"Science Stays True Here"
Journal of Mathematics and Statistical Science (ISSN 2411-2518), Vol.6, 9-17 | Science Signpost Publishing

The Role of Big Data and Surveys in Measuring and Predicting Inflation

Bruno Tissot
Head of Statistics and Research Support,
Bank for International Settlements (BIS), Basel, Switzerland,
and Head of the Secretariat of the Irving Fisher Committee on Central Bank Statistics (IFC),
Bruno.Tissot@bis.org

Abstract

Monitoring and forecasting price developments is an important issue for public authorities. Measuring inflation requires significant resources, and substantial methodological work has been developed over the past decades to support this task. The recent emergence of Big Data can provide many opportunities in this context, for instance to produce timelier indicators, enhance the collection of specific types of prices, and take due consideration of economic agents' expectations. However, there are important challenges when using big data-type of information, and traditional statistical surveys have continued to prove their usefulness. This suggests that measuring and forecasting inflation should continue to benefit from drawing on multiple, complementary approaches.

Keywords: Prices, Public policy, Internet, Inflation expectations, Nowcasting.

1. Introduction

Measuring and predicting the evolution of prices is key in a market economy, and it has therefore been traditionally an important objective for authorities. Indeed, past centuries were characterised by various episodes of strong money creation, leading to peaks in inflation with often important social and political consequences (see Graph 1 for an historical perspective on European inflation). Episodes of price deflation have been also particularly disruptive, and the Great Depression of the 1930s underlined the need for **properly measuring and anticipating both the real economic activity and the evolution of prices.** This was clearly a key factor driving the subsequent development of the Systems of the National Accounts (SNA; European Commission et al (2009)) framework. Today, most advanced economies and a growing number of emerging economies have complete and reliable statistics to measure consumer price inflation (CPI) based on detailed international guidance (ILO et al (2004)). Such statistics are often based on statistical, census-type surveys, through which statisticians collect the prices observed for a specific list of products in selected retail points and aggregated with proper weights to reflect the composition of households' consumption basket.

Nevertheless, **important limitations** still hinder a timely and comprehensive provision of reliable inflation indicators across the globe. Five key issues are worth highlighting from this perspective.

First, a **robust statistical infrastructure** is required to produce inflation data on a regular basis; this calls for sufficient staff resources, adequate statistical skills, effective IT support, the set-up of specific processes to capture prices observed in various market segments, etc.

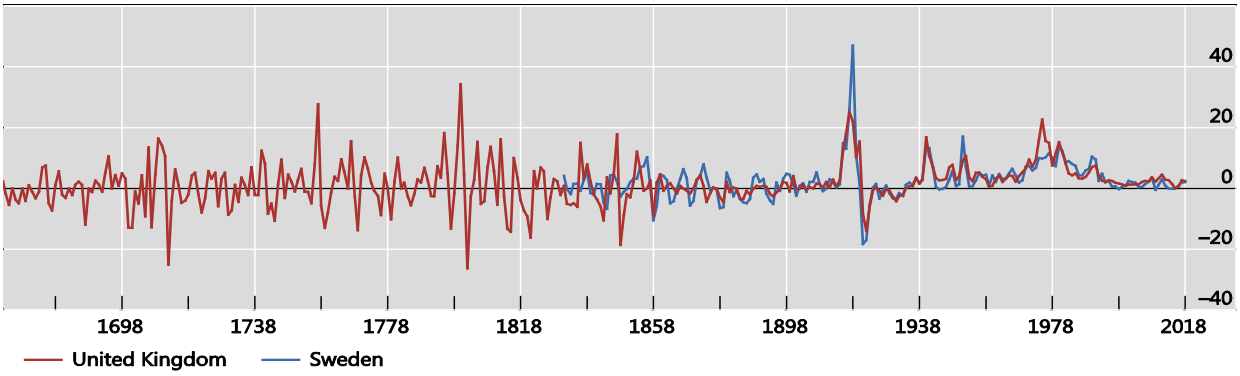
Second, the **impact of innovation and digitalisation** is posing practical and theoretical difficulties as regards the definition and the measurement of inflation (see Reinsdorf and P Schreyer (2019)). In particular,

it is increasingly difficult to capture instant prices offered during limited periods of time and/or for specific groups of buyers. Moreover, correctly measuring inflation requires the ability to capture, and correct for, quality changes. Yet, increased innovation has reduced the ability to observe the repeated sales of the same products, since their intrinsic characteristics are constantly evolving, hence making difficult the observation of prices for constant quality goods. The general solution for statisticians is to apply the so-called hedonic method.[1] But this requires capturing a wealth of product characteristics that have to be analysed using econometric techniques – making the data collection and compilation process more complex.

