N° 505-506 - 2018

Economie ET Statistique

Economics AND Statistics

Big Data et statistiques

1ère partie

Big Data and Statistics

Part 1

Insee

Economie ET Statistique

Economics AND Statistics

Directeur de la publication / Director of Publication:
Jean-Luc TAVERNIER
Rédactrice en chef / Editor in Chief:
Sophie PONTHIEUX
Rédacteur associé / Associate Editor: Clément CARBONNIER
Assistant éditorial / Editorial Assistant: Étienne de LATUDE
Traductions / Translations:
RWS Language Solutions
2, rue Sainte Victoire, 78000 Versailles, France
UBIQUS
Tour PB5, 1 avenue du Général-de-Gaulle, 92074 Paris La Défense Cedex
Maquette PAO et impression / CAP and printing: JOUVE
1, rue du Docteur-Sauvé, BP3, 53101 Mayenne

OÙ SE PROCURER
Economie et Statistique / Economics and Statistics

Les numéros sont en accès libre sur le site www.insee.fr. Il est possible de s'abonner aux avis de parution sur le site.

La revue peut être achetée sur le site **www.insee.fr** via la rubrique « Acheter nos publications ». La revue est également en vente dans 200 librairies à Paris et en province.

WHERE TO GET
Economie et Statistique / Economics and Statistics

All the issues and articles are available in open access on the Insee website **www.insee.fr.** Publication alerts can be subscribed on-line.

The printed version of the journal (in French) can be purchased on the Insee website www.insee.fr and in 200 bookshops in Paris and province.

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Directeur Général : Jean-Luc TAVERNIER
Direction Générale : 88, avenue Verdier, CS 70058, 92541 MONTROUGE Cedex
Tél : +33 (0)1 87 69 50 00

N° 505-506 - 2018

Economie ET Statistique

Economics AND Statistics

The views or opinions expressed by the authors engage only themselves,
and neither the institutions they work with, nor Insee.

Economie et Statistique / Economics and Statistics

Issue 505-506 – 2018

BIG DATA AND STATISTICS
Part 1

DATA AND METHODS

BIG DATA AND OFFICIAL STATISTICS

Introduction – The Contributions of Big Data

Philippe Tassi*

Abstract – The revolution, which is quite recent, brought about by digital convergence and connected objects, has enabled a homogenisation of data types which would historically have been considered as different, for example: digital data, texts, sound, still images, and moving images. This has encouraged the Big Data phenomenon, the volume of which includes two related parameters: quantity and frequency of acquisition; quantity can extend as far as exhaustivity and frequency can be up to and including real time. This Special Issue features a series of articles that examine its uses and implications, as well as the challenges faced by statistical production in general, and especially that of official statistics. Just like any innovation, Big Data offer advantages and raise questions. The obvious benefits include “added” knowledge – a better statistical description of the economy and the society. They are also a driver for development in computer science in the broadest sense, and in applied mathematics. However, we cannot do without some degree of vigilance, since data and how they are used can affect individuals, their freedoms and the preservation of their privacy.

JEL Classification: C1, C8
Keywords: digital data, Big Data, statistics, official statistics

* *Médiamétrie (ptassi@mediametrie.fr)*

Received on 21 March 2019
Translated from the original version: “Introduction – Les apports des Big Data”

To cite this article : Tassi, T. (2018). Introduction – The Contributions of Big Data. *Economie et Statistique / Economics and Statistics*, 505-506, 5–16.
https://doi.org/10.24187/ecostat.2018.505d.1963

Some History – and Histories

Although the term “data” may seem modern, especially when preceded by “big”, we ought to remember that data is no other than the plural form of the supine of the Latin verb “do”, “das”, “dare”, “dedi”, “datum”. Moving beyond the Latin origins of the word, mass or even exhaustive data collection is not a phenomenon from the present digital era; it is an activity that began as soon as writing emerged, since the latter was a necessary pre-requisite for it. Most historians and archaeologists believe that writing first appeared in Lower Mesopotamia (present-day Iraq) approximately 5,000 years ago, at a time when nomadism was in decline and the first settlements were being established, which lead to the birth of the cities of Sumer. Since memory alone would no longer suffice to understand, manage, and govern these cities, written marks had to be used. The site at Uruk (Erech in the Bible) has yielded many clay tablets dating from the fourth millennium BCE, tablets covered with signs traced using a reed stylus - the origin of cuneiform script, a structured writing system involving several hundred signs.

Data collection could then begin, starting with two main areas of interest: astronomy and the counting of populations. As Jean-Jacques Droesbeke wrote: “[…] the Mesopotamians were very early adopters, […] and in ancient Egypt also, from the end of the third millennium BCE […] to know how many men were available to participate in the construction of temples, palaces, pyramids […] or even […] for tax purposes.” Data collection was not limited to these city-states. China and India had systems covering large territories in the last millennium BCE. China had its “directors of the masses”. In India, the Maurya Empire covered a vast territory similar to present-day India, and its first emperor, Chandragupta, established a census in the 4th century BC.

Turning to data processing, and given that the expression “artificial intelligence” (AI) is now in common usage, let’s give it a definition and an historical perspective. Yann LeCun, who was the 2016 Chair of “Informatics and Computational Sciences” at the College de France as well as the first Director of the Facebook Artificial Intelligence Center in New York and later in Paris, and also a leading figure in AI and Deep Learning has defined AI as follows: “making machines complete tasks that would normally be assigned to people and animals”. For the historical perspective, we might look back to Babylon or the Chinese Empire, since even at that early stage, it seemed natural to try to model human brain behaviour and depict man as a machine as a precursor to the design of learning machines.

One forerunner of Artificial Intelligence was the Catalan philosopher and theologian, Ramon Llull (1232 - 1315), who invented “logic machines”. Predicates, subjects and theories were organised into geometric figures that were deemed perfect (circles, squares, triangles). With the aid of dials, cranks and levers to turn a wheel, the propositions and theses were moved into position to reveal their truth or falsehood. Llull was a major influence on his contemporaries and even beyond, since four centuries later, Gottfried Leibniz would find inspiration in his work.

From Sampling to Big Data: Complementary Paradigms

We could say that the world has lived under the near-total reign of exhaustive data collection, however, a few rare approaches at sampling did exist in the mid-17th century: John Graunt and William Petty's school of political arithmetic in England and Vauban's advances in France. The 20th century featured a slow decline in census and exhaustive data collection, and an ever more assertive rise of the sampling paradigm. The founding act was the speech by Anders N. Kiaer, Director of the Central Bureau of Statistics in Norway, during the International Statistical Institute's Berne Congress in August 1895: the first recognition for the *pars pro toto*.

In 1925, the ISI validated Kiaer's approach, and developments followed swiftly: the benchmark paper on surveys would appear in 1934 (Neyman, 1934). Operational applications rapidly ensued: in the economic field, following J. M. Keynes's articles in the early 1930s, the first consumer and distributor panels emerged in 1935, run by companies such as Nielsen in the United States, GfK in Germany. In 1935, George Gallup launched his company in the United States: the American Institute for Public Opinion (AIPO). He went on to achieve fame among the general public after he used a sample of voters to predict Franklin D. Roosevelt's victory over Arthur Landon in the 1936 presidential election. In 1937, Jean Stoetzel created the French equivalent organisation – l'Institut Français d'Opinion Publique (IFOP), the first opinion research company in France.

In the post-war years, sampling became the reference due to its operational speed and reduction of costs. With the advancement of probability, statistics and information technology, we also witnessed a general expansion into new fields such as economics, official statistics, health, marketing, sociology, media audiences, political science, etc. For most of the 20th century, therefore, the sampling paradigm prevailed and exhaustive censuses went into decline; we should note that in the 1960s, official statistical censuses of the population, agriculture and industry were still in existence.

Since the end of the 20th century and the start of the 21st century, digital convergence has favoured the automated collection of data for ever larger populations, generating databases that hold a growing mass of information, and heralding a return to exhaustive collection. Additionally, under the digital transition, it has become possible to harmonise information that was previously considered distinct and heterogeneous such as: quantitative data files, text files, (audio) sound files, photos and moving images (video). The two main parameters that help to define the volume of Big Data are: quantity and frequency of acquisition; quantity can extend to exhaustivity, and frequency can be up to and including real time.

Issues Raised by Big Data

Among the various issues raised by Big Data, some are old and some are new. They concern processing methods, storage, protection and security, property rights, etc. What statistical processing or algorithms should be applied to the data? What status does the data have and what is the status of the data author/owner? What is the regulatory or legislative framework like?

An enduring phenomenon

Big Data is clearly more than just a fad. We are only beginning to exploit it. Every day, new examples of Big Data appear across ever-expanding areas: medicine, epidemiology, health, insurance, sport, marketing, culture, and human resources, not to mention official statistics.

Digitalisation has lent weight to methodologies, modelling and technologies, as well as to their related professions. Innovations in algorithmics and machine learning when applied to Big Data represent a rapidly growing field, from the genius of Alan Turing to Arthur Samuel, Tom Mitchell, Vladimir Vapnik and Alexey Chernovenkis (Vapnik, 1995, 1998). The digital world is everywhere, investments are not fleeting, and the policy orientation of states is clear. In France, clear guidance was given in the thirty-four proposals for an industrial renaissance in France (François Hollande, September 2013). The report by the Innovation 2030 Commission chaired by Anne Lauvergeon placed particular emphasis on the excellent reputation of French training in mathematics and statistics. This was further demonstrated by the strong showing of "French Tech" at the Consumer Electronics Show (CES) in Las Vegas. As part of its strategic reflection "Insee 2025", Insee addressed access to private data and its use in official statistics. Connected objects and the "Internet of Things" are strengthening this phenomenon (Nemri, 2015).

Trust

In general, data and statistics produced by governments or companies are based on personal information, which raises questions about how to protect these sources, i.e. their privacy. Given the constant advances in science and data processing procedures, how can we establish and maintain the confidence of the general public, who are the leading stakeholder, whilst respecting the balance between the promise of confidentiality and the use of the collected data? The answer lies in two complementary approaches: one regulatory, since States have long been aware of the need to establish legal safeguards; and one technological, by erecting technical barriers to prevent the unwarranted dissemination of data.

Significant Regulatory Framework

In the field of statistics, many countries have a legislative framework, among them France which played a pioneering role with its Data Protection Act (*Loi Informatique et Libertés*) from 1978. An even earlier French law of 7th June 1951 related to obligation, coordination and secrecy in the statistical domain. It defined statistical secrecy as the "impossibility of identification" in the context of official statistics (censuses and surveys). Accordingly, any communication of personal, family or private data was prohibited for 75 years. The 23rd October 1984 Post and Electronic Telecommunications Code (*Code des Postes et Télécommunications électroniques*) and its various amendments addresses the processing of personal data in the context of electronic communication services, in particular via networks that support devices that collect and identify data. France's Conseil d'État also published a work entitled "Fundamental Rights in the Digital Age", containing fifty proposals to ensure that digital technology supports the rights of individuals and the public interest and including a chapter on "predictive algorithms" (Rouvroy, 2014). We should also mention professional codes of ethics, such as that of the European Society for Opinion and Market Research (ESOMAR), established in 1948, and regularly updated to clarify best practice when conducting market and opinion research.

The most famous law known to the general public in France is probably the law of 6th January 1978 on data processing, data files and individual liberties, known in

French as "*Loi Informatique et Libertés*" (Data Protection Act). It specifies the rules that can apply to personal data. The first article of the 1978 law clarifies: "personal data is taken to mean any information relating to a natural person who is or can be directly or indirectly identified by reference to an identification number or to one or more factors that are specific to them." These personal data may either be kept raw or may be processed and then stored. The law stipulates that processing means: "any operation or set of operations performed on such data using any mechanism whatsoever, and in particular the collection, recording, organisation, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or interconnection, as well as locking, deletion or destruction." The significance of this lies in the fact that Big Data is "massive" in both of the aforementioned senses: quantity and variety (the 6 Vs). The extent of the analysis that can be deduced from data calculated by inference is also important.

One special category of personal data is sensitive data, the collection and processing of which are prohibited as a matter of principle. Sensitive information is taken to mean information that directly or indirectly reveals a person's racial or ethnic origins, political, philosophical or religious opinions, trade union affiliations, or information that concerns their health or sexual life (Article 8). Finally, the GDPR (General Data Protection Regulation) enacted in 2016 and in force across EU Member States since May 2018, has been the focus of all attention; all the more so since it will be followed by the ePrivacy regulation, a *lex specialis* to the GDPR.

Technology and Data Confidentiality

The relationship between information technology, privacy, personal data and databases is a fairly long-established research area, having been formally addressed since the 1970s. Respect for privacy is also a principle on which the whole world seems to agree *a priori*. Is it possible to guarantee this respect in the technology arena?

Cybersecurity and methods of encryption have evolved a lot since their beginnings over three thousand years ago. These methods can be used to render a document (here used in the broadest sense of the term) unreadable – i.e. incomprehensible to anyone who does not possess the encryption key. Julius Caesar encrypted the messages he sent to his generals and the Rossignol des Roches family (Antoine Rossignol, his son Bonaventure and grand-son Antoine-Bonaventure) ran Louis XIV's "*Cabinet Noir*" (Black Chamber) and operated the world famous 17th century "Great Cipher". Lastly, we all know of Claude Chappe's telegraph coding from the late 18th century, as well as Samuel Morse's electric telegraph which arrived a few years later.

The vision of Tore Dalenius

In the context of the databases in existence prior to 1980, the Swedish statistician Tore Dalenius laid down some principles on ethics and respect for privacy. In his article (Dalenius, 1977), he set out the following principle: "Accessing a database should not allow you to learn more about an individual than could be learned without access to that database.*"* He added: $X(i)$ being the value of the variable X for an individual i, if the publication of a statistical aggregate T helps to determine $X(i)$ specifically, without access to T, then a breach of confidentiality has occurred. This principle seems acceptable. Unfortunately, we can demonstrate that it cannot be generalised: any third party who would like to collect personal data about individual *i* can do so by taking advantage of auxiliary information accessible from outside the database.

Anonymisation

One *a priori* intuitive technique of data protection consists of rendering the personal data anonymous. This is tantamount to removing all of the variables in the database that could identify a particular person. It is a reiteration of the personal data concept as expressed in French data protection law. Of course, a natural person is identifiable by his or her name, but also through other characteristic variables such as a registration code, an address (postal or IP), phone numbers, a PIN code (Personal Identification Number), photographs, or biometric components such as a fingerprint or DNA. More generally, identification is possible through any variable that can be used in crossing or cross-checking to find an individual in a group (e.g. their place of birth, date of birth or local polling station). This represents a less perfect and less immediate identification than using their name, however, it remains highly likely that the person would be identified, which is a far cry from complete ignorance.

For the last ten years or so, information and communication technologies have been generating lots of data that is useful for the previously mentioned analysis, for example, calls from a mobile device or connections to the internet. All of these “computer traces”, or “logs”, can easily be exploited thanks to advances in software and search engines. On the face of it, anonymisation is a simple concept to understand and to implement, however, it can become complex and also risks the deletion of useful or relevant variables from the database. Furthermore, as science progresses, the number of privacy breaches is increasing and the probability of identifying an individual from a database containing personal data is higher too, even after anonymisation.

Destruction or Aggregation of Data

Another method is to delete the data once a certain period of operational use has elapsed. However, deleted data can be valuable long after its “working life”, e.g. for historians and researchers. If we reuse the principle of France’s 1951 law on statistical secrecy in companies, it would then be possible to aggregate the individual data and only divulge these aggregated results once a certain amount of time has passed.

Data Masking

Obscuring data (also called data obfuscation or masking) involves maintaining the confidentiality of data by deliberately “altering” it. This can be done indirectly by burying the data in a bigger environment (along the same lines as diluting meaningful data) or else directly by transforming the data to make it insignificant. For the first of these methods, we might, for example, create additional variables that increase the data vector’s size, thereby creating a “fog” in which to hide our data. The second group of methods is characterised by techniques that are non-disruptive: masking the value of some cells in a table of results; removing variables concerning certain individuals; dividing a sample of data extracted from the database; or combining certain categories for variables with modalities, etc.

Particular methods also exist that directly intervene on the data to create noise, in the broadest sense of the term, and modify certain variables by rounding or blocking them *via* truncation at maximum or minimum thresholds. The transformation of variables can also be achieved by applying a homomorphism, swapping the value of the same variable between two individuals, or by means of data perturbation through random noise injection. Applied to the original data, some transformations (e.g. swapping, rotation) will leave the linear statistics invariant, whereas others will not. Arising out of work into missing data (Little, 1993; Rubin, 1993, 2003), this approach is especially relevant to synthetic data.

A New Approach: Differential Privacy

Since the mid-2000s, there has been another perspective on privacy protection (Dwork, 2004, 2006), and its philosophy owes a lot to Dalenius: “The probability of a negative consequence of any kind for individual *i* (for example: being refused credit or insurance) cannot significantly increase because of the representation of *i* in a database.”

We should nuance the adverb “significantly” here because it is very hard to predict what information – or what combination of information – might have negative consequences for the individual in question, were this information to be made public. All the more so since this information cannot be observed but rather is estimated using a calculation, and also because some consequences deemed negative by one person may on the contrary seem positive to someone else! This approach, which we could name “privacy” or “differential privacy” is based on probabilistic and statistical suppositions. Could this approach be expanded? The idea is to quantify the risk of a possible privacy breach, whilst at the same time measuring the impact of effective data protection on privacy in statistical terms. This opens up a new field of research that will analyse data post-obfuscation, alteration or modification of the original in order to maintain confidentiality.

Mathematical Statistics, Econometrics and Big Data: An Inevitable Convergence

Statisticians and econometricians have been slow to familiarise themselves with the volumetry and techniques derived from machine learning, which did not provide direct answers to classic problems such as the accuracy of estimates or causality. Change is under way with the creation of bridges to machine learning and artificial intelligence. In terms of data, it is pointless to pit *sampling data* against *Big Data*. Far better to try to bring them closer, hybridising these two information sources to obtain a third, richer source.

Similarly, in methods and tools, it serves no purpose to set econometrics against machine learning. These approaches have been developed in response to different yet complementary questions, and there is real convergence between these disciplines: econometrics borrows some machine learning methods and vice versa; the causality concepts that econometricians hold dear are among those themes that have been identified as ways to advance machine learning research. Data scientists now have a wider range of tools at their disposal: convolutional neural networks (deep learning), support vector machine approaches (SVM); random forests and boosting, not to mention adapted software and libraries. Nevertheless, we need to remain aware of the possible limitations of Big Data and new tools, and know that predictive machine learning technology could possibly predict what is observed in the data. This convergence has become all the more inevitable with the emergence of quantum computing.

A Special Issue on Big Data and Statistics

Economie et Statistique / Economics and Statistics is devoting two volumes of a special issue to Big Data. This first volume is wider in scope, featuring eight articles with a mix of areas for reflection, applications and methodology. The second volume (forthcoming) will address the theme of price indexes.

