4

Big data types for macroeconomic nowcasting

DARIO BUONO [1], GIAN LUIGI MAZZI [2], GEORGE KAPETANIOS [3], MASSIMILIANO MARCELLINO [4] AND FOTIS PAPAILIAS [5]

Abstract: In this paper we present a detailed discussion on various types of big data which can be useful in macroeconomic nowcasting. In particular, we review the big data sources, availability, specific characteristics and their use in the literature. We conclude this paper identifying the big data types which could be adopted for real applications.

JEL codes: C82, E27, E37, E47

Keywords: big data, nowcasting, data features

[1] Eurostat.
[2] retired from Eurostat.
[3] King's College London.
[4] Università Bocconi, Italy.
[5] Queen's University Management School, Queen's University Belfast, UK and quantf research.

1. Introduction

The 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, and there is nowadays an ever growing interest in the analysis of these big data. In this paper, we summarize the results of an investigation of the specificities of various big data sources relevant for macroeconomic nowcasting and early estimates, and we discuss the characteristics of big data sources that are relevant for their application in this context.

Perhaps not surprisingly given their different origins, the literature provides various definitions of big data. One possibility to obtain a general classification is to adopt the '4 Vs' classification, originated by the IBM, which relates to: (i) Volume (Scale of data), (ii) Velocity (Analysis of streaming data), (iii) Variety (Different forms of data) and (iv) Veracity (Uncertainty of data). However, this classification seems too general to guide empirical nowcasting applications.

A second option is to focus on numerical data only, which can either be the original big data or the result of a transformation of unstructured data. Once data have been transformed, and following, e.g. Doornik and Hendry (2015), we can distinguish three main types of big data. 'Tall' datasets include not so many variables, *N*, but many observations, *T*, with *T* >> *N*. This is for example the case with tick by tick data on selected financial transactions or search queries. 'Fat' datasets have instead many variables, but not so many observations, *N* >> *T*. Large cross-sectional databases fall into this category, which is not so interesting from an economic nowcasting point of view, unless either *T* is also large enough or the variables are homogeneous enough to allow proper model estimation (e.g., by means of panel methods) and nowcast evaluation. Finally, 'Huge' datasets, with very large *N* and *T*, are the most interesting type of data in a nowcasting context even if, unfortunately, they are not so often available for economic variables. However, 'complexity' can help in expanding the *N* dimension, for example allowing for dynamics (lead-lag relations), non linearities and the micro-structure of information (calendar issues for example).

A third possibility to classify big data is to resort to some sort of official definition. A particularly useful taxonomy is provided by the statistics division of the United Nations Economic Commission for Europe (UNECE), which identifies big data according to their source.

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 information). 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 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 include 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 scientific 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 macroeconomics there have so far been limited applications, mostly concentrated in the areas of nowcasting/forecasting, on which we also focus.

A first aim of this paper is to review the main existing macroeconomic applications and, in particular, to provide a discussion of various big data types which could be useful in macroeconomic nowcasting. A survey of the econometric methods for big data is instead left for future research.

We have identified ten categories of big data which are organised as: (i) financial markets data, (ii) electronic payments data, (iii) mobile phones data, (iv) sensor data, (v) satellite images data, (vi) scanner prices data, (vii) online prices data, (viii) online search data, (ix) textual data, and (x) social media data. For each of these categories, in Section 2, we analyse the specific characteristics of the data, review the relevant macroeconomic papers based on them, and provide details regarding data sources and availability.

Next, in Section 3, we consider the dual problem: nowcasting a specific macroeconomic variable of interest, such as gross domestic product (GDP) growth, inflation or unemployment, using big data. A summary table, reported at the end of the Section, lists the main studies, providing details on the specific type of big data adopted.

In Section 4, we propose a further classification, specifically designed for numerical big data and based on their relative cross-sectional and temporal dimensions, which has implications for the proper data pre-treatment and required econometric techniques to analyse them. As mentioned above, big data are now split into three main categories: Fat, Tall, and Huge.

The final goal of the paper is to use the extensive review of existing studies and data to provide a description of the main features of big data that make them useful for nowcasting and flash estimation, and an indication of which big data categories seem particularly promising for nowcasting specific variables and for further analysis. We deal with these issues in Section 5.

Section 6 summarizes the main findings and proposes a summary classification of big data, based on a set of key features that emerge from the detailed review: source, provider, availability, continuity, type, size and sample, meta data, feature, frequency, pre-treatment, link with target, previous use, and required econometrics. This classification is finally used to create a grid for assessing the need and potential usefulness of big data for macroeconomic nowcasting and forecasting, in general and in official institutions such as Eurostat, according to the outcome of a careful cost-benefit analysis.

2. Types of big data for macroeconomic nowcasting

This section is divided into subsections, each corresponding to a big data category which could be useful for macroeconomic nowcasting. To facilitate the reading process, and in order to provide a typology of big data, we have standardised the format of each subsection as follows. At first we describe the big data and discuss its characteristics. Then, we review the relevant literature which currently uses the specific type of big data in nowcasting/forecasting or other applications in economics. Finally, we present a detailed list of data sources along with their availability.

2.1. Financial markets data

Advances in computer technology and data storage has allowed for the collection and analysis of high-frequency financial data. The most widely observed forms of financial big data are the trades and quotes. New York Stock Exchange (NYSE) initiated the collection of this data in 1992. This intraday data potentially provides detailed information which could be used in the analysis of markets efficiency, volatility, liquidity as well as price discovery and expectations. Central banks 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;
- and generally all operations in financial markets.

The easiest form of the above data to obtain is intraday trades and quotes. However, this data is available for each security individually; therefore if we are interested in the construction of time series index based on intraday data, we would have to consider a large number of securities. Consequently, this increases the total cost for this data. Furthermore, data which refers to the monitoring of all market operations is very sensitive and only central banks and other regulatory bodies have access to, 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. We consider financial data as a major data source, therefore it forms the basis of our explanatory data used in the empirical application in future. Our main aim is to examine whether one of the other nine types of big data can improve the macroeconomic nowcasting and result in improved estimates.

2.1.1. APPLICATIONS

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).

2.1.2. INDICATIVE DATA SOURCES & EXAMPLES

Although most of the historical high frequency data is proprietary, end of day transaction prices (along with the daily open, high and low) are publicly available for most equities in major stock exchanges. Below, we provide an indicative list of data sources as well as live feeds which could be used to collect intraday data:

- Daily equity prices: Google Finance, Yahoo Finance, St. Louis FRED for major US indexes, etc. are some free online sources;
- Historical Intraday data: Kibot, PiTrading and Intradata provide proprietary historical intraday data for assets in major markets. Unstructured intraday data could be converted to structured time series format by constructing daily realised volatility estimators which could be included in the nowcasting exercise as risk indicators;
- Live feeds: Live-Rates.com, knoema and Oanda provide live feeds for most of the financial instruments for free. These feeds could be used as a source for a data collection which could then be used in an aggregated way in nowcasting.

2.1.3. CONCLUSIONS

Financial transactions data is a very important source of information for the analysis of markets as well as the forecasting of the economy. Given their high-frequency nature, financial markets illustrate a very fast discount of news, much faster than lower frequency macroeconomic indicators. Therefore, financial markets data is an essential source of information in macroeconomic nowcasting. We will investigate whether such data can be combined with another source of big data to provide more accurate nowcasts and short-term forecasts. Overall, if the researcher is interested in risk sentiment indexes or the construction of policy uncertainty indexes, realised volatility estimators might prove a more accurate and timely instrument compared to the standard measures, say VIX in the US.

2.2. 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) electronic bill payments in online stores, 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 follow the CEMEA markets with cards usage at 77 % of the non-cash transactions. Japan, Australia, Singapore and

South Korea markets 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 is the main form of non-cash payments. Card payments include online as well as offline POS [6] purchases making them very useful in the tracking of consumer behaviour and retail sales (among others).

Table 1: 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. We have 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.

2.2.1. APPLICATIONS

The literature in economics which uses credit cards data started very recently. Galbraith and Tkacz (2007) is one of the first papers which published the results of cards data in macroeconomics. In particular, they use Canadian debit card transactions [7] in order to provide real-time estimates of economic activity. Their predictive regression analysis provides information for 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 analysis 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 consumption (especially non-durable) growth can be partly explained by cards data.

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

[6] Point of Sale (POS). For example, card payment at a retail store.
[7] Obtained via the Canadian interbank network, Interac.

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 (2015) tackle directly with 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 X-11 methodology ([8]) in order to clean the data. Their main finding is that nowcasting using high frequency electronic payments improve by 65 % between the first and final estimates presenting supporting evidence in 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 augmented 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 nowcasting of the GDP compared to a benchmark model.

Finally, Aprigliano, Ardizzi and Monteforte (2016) use a mixed frequency dynamic 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 improves nowcasting. In a separate screening exercise using the Least Absolute Shrinkage and Selection Operator (LASSO), payment system variables are indicated as potential predictors of GDP growth.

2.2.2. INDICATIVE DATA SOURCES & EXAMPLES

Based on the literature review and an extensive online search, we discuss potential data sources as well as availability for cards data.

- At first we have all credit and debit card financial services corporations. These companies facilitate electronic funds transfers throughout the world. We briefly mention them here.
    - Visa, Inc. It is not known if VISA provides aggregated data to third-parties,
    - American Express. American Express is known to sell anonymised data to third-party marketing companies ([9]). American Express also offers anonymised data to their business partners as a way to help them analyse and promote their products ([10]),
    - Mastercard. Mastercard, along with American Express, is known to sell anonymised data to third-party marketing companies ([11]);

[8] See Kapetanios, Marcellino and Papailias (2016) for a full discussion.
[9] See online news articles at https://goo.gl/Co3oc6 and https://goo.gl/yrj74b.
[10] See online news article at https://goo.gl/mS9J0H.
[11] See online news article at https://goo.gl/Co3oc6 .

- Another way to acquire electronic payments data is through interbank networks. Below we summarise the interbank networks for major economies.
  - Australia: Electronic Funds Transfer at Point Of Sale (EFTPOS). Data is not publicly available. In 2010, ANZ BANK has announced the launch of an online tool that uses aggregated data from merchant EFTPOS transactions to illustrate estimated sales patterns, market share, turnover and to provide insights into customer behaviour. However, this is available to ANZ business customers only,
  - Canada: Interac Association (as mentioned in the papers above). Not publicly available. However, given the existence of current literature, data could be potentially purchased,
  - China: China Union Pay. Not publicly available and unknown if China Union Pay have an interest in selling anonymised data,
  - France: Groupement des Cartes Bancaires CB. Payment system: EBA Clearing (Euro1) and Trans-European Automated Real-time Gross Settlement Express Transfer System (TARGET2). Not publicly available. However, given the existence of current literature, data could be potentially purchased or come to an agreement,
  - Germany: Girocard. Euro1 and TARGET2. Not publicly available. However, given the existence of current literature, data could be potentially purchased or come to an agreement,
  - Italy: BI-COMP. Euro1 and TARGET2. Not publicly available. However, given the existence of current literature, data could be potentially purchased or come to an agreement,
  - Japan: Yucho. Not publicly available. It is not known if Yucho would be interested in selling anonymised data,
  - Portugal: Multibanco (as mentioned in the papers above). Not publicly available. However, given the existence of current literature, data could be potentially purchased or come to an agreement;
- Third-party data providers, such as Envestnet Yodlee ([12]), could also provide aggregated data on electronic transactions. In most of these cases, the data is proprietary;
- Finally, the last way to collect cards data would be via access to central bank databases which are not publicly available. All central banks could provide such data in aggregated or unstructured format. Some of the central banks which could provide useful data regarding this project are: the European Central Bank (ECB), Banque de France, Deutsche Bundesbank, Banca d'Italia, De Nederlandsche Bank, Banco de Portugal and Banco de Espana. However, Central Banks do not usually provide any data to third-parties.

Some data which is publicly available and might be useful for demonstration purposes, but not for an empirical analysis due to the short sample, include:

(1) UK Corporate credit card transactions 2014-2015 in monthly frequency. This is a structured series based on all transactions on corporate credit cards in the financial year 2014-2015;

(2) Sample dataset of European credit card transactions during two days in September 2013 of Dal Pozzolo, Caelen, Johnson, and Bontempi (2015). The data is publicly available on Kaggle website ([13]). Due to confidentiality issues, no original features are presented apart from the 'Time' of the purchase, the 'Amount' and a flag which indicates if the transaction is genuine

[12] https://goo.gl/eGaloU.
[13] https://goo.gl/G1saSD.

or fraudulent. In total, there are 284,808 transactions. This dataset might be useful in order to illustrate the daily aggregation process, however it is not sufficient to show the weekly seasonal pattern and its subsequent removal and data cleaning.

2.2.3. CONCLUSIONS

Electronic payments include credit and debit card transactions, credit transfers, direct debits, cheques, etc. Based on the nature of the data, it tracks well economic activity and particularly household purchases (via credit and debit card transactions). For this reason it would be very useful in the monitoring, nowcasting and forecasting of retail sales, private consumption, and other related variables. The literature provides empirical evidence in favour of cards data. However, unstructured or aggregated data is not publicly available and most of these studies are provided with confidential data from interbank networks or central banks.