The history of inflation – selected European countries

Graph 1

Estimated CPI inflation, year-on-year; in per cent



Source: BIS.

Third, inflation is a **multiform phenomenon** that varies across sectors and economic agents, leading to a multiplicity of inflation indicators. In general, one will refer to inflation as the evolution in the prices of goods and services consumed by an average household consumer, ie the CPI index. However, inflation will be perceived differently by other types of agents, for instance by producers, exporters or importers, or even within these groups themselves – say, between wealthy and poor consumers. Moreover, one could be more interested in the evolution of some sub-components of inflation – for instance in "core" inflation, which corrects for the volatility caused by movements in food and energy prices to reveal underlying inflation trends. The situation in the United States illustrates this variety of measures in the case of consumer inflation: the Bureau of Labour Statistics (BLS) produces monthly data on changes in the prices paid by urban consumers for a representative basket of goods and services, with separate indexes for two groups: all urban consumers (CPI-U), which is the index most often reported; and the urban wage earners and clerical workers (CPI-W), which is used more for wage negotiations (BLS (2019)). Several additional indicators are compiled, such as: the BLS chained CPI (designed as a "cost-of-living" index); the Personal Consumption Expenditures (PCE) price index developed by the Bureau of Economic Analysis in consistency with the SNA framework; the core PCE; etc.

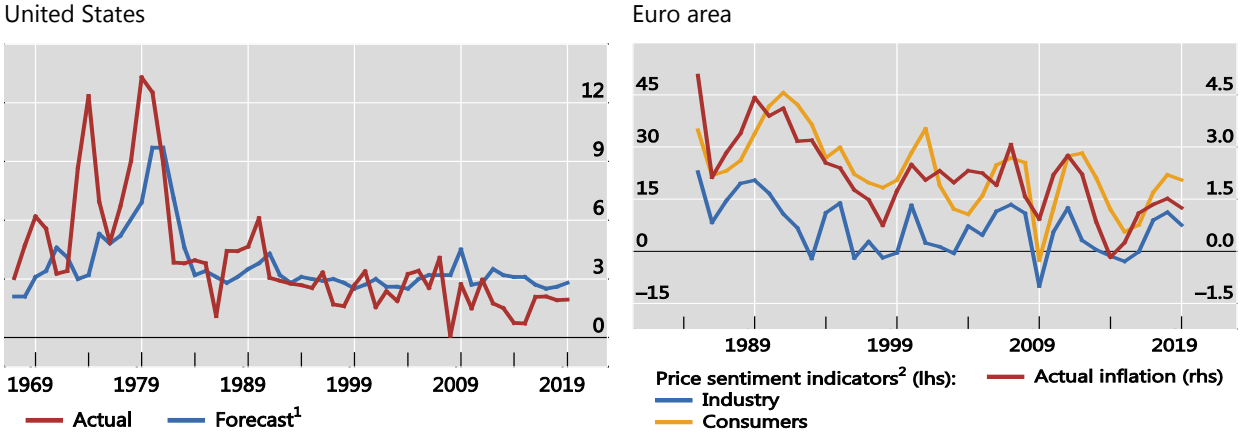
Fourth, an important factor driving price and wage developments relates to **expectations**. To address this point, many countries have set up a number of surveys to help predict short-term developments in inflation – eg in the context of the work organised since 1953 under the umbrella of the Centre for

[1] Defined as a "regression technique used to estimate the prices of qualities or models that are not available on the market in particular periods, but whose prices in those periods are needed in order to be able to construct price relatives"; cf *OECD Glossary of Statistical Term*, available at stats.oecd.org/glossary/index.htm.