The first article by **Clément Bortoli, Stéphanie Combes and Thomas Renault** deals with France's quarterly GDP growth forecast adjusted for seasonal variations and working days. The authors compared the use of a simple autoregressive model (AR) with an AR model featuring a "business climate" variable and a "media sentiment" variable. Elaborating a media sentiment indicator allows us to gauge the overall tone of a media base, and a press title in particular. The integration of this indicator in the model provided some promising results, whether we are looking at the advance GDP forecast (forecasting) or the immediate forecast (nowcasting).

François Robin's article examines the modelling of e-commerce turnover based on data sourced from FEVAD. The Banque de France traditionally uses a SARIMA (12) model, and the author's approach is to complement this model with data from the Monthly Business Survey and data from Google Trends. Illustrating a major advantage of Big Data, Google Trends data is available almost in real time. It analyses the mass of search queries on Google's search engine to construct monthly indices for the terms employed. As independent sources, they are available before the FEVAD results and make a nowcasting approach possible. The technique used stems from machine learning (adaptive lasso method).

Pete Richardson's paper focuses on short-term macroeconomic forecasting and immediate forecasting, (also called nowcasting) that was conducted using Big Data, internet search queries, social media, and financial transactions – i.e. a wider set of databases than the ones traditionally used by national statistical institutes. In a broad-spectrum piece, he analyses a variety of applied research: labour market, consumption, housing market, travel and tourism, and financial markets. The author explains the limitations on what data from web searches can contribute, and he seems to have a preference for data sourced from social media. In his conclusion, he focuses on four particular areas in which to improve these new models and new data: quality and accessibility, information extraction methods, comparison of measurement methods, and improvements to testing and modelling.

The next two articles analyse the contributions of a special type of big data – data sourced from mobile phone operators, which is all the more interesting in light of the penetration rate of these phones among the population. **Guillaume Cousin and Fabrice Hillaireau** tackle the tourism sector and, more particularly, attempt to estimate foreign tourist numbers by counting the number of foreign visitors and their overnight stays. Currently, the survey of visitors from abroad is based on traffic data according to their mode of transport. Using this as a starting point for their estimates, the authors also used counting and surveys. For the time being, this trial conducted since the summer of 2015 has come to the conclusion that mobile phone data is relevant as a complement to the current mechanism rather than a replacement for it. The experiment also identified the limitations of and areas for improvement in this new information source.

The use of mobile phone data to estimate the resident population is studied and analysed by **Benjamin Sakarovitch, Marie-Pierre de Bellefon, Pauline Givord and Maarten Vanhoof**. This exploratory article manages to construct a detailed overview of the current limitations and issues raised regarding this kind of data, as well as its significance and potential. One example of the difficulties they encountered was the uneven territorial coverage caused by variable antenna density which led them to resort to using a Voronoï tessellation (division of the area into polygons of varying sizes). A second problem was how to adjust the data to move away from the subscriber population to the total population. This first exploration has demonstrated that, at the current stage of development, it is still complex and premature to align exact counting statistics such as those currently produced by official statistics. Nevertheless, this mobile phone input source is of potential relevance to some approaches, such as the study of social and spatial segregations.

In the context of media audience measurement, **Lorie Dudoignon, Fabienne Le Sager and Aurélie Vanheuverzwyn** approached a concrete example of the complementarity of panel data and Big Data from a methodological perspective, thereby offering an illustration of the hybridisation of these two database types. Although traditionally based on data from individual sampling, nowadays, the mechanisms for measuring media performance, at least as far as the internet and potentially some television services are concerned, have integrated big data found in real time via equipment such as broadband routers. Once the Big Data found in the objects has been cleaned up ("big" does not necessarily mean "perfect"), the methodological basis for the hybridisation of the two data types is provided by a hidden Markov model, which is used to arrive at an equivalent level of granularity for the two sources, i.e. at the level of individuals, the state of an object such as a router that is supplying no information about the number of viewers or their socio-demographic variables.

The article by **Arthur Charpentier, Emmanuel Flachaire and Antoine Ly** illustrates the necessary convergence between econometric techniques and learning models. The proximity and differences between learning and econometrics are demonstrated. The authors introduce neural networks, the SVM approach, classification trees, bagging, and random forests, as well as sketching out the impact of big data on models and techniques in several application fields. They conclude that, although the two cultures – econometrics and learning – developed in parallel, an ever-increasing number of bridges exists between the two.

Finally, the article by **Evelyn Ruppert, Francisca Grommé, Funda Ustek-Spilda and Babi Cakici** examines the significant issue of trust in official statistics, in the current big data context. The authors return to the importance of respect of privacy, data protection, and especially of the need to rethink the relationship with the public, who supply the raw material used to produce statistical indicators, in particular through national institutes. Big Data, which are not from public sources, influence the notion of trust; the co-production of "citizen data", defined as the participation of citizens in all stages of production, is a fundamental principle. □

BIBLIOGRAPHY

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *StatistikTidskrift*, 15, 429–444.

Desrosières, A. (1993). *La politique des grands nombres. Histoire de la raison statistique*. Paris : La Découverte.

Droesbeke, J.-J., Saporta, G. (2010). Les modèles et leur histoire. In : Droesbeke, J.-J. & Saporta, G. (Eds), *Analyse statistique des données longitudinales*, pp. 1–14. Paris : Technip.

Droesbeke, J.-J., Tassi, P. (1990). *Histoire de la Statistique*. Paris : PUF.

Dwork, C. (2006). Differential Privacy. *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*, 1–12.
https://link.springer.com/chapter/10.1007/11787006_1

Executive Office of the President (2014). *Big Data: Seizing Opportunities, Preserving Value*.
https://obamawhitehouse.archives.gov

Fisher, R. A. (1922). *On the Mathematical Foundations of Theoretical Statistics*. *Philosophical Transactions of the Royal Society*, 222(594-604), 309–368.
https://doi.org/10.1098/rsta.1922.0009

France Stratégie & CNNum (2017). Anticiper les impacts économiques et sociaux de l'Intelligence Artificielle. Rapport du groupe de travail 3.2.
https://www.strategie.gouv.fr/publications/anticiper-impacts-economiques-sociaux-de-lintelligence-artificielle

Hamel, M.-P., Marguerit D. (2013). Analyse des big data. Quels usages, quels défis ? France Stratégie, *Note d'analyse* N 08.
https://strategie.gouv.fr/publications/analyse-big-data-usages-defis

Jensen, A. (1925). Report on the Representative Method in Statistics. *Bulletin de l'Institut International de Statistique*, 22(1), 359–380.

Kiaer, A. N. (1895). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique,* 9(2), 176–183.

Little, R. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics*, 9(2), 407–426.

Nemri, M. (2015). Demain l'internet des objets. France Stratégie, *Note d'analyse* N° 22.
https://strategie.gouv.fr/publications/demain-linternet-objets

Neyman, J. (1934). On the Two Different Aspects of Representative Method Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4), 558–625.
https://doi.org/10.2307/2342192

OPECST (2017). Pour une intelligence artificielle maîtrisée, utile et démystifiée. Rapport N°464.
https://www.senat.fr/notice-rapport/2016/r16-464-1-notice.html

PCAST (2014). Big Data and Privacy: A Technological Perspective. Report to the President.
https://bigdatawg.nist.gov/pdf/pcast_big_data_and_privacy

Rouvroy, A. (2014). Des données sans personne : le fétichisme de la donnée à caractère personnel à l'épreuve de l'idéologie des Big Data. In : *Étude annuelle du Conseil d'État : le numérique et les droits fondamentaux*, pp. 407–422. La Documentation Française

Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics*, 9(2), 461–468.
https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf

Rubin, D. B. (2003). Discussion on Multiple Imputation. *International Statistical Review*, 71(3), 619–625.
https://www.jstor.org/stable/1403833

Singh, S. (2000). *The Code Book*. London: Fourth Estate Ld.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.

Villani, C. (2018). Donner un sens à l'Intelligence Artificielle. Rapport public.
https://www.ladocumentationfrancaise.fr/rapports-publics/184000159/index.shtml.

Nowcasting GDP Growth by Reading Newspapers

Clément Bortoli*, Stéphanie Combes et Thomas Renault*****

Abstract – GDP statistics in France are published on a quarterly basis, 30 days after the end of the quarter. In this article, we consider media content as an additional data source to traditional economic tools to improve short-term forecast / nowcast of French GDP. We use a database of more than a million articles published in the newspaper *Le Monde* between 1990 and 2017 to create a new synthetic indicator capturing media sentiment about the state of the economy. We compare an autoregressive model augmented by the media sentiment indicator with a simple autoregressive model. We also consider an autoregressive model augmented with the Insee Business Climate indicator. Adding a media indicator improves French GDP forecasts compared to these two reference models. We also test an automated approach using penalised regression, where we use the frequencies at which words or expressions appear in the articles as regressors, rather than aggregated information. Although this approach is easier to implement than the former, its results are less accurate.

JEL Classification: E32, E37, C53
Keywords: economic analysis, nowcasting, GDP, media, Big Data, sentiment analysis, machine learning, natural language analysis

Reminder:

The opinions and analyses in this article are those of the author(s) and do not necessarily reflect their institution's or Insee's views.

* *Insee, Economics department, at the time of writing this article (clement.bortoli@gmail.com)*
** *Insee, Statistical methods department, at the time of writing this article (stephanie.combes@gmail.com)*
*** *Paris 1 Panthéon Sorbonne University, CES & LabEx ReFi, IÉSEG School of Management (thomas.renault@univ-paris1.fr)*

Received on 20 September 2017, accepted after revisions on 18 May 2018
Translated from the original version: "Prévoir la croissance du PIB en lisant le journal"

To cite this article: Bortoli, C., Combes, S. & Renault, T. (2018). Nowcasting GDP Growth by Reading Newspapers. *Economie et Statistique / Economics and Statistics*, 505-506, 17–33. https://doi.org/10.24187/ecostat.2018.505d.1964

Because macroeconomic data are only known after a certain time lag, it is vital for policy makers to have tools enabling them to forge a real-time analysis of the economic situation. For example, GDP statistics in France are published on a quarterly basis, with a delay of 30 days after the end of the quarter. To get a forecast (or a nowcast) of GDP growth before its publication, forecasters traditionally use business climate surveys, conducted by different institutes, as the main information source. These are questionnaires made up of qualitative questions sent every month to a sample ranging from several hundred to several thousand business leaders. The answers are summarised as "opinion balances", i.e. by calculating the difference between positive and negative responses. Some synthetic indicators are also calculated from these opinion balances, such as the business climate that considers the economic situation overall or by sector. These various indexes are sometimes called "leading indicators", since they are available before official figures are published. It is also possible to forecast the GDP for the current quarter (which is obviously not known before the end of the quarter): this type of forecast is referred to as "nowcast". It can also be interesting to forecast GDP for the coming quarter, which is also possible *via* business climate surveys that contain prospective balances.

The current explosion of Internet content, as well as the technical progress associated with the "Big Data" offer the possibility of building alternative economic indicators in real time. From this point of view, media content is particularly interesting, because its properties are quite similar to those of surveys on business climate. Indeed, this information is available instantaneously and contains qualitative details about the economic situation several weeks before official data are published.

Thus, the aim of this paper is to use the content of a major media Internet site to improve real-time GDP growth forecasting (or nowcasting). It will be particularly interesting to compare the predictive power of this information with that from business surveys traditionally used, in order to determine if the two approaches substitute for each other, complement each other, or if one of them appears more accurate than the other.

For the purpose of this article, we chose the French newspaper *Le Monde* website. The content of this website is covering a time scale rare in France, in particular including many articles published in the paper edition before the advent of the Internet. In addition, it is the leading information website in France. We therefore built a database containing more than a million articles published in this newspaper from 1990 to today. We first sorted this database by combining statistical models and textual analysis, retaining only articles covering the French economic situation; this results in a sample of approximately 200,000 news items. We then use the information contained in this reduced database, following two different strategies.

The first strategy requires the use of a sentiment dictionary, in other words a list of terms with positive or negative connotations from an economic viewpoint. Such dictionaries are widespread in English, less so in French. We therefore built one containing 548 positive terms and 1,295 negative terms. These terms are then identified in each article in the database, which are allocated a "sentiment score" based on the number of positive and negative terms it contains. In this way it is possible to summarise the information contained in the database as a unique numerical indicator, called media sentiment. This can then be used in simple regression models (autoregressive or AR, augmented).

We then carry out a real-time[1] forecasting exercise over the period 2000-Q2 - 2017-Q3, which means that we make forecasts for a given timescale every quarter from the second quarter of 2000 to the third quarter of 2017, each time only using the data available up to that date. We compare the accuracy of each model by calculating the RMSFEs (Root Mean Square Forecast Error) from the series of forecast errors relative to the real value calculated in this way. Thus, we find that a model combining "Media Sentiment" and "Business Climate" provides, over certain timescales, significantly greater accuracy compared to an augmented AR model of business climate surveys alone.

The use of a "handmade" dictionary may appear to be somewhat arbitrary, costly, and imprecise since all the available information is summarised in a single indicator. A way to deal with these drawbacks is to consider automatic

[1] *Strictly speaking, one should talk about* pseudo real time*, as the sentiment dictionary is constructed* ex-post *by experts. However, for ease of language, we will speak of* real time *in the rest of this article.*

methods, which would avoid prior judgment on the terms to be selected or their connotation, while also keeping the information in a disaggregated format. The automatic methods used here also have the advantage of being relatively inexpensive to implement. It involves constructing the series for the frequency of appearance (or a weighting similar to the concept of frequency) for each term and combinations of two terms (or bigrams): to do so, the terms are first stemmed to bring singular and plural to the same form. These time series are then used for forecasting as part of penalised regressions (Elastic-Net). Penalisation ensures a selection of regressors and therefore the parsimony of the model, which helps to prevent a risk of over-adjustment, especially high given the large number of variables available. However, the calculation of RMSFEs suggests that an automatic method for selecting words does not significantly improve the forecast compared to the autoregressive model augmented with the business climate indicator.

The structure of this article is organised as follows: A brief literature review is presented in the first section. The second section describes the data used and the way they are handled. The econometric models used are then described in the third section. The fourth section presents the results obtained. The fifth section concludes.

Literature Review

It is possible to separate the literature dealing with GDP nowcasting into two major categories. On the one hand, a part of the literature is dealing with techniques aimed at choosing the best forecasting model from a predefined "traditional" set of variables. These papers are generally devoted mainly to comparing the predictive performance of different approaches: bridge models, state space model, mixed-data-sampling, blocking, etc. Among others, we can refer to Baffigi *et al.* (2004) and Foroni & Marcellino (2014). More recently, Bec & Mogliani (2015), in a paper devoted to comparing combinations of models and combinations of information, offer an instructive summary of the different techniques that can be used to make a macroeconomic forecast. On the other hand, a part of the literature, using a predefined model, deals with improving the forecast by considering the addition of new explanatory variables. Here we focus our attention on this second segment of the literature.

Four main types of variables are used in the literature: (1) "quantitative" variables (industrial production, retail sales, etc.), published monthly with a time delay of 30 to 45 days; (2) "qualitative" variables (surveys, polls, etc.), available at the end of every month; (3) "financial" variables (interest rate, stock market index, etc.) available in real time; and (4) "alternative" variables (Google Trends, media sentiment, etc.) often available in near real time.

There is a consensus regarding the contribution of adding "qualitative" variables, mainly when the "quantitative" information about the current quarter is not yet available. For example, by analysing the contribution of each variable depending on the timing of the GDP forecast for the current quarter (1st month, 2nd month or 3rd month), Angelini *et al.* (2011) showed that "qualitative" information carried greater weight for the first estimates, while the predictive power of "quantitative" information becomes predominant for estimates in the 3rd month. This change is explained quite simply by the fact that "quantitative" information about the current quarter starts to become available during the 3rd month (e.g. industrial production for January 2016 was published on 15 March 2016 and can therefore be used to nowcast GDP for the 1st quarter of 2016 conducted during the 3rd month of the same quarter): this "quantitative" information is used by national accountants in order to construct the GDP on a quarterly basis. The contribution of qualitative information was confirmed by Darné (2008), among others, for the specific case of France.

The conclusions are more mixed regarding the contribution of financial variables. According to Andreou *et al.* (2013), adding financial variables improves the accuracy of the model, while the opposite findings are presented by Banbura *et al.* (2013). This difference is explained by the fact that Andreou *et al.* (2013) do not use the high frequency of indicators by extrapolating the monthly data over the quarter (unlike Banbura *et al.*, 2013), making it difficult to compare the two studies.

Finally, more recently, different studies have focused on the contribution of "alternative" variables. For example, Choi & Varian (2012), McLaren & Shanbhogue (2011), Fondeur & Karamé (2013), and D'Amuri & Marcucci (2017) showed that the change in the search volume on Google Trends for certain keywords ("jobless claims", "unemployment benefits") improved the forecast for the change in unemployment rate. Regarding the contribution of *Google Trends* to predicting the French economic situation, more mixed findings were put forward by Bortoli & Combes (2015).

Content published in the media has also been used extensively in finance to forecast the change in financial markets (Tetlock, 2007; Garcia, 2013). One possible approach is to calculate a sentiment score for a press article, then to construct a time series for "sentiment" by aggregating the scores for articles published over a given period (e.g. every month). To do so, a dictionary containing a list of "positive" keywords and a list of "negative" keywords, either generic (Harvard IV dictionary) or specific to the study area (e.g. Loughran & McDonald's financial dictionary, 2011), is used: the sentiment of each article is then simply defined using the frequency of words from the dictionary in the body text weighted by their score (in the simplest case, +1 for a word with positive connotations and -1 for a word with negative connotations).

The approach founded on a dictionary or sentiment score is not always based on a binary positive/negative approach: Baker *et al.* (2016) looked at the change in the number of articles containing at least one keyword linked to a sentiment of uncertainty and dealing with economic policy, in order to create a new index (Economic Policy Uncertainty Index).[2] The authors of this new index show that an increase in media uncertainty helps to forecast changes in GDP.

Another possible approach using "media" data consists of analysing the change in the frequency of appearance of different subjects detected automatically using an unsupervised approach such as Latent Dirichlet Allocation. Applying this methodology to Norway, Larsen & Thorsud (2015) showed that the variation in frequency of appearance of certain subjects may be used to improve the forecast of economic fluctuations.

In this article, we focus on the forecast at the end of the 1st month, the 2nd month and the 3rd month of the current quarter and the previous quarter. We compare the accuracy of a simple AR model with an AR model augmented by the business climate and an AR model augmented by alternative "media" data (summarised or disaggregated).

