Evaluation of Nowcasting / Flash Estimation based on a Big Set of Indicators

Dario Buono[1], George Kapetanios[2], Massimiliano Marcellino[3], Gian Luigi Mazzi[4] and Fotis Papailias[5]

Paper prepared for the 16th Conference of IAOS
OECD Headquarters, Paris, France, 19-21 September 2018

Session 1.A., Day 1, 19/09, 11:00: Use of Big Data for compiling statistics

[1] Eurostat, European Commission. E-mail: dario.buono@ec.europa.eu
[2] King's College London. E-mail: george.kapetanios@kcl.ac.uk
[3] Bocconi University. E-mail: massimiliano.marcellino@unibocconi.it
[4] GOPA. E-mail: gianluigi.mazzi@gopa.lu
[5] King's College London. E-mail: fotis.papailias@kcl.ac.uk

Dario Buono
dario.buono@ec.europa.eu
Eurostat, European Commission

George Kapetanios
george.kapetanios@kcl.ac.uk
King's College London

Massimiliano Marcellino
massimiliano.marcellino@unibocconi.it
Bocconi University

Gian Luigi Mazzi
gianluigi.mazzi@gopa.lu
GOPA

Fotis Papailias
King's College London
fotis.papailias@kcl.ac.uk

Evaluation of Nowcasting / Flash Estimation based on a Big Set of Indicators

DRAFT VERSION 18/12/2017

PLEASE DO NOT CITE *(optional)*

Prepared for the 16th Conference of the
International Association of Official Statisticians (IAOS)
OECD Headquarters, Paris, France, 19-21 September 2018

ABSTRACT

This paper aims at providing a primer on the use of big data in macroe-conomic nowcasting and early estimation, with a special focus on the use in official Statistical Agencies and similar institutions. We discuss: (i) a typology of big data characteristics relevant for Macroeconomic Nowcasting and early estimates, (ii) various methods for feature extraction of big data sources to usable time series format, (iii) econometric methodologies which could be used for nowcasting with big data, (iv) empirical nowcasting results and gains in terms of increased timeliness for three key target variables in four countries, and (v) various ways to evaluate and present nowcast estimates. We conclude by providing a set of recommendations to assess the pros and cons of the use of big data in a specific empirical context.

Keywords: big data, Nowcasting, Early Estimated, Econometric Methods.

Contents

1 INTRODUCTION

Rapid advancements of technology now allow to store every decision or action in our every day life, work and research. Mobile phones can indicate our position and activity via GPS and social media posts, browsing activity can be monitored -in the form of "cookies"- to provide personalised advertisements, various satellites take a vast amount of pictures and indicate activity during day and night. The transfor-mation of such digital traces results in a very large amount of data -henceforth, big data- which is used by government agencies, public institutions, banks, marketing companies and others.

While initially big data has been used mainly in the private sector, it also rep-resents an opportunity in other fields, possibly combined with more traditional data sources. In particular, Official Statistics could also benefit from big data, as indicated for example by the High-Level Group for the Modernisation of Statistical Production and Services (HLG) or the Eurostat Task Force on big data[1].

Nowcasting and the construction of early estimates are concerned with the pro-duction of a preliminary estimate for the contemporaneous value of an indicator, which has not yet officially been released. Leading examples[2] are the GDP and its components, deflators, and fiscal variables, which are typically re-leased at least 30- 45 days after the end of the reference month or quarter, and later revised. Nowcasts of monthly variables such as the HICP or confidence, sales, trade and labour market indicators could be also of interest.

In this paper we focus on the use of big data for macroeconomic nowcasting and the production of early estimates, by surveying, developing and applying proper data handling techniques combined with state of the art econometric methods. big data have substantial potential in this context, as timely/continuous/large sets of data should provide new or complementary information with respect to standard economic indicators.

Our aim is to provide a useful guide for the applied researcher in official Sta-tistical Agencies, or similar institutions, and present ways big data can be used in macroeconomic nowcasting to improve the quality of the early estimates, increase the timeliness of the releases,

[1] See `https://goo.gl/FzQB68` and `https://goo.gl/xQitj6` respectively

[2] See Buono et al. (2017) and the references therein for a comprehensive discussion and list of examples.

and complement the standard information with uncertainty and directional measures.

The rest of the paper is organised as follows: Section 2 presents a typology of big data, Section 3 discusses big data feature extraction, Section 4 briefly reviews some indicative methodologies which can handle big data, summarises various evaluation measures and the timely gains from a nowcasting experiment based on key variables for European countries, Section 5 discusses further evaluation measures and ways to communicate the results, and Section 6 offers brief overall conclusions and a set of recommendations for the practical use of big data for nowcasting and early estimation in the context of Statistical Agencies and similar institutions.

2 TYPOLOGY OF BIG DATA[3]

Advancements in computer technology during the last decades have allowed the storage, organisation, manipulation and analysis of vast amount of data from different sources and across different disciplines.

A traditional data source, revamped by the IT developments, is represented by Business Systems that record and monitor events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected by either private businesses (commercial transactions, banking/stock records, e-commerce, credit cards, etc.) or public institutions (medical records, social insurance, school records, administrative data, etc.) is highly structured and includes transactions, reference tables and relationships, as well as the metadata that sets its context. Traditional business data is the vast majority of what IT managed and processed, in both operational and BI systems, usually after structuring and storing it in relational database systems.

A novel data source is represented by Social Networks (human-sourced informa-tion). This information is the record of human experiences, by now almost entirely digitally stored in personal computers or social networks. Data, typically, loosely structured and often ungoverned, include those saved in proper Social Networks (such as Facebook, Twitter, Tumblr etc.), in blogs and comments, in specialized

[3]Below we present a general introduction to different types of big data, their advantages and characteristics. For more information we refer the reader to Buono et al. (2017).

websites for pictures (Instagram, Flickr, Picasa etc.) or videos (Youtube, etc.) or internet searches (Google, Bing, etc.), but also text messages, user-generated maps, e-mails, etc.

Yet another source of big data, and perhaps the fastest expanding one, is the so-called Internet of Things. Machine-generated data are derived from sensors and machines used to measure and record the events and situations in the physical world. It is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches. Examples in-clude data from sensors, such as fixed sensors (home automation, weather/pollution sensors, traffic sensors/webcam, etc.) or mobile sensors (mobile phones, connected cars, satellite images, etc.) but also data from computer systems (logs, web logs, etc.).

The resulting many types of big data have been already exploited in many scien-tific fields, such as climatology, oceanography, biology, medicine, and applied physics. Specific areas of economics have also seen a major interest in big data and business analytics, in particular marketing and finance. Instead, in conventional macroeco-nomics there have so far been limited applications, mostly concentrated in the areas of nowcasting/forecasting.

2.1 TYPES OF BIG DATA FOR MACROECONOMIC NOWCASTING

2.1.1 Financial Markets Data

Advances in computer technology and data storage have permitted the collection and analysis of high-frequency financial data. The most widely observed forms of financial big data are trades and quotes. The New York Stock Exchange (NYSE) initiated the collection of this data in 1992. This intraday data potentially pro-vides detailed information which could be used in the analysis of markets efficiency, volatility, liquidity as well as price discovery and expectations. Central banks also monitor activity across all financial markets and nowadays high-frequency financial data includes:

- Equities trades and quotes for all types of investors

- Fixed-Income trades and quotes for all types of investors
- Foreign Exchange trades and quotes for all types of investors
- NXL and OTC Derivatives and option transactions
- Generally, all other operations in financial markets.

This data is very sensitive and central banks do not make it publicly available, therefore most studies in the literature rely on anonymised data obtained by various third-party providers. Although financial big data is very important in the analysis of market microstructure, its use in macroeconomic nowcasting and forecasting is mainly on daily or weekly frequency.

Financial data in high-frequency form has been the main element in volatility and market microstructure studies. Moreover, the use of such data in macroeconomic forecasting, and subsequently in nowcasting, has been included in many studies, either after aggregating the data to the same frequency as the macroeconomic targets or, more recently, using mixed frequency models. Given that our focus in this project is in alternative big data types, we refer the reader to Stock and Watson (2002a), Stock and Watson (2002b), Giannone, Reichlin and Small (2008), Angelini, Camba-Mendez, Giannone and Reichlin (2011), Banbura, Giannone and Reichlin (2011), Banbura and Runstler (2011), Modugno (2013), Andreou, Ghysels and Kourtellos (2015) among others.

2.1.2 Electronic Payments Data

The term electronic payments is broad and considers all kinds of electronic funds transfer. In particular, forms of electronic payments include: (i) cardholder-initiated transactions, (ii) direct deposit payments initiated by the payer, (iii) direct debit payments initiated by businesses which debit the consumer's account for the purchase of goods or services, (iv) credit transfers, (v) online electronic bill payments, among others. The most heavily used form of electronic payments is the cardholder-initiated transactions, i.e. credit and debit card payments.

According to the Capgemini and BNP Paribas (2016) report, cards dominate the global non-cash market accounting for 65% of all non-cash transactions. In Table 1

we report the credit card usage in 2014 across six regions. Credit transfers follow with 17% share and direct debits with 12%. Finally, checks account for 6% globally. China, Hong Kong, India and other Asian markets rank first in the cards usage, with 84% share in the non-cash market. Second follows the Central European and Middle Eastern Area (CEMEA)[4], with cards usage at 77% of the non-cash transactions. Japan, Australia, Singapore and South Korea follow third with 75% cards share, and North America ranks fourth with 71% share. Latin America and Europe are last with 49% and 47% share of cards in the non-cash market. Even in the markets with the lowest usage, cards share is at least double of the other non-cash alternatives. For example, direct debits and credit transfers account for 23% and 26% in Europe compared to 47% of cards. The above statistics highlight that cards are the main form of non-cash payments. Card payments include online as well as offline Point of Sale (POS) purchases,[5] making them very useful in the tracking of consumer behaviour and retail sales (among others).

Non-cash payments mix (%)							
	Europe	**North America**	**JASS**	**CHI**	**Latin America**	**CEMEA**	**Global**
Cards	47%	71%	75%	84%	49%	77%	65%
Credit Transfers	26%	8%	17%	10%	32%	20%	17%
Direct Debits	23%	11%	7%	2%	15%	3%	12%
Checks	4%	11%	1%	5%	4%	0%	6%

Source: Capgemini and BNP Paribas (2016). JASS: Japan, Australia, Singapore and South Korea. CHI: China, Hong Kong, India and other Asian markets. CEMEA: Poland, Russia, Saudi Arabia, South Africa, Turkey, Ukraine, Hungary, Czech Republic, Romania and other Central European and Middle Eastern markets.

The cards payments are considered as a category of big data because of high frequency of transactions. There are in fact thousands of transactions throughout the day and, with the huge increase of e-commerce, also during the night. One specific characteristic of cards data is the weekly pattern in daily aggregated data (or intraday pattern in non-aggregated data). As indicated by the literature, consumers tend to purchase more goods and services towards the end of the week. Cards data usually is offered aggregated in order to ensure protection of personal details.

[4]CEMEA markets include: Poland, Russia, Saudi Arabia, South Africa, Turkey, Ukraine, Hungary, Czech Republic, Romania and other Central European and Middle Eastern markets.

[5]For example, card payment at a retail store.

The economic literature using credit cards data started recently. Galbraith and Tkacz (2007) is one of the first papers that published results based on cards data in macroeconomics. In particular, they use Canadian debit card transactions[6] in order to provide real-time estimates of economic activity. Their predictive regression analysis provides information on consumer behaviour, as well as improved nowcast estimates. At first, they find that household transactions have a weekly pattern (on average), peaking every Friday and falling every Sunday. The high-frequency anal-ysis of electronic transactions around extreme events explains expenditure patterns around the September 11 terrorist attacks and the August 2003 electrical blackout. Finally, consensus forecast errors for GDP and (especially non-durable) consumption growth can be partly explained by cards data.

Esteves (2009) uses Automated Teller Machines (ATM) and POS data to now-cast private consumption via predictive regression analysis. As in Galbraith and Tkacz (2007), he also finds that nowcasting of non-durable private consumption ben-efits from the use of ATM/POS data. One of the drawbacks of the paper is the short evaluation sample which consists of 18 and 10 temporal observations (2005Q1 - 2009Q2 and 2007Q1 - 2009Q2). However, this is a common feature in studies based on big data, and overall results are in favour of cards data use in macroeconomic nowcasting applications.

Carlsen and Storgaard (2010) use Dankort payments in order to nowcast the retail sales index in Denmark. Dankort is a debit card developed jointly by the Danish banks and introduced in 1983. The Dankort is free of charge to the customers, and the card is extensively used by households. This fact makes Dankort an ideal instrument for tracking household activity and thus, retail sales. Another advantage of using Dankort is the timing of publication. Dankort data is available one week after the reference month, whereas the retail sales index is published three weeks later. As mentioned in the previous studies, seasonal effects are also present which need extra care in order to end up with a clean dataset. The out-of-sample nowcast exercise is in favour of the two models which use cards data. However the evaluation period is again too narrow: monthly nowcasts between January 2007 and May 2008.

Galbraith and Tkacz (2011) build on their earlier work (Galbraith and Tkacz (2007)) in which they focus on the economic effects of extreme events using Canadian cards data. This paper is not

[6]Obtained via the Canadian interbank network, Interac.

related to the nowcasting/forecasting literature, however it demonstrates another area of application of high frequency cards data: the analysis of economic activity during periods when low frequency data, such as quarterly consumption, is unavailable. Examples of extreme events that are con-sidered in the paper are the September 11, 2001 events, the SARS epidemic in the spring of 2003 and the August, 2003 electrical blackout. Applications like this would also be useful in macroeconomic nowcasting as they could be used as early warning indicators of economic activity.