2.3. 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, already provides enough 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 as internet activity, mobile banking, GPS tracking and other sensors data ([14]). Overall, mobile phones data provides detailed information of human behaviour and therefore 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 ([15]).

2.3.1. APPLICATIONS

As in the case of card payments data, the literature which uses mobile phones data is also very recent. Smith-Clarke, Mashhadi and Capra (2014) employ call data in order to examine poverty in two developing countries where survey data is scarce. In particular, they use the aggregated call detail records of mobile phone subscribers and extract features that are strongly correlated with poverty indexes derived from census data. Their data consists of calls between 5 million Orange customers from Cote d' Ivoire and 928,000 calls between customers of an undisclosed network in an undisclosed region. The authors highlight the difficulty of obtaining call details records for other developed and emerging countries.

Deville, Linard, Martin, Gilbert, Stevens, Gaughan, Blondel and Tatem (2014) use mobile phone data for population mapping. They use more than 1 billion mobile phone calls from Portugal and France and show how spatially and temporarily explicit estimations of population densities can be produced at national scales. In the same topic also lies the work of Ricciato, Widhalm, Craglia and Pantisano (2015) who estimate population density distribution from network-based

[14] In that sense, mobile phones data can also be part of sensors data.
[15] See https://goo.gl/MYb21Q, https://goo.gl/e35lQQ and https://goo.gl/Dkdara.

mobile phone data in Europe. In the same context, De Meersman, Seynaeve, Debusschere, Lusyne, Dewitte, Baeyens, Wirthmann, Demunter, Reis and Reuter (2016) assess the quality of mobile phones data to estimate the actual population. They use Belgian mobile phone data from the major network operator, Proximus, and their results are consistent with the 2011 Census.

Mao, Shuai, Ahn and Bollen (2015) also investigate the case of Cote d' Ivoire to analyse the relations between its nation-wide communications network and the socio-economic dynamics of its regional economies. They introduce an indicator to quantify the relative importance of an area on the basis of call records, and show that a region's ratio of in- and out-going calls can predict its income level. Their results demonstrate the potential of mobile communication data to monitor the economic development and social dynamics of low-income developing countries in the absence of extensive econometric and social data.

2.3.2. INDICATIVE DATA SOURCES & EXAMPLES

Below we briefly mention some data sources as well as examples for demonstration.

- Obviously, the network operators are the main mobile phone data source as all information is sent and received through their network. However, the data is not publicly available and, as mentioned in the literature, this is due to regulatory limitations. It is not known if a private network operator company would be interested in selling anonymised data.
- Third-party software developers, such as HelloSpy ([16]), which offer their customers the ability to track their call history. This is proprietary data.

2.3.3. CONCLUSIONS

Mobile phone data also seems to be an interesting source. It could be used in order to provide timely estimates for population and census data, moreover, as we see in Toole *et al* (2015), the analysis of phone calls can provide details about behaviour which could be used for unemployment and economic activity forecasting. But, generally speaking, it seems difficult to find a detailed and long enough dataset to be tested in a nowcasting/forecasting application.

2.4. 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 facilitate the collection 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.

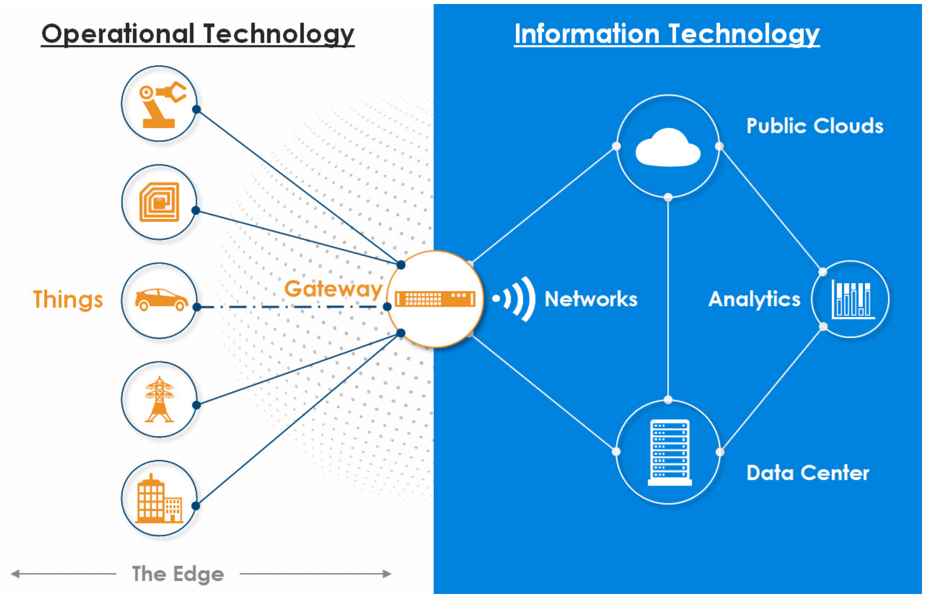
Mobile things, e.g. smart phones, computers, tablets, watches, home appliances, cars, drones, etc., as well as stationary things, e.g. wind turbines, temperature sensors, etc., all come attached with sensors which can send and receive information online. In the this category, we include

[16] https://goo.gl/iGvLJM.

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) ([17]), 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.

Apple and Samsung 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 ([18]). A recently start-up company, ThingSpeak ([19]), provides an open-source platform which allows the user to have his things transmit data to their servers and then offers his data via an API. Mathworks and MATLAB are integrated in the platform which can be used directly for analysis and visualisation of the data.

Figure 1: The internet of things



Source: Graham (2016).

2.4.1. APPLICATIONS

Unfortunately, it seems that there does not exist a literature on economic forecasting or nowcasting based on sensor data. We believe that this is due to the unavailability of data. Most of the literature is based on weather forecasting, where ample data exists, traffic and other geo-related applications. Some indicative sources include Jones (2010), Bazzani, Giovannini, Galotti and Rambaldi (2011), Chan, Khadem, Dillon, Palade, Singh and Cheng (2012), Mass (2012) and Fernandez-Ares, Mora, Arenas, de las Cuevas, Garcia-Sanchez, Romero, Rivas, Castillo and Merelo (2016). Also, Suryadevara, Mukhopadhyay, Wang and Rayudu (2013) forecast human

[17] https://goo.gl/9h3Zf8.
[18] See https://goo.gl/9h3Zf8.
[19] https://goo.gl/z4y9by.

behaviour using wireless sensors data in a smart home. For example, if this could be generalised for households in an economy, sensor data could be used for private consumption nowcasting/forecasting. Papadimitriou, Sun, Faloutos and Yu (2013) review dimensionality reduction and econometric techniques which could be applied in the analysis and filtering of time series sensor streams which could be used based on data availability.

However, IoT will be very useful in the future. According to Manyika, Chui, Bisson, Woetzel, Dobbs, Bughin, and Aharon (2015) IoT will mainly contribute to:

- Factories, e.g. operations management, predictive maintenance;
- Cities, e.g. public safety, traffic control, resource management;
- Human, e.g. improving wellness;
- Retail, e.g. self-checkout, layout optimisation, etc;
- Outside, e.g. logistics routing, autonomous (self-driving) vehicles [20];
- Homes, e.g. energy management, safety and security, chore automation;
- Offices, e.g. organisation redesign and worker monitoring, augmented reality for training.

In the future, IOT related data could become relevant also for nowcasting/forecasting.

For example, sensors could be used to monitor the number of people entering shops or of commercial vehicles moving along specific routes, or the intensity usage of machinery, which could be relevant for variables such as retail sales, exports, or IP.

2.4.2. INDICATIVE DATA SOURCES & EXAMPLES

Below we provide an indicative list of sources and examples. The majority is focused on temperature data and traffic control.

- Open Cities Project, https://goo.gl/XBnYuZ. Open Cities is a project co-founded by the European Union that aims to validate how to approach Open & User Driven Innovation methodologies to the Public Sector in a scenario of Future Internet Services for Smart Cities. Data is not publicly available but given that nature of the funding, data could be made available;
- Array of Things (AoT), https://goo.gl/qajel1. AoT is an urban sensing project, a network of interactive, modular sensor boxes that will be installed around Chicago to collect real-time data on the city's environment, infrastructure, and activity for research and public use. AoT will essentially serve as a 'fitness tracker' for the city, measuring factors that impact liveability in Chicago such as climate, air quality and noise. Visualised data is available online and downloadable nodes data will be available from the City of Chicago Data Portal in early 2017;
- Smart Santander, https://goo.gl/jw0rjU. The project envisions the deployment of 20,000 sensors in Belgrade, Guildford, Lübeck and Santander (12,000), exploiting a large variety of technologies;
- CityPulse Dataset Collection, https://goo.gl/9gBqxs. It includes road traffic data, weather data, cultural, social and library event data and parking data for Aarhus, Surrey and Brasov;

[20] Amazon.co.uk already performed the first drone flight delivery in December 2016 in Cambridge.

- ThingSpeak, https://goo.gl/QXDQS3. Publicly available channels with sensor data and examples. They are supplied by private users, therefore their accuracy is not guaranteed. Most of the data is meteorological.

2.4.3. CONCLUSIONS

As discussed in this section, sensor data and IoT will have a great impact in various business sectors in the next five years or sooner. At this point, there does not exist a literature on sensor data in economic nowcasting and forecasting, however this might change if, for example, retail sales related data becomes available. Currently, there is a big interest in marketing and tailor-made services as well as meteorological applications where sufficient data already exists.

2.5. Satellite images data

Satellite imagery consists of images of Earth or other planets collected by satellites. 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 [21], of course not all of them were imaging satellites.

Satellites images have many applications in meteorology, oceanography, agriculture, forestry, geology, intelligence, warfare and others. Recently, satellite imaging 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 in satellite data handling which eases the use of this data.

2.5.1. APPLICATIONS

The literature has recently started including satellite image data in economic applications, mostly for measuring long-term variables such as poverty, inequality and long-term growth.

Elvidge, Sutton, Ghosh, Tuttle, Baugh, Bhaduri and Bright (2009) use population count and satellite observed lighting in order to build a global poverty map. They claim that this construction of poverty maps is more accurate as it should improve over time through the inclusion of new reference data as improvements are made in the satellite observation of human activities related to economics and technology access.

Henderson, Storeygard and Weil (2011) first put the night lights satellite images in the economic growth context. A scatterplot of the average annual percentage change in GDP and the average annual percentage change in lights indicates a strong and positive relationship. Also, the paper suggests that night lights could be even more useful in developing countries where there is lack of accurate official statistics. This claim is further confirmed in Chen and Nordhaus (2011) who also use luminosity as a proxy for economic output. They find that luminosity has

[21] https://goo.gl/elqcWi.

informational value particularly for countries with low-quality statistical systems or no recent population or economic censuses. In their second paper on this topic, Henderson, Storeygard and Weil (2012) show that lights data can measure growth not only at country level but also for sub- and supranational regions. As an example, their empirical analysis shows that the coastal areas in sub-Saharan Africa are growing slower than the hinterland.

Mellander, Lobo, Stolarick and Matheson (2015) use night lights data for Sweden finding the correlation between luminosity data and economic activity is strong enough to make it a relatively good proxy for population and establishment density, but the correlation is weaker in relation to wages. Also, Keola, Andersson and Hall (2015) argue that nighttime lights alone may not explain value-added by agriculture and forestry, however by adding land cover data, the estimate of economic growth in administrative areas becomes more accurate.

Finally, Alesina, Michalopoulos and Papaioannou (2016) study how night lights can be used to explore the issue of ethnic inequality in a region. Donaldson and Storeygard (2016) is a very detailed review paper which discusses all the uses of satellite image data which include climate, weather, agricultural use, urban land use, buildings, natural resources and pollution monitoring.

2.5.2. INDICATIVE DATA SOURCES & EXAMPLES

As discussed above, the use of satellite image data in economics is very promising. However, its uses in macroeconomic nowcasting are more limited mainly due to the slowly changing nature of the underlying subjects. For example, monitoring the agricultural use could be used for long-term agricultural production value-added figures. Or, monitoring plants entry/exit traffic could provide some insights regarding long-term forecasts of industrial production. Following Donaldson and Storeygard (2016) we provide below a list of data sources.

- Landsat, https://goo.gl/Xhqya5. This dataset is publicly available and provides satellite images for urban land cover, beaches, forest cover, mineral deposits;
- MODIS, https://goo.gl/NhHU6x. This dataset is publicly available and provides satellite images for airborne pollution and fish abundance;
- NightLights, https://goo.gl/vdlksu. This is one of various luminosity datasets which is publicly available. It provides satellite images of electricity use;
- SRTM, https://goo.gl/6zKR4x. This dataset is publicly available and provides satellite images for elevation and terrain roughness;
- DigitalGlobe, https://goo.gl/0rL1nW. This dataset is not publicly available and provides satellite images for urban land cover and forestry. Data could be purchased;
- Publicly available datasets (including those mentioned above) are also provided by Google Earth [22].