Data

The Original Database

Among the different French media whose content can be used to construct a media sentiment indicator, *Le Monde* has interesting features. It is one of the leading French newspapers: the printed edition has the second highest national circulation behind *Le Figaro* (approximately 260,000 copies per day) and its website lemonde.fr is the most-visited information site in France, just ahead of the website of *Le Figaro*. In addition, the media content available online covers a remarkable time scale for France, including many articles published in the paper edition before the advent of the Internet. Thus, it was possible to create a database of 1,405,038 online articles published since 1990.

It might also be interesting to use articles from specialised economic newspapers such as *Les Echos* or *La Tribune*. In fact, *Les Echos* website also has interesting properties, as articles have been available since 1991. However, this newspaper has a lower media outreach than *Le Monde* (whether in terms of number of printed copies sold or visits to the Internet site): for this article, we chose to favour the "general public" source. It could therefore be interesting for a future study to estimate if "specialist" information has stronger predictive power than generalist media. On the other hand, the use of *La Tribune* seems more problematic: the risk of a break in the series over a long period is high in relation to the predictive power of online content, due to the radical change in editorial line that occurred in 2012.

The number of articles contained in the database varies strongly depending on the period, most of the time between 2,000 and

[2] *www.policyuncertainty.com. For France, the EPU index is based only on articles from the newspapers* Le Monde *and* Le Figaro *(which justifies the comparison we are making* infra *between this indicator and our media sentiment index). However, for the United States, the EPU indicator is based on three components, one of which relates to the media.*

6,000.[3] This limit is exceeded between 2000 and 2002, where the series reached its maximum (11,000 in March 2001), then more briefly in 2012. Since 2013, the number of articles per month has oscillated around 4,000.

Construction of a Restricted Database

First, it is necessary to sort this database to keep only the articles relevant to our study, i.e. those covering economic topics and dealing mainly with the situation in France. In fact, keeping more articles could interfere with summarising media information and its use in forecasting. It is also necessary to remove articles from the database dealing with information published by statistical institutes (Insee, Dares – the ministry of Labor statistical services, *Pôle emploi* – the French employment agency, etc.), because the information we are looking for in the media content has to be independent from these sources. Moreover, some articles are reserved for subscribers: in this case, only the title and first lines are freely available. We restrict our analysis to articles where there are at least 50 characters freely accessible.

We first discard all articles not dealing with economics. The more recent articles (since 2005) are already classified into categories (economics, international, politics, sport, etc.) by journalists at *Le Monde*. This classification is registered in the metadata for each article, and can therefore contribute to identifying articles dealing with economics among the older texts that have not been pre-classified by the journalists. A supervised learning algorithm is calibrated using a sample of 25,000 articles from the "economics" category and 25,000 articles from other categories: the algorithm calculates the probability of an article belonging or not belonging to the "economics" category, based on the frequency of appearance of words contained in it for the two sets of the training dataset. In this way, the presence of the word "employment" in an article will increase its probability of belonging to the "economics" category, as in the training dataset this word is most frequent in articles covering economics than in the others. Such an algorithm, which can be qualified as "naive Bayesian" (Kotsiantsis *et al.*, 2006), makes it possible to classify all the oldest texts in the database very rapidly. By analysing the accuracy of the classification on 10,000 articles (out-of-sample), we get a classification accuracy of 89.7%; this supports our use of this type of approach to categorise all the articles in our database.

In parallel, the articles that focus mostly on France are identified by another procedure. Two lists containing the names of geographical entities are used: one is made up of French toponyms (names of towns, *départements*, regions) and the other international toponyms (names of countries and capitals). We retain only the articles that include at least as many French entities as foreign entities.

The final sample contains 194,848 articles. The proportion of articles retained each month oscillates between 10% and 20%. This proportion seems to follow a falling trend over the recent period: it was 18% in 2009 and not more than 13% in 2016.

The Traditional Economic Indicators: Insee Business Surveys

One of the important ideas of the article is to compare the information contained in the media content with that synthetized in more traditional economic variables such as business surveys.

Business climate surveys are used to follow the recent and current economic situation, and to forecast short-term changes. They are conducted every month among company managers. They provide an overview of a given business sector, highlighting the fields that are not covered, or covered more belatedly, by traditional statistics. The information gathered in business climate surveys are referred to as qualitative, because the respondents are asked to assign qualities and not quantities to variables about which questions are asked.

For France, the three main producers of business climate surveys are Insee, *Banque de France* and Markit (PMI surveys). For this paper, we relied only on Insee business surveys, and more particularly on the synthetic Business Climate indicator. This is the common component, extracted using factor-based analysis techniques, of 26 business surveys in five different sectors (industry, services, construction, retail and wholesale trade). The Business Climate indicator is normalised: over the long period, its mean is 100 with a standard deviation of 10.

[3] *The database is visibly atypical in 2006, where the number of articles per month was highly discontinuous compared to prior and subsequent periods (barely more than 1,000 articles per month).*

The Variable to Be Forecast: GDP Growth

The variable we aim to forecast is the real quarterly volume growth of the French GDP (chain-linked), seasonally and working-day adjusted, published by Insee. There are three publications for each quarter (two before 2016): a first estimate 30 days after the end of the quarter, a second estimate 60 days after the end of the quarter and "detailed figures" 85 days after the end of the quarter.[4] The quarterly growth figures may still change further over three years, until the national accountants publish the final account for the year considered. After this date, GDP growth for a given quarter no longer changes beyond the normal fluctuations associated with corrections for seasonal variations.

Knowing whether it is best to measure the performance of a forecasting model over the series of first publications of GDP or over a given recent vintage ("final" series) is a question with no obvious answer. As stated by Bec & Mogliani (2015), it is possible to defend an economic forecast having the main purpose of giving political decision-makers the best possible estimate of business activity: from this viewpoint, it would be better to test our models using given historical data, preferably the most recent possible ("final" series). In fact, GDP growth values in this case closely match the best possible measurement of economic activity, once all the information is available. Thus, Mogliani & Ferrières (2016) show that, in the case of France, revisions of the GDP are generally not biased, but that the first growth estimates do not efficiently use all the available macroeconomic and financial information.

Nonetheless, from a pragmatic viewpoint, it is true that the performance of one forecasting method is *de facto* judged in the light of the first GDP figures published. In this article we have thus chosen to adopt a real-time approach, i.e. using historical first publication data. In particular, this is justified by the fact that we are using time lags in GDP as explanatory variables in our model: so, we are using the information that was available during the quarter to be forecast. Nonetheless, as a precaution, all the estimations in this paper have also been made using a recent vintage of GDP growth: the results are very similar to those presented here.

Models

We propose two different strategies for extracting media information from the database and for using it in forecasts. The first consists in constructing a "Media Sentiment" index that offers a numerical figure for the general tone of the articles, following a procedure similar to that applied in Bortoli *et al*. (2017). The second uses all the available information by calculating the change over time in the occurrence frequencies of the terms in the database. These time series are then used in forecasts as part of penalised regressions.

Constructing an Indicator of "Media Sentiment" and Using it in a Forecast Model

A first strategy for extracting the media information from the database consists in constructing a Media Sentiment indicator that gives a score for the general tone of the articles in the database. The main advantage of this method is that it provides a tool very similar to more traditional economic indicators, such as Business Climate indicators: thus, it will be possible to compare the predictive performance of our Media Sentiment indicator to the Business Climate constructed by Insee. In addition, this is an easily interpreted indicator: a simple reading discloses the cyclical position of the economy as established by the indicator.

Choice of Frequency for the Media Sentiment Indicator

The first strategic choice to be made for the media sentiment indicator is its frequency. Given the database created, it would be possible to create a quarterly, monthly, weekly or even daily index. We have chosen to ignore the last two possibilities:

- A daily indicator would risk appearing too volatile, even more since the number of articles published is likely to vary significantly from one day of the week to another (with a particular drop at weekends, especially Sunday);

- A weekly indicator would be problematic to use to forecast a quarterly variable such as GDP, given that these two frequencies do not "fit" one

[4] *Before 2016, the first results were published 45 days after the end of the quarter and there was no additional publication before the detailed figures.*

inside the other (a quarter does not contain a fixed whole number of weeks). In addition, such an indicator could present a risk of volatility that would still be too high.

It therefore remains to choose between quarterly and monthly frequencies. The first solution would have the advantage of minimising the noise contained in the indicator. However, it would be necessary to wait for the end of the quarter to calculate it. Conversely, a monthly indicator offers a forecasting model from the first month of the quarter, without waiting for its end. Thus, the monthly indicator appears to offer the best compromise between volatility and frequency/speed of availability (in theory, from the end of every month). Furthermore, this frequency is also chosen by major institutions to publish their main economic and business climate statements.

Construction of the Sentiment Dictionary

Calculating a media sentiment indicator requires being able to quantify the positive or negative tone of the articles selected; to do so, we use a "sentiment dictionary". This is a list of terms that may have positive or negative connotations. Many dictionaries already exist in English to analyse texts: the Harvard IV-4 Psychological Dictionary is the main one, but other dictionaries are used for specific research fields, such as the Loughran & McDonald dictionary (2011) in the field of finance. However, this type of pre-existing list is much rarer in French: it is therefore necessary to construct one for the needs of this study.

We began by stemming all the terms encountered in the corpus studied using the Snowball algorithm adapted for the French language (Porter, 2001). We then assigned a sentiment to all stems appearing more than 500 times in the corpus (i.e. 5,575 stems), based on three possible ratings: positive, neutral or negative.

However, building a dictionary comprising exclusively unique stems (or unigrams) could prove problematic. In reality, a stem such as "increase" does not have the same value depending on whether one is discussing an increase in growth or unemployment. To overcome this type of ambiguity, we supplemented the dictionary with a list of bigrams, i.e. pairs of stems. Similar to what we did for the unigrams, we identified the commonest 5,000 bigrams from the corpus, then we classified them according to the same three ratings. In total, the dictionary contains 840 terms, 281 positive and 559 negative.[5]

Allocating a Score to Each Article and Calculating the Media Sentiment Indicator

From the dictionary created, a "sentiment score" is attributed to each article *i*, depending on the number of positive and negative terms that it contains. Several scoring systems can be considered. The simplest coding consists in adopting a discreet score for each article (discrete coding). The attributed score is 1 if the article contains more positive than negative terms, -1 if it contains more negative than positive terms and 0 if the two categories are equal. Discrete coding has the merit of simplicity, but it does not distinguish articles where the overall connotation is very marked from those where it is more subtle. It may therefore be interesting to consider an alternative scoring system, where the score can be established on a continuous scale between 1 and -1 (continuous coding). To do this, we calculate for each article the difference between the number of positive words and number of negative words, then we normalise for the number of words in the article.

The value of the sentiment indicator for month *t* is then a simple arithmetic mean of the sentiment scores obtained for each article *i* published during the month. Labelling *n(t)* the number of articles published in month *t*, $S_{i,t}$ the sentiment associated with each article *i* published during month *t*, we therefore define a monthly sentiment variable $MediaSent_t$, such that:

$$MediaSent_t = \frac{1}{n(t)} \sum_{i=1}^{n(t)} S_{i,t}$$

It is thus possible to calculate two monthly media sentiment indicators: one based on continuous coding and the other based on discrete coding. These two indicators are obviously very similar over the period[6] (Figure I): this result is already reassuring in itself, as it shows that our method makes it possible to extract from the articles database an overall media sentiment that does not depend too much on the parameters chosen to do so. We also note that the indicator is always negative,

[5] The dictionary is available online: http://www.thomas-renault.com

[6] In Figures I, II, III and IV, the media sentiment indicators are smoothed for reasons of legibility (3rd order moving averages). However, the raw indicators are used in the forecasting models.

irrespective of the chosen coding, which denotes a generally pessimistic bias across the articles selected by the filter. Furthermore, we note that using continuous coding leads to obtain a less volatile indicator than discrete coding and better considers the nuances developed in the text of these articles. In the rest of this article, we select the continuous indicator as it provides better forecasting results.

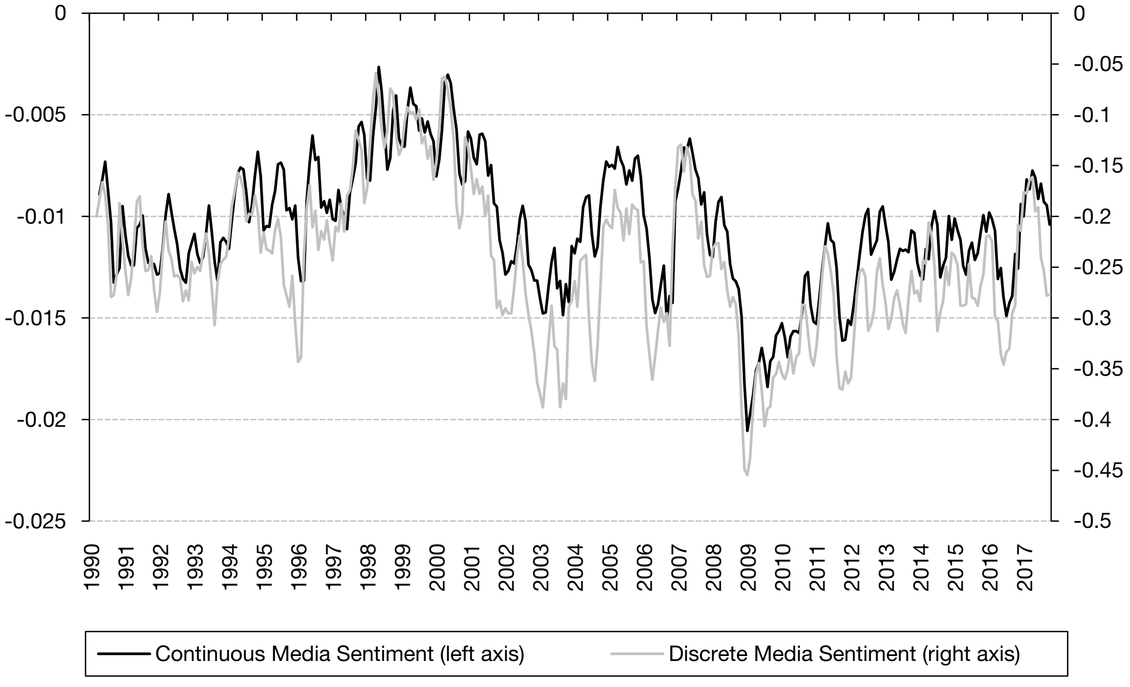
Over the whole period, the Media Sentiment indicator also appears to follow the main growth trends closely (Figure II), even if it does not fit very well the quarterly ups and downs, especially over the recent period. However, this does not disqualify it, as the sudden quarterly variations in GDP may be due to specific phenomena that an economic indicator does not always capture. Nonetheless we observe two significant divergences between our indicator and business activity. First, the indicator diverged abruptly in 2006, while business activity experienced no particularly noticeable deviation in that year (apart from a weak third quarter). Second, at the end of the crisis, the indicator only recovers gradually after having reached a low point in 2008-2009, although business activity rebounded vigorously over the same period. This created a divergence between the two series, which only disappears in 2011, when business activity slumped again following the Eurozone sovereign debt crisis.

In addition, our indicator is obviously quite similar to the Business Climate published by Insee (Figure III). However, we note that, while the two series follow identical major trends, the Insee Business Climate reveals short cycles lasting one or two years (particularly visible at the beginning of the period), absent from the Media Sentiment indicator. In the same way, the divergences already observed by comparing our indicator with business activity (in 2006 and post-crisis) are also visible here.

Finally, an overall similarity can be observed between our Media Sentiment indicator and (the opposite of) the "Economic Policy Uncertainty" (EPU) indicator described by Baker *et al.* (Figure IV).[7] Once again, two significant exceptions can be noticed. First, the media sentiment

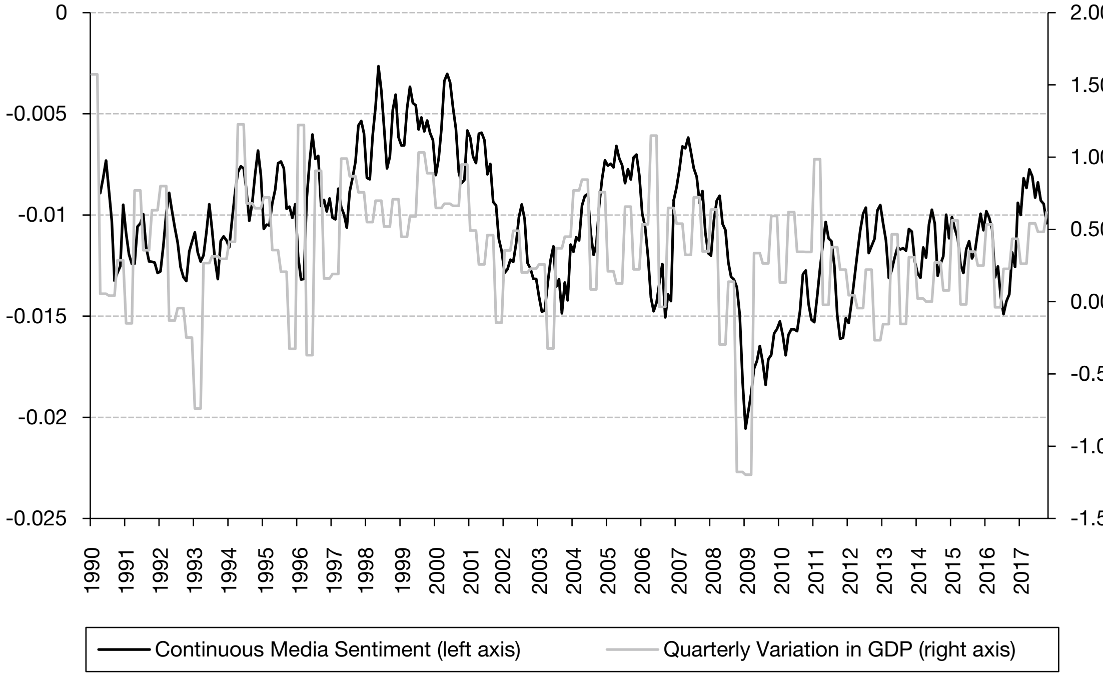
[7] *As the EPU indicator is an index of uncertainty, we have reversed the scale for the latter in order to compare it with our media sentiment, so as to make the graph easier to read (increasing uncertainty is actually consistent with decreasing sentiment)*

Figure I
Discrete and Continuous Media Sentiment Indicators – 3-Month Moving Average



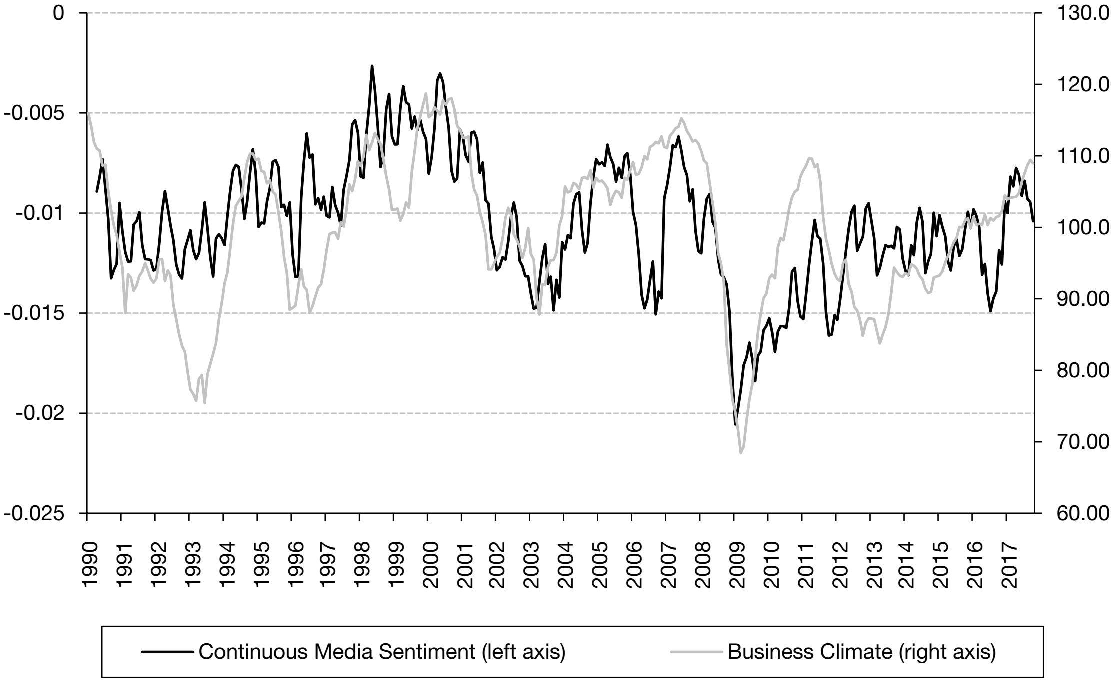
Note: This graph illustrates the change in the media sentiment indicator (3-month moving average), calculated on the basis of continuous coding and discrete coding.
Source: *Le Monde* authors' database.