Galbraith and Tkacz (2015) tackle directly the issue of nowcasting Canadian GDP growth using Canadian credit and debit cards transactions as well as checks. They find that, among the payments data, debit card transactions seem to produce the most improved estimates. The issue of seasonality is also present here. The authors suggest the use of the X-11 methodology to clean the data.[7] Their main finding is that nowcasting using high frequency electronic payments improves by 65% between the first and final estimates, presenting supporting evidence on the use of electronic payments data.

Duarte, Rodrigues and Rua (2016) use ATM and POS high frequency data for nowcasting and forecasting quarterly private consumption for Portugal. Their ATM data is provided by Multibanco, which is the Portuguese ATM and POS network. Their methodology is based on Mixed Data Sampling (MIDAS) models and builds on the earlier work by Esteves (2009), confirming that the use of electronic payments data improves nowcasting and forecasting accuracy. Weekly payment data produce particularly good results, while daily data are too noisy.

Barnett, Chauvet, Leiva-Leon and Su (2016) derive an indicator-optimized aug-mented aggregator function over monetary and credit card services using credit card transaction volumes. This new indicator, inserted in a multivariate state space model, produces more accurate nowcasts of GDP compared to a benchmark model.

Finally, Aprigliano, Ardizzi and Monteforte (2016) use a mixed frequency dy-namic factor model to predict the Italian GDP growth using standard business cycle indicators (such as electricity consumption, industrial production, inflation, stock market indexes, manufacturing indexes, etc.) as well as payment systems data (cheques, credit transfers, direct debits, payment cards). They find that monthly payment data helps in tracking the economic cycle in Italy and

[7]See Kapetanios, Marcellino and Papailias (2017a) for a full discussion.

improves nowcasting. A separate screening exercise using the Least Absolute Shrinkage and Selection Operator (LASSO) confirms payment system variables as potential predictors of GDP growth.

2.1.3 Mobile Phones Data

With the introduction of mobile phones, nearly thirty years ago, scientists across various fields were positive that mobile phones usage and information would be an important tool for statistics. The data collection from basic functions of a mobile phone, i.e., receiving and making phone calls and short text messages, provides relevant information about population density, location, economic development of particular geographic areas and use of public transport, among others. The rapid development and growth of mobile phones technology during the past twenty years allows for even more specific data collection, related to internet activity, mobile banking, GPS tracking and other sensors data[8]. Overall, mobile phones data provides detailed information on human behaviour and, therefore, it can be useful in social sciences too.

The Deloitte (2012) report, which uses data from the Cisco's VNI Index for 14 countries, states that a doubling of mobile data use leads to a 0.5% increase in the GDP per capita growth rate. Given the heavy use of mobile phones, the collected data is characterised as big data due to the massive volume. News coverage also provides evidence that mobile data is promising for the future[9].

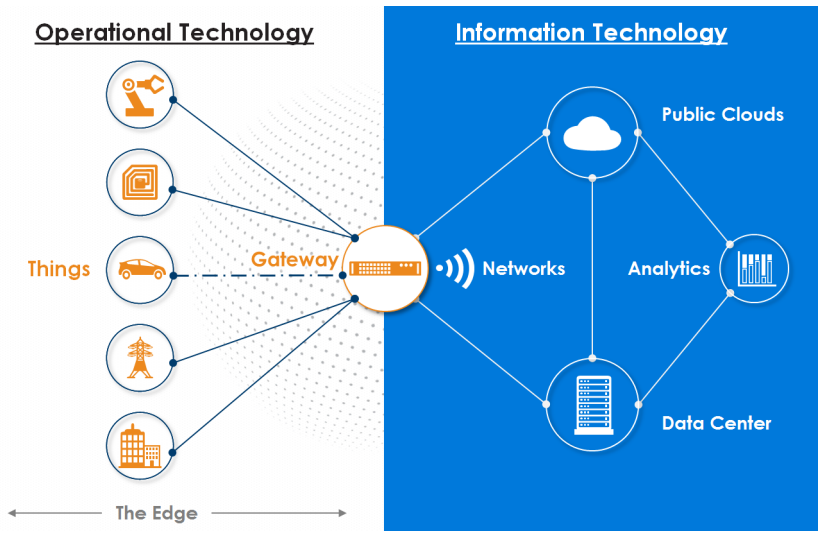
2.1.4 Sensor Data and the Internet of Things

Sensor data mainly refers to any kind of output of a device which detects and responds to input sources from the physical environment. One of the oldest examples is the temperature monitoring via climate sensors. Sensors have been long used in manufacturing of plants, cars, ships, military equipment, and sensor data has been used by operational technology engineers for quite many years.

Information technology and the rapid development of internet greatly facilitates the collec-

[8] In that sense, mobile phones data can also be part of sensors data.
[9] See `https://goo.gl/MYb21Q`, `https://goo.gl/e35IQQ` and `https://goo.gl/Dkdara`.

tion and distribution of sensor data and gives rise to the so-called "Internet of Things (IoT)". A "thing" includes any item which can be attached with sensors. Internet access allows the "thing" to automatically transmit the data via networks and store it to public clouds and databases. This, in-turn, provides easy access to sensor data for mining and analytics purposes.



Source: Graham (2016).

Mobile things, e.g., smart phones, computers, tablets, watches, home appliances, cars, drones, etc., as well as stationary things, e.g., wind turbines, temperature sensors, etc., all come attached with sensors which can send and receive information online. In the this category, we include datasets which are collected from sensors and other devices which do not track directly human activity as opposed to mobile phones data where all information is returned by human actions. According to Ebner (2015)[10], Gartner, a leading information technology research and advisory company, forecasts 25 billion connected things by 2020. The data generated by these sensors could impact almost every major industry, including healthcare, transportation, and energy. Cisco estimates that IoT could have an economic impact of $14 trillion by the early 2020s.

[10] https://goo.gl/9h3Zf8.

Smartphone manufacturers are already developing platforms which gather data from the watches and other instrumented objects of the IoT. This data is then made available to developers for the creation of new applications and analyses[11].

2.1.5 Satellite Images Data

Satellite imagery consists of images of the Earth or other planets collected by satel-lites. The satellites are operated by governments and businesses around the world. Satellite images are licensed to government agencies and businesses such as Apple and Google. One of the first image satellites was launched in 1946 taking one picture every 1.5 seconds. At the end of August, 2015 it was estimated that there were 4,077 satellites orbiting the Earth[12], of course not all of them were imaging satellites.

Satellites images have many applications in meteorology, oceanography, agricul-ture, forestry, geology, intelligence, warfare and others. Recently, satellite imagery has attracted the interest of economists as well. Photos of homes with metal roofs can indicate transition from poverty, night lights can show economic growth and monitoring of factory trucks and deliveries can be used for industrial production nowcasting; See Florida (2014) and Kearns (2015) for more details.

Satellite image data presents the following features: (i) the use of high quality images and the frames taken per second make satellite image databases very big and cumbersome, and (ii) in some cases, as in city night lights, the data is slowly changing and, thus, not useful for nowcasting. Lowe (2014) provides a brief guide on satellite data handling, which eases the use of this data.

2.1.6 Online Prices Data

The development of internet gave rise to online shopping. According to Abramovich (2014), online shopping retail sales are predicted to grow to $370 billion in 2017, up from $231 billion in 2012. Therefore, since online shopping substitutes, or at least supplements, offline shopping, online prices can also be used as a substitute, or supplement, of offline prices. Data collection over the internet is called web scraping. This technique provides flexibility and extreme automation. As with scanner data, scraped prices are a potentially useful instrument in nowcast-

[11] See `https://goo.gl/9h3Zf8`.
[12] `https://goo.gl/elqcWi`.

ing and short-term fore-casting CPI inflation and some of its subcomponents, as well as retail sales related variables.

Online prices are also characterised by seasonalities and specific irregularities that, as with scanner data, need to be taken into account by the researcher. Daily access to online super markets and retailers, which is publicly allowed, can lead to a mass collection of data. For example, a major UK retailer, Sainsbury's, offers 12 groceries categories for online shopping with about 50 products per category.[13] This leads to about 600 products online, thus 600 prices have to be collected from this retailer. Usually, there are 4 or more major retailers in a country which leads to about 2,400 prices to be collected. Over the course of a calendar year, this sums up to about 864,000 prices per year.

Academic papers in economics have started using web scraped data recently. Lunnemann and Wintr (2011) collected more than 5 million price quotes from price comparison websites for France, Italy, Germany, the UK and the US. Their data was collected daily for a year (December, 2004 - December, 2005). They find that for some product categories, prices change more frequently in the European countries. They also find that scraped prices are not more flexible than offline prices and, as mentioned in the scanner data section, there is heterogeneity in the frequency of price changes across online retailers.

Cavallo (2013) used web scraping to collect online prices from the largest su-permarket retailers in Argentina, Brazil, Chile, Colombia and Venezuela. The time frame spans from October, 2007 to March, 2011. The paper finds that for Brazil, Chile, Colombia, and Venezuela, indexes using the online prices approximate both the level and main dynamics of official inflation. This is evidence that scraped prices could be used for inflation nowcasting. However, this might not be true for all economies. Interestingly, the paper finds that, for Argentina, the online inflation rate is nearly three times higher than the official estimate, which in fact was not credible. This data collection is part of the MIT Billion Prices Project[14]. Rigobon (2015) and Cavallo and Rigobon (2016) provide a brief discussion of the project which is now expanded and prices are collected for European countries as well.

[13]This is a rough estimate as in specific categories there can be as much as 100 or more products.
[14]Available at `https://goo.gl/xb4H95`.

Boettcher (2015) describes in detail technological, data security and legal re-quirements of web crawlers focusing on Austria. The paper finds that web crawling technology provides an opportunity to improve statistical data quality and reduce the overall workload for data collection. Automatic price collection methods enable statisticians to react better to the increasing amount of data sources on the internet.

Cavallo (2016) uses again scraped prices to study the impact of measurement bias on three common price stickiness statistics: (i) the duration of price changes, (ii) the distribution of the size of price changes, and (iii) the shape of their hazard function over time. The paper finds that online prices have longer durations, with fewer price changes close to zero, and hazard functions that initially increase over time. The author claims that the differences with the literature is due to time-averaging and imputed prices in scanner and CPI data.

Metcalfe, Flower, Lewis, Mayhew and Rowland (2016) introduce the CLIP, which is an alternative approach to aggregating large data sets into price indices using clus-tering. The CLIP uses all the data available by creating groups (or clusters) of similar products and monitoring the price change of these groups over time. Unsupervised and supervised machine learning techniques are used to form these product clusters. The index is applied on web scraped data. The authors explicitly say that this index does not replace official statistics. However, it clearly shows the interest of official statistical agencies, the UK ONS in this case, in online prices. Also, Radzikowski and Smietanka (2016) try to construct a CPI for Poland based entirely on online prices.

Cavallo (2017) compares the online and offline prices of 56 large multi-channel retailers in 10 countries: Argentina, Australia, Brazil, Canada, China, Germany, Japan, South Africa, the UK and the US. He finds that price levels are identical about 72 percent of the time. Price changes are not synchronised but have similar frequencies and average sizes. These results show that, potentially, scanner prices, which are more difficult to collect on a daily basis, can be substituted by online prices.

2.1.7 Online Search Data

Online search data consists of searches for particular keywords on the world wide web. The user typically inserts a keyword or a phrase in the search field of a search engine website. Then, the web search engine returns the information which mostly relates to the keyword. The information may be a mix of web pages, images, and other types of files. Search engines maintain real-time information by running an algorithm on a web crawler, thus a newly uploaded website must be easily accessible by search engine robots in order to be included in the databases for future web searches.

Between 1993 and 1995, Lycos, Altavista and Yahoo were some of the first web search engines that gained popular attention and daily visits. However, the search results were based mainly on the web directory of each engine rather than its full-text copies of web pages. Some years later, in about 2000, Google was introduced. The company achieved better results for many searches with an innovative proce-dure called PageRank. This algorithm ranks web pages based on the number and PageRank of other web sites and pages that link there, on the premise that good or desirable pages are linked to more than others. Since then, Google search engine has dominated followed recently by Microsoft's Bing and Baidu[15].

Google launched a public web facility, Google Trends, which shows how often a particular keyword is searched relative to the total search-volume across various regions of the world, and in various languages. The procedure is simple for all internet users and Google Trends data is publicly available. At first, the user specifies the keyword or search-items she wants to look for. Then, Google Trends returns a time series line plot with time on the horizontal axis and search frequency on the vertical axis. The time series data is offered at weekly frequency starting in 2004 and can be downloaded in .csv format. Thinking about the behaviour of internet users, who might search for particular search-items multiple times throughout the day, it is easy to understand that the raw search data Google has is a type of big data. Therefore, Google Trends is a weekly aggregated, and thus structured, form of big data (even though it might not be "big" itself).

Google also offers another tool which aims to help the user with specified keywords and search-

[15]This is mainly used in China.

items, Google Correlate[16]. This service, which is part of Google Trends, finds search patterns which correspond with real-world trends. There are two ways a researcher can use this tool. First, if there exists a weekly or monthly time series data which is of interest, the researcher can upload this data on Google and identify search-items and keywords which are correlated with the time series. This, in principle, is useful as the keywords, which will then be used to extract Google Trends, are almost automatically selected. The second use of Google correlates would be between Google Trends and keywords. In case a researcher has already identified a particular search-item, Google Correlate can be used in order to provide a list of correlated keywords which could be used in the analysis in order to decrease selection bias[17]. However, given that this is an automatic procedure, Google Correlate is not able to filter non-appropriate keywords. For example, Google Correlate might return as a result celebrities' names which happens to be trending during the same time period.

Google Trends have been used in various applications in economics, finance, health sector, etc. with substantial success. See Askitas and Zimmermann (2009), Choi and Varian (2009), Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Bril-liant (2009), Da, Engelberg and Gao (2011), DAmuri and Marcucci (2012), Choi and Varian (2012), Koop and Onorante (2013), Yuan, Nsoessie, Lv, Peng, Chunara and Brownstein (2013), Tkacz (2013), Varian and Stephens-Davidowitz (2014), among others.