2.5.3. CONCLUSIONS

Satellite image data is very promising in economics. Satellite images can supplement official statistics and can be very important in developing countries [23] where official figures are difficult to estimate. Luminous data from satellite images has been studied by the literature indicating that night lights are a good proxy for economic activity. Therefore, satellite images

[22] Available at https://goo.gl/Xk0o16.

[23] An EU project is concerned with similar tasks.

are very useful in economic activity and structural analysis, poverty or inequality forecasting, however, due to the slowly changing nature of the captured subjects might not be useful in macroeconomic nowcasting or short-term forecasting on a monthly or quarterly basis. However, on a local economy level, such as regional industry forecasting or consumption, this data could also be used in nowcasting as it would illustrate the activity of particular businesses during the night.

2.6. Scanner prices data

Scanner prices data consists of bar-code scanned data mainly provided by retailers. Prices could be scanned daily allowing for a high-frequency measurement of the supply side in the retail market. Price changes can reveal information in macroeconomic level as well as industry-specific and, more particular, retailer-specific level. Small price changes might be due to measurement error, the analyst who deals with this data must take this into account. Slowly changing prices, on average, often indicate upcoming changes in inflation, however extreme relative prices typically reflect the retailer's conditions rather than changes in average prices. Also, scanner data allow examining different regions inside an economy. Therefore, scanner prices data can be very useful in macroeconomic nowcasting, particularly for inflation subcomponents, such as food prices.

2.6.1. APPLICATIONS

Silver and Heravi (2001) is one of the first papers which associates inflation estimation with scanner prices data. Scanner data provides improved coverage compared to data collected by price collectors and the availability of up-to-date weights at a highly detailed level. This facilitates the creation of superlative price indexes which can incorporate substitution effects. Another use of scanner data in micro-level is to investigate why prices do not rise during peak demand periods. Chevalier, Kashyap and Rossi (2003) find that this is largely due to changes in retail margins which fall during peak demand periods.

Ivancic, Diewert and Fox (2011) use scanner data from 100 stores of four super market chains focusing on 19 item categories. Time aggregation choices lead to a difference in price change estimates for chained and superlative indexes suggesting that traditional index number theory appears to break down when high-frequency data is used. Statistics Netherlands is about to start using regularly scanner data for the compilation of the Dutch CPI. de Haan (2015) proposes a framework which could be employed in order to use scanner data for CPI aggregation.

Lloyd, McCorriston, Morgan, Poen and Zgovu (2012) use scanner data of food retailers in the UK to examine retailer characteristics as well as price dynamics. As mentioned in the data description above, scanner data can reveal information in macro as well as micro level. Lloyd *et al* (2012) find that the frequency of price adjustment and the duration of prices vary across retailers. Also, price promotions vary across retailers as well; some use sales regularly as a promotion tool whereas others rarely use them. These findings can help researchers analyse the local industry and examine consequences for CPI inflation.

Statistics New Zealand (2014) use scanner data to measure price inflation in a particular sector: consumer electronics. Scanner data allows more accurate price measurement, reflects seasonalities in quantities and product substitution. Therefore, such data can be a very powerful tool in nowcasting and short-term forecasting subcomponents of CPI inflation. On the other hand, as electronic goods are more and more sold directly on the web, scanner prices could progressively lose their relevance in favour of internet based prices.

Pandya and Venkatesan (2015) reveal another use of retail scanner prices: consumer behaviour. They use data from 135 super market chains, which accounts for about 79 % of the US super market sales during that period. They show that during the 2003 US-France dispute over the Iraq War the market share of French-sounding, US supermarket brands declined due to consumers' boycott.

2.6.2. INDICATIVE DATA SOURCES & EXAMPLES

As expected, scanner data is not publicly available. Two sources seem to dominate in the literature, those mentioned below. Official Statistical Agencies (e.g. Statistics Netherlands) should own data which is used for in-house analysis.

- The Nielsen Datasets on Consumer Panel Data and Retail Scanner Data have been used by many papers in the literature. Nielsen also provides POS data. The pricing ranges from $4,000 to $7,000 for institutional subscription. More information is available at https://goo.gl/ZOIHP4;
- The second most used dataset on scanner prices is the IRi dataset. This includes 11 years of weekly store data (2001-2011) for chain grocery and drug stores in 47 markets. The academic license data is $1,000 and the size of dataset is more than 350GB. More information is available at https://goo.gl/P5fcda.

2.6.3. CONCLUSIONS

Based on the above discussion, we can conclude that there is ample evidence in favour of scanner data in macro and micro level analysis. Nowcasting and short-term forecasting of specific CPI sub-components could benefit from scanned prices of retailers. Also, industry analysis can be carried out in a both frequent and regional basis. However, access to scanner data, and especially EU data, is very limited.

2.7. 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 in the previous case with scanner data, scraped prices is a potentially useful instrument in nowcasting and short-term forecasting of CPI inflation as well as retail sales variables.

Online prices are also characterised by seasonalities and stylised facts, which 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 ([24]) products per category. 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

[24] This is a rough estimate as in specific categories there can be as much as 100 or more products.

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.

2.7.1. APPLICATIONS

Academic papers in economics have started using web scraped data during the last six to seven years. 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 supermarket 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. The paper finds that for Argentina, the online inflation rate is nearly three times higher than the official estimate. This data collection is part of the MIT Billion Prices Project [25]. 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.

Boettcher (2015) describes in detail technological, data security and legal requirements 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 are 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 clustering. 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 statistics 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.

[25] Available at https://goo.gl/xb4H95.

2.7.2. INDICATIVE DATA SOURCES & EXAMPLES

We list below a short list of ready databases with scraped prices. Alternatively, an official statistics agency would start collecting/scraping the online prices from scratch, i.e. using major super markets, retailers, etc. In a few years, this procedure could generate enough data to cross-validate the ability of scraped prices for nowcasting CPI inflation, retail sales and other price related variables.

- MIT Billion Prices Project, https://goo.gl/xb4H95. Most of the data mentioned in the work of Cavallo is available online. Particularly:
    - Daily prices for all goods sold by 7 large retailers in Latin America and the US: 2 in Argentina, 1 in Brazil, 1 in Chile, 1 in Colombia, 1 in Venezuela, and 4 in the US. Daily data from 2007 to 2010. Used in Cavallo (2016),
    - Daily prices for all goods sold by APPLE, IKEA, ZARA, and H&M. Daily data for 85 countries [26] from 2008 to 2013,
    - Online and offline prices for individual goods sold by 56 large multi-channel retailers in 10 countries: Argentina, Australia, Brazil, Canada,
    - China, Germany, Japan, South Africa, the UK, and the US. Mixed frequency data from 2014 to 2016. Used in Cavallo (2017);
- PriceStats, https://goo.gl/zqQoaS. This is a private company. Cavallo (2016) used scraped data for 181 retailers in 31 countries provided by PriceStats. The dataset could be purchased.

2.7.3. CONCLUSIONS

Online prices is definitely a source of big data which seems to have potential in CPI inflation nowcasting and forecasting and is something which has not been extensively studied before. Data for European countries does exist by private companies, however it is not publicly available. Official statistical agencies will likely experiment with this method of price collection in the near future. It might be useful in that case to complement price data with volumes of sales.

2.8. 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 procedure 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

[26] Coverage of countries varies by retailer and time.

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 ([27]).

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 he/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 specific keywords and search-items, Google Correlate available at https://goo.gl/Uj1mox. 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 ([28]). However, given that this is an automatic procedure, Google Correlate is not able to filter not appropriate keywords. For example, Google Correlate might return as a result celebrities' names which happen to be trending during the same time period.

As we briefly describe below, Google Trends have been used in various applications in economics, finance, health sector, etc. with substantial success. Some specific characteristics of online search data which need to be taken into account by the researcher are: (i) clear and intuitive keywords and search-items (see Ross, 2013), (ii) weekly patterns and seasonalities and (iii) extreme events, such as physical phenomena or politics, which might result to outliers (even though sometimes could be useful as they illustrate agents behaviour, see Vosen and Schmidt, 2011).

2.8.1. APPLICATIONS

The literature which incorporates the use of online search data is already vast, even though it started very recently. Below, we briefly mention some indicative papers in economics, finance, tourism industry, health and politics. However, the use of online search data expands to other interdisciplinary fields as well. Chamberlain (2010), Choi and Varian (2012), Varian and Stephens-Davidowitz (2014) are among the first to provide nowcasting and forecasting applications using Google Trends. Also, Kapetanios *et al* (2016) provide a detailed overview and application of Google Trends data in EU macroeconomic nowcasting.

Firstly, internet data has applications in health. Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant (2009) present a method of analysing large numbers of Google search queries

[27] This is mainly used in China.
[28] This could be done by aggregating or averaging the Google Trend time series, or by extracting their common factors.

to track influenza-like illness in a population. Also see Yuan, Nsoessie, Lv, Peng, Chunara and Brownstein (2013) for influenza epidemics monitoring in China using Baidu search results. Tefft (2011) uses Google Trends to create a depression search index to provide insights on the relationship between unemployment, unemployment insurance, and mental health. The results indicate a positive relationship between the unemployment rate and the depression search index.

Then, Schmidt and Vosen (2011) propose a new indicator for the US private consumption based on search query time series provided by Google Trends. They find that the Google indicator outperforms the relevant survey-based indicators, the University of Michigan Consumer Sentiment Index and the Conference Board Consumer Confidence Index. This suggests that incorporating information from Google Trends may offer significant benefits to forecasters of private consumption.

Koop and Onorante (2013) suggest to nowcast using dynamic model selection (DMS) methods which allow for model switching between time-varying parameter regression models. This is potentially useful in an environment of coefficient instability and over-parametrisation which can arise when forecasting with Google variables. They allow for the model switching to be controlled by the Google variables through Google probabilities. That is, instead of using Google variables as regressors, they allow them to determine which nowcasting model should be used at each point in time. In an empirical exercise involving nine major monthly US macroeconomic variables, they find that DMS methods provide large improvements in nowcasting. The use of Google model probabilities within DMS often performs better than conventional DMS. Their application includes various US macroeconomic variables.

Bontempi, Golinelli and Squadrani (2016) develop a new uncertainty indicator based on Google Trends with applications to US variables. Their results suggest that the Google Trends indicator shocks embody timely information about people's perception of uncertainty and sometimes earlier than other indexes. This is evidence that online search data could act as a leading indicator and used in macroeconomic nowcasting and forecasting.

Choi and Varian (2009) and Choi and Varian (2012) illustrate the ability of Google Trends to predict the present using daily and weekly reports of Google Trends. In particular, they claim that people who lose their jobs search the internet for job ads. Therefore, the increasing volume of Google search queries for job-related keywords potentially has an impact on forecasting/nowcasting the initial claims. Their applications include US variables.

Reis, Ferreira and Perduca (2015) analyse the web activity as a big data source. Electronic traces left by users while they use web services could be used as data either in real time or with very small time lags. As many human activities measured by official statistics are closely related to people's behaviour online, this data on people's web activity offers the potential to produce predictions for socio-economic indicators with the purpose to increase the timeliness of the statistics. Papers in the literature have shown evidence that these predictions can be made. However, this type of data should be further checked about its transparency, continuity, quality and potential to be integrated with official statistics traditional methods. The empirical application they implement is an improved nowcasting of French and Italian unemployment. More recently, Smith (2016) use a mixed-frequency model in UK unemployment forecasting.

Wirthmann and Descy try to understand demand of labour and skills with a particular focus to the job requirement evolution web scraping of job vacancies. The aim is explore ways to better utilise existing data to produce derived measures of skills demand, skills supply, mismatches and skills development. A prototype system, which was successfully tested in five countries (UK, DE, CZ, IT, IE). The prototype was designed to retrieve from job portals selected on the basis of quality and relevance. Adopting an ethical behaviour, authors only scraped portals for which permission was granted or, even better, obtained direct access to job portals' databases. Data processing transforms 'documents' — original job posting as read by by the crawler — into 'vacancies' — pre-structured single job openings. This is done by expanding job posting — one job posting can contain more than one job opening - and de-duplication — one job opening can be posted in more than one website. Vacancies are processed using text mining and machine learning algorithms to identify classify jobs into occupations and gather information about contract type, working hours, job location as well as skills and job requirements. The European skills, competence, qualification and Occupations taxonomy (ESCO) was used as multilingual taxonomy. Results are accessible via OLAP cube as well as through simple graphic interface.

2.8.2. INDICATIVE DATA SOURCES & EXAMPLES

Perhaps the most intuitive, publicly available and accurate source of internet search data is Google Trends: https://goo.gl/64mcg6.