Figure II
Continuous Media Sentiment Indicator and Quarterly Variation in French GDP



Notes: This graph illustrates the change in the media sentiment indicator (3-month moving average) and quarterly variation in French GDP.
Sources: *Le Monde* authors' database; Insee.

Figure III
Continuous Media Sentiment Indicator and Insee Business Climate



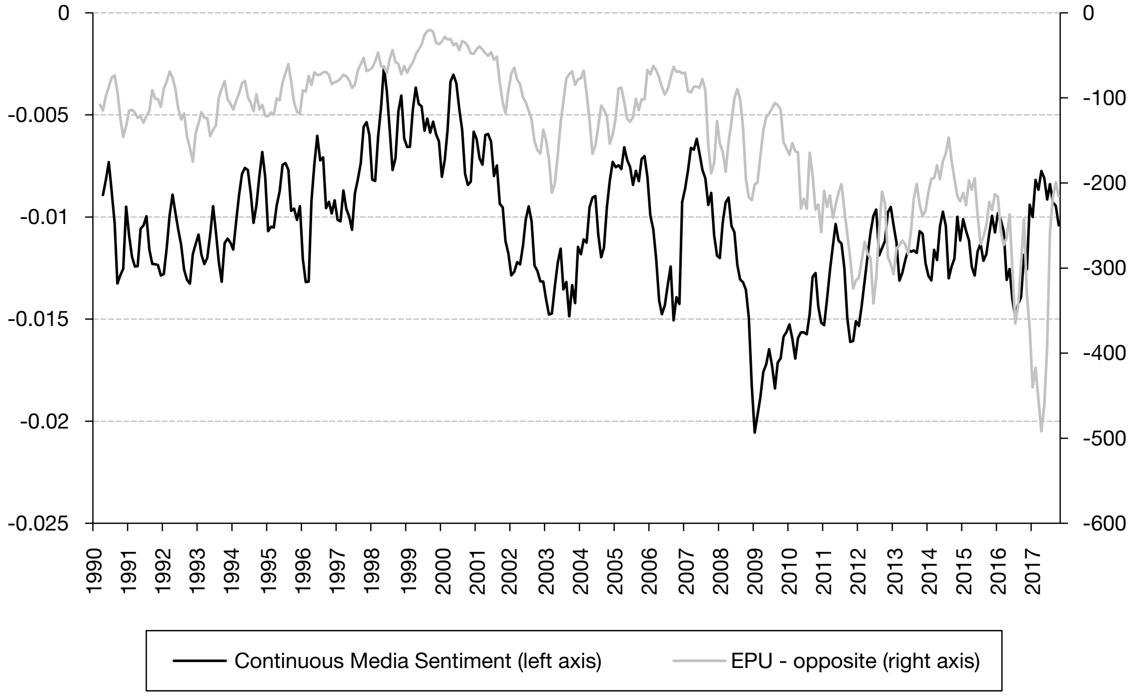
Notes: This graph illustrates the change in the media sentiment indicator (3-month moving average) and the Insee Business Climate indicator.
Sources: *Le Monde* authors' database; Insee.

indicator diverges more quickly and more significantly than the EPU of Baker *et al.* at the time of the 2009 financial crisis. Conversely, the latter shows a significant rise in uncertainty during 2016-2017, certainly due to the elections in France and rising influence of the *Front National* (perhaps with a Brexit effect), while our media sentiment indicator is fairly stable.

In both cases, our media sentiment indicator experiences changes more similar to economic activity than the EPU of Baker *et al.*: thus, we can expect *ex-ante* that the EPU is less effective than ours for forecasting.

Our graphical observations are confirmed by a simple analysis of correlations of the different series considered. The Insee Business Climate indicator is slightly more correlated to GDP growth than our media sentiment indicator, which may be a sign of better forecasting performance. Moreover, the Insee climate and sentiment indicator are fairly well correlated with each other. Finally, correlations of the EPU of Baker *et al.* with the other variables (and in particular with GDP growth) are weaker, which confirms our suggestion of lesser predictive capability (Table 1). Nonetheless, we can see that it is slightly better correlated to

Figure IV
Media Sentiment Indicator and (opposite) Economic Policy Uncertainty Indicator of Baker *et al.* for France



Note: This graph illustrates the change in the media sentiment indicator (3-month moving average) and the Economic Policy Uncertainty of Baker *et al.* (3-month moving average, opposite).
Sources: *Le Monde* authors' database; Baker *et al.* (2016).

Table 1
Correlations Between GDP Growth, Media Sentiment Indicator, Insee Business Climate and the EPU of Baker *et al.* (Opposite)

	Media Sentiment	Insee Business Climate	EPU (opposite)
GDP growth	0.469	0.547	0.268
Media Sentiment	-	0.575	0.389
Insee Business Climate	-	-	0.253

Note: The figure at the intercept of row *i* and column *j* corresponds to the correlation between the variable displayed in row i and that displayed in column j. For parsimony each correlation is shown only once.
Sources: *Le Monde* authors' database; Insee; Baker *et al.* (2016).

our media sentiment than to the other two variables, which seems to show a certain specificity of the media information. The statistics describing the different indicators are presented in an appendix.

Using Media Sentiment Indicators in Forecasting

The continuous monthly media sentiment indicator is used to forecast GDP growth for the current quarter. Several techniques can theoretically be considered to handle the difference in frequency between the variable to be forecast (quarterly) and the explanatory variables (monthly). A first possibility would be to use the MIDAS method (see, among others, the work of Ghysels *et al.*, 2005; 2007) that is designed to forecast a low frequency variable using high frequency explanatory variables. For this paper, we rather opted for an approach similar to "blocking", commonly used by forecasters (e.g. see Bec & Mogliani, 2015), which consists in using a different forecasting model (or "calibration") for each month of the quarter, each time using all the information available at the date considered. Thus, the "month 1", "month 2" and "month 3" calibrations use, respectively, all the information available at the end of the first, second and third months of the quarter. In practice, for example for the Business Climate (for which we consider the first difference) we will label $Climate_t$ the regressor that will correspond, in forecast "month 1", to the variation between the value of the business climate for the 1st month relative to the mean of the values taken for the three months of the previous quarter. In "month 2", we will consider the mean value for the two months of the current quarter relative to the value of the previous quarter. In "month 3", we then have all the information. The same logic is adopted for the variable $MediaSent_t$, except the fact that it is taken as level and not as first difference.[8] The first lag of GDP growth is also used as explanatory variable, when it is available (which is not the case, for example, in month 1).[9] However, we do not use the EPU indicator of BBD as explanatory variable: actually, our first graphic and correlation analyses were confirmed by the fact that this indicator does not improve the predictive performance of our models.

Since one of the aims of the article is to compare the respective performance of Insee Business Climate and the "Media Sentiment indicator", four models are considered for each month in the quarter: the first only uses the past variation in GDP (simple AR with the first lag of GDP growth when it is available, otherwise the second), the second includes the first lag of GDP growth and the Media Sentiment indicator, the third the first lag of GDP growth and the Business Climate, finally the fourth includes both the first lag of GDP growth, the Media Sentiment indicator and the Business Climate in France. The forecasting performance of these models are measured in real-time conditions. The models are estimated from the first quarter 1990 and up to a sliding date from the second quarter 2000 to the third quarter 2017, which supplies a list of forecasting errors from which we can calculate an RMSFE for each model.

To nowcast the current quarter, the models can be formalised as follows (to forecast the next quarter, only the index of the dependent variable changes).

$$\Delta GDP_t = \alpha + \beta_1 \cdot \Delta GDP_{t-1} + \varepsilon_t$$

$$\Delta GDP_t = \alpha + \beta_1 \cdot \Delta GDP_{t-1} + \beta_2 \cdot \Delta Climate_t + \varepsilon_t$$

$$\Delta GDP_t = \alpha + \beta_1 \cdot \Delta GDP_{t-1} + \beta_2 \cdot MediaSent_t + \varepsilon_t$$

$$\begin{aligned}\Delta GDP_t = \alpha + \beta_1 \cdot \Delta GDP_{t-1} + \beta_2 \cdot \Delta Climate_t \\ + \beta_3 \cdot MediaSent_t + \varepsilon_t\end{aligned}$$

We present the estimates in full sample for equations 1 to 4 in Appendix 2. The media sentiment variable is significant at the 1% threshold in all models.

Using Penalised Regression for Forecasting

Constructing a media sentiment indicator provides a simple and readable tool comparable with more traditional economic indicators such as the business climate. However, it also has disadvantages. Firstly, it depends largely on the researcher's preconceptions: on the one hand, the terms in the sentiment dictionary are classified by experts and therefore based on presuppositions, on the other hand choices

[8] *This choice provides the best fit of the data in-sample and offers the best forecasting performance out-of-sample.*

[9] *Longer time lags of the growth of GDP were rarely significant in samples and did not substantially improve the performance of forecasting models. In general, adding them only modified the models at the margin: in the end, we therefore chose not to include them and to keep the models lean.*

have to be made about scoring the articles and aggregating the scores, for which there is no "natural" method. In addition, calculating a simple summary indicator does not allow full use of the richness of the database and therefore creates the risk of ignoring part of the information that could turn out to be useful in forecasting.

Thus, we offer a second forecasting method, leaving less space for the researcher's preconceptions and making better use of the diverse information contained in the database. The regressors used in this approach are the weightings of each term of the vocabulary (i.e. all the terms used at least once in the corpus of articles): however, we exclude the so-called "stopwords", i.e. words very often used (determinants, certain adverbs) and therefore in principle not discriminating. Similarly, we have also eliminated the commonest terms (present in more than 90% of documents) and the rarest terms (less than 5% of the time). Furthermore, as previously, the terms are stemmed and the combinations of two consecutive terms, or bigrams, are also considered to take better account of expressions such as "labour market".

We calculate the weightings associated with each term of the vocabulary using the tf-idf approach (term frequency-inverse document frequency) used extensively in the literature on information retrieval (e.g. see Breitinger *et al.*, 2015).[10] This weighting has proven more relevant than the frequency of terms when the documents handled (here, articles) are long. Using the frequency of the word in the document and the inverse of the frequency of documents containing this word, it is possible to make better use of a frequent word within an article if it is little used elsewhere. The weightings for each word from each article of the corpus can then be averaged by month or quarter, so that regressors are available at the same frequency as the dependent variable.

Once these variables have been obtained, we can apply the usual transformations to them: thus, we also retain their first lag, their growth rate and moving average over two quarters. In total, we obtain approximately 6,000 potential regressors. As this is a very large number, even greater than the number of points in the series to be forecast, it is necessary to select a sub-set of regressors. Actually, it is better for the forecast to focus on parsimonious models, i.e. that only use a limited number of variables. This is necessary to avoid overlearning phenomena: selecting too many explanatory variables generally degrades the predictive performance of the model outside the estimation sample. To do so, we use one of the most commonly-used techniques for automatic variable selection: penalised regression.

Penalised regression is a simple linear regression, to which we add a constraint (or penalty) regarding the amplitude of the coefficients associated with each regressor. This amplitude can be measured using different norms: we talk about Lasso regression when the amplitude is measured using norm L1 (sum of absolute values of coefficients) and Ridge regression when norm L2 (Euclidean) is used. As the Lasso penalty has the property of being quite abrupt and often leads to models that are too parsimonious, we use a combination of the Lasso penalty and the Ridge penalty: this is referred to as Elastic-Net regression.

Penalised regressions offer greater robustness than iterative techniques such as *stepwise*, and they have the advantage of being configurable, the hyper-parameters corresponding to the size of the penalty. By seeking parameters optimising forecasting performance, we can favour the selection of regressors with better predictive power. More precisely, hyper-parameters are optimised by "grid search": for different values of the parameters, we use a sliding window and produce a listing of forecasting deviations, from which we calculate an RMSFE. We then select hyper-parameters minimising the RMSFE.[11]

Results

In this section, we present the results using the Media Sentiment indicator computed from our dictionary with a continuous coding as well as those supplied by the automatic penalised regression method.

We present the RMSFEs of the different models depending on the month of the quarter at which

[10] In information retrieval, tf-idf weighting is used to represent documents (e.g. web pages) in the form of numerical vectors that can then be compared with the numerical vector corresponding to a query; it is then possible to put documents in order based on their relevance to the query (e.g. a query from a user in the search engine).

[11] So as not to bias the results towards this approach, the sliding window used is not the same as that from which RMSFEs are produced from the different methods compared in this study. The RMSFEs are therefore produced over the period from the 1st quarter of 1998 to the last quarter of 1999.

Table 2
RMSFE of Models for Forecasting GDP Growth Rate in Quarter Q for Different Forecast Time Scales

	Forecast month	Month 1 (Q-1)	Month 2 (Q-1)	Month 3 (Q-1)	Month 1 (Q)	Month 2 (Q)	Month 3 (Q)
	Month before publication	6	5	4	3	2	1
[1]	AR(1)	0.4057	0.3941	0.3941	0.3927	0.4039	0.4039
[2]	AR(1) + Sentiment	0.3968	0.3951	0.3931	0.3798	0.3727	0.373
[3]	AR(1) + Elastic-Net	0.3781	0.3955	0.3904	0.3793	0.3672	0.3820*
[4]	AR(1) + Climate	0.3434*	0.3475*	0.3459*	0.3406*	0.3689	0.3712
[5]	AR(1) + Elastic-Net + Climate	0.3642	0.3879	0.3835	0.3755	0.3552	0.3749
[6]	AR(1) + Sentiment + Climate	**0.3357**	**0.3446**	**0.3403**	**0.3281**	**0.3331***	**0.3326***

Note: This table presents the RMSFEs from models [1] to [6]. For each time scale (each column), the lowest RMSFE is shown in bold. For each month of the quarter and each model, the asterisk * indicate that, according to the Harvey *et al.* (1997) test, the Root Mean Square Forecast Error (RMSFE) of the model is significantly less than for the benchmark model (at the threshold of 10%). Models [2], [3] and [4] are compared to model [1]. Models [5] and [6] are compared to model [4]. For example at month 2 in Q, the RMSFE of model [6] (AR(1) + Sentiment + Climate) is significantly lower than that of model [4] (AR(1) + Climate).
Sources: *Le Monde* authors' database; Insee; authors' calculation.

the forecast is made (Table 2). We test the assumption that the model combining Media Sentiment and Business Climate provides a significantly better forecast than the other models using the Harvey *et al.* (1997) test.

Individually, model [2] (AR + sentiment) provides slightly better accuracy than model [1] (simple AR) for the current quarter (*nowcasting*), but this improvement is not significant. Model [4] (with climate) has superior properties. Nonetheless, when we combine climate and sentiment, the predictive performance of the model is superior (model [6]) to that for climate use alone (model [4]). This is particularly sensitive with effect from month 2 of the current quarter. For all time scales, the forecast from model [6] is more accurate than the other models. The Harvey *et al*. (1997) test shows us that this difference is significant for months 2 and 3 of the current quarter at a 10% threshold.

This result tends to show that individually, the Insee Business Climate remains a more reliable economic indicator than our Media Sentiment. Nonetheless, the Media Sentiment contains information in addition to that contained in the business climate, improving the forecast of French GDP.

Model [3] (penalised regression) also demonstrates superior performance compared to the autoregressive model [1] for some time scales. However, when we add business climate, a variable already having great predictive power, the disaggregated approach [5] does not give better performance than the simple autoregressive model augmented by the Insee Business Climate [4]. It should be stressed that despite its robustness when using large-scale data, this approach doubtless suffers here from the very small number of observations in comparison (one hundred for 60 times more variables). However, this disaggregated approach remains interesting, in the sense that it is easier to implement, automatically calibrated, and not involving compiling lists of terms, which is both laborious and debateable.

* *
*

We have therefore shown that media information was a promising tool for economic analysis. The systematic treatment of articles published by *Le Monde* since 1990 using textual analysis techniques enabled us to measure this potential for forecasting or nowcasting French GDP. More precisely, we considered two different strategies: the first consisted in constructing a synthetic indicator, the second in using more extensively all the information available in the database. These two approaches each have their advantages and drawbacks. The first offers the possibility to construct a readable media sentiment indicator with theoretical properties similar to other more traditional economic tools (business climate). However, such an indicator takes into account only a fraction of the information contained in the database and, in addition, its construction is based on a certain number of choices and questionable bias. Conversely, a variables selection technique (penalised regression) has the advantage of using all the information from the database in an exhaustive and “agnostic” way: it is easy to implement and does not rely on any preconception. However, it provides inferior results to the approach using a predefined sentiment dictionary.

