Big data, official statistics and measuring the economy

*Didier Blanchet, Pauline Givord**

The proliferation of digital traces generated by the activity of individuals or companies and the growing capacity to store and analyse them have given rise to what is known as the "Big Data" phenomenon. Using large quantities of individual data is obviously nothing new for official statistics, which already process data from surveys, censuses and a wide variety of administrative sources. However, the advent of big data does introduce two major shifts in the form of much larger volumes and almost-immediate access. There are a number of obstacles, nonetheless, to making the most of these advantages, as these data are themselves not always without their defects: they come in complex or unstructured formats, using them can require costly investments in specific technologies, they are not always guaranteed to be representative, and the same applies to their predictive power.

This focus proposes a review of the contribution of these data to three aspects of the measurement of the economy. First of all, short-term monitoring: does analysis of internet search behaviour or of the online press enable us to anticipate the short-term economic climate more effectively than survey data? As things stand, the answer to this question is still somewhat guarded.

The second aspect is tracking prices. In this respect, the contribution of big data is already shown to be much more tangible, whether in the form of prices collected online or scanner data provided by retail chains. Finally, we present some attempts to use big data to better capture the general phenomenon they are stemming from, the boom in the digital economy sector.

New data sources

The term "big data" has become widely publicised in recent years. This term was first coined as a result of the explosion in the volume of data produced by the internet giants, and in some scientific fields (genomics and astronomy in particular). It has been accompanied by some impressive advances in the techniques available for storing and processing these data which are not only very voluminous and varied (in text or image format), but are also produced in continuous flows.

These characteristics are often defined as the "three Vs" for Volume, Variety and Velocity, as first suggested in a report by McKinsey in 2011. Some also add a fourth V for "Veracity", considering the information collected as objective by nature, and even a fifth V for "Value", highlighting the economic interest in exploiting these data.

National Statistical Institutes are also interested in the potential of these big data. Better exploring their possibilities was one of the recommendations of the Bean review of official UK economic statistics, published at the beginning of 2016 (Bean, 2016). Various actions are currently underway at the international level. In particular, a network has been set up under the authority of Eurostat to promote the sharing of experiences among European Statistical

* Didier Blanchet, Pauline Givord, INSEE.

Institutes, of which INSEE is one.[1] These experiments cover a variety of fields: using scanner data to improve the price index, using satellite data to describe land use or forecast crop yields, exploiting social network data to forecast household confidence, using data from credit cards or mobile phones to improve tourism statistics, and data from smart electricity meters to measure energy consumption, etc.

There are several incentives for using this type of data. One first aim is to improve and supplement existing statistical production. By using high-frequency data, it is hoped that some indicators can be published still earlier than is presently the case. The mass of available data can also be used to produce indicators at a higher level of granularity (on sub-categories or sub-populations) or at more detailed levels, without placing any extra burden on respondents. With these new sources it is also possible to envisage reductions in the costs of collection, even if expected gains have to be offset against investments needed to process the data. Finally, using original data can complement the description of the economy provided by official statistics in "emerging" areas, such as the digital economy or the implementation of sustainable development indicators.

However, exploiting these data for official statistics raises several questions. The first relates to their quality. They often focus only on a limited field (internet users, customers of a specific chain of stores or mobile phone operator) and assessing how representative they are of the general population is not necessarily easy to do. Second, the information obtained never corresponds exactly to the concept that it is hoped to measure, unlike with data obtained from a survey whose questions are designed to get as close as possible to the definition of the phenomenon being investigated, such as a company's sector of activity or a person's employment situation. A third concern, which is just as important for official statistics, is that of durability. The indicators that are produced must be comparable over time. By using external data, there is a risk of breaks over which we have no control as a result of changes in their format or in their collection method. There is also the question of an ethical and legal framework that can guarantee long-term access to these data and which respects personal privacy and commercial confidentiality.

Three examples are presented to illustrate these different issues. The first is that of short-term macroeconomic monitoring: are big data available in real time capable of "outperforming" the predictive power of business tendency surveys? The second relates to measuring prices: can big data replace traditional survey-based methods? The third is measuring the digital economy. Traditional sources are not always suitable for quantifying emerging activities. Big data are one of the outputs of the digital economy, and therefore potentially well placed to contribute to measuring its impact.

Using information as close to events as possible: can "nowcasting" live up to its promise?

Steering public policy requires an accurate and rapid diagnosis of the health of the economy. Reducing deadlines for publication of the main economic indicators is therefore an important issue for official statistical institutes. The annual national accounts give a complete picture of the economy but they are only available after the necessary lapse of time for gathering and comparing all the sources on which they are based, tax sources in particular. In France, full annual accounts for year n are not published until May of year *n*+1, and they continue to be revised during the following two years. In order to obtain estimates more quickly, quarterly accounts are based on advance quantitative indicators, such as the industrial production index

1. *https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data*

or turnover indices: INSEE recently reduced the delay in producing the first aggregates of these accounts, which is now only thirty days after the end of the quarter. For even earlier assessments, the qualitative information collected on a monthly basis for the business tendency surveys has to be used. This is moving towards what is called "nowcasting", or the use of indicative data to "forecast" a present for which all the details will only be known at a much later date. The same problem arises when monitoring employment or unemployment. Quarterly employment is not known until 45 days after the end of the quarter. The number of jobseekers registered with the French unemployment agency *(Pôle emploi)* is monitored monthly and published at the end of the following month, but ILO unemployment, whose definition is harmonised and which is more stable over time, is collected by the Employment Survey, whose sampling is such that only quarterly monitoring is possible and whose first results are not published until one and a half month after the end of the quarter under consideration.

Certain types of big data can claim to contribute to this nowcasting. As these data are available virtually instantaneously, it seems possible to monitor economic or social news almost in real time. But this will only happen if these data are sufficiently closely linked with the phenomena in which we are interested. Here we consider two cases: using data from internet searches, as collected and made available by Google, and data from a "media sentiment" indicator constructed from online press articles.

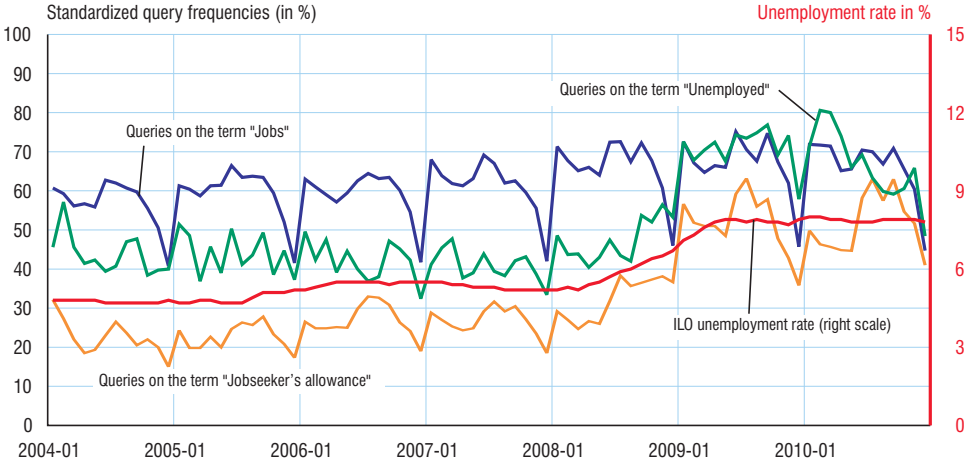
Using internet searches to forecast economic fluctuations: an avenue that is still limited