2.8.3. CONCLUSIONS

Internet search data and particularly Google Trends is a very promising source of structured big data. The literature provides ample scientific evidence that web search data has predictive abilities in various fields such as economics, finance, politics and health. This is not surprising as web search data is input of humans and thus, reflects agents' behaviour. However, it must be noted that online search data must be carefully used as the value it adds depends on the nowcasting exercise. For example, unemployment rate benefits from online search data whereas GDP growth might not, unless perhaps a large universe of manually selected keywords is employed.

2.9. 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.

2.9.1. APPLICATIONS

The literature on textual data in economics mainly uses newspaper text and text data from the FOMC minutes of the FED.

General information about text mining is provided in Smith (2010) and Bholat, Hansen, Santos and Schonhardt-Bailey (2015). Predictive Analytics [29] website lists a number of software which can be used in text analysis.

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

[29] https://goo.gl/SkNnVs.

closeness to the actual future stock price, the same direction of price movement 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 firecessions.

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 (2015) [30] develop an index of economic policy uncertainty based on newspaper coverage frequency.

Using firm-level data, the authors find that policy uncertainty raises stock price volatility and reduces investment and employment in policy-sensitive sectors like defense, healthcare, and infrastructure construction. 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 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. Steckler and Symington (2016) quantify the minutes 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 (2015) 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.

Textual analysis can also be used in political economy. Acemoglu, Hassan and Tahoun (2015) use textual data from GDELT project 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.

2.9.2. INDICATIVE DATA SOURCES & EXAMPLES

Below we provide an indicative list of textual data sources.

- First, an obvious source of newspaper headlines and media coverage is access to newspaper archives and other archive websites. There are two difficulties associated with archives: (i) the first is that archives offer scanned images or photographs of newspaper pages, thus a transformation of newspaper page images to text is necessary in order to create a textual

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

database [31], (ii) language is the second difficulty as national newspapers use the official language in each country:

- The British Newspaper archive, https://goo.gl/gdnt0s. It provides access to a historical archive which spans 200 years for 717 UK newspapers titles. It is publicly available but unlimited access requires a £12.95/month subscription,
- National Library of Australia — Trove, https://goo.gl/4yFmFC. This database offers access to 500 newspaper titles covering the period 1803–2011,
- Google News Archive, https://goo.gl/7IPVwL. Google provides free access to scanned archives of newspapers. Some of the news archives date back to 1700s and include many national and regional newspapers in multiple languages,
- Wikipedia: List of online newspaper archives, https://goo.gl/Ig2WPl. Wikipedia provides a detailed list of available online sources for each country;

- Reuters: Reuters Online News Archive — US Edition: https://goo.gl/G3lb30. The Reuters Online News Archive is a collection of news articles published on the United States edition of the Reuters website, starting from January 1st 2007. The archive is continuously updated as content is published on the website. The online repository is structured as a plain HTML web-page and can be browsed by date: subpages simply list articles published on a single day in inverse chronological order. Although the archive is not searchable, its simple design allows for easy web-scraping. Since articles are listed as they are published and/or updated, the archive contains duplicates and a large number of links return a 404 error (page not found). The Reuters News Archive has been used mainly to investigate interrelations among financial institutions (see Rönnqvist and Sarlin, 2015);
- GDELT Project [32], https://goo.gl/6FWcQl. The Global Database of Events, Language, and Tone (GDELT) is 'an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day'. It is mainly used for political conflict (see, e.g. Acemoglu, Hassan and Tahoun (2014), Heaven (2013) among others), however it could be used potentially in the building of an uncertainty index as it provides data every 15 mins. It must be highlighted that the data volume is very big and noise could be present (see Beieler, 2013).

2.9.3. CONCLUSIONS

Financial applications using textual information are mostly concerned with uncertainty indexes and market forecasting based on newspaper headlines and quantified FOMC minutes. Newspaper headlines could be potentially used for macroeconomic nowcasting on a daily basis. The GDELT project could also be a potential instrument in the creation of an uncertainty index based on big data. This, in turn, could be used in macroeconomic nowcasting and forecasting (provided that its predictive abilities have been identified).

[31] There exist, at least, two Optical Character Recognition (OCR) packages available in R: abbyyR and Tesseract. In both cases, the use loads a vector with image data (.png, .jpeg, .tiff) and R recognizes and scraps the main text.

[32] GDELT provides an online interface to make queries about particular keywords. However, GDELTTools, https://goo.gl/lqDnl9, can also be used to load and normalise data inside R without the need of other user interfaces.

2.10. 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 ([33]). 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', restricted to 140 characters. Tumblr, launched in 2007, is a service which allows users to post multimedia and other content to a short-form blog (or 'micro-blogging'). Instagram, going online in 2010, started as a photo-sharing site which now allows the post 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 with Google Trends, social media data could have a 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:

- 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 (#) makes Twitter discussions easier for monitoring and tracking events. This allows identifying 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 ([34]), 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 organisations, government agencies and key reporters and analysts and, then, filter the feeds for particular hashtags or keywords.

[33] See https://newsroom.fb.com/company-info/.
[34] See Chi and Yang (2010) for more information about the use of Twitter in the Congress.

2.10.1. APPLICATIONS

Social media data use in finance and economics has rapidly increased. The University of Michigan has established economic indicators based on social media with encouraging results on forecasting. They publish online their nowcasts on initial claims for unemployment insurance using their Social Media Job Loss Index [35]. Ramli (2012) and Griswold (2014) are just some examples of the media interest on the use of social networks data in economic analysis. The literature indicates three main areas of application: (i) financial markets; mainly stock market and foreign exchange, (ii) politics [36], and (iii) public mood [37]. Below, we provide some additional information focusing more on the financial markets.

Economics and financial researchers have realised the usefulness of social networks since 2004. Antweiller and Frank (2004) are among the first to use a big data of online message boards to examine the impact of social networks on stock returns. They analyse 1.5 million messages from Yahoo!Finance and Raging Bull for the 45 Dow Jones Industrial Average. They confirm that stock messages help predict market volatility and their effect on stock returns is statistically significant (although economically small).

The stock market forecasting greatly benefits from the use of Twitter. The key reason for this is that the hashtag keyword is the stock's ticker, which allows for a clear and precise feed. Many papers in the academic and professional literature demonstrate the usefulness of Twitter data in stock directional movements or returns. Bollen, Mao and Zheng (2011) and Mittal and Goel (2012) extract investors' sentiments from Twitter data to predict the Dow Jones Industrial Average. Both papers, although they follow different methodologies, report an improvement in DJIA forecasting accuracy by 86.7-87 %. Makrehchi, Shah and Liao (2013) also create Twitter sentiment data to predict the S&P500. They use two rating scale systems for moods: OpinionFinder and Google-Profile-of-Mood States. The authors build a trading strategy using the signals from the sentiment data outperforming the standard buy-and-hold by about 20 %. Due to the proprietary nature of the Google-Profile of Mood States algorithm, Chen and Lazer experiment with simpler methods which could provide similar results.

Arias, Arratia and Xuriguerra (2013) present empirical work using Twitter forecasting in (i) US box-office sales, and (ii) Apple, Google, Yahoo, Microsoft, S&P100 implied volatility index, S&P500 index and S&P500 implied volatility index closing prices. As in the previous cases, sentiment indicators are created based on Twitter data. As a general result, they show that nonlinear models do take advantage of Twitter data when forecasting trends in volatility indices, while linear ones fail systematically when forecasting any kind of financial time series. In the case of predicting box office revenue trend, it is support vector machines that make best use of Twitter data.

Sprenger, Tumasjan, Sandner and Welpe (2014a) also use Twitter sentiment to predict stock returns. The authors use data for an undisclosed S&P 100 company and find an association between tweet sentiment and stock returns, message volume and trading volume, as well as disagreement and volatility. Their results demonstrate that users providing above average investment advice are retweeted more often and have more followers, which amplifies their share of voice. Sprenger, Tumasjan, Sandner and Welpe (2014b) use computational linguistics to a dataset of more than 400,000 S&P 500 stock-related Twitter messages, and distinguish between good and bad news. The results indicate that the returns prior to good news events

[35] Visit https://goo.gl/jfvpDm for more information.

[36] See Tumasjan, Sprenger, Sandner and Welpe (2010), Conover, Goncalves, Ratkiewicz, Flammini and Menczer (2011), Wang, Can, Kazemzadeh, Bar and Narayanan (2012) and Makhazanov and Rafiei (2013) among others.

[37] See Bollen, Mao and Pepe (2011) and Lansdall-Welfare, Lampos and Cristianini (2012) among others.

are more pronounced than for bad news events. Pineiro-Chousa, Vizcaino-Gonzalez and Perez-Pico (2016) use data from Stocktwits, a social network similar to Twitter, where users share posts about stocks, indexes, and financial markets. They focus on investors' activity through social media and these media's influence over the Chicago Board Options Exchange Market Volatility Index. The results show that social media sentiment influences stock markets.

2.10.2. INDICATIVE DATA SOURCES & EXAMPLES

Following the discussion above, we can claim that the most obvious, and possibly useful, source of social media data for economic applications is Twitter:

- Twitter public streams data can be downloaded for free using the Twitter API, see https://goo.gl/YYaTaA. An R package, twitteR [38], can be integrated and Twitter data can be downloaded automatically in R for further editing. However, the Twitter API data can go back 1400 days;
- Alternative services which offer Twitter data which can go back to the first day Twitter was online are provided by third-party agents. A reliable company is gnip, https://goo.gl/dmKJzb, and sifter, https://goo.gl/uHh1oc. Both these services are proprietary;
- For financial market instrument with ticker symbols, StockTwits could be used as well. However, Stocktwits social media market share is much smaller compared to Twitter's.

2.10.3. CONCLUSIONS

There has been documented by now plenty of evidence in favour of Twitter use in the creation of sentiment indexes. Most applications are in financial markets instruments, elections and public mood. There does not seem to exist empirical work in macroeconomic nowcasting or forecasting, however Twitter could be one of the ingredients in a sentiment indicator, possibly with Google Trends as well.

3. Nowcasting specific macroeconomic variables using big data

In Section 2, we have considered a typology of big data potentially relevant for macroeconomic analysis and, in particular, for nowcasting/forecasting, and for each of them discussed some relevant applications. As mentioned in the Introduction, we now consider the dual problem: for a specific macroeconomic variable of interest, such as unemployment, GDP and components, inflation, surveys, financial variables, we list studies based on nowcasting them using big data. More specifically, for each study we provide: (i) author(s) and their affiliation, (ii) brief description of the paper, (iii) data characteristics (big data and standard macro/financial variables), (iv) econometric methodology. A summary of the most useful papers and their characteristics is described in Table 2.

[38] See https://goo.gl/cuLZTL for more information.

Table 2: Indicative list of most important work in macroeconomic nowcasting using big data

	Title	Auhor(s)	Data	Method(s)	Country	Big data Added Value	Limitation(s)
Unemployment	Predicting Initial Claims for Unemployment Benefits and Predicting the Present with Google Trends	Choi, Varian	Google	Linear Regression	US	Improvement of fore(now-) casting	Limited number of other indicators, Methodology
	Google Econometrics and Unemployment Forecasting	Askitas, Zimmermann	Google	VECM	DE	Improvement of fore(now-) casting	Limited number of other indicators, Methodology
	The Predictive Power of Google Searches in Forecasting Unemployment	D'Amuri, Marcucci	Google	ARMA	US	Improvement of fore(now-) casting	Methodology
	Nowcasting with Google Trends: a Keyword Selection Method	Ross	Google	Linear Regression	UK	Improvement of fore(now-) casting	Methodology
	The Use of Web Activity Evidence to Increase the Timeliness of Official Statistics Indicators	Reis, Ferreira, Perduca	Google	Linear Regression	FR, IT	Big data in official statistics, Improvement of fore(now-) casting	Methodology
	Improving Prediction of Unemployment Statistics with Google Trends: Part 2	Ferreira	Google	Linear Regression, Factor Models	PT	Latent variables, Improvement of forecasting	Methodology
GDP and components	Nowcasting GDP with Electronic Payments Data	Galbraith, Tkacz	Debit card and Credit Card transactions, Cheques, housing index, employment, stock exchange index, money supply, average work week (hours), new orders, durables, inventories, retail trade	Linear Regression, Factor Models	CA	Nowcasting improvement using big data	Data, Methodology
	Forecasting Private Consumption: Survey-based Indicators vs Google Trends	Schmidt, Vosen	Google, Consumer Sentiment, Consumer Confidence	Linear Regression	US	Activity Indicator, forecasting of macro variables	Methodology
	Are ATM/POS data relevant when nowcasting private consumption?	Esteves	ATM/POS	Linear Regression	PT	Nowcasting improvement using big data	Data, Methodology
	Dankort payments as a timely indicator of retail sales in Denmark	Carlsen, Storgaard	Debit card data		DK	Nowcasting improvement using big data	Data, Methodology
	A mixed frequency approach to forecast private consumption with ATM/POS data	Duarte, Rodrigues, Rua	ATM/POS	MIDAS	PT	Nowcasting improvement using big data	Data, Methodology

Table 2: Indicative list of most important work in macroeconomic nowcasting using big data

	Title	Auhor(s)	Data	Method(s)	Country	Big data Added Value	Limitation(s)
GDP and components	Using the payment system data to forecast the Italian GDP	Aprigliano, Ardizzi, Monteforte	Payments data	MIDAS, LASSO	IT	Nowcasting improvement using big data	
	Macroeconomic Nowcasting Using Google Probabilities	Koop, Onorante	Inflation, Wage Inflation, Unemployment, Term Spread, FCI, Commodities Price Inflation, Industrial Production, Oil Price Inflation, Money Supply	Model Averaging	US	Use of Google Data as proxy for improved weighting scheme	
Inflation	The Billion Prices Project: Research and Inflation Measurement Applications	Cavallo	Google, CPI, Gas	VAR	Argentina, US, EA	Use of big data in inflation forecasting	Data
	Automatic Data Collection on the Internet (Web Scraping)	Boettcher	Prices	-	AT	Use of Web scraping	Algorithms, Storage, Quality of data
	Collecting Clothing Data from the Internet	Griffioen, de Haan, Willenborg	CPI, Apparel	-	NL	Use of Web scraping	Algorithms, Storage, Quality of data
	Using Web Scraped Data to Construct Consumer Price Indices	Breton, Swier, O'Neil	Inflation, CPI, RPI	Aggregation	UK	Use of Web scraping in aggregate index construction	Quality of data

3.1. Unemployment

3.1.1. PREDICTING INITIAL CLAIMS FOR UNEMPLOYMENT BENEFITS AND PREDICTING THE PRESENT WITH GOOGLE TRENDS

Author(s): Choi, H., Varian, H. (Google Inc.)