Nonetheless, this generally favourable observation should be somewhat tempered. At all time scales, the Insee Business Climate indicator appears to be a more effective tool than media information. Similarly, adding media information does not always enable a significant gain in predictive power: it therefore currently appears to play a greater role as complement than substitute. Finally, it should be recalled that economic institutes have to continue to develop their activity producing indicators: media sentiment indicators would not be able to replace them since economists and public authorities need an independent and controlled source to measure the business climate. □

BIBLIOGRAPHY

Andreou, E., Ghysels, E. & Kourtellos, A. (2013). Should Macroeconomic Forecasters Use Daily Financial Data and How? *Journal of Business & Economic Statistics*, 31(2), 240–251.
https://doi.org/10.1080/07350015.2013.767199

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L. & Rünstler, G. (2011). Short-term forecasts of euro area GDP growth. *The Econometrics Journal*, 14(1), C25–C44.
https://doi.org/10.1111/j.1368-423X.2010.00328.x

Baffigi, A., Golinelli, R., & Parigi, G. (2004). Bridge models to forecast the euro area GDP. International Journal of forecasting, 20 (3), 447–460.
https://doi.org/10.1016/S0169-2070(03)00067-0

Baker, S. R., Bloom, N. & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593–1636.
https://doi.org/10.1093/qje/qjw024

Bańbura, M., Giannone, D., Modugno, M. & Reichlin, L. (2013). Now-Casting and the Real-Time Data Flow, *Handbook of Economic Forecasting*, vol. 2 (Part A), 195–237.
https://doi.org/10.1016/B978-0-444-53683-9.00004-9

Bec, F. & Mogliani, M. (2015). Nowcasting French GDP in real-time with surveys and “blocked” regressions: Combining forecasts or pooling information? *International Journal of forecasting*, 31 (4), 1021–1042.
https://doi.org/10.1016/j.ijforecast.2014.11.006

Bortoli, C. & Combes, S. (2015). Apports de Google trends pour prévoir la conjoncture française: des pistes limitées. Insee, *Note de conjoncture*, mars 2015.
https://www.insee.fr/fr/statistiques/1408926?sommaire=1408931

Bortoli, C., Combes, S. & Renault, T. (2017). Comment prévoir l’emploi en lisant le journal. Insee, *Note de conjoncture*, mars 2015.
https://www.insee.fr/fr/statistiques/2662520?sommaire=2662600

Breitinger, C., Gipp, B. & Langer, S. (2015). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4), 305–338.
https://doi.org/10.1007/s00799-015-0156-0

Choi, H. & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, 88 (1), 2–9.
https://doi.org/10.1111/j.1475-4932.2012.00809.x

Darné, O. (2008). Using business survey in industrial and services sector to nowcast GDP growth: The French case. *Economics Bulletin,* 3(32), 1–8.
https://ideas.repec.org/a/ebl/ecbull/eb-08c50137.html

D’Amuri, F. & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801–816.
https://doi.org/10.1016/j.ijforecast.2017.03.004

Fondeur, Y. & Karamé, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, 30, 117–125.
https://doi.org/10.1016/j.econmod.2012.07.017

Foroni, C. & Marcellino, M. (2014). A comparison of mixed frequency approaches for nowcasting Euro

area macroeconomic aggregates. International Journal of Forecasting 30(3), 554–568.
https://doi.org/10.1016/j.ijforecast.2013.01.010

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3), 1267–1300.
https://doi.org/10.1111/jofi.12027

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2005). There is a risk-return trade-off after all. *Journal of Financial Economics*, 76(3), 509–548.
https://doi.org/10.1016/j.jfineco.2004.03.008

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS Regressions: Further Results and New Directions. *Econometric Reviews*, 26(1), 53-90.
http://dx.doi.org/10.2139/ssrn.885683

Harvey, D., Leybourne, S. & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2), 281–291.
https://doi.org/10.1016/S0169-2070(96)00719-4

Kotsiantis, S. B., Pintelas, P. E. & Zaharakis, I. D. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3), 159–190.
https://doi.org/10.1007/s10462-007-9052-3

Larsen, V. H. & Thorsrud, L. A. (2015). The value of news. BI Norwegian Business School, *Working Papers* N° 6/2015.
ttps://ideas.repec.org/p/bny/wpaper/0034.html

Loughran, T. & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66 (1), 35–65.
https://doi.org/10.1111/j.1540-6261.2010.01625.x

McLaren, N. & Shanbhogue, R. (2011). Using Internet search data as economic indicators. *Bank of England Quarterly Bulletin* N° 2011-Q2.
http://dx.doi.org/10.2139/ssrn.1865276

Mogliani, M., Darné, O. & Puyaud, B. (2017). The new MIBA model: Real-time nowcasting of French GDP using the Banque de France's monthly business survey. *Economic Modelling*, 64, 26–39.
https://doi.org/10.1016/j.econmod.2017.03.003

Mogliani, M. & Ferrière, T. (2016). Rationality of announcements, business cycle asymmetry, and predictability of revisions. The case of french GDP. *Banque de France, Working Papers Series* N° 600.
https://publications.banque-france.fr/en/economic-and-financial-publications-working-papers/rationality-announcements-business-cycle-asymmetry-and-predictability-revisions-case-french-gdp

Porter, M. F. (2001). Snowball: A language for stemming algorithms.
http://snowball.tartarus.org/texts/introduction.html

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3), 1139–1168.
https://doi.org/10.1111/j.1540-6261.2007.01232.x

APPENDIX 1

DESCRIPTIVE STATISTICS

Table A1

	Frequency	Average	Median	Min	Max	Standard deviation	Skewness	Kurtosis
GDP growth	Quarterly	0.3383	0.3456	-1.1967	1.2270	0.4218	2.0606	-0.7953
Media Sentiment	Monthly	-0.0105	-0.0104	-0.0228	-0.0011	0.0037	0.1955	-0.2251
Insee Business Climate	Monthly	99.47	100.35	68.43	118.71	10.13	-0.0877	-0.4747

Sources: *Le Monde* authors' database; Insee.

APPENDIX 2

COEFFICIENTS OF ECONOMETRIC MODELS

Table A2-1

	Month 1 (Q)	Month 1 (Q)	Month 1 (Q)	Month 1 (Q)
α	0.2537***	0.7207***	0.2514***	0.6228***
ΔGDP_{T-2}	0.2700***	0.1456	0.2942***	0.1935**
ΔGDP_{T-1}				
$\Delta Climate_T$			0.0605***	0.0560***
$\Delta MediaSent_T$		40.4608***		32.1605***
Adjusted R^2	0.070	0.145	0.258	0.303

Table A2-2

	Month 2 (Q)	Month 2 (Q)	Month 2 (Q)	Month 2 (Q)
α	0.2642***	0.8672***	0.2980***	0.8402***
ΔGDP_{T-2}				
ΔGDP_{T-1}	0.2430*	0.0908	0.1593	0.0283
$\Delta Climate_T$			0.0467***	0.0431***
$\Delta MediaSent_T$		51.95***		46.9353***
Adjusted R^2	0.055	0.169	0.196	0.288

Table A2-3

	Month 3 (Q)	Month 3 (Q)	Month 3 (Q)	Month 3 (Q)
α	0.2761***	1.0301***	0.3118***	0.9987***
ΔGDP_{T-2}				
ΔGDP_{T-1}	0.2139*	0.0036	0.1190	- 0.0645
$\Delta Climate_T$			0.0423***	0.0384***
$\Delta MediaSent_T$		64.4305***		58.9808***
Adjusted R^2	0.037	0.206	0.190	0.331

Note: The table shows the results from the equation $\Delta GDP_T = \alpha + \beta_1 * \Delta GDP_{T-1} + \beta_2 * \Delta Climat_T + \beta_3 * MediaSent_T + \varepsilon_t$ (ΔGDP_{T-2} at month 1, as the GDP for the next quarter has not been published yet) over the whole sample (1990-Q1 to 2017-Q4). ***, **, * indicate significance of the coefficients at 1%, 5% and 10%, respectively. The standard deviations are robust to heteroscedasticity.
Sources: *Le Monde* authors' database; Insee; authors' calculation.

Use of Google Trends Data in Banque de France Monthly Retail Trade Surveys

François Robin*

Abstract – Under its partnership with the Banque de France, the Federation of E-Commerce and Distance Selling (*Fédération du e-commerce et de la vente à distance* - FEVAD) has provided monthly consumer online retail sales data since 2012. Pending the release of new data, the Banque de France carries out estimations, a task complicated by the growth of online retail. The autoregressive model (SARIMA(12)) used up to now can now be complemented by other statistical models that draw on exogenous data with a longer historical time series. This paper details the system of choices that results in the final forecast: data conversion, variable selection methods and forecasting approaches. In particular, Google queries, as measured by Google Trends, help enhance the predictive accuracy of the final model, obtained by combining single models.

JEL Classification: C51, C53, C11, E17
Keywords: Google Trends, nowcasting, trends, e-commerce, distance selling, Big Data, Bayesian averaging, variable selection, lasso

* *Banque de France, Economic Surveys Department, Directorate General Statistics (francois.robin@banque-france.fr)*

The author would like to thank in particular those who took part in an internal research workshop at the Banque de France, Patrick Kwok for his regular follow-up, contributions and support for this work. A big thank you also to Martial Ranvier and Valérie Chauvin for their helpful comments and valuable methodological input. Lastly, thank you to François Guinouard, Léon Ipdjian and Mathilde Gerardin; and to François Brunet, who proposed the topic.

Translated from the original version: "Utilisation de Google Trends dans les enquêtes mensuelles sur le Commerce de Détail de la Banque de France"

To cite this article: Robin, F. (2018). Use of Google Trends Data in Banque de France Monthly Retail Trade Surveys. *Economie et Statistique / Economics and Statistics*, 505-506, 35–63. https://doi.org/10.24187/ecostat.2018.505d.1965

Under its partnership with the Banque de France, the Federation of E-Commerce and Distance Selling (Fédération du e-commerce et de la vente à distance – FEVAD) provides monthly B2C (Business-to-Consumer) online retail sales data.[1] However, data releases are too late to be included in the first publication of the monthly retail trends survey, which instead uses estimates of the data.

Until now, the paucity of historical time series data limited the range of possible forecasting methods. The autoregressive model used until now may now be complemented by using exogenous data available at the time of carrying out estimations: quantitative indices for traditional retail trends (from the monthly retail trends survey) and Google Trends data. Estimation of FEVAD data for month *M* takes place during the survey period (at the beginning of month *M* + 1). Quantitiative monthly survey indices *M* are at this time under construction, while monthly Google Trends data for month *M* are final. These estimations fall fully within the scope of a nowcasting exercise.

This development runs into two issues. First, Google Trends provides a range of explanatory variables, from which those most suitable must be selected. The "adaptive lasso" machine learning-based approach developed by Zou (2006) addresses the twin challenges posed by a lack of historical FEVAD time series data (dating back to 2012 only) and the huge range of possible Google queries. Second, given the number of available models with exogenous variables from different sources, it is helpful to confirm whether the combination of models can produce better output. This topic has been the subject of much debate, as outlined in Bec & Mogliani (2015).

Following a review of the relevant literature, the second part looks at the datasets in greater detail, including an overview of the retail monthly survey data, FEVAD data and Google Trends data. Due to the particular nature and lack of clarity around the methodology of construction of Google Trends data, robustness checks and automated error corrections linked to breaks in series are required. The third part addresses the choice of models, looking at how stationarity is managed in time series, then at the stages of the model testing process. The fourth part examines the results and how they are interpreted. The final part concludes.

Literature Review

Google Trends Data

Available almost in real time, Google Trends indices show the evolution in queries made by users of the Google search engine over time. These indices represent an information flow and a source of big data. While there is no record of their use by public sector institutions in recurring studies, they have been the subject of a number of publications. Research by Ettredge *et al.* (2005) and Askistas & Zimmerman (2009), which look at the unemployment rate forecasting using keywords used in Google searches, indicate the potential benefit of such indices. Choi & Varian (2009, 2011) are more cautious about input from Google Trends. However, their literature review contains a number of papers using Google searches, mainly in the field of epidemiology; the tool used at the time and developed by Google (Google Flu) was discontinued on 20 August 2015, in light of shortcomings already highlighted by Bortoli & Combes (2015).

These tools are operated entirely by Google: the construction methodology is vague, which presents further risks for users. Methodological changes to Google Trends may involve breaks in series. Furthermore, the introduction of new actors has an impact on how user queries are formulated. McLaren & Shanbhogue (2011) warn about the mechanical fall in popularity of certain requests in their application to unemployment (e.g. in France, when the ANPE became "*Pôle emploi*" after the restructuring of the employment agency and unemployment insurance, then the ANPE and Assedic, respectively).

Selection of Variables

Machine learning methods offer a solution to variable selection, in particular the adaptive lasso approach developed by Zou (2006). As a reminder, the standard lasso[2] function introduced by Tibshirani (1996) is:

$$\widehat{\beta}_{lasso} = \operatorname{argmin}_{\beta} \left\| Y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|, \lambda \geq 0$$

[1] *According to the FEVAD, online retail as a share of total retail (excluding food and drink, in line with the scope of the Banque de France retail trends survey) was 7% in 2013, 8% in 2014 and 9% in 2015.*

[2] *Lasso stands for* Least Absolute Shrinkage and Selection Operator.

In a lasso regression, the same penalty function λ is applied to all variables. Zou (2006) proposes adjusting the penalty based on the variables in the adaptive lasso (*adalasso*) :

$$\widehat{\beta}_{adalasso}=\text{argmin}_{\beta}\left\| Y-\sum_{j=1}^{p}x_j\beta_j\right\|^2+\lambda\sum_{j=1}^{p}w_j\left|\beta_j\right|,\begin{cases}\lambda\geq 0\\ w_j\geq 0\end{cases}$$

The adaptive lasso is a weighted lasso. Its Oracle properties, as demonstrated by Zou (2006), offer the adaptive lasso two advantages over the standard lasso. The first is the consistency in its variable selection, i.e. the best sub-set of variables (from the initial set) is chosen; which is not always the case with a standard lasso (see Zou, 2006). The other Oracle property is the consistency of parameter estimation (asymptotic convergence of the estimator in normal distribution).

While Zou (2006) defines individual penalties as $\widehat{w}=1/\left|\widehat{\beta}\right|^{\gamma}$, with $\widehat{\beta}$ the ordinary least squares estimator and $\gamma>0$ (in practice, $\gamma\in\{0.5\,;1\,;2\}$), an alternative approach involves using the estimator from the ridge regression,[3] introduced by Hoerl & Kennard (1970), to define the vector of individual penalties. Its use helps prevent errors in estimation of penalties due to multicolinearity in the regressors.

The adaptive lasso is optimised in two stages. First, the individual penalties are obtained from a ridge regression:

$$\widehat{\beta}_{ridge}=\text{argmin}_{\beta}\left\| Y-\sum_{j=1}^{p}x_j\beta_j\right\|^2+\kappa\sum_{j=1}^{p}\left\|\beta_j\right\|^2,\ \kappa\geq 0$$

The penalty value κ is then obtained by leave-one-out cross-validation (Hyndman & Athanasopoulos, 2018).[4] Then, $\widehat{w}=\widehat{\beta}_{ridge}$ results in the lasso function (for which the penalty λ is also optimised by leave-one-out cross-validation):

$$\widehat{\beta}_{adalasso}=\text{argmin}_{\beta}\left\| Y-\sum_{j=1}^{p}x_j\beta_j\right\|^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|/\left|\widehat{w}_j\right|,\lambda\geq 0$$

The advantage of the adaptive lasso is its large-scale operation (greater number of variables than observations, i.e. to the size of the temporal window in this case). It is also considered parsimonious. Both of these properties address the twin challenges posed by the large number of possible Google queries and the short historical time series for FEVAD releases.

Combination of Models or Global Model?

Three individual models have been used: the Google Trends model, the retail model from the retail trends survey, and the SARIMA model that has been used up to now.[5]

Bec & Mogliani (2015) document the most common methods of combining data. In their view, Bates & Granger (1969) were the first to support the aggregation of forecasts from different models. Subsequently, Diebold (1989) recommends the use of a single model, combining multiple heterogeneous data sources. More recently, Huang & Lee (2010) argue that a global model with sound specifications is preferable. Moreover, Clements & Galvão (2008) and Kuzin *et al.* (2013) argue in favour of aggregation for empirical applications. Bec and Mogliani (2015) find that aggregation performs better when forecasting movements in consumption indices. The test designed by Diebold & Mariano (1995), whose null hypothesis is that two forecasts generated by different models are not significantly different, is a critical indicator when opting for a model.

This paper seeks to contribute to the debate around a new application by comparing output from a combination of models with that from a global model with the same specifications as the individual models, in this case the adaptive lasso applied to all regressors simultaneously (Google Trends, retail indices and SARIMA). De Gooijer & Hyndman (2006) highlight the benefits of aggregation, in particular comprehensibility where aggregated models can be easily interpreted. Here, aggregation applies to three individual models, each with their own effects:

- The SARIMA model reproduces the past time series pattern;
- The retail model exploits traditional retail data;
- The data of interest is extracted from the model based on Google Trends indices.

The issue is weighting each forecast:

$$\widehat{Y}_{t+1}=\gamma\widehat{Y}_{t+1}^{SARIMA}+\mu\widehat{Y}_{t+1}^{gTrends}+\vartheta\widehat{Y}_{t+1}^{CD}$$

There are a number of possible approaches to aggregation: from the most straightforward, such

[3] Ridge and lasso regressions are penalised by L2 and L1 norms respectively.

[4] Specifically, the validation sample is made up of one observation; the training sample is made up of the $n-1$ other observations (for sample size n). The n values for κ, obtained for each training sample (each minimising the RMSE) give a mean to obtain the final value of κ.

[5] The retail model is an adaptive lasso function for which explanatory variables are quantitative retail indices.

as weighting by the mean ($\gamma=\mu=\vartheta=1/3$) or by the inverse of errors – in-sample or out-of-sample (see Aiofli & Timmerman, 2006), to the most elaborate. For example, Bayesian inference, based on Bayes' theorem[6] (see Marin & Robert, 2010), determines the probability of an event from prior measured events. Bayesian statistics, commonly used for small sample sizes, produces methods of classification, or aggregation in this case. Hoeting *et al.* (1999) highlight the effectiveness of Bayesian aggregation. Zeugner (2011) developed an R package on this subject. The purpose is to test models of a given category M and weight them according to their probability of being the correct model. The category M is that for linear models. Usually, the large number of models complicates Bayesian aggregation (see Hoeting *et al.*, 1999). This is not the case here: with three regressors (for the Google Trends, retail trends and SARIMA model estimates), eight linear models are possible. By denoting data as D and a given model as $M_j\left(1\le j\le 8\right)$ un modèle donné, le théorème de Bayes donne Bayes theorem gives:

$$P\left(M_j|D\right)=\frac{P\left(D|M_j\right)P\left(M_j\right)}{\sum_{1\le i\le 8}P\left(D|M_i\right)P\left(M_i\right)}$$

Both terms of the numerator used to measure *a posteriori*[7] probability are as follows:

- $P\left(M_j\right)$ corresponds to the *a priori*[8] probability that model M_j is the correct one;

- $P\left(D|M_j\right)=\int pr\left(D|\beta_j,M_j\right)pr\left(\beta_j|M_j\right)d\beta_j$ with β_j the parameters of the model: $\beta_j=\left\{\gamma_j,\mu_j,\vartheta_j\right\}$ estimé sur le modèle M_j. Here, we are interested in the values for the parameters.