2.1.8 Textual Data

This includes any kind of dataset providing summarised information in the form of text. Examples of textual data include news and media headlines, information related to specific events, e.g., central banks' board meetings, Twitter data (this is analysed in more details in the next section) and Wikipedia information.

Schumaker and Chen (2006) investigate 9,211 financial news articles and 10,259,042 stock quotes covering the S&P 500 stocks during a five week period. They show that the model containing both article terms and stock price at the time of article release had the best performance in closeness to the actual future stock price, the same direction of price move-

[16] Available at `https://goo.gl/Uj1mox`

[17] This could be done by aggregating or averaging the Google Trend time series, or by extracting their common factors.

ment as the future price and the highest return using a simulated trading engine.

Moat, Curme, Avakian, Stanley and Preis (2013) use the frequency of Wikipedia page views. The paper presents evidence that financially-related Wikipedia page views have predictive ability over financial recessions.

Levenberg, Pulman, Moilanen, Simpson and Roberts (2014) present an approach to predict economic variables using sentiment composition over text streams of Web data. Their results show that high predictive accuracy for the Nonfarm Payroll index can be achieved using this sentiment over big text streams.

Baker, Bloom and Davis (2016) develop an index of economic policy uncertainty based on newspaper coverage frequency.[18] Using firm-level data, the authors find that policy uncertainty raises stock price volatility and reduces investment and em-ployment in policy-sensitive sectors like defense, healthcare, and infrastructure con-struction. At the macro level, policy uncertainty innovations foreshadow declines in investment, output, and employment in the United States and, in a panel VAR setting, for 12 major economies. Using the same indicators for policy uncertainty, Bacchini, Bontempi, Golinelli and Jona-Lasinio (2017), provide similar results for the slowdown of Italian investments. Ericsson (2015) and Ericsson (2016) construct indexes that quantify the FOMC views about the U.S. economy, as expressed in the minutes of the FOMC's meetings. Stekler and Symington (2016) quantify the min-utes of the FOMC and show that the FOMC saw the possibility of a recession but did not predict it. Using textual analysis, the authors are able to determine which variables informed the forecasts.

Thorsrud (2016) constructs a daily business cycle index based on quarterly GDP and textual information contained in a daily business newspaper. The newspaper data is decomposed into time series representing newspaper topics. The textual part attributes timeliness and accuracy to the index and provides the user with broad based high frequent information about the type of news that drive or reflect economic fluctuations. Eckley (2016) develops a news-media textual measure of aggregate economic uncertainty using text from the Financial Times. This index is documented to have a strong relationship with stock volatility on average.

[18] Their index is available online with real-time information and updates at: `https://goo.gl/ZCWx92`.

Textual analysis can also be used in political economy. Acemoglu, Hassan and Tahoun (2015) use textual data from the GDELT project[19] to proxy street protests in Arab countries and investigate the relationship between protests and stock market returns. Using daily variation in the number of protesters, they document that more intense protests in Tahrir Square are associated with lower stock market valuations for firms connected to the group in power relative to non-connected firms, but have no impact on the relative valuations of firms connected to other powerful groups.

2.1.9 Social Media Data

Since the introduction of the internet, users were finding ways to communicate with each other. Message boards, guestbooks, chat platforms and personal sites have been online for many years. These services, which could be considered as primal social networks, set the grounds for modern social platforms and the new age of online social interaction. Facebook was officially launched in 2004, however it was not until 2006 that it was widely open to all internet users aged 13 or older. It has about 1.86 billion monthly active users as of December 31, 2016[20]. Since its introduction, social networks have been rapidly expanded and become an integral part our lives. The social networks are accessible via computers as well as mobile devices allowing for continuous connectivity and interaction to news and events. Following Facebook, Twitter was introduced in 2006 as an online news and social networking service where users post and interact with messages, or "tweets", originally restricted to 140 characters. Instagram, going online in 2010, started as a photo-sharing site which now allows posting videos as well.

Social media, as in the case with online search data, illustrate human activity and reactions. Discussions or posts on Facebook, Twitter or Instagram include a variety of topics from personal issues to politics and breaking news. Therefore it would be reasonable to assume that, as in the case of Google Trends, social media data could have predictive ability towards social and economic variables. Based on the types of social media, Twitter seems to be the most appropriate to use for scientific analysis. Below we list some key reasons:

[19] See `https://goo.gl/QLAosJ` for more info.
[20] See `https://newsroom.fb.com/company-info/`.

- First, Twitter mainly uses short text streams which are often very specific about an event. In this sense, Twitter could also be part of the Textual Analysis data described in the previous section. The use of hashtags (#) make Twitter discussions easier for monitoring and tracking events. This allows to identify which discussions are "trending".

- Twitter data, due to its "higher"-frequency nature can offer more information. See Paul, Dredze and Broniatowski (2014) who argue that influenza forecasting using Twitter data is improved compared to Google Flu Trends.

- Politicians[21], reporters and analysts have included Twitter as one of their main means of communication. For example, reports correspondence during Federal Open Market Committee meetings are often on Twitter with multiple tweets for breaking news. The Guardian newspaper uses Twitter feeds on their website. Particularly, the use of Twitter during the Brexit referendum and US elections was very successful.

However, as in all big data types, Twitter data might include a lot of noise. Therefore a careful selection of topics/hashtags must be done. An alternative way to using hashtags would be to follow specific users or organisations. For example, a researcher could monitor the twitter feeds from various newspaper and media or-ganisations, government agencies and key reporters and analysts and, then, filter the feeds for particular hashtags or keywords.

2.1.10 Summary

We have provided a typology including nine main big data types: (i) financial markets data, (ii) electronic payments data, (iii) mobile phones data, (iv) sensor data, (v) satellite images data, (vi) online prices data, (vii) online search data, (iix) textual data, and (ix) social media data. This detailed analysis of existing big data and nowcasting/forecasting applications based on them already provides some indication of which types of data seem more promising. In particular, electronic payments data, scanner price and online price data, online search data, textual data and social media data all have substantial potential. Instead, mobile phones data, sensor data and sat-

[21] See Chi and Yang (2010) for more information about the use of Twitter in the Congress.

ellite images data seem less promising for macroeconomic nowcasting /forecasting, while they could be relevant for other types of macroeconomic analysis.

2.2 Types of big data by Dominant Dimension

Following, e.g., Doornik and Hendry (2015), we can distinguish three main types of big data according to their dominant dimension: Fat (big cross-sectional dimension, N, small temporal dimension, T), Tall (small N, big T), or Huge (big N, big T).

Huge numerical datasets, possibly coming from the conversion of even larger but unstructured big data, pose substantial challenges for proper econometric analysis, but also offer potentially the largest informational gains for nowcasting. Fat or Tall datasets can be also relevant in specific applications, for example for microeconomic or marketing studies in the case of Fat data or for financial studies in the case of Tall data. Tall data resulting from targeted textual analysis could be also relevant for macroeconomic nowcasting. The following table attempts to classify, - generally -, the ten big data categories as fat, tall or huge. However, the classification depends heavily on the application needs for big data.

Doornik and Hendry (2015) classification.

Type	Fat	Tall	Huge
Financial Markets			X
Electronic Payments			X
Mobile Phones	X		X
Sensor Data / IoT	X		X
Satellite Images		X	
Online Prices			X
Online Search		X	
Textual		X	
Social Media		X	

It must be noted that the classification depends heavily on the application needs for big data. Also, the last three types could be "Huge" if disaggregated data is available.

Table 1: big data Classification.

2.3 Data Issues

The previous discussion has already emphasized several advantages of big data in a nowcasting context, starting with the fact that they provide potentially relevant complementary information with respect to standard data, being based on rather different information sets. Moreover, big data are timely available and, generally, they are not subject to subsequent revisions, all relevant features for potential coin-cident and leading indicators of economic activity. Furthermore, big data could be helpful to provide a more granular perspective on the indicator of interest, both in the temporal and in the cross-sectional dimensions. In the temporal dimension, they can used to update nowcasts at a given frequency, such as weekly or even daily, so that the policy and decision makers can promptly update their actions according to the new and more precise estimates. In the cross-sectional dimension, big data could provide relevant information on units, such as regions or sectors, not fully covered by traditional coincident and leading indicators.

Besides all these actual and potential benefits of big data, it is however important to also consider some possible drawbacks, in particular related to internet data. The use of internet data for the production of official statistics has been evaluated in details, e.g., in the report "Analysis of methodologies for using the Internet for the collection of information society and other statistics"[22]. We are instead mainly interested in those aspects related to the use of big data for nowcasting economic indicators.

A first issue concerns data availability. As it is clear from the data categorization presented above, most data pass through private providers and are related to personal aspects. Hence, continuity of data provision could not be guaranteed. For example, Google could stop providing Google Trends, or at least no longer make them available for free or even totally shut down the service as they did for their Maps Engine. Indeed, they have recently decided to substantially restrict the availability of weekly Trends. Or online retail stores could forbid access to their websites to crawlers for automatic price collection. Or individuals could extend the use of softwares that prevent tracking their internet activities, or tracking could be more tightly regulated by law for privacy reasons. Similarly, there are Twitter data limitations through

[22] See `https://goo.gl/skaXxJ` for more information

Firehose.

Continuity of data availability is more an issue for the use of internet data in official statistics than for a pure nowcasting purpose, as it often happens in nowcasting that indicators become unavailable or no longer useful and must be replaced by alternative variables. That said, continuity and reliability of provision are important elements for the selection of a big data source.

Another concern related to data availability is the start date, which is often quite recent for big data, or the overall number of temporal observations in low frequency (months/quarters), which is generally low, even if in high frequency or cross-sectionally there can be thousands of observations. A short temporal sample is problematic as the big data based indicators need to be related to the target low frequency macroeconomic indicators and, without a long enough sample, the param-eter estimators can be noisy and the ex-post evaluation sample for the nowcasting performance too short. On the other hand, several informative indicators, such as surveys and financial condition indexes, are also only available over short samples, starting after 2000, and this feature does not prevent their use.

One more issue for internet based big data is related to the "digital divide", the fact that a sizable fraction of the population still has no or limited internet access. This implies that the available data are subject to a sample selection bias, and this can matter for their use. Suppose, for example, that we want to nowcast unem-ployment at a disaggregate level, either by age or by regions. Internet data relative to older people or people resident in poorer regions could lead to underestimation of their unemployment level, as they have relatively little access to internet based search tools, e.g. referred to as sample selectivity and representativeness issue.

For nowcasting, the suggestion is to carefully evaluate the presence of a possible selection bias, but this is likely less relevant when internet based data are com-bined with more traditional indicators and are therefore used to provide additional marginal rather than basic information. There can also be other less standard big data sources for which the digital divide can be less relevant. For example, use of mobile phones is quite widespread and mobility of their users, as emerging from calls and text messages, could be used to measure the extent of commuting, which is in turn typically related to the employment condition.

Another issue is that both the size and the quality of internet data keeps changing over time,

in general much faster than for standard data collection. For example, applications such as Twitter or WhatsApp were not available just a few years ago, and the number of their users increased exponentially, in particular in the first period after their introduction. Similarly, other applications can be gradually dismissed or used for different uses. For example, the fraction of goods sold by EBay through proper auctions is progressively declining over time, being replaced by other price formation mechanisms.

This point suggests that the relationship between the target variable and the big data (as well as that among the elements of the big data) could be varying over time, and this is a feature that should be properly checked and, in case, taken into consideration at the modelling stage.

Yet another issue, again more relevant for digital than standard data collection, is that individuals or businesses could not report truthfully their experiences, assess-ments and opinions. For example, some newspapers and other sites conduct online surveys about the feelings of their readers (happy, tired, angry, etc.) and one could think of using them, for example, to predict election outcomes, as a large fraction of happy people should be good for the ruling political party. But, if respondents are biased, the prediction could be also biased, and a large fraction of non-respondents could lead to substantial uncertainty[23].

As for the case of the digital divide, this is less of a problem when the internet data are complementary to more traditional information, such as phone or direct interviews, or indicators of economic performance.

Data could also not be available in a numerical format, or not in a directly usable numerical format. A similar issue emerges with standard surveys, for example on economic conditions, where discrete answers from a large number of respondents have to be somewhat summarized and transformed into a continuous index. However, the problem is more common and relevant with internet data.

A related problem is that the way in which the big data measure a given phe-nomenon is not necessarily the same as in official statistics, given that the data are typically the by-product of different activities. A similar issue arises with the use of proxy variables in econometric studies, e.g. measures of potential output or inflation expectations. The associated measurement error can

[23] See `https://goo.gl/PAkU2x` for more information about types of "response bias" in standard surveys

bias the estimators of structural parameters but is less of a problem in a forecasting context, unless the difference with the relevant official indicator is substantial.

Clearly, the collection and preparation of big data based indicators is far more complex than that for standard coincident and leading indicators, which are often directly downloadable in ready to use format from the web through statistical agen-cies or data providers. The question is whether the additional costs also lead to additional gains, and to what extent, and this is mainly an empirical issue. The literature review we have presented suggests that there seem to be cases where the effort is worthwhile.

A final issue, again common also with standard data but more pervasive in inter-net data due to their high sampling frequency and broad collection set, relates to data irregularities (outliers, working days effects, missing observations, etc.) and presence of seasonal / periodic patterns, which require properly de-noising and smoothing the data.

As for the previous point, proper techniques can be developed and the main issue is to assess their cost and effectiveness, which is again an empirical issue (which will be shortly considered later on).

A few other, more methodological, potential problems associated with big data are discussed in the next subsection.

2.4 Methodological Issues

As Hartford (2014) put it: " 'Big data' has arrived, but big insights have not. The challenge now is to solve new problems and gain new answers – without making the same old statistical mistakes on a grander scale than ever."