The idea of enhancing the short-term economic diagnosis by looking at the frequency of certain internet search terms was popularised by Choi and Varian (2009). Intuitively, the belief is that, given the widespread use of the internet, internet search queries must reflect the concrete activity of most economic actors. For example, it has become common to search on the internet before making a purchase, especially for major items like a new car or a household appliance. An increase in the number of searches corresponding to terms like "car" or "washing machine" therefore suggests an increase in the consumption of these goods. Similarly, a rise in unemployment is likely to be linked with an increase in searches for terms like "employment" or "unemployment insurance", even before this rise in unemployment appears in the *Pôle Emploi* figures and certainly before it is seen in the Labour Force Survey.

Testing this intuition was facilitated by the availability of the Google Trends tool which shows up changes in internet searches over time, from 2004, for precise terms or terms grouped together into larger categories and for various geographic areas. For example, in the United Kingdom, McLaren and Shanbhogue (2011) observed that searches for the keywords "unemployed" and "jobseeker's allowance" showed up fairly accurately the surge in unemployment during the 2008 crisis *(Figure 1)*. The same type of correlation can be seen for France *(Figure 2)*. However, this similarity in overall movement is not enough to guarantee a high predictive power in real time: searches for "unemployment insurance" may reflect other movements than simply ILO unemployment, such as the different categories of end-of-month jobseekers (DEFM); they may also increase independently of any real change in unemployment, for example in the case of changes in the rules on compensation.

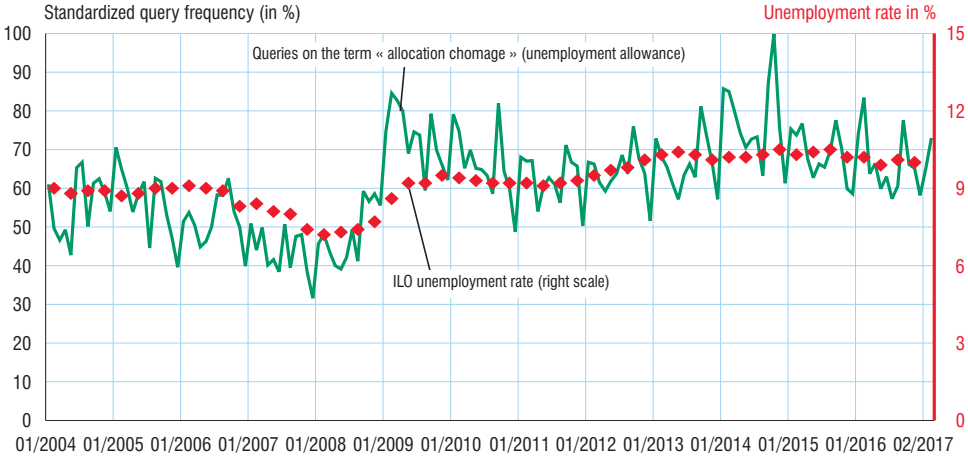
To analyse the predictive power of these series more precisely, they must be inserted into explanatory models of the variable we are trying to predict and tested in anticipation. In the case of France, Fondeur and Karamé (2013) confirmed the predictive power of the Google Trends series for end-of-month jobseekers in the 15-24 year-old age bracket, but only in comparison with an autoregressive model with no other explanatory variable. In the case of the United Kingdom, McLaren and Shanbhogue (2011) showed that the search indicator for

1. Google Trends queries and unemployment rate in the United Kingdom



Note: As in the original study, each series is the average of seven successive Google Trends independent consultations, however, the series were not seasonally adjusted.
Source: Google Trends and Office for National Statistics, according to McLaren and Shanbhogue (2011).

2. Google Trends queries and unemployment rate in France



Source: Google Trends and Insee.

the keyword "jobseeker's allowance" accurately reflected past changes in ILO unemployment in the UK, but the performance was slightly inferior to the administrative count of the number of unemployed receiving benefits and to the opinion on prospects for change in unemployment expressed in a household survey. In anticipation, its predictive ability fell midway between the other two variables. There is therefore no decisive advantage to be gained for economic forecasters or analysts.

In an edition of *Conjoncture in France* published by INSEE, Bortoli and Combes (2015) looked in more depth at this question of the predictive value of Google Trends series for one of the most important items of gross domestic product, monthly household consumption expenditures on goods or services. According to their conclusions, it is certainly possible to highlight positive correlations between search frequency for certain keywords and the

purchasing behaviour ultimately observed, but only for a few highly targeted items, such as clothing, sports articles or housing equipment, and with gains in forecasting which remain very modest, with a fall in mean forecast errors of around 5 to 10% compared with simple autoregressive models which forecast these expenditures based merely on past changes.

The low level of these gains must also be balanced against the risks inherent in using these sources. These risks were clearly highlighted in a case that is outside the scope of economics but from which lessons can certainly be transposed, the case of the Google Flu indicator, set up in 2008 in order to track the seasonal influenza epidemic. This indicator was also based on internet searches with the same initial idea that, with the appearance of flu-like symptoms (fever, headaches, other pains, etc.), a common reflex is to search for information on the internet, even before consulting a doctor. When it was published, this indicator appeared to be actually ahead of the official figures produced by the American Centers for Disease Control and Prevention. However, despite this promising start, Google Flu proved to be rather unreliable in forecasting: it very often produced overestimates of the epidemic peaks compared with what was finally observed. The indicator has not been updated since 2015.

This failure was studied in detail by Lazer *et al*. (2014). One of the limitations of exploiting the Google Trends data is their instability. The search engine is constantly being modified - to improve the service to users - for example, by providing automatic search suggestions. These suggestions influence users' queries. Users are also influenced by external events. A media frenzy about an ongoing influenza epidemic increases the likelihood of queries around search terms associated with the flu but in no way reflects the real severity of the epidemic. There is another source of time inconsistency in that series supplied by Google Trends do not correspond directly to an exhaustive count of the terms selected: they are obtained by taking samples across all the search terms, then applying a number of reprocessing procedures. For example, to avoid producing a trend increase as a result of the rise in internet use since the tool was created, the series are standardised. These are legitimate reprocessing techniques, but they may have important consequences. As illustrated in Bortoli and Combes (2015), Google Trends may produce two different time series, one week apart, for an identical search term, without the internet user having any visibility over these changes: there is very little documentation available on this tool to identify these changes and have control over the consequences.