Specifically, the values of coefficients obtained in each model M_j for category M are weighted by the probability that each model M_j is the correct one: $\gamma=\sum_i\gamma_i P(M_i\,|\,D)$ with $\gamma_i=E(\gamma\,|\,D,M_i)$ the value of the coefficient in model M_i. The same applies to μ and ϑ.

Data

The Monthly Retail Trends Survey

One of the monthly trend surveys undertaken by the Banque de France covers the retail sector.[9] The survey tracks changes in sales including tax for a sample population of 6,800 (divided among more than 4,000 businesses); each month, the response rate is approximately 90%. Each entity provides its total sales figure and the respective shares of its main products (where it is not a "single-good" retailer). Individual data are then grouped according to characteristics common to retail businesses: by method of distribution (physical: small traditional retailer, large specialist and chain retailer, hypermarkets and supermarkets, department store and variety store and mail-order store: distance selling) and by product (e.g. household appliances, shoes, etc.). Quantitative indices are established for these groupings (e.g. small traditional furniture retailers).

Construction of Quantitative Survey Indices

Each sales index Y from the survey is constructed as follows (with X the relevant sales figure):

$$Y_M=Y_{M-12}\frac{X_M}{X_{M-12}}$$

All quantities for the above equation apply to the same companies. Under the survey methodology, sales samples are "balanced", i.e. the scope of X_M is the same as for X_{M-12}. In other words, the same companies are measured in the case of both sales values. This approach prevents extreme variations not representative of the sample (outliers). The closure (or opening) of a store is the most common extreme event that is problematic for the survey: the resultant reduction (respectively increase) in sales is offset by opposite movements for all of its competitors, which will not be fully captured by the sample. Furthermore, it is easier to track a store closure than a store opening (store or new brand not yet included in the sample), which would present a measurement bias in the case of unbalanced data.

Availability of Indices

Only a portion of indices cross-referencing products and physical distribution systems (in-store sales) are measured. This is due to the lack of an adequate sample size and for data

[6] *The theorem is commonly formulated as:* $P(A\,|\,B)=\frac{P(B\,|\,A)P(A)}{P(B\,|\,A)P(A)+P(B\,|\,\bar{A})P(\bar{A})}$ *with P measuring probability,* A *and* B *two events.*

[7] *The* a posteriori *probability is obtained using in-sample data.*

[8] *There are a number of ways to obtain* a priori *probabilities, as shown in Zeugner (2011). In our case, a number of tests were carried out (e.g. prior binomial, uniformity, deterministic, etc.) with no significant effect on output.*

[9] *The latest survey results are available at (French only): https://www.banque-france.fr/statistiques/chiffres-cles-france-et-etranger/enquetes-de-conjoncture/conjoncture-commerce-de-detail.*

Table 1
Mean and Standard Deviation of Quantitative Survey Indices for Physical Retail Sales (Retail Trends Survey)

	Small traditional retailers	Large specialist retailers and chains	Hypermarkets and supermarkets	Department stores and variety stores	All physical sales
Total industrial products excl. cars					91.2 17.5
Shoes	91.2 21.1	94.3 27.0			
Consumer electronics	81.5 30.3	98.1 46.3	82.5 33.1		
Household appliances	96.1 15.6	103.6 15.4	97.6 23.9		
Furniture	105.1 18.4	108.4 20.5	126.1 38.7		
Clothing	102.1 28.4	101.6 28.8	99.0 19.1	87.4 24.3	

Reading Note: An empty cell denotes the absence of the indicator for the cross-reference in question. The mean values and standard deviations calculated over the period January 2012 - December 2017 are indicated on the first and second line.
Sources: Banque de France DGS SEEC.

protection reasons (see empty cells in Table 1). These indices have been available since 1990, whereas those for remote selling have been available since 2012. This paper examines raw indices (see the section on Models). Table 1 shows the quantitative survey indices for physical sales of products covered by the FEVAD data releases.[10]

FEVAD

Online retail data are not collected directly. The Federation of E-Commerce and Distance Selling (*Fédération du e-commerce et de la vente à distance* – FEVAD) has been providing the Banque de France with monthly aggregated sales data for its largest members since January 2012. There are currently around 70 members, changing over time. For the survey, these data are used to construct sales indices (defined above) applied to remote selling. In line with the survey methodology, sales data for month M and the revised panel figure for month $M-12$ are released every month. These releases concern total sales ("total industrial goods excluding cars") and those for five products: household appliances, textiles (clothing and household textiles – hereafter referred to as clothing), shoes (including leather goods), consumer electronics and furniture (furniture only). As the total covers more than the five products combined, its sales figure is higher than that for all five product sales combined. On average (for the time series history), sales for the five products account for 68% of total sales. Furthermore, Table 2 gives the share (in %) of remote selling for each product as captured by FEVAD.

Approximation of FEVAD Data with Retail Survey Data

Data taken from the survey are used for each of the six estimations (indices for total sales by remote selling and for the five products by remote selling). Figure I presents the indices for each distribution channel for consumer electronics (physical and remote sales).

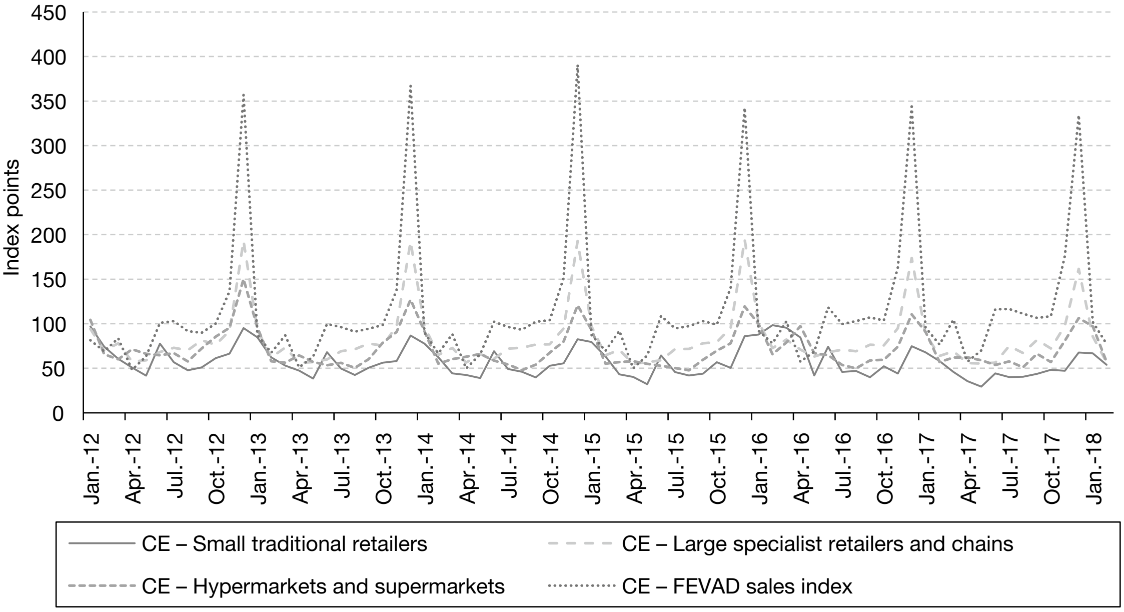
For consumer electronics, the December sales peak obtains for all distribution channels. The correlations[11] between FEVAD sales data and physical retail sales indices for consumer electronics (as a %) complement the information on the graph (Table 3).

The correlation between the remote sales index and that for large specialist and chain retailers prompts us to use physical sales data to estimate FEVAD data.

Generally, approximating these data allows us to observe straightforward economic processes. For example, over the long term, a substitution effect can be observed through

[10] *NB: products other than shoes, consumer electronics, household appliances, furniture and clothing make up the total.*
[11] *Correlation is measured for the differentiated indices on a monthly basis, in line with data used for modelling (see below).*

Figure I
Raw Indices for the Various Consumer Electronics (CE) Distribution Channels



Sources: FEVAD, Banque de France DGS SEEC.

Table 2
Share of Remote Sales for Each Product

Product	Remote sales weighting (in %)
Shoes	11
Consumer electronics	23
Household appliances	18
Furniture	13
Clothing	13
Total	10

Reading Note: According to the FEVAD, remote sales represent 11% of shoe sales in 2017.
Source: FEVAD.

Table 3
FEVAD Sales Index Correlations with Traditional Retail Indices from the Consumer Electronics (CE) Survey

Distribution channels	Correlation with remote sales index (in %)
CE – Small traditional retailers	44
CE - Large specialist retailers and chains	96
CE – Hypermarkets and supermarkets	48

Notes: Correlations are calculated for the period 01/2012 - 01/2018.
Sources: Banque de France DGS SEEC, FEVAD.

a reduction in sales at physical retail outlets; the corollary is an increase in remote sales. On the other hand, in the short term, an increase (or decrease) in physical sales may predict an increase (decrease, respectively) in remote sales: such collective movements reflect an increase in household consumption.

Google Trends

Google Trends provide monthly indices for terms queried via the Google search engine by users. Developed by Google using a methodology that has not been made public, indices are created by user-defined fields based on geography (in this case, France), time period (series date back no further than 2004), frequency (in this case, monthly) and belonging to a category (e.g. “Shopping”, see below). Available where search volumes are “sufficient” (as defined by Google), these indices are made up of whole values between 0 and 100 and are produced for samples of all completed searches. Aside from the vagueness of Google Trends’ index construction methodology, some earlier points raised call for robustness tests to be carried out.

Google Sampling

Constructed from a random sample of searches, a Google Trends index will differ between two samples. Comparing the series for the same term, queried multiple times, helps to verify the robustness of the tool. To illustrate this, Table 4 provides the correlations obtained for two separate samples taken a few days apart (i.e. with constant Google and Google Trends methodologies, *a priori*).

This was repeated several times, without obtaining a rate of correlation below 90% for differentiated monthly indices. Under these conditions, the sampling method appeared sufficiently reliable to periodically query Google Trends indices. The impact of sampling on the output will be discussed in the relevant section.

Whole-Value Indices

The simultaneous extraction of Google Trends indices is subsequently problematic. When making a common extraction of indices (between two and five) using the tool, the value 100 is attributed to the index experiencing a peak in searches for the period under query; the maximum values for other indices are given as a proportion. Where search volumes differ significantly, indices for less popular queries take on a limited number of values – as they are made up of whole values – which do not fully capture their fluctuations. However, in a statistical model, the number of decimal points for variables and, more generally their precision, can have an influence on the final estimation, according to Kozicki & Hoffman (2004). In order to obtain the most precise values, each Google Trends series is extracted individually. Taken together, the latter two points – sampling and the fact that indices consist of whole values – do not facilitate precision in Google Trends data.

Category

The Google Trends tool lists Google queries by category, corresponding to the context in which the search is made.[12] The example of the "iPhone" query urges caution when extracting data (Figure II).

Table 4
Correlations Between Google Trends Indices Taken Several Days Apart

(In %)

Amazon	Cdiscount	Fnac	E. Leclerc	eBay
98.1	97.4	98.9	95.5	90.2

Notes: Correlations calculated for the differentiated indices on a monthly basis, from January 2004 to February 2018 (170 points).
Sources: Google Trends, Banque de France DGS DESS SEEC.

While the "Commercial and industrial markets" category is not useful for analysis of remote sales, the line chart underlines the importance of category selection: its maximum level, reached in September 2013, does not equate to an explosion in sales. In the absence of more information about the categories, all subsequent queries referred to in the paper belong to the "Shopping" category, most closely corresponding *a priori* to online retail.

Breaks in Series

While Google does not share a great deal of information regarding changes to its methodology in constructing indices, the extraction page includes two observations:

- "The feature for determining geographic position has been updated. This update was applied as of 1 January 2011."

- "Our system for collecting data has been updated. This update was applied as of 1 January 2016."

Users are therefore notified of major changes to the tool. In addition, these are in effect several months later. As FEVAD sales indices begin from January 2012, the second observation requires particular attention.[13]

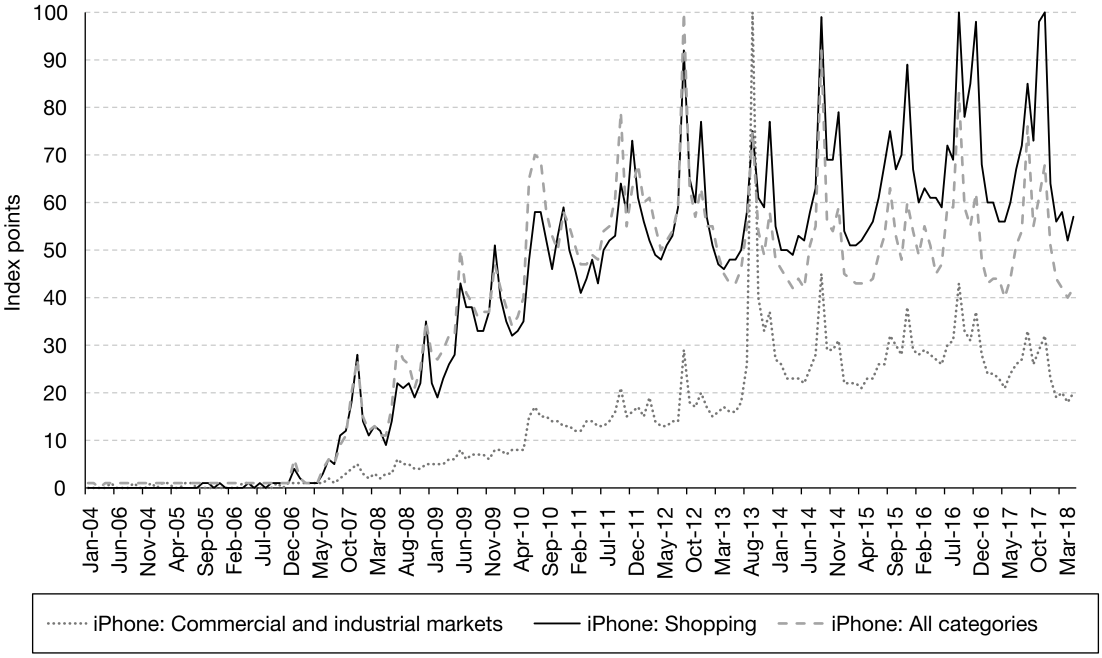
Analysis of Google Trends using the X-13 method detects a greater number of outliers, in particular for January 2016. Due to the vagueness of the methodology for constructing Google Trend indices and their substantial number (more than 150) – likely to increase further with the growth of online retail – outliers are now treated systematically. Using the Google Trends index for Amazon as an example, the various steps can be explicitly set out. Here, a level shift is detected in January 2016; following evaluation, the series can be corrected (Figure III).

The first step in treatment is seasonal adjustment of the index, because two indices are used in detection (raw and seasonally-adjusted)

[12] *For example, "jaguar" may refer to the animal or the car manufacturer. Google queries are most likely listed in categories based on post-query browsing activity (i.e. websites visited after the query).*

[13] *In order to improve the robustness of calculations, Google Trends indices have been extracted since January 2011. Although it is generally accepted that seasonal adjustment is not possible for historical time series of less than 3 years, adding one year of series data helps stabilise seasonally adjusted time series and, in so doing, improve outlier detection.*

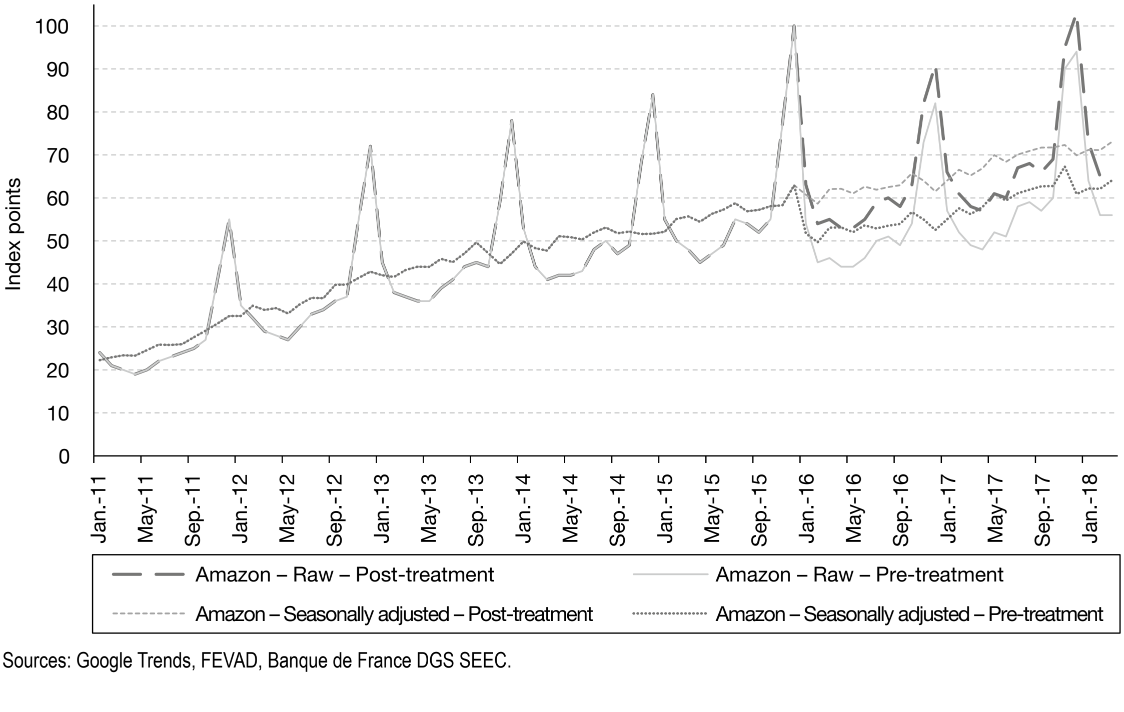
Figure II
Google Trends “iPhone” Queries



Sources: Google Trends.

in order to identify the maximum number of outliers. The nature of the outlier is then identified as a level shift, transitory change or additive outlier. In the case of Amazon – and outliers detected in January 2016 more generally (see Cdiscount, Appendix 1) – this is a level shift. Lastly, the extent of the break in series is estimated by the deviation between the January 2016 value on the seasonally adjusted series and the same truncated series forecast

Figure III
Treatment of the Outlier Detected on the Google Trends Amazon Index



Sources: Google Trends, FEVAD, Banque de France DGS SEEC.

in December 2015.[14] This adjustment is then applied to the rest of the series (unlike single outliers, which are treated *ad hoc*).

With the improving quality of series in quasi-real time, detecting outliers is less reliable in real time, i.e. for the latest series value: not adjusting too soon for the outlier enables more accurate classification,[15] thereby improving the precision of its estimation. The only outliers not treated are those reflecting the emergence of new queries (new company, new brand, etc., see shoes example below). The ever-changing nature of online retail also requires caution to be exercised when selecting queries.