The statistical mistakes he refers to are well summarized by Doornik and Hendry (2015): " ... an excess of false positives, mistaking correlations for causes, ignoring sampling biases and selecting by inappropriate methods."

An additional critic is the "big data hubris", formulated by Lazer et al. (2014): ""Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. They also identify "Algorithm Dynamics" as an additional potential problem, where algorithm dynamics are the changes made

by engineers to improve the commercial service and by consumers in using that service. Specifically, they write: “All empirical research stands on a foundation of measurement. Is the instrumentation actually capturing the theoretical construct of interest? Is measurement stable and comparable across cases and over time? Are measurement errors systematic?”.

Yet another caveat comes from one of the biggest fans of big data. Hal Varian, Google’s chief economist, in a 2014 survey wrote: “In this period of big data it seems strange to focus on sampling uncertainty, which tends to be small with large datasets, while completely ignoring model uncertainty which may be quite large. One way to address this is to be explicit about examining how parameter estimates vary with respect to choices of control variables and instruments.”

In our nowcasting context, we can therefore summarize the potential method-ological issues about the use of big data as follows.

First, do we get any relevant insights? In other words, can we improve nowcast precision by using big data? As we mentioned in the previous subsection, this is mainly an empirical issue and, from the studies reviewed in the previous sections, it seems that for some big data and target variables this is indeed the case.

Second, do we get a big data hubris? Again as anticipated, we think of big data based indicators as complements to existing soft and hard data-based indicators, and therefore we do not get a big data hubris (though this is indeed the case some-times even in some of the nowcasting studies, for example those trying to anticipate unemployment using only Google Trends).

Third, do we risk false positives? Namely, can we get some big data based indicators that nowcast well just due to data snooping? This risk is always present in empirical analysis and is magnified in our case by the size of the dataset since this requires the consideration of many indicators with the attendant risk that, by pure chance, some of them will perform well in sample. Only a careful and honest statistical analysis can attenuate this risk. In particular, as mentioned, we suggest to compare alternative indicators and methods over a training sample, select the preferred approach or combine a few of them, and then test if they remain valid in a genuine (not previously used) sample.

Fourth, do we mistake correlations for causes? Again, this is a common problem in empirical analysis and we will not be immune for it. For example, a large number of internet

searches for “filing for unemployment” can predict future unemployment without, naturally, causing it. This is less of a problem in our nowcasting context, except perhaps at the level of economic interpretation of the results.

Fifth, do we use the proper econometric methods? Here things are more com-plex because when the number of variables N is large we can no longer use standard methods and we have to resort to more complex procedures. Some of these were developed in the statistical or machine learning literatures, often under the assump-tion of i.i.d. observations. As this assumption is likely violated when nowcasting macroeconomic variables, we have to be careful in properly comparing and selecting methods that can also handle correlated and possibly heteroskedastic data. This is especially the case since these methods are designed to provide a good control of false positives, but this control depends crucially on data being i.i.d. To give an exam-ple, it is well known that exponential probability inequalities, that form the basis of most methods that control for false positives, have very different and weaker bounds for serially correlated data leading to the need for different choices for matters like tuning parameters used in the design of the methods[24]. Overall, as we will see, a variety of methods are available, and they can be expected to perform differently in different situations, so that also the selection of the most promising approach is mainly application dependent.

Sixth, do we have instability due to Algorithm Dynamics or other causes (e.g., the financial crisis, more general institutional changes, the increasing use of internet, discontinuity in data provision, etc.)? Instability is indeed often ignored in the current big data literature, while it is potentially relevant, as we know well from the economic forecasting literature. Unfortunately, detecting and curing instability is complex, even more so in a big data context. However, some fixes can be tried mostly borrowing from the recent econometric literature on handling structural breaks.

Finally, do we allow for variable and model uncertainty? As we will see, it is indeed important to allow for both variable uncertainty, by considering various big data based indicators rather than a single one, and for model uncertainty, by comparing alternative procedures and then either selecting or combining the best performing ones. Again, all issues associated with model selection and uncertainty are likely magnified due to the fact that large data also allow for larger classes of models to be considered, and model selection methods, models to be considered, and

[24]See Roussas (1996) and the references therein for more information

model selection methods, such as information criteria, may need modifications in many respects.

3 A GENERAL DATA CONVERSION FRAMEWORK FOR UNSTRUCTURED NUMERICAL BIG DATA

A rather common feature of big data is the lack of a structure that makes them directly suitable for econometric or statistical analysis. This section aims to offer a general unstructured data conversion framework.

Data structuring can be made in different ways, producing different results ac-cording to the targeted phenomenon, the granularity of the phenomenon and the characteristics of the unstructured data. Also, the dimensions of the structured dataset can have relevant impact on the modelling process. In particular, if the out-put of the structuring process is just few to many time-series, then common econo-metric methods for small or large datasets can be used, while specific techniques are needed for really big, and possibly sparse, datasets. See Kapetanios, Marcellino and Papailias (2017a) for more details.

3.1 Conceptual Setting

We start by considering an unstructured dataset which consists of N events, each described by a vector of the form $y_i = (y_{1,i}, ..., y_{m,i}, t_i)' = (\tilde{y}_i, t_i)'$, $i = 1, ..., N$, where N is potentially very large. The vector $\tilde{y}_i'$ contains information on the event, while t_i denotes the time the event has occurred, and is referred to as a time stamp. This definition of an event in terms of a vector of characteristics and time stamp is very general. It is difficult to envisage any big dataset with a temporal dimension that cannot be accommodated by this definition. Time is defined continuously in an interval $(0, T]$. Our aim is to describe a mapping between these events and potential time series defined in discrete time, $t = 1, ..., T$, which can then be used for nowcasting or other econometric analyses.

We first need to define a mapping between the set $\mathcal{T}=\{t_i\}_{i=1}^{N}$ and time series observations. That is, a set function given by $\mathcal{T}_t = f(\mathcal{T}, t)$ where $\mathcal{T}_t$ is a subset of $\mathcal{T}$ containing the event time stamps that relate to time period t. The obvious mapping is for $\mathcal{T}_t$ to contain the

events whose time stamp satisfies $t-1<t_i\leq t$ but more complex setups, e.g. with lags, are possible. For example, it may be that $\mathcal{T}_t$ contains the events whose time stamp satisfies $t-p<t_i\leq t-p+1$.

Then, we can define a time series as

$$x_t=\sum_{t_i\in\mathcal{T}_t} g_t\left(y_i,\gamma\right), \tag{1}$$

where $g_t\left(y_i,\gamma\right)$ is a function possibly depending on t (to reduce the notational burden we sometimes suppress the t subscript), and parameterized by γ, mapping information on the event (contained in $\tilde{y}_i$) into a single number. This is a very general formulation and our discussion below provides ways in which we can better interpret its applicability. Before proceeding, it is important to note that the only restriction imposed here is a form of linear separability across events. The alternative would be a formulation of the type $x_t=g_t\left(y_1,...,y_{\mathcal{T}_t};\gamma\right)$. This would provide the ability to aggregate events in a nonlinear fashion, but we feel this might be unwieldy in practice. We also note that there is the potential for different time scales to be used in defining the mapping from events to time series. In particular, we can use the above mapping framework to construct, for example, both quarterly and monthly time series which can then be used by considering mixed frequency methods. Finally, we note that missing fields or elements in y_i can simply be treated by removing the events or, if the incidence of missing values is great, by setting indicator functions, associated with selecting particular features, to zero.

3.2 Aggregation

We suggest to consider g functions of the type:

$$g_t\left(y_i,\gamma\right)=\sum_{j=1}^{m} w_{tj}g_{tj}\left(y_{j,i},\gamma_j\right), \tag{2}$$

where the w_{tj} are weights generally depending on t. For example, the function could be $g_{tj}\left(y_{j,i},\gamma_j\right)=1$, or $g_{tj}\left(y_{j,i},\gamma_j\right)=y_{j,i}^{\gamma_j}$ or $g_{tj}\left(y_{j,i},\gamma_j\right)=y_{j,i}I(|y_{j,i}|>\gamma_j)$, with $(w_{t1},w_{t2},...,w_{tm})$ equal to, e.g., $(1,1,...,1)$ or $(1/m,1/m,...,1/m)$ or even $w_{tj}=\frac{\tilde{w}_{tj}}{\mathcal{T}_t}$, thereby providing a useful normalisation for the summation.

The function in (2) is a linear transformation of (possibly non-linear) g_{tj} functions. As an alternative, a non-linear transformation could be considered. Moreover, the w_{tj} weights could be also optimally computed given a certain error loss function.

An interesting possibility is to cast in this set-up the construction of internet search-based indicators, such as Google trends. In this context, $(y_{1,i}, ..., y_{m,i}) = (y_1, ..., y_m)$ denotes the numerical representation of alphanumeric sequences that de-scribe generic internet searches, for example y_1 corresponds to "recession" and y_2 t o "income" (so $m = 2$ i n this example).[25] Then, we can define

$$g(y_i, \gamma) = I(y_i \in \mathcal{G}_{t_i}(\gamma)), \tag{3}$$

where $\mathcal{G}_{t_i}(\gamma)$ denotes (possibly a subset of) all Google searches during period t_i (transformed into numerical representation), and may depend on the parameter vector γ (for example, γ can select searches for only a specific country or in a specific language), so that the function $g(y_i, \gamma)$ counts how many times y_i (a search for both "recession" and "income") happened in period t_i. Computing $g(y_i, \gamma)$ for $i = 1, ..., N$ and plugging the results into (1) returns a "Google trend" time series for period $t = 1, 2, ..., T$. Naturally, if $m = 1$ only single key-word searches are considered, so we would get two separate Google trends for "recession" and "income".

Once parameterised by, e.g., setting a subset of w_j to zero and choosing the functional form for g_j, the parameters , $w_j, \gamma_j, j = 1, ..., m$ can be either fixed a priori or calibrated.

The most obvious approach is to consider many different events $\tilde{y}_i$ and/or functions g, by, e.g., defining a grid for γ and w, and so obtain a large number of time series indicators, which can then be handled by appropriate econometric methods.

An interesting alternative is to pin down the function g by considering which choice of g has the best forecasting ability for the target variable when used on its own or potentially,

[25] It is important to note that the list of key words that should be entered into Google Trends is a matter for consideration. However, in our view the wider the list the better, since our recommendation is that at a later stage all these variables constructed via Google Trends (or other Google facilities such as Google Correlate) can be analysed in a regression setting using various data rich methodologies. As a result one can design a wide variety of relevant keywords (the use of a dictionary or a thesaurus may be useful here, together with an automatic translation service such as Google translate) and these can then be used on their own or combined through the use of logical relationships based on "or" or "and" to provide the final list.

in combination with a lag of the target variable.

3.3 Features Extraction

Apart from the above function, we can extract features from the distribution of the events which occur at time t. Depending on the nature of the underlying data, we could extract the variance, various percentiles, the skewness and kurtosis, or even higher moments, and then use the resulting time series of big data features.

Formally, and assuming that N events occur at a particular time period, t, we can extract the features using:

$$\text{Variance: } g_t(y_i) = \frac{1}{N}\sum_{i=1}^{N}(y_{it} - \overline{y}_t)^2, \tag{4}$$

$$\text{P-Percentile: } g_t(y_i) = y_{\lceil \frac{P}{100}N \rceil t}, \text{ where } \lceil \cdot \rceil \text{ denotes ordinal ranking,} \tag{5}$$

$$\text{Skewness: } g_t(y_i) = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_{it} - \overline{y}_t)^3}{\left[\frac{1}{N-1}\sum_{i=1}^{N}(y_{it} - \overline{y}_t)^2\right]^{3/2}}, \tag{6}$$

$$\text{Kurtosis: } g_t(y_i) = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_{it} - \overline{y}_t)^4}{\left[\frac{1}{N}\sum_{j=1}^{N}(y_{it} - \overline{y}_t)^2\right]^2} - 3, \tag{7}$$

where $\overline{y_t}$ denotes the sample mean of N events at time t.

These measures are rather intuitive, with the specific choice depending on the nature of the big data and the purpose of the exercise.

3.4 Data Mining

Another approach to reduce dimensionality is data mining and, in particular, clustering. To keep things simple and intuitive, we illustrate this approach by means of a basic clustering method, which is the $k-means$ clustering.

$k-means$ clustering aims to partition N observations into k clusters, with generally $k << N$, in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Given a set of observations, $(y_1, ..., y_i)$, $k-means$ clustering aims to split the data into k sets, e.g. a partition $S = (S_1, ..., S_k)$,

so as to minimise the within-cluster sum of squares which is the sum of distance func-tions of each point in the cluster to the K center. The relevant objective function is:

$$\arg_s \min \sum_{j=1}^{k} \sum_{y_i \in S_j} \|y_i - \mu_j\|^2 , \tag{8}$$

where μ_j is the mean of points in S_j. This will return k centers. In this way, the researcher succeeds in the reduction of the number of events to be considered. This method is particularly useful when similar observations repeatedly appear in the sample. In this case simple aggregation could be suboptimal, as it could hide substantial heterogeneity. See the seminal work of Lloyd (1982) for a starting point.

3.4.1 Random Subsampling

The above time series conversion techniques are expected to work well in applications where the researcher is interested in summaries and/or features of the data and where the size of the dataset allows the required computations. However, how should we approach cases where data is huge? For example, consider datasets which are measured in terabytes, petabytes, exabytes, zettabytes or even yottabytes. Can these conversion techniques still be applied? In cases where even database managers struggle to work well with software to construct structured time series, we can adopt the random subsampling approach; see Ng (2016) and references therein.