Using Google Trends series for forecasting can be further hampered by problems of a more technical nature. Although the raw material (search queries by internet users) is vast, the series recovered on the site are in fact small in size as they date back to 2004 at the earliest. On the other hand, the number of terms that can be used to account for the time fluctuations in the phenomenon being considered is very high. In the absence of any initial expertise on the subject, one is tempted to keep the largest possible set of categories of search terms. There is a risk that, among these many variables, some have a correlation with the relevant variable which is purely coincidental and does not reflect a "real" link. Relying on correlations of this kind can distort the forecast, and the more variables used for the forecast compared with the time depth of the series being considered, the greater the risk. The model would certainly be able to explain the past perfectly, but would be incapable of correctly anticipating future variations in the variable of interest: this is called "overfitting". For example, in the case of Google Flu, Lazer *et al*. (2014) observed that the model selected "seasonal" terms such as basketball games, which are very popular in the United States and which take place during the winter. These events often coincided with flu epidemics due simply to the calendar, but this in no way suggested a causal link between the two. If the flu epidemic happens a little earlier or later than usual, selecting these seasonal terms in the forecast model has an adverse effect on performance.

This type of problem is well known to forecasters. They can reduce its impact by applying rigorous selection protocols to the explanatory variables, as described by Bortoli and

Combes (2015). However, the problem of the instability of the series produced by Google Trends still remains. One basic principle of a forecast based on an estimate from a model is that what is measured by a variable today is identical to what it measured in the past. This is a necessary condition for the correlation estimated in the past to be extrapolated to the current period. It cannot be satisfied if the time variations in a series are altered by technical changes introduced into the search engine. Basing a forecasting model on a source whose construction cannot be controlled or traced creates a significant risk of obtaining unreliable estimates.

Can we predict the present by analysing the online press?

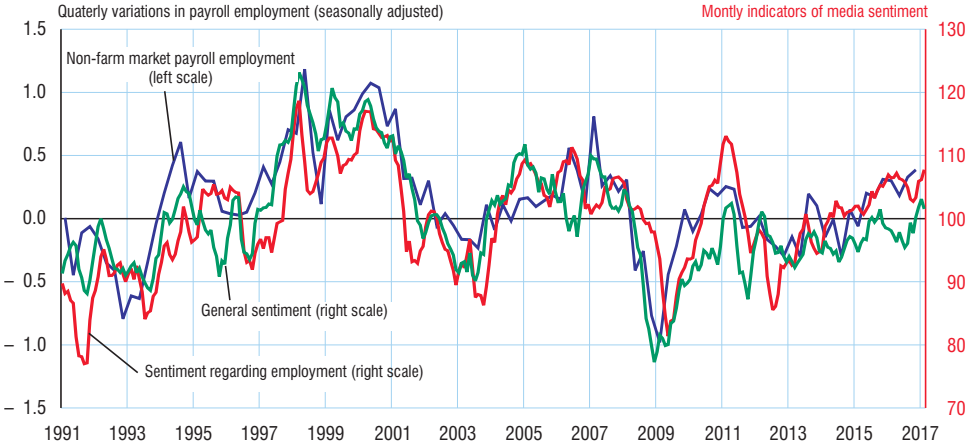
Another obstacle to using Google Trends data is that they are only very indirectly related to the phenomenon under investigation. The really useful information (e.g. the intention to actually purchase a new vehicle when forecasting the consumption of goods) is drowned out by the mass of queries that are unrelated to economic activity (e.g. a scandal in the automobile sector).

Using data from the economic press is one way to respond to this criticism, in part at least, by refocusing on sources more directly related to what we hope to measure. It also allows us to regain control over the entire data construction process. In a report in another edition of *Conjoncture in France*, Bortoli *et al*. (2017) tested this method for constructing advance indicators of payroll employment or of the general economic sentiment, based on a corpus of online articles from the *Le Monde* newspaper. The basic idea was that the economic climate is reflected in the level of optimism expressed in the tone of articles. It should therefore be possible to anticipate economic fluctuations (e.g. the labour market situation) from the accumulation of terms describing a favourable or unfavourable situation.

If this assumption is correct, this information could be mobilised very quickly. Another advantage of this process is that it gives a better time perspective than the Google Trends series: the articles analysed dated back to 1991, giving a total of 1.3 million articles. In practice, a considerable amount of data processing was required using text analysis, or "text mining". First, articles relating to the economic situation were singled out, bringing the number to about 200,000, then the general tone of each of them was assessed by first identifying significant terms that could be correlated with the economic situation (excluding linking words or words related very specifically to the subject of the article). Finally, these terms were assessed as to whether they corresponded to a positive assessment (e.g. with terms like "improvement", "favourable", "stability") or a negative assessment (e.g. "weakening", "instability", "problem"). Each article was then assigned a "sentiment score" based on the number of positive or negative terms and was examined to see to what extent the average score was correlated with the objective economic situation.

There is indeed a link between the quarterly variation in non-farm market employment and two indices obtained using this method *(Figure 3)*: a sentiment index on employment and a general sentiment index. Observing that the textual analysis successfully produces indices that follow the economic movement fairly closely overall is a remarkable result in itself, showing the power of these textual analysis methods, with this time the advantage of controlling them from start to finish. Nevertheless, as with the Google Trends data, extracting relevant information from the mass of available articles requires complex processing techniques which, at this stage, do not yet make any decisive contribution to forecasting compared with conventional employment forecasting methods. While the resulting series showed a good predictive performance for the past period, using them in real conditions would require being sure of the stability of the editorial lines for all the titles that we would choose to follow. Basing statistical findings on press articles could also pose problems of circularity between the statistical finding and reactions to this finding *(Box 1)*.

3. Quaterly variations in payroll employment in France and monthly media sentiment indicator



Note: a moving average of order 5 has been applied to the media sentiment indicators.
Source: Bortoli, Renault and Combes (2017).

Box 1

Internet data and the media, nowcasting and self-fulfilling prophecies

Using data from the internet or from the online press may pose problems of circularity, both in nowcasting and in forecasting. One reason is that data or forecasts published by the bodies in charge of the current climate are usually picked up in press articles. Using these press data without the necessary care would result in these bodies basing part of their current observations or forecasts on their own past observations or forecasts. However, in the case of the study by Bortoli *et al.* (2017), this problem did not appear to be invalidating: the results were robust when articles containing the names "INSEE", "DARES" or "*Pôle Emploi*" were excluded.

But there can also be two-way links with real activity. The events that the press refers to illustrate an objective economic situation. However, in return, the level of optimism shown by the media influences economic behaviour. These mechanisms are well known on the financial markets: a media announcement usually has immediate repercussions on stock market indices. Using these same methods to analyse sentiment, Tetlock (2007) and Engelberg and Parsons (2011) show that press articles have a specific effect on stock price fluctuations. Soo (2015) observes a similar phenomenon in the formation of real estate prices in several large cities in the United States between 2000 and 2011: all those involved adapted their behaviour by mimicking what they believed to be the general movement, thus creating this movement.

This risk of self-fulfilling prophecies is in principle less pronounced in the case of real economic activity than it is in highly speculative markets, but it may nevertheless exist (Blanchard *et al.*, 2017). This phenomenon does not detract from the performance of the prediction models, in fact it even tends to enhance it, but at the same time it would strengthen the natural instability of the economy. Hence the importance of anchoring forecasts to the most objective and the most independent information possible, to minimise the risk of self-fulfilling cycles with no real motivation. Ultimately, we cannot exclude the possibility of manipulation. As the use of some terms may influence the level of the economic indicator, some actors may be tempted to increase the presence of these terms on the internet or in the media, with a positive or negative connotation, depending on the desired results.