Lists of Variables

On the one hand, the emergence of online retail has introduced new actors. In the case of shoes, for example, the three "pure players" (online-only retailers) dominating the French online retail market are relative newcomers (Figure IV).

The movements in the "Shoes" index between 2004 and 2011 point to an emergence of online retail for shoes. The launch of Zalando in France in December 2010 is very clear on the graph (the index increases from 1 to 19 inside in the two months 11/2010 - 01/2011).

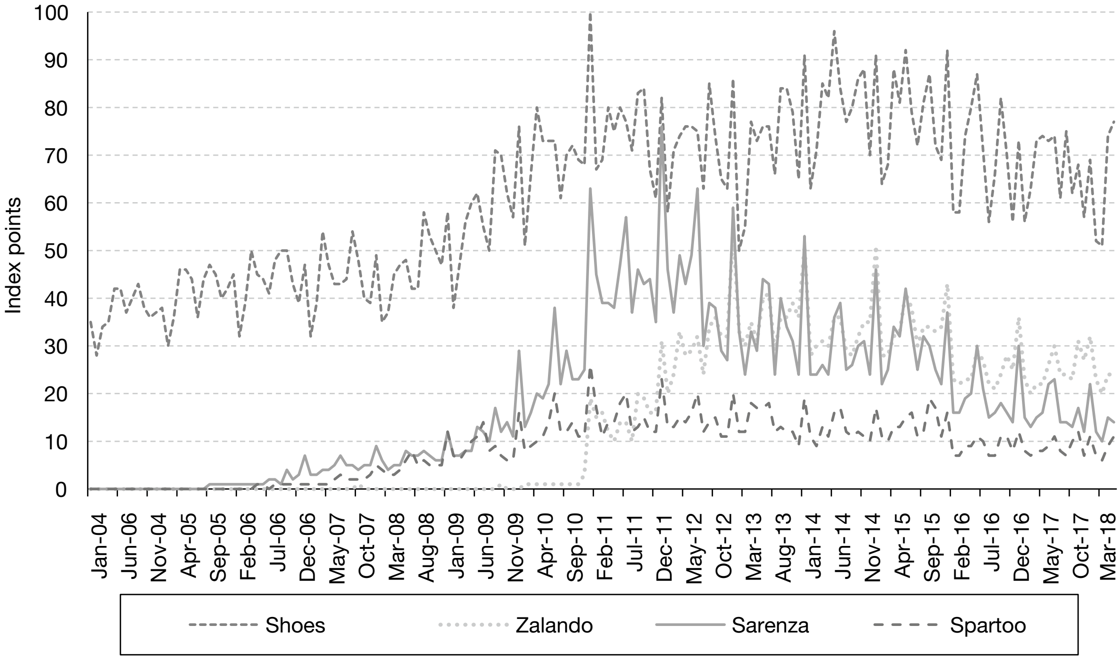
On the other hand, some erstwhile highly visible online retailers experience decline. In terms of household appliances, the Google Trends index for GrosBill serves as proof of this (Figure V).

While the query demonstrated a level of interest some years ago, this online electrical goods and consumer electronics retailer has lost market share by comparison with Boulanger, for example. Another example of the ever-changing face of online retail is the merger of Fnac and Darty: the related Google Trends index is now "Groupe Fnac Darty". In general, online retail has been in constant evolution, which Google Trends has managed to relay. For example, the fall in popularity in Google queries for one online retailer can be accompanied by an increase in queries for rival firms. In this space, it is essential to frequently review the variables used, particularly those applicable to online retailers for the various products. In order not to ignore the changing nature of search terms, it is possible to backward-extrapolate results with other variables through double collection (i.e. by testing the model on two sets of variables).

[14] *In this example, the level shift is estimated by adjusting for seasonal variations as the estimate provided by raw data appeared less consistent.*

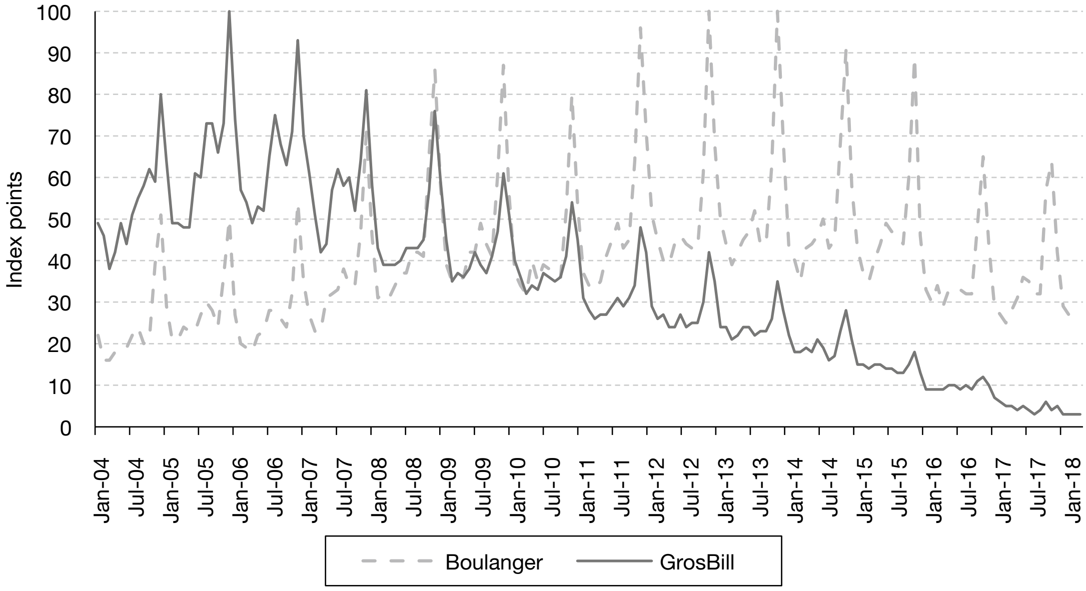
[15] *For example, a level shift can only be detected* a posteriori*: when it appears, the outlier may be characterised (at best) as a single outlier before being reclassified as a level shift (following the appearance of further observations).*

Figure IV
Google Trends Indices for "Shoes"



Sources: Google Trends.

Figure V
Google Trends Indices for "Boulanger" and "GrosBill"



Sources: Google Trends.

However, one of the limits of this approach lies in the initial lists of variables (see Appendix 2). For the overall index, such pre-selection corresponds for the most part to the major online retailers in France. Pre-selections for the five products are a mix of pure players (e.g. Sarenza, in the case of shoes), major retailers (La Halle), generic terms (ladies' shoes) and brands (Converse). Preliminary research, such as a document search for the products in question by appraising or visiting specialist websites, was carried out in order to predict behaviour prior to making a purchase. This resulted in the compilation of lists of heterogenous variables (see full table in Appendix 2).

Moreover, the trend in a website's popularity is not necessarily the same as that for the Google Trends index, as not all internet users visit Google: internet browsing patterns change, in particular with the emergence of m-commerce,[16] where applications eliminate the need to use a search engine.

Models

Treatment of Stationarity and Seasonality

Most series are not stationary but are instead integrated to the order one: differentiation is required. This standard operation helps prevent spurious regressions (see Phillips, 1986), a common occurrence in time series regressions, which produce overly optimistic output reflected in an unusually high R^2 (see Granger & Newbold, 1974). Introducing a variable measuring the trend (Phillips & Perron, 1988) or autoregressive terms, also play a role.

In order to better measure trends in online retail, reference has been made to working with seasonally adjusted series. This solution has not been adopted. First, the short time series do not facilitate meaningful seasonal adjustment across all series,[17] chiefly for the initial estimations (36 points in the first iteration; more than 70 at present); particularly given that online retail itself has seen shifts in seasonality (Figure VI).

Figure VI represents the raw series for two sales indices for clothing (remote sales and small traditional retailers) and two product-related Google Trends queries: Kiabi and Zara. The remote sales index has seen changes in seasonality; for example in the early years, July figures were far higher than those for August. In 2015, the gap between both months narrowed and in 2016, the figures for July were lower. An overview of the series aptly demonstrates changes in seasonality. This phenomenon is

[16] According to FEVAD, 36.6 million people in France shop online, of whom 9.3 million have made a purchase using their mobile phone in the past (2017).
[17] More than 150 series are used to complete six estimations.

common to Google Trends indices. For example, in the case of Zara, the annual maximum is reached in January in the years 2013 to 2016; however, the value for November 2017 exceeds that for both January 2017 and January 2018. In addition, the Kiabi series does not exhibit any noticeable seasonality. Under such conditions, seasonal adjustment of multiple series becomes problematic. On the other hand, seasonal trends for the small traditional retailer index remain stable. Changes are more rare for well-established series (the index dates back to 1990). More generally, survey time series systematically pass seasonality tests (autocorrelation, Friedman, Kruskall-Wallis, spectral peaks, periodogram), which is not always the case for Google Trends series.

In addition, the latest values of a seasonally adjusted series are more likely to be revised in light of subsequent data releases (see Eurostat, 2018). At each FEVAD release, when a prior forecast can be evaluated, the most recent values of the seasonally adjusted series change, which can have a substantial impact on the model. The instability of seasonal adjustment on the most recent values is particularly pronounced for online retail series, notably due to poorly established seasonal trends and short historical time series. While the extent of instability from seasonal adjustments is on average 0.2 points between 01/2015 and 01/2018 for the large retailer index, it averages 1.6 for total remote sales (see Appendix 3) – the same order of magnitude as forecasting errors (see below). These arguments tend to favour a differentiated raw data model.

Process of Performance Estimation and Evaluation

Models

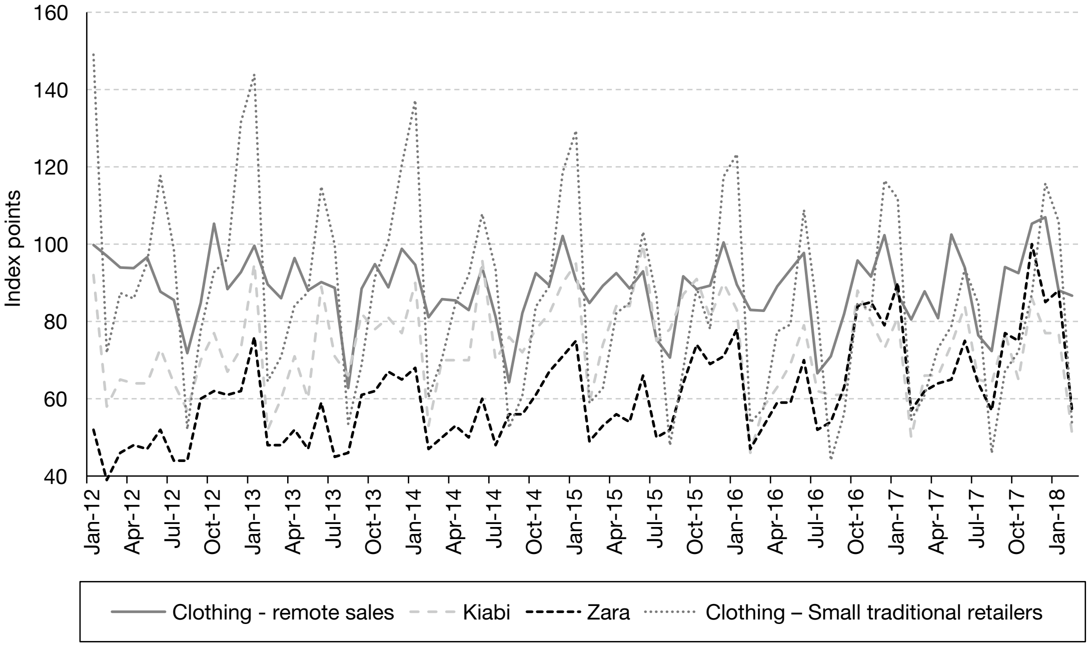
Until now, a SARIMA model has been used for each product. It is always updated and serves as a subsequent reference. Furthermore, the adaptive lasso is used in three models, implemented for each product:

- The Google Trends model, using Google Trends (see Appendix 4);

- The retail model, based on quantitative survey data for physical sales from retail surveys[18] (see Table 1);

[18] To recap, for the overall index, sales indices for the five products are also used.

Figure VI
Indices for Clothing



Sources: Google Trends, FEVAD, Banque de France DGS SEEC.

- Global model, which is a selection of all available variables.[19]

In addition to exogenous variables, a trend and an autoregressive component are also included in the set of initial variables for these three models. The incorporation of the trend variable addresses the (*a priori* non-linear) rapid growth of online retail. As well as being an autoregressive component, it is the SARIMA model for the index, which becomes a variable potentially selected by the adaptive lasso algorithm, in the same way as the trend and exogenous variables (Google Trends and/or quantitative survey indices). Lastly, the fifth model is a Bayesian aggregation ("model combination") of SARIMA, Google Trends and retail trends models. The comparison of its output with that of the global model contribute to the debate regarding data combination.

Test Protocol

For each iteration of the test protocol, i.e. each month, the actual conditions are replicated. Specifically, the values of Google Trends data and quantitative survey data for physical sales for month M are available, which is not the case for FEVAD data.

Estimation takes place in two stages: the first involves modelling the index using an autoregressive process (SARIMA). As well as obtaining its own forecast, this action also helps determine the variable used in adaptive lasso models. In the second stage, the three variable selection models (Google Trends, retail trends and global) are formulated. The model combination can only be constructed after the SARIMA, Google Trends and retail trends models.

The model quality can be determined following the release of FEVAD data as the evaluation criterion adopted for the nowcasting process, is predictive capacity. The predictor is therefore the RMSFE (Root Mean Squared Forward Error), the standard deviation of forecasting errors, measuring the out-of-sample error. The RMSE (Root Mean Squared Error), measuring the in-sample error, is also provided as it can be used to ascertain the weighting in the model combination and identify any overfitting.

Furthermore, each month the estimation window for the models expands by one observation. Due to the limited sample sizes available, working with an expandable window rather than a rolling window for the sample estimation contributes to the models' stability. FEVAD data releases began in January 2012. Differentiation of data brings the series to February 2012. With a minimum time series history of three years required to ensure the robustness of the estimation, the initial forecast is that for February 2015.

Results

Only the results for the total will be set out in detail; those applicable to products will be summarised.

Total

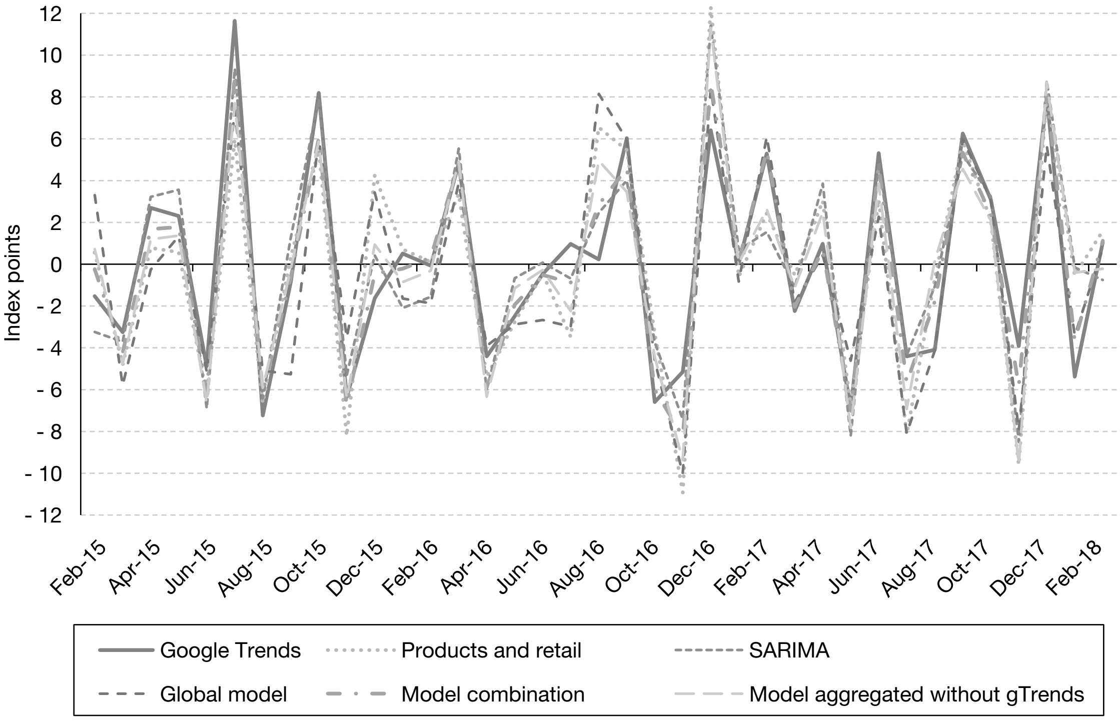
In line with the aim of the study, forecasting errors (out-of-sample) represent an important result (Figure VII).

Figure VII shows the forecasting errors for each model. The results are visually close: overall, the output from the Diebold-Mariano test (see Diebold & Mariano, 1995) does not conclude that the model forecasts are significantly different. The RMSFEs and average forecasting errors (in absolute terms) offer better insight into the forecast output, while the RMSEs attest to the responsiveness to in-sample data (Table 5).

For the purpose of the RMSFE, which remains the preferred indicator, the Google Trends model performs best with the model combination (4.8), for the Google dataset selected (deemed representative of simulations carried out, see Box). In this case, the poorer performance of the model combination without Google data justifies the use of Google Trends. The model combination also performs best in terms of the average of absolute errors. This error measurement is relevant, as one of the purposes of aggregation is also to minimise large forecasting errors. The result from individual models is relatively close. In terms of the RMSE, both models with all available information (the model combination and the global model) are a better fit for the sample data.

[19] *This model may, for some iterations, be identical to one of the individual models (e.g. if no Google Trends query is selected).*

Figure VII
Forecasting Errors of Models in Estimation of Total Index



Reading Note: The forecasting error for August 2016 by the global model is 8.1 index points: following the release of FEVAD data, the observed value of the sales index for the total was higher (by 8 points) to the global model forecast.
Sources: Google Trends, FEVAD, Banque de France DGS SEEC.

Before comparing both models, it is worthwhile to detail the output from individual models in detail – in particular the Google Trends model, for which parsimony and stability must be ensured.

SARIMA Model

Serving as a reference in this paper and in the literature, the SARIMA model offers sound forecasting performance (RMSFE=5.0), despite a less impressive fit with sample data (RMSE=4.2). However, it emerges as less suitable than the other models. For example, in December 2016, exogenous data provide real information.

Google Trends Model

In line with test protocol, variable for the adaptive lasso are selected for each iteration. Model coefficients therefore change over time (Figure VIII). For greater clarity, 30 variables for estimation of total sales have been included in six graphs. In the secondary axis for each one, the change in the lasso penalty.

