstash

Pig

IBM TOM

talend

syslog-ng

Streaming

Storm

kafka

Flume

Spark Streaming

syslog-ng

DATA USAGE

Batch analytics

HUE

SQL

Pig

Hunk

Interactive analytics

HUE

Kibana

SQL

Hunk

Predictive analytics

Spark MLlib

H2O

sas

R Studio

API

d3.js

node JS

WebHDFS

BUSINESS VIEWS

Batch

Hive

MapReduce

Interactive NoSQL

mongoDB

APACHE HBASE

Interactive Search Engine

hp IDOL OnDemand

elastic

Real time

Spark

DATA LAKE - DATALAB

HDFS : distributed file system

hadoop HDFS

TRANSVERSAL : DATA MANAGEMENT, SECURITY, TRACEABILITY

Data Management

Authorizations : **Ranger, Shield, LDAP, Knox**

Audit : **Ranger**

Encryption hp **Encryption**

PRODUCTION

Supervision : **Ambari, Nagios, Ganglia, Marvel**

Sequencing OOZIE

DEVELOPMENT

talend

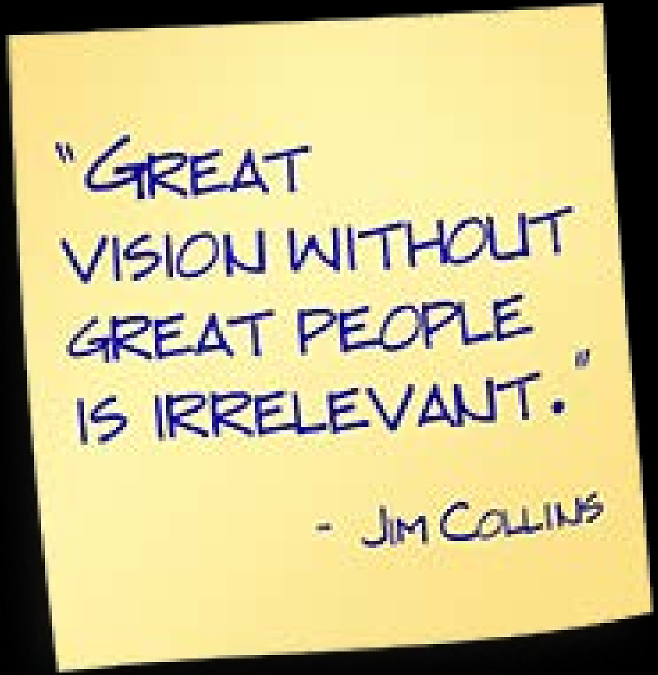
DEV OPS

git

JIRA

Jenkins

Nexus



"GREAT
VISION WITHOUT
GREAT PEOPLE
IS IRRELEVANT."

- JIM COLLINS

pur•pose

/ˈpərpəs/

Noun

The reason for which something is done or created
or for which something exists.

BearingPoint®

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Issues on Big Data Governance – Big Data work in Central Banks, HR and IT issues[1]

Rhys Mendes,
Managing Director / Chief, Economic and Financial Research, Bank of Canada

[1] This presentation was prepared for the meeting. 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 meeting.

BANK OF CANADA
BANQUE DU CANADA

Panel Discussion at IFC-BI Seminar on “Big Data”

21 March 2017

Rhys Mendes
Managing Director/Chief, Economic and Financial Research
Bank of Canada

Agenda for Discussion

- Big Data’s impact on:
  - Data Governance
  - Organizational Structure, Human Resources, Information Technology
  - Cooperation/Partnerships
  - Business Models

Data Governance

- Traditionally, CBs have been data users
- Big Data is increasingly making us data collectors/generators
- Data governance needs to evolve with this shift
- Illustrative Example: Web scraping retail price data

Data Governance

- The example raises several questions:
  - Legal/ethical questions
  - Who should have responsibility for authorizing collection?
  - Need for legislative change?
  - Is greater coordination with statistical agency a better route?
  - How should processes change when going from experimental phase to production phase?