In this approach, random subsamples of the data are selected to create a new big dataset which is significantly smaller. This can be related to Equation (1) using indicator functions. For example, we could select observations of random seconds in an hour, or random hours in a day or even random days in a month. Then, the conversion techniques as described above can be applied and continue in the standard way. In this cost-efficient way, the researcher can randomly extract a large number of subsamples, which consequently leads to a large number of structured time series. At this point, an average of the resulting time series can be used or principal components can be extracted. The issue here is mostly one of data management and less of econometric difficulty, since a number of methods stay, in principle, valid for any dataset size while the informational loss associated with subsampling has an unknown cost.

4 ECONOMETRIC METHODS, EVALUATION CRITERIA & NOWCASTING EXAMPLES

In this section we provide a short overview of approaches that could be used for macroeconomic nowcasting using big data based indicators. This may seem restric-tive but, as we have discussed, it is typically convenient to structure and reduce the dimensionality of big data prior to using them for macroeconomic forecasting. We take an applied perspective, starting with a description of the data used in the nowcasting exercise, and then discussing reasonable econometric methods for this application, evaluation criteria, and results. Our perspective is to provide a compre-hensive illustration and applicability of some of these methods.

4.1 Data

4.1.1 Targets

We consider the four largest EU countries: Germany (DE), France (FR), Italy (IT) and the UK. For each of them, we have collected data on the quarterly GDP growth rate and three other key monthly economic indicators: industrial production (IP), harmonised index of consumer prices (HICP) and unemployment rate (UR). The datasets considered are made available by Eurostat and have been downloaded from their online dissemination database[26].

4.1.2 Predictors

Our set of monthly macroeconomic predictors includes various coincident and lead-ing indicators plus additional key economic variables for each country. Specifically, we consider: Bank Lending Rate, Bankruptcies, Building Permits, Capital Flows, Car Registrations, Construction Output, Consumer Credit, Core Consumer Prices,

various CPI components, Crude Oil Production, Export Prices, Exports, Factory Orders, Gasoline Prices, House Price Index, Import Prices, Imports, Job Vacan-cies, Manufacturing Production, Mining Production, Money Supply M1, M2 and M3, New Or-

[26]The Eurostat codes we use are: target.ei_isin_m.XX, target.prc_hicp_midx.XX, target.ei_lmhr_m.XX and target.namq_10_gdp.XX for IP, HICP, UR and GDP respectively. XX denotes the corresponding country code. Data can be downloaded from `https://goo.gl/pThhqB`.

ders, Private Sector Credit, Producer Prices, Steel Production, Youth Unemployment Rate, Consumer Confidence Indicators and various surveys.

Our set of weekly variables includes mainly financial indicators: interest rates at various maturities and spreads, equity indexes, volatility indexes.

Finally, we have constructed two big data-based uncertainty indicators, one relying on Reuters news and the other on Google searches, which are discussed in more details below.

4.1.3 The Reuters Uncertainty Index

Using web-scraping procedures, we downloaded data from the Reuters news database and construct the uncertainty indexes.[27] In particular, we used the following key-words:

- Germany: at least one of {uncertainty uncertain, uncertainty, uncertainties} and at least one of {Germany, German, Germans}.
- France: at least one of {uncertainty uncertain, uncertainty, uncertainties} and at least one of {France, French}.
- Italy: at least one of {uncertainty uncertain, uncertainty, uncertainties} and at least one of {Italy, Italian, Italians}.
- UK: at least one of {uncertainty uncertain, uncertainty, uncertainties} and at least one of {UK, Britain, British, United Kingdom, Briton}.

In Kapetanios, Marcellino and Papailias (2017b), it has been shown that Reuters Uncertainty is strongly correlated with other uncertainty measures such as the Eco-nomic Policy Uncertainty index of Baker, Bloom and Davis (2016) and volatility indexes such as VIX.

4.1.4 The Google Uncertainty Index

The Google Uncertainty Index is based on the use of Google trends. For the general uncertainty and risk indexes, we consider four keywords, given that the searches are di-

[27] We are grateful to Mattia Serrano' for help with the construction of these indexes.

rected worldwide and the language barrier might jeopardise the result, and two keywords for the country-specific indexes using the domestic languages. In particular, we included the following Google Trends:

- Germany: for the German Google Uncertainty Index we use the keywords "unsicherheit" and "risiko" across web and news searches in the region of Ger-many.
- France: for the French Google Uncertainty Index we use the keywords "incertitude" and "risque" across web and news searches in the region of France.
- Italy: for the Italian Google Uncertainty Index we use the keywords "incertezza" and "rischio" across web and news searches in the region of Italy.
- UK: for the UK Google Uncertainty Index we use the keywords "uncertainty" and "risk" across web and news searches in the region of the UK.

For the separate countries, we used both "uncertainty" and "risk" keywords in order to cover more user profiles and obtain a more robust result.

4.1.5 Transformations

To ensure that the variables under analysis are stationary, a pre-requisite for several of the econometric methods we will implement, we use a set of transformations, which include: (i) log, (ii) first difference, (iii) percentage change, (iv) log difference, (v) second log difference. The specific transformation for each variable follows standard practice in the literature, see, e.g., McCracken and Ng (2015) as well as Stock and Watson (2002a and 2002b) among others.

We nowcast the period-to-period percentage change of IP, HICP and GDP and the period-to-period first difference of UR. The nowcasts for most models are first produced using growth rates, and then translated to levels, as Eurostat and other official statistical agencies typically publish their nowcasts in levels.

4.1.6 Timing, Mixed-Frequency & Unbalancedness

For each variable, we mark the publication delay and the day of the month that the release for this variable is due. For example, suppose that industrial production for month T is released

on the 25th day of the next month, month T+1. In that case, we note one time period publication lag (i.e., -1) and the 25th day as publication re-lease. Repeating this procedure for each variable allows us to construct a nowcasting exercise that is accurate (on average) in terms of the information which is available at each point in time. For example, when we are about to construct nowcasting es-timates five weeks prior to the official release, we use only the available information up to that point.

In our nowcasting exercise we transform weekly observations to monthly (or quar-terly) by averaging the available information in each month (or quarter), as routinely done when constructing bridge models. This creates an unbalanced panel of variables in which weekly-to-monthly variables have zero publication lag, i.e., information is up-to-date, but macroeconomic variables might have 1, 2, or more periods of publi-cation lags. In such cases of unavailable information, we assign missing values.

The U-MIDAS approach could also be an attractive alternative, in general, to deal with mixed frequency data. However, in our specific context the available sample is too short and the difference in sampling frequency can be rather high (when going from weekly to quarterly)[28]. The MIDAS approach could reduce the number of parameters to be estimated but it would introduce non-linearity, which would make an estimation with a large number of explanatory variables practically unfeasible. Hence, our preference for simple temporal aggregation. See Ghysels et al. (2004) and Foroni, Marcellino and Schumacher (2015), among others, for MIDAS and U-MIDAS models.

To deal with the missing values at the end of the sample due to publication delays, we adjust each series by moving it forward so that the last observed data point matches the observed value in the dependent variable. Other solutions could be to replace the missing values with the median or mean, or to use extrapolation employing a simple autoregressive (AR) or other model. However, the median or mean could be different from the recent trend of the variables, and extrapolation is complex when dealing with a very large dataset.

[28]In particular, Reuters and Google data are only available from 2007 onwards, so there would be only about 40 quarters in the U-MIDAS regressions for GDP growth with a large number of regressors, which would lead to overfitting.

4.1.7 Time Span

Because the Reuters Uncertainty Index starts in 2007, our monthly variables (in-cluding monthly targets) span from 2007-01-31 to 2016-10-31 (118 months). The nowcasting exercise starts in 2014-01-31, to ensure that there is sufficient data to be used in the estimation also for the complex econometric models. A lag of the target variable is included in most specifications, i.e., an autoregressive term. However, lags of the predictors are not included, to avoid overfitting due to the short sample span.

For the monthly variables, we have 34 evaluation periods (2014-01-31 to 2016- 10-31). For the quarterly GDP, we have 12 periods, still for 2014-2016. The short evaluation sample should be kept in mind when assessing the empirical results.

4.2 Nowcasting Exercise

The nowcasting exercise is based on the algorithm described in the following steps.

1. First, we leave a number of observations, T^{OUT}, out-of-sample, in order to use them in the evaluation of the nowcasting performance of different models. In our experiments, $T^{OUT} = 34$ for the monthly targets and $T^{OUT} = 12$ for the quarterly target.

2. The initial sample we use in the first round of estimation and nowcasting is $T_1^{IN} = \{1, .., (T - T^{OUT}) + 1\}$. Then, we estimate the parameters and produce the nowcasts from the various models. We construct nowcasts for $h = \{-5, -4, -3, -2, -1\}$ weeks prior to the target date when the official release is due. For each h, we keep the same target date, however we re-estimate and produce different nowcasts (updates) using all available information up to that time point. This produces five estimates for each corresponding target date.

3. We repeat Step 2 in a recursive manner , i.e. $T_2^{IN} = \{1, .., \left(T - T^{OUT} + 2\right)\}$ and generally $T_j^{IN} = \{1, .., (T - T^{OUT} + j)\}$. We stop when $T_j^{IN} = \{1, .., (T - 1)\}$, as we always need the true value of the next period to evaluate the nowcasts.

At the end of the above recursive procedure we end up with T^{OUT} nowcasts for each model under consideration.

4.3 Econometric Models

In the nowcasting exercise we employ several methodologies to assess the relative gains from more complex methods that can handle large datasets with respect to simpler procedures. Specifically, we consider:

- (7) Naive and AR models. The first group of models consists of some simple models which assume that the best nowcast is given by: the average of the last four periods $(Ave4)$, the average of the last twelve periods $(Ave12)$ and the average of the last twenty-four periods $(Ave24)$. The Naive model uses the last observed value as the best nowcast (which coincides with that from a pure random walk model). Then, we include an $AR(1)$, $AR(4)$ and an $AR(p_{AIC})$ model where the p_{AIC} is determined via the Akaike's Information Criterion.

- (6) Simple Linear Regressions. These are simple specifications using the Google and Reuters Uncertainty Indexes with zero, one and three lags of the target variable.

- (8) Various other univariate models. The following models have been shown to work well in various forecasting competitions by the International Journal of Forecasting. More details can be found in Hyndman and Khandakar (2008).

  - AutoArima: chooses the best ARIMA(p,d,q) model using AIC. Transformation of the univariate series not necessary as the model handles integrated series.

  - ETS and BaggedETS: Exponential smoothing methods as in Hyndman et al. (2002) and Bergmeir et al. (2016). The methodology is fully automatic and performed extremely well on the M3-competition data. The bootstrapped series are obtained using the Box-Cox and Loess-based decomposition (BLD) bootstrap; see Bergmeir et al. (2016).

  - BATS and TBATS: This class of models are exponential smoothing state space models with Box-Cox transformation, ARMA errors, Trend and Seasonal components. It is part of automatic procedure forecasting. See De Livera et al. (2011).

  - Neural Networks (NN): This is a simple feed-forward neural network with a single hidden layer and lagged inputs.
  - Spline forecasts: this model produces local linear nowcasts using cubic smoothing splines.
  - Theta method: The theta decomposition method of Assimakopoulos and Nikolopoulos (2000).

- (16) Dynamic Factor Analysis (DFA). We use the default setup of Giannone, Reichlin and Small (2008) with (q, r, p) = (2, 2, 1) where q is the dynamic rank, r is the static rank and p is the AR order of the state vector. We also try three more settings with (q, r, p) = (3, 3, 1), (q, r, p) = (4, 4, 1) and (q, r, p) = (5, 5, 1). For each setting we use: (i) the set of macroeconomic and financial indicators only (MacroFin), (ii) the set of macroeconomic and financial indicators including the Google Uncertainty Indexes (MacroFin-Google),(iii) the set of macroeconomic and financial indicators including the Reuters Uncertainty Indexes (MacroFin-Reuters), and (iv) the set of macroeconomic and financial indicators including both the Google and the Reuters Uncertainty Indexes (MacroFin-GoogleReuters).

- (20) Partial Least Squares (PLS). We use PLS to extract one, two, three, four and five factors; the resulting models and nowcasts are labeled, respec-tively, P LS(1), P LS(2), P LS(3), P LS(4) and P LS(5). As above, for each model we use MacroFin, MacroFin-Google, MacroFin-Reuters and MacroFin-GoogleReuters.

- (20) Sparse Principal Components (SPC). We use SPC to extract one, two, three, four and five factors; the resulting models and nowcasts are labeled, respectively, SP C(1), SP C(2), SP C(3), SP C(4) and SP C(5). As above, for each model we use MacroFin, MacroFin-Google, MacroFin-Reuters and MacroFin-GoogleReuters.

- (8) LASSO and Elastic Net (EN). We use the standard 10-fold cross-validation to determine the value for λ in LASSO. The chosen value is the one which minimises the in-sample MSE. As above, we use MacroFin, MacroFin-Google, MacroFin-Reuters and MacroFin-GoogleReuters.

- (4) Spike and Slab (SS) regressions using MacroFin, MacroFin-Google, MacroFin-Reuters and MacroFin-GoogleReuters.
- (4) Data-Driven Automated Forecasting Strategies. On top of the above methodologies we introduce some automated data-driven "*forecasting strategies*". Our idea is simple and intuitive: we suggest the use of a "*model rotation*" strategy which chooses the model with the smallest cumulative nowcast error. Fur-thermore, we use an equally-weighted average of the top three, five and ten models.

In addition to the above procedures, we also consider the inclusion of a lag in the set of predictors and let each method choose the necessary variables unconditionally (even though in some cases a lag might not be chosen as significant in terms of nowcasting/forecasting).