Brief Description: Choi and Varian (2009) and Choi and Varian (2012) illustrate the ability of Google Trends to predict the present (nowcasting) using daily and weekly reports of Google Trends. In particular, they claim that people who lose their jobs search the internet for job ads. Therefore, the increasing volume of Google search queries for job-related keywords potentially has an impact on forecasting/nowcasting the initial claims.

Data: Google Search Insights and Google Trends, Retail Sales, Automotive Sales, Home Sales, Travel

Methodology: Linear Regression Models with/without lags and independent variables

3.1.2. GOOGLE ECONOMETRICS AND UNEMPLOYMENT FORECASTING

Author(s): Askitas, N. (IZA), Zimmermann, K. (Bonn University, DIW Berlin, IZA)

Brief Description: The paper suggests an innovative method of using data on internet activity to predict economic behaviour in a timely manner, which is difficult at times of structural change. They show a strong correlation between keyword searches and unemployment rates using monthly German data.

Data: Seasonally unadjusted monthly unemployment rate of Germany (01/01/2004- 01/04/2009). Google Insights. German Keywords (in English): unemployment office/agency', 'unemployment rate', Personnel Consultant', popular job search engines in Germany (stepstone, jobworld, jobscout etc.).

Methodology: Time series causality analysis using Error Correction Model (ECM)

3.1.3. THE PREDICTIVE POWER OF GOOGLE SEARCHES IN FORECASTING UNEMPLOYMENT

Author(s): D'Amuri, F., Marcucci, J. (Bank of Italy, Economic Research and International Relations)

Brief Description: D'Amuri and Marcucci (2012) suggest the use of an index of Internet job-search intensity (the Google Index, GI) as the best leading indicator to predict the US monthly unemployment rate. They perform a deep out-of-sample forecasting comparison analysing many models that adopt their leading indicator, the more standard initial claims or combinations of both. They find that models augmented with the GI outperform the traditional ones in predicting the unemployment rate for different out-of-sample intervals that start before, during and after the Great Recession. Google-based models also outperform standard ones in most state-level forecasts and in comparison with the Survey of Professional Forecasters. These results survive a falsification test and are also confirmed when employing different keywords.

Data: Google Data (Unit Root testing and transformations)

Methodology: ARMA, ARMAX (various combinations)

3.1.4. NOWCASTING WITH GOOGLE TRENDS: A KEYWORD SELECTION METHOD

Author(s): Ross, A. (Fraser of Allander Institute, University of Strathclyde)

Brief Description: Ross (2013) investigates the issues of identifying and extracting keywords from Google Trends relevant to economic variables. He suggests the backward induction method which identifies relevant keywords by extracting these from variable relevant websites. This backward induction method was applied to nowcast UK unemployment growth using a small set of keywords. The majority of keywords identified using the backward induction method outperformed the competing models in terms of in-sample and out-of-sample tests of predictability indicating that the backward induction method is effective in identifying relevant keywords.

Data: Google Data, Unemployment

Methodology: Linear Regressions

3.1.5. THE USE OF WEB ACTIVITY EVIDENCE TO INCREASE THE TIMELINESS OF OFFICIAL STATISTICS INDICATORS

Author(s): Reis, F. (Eurostat), Ferreira, P. (Eurostat), Perduca, V. (Universite Paris Descartes, CNRS)

Brief Description: Reis *et al* (2015) analyse the web activity as a big data source. Electronic traces left by users while they use web services could be used as data either in real time or with very small time lags. As many human activities measured by official statistics are closely related to people's behaviour online, this data on people's web activity offers the potential to produce predictions for socio-economic indicators with the purpose to increase the timeliness of the statistics. Papers in the literature have shown evidence that these predictions can be made. However, this type of data should be further checked about its transparency, continuity, quality and potential to be integrated with official statistics traditional methods. The empirical application they implement is an improved nowcasting of French and Italian unemployment.

Data: French and Italian Job related keywords

Methodology: Linear Regression

3.1.6. IMPROVING PREDICTION OF UNEMPLOYMENT STATISTICS WITH GOOGLE TRENDS: PART 2

Author(s): Ferreira, P. (Eurostat)

Brief Description: Ferreira (2015) uses a dynamic factor model to extract a latent variable from Google Trends data which is a good proxy for the unemployment dynamics. Prediction models for unemployment that make use of the estimated latent variable have performed better than the proposed approaches in previous works, in particular during a period where there was an abrupt change in the trend.

Data: Google Trends

Methodology: Linear Regression, Dynamic Factor Models

3.2. GDP and components

3.2.1. NOWCASTING GDP WITH ELECTRONIC PAYMENTS DATA

Author(s): Galbraith, J. W., Tkacz, G. (ECB)

Brief Description: Galbraith and Tkacz (2015) assess the usefulness of a large set of electronic payments data comprising debit and credit card transactions, as well as cheques that clear through the banking system, as potential indicators of current GDP growth in Canada. These variables capture a broad range of spending activity and are available on a very timely basis, making them suitable current indicators. While every transaction made with these payment mechanisms is in principle observable, the data are aggregated for macroeconomic forecasting. Controlling for the release dates of each of a set of indicators, they generate nowcasts of GDP growth for a given quarter over a span of five months, which is the period over which interest in nowcasts would exist. They find that nowcast errors fall by about 65 per cent between the first and final nowcast. Among the payments variables considered, debit card transactions appear to produce the greatest improvements in forecast accuracy.

Data: Debit card and Credit Card transactions, Cheques, lagged GDP, housing index, business and personal sales employment, stock exchange index, money supply, average work week (hours), new orders, durables, inventories, retail trade.

Methodology: Mostly based on linear regressions

3.2.2. FORECASTING PRIVATE CONSUMPTION: SURVEY-BASED INDICATORS VS GOOGLE TRENDS

Author(s): Schmidt, T. (RWI), Vosen, S. (RWI)

Brief Description: Schmidt and Vosen (2011) introduce an indicator for private consumption based on search query time series provided by Google Trends. The indicator is based on factors extracted from consumption-related search categories of the Google Trends application Insights for Search. The forecasting performance of this indicator is assessed relative to the two most common survey-based indicators — the University of Michigan Consumer Sentiment Index and the Conference Board Consumer Confidence Index. The results show that in almost all conducted in-sample and out-of-sample forecasting experiments the Google indicator outperforms the survey-based indicators.

Data: Google Insight/Trends, Michigan Consumer Sentiment, Conference Board Consumer Confidence

Methodology: Linear Regression Models

3.2.3. MACROECONOMIC NOWCASTING USING GOOGLE PROBABILITIES

Author(s): Koop, G. (University of Strathclyde), Onorante, L. (ECB)

Brief Description: Koop and Onorante (2013) suggest to nowcast using dynamic model selection (DMS) methods which allow for model switching between time-varying parameter regression models. This is potentially useful in an environment of coefficient instability and over-parametrisation which can arise when forecasting with Google variables. They allow for the model switching being controlled by the Google variables through Google probabilities. That is, instead of using Google variables as regressors, they allow them to determine which nowcasting model should be used at each point in time. In an empirical exercise involving nine major monthly US macroeconomic variables, they find that DMS methods provide large improvements in nowcasting; the variables are: inflation, industrial production, unemployment, wage inflation, money, supply, financial conditions index (FCI), oil price inflation, commodity price inflation and the term spread. The use of Google model probabilities within DMS often performs better than conventional DMS.

Data: Inflation, Wage Inflation, Unemployment, Term Spread, FCI, Commodities Price Inflation, Industrial Production, Oil Price Inflation, Money Supply (see Table 1 of the paper for more details and transformations)

Methodology: Dynamic Model Averaging

3.3. Inflation

3.3.1. THE BILLION PRICES PROJECT: RESEARCH AND INFLATION MEASUREMENT APPLICATIONS

Author(s): Cavallo, A. (MIT), Rigobon, R. (MIT)

Brief Description: Cavallo and Rigobon (2016) examines ways to deal with price data. Potential sources for micro price data include: Statistical Offices, Scanner Data (e.g. Nielsen), Online data (e.g. Billion Prices Project) etc. CPI data is useful in measuring inflation whereas Scanner and Online data can be used in marketing analytics (e.g. market shares). The Billion Prices Project is an automated web-scraping software where a robot downloads a public page, extracts the prices information and stores it in a database. A direct outcome from the papers is that online data is also useful for nowcasting inflation in the US, Latin America and Euro Area. Links between online data and CPIs are tracked using VAR models and calculating the cumulative Impulse Response Functions. The forecasting examples use predictive regressions.

Data: Google Data, CPI, Gas Prices

Methodology: VAR (for Impulse Responses), Linear Regressions for Forecasting

3.3.2. COLLECTING CLOTHING DATA FROM THE INTERNET

Author(s): Griffioen, R., de Haan, J., Willenborg, L. (Statistics Netherlands)

Brief Description: The paper is concerned with the usability of online apparel prices for CPI analysis. This study falls in the web scraping category and reports the findings and difficulties of online price collection during a two years period. The advantages of web scraping clothing prices are: (i) online price collection is cheaper than price collection in physical stores, (ii) given the relatively low collection costs, there is an incentive to rely on 'big data' and circumvent small sample problems (e.g. high sampling variance), (iii) the quality of online data tends to be very good and (iv) some item characteristics can be easily observed. The main disadvantages of conducting a data collection of this type are: (i) website changes can lead to data problems, (ii) the choice of web scraping strategy can affect the information collected and item representativeness, (iii) weighting information is unavailable, and (iv) the available information on characteristics may be insufficient, depending on the need for quality adjustment.

Data: CPI, Clothing Prices

3.3.3. USING WEB SCRAPED DATA TO CONSTRUCT CONSUMER PRICE INDICES

Author(s): Breton, R., Swier, N., O'Neil, R. (Office for National Statistics (ONS), UK)

Brief Description: The purpose of this paper is to provide an overview of ONS research into the potential of using web scraped data for consumer price statistics. The research covers the collection, manipulation and analysis of web scraped data. As before, the main benefits of web scraped data are identified as follows: (i) reduced collection costs, (ii) increased coverage (i.e. more basket items), (iii) increased frequency, (iv) production of new or complimentary outputs/indices, and (v) improved ability to respond to new challenges. ONS use web scraped data to calculate price indices which: (i) expand the number of items used, (ii) expand the number of days considered, and (iii) expand both the number of items and days considered. The construction of this sort of indices can be useful for economists and policymakers.

Data: Inflation, CPI, RPI, Web Scraping

3.4. Surveys

3.4.1. NEWS AND NARRATIVES IN FINANCIAL SYSTEMS: EXPLOITING BIG DATA FOR SYSTEMIC RISK ASSESSMENT

Author(s): Nyman, R. (UCL), Gregory, D. (Bank of England), Kapadia, S. (Bank of England), Smith, R. (UCL), Tuckett, D. (UCL)

Brief Description: Nyman *et al* (2014a) investigate ways to use big data in systemic risk management. News and narratives are key drivers behind economic and financial activity. Their news data consists of (i) daily comments on market events, (ii) weekly economic research reports and (iii) Reuters news. Machine Learning and Principal Components are included in the methodology in order to calculate the consensus indexes based on the above sources. Their findings include that weekly economic research reports could potentially forecast the Michigan Consumer Index and daily comments on market events could potentially forecast market volatility.