Organizational Structure

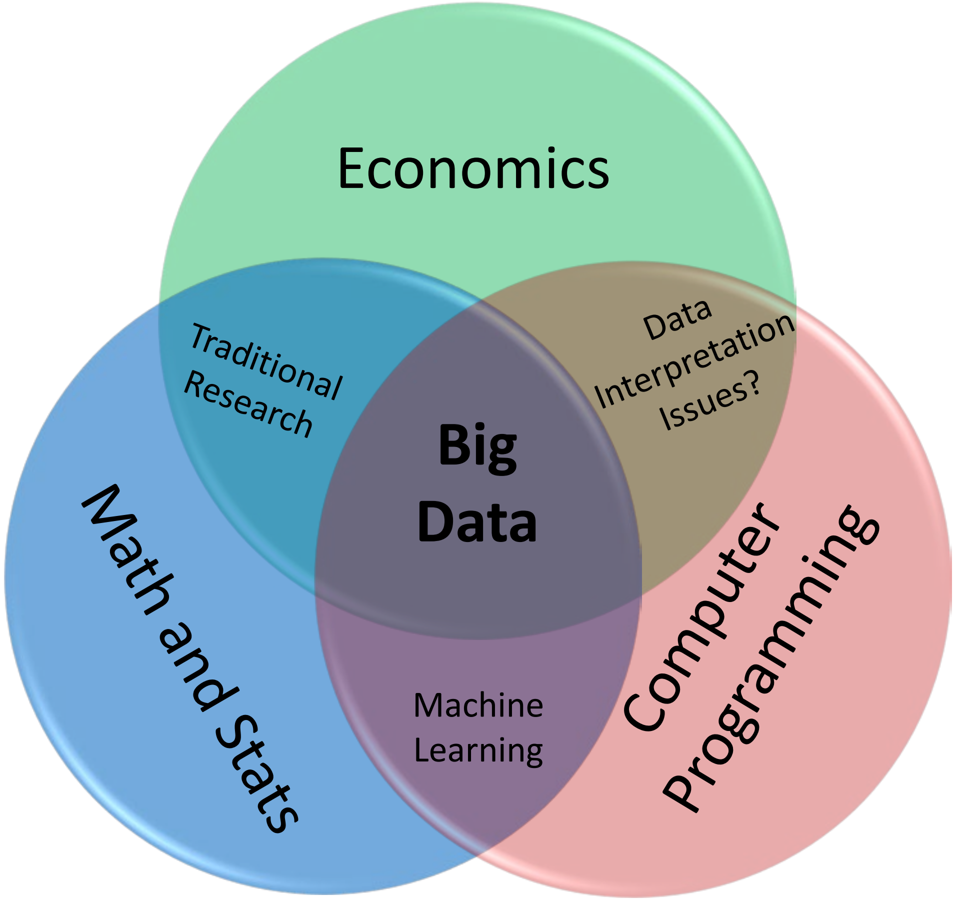
- Where should big data resources reside?

- Many departments pursuing big data projects.
  - Should each build its own technical expertise?
  - Should the technical expertise reside with some central office to maximize cross-pollination?

- Some departments generate operational data that others want to use. How to share costs?

BANK OF CANADA
BANQUE DU CANADA

Human Resources:
Challenges Created by Multidisciplinary Nature of Big Data Analytics



Human Resources

- Hiring challenges depend on existing HR structure
  - Job profiles
  - Pay lines
- Impact of organizational structure on hiring/retention:
  - Economists managing data scientists?
  - Career paths for data scientists, etc.
- Critical mass for hiring

Information Technology

- Critical mass
- Flexibility of IT funding
- Ability to experiment

Cooperation/Partnerships

- Right now lots of learning from one another
  - Very valuable as we all ramp up big data capacity

- Not clear that increase in cooperation will persist as this becomes a more mature area of work

Business Models

- Won’t fundamentally change what central banks do

- But has potential to eventually change how certain areas work. Examples:
  - Short-term forecasting
  - IT Security

Conclusions

- Still in a learning phase
- Cooperation/partnerships will accelerate learning
- Experimentation is key

BANK OF CANADA
BANQUE DU CANADA

Thank you