4.4 Evaluation Criteria

Once we have computed T^{OUT} nowcasts for 5 to 1 weeks prior to the release, and transformed them in levels, we evaluate their performance using the mean absolute error and the root mean squared forecast error statistics defined as:

$$MAE_{i,h} = \frac{1}{T^{OUT}} \sum_{t=1}^{T^{OUT}} |e_{i,t}| \,,$$

$$RMSFE_{i,h} = \left(\frac{1}{T^{OUT}} \sum_{t=1}^{T^{OUT}} e_{i,t}^2 \right)^{\frac{1}{2}} \,,$$

where e_i is the out-of-sample forecast error (in levels) for model i and weekly nowcast h weeks prior to the release.

All our tables present the actual MAE and RMSFE in order to illustrate how nowcast errors change as we approach the release date.

We further calculate the Diebold and Mariano (1995) statistic for predictive ac-curacy as follows:

$$DM = \frac{\overline{d}}{\left(\widehat{LRV}_{\overline{d}}/T \right)^{1/2}}$$

where

$$\overline{d} = \frac{1}{T^{OUT}} \sum_{t=1}^{T^{OUT}} d_t,$$

$$d_t = e_{1,t}^2 - e_{2,t}^2,$$

$$LRV_{\overline{d}} = \gamma_0 + 2 \sum_{j=1}^{\infty} \gamma_j, with\ \gamma_j = cov(d_t, d_{t-j}),$$

for candidate models 1 and 2. The null hypothesis of the DM statistic states equal predictive ability between models. In the tables we report the p-value of the test.

4.5 Summary

The main goal of this exercise is to assess whether the use of big data, as proxied by Reuters news and Google searches in this research, can, either in isolation or in combination with high frequency economic and financial indicators, improve the precision of nowcasts and flash estimates. We are interested in gains in terms of both standard measures such as MAE and MSE and in increased timeliness.

We find that the nowcast error decreases significantly when we estimate three, two and one weeks prior to the official release. The inclusion of big data-based uncertainty indexes results in improved nowcasting performance. See Figure 1, as an example. In some cases even a simple linear regression model using the Reuters index and three lags of the target variable results in accurate and robust nowcasts. Various univariate models also seem to perform well. However, it must be highlighted that the evaluation sample is short (2014-2016) and there is a strong trend in the series, which works in favour of these simple models.

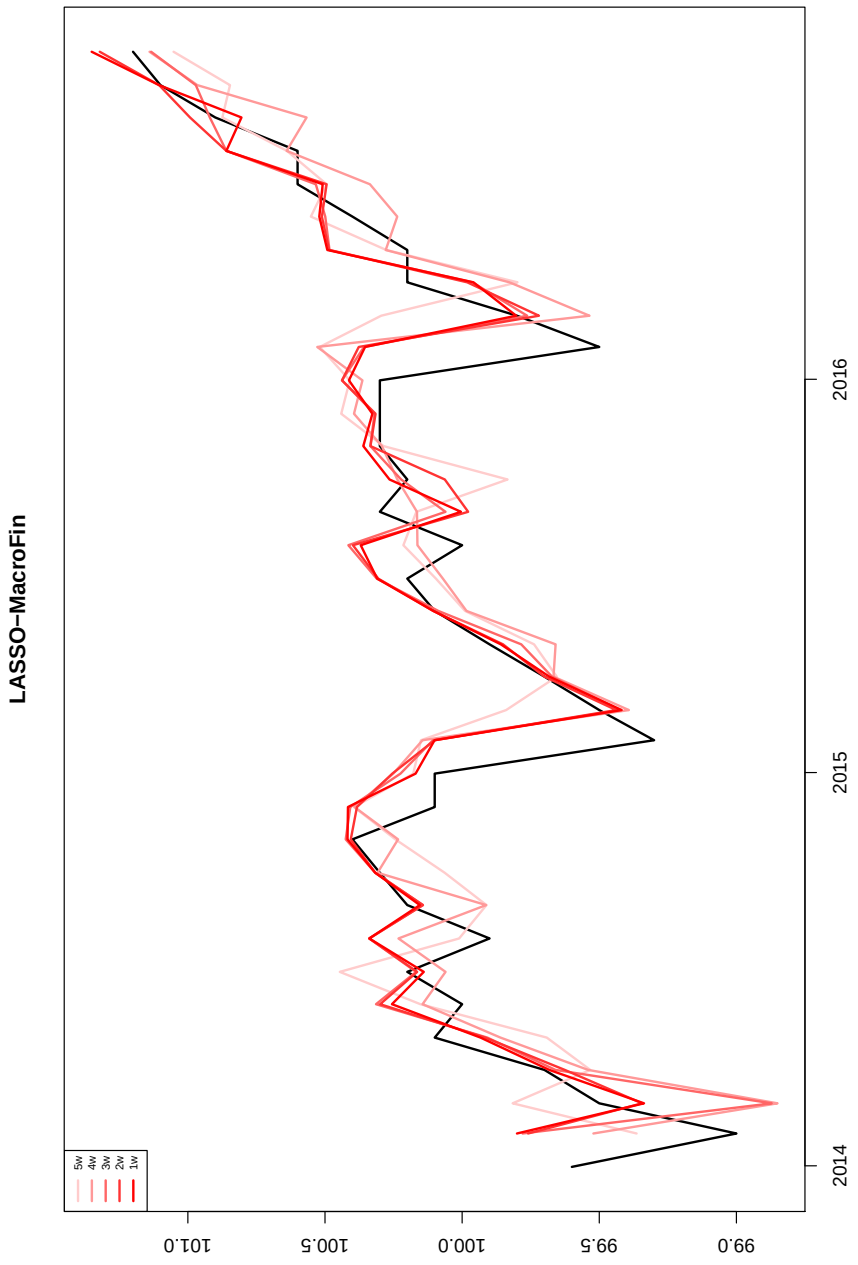


Figure 1: UK, HICP

5 FURTHER EVALUATION & METRICS ROBUSTNESS

While point nowcasts and forecasts are routinely computed and reported, also in of-ficial publications, there is growing interest in providing also measures of uncertainty around the point forecasts, and possibly other complementary information, such as directional forecasts. Hence, in this Section, we review density and interval forecasts and methods for their evaluation. Then, we apply them empirically, as a contin-uation of the nowcasting exercise, to assess whether the big data based indicators can also yield gains in terms of reducing uncertainty and/or improving directional accuracy.

5.1 Interval and density forecasts

Assuming that the model used to generate the flash estimates is correctly speci-fied and with normal errors, and that the sample size *T* is large so that parameter estimation error can be ignored, we have

$$\left(\frac{y_{T+h} - \hat{y}_{T+h}}{\sqrt{V(e_{T+h})}} \right) \sim N(0,1), \tag{9}$$

where $\hat{y}_{T+h}$ indicates the forecast and e_{T+h} the forecast error, $e_{T+h} = y_{T+h} - \hat{y}_{T+h}$, and $V(\bullet)$ denotes the variance function. Equation (9) implies

$$y_{T+h} \sim N(\hat{y}_{T+h}, V(e_{T+h})), \tag{10}$$

which is the expression for the *density forecast* of y_{T+h}.

The density forecast can be used to assign probabilities to specific events of interest concerning the future behaviour of the variable y. For example, if y is inflation, with the formula in (10) we can compute the probability that inflation in period $T+h$ will be higher than 2%.

Another use of the forecast density is to construct *interval forecasts* for y_{T+h}. A $[1-\alpha]\%$ forecast interval is represented as

$$\hat{y}_{T+h} - c_{\alpha/2}\sqrt{V(e_{T+h})}; \hat{y}_{T+h} + c_{\alpha/2}\sqrt{V(e_{T+h})}, \tag{11}$$

where $c_{\alpha/2}$ is the $(\alpha/2)\%$ critical value for the standard normal density. For example, a 95% confidence interval is given by

$$\hat{y}_{T+h} - 1.96\sqrt{V(e_{T+h})}; \hat{y}_{T+h} + 1.96\sqrt{V(e_{T+h})}. \tag{12}$$

As an example, if y is again inflation, the optimal linear point forecast is $\hat{y}_{T+h} = 2$, and $V(e_{T+h}) = 4$, the formula implies that a 95% interval forecast for inflation in period $T + h$ is $[-1.92, 5.92]$.

The interpretation of the interval forecast is the following. Suppose that we could generate a very large number of samples of data for y and X, each of size $T + h$, and for each sample construct the interval forecast for y_{T+h} as in (11). Then, in $[1 - \alpha]\%$ of the samples the realization of y_{T+h} will fall in the interval described in (11). A more common interpretation is that there is a $[1 - \alpha]\%$ probability that the future realization of y_{T+h} will fall in the interval in (11). Hence, continuing the example, there is a 95% probability that inflation at $T + h$ will be lower than 5.92 and higher than −1.92. However, strictly speaking, only the former interpretation of the confidence interval is correct.[29]

Finally, density forecasts and confidence intervals can be also constructed with different assumptions on the distribution of the error term, though the derivations are more complex. Also, as long as the distribution of the error is symmetric, the density (and the interval forecasts) will be centered around the optimal point forecast that coincides, as said, with the future expected value of the dependent variable, conditional on the available information set. Finally, in the case of nonlinear models, simulation methods are generally required to approximate the density forecasts.

5.2 Evaluation of interval and density forecasts

Let us indicate the density forecast by f_t and its cumulative distribution function (CDF) by F_t. Similarly, we indicate the true density of the target variable by g_t and its CDF by G_t. For example, in the case of the linear regression model, under the assumption of normal errors, we have seen that the (optimal) density forecast (f_t)

[29] The expressions for density and interval forecasts should be considered as only approximate, as typically there are finite estimation samples or non normal errors.

can be express through Equation (10) with

$$y_{T+h} \sim N(\hat{y}_{T+h}, V(e_{T+h})),$$

where $\hat{y}_{T+h} = X_{T+h}\hat{\beta}$ and $V(e_{T+h})$ indicates the variance of the forecast error. The true density (g_t) is instead

$$y_{T+h} \sim N(X_{T+h}\beta, \sigma_\epsilon^2).$$

For the evaluation of point forecasts we just compare the forecast and actual val-ues, for the density forecasts we must instead compare the entire forecast and actual densities, or the corresponding CDFs, which makes the evaluation more complex.

It is convenient to introduce the *Probability Integral Transformation (PIT)*, defined as

$$(PIT) \equiv p_t = F_t(x_t), \tag{13}$$

where x_t denotes the forecast value. In practice, the PIT associates to each possible forecast value x_t its probability computed according to the density forecast f_t.

It can be shown (see Diebold et al. (1998)) that if $F_t = G_t$, $t = 1, 2, ...,$ then the p_ts are independent $U[0, 1]$ variables, where U denotes the uniform distribution. Therefore, to assess the quality of density forecasts we can check whether their associated $PIT s$ are independent and uniformly distributed.

Uniformity (typically defined probabilistic calibration) can be evaluated qualitatively, by plotting the histogram of the p_ts for the available evaluation sample. For a more formal assessment of probabilistic calibration, let us consider the inverse normal transformation:

$$z_t = \Phi^{-1}(p_t), \tag{14}$$

where Φ is the CDF of a standard normal variable. If p_t is $i.i.d. \sim U(0, 1)$ then z_t is $i.i.d. \sim N(0, 1)$. Let us define z_t as PIT-N. For example, in the case of the linear regression model with normal errors considered above, the z_ts are just the (point) forecast errors (e_t) divided by the forecast standard deviations.

It is more convenient to assess probabilistic calibration using z_ts rather than p_ts, since there are many more tests for normality than for uniformity (see e.g. Mitchell and Wallis

(2011) for a list of tests for uniformity and normality).

The combination of independent and uniform PITs (or normal $PIT-N$s) is typically defined *complete calibration*. Several procedures can be used to test for independence. For example, if the z_ts are indeed normally distributed and therefore independence and lack of correlation are equivalent, any test for no correlation in the errors can be used, see, e.g., Mitchell and Wallis (2011) for other procedures.

A related approach to probabilistically evaluate density forecasts is based on likelihood ratio tests, see Berkowitz (2001). Specifically, let us assume the model:

$$z_t - \mu = \rho\left(z_{t-1} - \mu\right) + \epsilon_t,$$

with $\epsilon_t \sim N\left(0, 1\right)$ and, as before, $z_t = \Phi^{-1}\left[\int_{-\infty}^{y_t} f\left(u\right) du\right]$ is the inverse standard normal normal distribution function with $f\left(y_t\right)$ being the probability density of y_t. Then, three alternative likelihood ratio tests can be used:

$$\begin{aligned} LR_1 &= -2(L\left(0, 1, \widehat{\rho}\right) - L\left(\widehat{\mu}, \widehat{\sigma}_\epsilon^2, \widehat{\rho}\right)) \\ LR_2 &= -2(L\left(\widehat{\mu}, \widehat{\sigma}_\epsilon^2, 0\right) - L\left(\widehat{\mu}, \widehat{\sigma}_\epsilon^2, \widehat{\rho}\right)) \\ LR_3 &= -2(L\left(0, 1, 0\right) - L\left(\widehat{\mu}, \widehat{\sigma}_\epsilon^2, \widehat{\rho}\right)) \end{aligned}$$

where $L\left(\cdot\right)$ is the likelihood as a function of the unknown parameters; for more information see Berkowitz (2001). LR_1 tests for the zero mean of the series and unity of the residuals variance. LR_2 tests for independence and LR_3 is the joint test.

While assessing density forecasts is relevant, in practice interval forecasts are more commonly used, so we now turn to their evaluation, which is much simpler. In particular, let us assume that we construct forecast intervals at the 90%, 80%and 60% confidence levels, which correspond, respectively, to the (0.05, 0.95), (0.1, 0.9) and (0.2, 0.6) quantiles. We can then use the coverage rates at the previously mentioned significance levels, which measure the number of times the true values lie inside the specified intervals.

5.3 Directional forecasts

In forecasting models the dependent variable is typically continuous, for example GDP growth or inflation[30]. However, we might be interested in forecasting not the level of the variable but whether it will be positive or negative, or higher or lower than the current value. In this case, the target becomes a binary variable.