International Research on Economic Tendency Surveys (CIRET), a global forum for institutions conducting business and/or consumer surveys. These comprise the well-known consumer survey conducted in the United States by the University of Michigan (Curtin (1996)), the business and consumer surveys conducted under the aegis of the European Commission in Europe (European Commission (2019)), and the consensus forecasts collected from various economists and forecasters – cf for instance the short- and long-term inflation forecasts of the US Survey of Professional Forecasters published by the Federal Reserve Bank of Philadelphia (Croushore (1993)) as well as the similar survey set up by the ECB (Garcia (2003)). Yet how useful such measures of inflation expectations can be to help predict actual inflation numbers is an open issue, as suggested by Graph 2 for the Euro Area and the United States.

Year-on-year inflation: expectations and actual observations — Graph 2

In per cent



[1] Inflation rate changes expected one year earlier. [2] Annual averages, balances (%) of the price expectations for the next 12 months questions; not seasonally adjusted.

Sources: European Commission; University of Michigan; national data; BIS calculations.

Fifth, inflation is a general concept that is not limited to traded goods and services. One important area relates to **asset prices**, which can play a powerful role as was seen during the Great Financial Crisis of 2007-09, with the strong impact of the ups and downs observed in housing prices (Tissot (2014)). In addition, developments in financial markets can also be useful to measure inflation expectations based on market-based indicators of asset prices (eg index-linked bonds).

2. Opportunities provided by Big Data

Given the challenges mentioned above, what are the opportunities offered by big data – described by some as the new oil of the 21st century (The Economist (2017))? The main sources of big data are social networks, traditional business systems, and the internet of things,[2] and **four types of data sets** appear of particular importance in the economic and financial area: internet-based indicators, commercial data sets,

[2] Following the work conducted under the aegis of the United Nations (see Meeting of the Expert Group on International Statistical Classifications (2015)).

financial market indicators and administrative records (IFC (2017)).[3] These data are often available in an "organic" way, unlike statistical surveys and censuses: the reason is that they are usually not collected ("designed") for a specific statistical purpose, being the by-product of other activities (Groves (2011)). Hence, there is a clear interest to use them, considering that the cost of launching traditional statistical surveys can be significant. Moreover, the new type of information they provide can help addressing some of the challenges faced when measuring and predicting inflation, especially the five main ones mentioned above.

As regards the first issue related to the need for a robust statistical infrastructure, big data can **facilitate the work of public authorities** compiling inflation measures. It can certainly be an innovative source for the current production of official statistics, offering access to a wider set of prices (eg prices recorded online, credit card operations); it can also facilitate statistical compilation work when the data are easier to collect compared to traditional approaches. Moreover, big data-related IT innovation can help to better process and use the information collected. Indeed, data volumes have surged hand in hand with the development of specific techniques for their analysis, with the emergence of "big data analytics" – broadly referring to the general analysis of large data sets – and "artificial intelligence" (AI); cf IFC (2019). Modern computing tools can now be used to collect data, correct them, improve coverage (eg web-scraping), process textual information (text-mining), match different data sources (eg fuzzy-matching), extract relevant information (eg machine learning) and communicate or display pertinent indicators (eg interactive dashboards). All these elements can help to address the resource issues posed by the compilation of official price statistics, especially in developing economies where statistical systems are still in infancy and staff skills are limited. One example is the Billion Prices Project at the Massachusetts Institute of Technology (MIT), which allows inflation indices to be constructed for countries that lack an official and/or comprehensive index and that can be used for enhancing international comparisons of price indexes in multiple countries and for dealing with measurement biases and distortions in international relative prices (see www.thebillionpricesproject.com and Cavallo and Rigobon (2016)). Similarly, a number of central banks in emerging market economies have compiled quick price estimates for selected goods and properties, by directly scraping the information displayed on the web, instead of setting up specific surveys that can be quite time- and resource-intensive. One notable situation relates to those developing economies as big as India, where collecting internet-based data is seen as a potentially useful alternative to the organization of large surveys that would have to cover millions of reporters. Yet, as indeed noted by Hill (2018)[4] in the case of the United States, and in contrast to what is observed in the research and academic community, the use of big data in more mature statistical systems has been relatively incremental and limited. It is often targeted at methodological improvements (for instance quality adjustment) and at reducing reporting lags.

As regards the measurement **challenges posed by rapid innovation** (the second main issue highlighted in section 1), the high velocity of big data sources can be particularly useful when prices change rapidly. For instance, direct web-scraping allows extracting almost in real time retailers' prices from online advertisements. This can support a timelier publication of official data, by bridging the time lags before official statistics become available – ie through the compilation of advance estimates or "nowcasting

[3] It is important to note that the category of internet-based indicators is not necessarily the most useful; see for instance the growing importance for public statisticians of the large datasets derived from administrative records (Bean (2016)).