Big data and price measurement

The examples described above certainly open up some very interesting areas for further research. Measuring internet search behaviour or media sentiment are relevant subjects that deserve being explored, either for themselves or for their link to other questions, such as the measurement of subjective well-being (Algan *et al.*, 2016). However, we need to be much more cautious on their contribution to the short-term economic diagnosis. Their predictive performance is at best similar to that of the traditional sources, without offering the same guarantees of stability. The same message emerges from other attempts at nowcasting or short-term forecasting using other types of big data, such as bank transaction data (Gill *et al.*, 2012; Galbraith and Tkacz, 2015).

However, the improvement in short-term diagnosis is only one of the possibilities opened by these big data. Another important field is price measurement. At present, the majority of price monitoring is done by direct collection at points of sale. Measured price changes are then weighted according to the budget coefficients of the different types of products. This method of price collection has the advantage of being applicable to all types of goods but it is time consuming and costly. Two other collection methods are possible. The first is to recover prices online, automatically and in real time, from the websites of distributors (a process called "web scraping"). The second is to use scanner data, i.e. the records of sales receipts produced automatically and centralised by the major retail chains when their customers make purchases.

In this case, it is not velocity that is most interesting, because these data are already available very quickly.[2] The key advantage is volume. First, we look at the web scraping approach, as implemented by an international project conducted outside the scope of official statistics, the Billion Prices Project (BPP).

This project originated with a case where official statistics were challenged over the measurement of inflation, the case of Argentina at the end of the 2000s. Despite being based on stable and proven protocols, measuring inflation by direct collection at the point of sale is often called into question. This happened in France during the change over to the euro. It is not unusual to observe differences between measured inflation and perceived inflation: perceived inflation overweights price movements in the most frequently consumed goods, and can also place more weight on increases than on decreases (Accardo *et al.*, 2011). However, in the Argentinian case, this mistrust was also fed by evaluations carried out by independent local authorities and by studies produced by economists: official inflation was around 7% per year and alternative estimates were around 20%. Web scraping of the websites of the major retailers turned out to confirm this discrepancy (Cavallo, 2013), illustrating the potential interest of this data collection method for other countries. The BPP is the direct result of this experiment. It was launched in 2008 as an academic project,[3] and aimed to cover as many countries as possible. The symbolic target of a billion prices, which gave the project its name, was reached in annual flows in 2010. The change in scale meant that funding had to be sought and this led to the creation of a dedicated company[4] which currently monitors 15 million products for 900 retailers in 50 countries (Cavallo and Rigobon, 2016).

The first step in web scraping is to select the retailers that are to be monitored. This choice is limited by the fact that they must sell online, but representativeness is ensured by selecting only those that sell both online and in store in the traditional way. Search robots collect all the information on their websites concerning the products covered: names and identifiers, varieties, packaging and other characteristics, and of course the price. It is necessary to check that, for the same item and the same retailer, the prices online are representative of

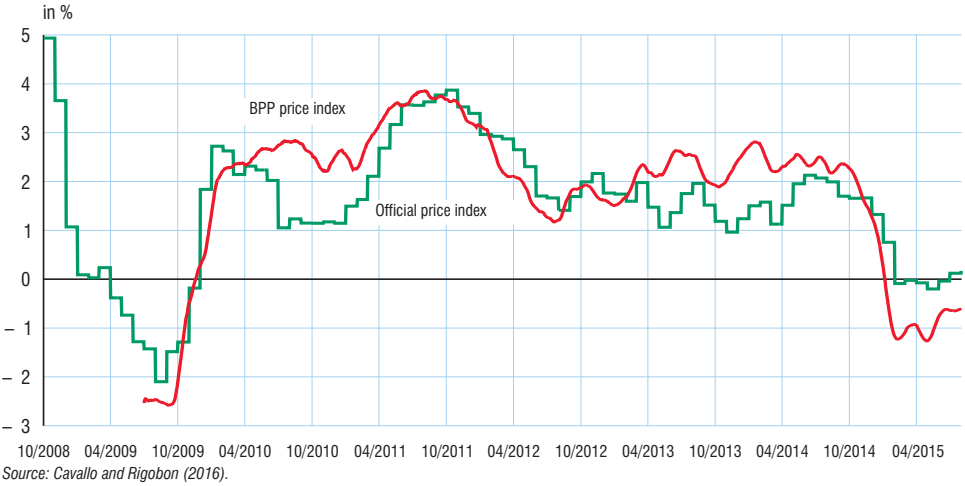
2. Since January 2016, INSEE has offered an initial estimate of the monthly consumer price index at the end of the month concerned.
3. *bpp.mit.edu*
4. *www.pricestats.com*

prices in store, which does indeed seem to be confirmed by comparison with direct price readings (Cavallo, 2017). However, this technique, by its very nature, does not collect all the information needed to calculate a price index: it provides only the prices of the products considered but not the corresponding quantities, which are needed for weighting price changes. These must be obtained from the same sources used for the official price indices.

Although the original project did indeed confirm a failure in the Argentine official statistics, its extension to other countries showed that this failure was the exception rather than the rule. In the case of the United States and the Eurozone, no trend bias is observed: differences are alternately positive or negative, and by about one point in absolute value at the most, which seems small given the large differences between the collection methods and also the difference in scope, as the BPP index covers only around 70% of the scope of the official index *(Figures 4a and 4b)*. In particular, the BPP data confirmed the very low level of inflation in the developed countries since 2015. The two series are virtually identical in the Eurozone

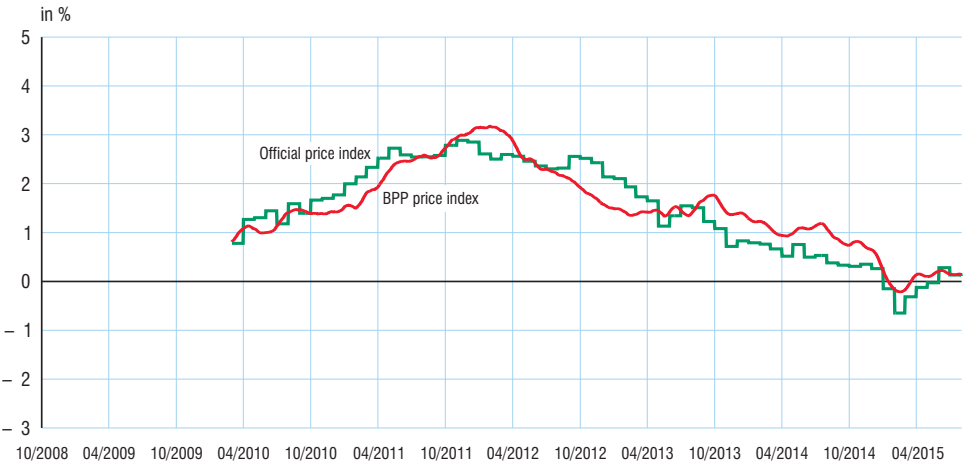
4. Annual inflation rate: official and BPP indexes

a. United states



Source: Cavallo and Rigobon (2016).

b. Eurozone



Source: Cavallo ans Rigobon (2016).

for recent years, with these scraped data even suggesting that the traditional index might have overestimated real inflation in the United States since the start of 2015.