Regression models can still be used to forecast binary events: we construct point forecasts for the variable in levels, and transform them into binary forecasts depend-ing on whether the point forecasts are positive or negative, or higher or lower than the current value.

As an alternative, the variable itself can be turned into a binary indicator, e.g., taking value 0 if growth or inflation is negative, and value 1 for positive values. Logit or Probit specifications can be then used to model and forecast the resulting binary variable.

In particular, let us assume that the variable of interest is negative in period t, $R_t = 1$, if the unobservable variable s_t is larger than zero, where the evolution of s_t is assumed to be governed by the following linear model

$$s_t = \beta' y_{t-1} + e_t. \tag{15}$$

Therefore,

$$\Pr(R_t = 1) = \Pr(s_t > 0) = F(\beta' y_{t-1}), \tag{16}$$

where $F(.)$ is either the cumulative normal distribution function (Probit model), or the logistic function (Logit model). The model can be estimated by maximum likelihood, and the estimated parameters combined with current values of the leading indicators y to provide an estimate of the probability of observing a negative value in period $t+1$, i.e.,

$$\widehat{R}_{t+1} = \Pr(R_{t+1} = 1) = F(\widehat{\beta}' y_t). \tag{17}$$

Note that, as in the case of dynamic estimation, a different model specification is required for each forecast horizon. For example, if a h-step ahead prediction is of

[30] See also the Eurostat flash estimates on income and poverty at `https://goo.gl/t9vAJZ`.

interest, the model in (15) should be substituted with

$$s_t = \gamma_h' y_{t-h} + u_{t,h}. \tag{18}$$

This approach typically introduces serial correlation and heteroskedasticity into the error term $u_{t,h}$, so that the Logit specification combined with nonlinear least squares estimation and robust estimation of the standard errors of the parameters can be preferred over standard maximum likelihood estimation. As an alternative, the model in (15) could be complemented with an auxiliary specification for y_t, say,

$$y_t = Ay_{t-1} + v_t \tag{19}$$

so that

$$\Pr(R_{t+h} = 1) = \Pr(s_{t+h} > 0) = \Pr(\beta' A^{h-1} y_t + \eta_{t+h-1} + e_{t+h} > 0) = F_{\eta+e}(\beta' A^{h-1} y_t) \tag{20}$$

with $\eta_{t+h-1} = \beta' v_{t+h-1} + \beta' A v_{t+h-2} + ... + \beta' A^{h-1} v_t$. In general, the derivation of $F_{\eta+e}(.)$ is quite complicated, and the specification of the auxiliary model for y_t can introduce additional noise, so that the direct approach seems empirically preferable.

Finally, the estimated probability of a negative value, $\widehat{R}_{t+1}$ or $\widehat{R}_{t+h}$, should be transformed into a 0/1 variable using a proper rule. The common choices are of the type $\widehat{R}_t \geq c$ where c is typically 0.5.

5.4 Evaluation of directional forecasts

When the target variable, R_t, is a binary indicator while the (out of sample) forecast is a probability of recession, $\widehat{R}_t$, evaluation criteria similar to the MSFE can be introduced. Specifically, Diebold and Rudebusch (1989) defined the accuracy of the forecast as

$$QPS = \frac{1}{T} \sum_{t=1}^{T} 2(R_t - \widehat{R}_t)^2, \tag{21}$$

where QPS stands for quadratic probability score. The range of QPS is $[0, 2]$, with 0 for perfect accuracy.

A similar loss function that assigns more weight to larger forecast errors is the log probability

score,

$$LPS = -\frac{1}{T}\sum_{t=1}^{T}\left((1-R_t)\log(1-\widehat{R}_t) + R_t \log \widehat{R}_t\right). \tag{22}$$

The range of LPS is $[0, \infty]$, with 0 for perfect accuracy.

A third option, is the Sign Success Ratio (SSR) defined as:

$$SSR_{i,h} = \frac{\sum_{j=1}^{Eval} I\left(\text{Sign}\left(\widehat{y}^{f}_{i,T_j+h}\right)\right)}{Eval}, \tag{23}$$

for h is the forecast horizon, $I\left(\text{Sign}\left(\widehat{y}^{f}_{i,T_j+h}\right)\right)$ is an indicator function that receives and the value 1 if $\text{Sign}\left(\widehat{y}^{f}_{i,T_j+h}\right)=\text{Sign}\left(y_{i,T_{j+h}}\right)$ 0 otherwise, and $Eval$ indicates the number of the evaluation periods. A large SSR indicates that the specified model correctly predicts the "direction" of the target during the cross-validation period. This statistic is a percentage and, as such, it is easier to interpret and communicate than QPS or LPS. For example, a 95% SSR indicates that the underlying model has correctly predicted the direction of the target over 95% of the evaluation period.

Of course, the drawback of SSR is that it does not account for the level of the point forecast. For example, we might have a model which correctly predicts the direction 95% or even 100% of the times but it is far away from the actual values. Therefore, SSR should be used as a complement to the standard forecast error statistics.

5.5 Empirical Gains

Overall, the results for GDP are rather heterogeneous in terms of best forecasting procedures, but sequential selection and averaging of the best models over a training simple often delivers good directional and interval forecasts.

Models with the big data based uncertainty indicators have a mixed performance, as well as those based on large macro and financial information sets. Simple univari-ate models tend to do well in terms of directional forecasts, but often produce too wide interval forecasts, which contain 100% of the realizations rather than only 60% of them.

As an illustration, we continue the example with the UK HICP. The best performing mo-

dels in terms of directional accuracy for this variable are linear regression using Google uncertainty and 1 to 3 lags of the target, with an SSR of 70.59%, and also LASSO using Macro, Financial and big data uncertainty indexes; see Figure 2.

Figure 2: UK, Google with first lag of target, 60% Intervals (dark blue), 90% Intervals (light blue)

5.6 Data Uncertainty & Metrics Robustness

Statistical agencies pay attention to data uncertainty measures. Various metrics published by these agencies are based on the statistical analysis of data that are prone to errors, outliers or might be incomplete, which requires some form of pre-filtering and data adjustment. Big data can also include errors or measurement inaccuracies. This might not always be a problem when the big data are aggregated and the magnitude of the observations is much larger than that of the errors. However, even in this case a careful check is needed before the statistics are made public.

A key feature to improve nowcasting and flash estimation, and therefore the associated evaluation metrics, lies in the identification of the sources of data errors. Measures of data spread, e.g., the interquartile range, variance, etc. are well known to provide important information regarding the distribution of the underlying data, which could be used by the researcher prior to the forecasting experiment. However, bias in the results is still not eliminated.

Iaccarino and Buono (2017) and Iaccarino, Frankel, Sharp and Buono (2017) suggest the use of global sensitivity metrics to identify and rank potential sources of uncertainties, which explain the inaccuracy of estimates. In particular, the Sobol indices and the active subspace variables can be used as tools to describe the variance errors.

In the context of macroeconomic nowcasting and flash estimation, this idea could be used to identify the score of uncertainty in each variable, which could be later further used in the forecasting context. In particular, there are two ways a researcher can apply this procedure empirically:

- Apply the uncertainty scores first, rank the variables and identify a subset of all variables which has low uncertainty. Then, produce the nowcasts from various models based on this subset only.

- Alternatively, one could first estimate a model, e.g. by LASSO, using the full dataset. Then, one produces the nowcast using the variables both as indicated by the model and using a smaller subset of selected variables ranked according to their uncertainty score. Finally, a comparison of the two nowcasts can provide information on whether uncertainty scores can improve the nowcasting performance when used as a filter.

6 CONCLUSIONS AND OVERALL RECOMMENDATIONS

This paper provides a primer for applied researchers of Official Statistical Agencies and similar institutions who aim to exploit big data for macroeconomic nowcasting and the production of early estimates.

We have developed a typology of big data and discussed various ways to move from unstructured big data to time series. We have also introduced ways the re-searcher can deal with outliers before and after the transformation. We have consid-ered a variety of econometric methods suited for large (though not huge) datasets, and implemented them in an empirical nowcasting exercise for key economic variables for the four largest European countries. The exercise has shown the timeliness gains that can be obtained by adding big data based indicators to the usual set of explanatory variables. Particular series, such as consumer prices and unemployment, tend to benefit more compared to industrial production. We have also discussed stan-dard evaluation measures, and some extensions related to density and directional forecasting.

We believe our analysis has highlighted the potential benefits associated with the use of big data. However, there are also costs that should be considered, for example, for data collection, storage and handling, and there are also potential issues in terms of data quality, confidentiality, and reliability of provision. Overall, our suggestion is to take a pragmatic approach that balances potential gains and costs from the use of big data for nowcasting macroeconomic indicators, in addition to standard indicators. The following step-wise procedure could be helpful in the assessment of the expected net gains from the use of big data in a nowcasting context.

A preliminary step should be an a priori assessment of the potential usefulness of big data for a specific indicator of interest, such as GDP growth, inflation or unem-ployment, for a specific country or region. This requires to evaluate the quality of the existing nowcasts and whether any identified problems, such as bias or inefficiency or large errors in specific periods, can be fixed by adding information as potentially available in Big Data based indicators. Similarly, it should be considered whether these additional indicators could improve the timeliness, frequency of release and extent of revision of the nowcasts. Relevant information can be gathered by looking at existing empirical studies focusing on similar variables or countries.

Once big data passes the “need check” in the preliminary step, the first proper step of the big data based nowcasting exercise is a careful search for the specific big data to be collected. As we have mentioned, there are many potential providers, which can be grouped into Social Networks, Traditional Business Systems, and the Internet of Things. Naturally, it is not possible to give general guidelines on a preferred data source, as its choice is heavily dependent on the target indicator of the nowcasting exercise.

Having identified the preferred source of big data, the second step requires to assess the availability and quality of the data. A relevant issue is whether direct data collection is needed, which can be very costly, or a provider makes the data available. In case a provider is available, its reliability (and cost) should be assessed, together with the availability of meta data, the likelihood that continuity of data provision is guaranteed, and the possibility of customization (e.g., make the data available at higher frequency, with a particular disaggregation, for a longer sample, etc.). All these aspects are particularly relevant in the context of applications in fficial statistical offices. As the specific goal is nowcasting, it should be also carefully checked that the temporal dimension of the big data is long and homogeneous enough to allow for proper model estimation and evaluation of the resulting nowcasts.

The third step requires an analysis of specific features of the collected Big Data. A first issue that is sometimes neglected is the amount of the required storage space and the associated need of specific hardware and software for storing and handling the big data. A second issue is the type of the big data, as it is often unstructured and may require a transformation into cross-sectional or time series observations. Even when already available in numerical format, pre-treatment of the big data is often needed to remove deterministic patterns and deal with data irregularities, such as outliers and missing observations. While standard methods can be usually applied, the size of the datasets suggests to resort to robust and computationally simple approaches, applied variable by variable.

The fourth step requires to assess the presence of a possible bias in the answers provided by the big data, due to the “digital divide” or the tendency of individuals and businesses not to report truthfully their experiences, assessments and opinions. A related problem, particularly relevant for nowcasting, is the possible instability of the relationship with the target variable. This is a common problem also with standard indicators, as the type and size

of economic shocks that hit the economy vary over time. Both issues can be however tackled at the modelling and evaluation stages.

The fifth step when nowcasting with big data requires to select the proper econometric technique. Here, it is important to be systematic about the correspondence between the nature of the big data setting and use under investigation and the method that is used. There is a number of dimensions along which we wish to differentiate. The first choice is between the use of methods suited for large but not huge datasets, and therefore applied to summaries of the big data (such as Google Trends, commonly used in nowcasting applications), or of techniques specifically designed for big data. For example, nowcasting with large datasets can be based on factor models, large BVARs, or shrinkage regressions. Huge datasets can be handled by sparse principal components, linear models combined with heuristic optimization, or a variety of machine learning methods (which, though, are generally developed assuming i.i.d. variables). As we have seen, it is difficult to provide an a priori ranking of all these techniques and there are few empirical comparisons and even fewer in a now-casting context, so that it may be appropriate to apply and compare a few of them for nowcasting the specific indicator of interest. A second dimension is the frequency of the available data. If this frequency is mixed then specific techniques for mixed frequency data become relevant. Chief among them is unrestricted MIDAS which provides a very flexible framework of analysis and can be adapted to work together with most if not all big data methods be they machine learning of econometric. Yet another dimension relates to the purpose for which large datasets are considered. Possibilities include model or indicator selection, forecasting or a more structural analysis. In this case of course each purpose is best served by different methods and the choice of method crucially depends on the purpose. Most methods can be used for forecasting and so the choice has to be case dependent. We recommend that as many methods as possible are evaluated in a forecasting context although past experience suggests that factor analysis and shrinkage methods can be of great use. For model or indicator selection penalised regression and the Multiple Testing methods seem to be appropriate and also have been reported to have good potential. Finally, for more structural analysis it is clear that it is likely that huge datasets are more difficult to accommodate. In this case, system methods that analyse the whole or a large proportion of the available data simultaneously, seem necessary for a satisfactory analytical outcome. Bayesian VAR models stand out as an appropriate method in this context.

The final step consists of a critical and comprehensive assessment of the contri-bution of big data for nowcasting the indicator of interest. In order to avoid, or at least reduce the extent of, data and model snooping, a cross-validation approach should be followed, whereby various models and indicators are estimated over a first sample and they are selected and/or pooled according to their performance, but then the performance of the preferred approaches is re-evaluated over a second sample. This procedure, which we have implemented in the empirical evaluation, provides a reliable assessment of the gains in terms of enhanced nowcasting performance and timeliness from the use of big data.

To conclude, we are very confident that big data are precious also in a nowcasting and early estimation context, not only to reduce the errors but also to improve the precision, directional accuracy, timeliness, frequency of release and extent of further revisions. We hope that this primer may be useful for many users willing to experiment with this fascinating approach.