Data: Broker Reports, Bank of England Internal Reports, Reuters News Archive

Methodology: Emotion Dictionary Words

3.4.2. BIG DATA AND ECONOMIC FORECASTING: A TOP-DOWN APPROACH USING DIRECTED ALGORITHMIC TEXT ANALYSIS

Author(s): Nyman, R. (UCL), Smith, R. (UCL), Tuckett, D. (UCL)

Brief Description: Nyman *et al* (2014b) introduce the Directed Algorithmic Text Analysis and show that this methodology can improve considerably on consensus economic forecasts of the Michigan Consumer Index Survey. The approach is based upon searching for particular terms in textual data bases. In contrast to econometric approaches, their methodology is based upon a theory of human decision making under radical uncertainty. The search is directed by the theory. This direction dramatically reduces the dimensionality of the search. They look for words which convey a very limited number of emotions. As in other approaches, they also use regression analysis, but the choice of variables comes from the underlying theory of decision making.

Data: Text Data, Michigan Cons. Index

Methodology: Linear Regression, Building regressors upon a theory

3.5. Financial variables

3.5.1. HOW TO MEASURE THE QUALITY OF FINANCIAL TWEETS

Author(s): Cerchiello, P., Giudici, P. (University of Pavia)

Brief Description: Apart from Google Trends, economic and financial researchers have also started using Twitter posts about various economics and financial news. Cerchiello and Giudici (2014) investigate how the quality of financial tweets can be measured. They suggest that a Google Scholar 'h-index' type measure allows for improved nowcasting of financial variables using Twitter texts. The Twitter users are ranked according to their 'h-index' and confidence intervals are constructed to decide whether top Twitter users are significantly different. Twitter data are collected and R language's TwitteR package is adopted. Their methodology lies in the field of loss data modelling.

Data: Twitter data

Methodology: Proposal of an h-index (similar to Google Scholar) using tweets.

3.5.2. NEWS VERSUS SENTIMENT: COMPARING TEXTUAL PROCESSING APPROACHES FOR PREDICTING STOCK RETURNS

Author(s): Heston, S. L. (University of Maryland), Sinha, N. R. (Board of Governers of the Federal Reserve System)

Brief Description: Heston and Sinha (2014), even though it is not a macroeconomics application, use a dataset of over 900,000 news stories to test whether news can predict stock returns. They find that firms with no news have distinctly different average future returns than firms with news. Confirming previous research, daily news predicts stock returns for only 1-2 days. But weekly news predicts stock returns for a quarter year. Positive news stories increase stock returns quickly, but negative stories have a long-delayed reaction.

Data: News stories, Stock Returns

Methodology: News Sentiment, Regressions (Cross-Sectional)

3.6. Other studies

3.6.1. STATISTICAL MODEL SELECTION WITH 'BIG DATA'

Author(s): Doornik, J. A., Hendry, D. F. (Institute for New Economic Thinking, Economics Department, Oxford University)

Brief Description: big data offers potential benefits for statistical modelling, but confronts problems like an excess of false positives, mistaking correlations for causes, ignoring sampling biases, and selecting by inappropriate methods. Doornik and Hendry (2015) consider the many important requirements when searching for a data-based relationship using big data. Paramount considerations include embedding relationships in general initial models, possibly restricting the number of variables to be selected over by non-statistical criteria (the formulation problem), using good quality data on all variables, analysed with tight significance levels by a powerful selection procedure, retaining available theory insights (the selection problem) while testing for relationships being well specified and invariant to shifts in explanatory variables (the evaluation problem), using a viable approach that resolves the computational problem of immense numbers of possible models.

Data: Artificial Data

Methodology: Multiple Testing, Autometrics, Lasso

3.6.2. MEASURING ECONOMIC POLICY UNCERTAINTY

Author(s): Baker, S. R. (Northwestern), Bloom, N. (Stanford), Davis, S. J. (The University of Chicago)

Brief Description: This paper develops a new index of economic policy uncertainty (EPU) based on newspaper coverage frequency. Several types of evidence — including human readings of 12,000 newspaper articles — indicate that this index proxies for movements in policy-related economic uncertainty. The index spikes near tight presidential elections, Gulf Wars I and II, the 9/11 attacks, the failure of Lehman Brothers, the 2011 debt-ceiling dispute and other major battles over fiscal policy. Using firm-level data, they find that policy uncertainty raises stock price volatility and reduces investment and employment in policy-sensitive sectors like defence, healthcare, and infrastructure construction.

Data: Newspaper data with selected keywords such as: regulation, budget, spending, deficit, tax etc.

Methodology: Vector Autoregressions

4. Types of big data by dominant dimension

In this Section we focus on numerical data only, which can either be the original big data or the result of a transformation of unstructured data, and focus on the specific dimensions of the dataset, which are also important to identify the required econometric techniques. This classification is also relevant as will deal, respectively, with how to transform unstructured data into structured numerical data, and how to pre-treat big numerical data to eliminate outliers, recurrent temporal patterns and other data irregularities. It is also a convenient classification for the review of econometric techniques for numerical big data.

Following, e.g. Doornik and Hendry (2015), and as mentioned in the Introduction, we can distinguish three main types of big data: Fat (big cross-sectional dimension, N, small temporal dimension, T), Tall (small N, big T), or Huge (big N, big T). We discuss their main features in the following subsections.

4.1. Fat datasets

Fat datasets are characterized by a huge cross-sectional dimension but a limited temporal dimension, often it is just $T = 1$. Large cross-sectional databases (for example, coming from census or administrative records or medical analyses) fall into this category, which is not so interesting from an economic nowcasting point of view, unless either T is also large enough or the variables are homogeneous enough to allow proper econometric model estimation (e.g. by means of panel methods) and nowcast evaluation. However, Fat datasets can be of interest in many other applications of big data, both inside official statistics, e.g. for surveys construction, and outside, e.g. in marketing or medical studies.

As the collection of big data started only rather recently, Fat datasets are perhaps the most commonly available type. Actually, statistical methods for big data are mainly meant for Fat datasets, e.g. those developed in the machine learning literature, as they only require a large cross-section of i.i.d. variables.

When a (limited) temporal dimension is also present, panel estimation methods are typically adopted in the economic literature, but factor based methods can be also applied (possibly in their 'sparse' version). Classical estimation methods are not so suitable, as their finite (*T*) sample properties are generally hardly known, while Bayesian estimation seems more promising, as it can easily handle a fixed *T* sample size and, with proper priors, also a large cross-sectional dimension.

4.2. Tall datasets

Tall datasets have a limited cross-sectional (*N*) dimension but a big temporal dimension, *T*. This is for example the case with tick by tick data on selected financial transactions, and indeed high frequency data have been long used in financial econometrics. Most of the data types considered in Section 2, if aggregated somewhat across the cross-sectional dimension, could be of the Tall type. For example, total daily cash withdrawals from ATM machines, second by second phone traffic in a given geographical location, hourly cloud coverage of a set of locations resulting from satellite images, or second by second volume of internet searches for a specific keyword. In all these cases *T* is indeed very large in the original time scale, say seconds, but it should be considered whether it is also large enough in the time scale of the target macroeconomic variable of the nowcasting exercise, say quarters. In other words, for nowcasting applications, it is not enough to have a huge *T* in terms of seconds if *T* is instead small when measured in months or quarters, as the evaluation will be typically conducted in the low frequency of the target macroeconomic variable [39].

Tall datasets at very high frequency are not easily obtained, as they are generally owned by private companies. An interesting exception is represented by textual data. For example, using a web-scraper, it is possible to download all the financial news articles appearing on the Reuters terminal over the past 10 years on a daily basis. Next, using proper software for textual analysis, it is possible to transform the unstructured textual data into daily numerical sentiment indexes, which can then later be used as coincident or leading indicators of economic conditions.

Tall datasets generally require substantial pre-treatment, as indicators typically present particular temporal structures (related, e.g. to market micro-structure in the case of financial indicators) and other types of irregularities, such as outliers, jumps and missing observations. Apart from these issues, when *N* is small and *T* is large, classical time series econometric models and methods can be largely used, even though specific features such as changing volatility and possibly parameter time variation should be carefully assessed.

The possible frequency mismatch between (low frequency) target and (high frequency) indicators should be also properly handled, and MIDAS type approaches are particularly promising in a nowcasting context. Moreover, the choice of the proper high frequency scale

[39] Also, the classification of variables such as financial indicators or commodity prices depends on the sampling frequency. If tick by tick data are collected, then the *T* dimension dominates but with daily, and even more with monthly, data the *N* dimension can become dominant, even more so when indicators for many countries are grouped together.

(extent of temporal aggregation) should be also considered, as in general there can be a trade-off between timeliness and precision of the signal.

4.3. Huge datasets

Huge datasets are characterised by very large *N* and *T*. This is perhaps the most interesting type of data in a nowcasting context, and all the data types surveyed in Section 2 are potentially available in Huge format. For example, all the POS financial transactions in a country over a given temporal period, the activity records of all the subscribers to a mobile operator, or all scanned prices for all products in a 24/7 supermarket chain.

In practice, however, unfortunately Huge datasets are not so often available because, as we mentioned, big data collection started only recently, while the collection of the target economic indicators started long ago, generally as far back as the 1950s or 1960s for many developed countries. Moreover, even when available, public access to the Huge data is often not granted, only some cross-sectionally and temporally aggregated measures are made public. Google Trends, publicly available weekly summaries of a huge number of specific search queries in Google, are perhaps the best example in this category, and not by chance the most commonly used indicators in economic nowcasting exercises (see Sections 2 and 3).

Contrary to basic econometrics and statistics, in Huge datasets both *T* and *N* diverge, and proper techniques must be adopted to take this feature into account at the level of model specification, estimation and evaluation. For example, in principle it is still possible to consider information criteria such as BIC or AIC for model specification (indicator selection for the target variable in the nowcasting equation), although it is the case that modifications may be needed to account for the fact that *N* is comparable or larger than *T*, as opposed to much smaller as assumed in the derivations of information criteria. Further, in the case of a linear regression model with *N* regressors, 2^N alternative models should be compared, which is not computationally feasible when *N* is very large, so that efficient algorithms that only search specific subsets of the 2^N possible models have been developed. Moreover, the standard properties of the OLS estimator in the regression model are derived assuming that *N* is fixed and (much) smaller than *T*. Some properties are preserved, under certain conditions, also when *N* diverges but empirically the OLS estimator does not perform well due to collinearity problems that require a proper regularization of the second moment matrix of the regressors. This is in turn relevant for nowcasting, as the parameter estimators are used to construct the nowcast (or forecast). As a result of these problems for OLS, a number of regularisation or penalisation methods have been suggested.

An early approach, referred to as Ridge regression, uses shrinkage to ensure a well behaved regressor sample second moment matrix. More recently, other penalisation methods have been developed. A prominent example is LASSO where a penalty is added to the OLS objective function in the form of the sum of the absolute values of the coefficients. Many related penalisation methods have since been proposed and analysed.

As an alternative to variable selection, the indicators could be summarized by means of principal components (or estimated factors) or related methods such as dynamic principal components or partial least squares. However, standard principal component analysis is also problematic when *N* gets very large, but fixes are available, such as the use of sparse principal component analysis.

Also, rather than selecting or summarizing the indicators, they could be all inserted in the nowcasting regression but imposing tight priors on their associated coefficients, which leads to specific Bayesian estimators.

Finally, in the case of both Fat and Huge datasets, it could be interesting to split them into smaller ones with a limited *N* dimension, apply standard techniques to each sub-dataset, and then re-group the results. For example, one could extract principal components from each sub-dataset, group all of them, and extract their principal components, which provide a summary of the entire big dataset.

Some of these methods can be also of interest for the analysis of Fat datasets, in particular those related to indicator selection or Bayesian estimation, and can therefore be used also outside the nowcasting context.

4.4. Overall conclusions

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. Table 3 attempts to classify — generally — the ten big data categories as fat, tall or huge. However, the researcher must notice that the classification depends heavily on the application needs for big data.

Table 3: 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	
Scanner Prices	X		X
Online Prices	X		X
Online Search		X	

Before drawing a definite conclusion, it is however important to consider more generally the pros and cons of big data for macroeconomic nowcasting or flash estimation, as this can provide additional details for an informed decision, which is particularly relevant in an institutional context. We tackle this issue in the next section.

5. Big data: pros and cons for macroeconomic nowcasting

In this Section, as mentioned above, we evaluate the general advantages and disadvantages of big data for macroeconomic nowcasting or flash estimation.

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.

For the sake of exposition, it is convenient to distinguish data related issues and more methodological potential problems with the use of big data. We consider them, in turn, in the following two subsections, both in general terms and specifically for nowcasting.

5.1. Data issues

A first issue concerns data availability. As it is clear from the data categorization described in Section 2, 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 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.

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 parameter 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.

A second 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 unemployment 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.

For nowcasting, the suggestion is to carefully evaluate the presence of a possible (keyword) selection bias, but this is likely less relevant when internet based data are combined 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.

A third 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.

A fourth issue, again more relevant for digital than standard data collection, is that individuals or businesses could not report truthfully their experiences, assessments 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.