This result is somewhat comforting for the traditional collection method, whose quality is not questioned, but it could also be an argument for its gradual replacement by this new technique. However, it is not this route that is generally favoured by national statistical institutes. Scraping is certainly being studied in some national institutes and, in France, some prices are actually collected from the internet, such as those for air or sea transport. For goods, however, the preference is for scanner data from the checkouts, mainly because these data provide information directly on both prices and quantities purchased, the two types of information required to construct the price index. In France, the use of these scanner data is being studied in a project which was launched in 2015 and is due to finish in 2020 *(Box 2)*. The project will be based on the Law for a Digital Republic (*Loi pour une république numérique*) which was passed at the end of 2016, whose article 19 sets out the conditions for making this type of data available to statisticians.

Whether using "scraped" data or scanner data, another expected result from large-scale price collection is a better management of what is one of the most sensitive issues when measuring purchasing power, that of the disappearance of products and the appearance of new ones (Boskin *et al.*, 1996).

Box 2

Scanner data and calculating the CPI

*Pascal Chevalier and Marie Leclair**

To measure the Consumer Price Index, INSEE survey workers currently collect around 200,000 prices each month in almost 30,000 points of sale. These readings are supplemented by prices collected centrally.

The major retailers, for their part, collect 1.3 billion prices each month using scanned information collected as consumers pass through the checkout. National statistical institutes were very quick to see the potential of using these scanner data for statistical purposes. Six European countries now use them to calculate their Consumer Price Index. In 2015, INSEE launched a project to produce a Consumer Price Index based in part on these scanner data. Given the time required to complete experiments and define the legal framework of this data collection, it is hoped that it will be operational by 2020. It will cover the prices of industrially produced food products and household maintenance-hygiene-beauty, which are currently monitored by survey workers in supermarkets and hypermarkets. Control surveys will be carried out in the field to ensure the quality of the data transmitted. For other products and points of sale, data will continue to be collected by survey workers in the traditional way. In the longer term, the project may be gradually extended to other products and other types of store.

INSEE took the decision not to modify the methodology and concepts of the price index with the introduction of this new data source, but to take advantage of the progress made possible within the same framework. Due to the mass of data collected, the scanner data should substantially improve the accuracy of the indices calculated. Using the scanner data, quantities sold are known to a very detailed level using barcodes, days and points of sale. They can be used as a sampling frame in places where, until now, the lack of information led statisticians to use quota methods. Another advantage is that it is effective sales prices that can be monitored whereas survey workers can only collect information on displayed prices. Lastly, scanner data provide information that is useful for correcting quality differences in product replacements when a product in the basket being tracked for the price index disappears.

Collecting scanner data is less costly than the traditional data collection by a survey worker, and ultimately it should also be possible to produce new statistics thanks to the detail and volume of the information collected: indices for particular segments of consumption, regional indices, etc.

* Pascal Chevalier and Marie Leclair, INSEE.

Let us first consider the case of a product that is no longer marketed. When a price survey worker finds that a product has disappeared on date *t*, he looks for a replacement product which is as similar as possible to take its place in the list of goods observed monthly. The item that has disappeared will be included in the index until the interval [*t*-2, *t*-1], and the new item will be included from the period [*t*, *t*+1]. This leaves a problem of missing data for the period [*t*-1, *t*]. One way to manage this is to assume that, if the missing item had still been present at *t*, then its price would have evolved in the same way as the average price of goods in the same category. This is not an unrealistic assumption, but scraped or scanner data allow us to do better, since they are able to monitor all products across the entire period that they are present on the shelves or in the catalogue, and hence go back to retrieve past changes in the price of the replacement product. We can then take advantage of the overlap between observation periods for the different items to better articulate their respective price changes. This question has begun to be explored at INSEE using information collected for the very first scanner data experiments (Sillard, 2013; Léonard *et al.*, 2015).

In the same way, these data can help to better manage products that are completely new and which are added to the list of existing products, typically the appearance of a new electronic product or a new service. The two-fold problem with new products is incorporating them into the index as soon as possible, then knowing how to situate them in relation to the products in the initial basket of goods. Incorporating products using the traditional methods necessarily takes some time, first to note the appearance of these new products, then to add them to the list given to the survey workers. Scanner data and scraped data, on the other hand, pick up these products automatically, as soon as they go on sale.

What remains to be seen is the degree to which this new product contributes to improving purchasing power. This depends on the quality of the service provided. One method that is often recommended to manage this problem is to use hedonic prices: using this method, product quality is objectified based on a few measurable characteristics, such as storage capacity or processor speed for computers. However, this method is costly to implement and cannot be applied to all types of goods, so that it is only used in specific cases. An easy alternative to put in place using scanner data or scraped data is to assume that, over the period where products of the same kind coexist, the price difference does measure this difference in quality. The assumption is that the new item would not be taken up if the quality difference did not justify the price difference. Once again, this is only an approximation. It could be, for instance, that the manufacturer or the distributor is taking advantage of a passing fad and is applying a price to the new product that is higher than the real gain in service provided, in which case the method will overestimate the contribution to standard of living. It is also possible that they will select an undervalued entry price in order to impose the product on the market, after which this price will gradually be readjusted upwards. In this case, the gain in standard of living generated by the quality effect will be underestimated. The method is therefore not infallible, but in both cases there is the advantage of more comprehensive information covering the entire life cycle of the products.

Big data and measuring the digital economy

This issue of product renewal brings us to our last topic, that of measuring the digital economy. It is in this area that the renewal of goods and services appears to be most important at the present time and is suspected of being poorly accounted for by usual statistics. This is the so-called "mismeasurement" question, according to which the current problems in regaining pre-crisis growth rates could be more apparent than real and stem from the fact that traditional tools are unable to detect that growth is changing in nature.

This notion of mismeasurement may reveal some misunderstandings over what the Gross Domestic Product is meant to measure: its goal is not to measure all gains in well-being generated by new products or services, it focuses on the share of these gains that have an explicit monetary equivalent (Bellego and Mahieu, 2016). Nevertheless, it cannot be denied that current developments pose many challenges: the internet promotes new consumer behaviour or the development of a collaborative economy which blurs the boundaries between market and non-market activities and between wage-earners and non-wage-earners. It is both the subject of the measurement and the tools used to measure it that are affected and big data can be expected to bring some natural solutions to this double issue.

Several studies have started to explore this line of enquiry. Some examples are given here, without pretending to be exhaustive. The first problem is to assess the share of business activities represented by the digital economy. This means selecting a definition. This type of question emerges with every great wave of innovation. Classifications of activity that exist on a given date reflect a certain state of the productive system and the nature of the goods and services it produces. This state is inherited from history and innovative activities, by definition, are therefore not easy to place in existing classifications. This problem has already been faced in the transition from an agricultural to an industrial economy.