7 REFERENCES

1. Abramovic, G. (2014). "15 Mind-Blowing Stats About Online Shopping", CMO.com, available at: https://goo.gl/xNZvoE.

2. Acemoglu, D., Hassan, T.A., Tahoun, A. (2014). "The Power of the Street: Evidence from Egypt's Arab Spring". NBER Working Paper No. 20665.

3. Ailon, N., Chazelle, B. (2006). "Approximate Nearest Neighborhood and the Fast Johnson-Lindenstrauss Transform". Proceedings of the 38st Annual Symposium on the Theory of Computing(STOC), 557-563.

4. Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L. (2011). 11Short-Term Forecasts of Euro Area GDP Growth", The Econometrics Journal, 14(1), C25-C44.

5. Andreou, E., Ghysels, E., Kourtellos, A. (2015). "Should Macroeconomic Forecasters Use Daily Financial Data and How?", Journal of Business & Economic Statistics, 31(2), 240-251.

6. Aprigliano, V., Ardizzi, G., Monteforte, L. (2016). "Using the payment system data to forecast the Italian GDP", Bank of Italy, Working Paper.

7. Askitas, N., Zimmermann, K. F. (2009). "Google Econometrics and Unemployment Forecasting". Applied Economics Quarterly, 55(2), 107-120.

8. Assimakopoulos, V., Nikolopoulos, K. (2000). "The theta model: a decomposition approach to forecasting." International Journal of Forecasting, 16, 521-530.

9. Bacchini, F., Bontempi, M.E., Golinelli, R., Jona-Lasinio, C. (2017). "Short-and long-run heterogeneous investment dynamics". Empirical Economics, DOI: 10.1007/ s00181-016-1211-4.

10. Baker, S.R., Bloom, N., Davis, S.J. (2016). "Measuring Economic Policy Uncertainty". The Quarterly Journal of Economics, 131(4), 1593-1636.

11. Banbura M,. Giannone D., Reichlin L. (2011). “Nowcasting”. In Oxford Handbook on Economic Forecasting, Clements MP, Hendry DF (eds). Oxford University Press: Oxford.

12. Banbura, M., Runstler, G. (2011). “A Look into the Factor Model Black Box: Publication Lags and the Role of Hard and Soft data in Forecasting GDP”, International Journal of Forecasting, 27, 333-346.

13. Barnett, W.A., Chauvet, M., Leiva-Leon, D., Su, L. (2016) ”Nowcasting nominal gdp with the credit-card augmented Divisia monetary aggregates”. MPRA Paper No. 73246.

14. Bergmeir, C., Hyndman, R.J., Benitez, J.M. (2016). “Bagging Exponential Smoothing Methods using STL Decomposition and Box-Cox Transformation.” International Journal of Forecasting, 32, 303-312.

15. Berkowitz, J. (2001). “Testing Density Forecasts, With Applications to Risk Management”, Journal of Business & Economic Statistics, 19, 465-474.

16. Boettcher, I. (2015). “Automatic Data Collection on the Internet (Web Scrap-ing)”, New Techniques and Technologies for Statistics, Eurostat Conference, 9-13 March 2015.

17. Buono, D., Mazzi, G.L., Kapetanios, G., Marcellino, M., Papailias, F. (2017). “big data Types for Macroeconomic Nowcasting”, EURONA – Eurostat Review on National Accounts and Macroeconomic Indicators, 93-145.

18. Capgemini, BNP Paribas (2016). “World Payments Report”. Available online at https:// goo.gl/niAklI.

19. Carlsen, M., Storgaard, P.E. (2010). “Dankort payments as a timely indicator of retail sales in Denmark”. Danmarks Nationalbank, Working Paper 2010-66.

20. Cavallo, A. (2013). “Online and official price indexes: Measuring Argentina’s inflation”. Journal of Monetary Economics, 60, 152-165.

21. Cavallo, A. (2016). “Scraped Data and Sticky Prices”. Review of Economics & Statistics, Forthcoming.

22. Cavallo, A. (2017). “Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers”. The American Economic Review, 107(1), 283-303.

23. Cavallo, A., Rigobon, R. (2016). “The Billion Prices Project: Using Online Prices for Measurement and Research”. The Journal of Economic Perspectives, 30(2), 151-178.

24. Chi, F., Yang, N. (2010). “Twitter Adoption in Congress”. SSRN Working Paper. Available at https://goo.gl/lub2Lr.

25. Choi, H. Varian, H. (2009). “Predicting initial claims for unemployment ben-efits”, Google Working Paper.

26. Choi, H. Varian, H. (2012). “Predicting the Present with Google Trends”. Economic Record, 88(1), 2-9.

27. Da, Z., Engelberg, J., Gao, P. (2011). “In Search of Attention”. Journal of Finance, 66(5), 1461-1499.

28. D’Amico, S., Orphanides, A. (2008). “Uncertainty and Disagreement in Eco-nomic Forecasting”. Federal Reserve Board Finance and Economics Discussion Series 2008–56.

29. D’Amuri, F., Marcucci, J. (2012). “The Predictive Power of Google Searches in Predicting Unemployment”. Banca d’Italia Working Paper, 891.

30. De Livera, A.M., Hyndman, R.J., Snyder, R. D. (2011). “Forecasting time series with complex seasonal patterns using exponential smoothing.” Journal of the American Statistical Association, 106(496), 1513-1527.

31. Deloitte (2012). “What is the impact of mobile telephoy on economic growth?A report for the GSM association”. November 2012.

32. Diebold, F.X., Mariano, R.S. (1995). Comparing Predictive Accuracy. Journal of Business & Economics Statistics, 13(3), 253-263.

33. Diebold, F.X., Gunther, T.A., Tay, A.S. (1998). “Evaluating Density Forecasts with Applications to Risk Management”, International Economic Review, 39(4), 863-883.

34. Doornik, J. A., Hendry, D. F. (2015). "Statistical Model Selection with big data". Cogent Economics & Finance, 3(1), 2015.

35. Drineas, P., Mahoney, M.W., Muthukrishnan, S. (2008). "Relative Error CUR Matrix Decompositions", Siam Journal of Matrix Analysis and Applicatons,(30), 844–811.

36. Duarte, C., Rodrigues, P.M.M., Rua, A. (2016). "A mixed frequency approach to forecast private consumption with ATM/POS data". Banco de Portugal, Working Paper 1-2016.

37. Eckley, P. (2015). "Measuring economic uncertainty using news-media textual data". MPRA Paper No. 69784.

38. Ericsson, N.R. (2015). "Eliciting GDP Forecasts from the FOMC's Minutes Around the Financial Crisis". International Finance Discussion Papers, Board of Governors of the Federal Reserve System, Working Paper 1152.

39. Ericsson, N.R. (2016). "Predicting Fed Forecasts". IFDP Notes, Board of Governors of the Federal Reserve System, February 12, 2016. Available at: https://goo.gl/nOl77h.

40. Esteves, P.S. (2009). "Are ATM/POS data relevant when nowcasting private consumption?". Banco de Portugal, Working Paper 25-2009.

41. Foroni, C., Marcellino, M., Schumacher, C. (2015). "Unrestricted Mixed Data Sampling (MIDAS): MIDAS Regressions with Unrestricted Lag Polynomials". Journal of the Royal Statistical Society: Series A, 178(1), 57-82.

42. Frey, B.J., Dueck, D. (2007). "Clustering by passing messages between data points". Science, 315, 972-976.

43. Galbraith, J.W., Tkacz, G. (2007). "Analyzing Economic Effects of Extreme Events using Debit and Payments System Data". CIRANO Scientific Series, Working Paper 2011s-70.

44. Galbraith, J.W., Tkacz, G. (2011). "Electronic Transactions as High-Frequency Indicators of Economic Activity". Bank of Canada, Working Paper 2007-58.

45. Galbraith, J.W., Tkacz, G. (2015). "Nowcasting GDP with electronic payments data". European Central Bank, Working Paper No 10 / August 2015.

46. Gandomi, A., Haider, M. (2015). "Beyond the hype: Big data concepts, methods, and analytics". International Journal of Information Management, 35, 134-144.

47. Ghysels, E., Santa-Clara, P., Valkanov, R. (2004). "The MIDAS Touch: Mixed Data Sampling Regression Models", CIRANO Working Paper, 2004s-20.

48. Giannone, D., Reichlin, L., Small, D. (2008). "Nowcasting: The Real-Time Informational Content of Macroeconomic Data", Journal of Monetary Economics, 55, 665-676.

49. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L. (2009). "Detecting influenza epidemics using search engine query data". Nature, 457, 1012-1014.

50. Graham, D. (2016). "How the Internet of Things Changes big data Analytics". DataInformed, published online: August 9, 2016. Available at https://goo. gl/M22uje.

51. Hartford, T. (2014). "Big data: Are we making a big mistake?". Financial Times, March, 28, 2014. Available at: https://goo.gl/SMOL2L.

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

53. Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S. (2002). "A state space framework for automatic forecasting using exponential smoothing methods." International Journal of Forecasting, 18(3), 439–454.

54 Iaccarino, G., Buono, D. (2017). "Reporting Uncertainties - Too Much Infor-mation?", 61st ISI World Statistics Congress ISI2017, Marrakech, submitted paper.

55. Iaccarino, G., Frankel, A., Sharp, D., Buono, D. (2017). "Reporting Uncertainties - Too Much Information?", 61st ISI World Statistics Congress ISI2017, Marrakech, presentation.

56. Kapetanios, G., Marcellino, M., Papailias, F. (2017a). “big data Conversion Techniques including their Main Features and Characteristics”. Eurostat Sta-tistical Working Papers, 2017. Available at: https://goo.gl/X4P71T.

57. Kapetanios, G., Marcellino, M., Papailias, F. (2017b). “Filtering techniques for big data and big data based uncertainty indexes”. Eurostat Statistical Working Papers, 2017. Available at: https://goo.gl/UoAeVk.

58. Koop, G., Onorante, L. (2013). “Macroeconomic Nowcasting Using Google Probabilities”. European Central Bank Presentation.

59. Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014). “The Parable of Google Flu: Traps in big data Analysis”. Science, 143, 1203-1205.

60. Levenberg, A., Pulman, S., Moilanen, K., Simpson, E., Roberts, S. (2014). “Predicting Economic Indicators from Web Text Using Sentiment Composition”. International Journal of Computer and Communication Engineering, 3(2), 109-115.

61. Lloyd, S. P. (1982). “Least squares quantization in PCM”. Information Theory, IEEE Transactions, 28(2), 129-137.

62. Lunnemann, P., Wintr, L. (2011). “Price Stickiness in the US and Europe Revisited: Evidence from Internet Prices”. Oxford Bulletin of Economics & Statistics, 73(5), 0305-9049.

63. McCracken, M.W, Ng, S. (2015) FRED-MD: A Monthly Database for Macroeconomic Research, *Research Division, Federal Reserve Bank of St. Louis, Working Paper Series*, Working Paper 2015-012B.

64. Metcalfe, E., Flower, T., Lewis, T., Mayhew, M., Rowland, E. (2016). “Research indices using web scraped price data: clustering large datasets into price indices (CLIP)”. Office for National Statistics, Release date: 30 November 2016

65. Mitchell, J., Wallis, K.F., (2011), ”Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness”, Journal of Applied Econometrics 26, 1023–1040.

66. Moat, H.S., Curme, C., Avakian, A., Kenett, D.Y., Stanley, H.E., Preis, T.(2013). “Quantifying Wikipedia Usage Patterns Before Stock Market Moves”. Scientific Reports, 3:1801, 1-5.

67. Modugno, M. (2013). “Nowcasting Inflation using High-Frequency Data”,International Journal of Forecasting, 29, 664-675.

68. Ng, S. (2016). “Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data”. Working Paper.

69. Rigobon, R. (2015). “Presidential Address: Macroeconomics and Online Prices”. Economia, 15(2), 199-213.

70. Roussas, G. (1996). “Exponential Probability Inequalities with Some Applications”. Statistics, Probability & Game Theory, IMS Lecture Notes – Mono-graph Series, 30.

71. Sarlos, T. (2006). “Improved Approximation Algorithms for Large Matrices via Random Projections”. Proceedings of the 47 IEEE Symposium on Foundations of Computer Science.

72. Schumaker, R.P., Chen, H. (2006). “Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System”. Working Paper.

73. Stekler, H.O., Symington, H. (2016). “Evaluating Qualitative Forecasts: The FOMC Minutes, 2006-2010”. International Journal of Forecasting, 32(2), 559-570

74. Stock, J., Watson, M. (2002a). “Forecasting Using Principal Components from a Large Number of Predictors”, Journal of the American Statistical Associa-tion, 297, 1167-1179.

75. Stock, J., Watson, M. (2002b). “Macroeconomic Forecasting using Diffusion Indexes”, Journal of Business & Economics Statistics, 20, 147-162.

76. Tkacz, G. (2013). “Predicting Recessions in Real-Time: Mining Google Trends and Electronic Payments Data for Clues”. C.D. HOWE Institute, Commentary No. 387, Financial Services, September 2013.

77. Thorsrud, L.A. (2016). “Words are the new numbers: A newsy coincident index of business cycles”. Norges Bank Working Paper Series, Working Paper 21-2016.

78. Varian, H., Stephen-Davidowitz, S. (2014). “Google Trends: A primer for social scientists”. Google Working Paper.

79. Venkatasubramanian, S., Wang, Q. (2011). “The Johnson-Lindenstrauss Trans-form: An Empirical Study”. Proceedings of the Thirteenth Workshop on Al-gorithm Engineering and Experiments, 148–173.

80. Yuan, Q., Nsoessie, E.O., Lv, B., Peng, G., Chunara, R., Brownstein, J.S.(2013). “Monitoring Influenza Epidemics in China with Search Query from Baidu”. PLoS ONE 8(5), e64323.