[4] "… while nearly 15% of the price quotes in the Consumer Price Index are now collected online (…) the size of the CPI sample has not increased to reflect the lower cost of online data collection".

exercises". In addition to the lag issue, the information provided by the wide range of web and electronic devices is often available with a higher frequency; changes in price developments can thus be tracked more promptly, compared to official CPI numbers that are usually available on a monthly basis. This can be particularly useful when analysing early warning indicators and assessing turning points. Indeed, an important objective of the Billion Prices Project mentioned above is to provide advance information on inflation in a large number of countries, including advanced economies, and with greater frequency.

Third, the high granularity of big data sets can support to measure of **various dimensions of inflation,** allowing for a better understanding of the dispersion of prices aggregates – eg across markets and/or locations, a type of distribution information that is generally missing in the SNA framework (Tissot (2018)). As regards market differentiation, web-scraped prices can be useful for measuring inflation in very volatile sectors, such as fresh food prices, allowing for a better measure of some specific components of the CPI. As regards geographical factors, large and granular data sets collected from commercial advertisements can help to capture local patterns with sufficient precision, say for instance to measure rents or property prices depending on zip codes or even street names.

Fourth, various big data sources such as numbers on internet search queries (eg Google Trends) and "soft" indicators computed from digitalised textual information (eg displayed by social media posts like Twitter) can provide interesting insights on **economic agents' sentiment and expectations** (Wibisono and Zulen (2019)). Traditional statistical surveys can also offer this kind of information, but they typically focus on specific items, eg firms' production expectations and consumer sentiment. In contrast, internet-based sources allow a wider range of indicators to be used (Rigobon (2018)). In addition, they can be less intrusive than face-to-face statistical surveys, and may therefore better reflect true behaviours and expectations.

Fifth, new big data sources appear of increasing interest for **measuring the wide range of asset prices** that are not easily covered by traditional surveys because of the lack of statistics available and/or methodological clarity. Cases in point relate to residential and commercial property markets, which are often lacking reliable statistics, while alternative sources can be easily found using big data (eg advertisements from property websites and newspapers). In addition, these markets are characterised by a low and infrequent number of transactions (compared to stocks) and by significant heterogeneity across tangible assets, making the compilation of quality-adjusted house price indices difficult. These challenges can be overcome by capturing the various characteristics of the properties displayed in web-based advertisements and the application of hedonic methods. Moreover, the information collected, being very granular, can more easily be matched with other datasets, say census survey-based information for similar homes and tax registries. Furthermore, the new type of information collected can provide additional insights that are not covered by "traditional" statistics, for instance to analyse housing market liquidity and tightness (eg by assessing demand intensity through the number of clicks on specific ads), discounting practices (eg by comparing asking and transactions prices, which can differ markedly for instance during turning points), and detailed geographical factors – see for instance Loberto et al (2018).

3. Challenges

Despite the various opportunities allowed by big data sources, there are important challenges in using this information when measuring and forecasting prices. First, there are **practical difficulties in collecting the data**. This challenge can be reinforced by the large variety of big data formats, especially when the

information collected is not well structured. Apart from the technical aspects (eg proper IT equipment, access rights etc), a key issue is data quality. For instance, references displayed on the web can be incorrect, or may not really reflect true transaction prices (eg in case customers benefit from discounts for other services); and the characteristics of the products may not be standardised properly. As a result, statisticians have to deal with duplicated information, since the same product may be sold in different places but be identified with different characteristics – for instance, a common feature for property markets is that several (different) advertisements can be associated with the same dwelling. Alternatively, a product may still be displayed on a website even though it is no more available for sale, hence the risk of measuring outdated prices. Dealing with these challenges requires significant work when cleaning and processing the data. In addition, the usefulness of the information collected is limited if the data sources and/or their market coverage change over time, and of course if its access is hindered by privacy laws and/or copyright issues.

There are also important **methodological limitations**. First, estimating price indices requires defining a basket of goods that are representative of the spending of the economic agents considered. As regards CPI, for instance, a significant part of the consumption basket is related to goods that are either not traded (eg self-consumption of housing services by homeowners) or that have an administrative nature and are therefore not quoted on the internet. So compiling a CPI indice using only web-based information will not be fully representative; one way to go is to complement this approach with other type of (non web-based) information.