This subject is covered in two attempts to use big data, in the United Kingdom and the Netherlands, respectively. The first study, by the National Institute of Economic and Social Research (NIESR), uses big data to question some popular views about the role of the digital economy in the UK economy (Nathan and Rosso, 2013): the fact that it is small in size, that it is dominated by start-ups, that it generates low revenue and low employment and that it is entirely concentrated in London. The study points out that the problem of identifying companies in the new economy is two-fold. First is the inadequate nature of the Standard Industrial Classification (SIC) codes normally used by the Office for National Statistics (ONS), equivalent to the French classification Nomenclature d'Activité Française (NAF). Second, the fact that activity declared at the Companies Houses is not necessarily updated when the firm changes its activity. Many companies in existence before digitisation have since become digitised, sometimes on a very large scale.

The study managed these two problems via a partnership with a UK technology company, Growth Intelligence, specialised in web scraping applied to predictive marketing. In contrast to the BPP, web scraping is not restricted to exploring the websites of companies of interest. It gathers everything that is said about them on the web, whether it is information disseminated by themselves or by third parties, such as references in the press. Thus the authors always had up-to-date information. They used it to introduce a specific classification, combining sector and product, so that within a global "architecture" sector for example, they were able to identify those companies specialising in computer-aided design.

This database was then matched to various sources describing the other characteristics of these companies. After applying filters, 1,676 million units were analysed, of which 14% were classified in the digital economy, which represented 11% of total employment. On average, these were not young companies: this is explained by the fact that the analysis picked up progress in digitisation in traditional companies. Lastly, geolocalised data was able to show that these companies were not all clustered in the London area.

The Dutch study has a lot in common with the previous study but also a significant number of differences. It involved the Dutch National Statistical Institute, the Central Bureau of Statistics (CBS) which, like the NIESR, worked in association with a company specialising in web scraping, Dataprovider, which indexes websites according to their content. The study was based on a list of 2.5 million sites, including those with the ".nl" domain name and those with a ".com" domain name and identified as being Dutch. As in the other study, a textual analysis of the content of these sites created a typology specifically for this study. The aim was to determine internet penetration even where it was a supplementary activity and not the main

activity. The typology defines five types of companies. Three categories correspond to the core of the digital economy: Category C is online stores, whether these are pure players or retailers also selling in the traditional way through their shops; category D consists in online services, such as dating websites, price comparison sites, information sites, and lastly Category E consists in internet-related ICT businesses whose job is to make the internet work, such as hosting companies and web designers. The other two categories are Category A, businesses without a website and Category B, businesses with only a totally passive website or businesses offering the visitor only minimal internet actions such as ordering a brochure.

Allocating a website to Category C was the most straightforward. To determine whether a website has an e-commerce functionality, Dataprovider shows, for example, whether it has a "shopping cart" function or offers online payment methods. Classification algorithms were used to recognise whether a website was an e-commerce site after applying a machine-learning stage to a sub-set of sites that had been identified as being definitely or definitely not e-commerce. Allocating websites to Categories D and E was more complex as there are fewer common denominators for a very wide range of activities. Sites were identified using keywords such as "hotel", "reservation", "news" which generated a large volume of manual adjustments, with a systematic inspection of the 100 websites with the largest turnover in each sub-category.

Once all classified, the sites have been matched to the business register managed by the CBS based on company identifier, telephone number or e-mail address. In all, out of around 2.5 million sites, 840,000 were matched against the business register, with the rest of the sites being mainly the sites of private individuals. 36% of companies were present on the internet. They represented 87% of the total turnover of the economy and 86% of jobs. 3% were part of the core of the digital economy, representing 8% of total turnover. This core of the internet economy included 28,500 units in Category C, 5,700 units in Category D and 16,000 units in Category E.

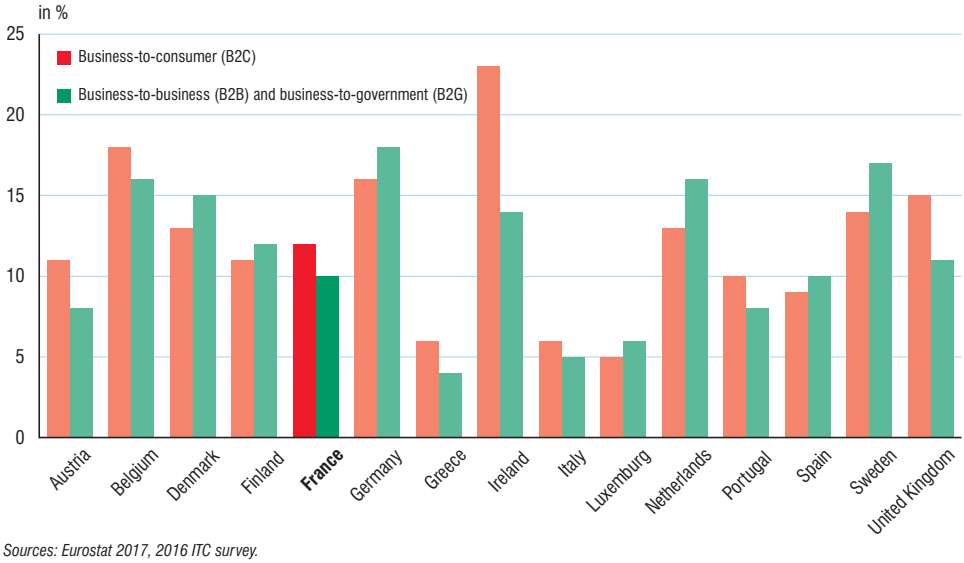
Comparing these two studies is very informative both in terms of the contribution of the data and the problems involved in using them. Analysis of internet content provides a potentially very rich and very up-to-date source of information on what companies do, whether from data collected from the entire internet, as in the British study, or information collected only from their own websites, as in the Dutch study. The value of this type of information is not limited to measuring the digital economy: it can also be used to measure all sorts of emerging activities or practices, for example in the field of the environment, social responsibility, etc. However, ordering this information requires agreement on what one is trying to measure, and it is clear that two uncoordinated studies do not measure exactly the same thing. The NIESR study focuses on digital products, the Dutch study on using digital tools applied to market goods that may have nothing to do with the digital field.

Thus the two definitions result in figures that are not better comparable with each other than they are with the definition of the usual classifications. Before using any new sources, it is essential to carry out a preliminary stage required to produce figures that will be comparable over time and space, the development of shared classification standards, like those produced and used by the National Statistical Institutes. In fact, traditional statistics are already producing data of this type at European level, through the Community survey on ICT usage and e-commerce, conducted annually since 2002 by the statistical institutes of each Member State. According to the 2014 wave of this survey, in the Netherlands, 16% of companies sell online, compared with 11% in the United Kingdom and 10% in France *(Figure 5)*. For the Netherlands, this result is not directly comparable with that given in the CBS study, as is not possible to say what is due to a difference in concept, a difference in scope (the European survey covers only enterprises with at least 10 occupied employees), bias in survey responses or in the ability of the scraping process to properly identify what is found on the company websites. A more systematic comparison between results from the different scraping methods

and responses to surveys would be useful. It would help assess how far the two methods can be considered substitutable, opening the possibility to reduce the response burden of the survey. This is something that Italy has already begun to explore (Barcaroli *et al.*, 2015).