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.

A fifth issue is that data could 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 issue is that the way in which the big data measure a given phenomenon 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 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 agencies 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 internet 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. The initial selection of keywords or a universe of variables is a qualitative issue and must be addressed by the researcher before the empirical investigation. Then, selecting the number of variables or keywords in the underlying universe is an empirical issue.

We have so far focused on the possible cons of big data, as the pros have been widely emphasized both in the general press and in more specialized journals and meetings. Among the main advantages in a nowcasting context, we believe that big data 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 coincident and leading indicators of economic activity.

Finally, 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 be 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.

Overall, our suggestion is to take a pragmatic approach that balances potential gains and costs from the use of big data for nowcasting. Hence, for a specific target variable of interest, such as GDP growth or unemployment, it is worth assessing the marginal gains of big data based indicators that are rather promptly available (such as Google Trends or other variables used in previous studies and made publicly available) with respect to more standard indicators based on soft and hard data.

5.2. 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 methodological 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 comparing alternative indicators and methods over a training sample, selecting 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 complex 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 assumption 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 example, 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. 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 bigger classes of models to be considered and model selection methods, such as information criteria, may need modifications in many respects.

6. Conclusions

In this paper we have provided an overview of the types of available big data potentially useful for nowcasting macroeconomic variables, focusing on: (i) financial markets data, (ii) electronic payments data, (iii) mobile phones data, (iv) sensor data, (v) satellite images data, (vi) scanner prices data, (vii) online prices data, (viii) online search data, (ix) textual data, and (x) social media data. We have also discussed the main empirical applications in a nowcasting/forecasting context, either based on a specific type of big data or with a specific macroeconomic indicator as a target. And we have classified big data based on the relative size of their cross-sectional and temporal dimensions, which is also informative for the proper econometric techniques to be used. Finally, we have discussed the a priori pros and cons of the use of big data, focusing on both data and methodological issues. While there are many pros, and these have been emphasized in many contexts, there are also a few cons that should not be ignored.

Table 4 proposes a summary classification of big data, based on 13 key features that we have identified according to the previous considerations: source, provider, availability, continuity, type, size and sample, meta data, feature, frequency, pre-treatment, link with target, previous use, and required econometrics. For example, online search data are:

- available from private providers (Google, Microsoft, etc.);
- easily available in aggregated form (freely downloadable from internet);
- continuity of provision cannot be guaranteed;
- their type is already numerical (time series);
- the size is limited and user dependent (dependent on the number of inserted search queries or selected number of correlates);
- the sample period (since 2004) is long enough to guarantee a rather reliable econometric analysis and evaluation;
- meta data are not easily available;
- raw data and details on the aggregation process are generally not publicly available;
- the frequency is generally weekly or higher (with daily data possibly available upon request), the release is timely, and the data is not further revised;
- there is not much information about pre-treatment but further treatment to remove outliers and temporal patterns is likely needed;

- the link with specific macroeconomic indicators has been established in previous empirical analyses
- the required econometrics is standard;

Table 4: Taxonomy of big data

Source	(Social Networks/Traditional Business Systems/Internet of Things)
Provider	(Private/Public,National/International,…)
Availability	(Open/Restricted,Free/Fee Based,Standard/Customized, …)
Continuity	(Yes/No/Uncertain)
Type	(Pictures/Binary/Text/Numerical…)
Size and Sample	(Gb and/or *N/T* if numerical)
Meta Data	(Available or not)
Features	(Raw/Transformed/Selected/Aggregated/…)
Frequency	(Low/high, continuous or undetermined, …)
Pre-Treatment	(Outliers, missing observations, seasonal adjustment,measurement errors, …)
Link with Target	(Definition/Economic theory/Empirical/…)
Previous use	(No/Yes, Where/What)
Econometrics	(Standard Methods/big data Specific Methods)

Table 5: Evaluation grid for use of big data for macroeconomic forecasting

A priori assessment of potential usefulness of Big data	Are available nowcasts/forecasts for the target variable biased or inefficient? Is this due to missing information?
	Could timeliness, frequency of release and extent of revision be improved with additional information?
	Is the required missing or additional information available from some type of big data? Is it not available from traditional sources?
	Are there any studies showing the usefulness of big data for forecasting a similar target? How large are the gains?
	Is the temporal dimension of the big data long and homogeneous enough to allow proper evaluation of the resulting nowcasts/forecasts?
Big data Sources	Is big data directly available, perhaps with some investment in data collection, or is a provider needed?
	If a provider is needed, is it public or private? Is it national or international?
	Is access free or fee based? Can it be customized?
	Can data production and access be expected to continue in the future?
	Are there any confidentiality and/or legal issues?
Big data Features	How big is the relevant big data? Does it require specific hardware and software for storing and handling?
	Is it in numerical or non-numerical (text / pictures / binary / etc) format?
	If non-numerical, can it be transformed into numerical format? Does the transformation require specific software? How expensive in terms of time and resources is it expected to be?
	If numerical, is it accessible in clean format or does it requires pre-treatment? How expensive in terms of time and resources is pre-treatment expected to be?
	Is the corresponding big data available for all the units of interest (e.g. countries, sectors, disaggregation levels, various target variables, etc.)?

Table 5: Evaluation grid for use of big data for macroeconomic forecasting

Big data Quality	Does data production and collection satisfy requirements for use from an official institution?
	Is the underlying sampling scheme sufficient to be representative of the entire relevant population?
	Is Meta Data available? Could data collection be replicated?
	If raw data is not available, is the process generating the available summarized or transformed big data clearly explained and reliable?
Big data Econometrics	If big data is in non-numerical format, is there a reliable mapping into numerical format?
	Can data pre-treatment (e.g. outliers removal, missing values imputation, seasonal and other types of adjustment, filtering etc.) be conducted with standard methods?
	Can econometric methods for large datasets (e.g. principal components, shrinkage regressions, large VARs) be adopted or are big data specific methods required (e.g. sparse principal components, special Bayesian priors, genetic algorithms, methods for infinite dimensional matrices, machine learning, etc.)?
	Is it better to work with the entire big data, if available, and related big data econometrics, or with a proper summary of the big data and standard econometric methods?
	How large and robust are, in the end, the forecasting gains when adding big data to traditional indicators? And how large are the gains in terms of timeliness, frequency of release and extent of revision?

It must be noted that a particular caveat of the above is that a dataset may move across these thirteen dimensions depending how a modelling step is done. For example, online search data requires standard econometrics, however this may or may not be needed depending on the choices of the researchers.

The classification in Table 4, as applied to specific big data types as in the example we have provided, can be then also used to create a grid for assessing the need and potential usefulness of big data for macroeconomic forecasting, in general and in official institutions such as Eurostat, according to the outcome of a careful cost-benefit analysis. The proposed grid is presented in Table 5.

As an example, let us assume that we are considering the use of Google trends (online search data) to nowcast unemployment and, according to the analysis based in Table 4, we have concluded that they are indeed potentially useful. We should then consider issues such as:

- the quality of the current unemployment nowcasts and the need for improvement (which depends on the specific country under analysis);
- the possibility of improving the timeliness, frequency of release and extent of revision of the nowcasts when using the online search data (all features can be potentially improved, as online search data are timely available, at least at weekly frequency and not revised);
- the availability of other traditional sources of information (e.g. business surveys with also questions related to future employment perspectives);
- the availability of previous studies showing the usefulness of the specific selected big data type for unemployment nowcasting and the extent of the reported gains;
- the availability of a long enough sample span for an econometric analysis;
- the availability of a sufficient level of disaggregation, e.g. sectors or regions, or age groups, or qualifications (which is not the case, at least not for the freely available data);
- the representativeness of the underlying sample (which only covers internet users);

- the resources required for the big data collection, storage, treatment and analysis (which are limited in this example);
- the possibility that access to the big data could be discontinued or no longer be free (which could indeed happen);
- the presence of confidentiality or legal issues (which could emerge, in particular with private data used by a public institution);
- the need of big data specific econometric techniques (which is not the case in this example as the big data have been pre-aggregated by the provider);
- the type and extent of the marginal gains emerging from the use of the big data from a proper econometric evaluation (for example, the out of sample increase in the unemployment nowcasts precision but also in their timeliness, frequency, and extent of revision).

Overall, a careful selection of the relevant big data, combined with the use of proper econometric methods to formally handle and analyse their relationship with the target macroeconomic indicator, has substantial potential to sharpen predictions and make them more timely and frequent.

In terms of specific big data types, electronic payments data, scanner price and online price data, online search data, textual data and social media data are all promising for nowcasting.

Acknowledgements

The authors thank two anonymous referees for their helpful and constructive comments. The authors would like to thank their Eurostat colleagues Emanuele Baldacci and Martina Hahn for their support, Fernando Reis and Dan Rieser for their feedback. They thank also Diego Kuonen for the fruitful discussion around the use of Big Data in Official Statistics. They are also grateful to the members of the Working Group Methodology for their constructive comments on earlier version of the paper. A special thanks is also extended to Sabine Bankwitz. The views expressed are those of the authors and do not necessarily reflect the views of their current or former employer.

References

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

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

Alesina, A., S. Michalopoulos and E. Papaioannou (2016), 'Ethnic Inequality', *Journal of Political Economy*, 124(2), pp. 428–488.

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

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

Antweiler, W. and M.Z. Frank (2004), 'Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards', *The Journal of Finance*, 59(3), pp. 1259–1294.

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

Arias, M., A. Arratia and R. Xuriguera (2013), 'Forecasting with Twitter Data', *ACM Transactions on Intelligent Systems and Technology*, 5(1), p. 8.

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

Baker, S.R., N. Bloom and S.J. Davis (2015), 'Measuring Economic Policy Uncertainty', *NBER Working Paper Series*, Working Paper 21633.

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

Banbura, M. and G. Runstler (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, pp. 333–346.

Bangwayo-Skeete, P.F. and R.W. Skeete (2015), 'Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach', *Tourism Management*, 46, pp. 454–464.

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

Bazzani, A., L. Giovannini, R. Gallotti and S. Rambaldi (2011), 'Now casting of traffic state by GPS data. The metropolitan area of Rome', Conference paper, January 2011, available at https://goo.gl/kcpPhm.

Beiler, J. (2013), 'Noise in GDELT', Personal Blog, October 28, 2013, available at: https://goo.gl/KC9iaJ.

Bendler, J., S. Wagner, T. Brandt and D. Neumann (2014), 'Taming Uncertainty in Big Data: Evidence from Social Media in Urban Areas', *Business & Information Systems Engineering,* 05-2014, pp. 279–288.

Bholat, D., S. Hansen, P. Santos and C. Schonhardt-Bailey (2015), 'Text mining for central banks', *Centre for Central Banking Studies*, (33), pp. 1–19.

Boettcher, I. (2015), 'Automatic Data Collection on the Internet (Web Scraping)', New Techniques and Technologies for Statistics, Eurostat Conference, 9–13 March 2015.

Bollen, J., H. Mao and X. Zeng (2011), 'Twitter mood predicts the stock market', *Journal of Computational Science*, 2(1), pp. 1–8.

Bontempi, M.E., R. Golinelli and M. Squadrani (2016), 'A new index of uncertainty based on internet searches: a friend or a foe of other indicators?', *Quaderni*, Working Paper DSE No 1062.

Breton, R., N. Swiel and R. O'Neil (2015), 'Using Web Scraped Data to Construct Consumer Price Indices', New Techniques and Technologies for Statistics, Eurostat Conference, 9–13 March 2015.

Campbell, J.R. and B. Eden (2014), 'Rigid Prices: Evidence from U.S. Scanner Data' *International Economic Review*, 55(2), pp. 423–442.

Capgemini and BNP Paribas (2016), 'World Payments Report', Available online at https://goo.gl/niAkIl.

Carlsen, M. and P.E. Storgaard (2010), 'Dankort payments as a timely indicator of retail sales in Denmark', Danmarks National bank, Working Paper 2010-66.

Carriere-Swallow, Y. and F. Labbe (2013), 'Nowcasting with Google Trends in an Emerging Market', *Journal of Forecasting*, 32, pp. 289–298.

Cavallo, A. (2013), 'Online and official price indexes: Measuring Argentina's inflation', *Journal of Monetary Economics*, 60, pp. 152–165.

Cavallo, A. (2016), 'Scraped Data and Sticky Prices', *Review of Economics & Statistics*, forthcoming.

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

Cavallo, A. and R. Rigobon (2016), 'The Billion Prices Project: Using Online Prices for Measurement and Research', *The Journal of Economic Perspectives*, 30(2), pp. 151–178.

Cerchiello, P. and P. Giudici (2014), 'How to Measure the Quality of Financial Tweets', Working Paper, ECB Workshop on using big data for forecasting and statistics, 07-08/04/2014, Frankfurt.

Chamberlain, G. (2010), 'Googling the present', *Economic and Labour Market Review*, Office for National Statistics, 4(12).