Even if one only focusses on the part of the consumption basket that can indeed be traded on the internet, another concern is that big data samples are often far from representative, so the veracity of the information collected may not be as good as it seems. Certainly, big data sets often cover entire populations, so by construction there is little sampling error to correct for, unlike with traditional statistical surveys. But a common public misperception is that, because big data sets are extremely large, they are automatically representative of the true population of interest. Yet this is not guaranteed, and in fact the composition bias can be quite significant, in particular as compared with much smaller traditional probabilistic samples (Meng (2014)). For example, when measuring prices online, one must realise that not all transactions are conducted on the internet. The measurement bias can be problematic if online prices are significantly different from the prices observed in physical stores, or if the products consumers buy online are different to those they buy offline.

Lastly, there are also **challenges when using big data sources**. Ideally, statistics based on big data should have the same quality of standards and frameworks that govern official statistics,[5] such as transparency of sources, methods and processes. But in practice they can be collected in an opaque way, arguably not in line with these recognised principles. "Misusing" such information could thus raise ethical, reputational as well as efficiency issues. In particular, if the confidentiality of the data analysed is not carefully protected, this could undermine public confidence, in turn calling into question the authorities' competence in collecting, processing and disseminating information derived from big data (Tissot (2019)).

Using big data for anticipating future developments is also challenging. While related applications such as machine learning algorithms can excel in terms of predictive performance, they can lend themselves more

[5] Cf the *Fundamental Principles of Official Statistics* adopted by the United Nations in 2014, available at unstats.un.org/unsd/dnss/gp/FP-New-E.pdf. For an overview of the challenges posed by using big data for official statistics more generally, see Hammer et al (2017).

to explaining what is happening rather then why. Indeed big data analytics rely frequently on correlation analysis, which can reflect coincidence as well as causality patterns. As such, they may be exposed to public criticism when insights gained in this way are used to produce official statistics and forecasts and/or justify policy decisions.

4. Conclusion

Big data sources and techniques can provide **new and useful insights that complement "traditional" data sets** and facilitate the compilation of price statistics as well as forecasting exercises. They can also provide new sorts of signals that can be useful especially for policy makers, for instance for analysing market liquidity as well as geographical patterns.

Central banks' experiences have underlined the **benefits of having an encompassing framework for making the most of all the data available** to enhance their regular monitoring of economic activity in general and of price developments in particular. These approaches draw on big data analytics to effectively summarise the information contained in large data sets through a small number of common factors and the capturing of time-varying effects to incorporate continuously the new information becoming available (Bok and al (2017)). But, interestingly, they primarily rely on "traditional" survey statistics, while the use of web-based indicators has remained limited, suggesting that they may work less well in nowcasting/forecasting exercises compared to "traditional" statistics and confidence surveys. In the United States, for instance, the price nowcasting exercises conducted at the Federal Reserve Bank of Cleveland (Knotek and Zaman (2014)) mainly rely on past observations of inflation (eg monthly developments in CPI, core CPI, CPI for food, etc) complemented by higher-frequency information on oil prices (eg retail gasoline prices and daily crude prices); this reflects the fact that core inflation tends to be relatively stable in the short-run and that most of the volatility observed in headline inflation is driven by changes in food and energy prices. The parallel approach set up at the ECB relies on similar mechanisms when trying to predict short-term developments in inflation (Modugno (2011): the high-frequency data retained for nowcasting inflation are mainly related to the daily prices of raw materials including energy. Such mixed approaches (eg using big data analytics to digest the large amount of incoming information while still having as input relatively "traditional" statistics) are reported to perform relatively well compared to, for instance, consensus forecasts. From a different perspective, the GDP nowcasting exercises conducted by the Federal Reserve Bank of New York have highlighted the important information content of past inflation developments when estimating economic activity in advance (Bok and al (2017)).

Whether such approaches will allow for the integration of more "big data-type" information when measuring and predicting inflation remains to be seen. Certainly, there are important challenges to be considered, related in particular to: the methodological choices to be retained; the need to clean the vast of amount of the new information available, not least to deal with duplicates, outliers and other quality issues; the difficulty to measure real transaction prices and avoid capturing obsolete information (since internet-based prices can remain on the web for a long time); and the issues posed by matching different datasets at a quite granular level. Looking forward, key is to **make more, sometimes untested information available**, so that researchers and policy analysers can "play with the data".