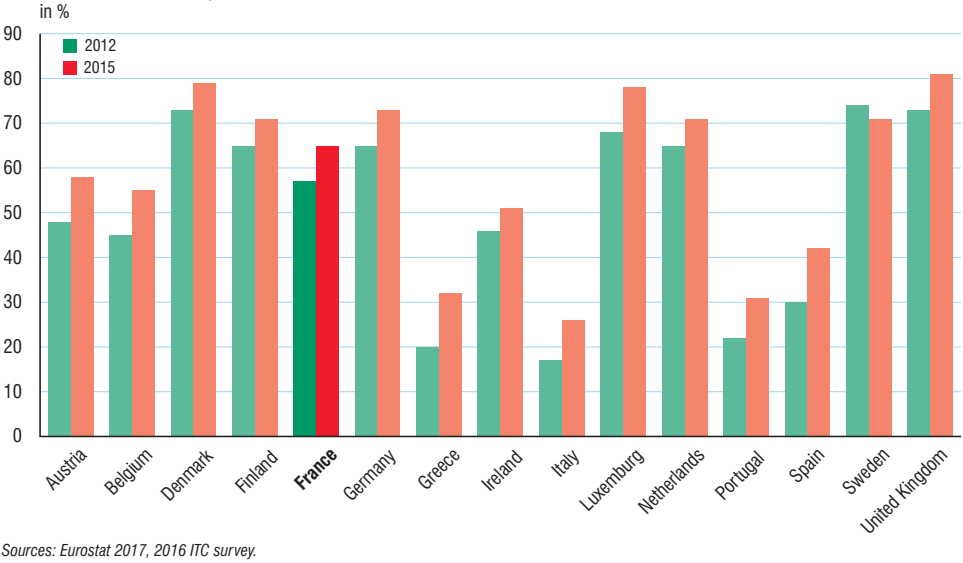
Finally, the same survey also provides information on the use that enterprises themselves make of big data or cloud computing techniques: in France, in 2015, 11% of enterprises with 10 or more employees in the mainly market sector excluding the agriculture, finance and insurance

5. Enterprises which sold via a web site in 2014



Sources: Eurostat 2017, 2016 ITC survey.

6. Individuals using the internet for ordering goods or services within the last 12 months before the survey



Sources: Eurostat 2017, 2016 ITC survey.

sectors used big data. This rate rose to 24% for enterprises with 250 or more employees; in 2016, 17% of enterprises paid for cloud computing, against an average of 21% for the countries of Europe (Vacher and Pradines, 2017).

The growing importance of the digital economy must then be assessed from the point of view of individuals and households. Traditional statistics also provide information in this field, especially, still at the European level, from the household equivalent of the ICT survey of businesses. According to the 2015 edition of this survey, in France, 65% of individuals had used the internet to purchase goods and services in the 12 months preceding the survey *(Figure 6)*. What traditional statistics measure less well at this stage is the way in which the collaborative online economy favours new forms of households' productive activity and the associated income gains, for example the temporary renting of one's apartment through *Airbnb* or the online selling of items through eBay. It is difficult to measure this phenomenon through surveys as it concerns only a small segment of the population. It may also be difficult to capture or impossible to isolate in administrative sources, either because they do not differentiate different forms of income, or because the income in question eludes them completely. One alternative is to mobilise information recorded by the operators of this type of service. INSEE has started doing this to estimate the supply of tourist accommodation offered by individuals *via* internet platforms (Franceschi, 2017). The Bean report (2016) also cites an example of an experimental use of big data which is completely outside the scope of official

Box 3

Big data and consumer surplus evaluations

In addition to general insights into the digital economy, big data can also provide more local but original highlights on some of its segments. A very recent example is provided by Cohen *et al.* (2016) who used some very detailed data on Uber to analyse the price sensitivity of the demand for rides and assess consumer surplus generated by this service. This question of consumer surplus is one of the possible keys to the divergence between change in GDP and gains in well-being obtained by the new economy. National accounting evaluates goods and services at prices that reflect their marginal utility, i.e. the gain in well-being provided by the last unit consumed. In general, marginal utility diminishes and thus underestimates utility derived from overall consumption. It is this difference that is defined as the consumer surplus. In order to reconstitute it, it is necessary to know the prices that the consumer would have been willing to pay for each unit consumed, starting with the first, and hence his entire demand profile according to price.

Estimating this profile is usually difficult due to a problem of circularity. Demand depends on price (negatively) and price depends on demand (positively). It is this loop which is supposed to bring equilibrium to the market, but with the result that we do not know exactly what is measured by the apparent relationship between price and effective consumption: it is a mix of these two relationships in opposite directions. Consumption of Uber services does not escape this problem, in fact it is all the more exposed to it as the system fine tunes this adjustment process by increasing the tariff offered for deliveries in real time according to the ratio between supply and local demand. However, the study exploits a specific feature of this pricing technique, the fact that prices adjustments are discrete rather than continuous. On either side of a price discontinuity, local conditions of supply and demand are very similar but the customer is offered a tariff that may be higher or lower. It can be assumed that the deviation in the customer's rate of acceptance around such a discontinuity does indeed measure a pure effect of the price offered.

The data consist of 54 million customer interactions for the period January to June 2015. For each interaction, the price offered after applying the discontinuity and the price that would have been offered if the discontinuity had not been applied are known, and also whether the transaction was accepted or not, which is the demand variable. Several other features of the delivery can be monitored. The authors estimated a surplus of 1.6 dollars per dollar actually spent on deliveries, which must of course be compared with all the effects for the other players in the system.

statistics: a major US bank used anonymised data on movements in its customers' bank accounts, enabling it to measure monthly variations in their income, and their ability to absorb these fluctuations by drawing resources from the "gig economy", the economy of small jobs or occasional incomes obtained through collaborative platforms. The participation rate in this collaborative economy is estimated to be 1% over one month and 4% when cumulated over three consecutive years, with a total contribution of 15% to labour income over the months of active participation by the individuals concerned, but the scope is that of customers of the institution, hence not necessarily representative (Farrell and Greig, 2016).

From the point of view of the consumer, a final question is that regarding the value created by these new services. This is at the heart of the mismeasurement question. National accounts only evaluate exchanged goods or services at their marginal value, that of the last unit consumed, and ignore what is known as consumer surplus, which corresponds to the difference between this marginal utility and the utility taken from the entire consumption. It has recently been suggested that big data can be used to evaluate this surplus for one of the players in this new economy -*Uber*- by taking advantage of the specific features of its pricing policy *(Box 3)*. This study illustrates a paradoxical impact of the digital economy: increasing the possibilities of tiered pricing makes price measurement much more complex, but it allows a more precise approach to the willingness to pay of different categories of individuals and hence a closer approach to a true measurement of the service provided. However, this example is very specific and is a long way from providing a global response to the question of quantifying what the digital economy brings to the standard of living, whether or not this contribution is to be tracked in the GDP.