Chan, K.Y., S. Khadem, T.S. Dillon, V. Palade, J. Singh and E. Chang (2012), 'Selection of Significant On-Road Sensor Data for Short-Term Traffic Flow Forecasting Using the Taguchi Method', IEEE *Transactions on Industrial Informatics*, 8(2), pp. 255–266.

Chen, X. and W.D. Nordhaus (2011), 'Using luminosity data as a proxy for economic statistics', *PNAS*, 108(21), pp. 8589–8594.

Chevalier, J.A., A.K. Kashyap and P.E. Rossi (2003), 'Why don't prices rise during periods of peak demand? Evidence from scanner data', *American Economic Review*, 93(1), pp. 15–37.

Choi, H. and H. Varian (2009), 'Predicting initial claims for unemployment benefits', *Google Working Paper*.

Choi, H. and H. Varian (2012), 'Predicting the Present with Google Trends', *Economic Record*, 88(1), pp. 2–9.

Conover, M.D., B. Goncalves, J. Ratkiewicz, A. Flammini and F. Menczer (2011), 'Predicting the Political Alignment of Twitter Users', 2011 IEEE International Conference on Privacy, Security, Risk ,and Trust, and IEEE International Conference on Social Computing.

Dal Pozzolo, A., O. Caelen, R.A. Johnson and G. Bontempi (2015), 'Calibrating Probability with Undersampling for Unbalanced Classification', In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015.

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

de Haan, J. (2015), 'A Framework for Large Scale Use of Scanner Data in the Dutch CPI', Report from Ottawa Group 14th meeting, International Working Group on Price Indices, Tokyo (Japan), May, 2015.

De Meersman, F., G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis and H.I. Reuter (2016), 'Assessing the Quality of Mobile Phone Data as a Source of Statistics', European Conference on Quality in Official Statistics, Madrid, 31 May–3 June 2016.

Deloitte (2012), 'What is the impact of mobile telephony on economic growth? A report for the GSM association', November 2012.

Deville, P., C. Linard, S. Martin, M. Gilbert, F.R. Stevens, A.E. Gaughan, V.D. Blondel and A.J. Tatem (2014), 'Dynamic population mapping using mobile phone data', PNAS, 111(45), pp. 15888–15893.

Donaldson, D. and A. Storeygard (2016), 'The View from Above: Applications of Satellite Data in Economics', *Journal of Economic Perspectives*, 30(4), pp. 171–198.

Doornik, J. A. and D.F. Hendry (2015), 'Statistical Model Selection with Big Data', *Cogent Economics & Finance*, 3(1), 2015.

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

Ebner, J. (2015), 'How Sensors will Shape Big Data and the Changing Economy', Dataconomy, published online: January 27, 2015, available at https://goo.gl/9h3Zf8.

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

Elvidge, C.D., P.C. Sutton, T. Ghosh, B.J. Tuttle, K.E. Baugh, B. Bhaduri and E. Bright (2009), 'A global poverty map derived from satellite data', Computers & Geosciences, 35, pp. 1652–1660.

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.

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.

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

Fernandez-Ares, A.J., A.M. Mora, M.G. Arenas, P. de las Cuevas, P. Garcia-Sanchez, G. Romero, V. Rivas, P.A. Castillo and J.J. Merelo (2016), 'Nowcasting Traffic', Working Paper.

Ferreira, P. (2015), 'Improving Prediction of Unemployment Statistics with Google Trends: Part 2', Eurostat Working Paper.

Florida, R. (2014), 'The Economic Data Hidden in Nighttime Views of City Lights', Citylab, published online: May 29, 2014, available at https://goo.gl/uaOYdV.

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

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

Giannone, D., L. Reichlin and D. Small (2008), 'Nowcasting: The Real-Time Informational Content of Macroeconomic Data', *Journal of Monetary Economics*, 55, pp. 665–676.

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

Graham, D. (2016), 'How the Internet of Things Changes Big Data Analytics', DataInformed, published online: August 9, 2016, available at https://goo.gl/M22uje.

Griffioen, R., J. de Haan and L. Willenborg (2014), 'Collecting Clothing Data from the Internet', Statistics Netherlands Technical Report.

Griswold, A. (2014), 'Can Twitter Predict Economic Data?', Slate, Moneybox: A blog about business and economics, April 4, 2014, Available at: https://goo.gl/DHXOJR.

Heaven, D. (2013), 'World's largest events database could predict conflict', *New Scientist*, May 8, 2013, available at: https://goo.gl/U0VOLI.

Henderson, J.V., A. Storeygard and D.N. Weil (2011), 'A Bright Idea for Measuring Economic Growth', *American Economic Review: Papers & Proceedings*, 101(3), pp. 194–199.

Henderson, J.V., A. Storeygard and D.N. Weil (2012), 'Measuring Economic Growth from Outer Space', *American Economic Review*, 102(2), pp. 994–1028.

Heston, S.L. and N.R. Sinha (2014), 'News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns', Working Paper.

Ivancic, L., W.E. Diewert and K.J. Fox (2011), 'Scanner data, time aggregation and the construction of price indexes', *Journal of Econometrics*, 111, pp. 24–35.

Jones, J. (2010), 'Nowcasting Applications of GPS Water Vapour Networks', UK Met Office, presentation.

Kapetanios, G., M. Marcellino and F. Papailias (2016), 'Big Data and Macroeconomic Nowcasting', Eurostat Working Paper, ESTAT No 11111.2013.001-2015.278.

Kearns, J. (2015), 'Cheap orbiting cameras and big-data software reveal hidden secrets of life down below', Bloomberg Business Week, published online: July 9, 2015, Available at https://goo.gl/fjhPnm.

Keola, S., M. Andersson and O. Hall (2015), 'Monitoring Economic Development from Space: Using Nighttime Light and Land Cover Data to Measure Economic Growth', *World Development*, 66, pp. 322–334.

Koop, G. and L. Onorante (2013), 'Macroeconomic Nowcasting Using Google Probabilities', European Central Bank Presentation.

Lansdall-Welfare, T., V. Lampos and N. Cristianini (2012), 'Nowcasting the mood of nation', Significance, *The Royal Statistical Society*, August 2012.

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

Lloyd, T.A., S. McCorriston, C.W. Morgan, E. Poen and E., Zgovu (2012), 'Retailer Heterogeneity and Price Dynamics: Scanner Data Evidence from UK Food Retailing', Working Paper No. 8, *Transparency of Food Pricing*, TRANSFOP.

Lowe, M. (2014), 'Night Lights and ArcGIS: A Brief Guide', *MIT Economics Working Paper.*

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

Makhazanov, A. and D. Rafiei (2013), 'Predicting Political Preference of Twitter Users', 2013 IEEE/ ACM International Conference on Advances in Social Networks Analysis and Mining, August 25–29, 2013.

Makrehchi, M., S. Shah and W. Liao (2013), 'Stock Prediction Using Eventbased Sentiment Analysis', 2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT).

Manyika, J., M. Chui, P. Bisson, J. Woetzel, R. Dobbs, J. Bughin and D. Aharon (2015), 'The Internet of Things: Mapping the Value beyond the Hype', McKinsey Global Institute, June 2015.

Mao, H., X. Shuai, Y.-Y. Ahn and J. Bollen (2015), 'Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to Cote d'Ivoire', EPJ Data Science, 4(15).

Mass, C. (2012), 'Nowcasting: The Promise of New Technologies of Communication, Modeling and Observation', *Bulletin of the American Meteorological Society*, 93(6), pp. 797–809.

Mellander, C., J. Lobo, K. Stolarick and Z. Matheson (2015), 'Night-Time Light Data: A Good Proxy Measure for Economic Activity?' PLoS ONE, 10(10).

Metcalfe, E., T. Flower, T. Lewis, M. Mayhew and E. Rowland (2016), 'Research indices using web scraped price data: clustering large datasets into price indices (CLIP)', Office for National Statistics, Release date: 30 November 2016.

Mittal, A. and A. Goel (2012), 'Stock Prediction Using Twitter Sentiment Analysis', Standford University Working Paper.

Moat, H.S., C. Curme, A. Avakian, D.Y Kenett, H.E. Stanley and T Preis (2013), 'Quantifying Wikipedia Usage Patterns Before Stock Market Moves', *Scientific Reports*, 3:1801, pp. 1–5.

Modugno, M. (2013), 'Now-casting Inflation using High-Frequency Data', *International Journal of Forecasting*, 29, pp. 664–675.

Nyman, R., D. Gregory, S. Kapadia, R. Smith and D. Tuckett (2014a), 'Exploiting Big Data for Systemic Risk Assessment: News and Narratives in Financial Systems', Working Paper, ECB Workshop on using big data for forecasting and statistics, 07-08/04/2014, Frankfurt.

Nyman, R., P. Ormerod, R. Smith and D. Tuckett (2014b), 'Big Data and Economic Forecasting: A Top-Down Approach Using Directed Algorithmic Text Analysis', Working Paper, ECB Workshop on using big data for forecasting and statistics, 07-08/04/2014, Frankfurt.

Pandya, S.S. and R. Venkatesan (2015), 'French Roast: Consumer Response to International Conflict — Evidence from Supermarket Scanner Data', *Review of Economics & Statistics*, 98(1), pp. 42–56.

Papadimitriou, S., J. Sun, C. Faloutos and P.S. Yu (2013), 'Dimensionality Reduction and Filtering on Time Series Sensor Streams', In: *Managing and Mining Sensor Data*, Aggarwal, C.C. (Eds.), Springer, 2013.

Paul, M.J., M. Dredze and D. Broniatwoski (2014), 'Twitter Improves Influenza Forecasting', *PLOS Currents Outbreaks*, October 28, 2014.

Pineiro-Chousa, J., M. Vizcaino-Gonzalez and A.M. Perez-Pico (2016), 'Influence of Social Media over the Stock Market', *Psychology & Marketing*, 34(1), pp. 101–108.

Radzikowski, B. and A. Smietanka (2016), 'Online CASE CPI', First International Conference on Advanced Research Methods and Analytics, CARMA2016.

Ramli, D. (2012), 'Treasury to mine Twitter for economic forecasts', Financial Review, October, 30, 2012, Available at: https://goo.gl/xHUhE2.

Reis, F., P. Ferreira and V. Perduca (2015), 'The Use of Web Activity Evidence to Increase the Timeliness of Official Statistics Indicators', Eurostat Working Paper.

Ricciato, F., P. Widhalm, M. Craglia and F. Pantisano (2015), 'Estimating population density distribution from network-based mobile phone data', European Commission, JRC technical report, EUR 27361 EN.

Rigobon, R. (2015), 'Presidential Address: Macroeconomics and Online Prices', *Economia*, 15(2), pp. 199–213.

Rönnqvist, S. and P. Sarlin (2015), 'Bank Networks from Text: Interrelations, Centrality and Determinants', *Quantitative Finance*, 15(10), pp. 1619–1635.

Ross, A. (2013), 'Nowcasting with Google Trends: a keyword selection method', Fraser of Allander Economic Commentary, 37(2), pp. 54–64.

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

Silver, M. and S. Heravi (2001), 'Scanner Data and the Measurement of Inflation', *The Economic Journal*, 111, F383-F404.

Smith, N.A. (2010), 'Text-Driven Forecasting', Carnegie Mellon University, Working Paper.

Smith, P. (2016), 'Google's MIDAS Touch: Predicting UK Unemployment with Internet Search Data', *Journal of Forecasting*, 35, pp. 263–284.

Smith-Clarke, C., A. Mashhadi and L. Capra (2014), 'Poverty on the cheap: estimating poverty maps using aggregated mobile communication networks', CHI 2014 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 511–520.

Sprenger, T.O., A. Tumasjan, P.G. Sandner and I.M. Welpe (2014a), 'Tweets and Trades: the Information Content of Stock Microblogs', European Financial Management, 20(5), pp. 926–957.

Statistics New Zealand (2014), *Measuring price change for consumer electronics using scanner data*, ISBN 978-0-478-42940-4.

Stock, J. and M. Watson (2002a), 'Forecasting Using Principal Components from a Large Number of Predictors', *Journal of the American Statistical Association*, 297, pp. 1167–1179.

Stock, J. and M. Watson (2002b), 'Macroeconomic Forecasting using Diffusion Indexes', *Journal of Business & Economics Statistics*, 20, pp. 147–162.

Suryadevara, N.K., S.C. Mukhopadhyay, R. Wang and R.K. Rayudu (2013), 'Forecasting the behavior of an elderly using wireless sensors data in a smart home', *Engineering Applications of Artificial Intelligence*, 26, pp. 2641–2652.

Tefft, N. (2011), 'Insights on unemployment, unemployment insurance, and mental health', *Journal of Health Economics*, 30(2), pp. 258–264.

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.

Tumasjan, A., T.O. Sprenger, P.G. Sandner and I.M. Welpe (2010), 'Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment', Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

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

Vosen, S. and T. Schmidt (2011), 'Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends', *Journal of Forecasting*, 30, pp. 565-578.

Wang, H., D. Can, A. Kazemzadeh, F. Bar and S. Narayanan (2012), 'A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle', Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 115-120.

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