References

[1]. Bean, C (2016): *Independent review of UK economic statistics*, March.

[2]. Bok, B, D Caratelli, D Giannone, A Sbordone and A Tambalotti (2017): "Macroeconomic nowcasting and forecasting with Big Data", *Federal Reserve Bank of New York Staff Reports*, no 830, November.

[3]. Bureau of Labor Statistics (BLS) (2019): *Handbook of Method*, Chapter 17. The Consumer Price Index, April.

[4]. Cavallo, A and R Rigobon (2016): "The Billion Prices Project: Using Online Prices for Measurement and Research", *Journal of Economic Perspectives*, Spring 2016, vol 30, no2, pp 151-78.

[5]. Croushore, D (1993): "Introducing: The Survey of Professional Forecasters", *Federal Reserve Bank of Philadelphia Business Review*, November/December.

[6]. Curtin R (1996): *Procedure to estimate price expectations*, University of Michigan Surveys of Consumers, January.

[7]. European Commission (2019): *The joint harmonised EU programme of business and consumer surveys - User Guide*, Directorate-General for Economic and Financial Affairs, January.

[8]. European Commission, IMF, OECD, United Nations and World Bank (2009): *System of National Accounts 2008*.

[9]. Garcia, J A (2003): "An introduction to the ECB's survey of professional forecasters", *ECB Occasional Paper Series*, no 8, September.

[10]. Groves, R (2011): "Designed data and organic data", in the Director's Blog of the US Census Bureau.

[11]. Hammer, C, D Kostroch, G Quirós and staff of the IMF Statistics Department (STA) Internal Group (2017): "Big data: potential, challenges, and statistical implications", *IMF Staff Discussion Notes*, no 17/06, September.

[12]. Hill, S (2018): "The Big Data Revolution in Economic Statistics: Waiting for Godot... and Government Funding", *Goldman Sachs US Economics Analyst*, 6 May.

[13]. ILO, IMF, OECD, UNECE, Eurostat and The World Bank (2004): *Consumer price index manual: Theory and practice*, Geneva, International Labour Office.

[14]. Irving Fisher Committee on Central Bank Statistics (IFC) (2017): "Big data", *IFC Bulletin*, no 44, September.

[15]. ——— (2019): "The use of big data analytics and artificial intelligence in central banking", *IFC Bulletin*, no 50, May.

[16]. Knotek E and S Zaman (2014): "Nowcasting US headline and core inflation", *Federal Reserve Bank of Cleveland Working Paper*, no 14-03, May.

[17]. Loberto M, A Luciani and M Pangallo (2018): "The potential of big housing data: an application to the Italian real-estate market", *Temi di Discussione*, no 1171, April.

[18]. Meeting of the Expert Group on International Statistical Classifications (2015): *Classification of Types of Big Data*, United Nations Department of Economic and Social Affairs, ESA/STAT/AC.289/26, May.

[19]. Meng, X (2014): "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, pp 537–62.

[20]. Modugno, M (2011): "Nowcasting inflation using high frequency data", *ECB Working Paper Series*, no 1324, April.

[21]. Reinsdorf M and P Schreyer (2019): "Measuring consumer inflation in a digital economy", *OECD Statistics and Data Directorate Working Paper*, No 101, February.

[22]. Rigobon, R (2018): "Promise: measuring from inflation to discrimination", presentation given at the workshop on "*Big data for central bank policies*", Bank Indonesia, Bali, 23–25 July.

[23]. The Economist (2017): "The world's most valuable resource is no longer oil, but data", 6 May edition.

[24]. Tissot, B (2014): "Monitoring house prices from a financial stability perspective - the BIS experience", International Statistical Institute Regional Statistics Conference, November.

[25]. ——— (2018): "Providing comparable information to assess global financial stability risks", *Eurostat Statistical Reports*, KS-FT-18-001, January.

[26]. ——— (2019): "Making the most of big data for financial stability purposes", in S Strydom and M Strydom (eds), *Big Data Governance and Perspectives in Knowledge Management*, IGI Global, pp 1–24.

[27]. Wibisono O and A Zulen (2019): "Measuring stakeholders' expectations for the central bank's policy rate", *IFC Bulletin*, no 50, May.