*
* *

To conclude, what are the main messages of this overview? The term big data covers a very wide range of different sources. There is sometimes a temptation to see it as a miracle solution to the growing demand for faster and more abundant statistics. Reality is not that straightforward and the question of the contribution of big data to official statistics must be looked at on a case by case basis. The area of prices is the one where these data seem to be most promising. This is an area where these data come in relatively structured formats, fairly similar to that of administrative data which official statistics are used to dealing with, and the subject of the measurement is conceptually straightforward. The response is less immediate in other areas, especially when using very qualitative sources: using such data to extract stable and conceptually consistent pieces of information appears much more challenging and it is an area where current explorations remain much more experimental. What can be expected at best is a complementarity with existing productions, with either the aim to shorten publication delays or to reduce response burdens on units that are surveyed.

One aspect of this complementarity that this overview has not explored in much detail is the contribution of big data to a higher level of granularity in the description of the economy, as our emphasis has been on applications to macro- or at most meso-economic observation. Of the three "Vs", it is volume that would represent the most obvious asset here: access to virtually exhaustive data makes it possible to envisage producing very localised statistics or statistics centred on very specific categories of population. This is the aim of a number of other experiments currently underway at INSEE or some of the other national statistical institutes, but these are not developed in this focus: they include satellite data, road sensor data, mobile phone or bank card data and other information generated by the increasing numbers of connected objects.

However, two problems also emerge. First is the question of the protection of private life and respecting business confidentiality. The greater the degree of detail in the statistics, the higher the risk of indirect re-identification, however much care is taken in anonymising the

data (de Montjoye *et al.*, 2015). Second comes the question of the ownership of these data: in the vast majority of cases, they derive from the activity of private enterprises. Access to these data has to fit within a clear and sustainable legal framework. For example, it would be impossible to ensure the continuity of the price index based on scanner data if there were no guarantee of their continuing availability. In the case of France, the Law for a Digital Republic has established the necessary legal framework for this guarantee. ■

For more information

Accardo J., Célérier C., Herpin N., Irac D., « L'inflation perçue », *Économie et Statistique* n° 447, p. 3-31, 2011.

Algan Y., Beasley E., Guyot F., Higa K., Murtin F., Senik C., "Big data measures of well-being: evidence from a Google well-being index in the United States", *Document de travail, Cepremap* n° 1605, 2016.

Barcaroli G., Nurra A., Salamone S., Scannapieco M., Scarno M., Summa D., "Internet as data source in the Istat survey on ICT in entreprises", *Austrian journal of statistics*, vol. 44, pp. 31-43, 2015.

Bean C. R., *Independent review of UK economic statistics*, 2016.

Bellego C., Mahieu R., « L'internet et la mesure de l'économie », *L'économie française*, coll. « Insee Références », édition 2016.

Blanchard O., Lorenzoni G., L'Huillier J.P., "Short run effects of lower productivity growth : a twist in the secular stagnation hypothesis", *NBER working paper* n° 23160, 2017.

Bortoli C., Combes S., « Apports de Google Trends pour prévoir la conjoncture : des pistes limitées », *Note de conjoncture*, Insee, pp. 43-56, mars 2015.

Bortoli C., Renault T., Combes S., « Peut-on prévoir l'emploi en lisant le journal ? », *Note de conjoncture*, pp. 35-43, Insee, mars 2017.

Boskin M., Dulberger E., Gordon R., Griliches Z., Jorgensen D. *Toward a more accurate measurement of inflation*, Advisory commission to study the consumer price index, US Senate, 1996.

Cavallo A., "Are online and offline prices similar? Evidence from large multi-channel retailers", *American Economic Review*, vol. 107, n° 1, p 283-303, 2017.

Cavallo A., "Online vs official price indexes: measuring argentina's inflation", *Journal of Monetary Economics*, vol. 60, n° 2, pp. 152-165, 2013.

Cavallo A., Rigobon R., "The billion prices project: using online prices for measurement and research", *Journal of Economic Perspectives*, vol. 30, n° 2, pp. 151-178, 2016.

Choi H, Varian H., "Googling the present with Google Trends", Google Inc, 2009.

Cohen P., Hahn R., Hall J., Levitt S., Metcalfe R., "Using big data to estimate consumer surplus: the case of uber", *NBER working paper* n° 22627, 2016.

Engelberg J. E., Parsons C. A., "The causal impact of media in financial markets", *The Journal of Finance*, vol. 66, pp. 67-97, 2011.

Eurostat, "Digital economy and society statistics - enterprises", *Statistics explained*, (http://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_enterprises), 2017a.

Eurostat, " Digital economy and society statistics - households and individuals", *Statistics explained*, (http://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_households_and_individuals), 2017b.

Farrell D., Greig F., "Paychecks, paydays and the online platform economy", JPMorgan Chase Institute, 2016.

For more information (continuation)

Fondeur Y., Karamé F., "Can Google data help predict French youth unemployment?", *Economic modelling*, vol. 30, pp. 117-123, 2013.

Franceschi P., « Les logements touristiques de particuliers proposés par internet », *Insee Analyses* n° 33, février 2017.

Galbraith J. W., Tkacz G., "Nowcasting GDP with electronic payments data", *ECB Statistics Paper Series* n° 10, 2015.

Gill T., Perera D., Sunner D., "Electronic indicators of economic activity", *Reserve Bank of Australia Bulletin*, pp. 1-12, juin 2012.

Lazer D., Kennedy R., King G., Vespignani A., "The parable of Google Flu: traps in big data analysis", *Science*, vol. 343 (6176), pp. 1203-1205, 2014.

Léonard I., Sillard P., Varlet G., Zoyem J.P., "Scanner data and quality adjustment", miméo, Insee, 2015.

McLaren N., Shanbhogue R., "Using internet search data as economic indicators", Bank of England Quarterly Bulletin , vol. 51, n° 2, pp. 134-140, 2011.

de Montjoye Y.-A., Radaelli L., Singh V. K., Pentland A. S., "Unique in the shopping mall: On the reidentifiability of credit card metadata". *Science*, vol. 347, n° 6221, pp. 536-539, 2015.

Nathan M., Rosso A., "Measuring the UK's digital economy with big data", rapport Growth Intelligence/NIESR, 2013.

Oostrom *et al.*, "Measuring the internet economy in the netherlands: a big data analysis", CBS working paper n° 2016-14, 2016.

Sillard P., « Les données de caisse : vers des indices de prix à la consommation à utilité constante », *Document de travail,* Insee/DSDS n° F1305, 2013.

Soo C.K., "Quantifying animal spirits: news media and sentiment in the housing market", Ross School of Business Paper, n° 1200, 2015.

Tetlock P. C., "Giving content to investor sentiment: the role of media in the stock market", *Journal of Finance*, vol. 62, n° 3, pp. 1139-1168, 2007.

Vacher T., Pradines N., « Cloud computing, big data : de nouvelles opportunités pour les sociétés », *Insee Première* n° 1643, 2017.