Table 5
RMSFE and Mean RMSE for Models in Estimation of the Total Index

Total	Google Trends	Retail	SARIMA	Global model	Model combination	Model combination without gTrends
RMSFE	4.8	5.2	5.0	5.5	4.8	5.0
Mean average forecasting error	3.9	4.0	3.9	4.5	3.8	3.9
Mean RMSE	3.3	3.7	4.2	2.3	2.6	2.8

Notes: The model combination without Google Trends corresponds to the aggregation of the retail and SARIMA models. It allows us to determine the input of Google Trends data. However, as the SARIMA variable is present in all retail models, the aggregation becomes less significant; it will therefore not be presented in results obtained for the products.
Sources: Google Trends, Banque de France DGS SEEC.

Box – **Sensitivity of Models to Google Sampling**

Google Trends variables may be modified from one month to the next, due to Google's sampling method. The standard deviations of RMSFEs, obtained from thirty simulations[a], of modules using Google Trends variables:

As can be seen in Table A, the impact of Google sampling is significant.

Furthermore, each simulation corresponds to the linked estimation of indices for the five products and for the total. However, some Google Trends variables are common to the total and one product. It was therefore not possible to extract a set of Google Trends variables for which the model results, in terms of RMSFE, are all at the median. Those set out in the main text of the paper correspond to one of the most representative simulations (for the six estimations), i.e. the RMSFE for the Google Trends models are very close to the median.

[a] Each simulation applies to the total and the individual products across 150 Google Trends series. As Google restricts bulk extraction of series, it is difficult to substantially increase the number of simulations.

Table A
Impact of Google Sampling on Output in Terms of Standard Deviation of RMSFE

	Google Trends	Global model	Model combination
Total	0.4	0.3	0.4

Reading Note: The standard deviation of RMSFEs obtained for the 30 simulations for the Google Trends model is 0.4.
Sources: Google Trends, FEVAD, Banque de France DGS SEEC.

The graphs in Figure VIII show the change over time in coefficients on the primary axis and that in the lasso penalty in the secondary axis. While it is not common to observe changes in the lasso penalty over time, as it is a different optimisation for each iteration, it helps explains the change in the number of variables selected: the lower the lasso penalty, the higher the number of Google Trends queries selected. With respect to the SARIMA variable, it was expected that its coefficient would be close to 1 as it corresponds to the autoregressive model for the variable. Moreover, changes in coefficients of Google Trends variables – highlighted with those of the lasso penalty – are stable, indicating that these variables model a portion of data not captured by the SARIMA component. The table of mean, minimum and maximum values obtained for each variable is included in Appendix 5.

With respect to selection, almost 9 variables are selected on average for each iteration, which is acceptable in light of the sample sizes (36 observations for the first iteration; 72 for the most recent). The most frequently selected Google Trends variables are eBay, PriceMinister, Groupon, Showroomprivé and Leroy Merlin (see Appendix 5).

Retail Model

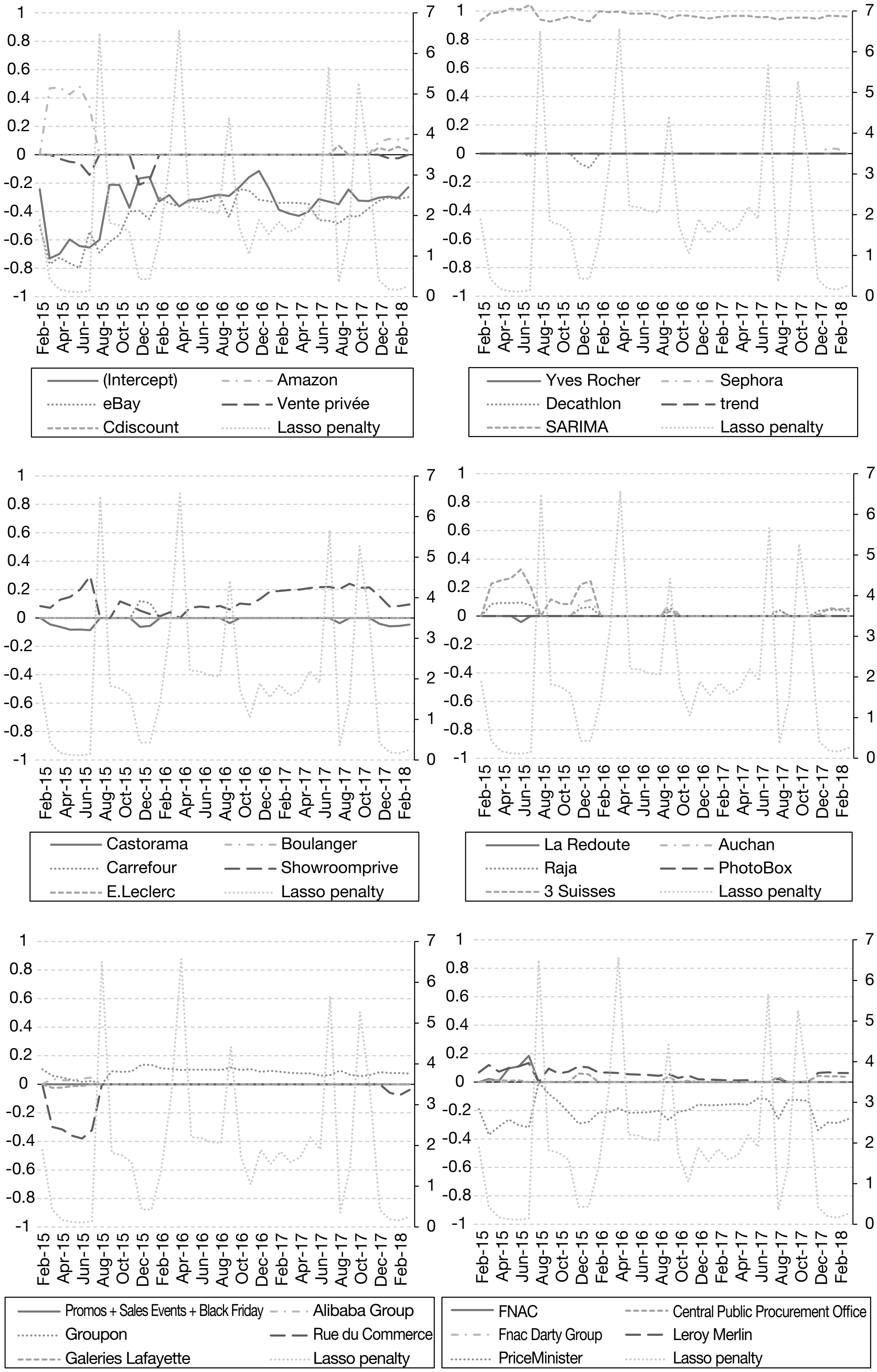
As well as the sales index for physical sales (cf. Table 1), remote sales indices for five products are also used. The sales for the five products contribute by design to the total sales figure. However, as all FEVAD data are released simultaneously, these indices extend to the latest observation using the SARIMA model. Movements in coefficients are detailed in Appendix 6. The model is parsimonious, selecting one to two variables in addition to the autoregressive component (SARIMA estimation). The remote sales index for clothing is selected systematically and logically – on average, the value of sales for clothing represents 22% of the total, the largest of the five products. Its out-of-sample output (RMSFE and mean absolute forecasting errors) is inferior to the Google Trends and SARIMA models (cf. Table 5). For in-sample output, it places between the Google Trends and SARIMA models.

Model Combination

The model combination offers the best forecasting performance under both indicators, RMSFE and the mean absolute error. Over time, it never produces the least accurate forecast. The weight of the models allows us to determine its stability (Figure IX).

Since the end of 2016, the weight of the Google Trends model has increased. On average, it is greater (0.55) than that of the other two models, SARIMA (0.21) and retail (0.18).

Figure VIII
Change in Coefficients for Google Trends Models and the Lasso Penalty



Sources: Google Trends, Banque de France DGS SEEC.

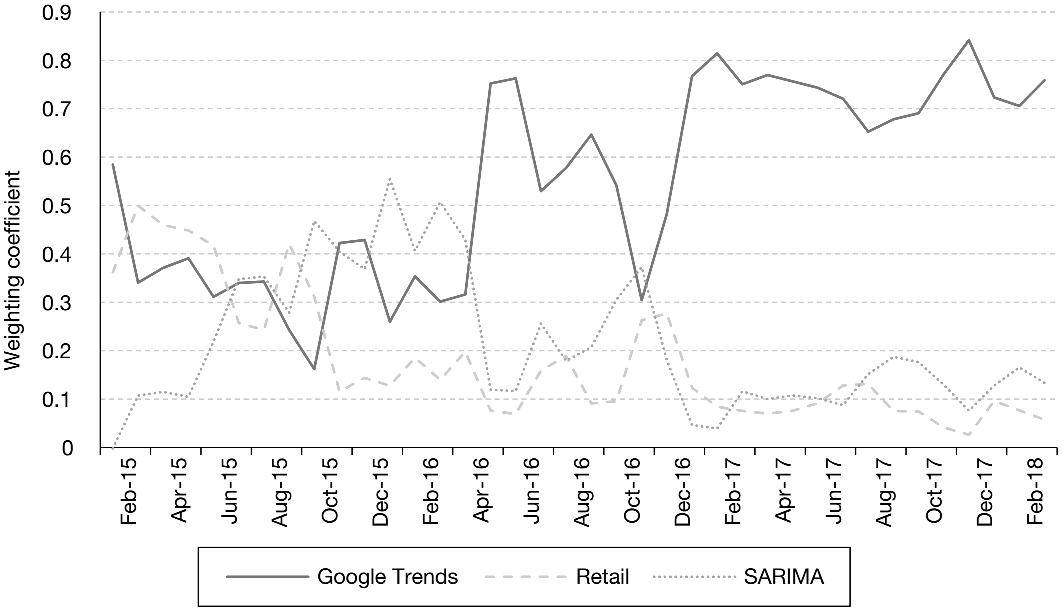
The movements in forecasting errors for the Google Trends, retail and SARIMA models highlight the change in weighting. While the forecasting errors for the three models are relatively close, which can be explained above all by the presence of the SARIMA variable in the Google Trends and retail models, some differences merit particular attention. For example, in October 2016, the Google Trends model had the largest weighting in aggregation, with 0.54, while that for the SARIMA and retail models were 0.31 and 0.10 respectively. For the FEVAD data release at the end of November 2016, it is possible to compare forecasts with their actual value. Figure VII, which sets out forecasting errors, shows that the Google Trends model is the least accurate of the three with an error of 6.0 index points, against 3.3 and 3.7 for the retail and SARIMA models. After the error was "learned", the weighting changed dramatically the following month: the weighting for Google Trends fell to 0.30, with 0.37 for the SARIMA model and 0.26 for the retail model.

Figure IX provides an occasion to detail the Bayesian aggregation formulas. As per the literature review, eight models are possible using three regressors (corresponding in this case to values estimated by Google Trends, retail and SARIMA models). Table 7 sets out the coefficients for each regressor in models M_i $(1 \leq i \leq 8)$ and probability $P(M_i \mid D)$ that each model M_i is the correct one. Lastly, the final column refers to the Bayesian model, whose coefficients are obtained by weighting those for models M_i by probabilities $P(M_i \mid D)$. The values for Table 6 are those for September 2016.

Values for the final column are consistent with Figure IX (September 2016). The Bayesian aggregation is incorporated in the machine learning algorithms; the in-sample error is used to determine the weightings. Figure X shows the movements in the RMSE for the various models.

For each iteration, it is possible to calculate the RMSE obtained for the model's estimation sample. The SARIMA model provides the greatest in-sample error over the period, in contrast to its high predictive capacity. Figure X also shows that the aggregation of data reduces the model's estimation sample error, by comparison with its components. Subsequently, where movements in the in-sample errors for the aggregated models (with and without Google Trends) are similar, the predictive performance is improved with the incorporation of Google Trends data. Lastly, the Google Trends model produces errors close to the other models, supporting the view that the number of variables selected by the adaptive lasso is suitable and that there is no overfitting. To be sure of this,

Figure IX
Change of Weightings in the Model Combination



Sources: Google Trends, FEVAD, Banque de France DGS SEEC.

Table 6
Detailed Calculation of the Weightings in Bayesian Aggregation (September 2016)

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	Bayesian Model
Google Trends	0.96		0.69		0.78		0.82		0.65
Retail			0.31	1.01		0.34	0.38		0.10
SARIMA		0.99			0.21	0.66	- 0.20		0.22
$P(M_i \mid D)$	0.57	0.19	0.09	0.06	0.05	0.02	0.01	0.00	

Sources: Google Trends, Banque de France DGS SEEC.

the in-sample errors (RMSE) may be compared with those produced out-of-sample (RMSFE) (cf. Table 5).

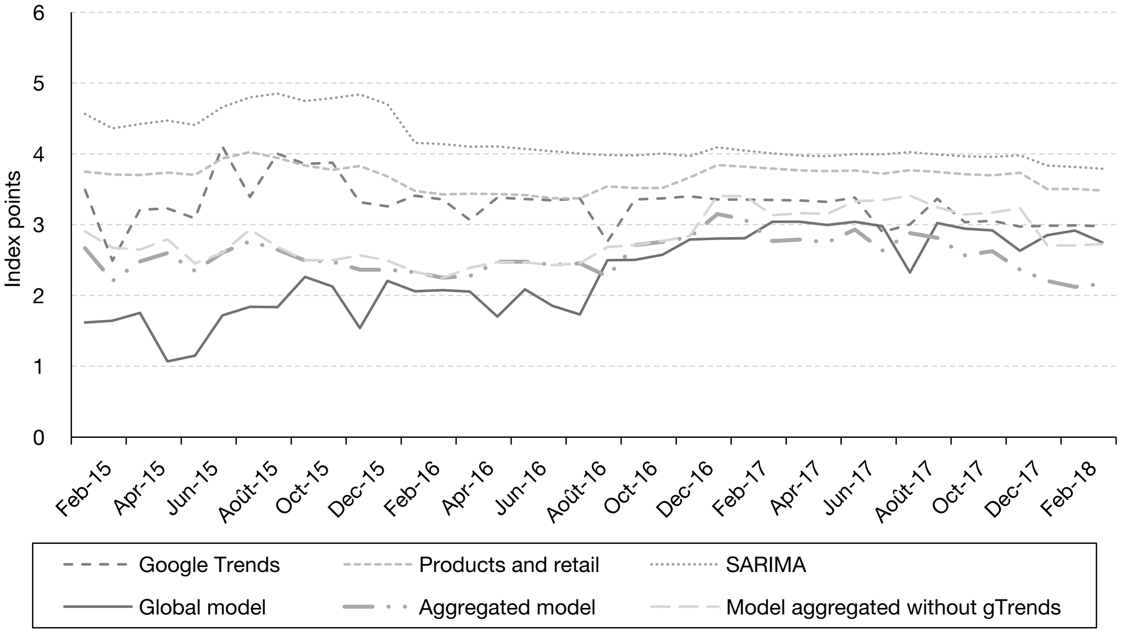
Logically, the forecasting errors are larger. The classifications of models are followed when changing from RMSE to RMSFE, except for the global model, whose out-of-sample error more than doubles.

Global Model

This phenomenon can likely be explained by overfitting. Although the adaptive lasso process is the same as for the Google Trends and retail models, the global model is less parsimonious: on average, 13 variables are selected, which is relatively high by comparison with the number of observations (36 at the first iteration). It selects more variables than the Google Trends and retail models combined. Specifically, over the test protocol period, 82% of variables selected in the global model are selected in one of the other two models; 12% are selected by the global model only and the remaining 6% refer to variables selected by the Google Trends or retail models but not by the global model. In summary, the selection of variables for the global models is too broad, which leads to overfitting. Indeed, movements in coefficients are less stable.

In the case of the overall index, the global model does not perform as well as the model

Figure X
Change in Model RMSEs for Estimation of the Total Sales Index



Reading Note: When estimating the models for the December 2015 forecast, the lowest RMSE was that of the global model (1.5). The highest RMSE was that of the SARIMA (4.8).
Sources: Google Trends, Banque de France DGS SEEC.

combination. In addition, the clarity of the model combination, although limited (output from the Diebold-Mariano test does not conclude that the forecasts of the three models are significantly different), remains better than that of the global model, in which changes in coefficients complicate interpretation. The model combination is therefore preferred. While the output obtained for the overall index has been set out in detail, those for the individual products are summarised below.

Products

Parsimony

The adaptive lasso aims to ensure that models are parsimonious. For each product, Table 7 shows the average number of variables selected per model (covered by the selection of variables).

The retail models are the most parsimonious; one survey variable is selected most frequently, in addition to the SARIMA component. The Google Trends models are less parsimonious; the number of variables selected remains correct in light of the sample sizes, with the possible exception of clothing.

In the Google Trends product models, in addition to the SARIMA component and the constant (systematically selected), the five most selected variables (from 38 iterations) are included in Table 8.

Table 8 illustrates the heterogeneity of the most selected Google queries in Google Trends models: items (ovens, televisions), brands (Cinna, Samsung), general queries (women's clothing, football boots), pure players (Spartoo, GrosBill) and remote retail specialists (3 Suisses). The variety of Google search engine user actions is well-captured here. Note that the trend variable is never selected.

In contrast to the overall index, the global model is more parsimonious for each product than the Google Trends model, thereby reducing one of the risks of overfitting. Table 9 below illustrates the mean values of RMSE.

As expected, the models with full information are better overall, in terms of the RMSE, than models with a single source of information (Google Trends, retail trends, SARIMA). The second finding from Table 9 is that the Google Trends is systematically a better fit for sample

Table 7
Number of Variables Selected by Model Using Adaptive Lasso and by Product

	Google Trends	Retail	Global model
Shoes	9.7	2.2	10.0
Furniture	10.3	2.4	10.9
Household appliances	9.0	2.1	5.9
Consumer electronics	8.8	2.0	10.5
Clothing	12.8	3.6	8.7

Sources: Google Trends, Banque de France DGS SEEC.

Table 8
Most Frequently Selected Variables in Google Trends Models, by Product

Shoes	Spartoo (38)	Sarenza (36)	Converse (36)	Dress shoes (32)	Football boots (28)
Furniture	Cinna (38)	Roset (33)	Wooden furniture (32)	Dresser + cupboard + cabinet (31)	IKEA (26)
Household appliances	Washing machine (37)	Oven (28)	Cooker (28)	Conforama (26)	GrosBill (24)
Consumer electronics	SLR digital camera (37)	Television (35)	JBL (35)	Sony (27)	Samsung Electronics (23)
Clothing	Suit (34)	Decoration (29)	Jennyfer (28)	Lingerie (27)	3 Suisses (25) and Women's clothing (25)

Sources: Google Trends, FEVAD, Banque de France DGS SEEC.

data than the retail and SARIMA models; which may be explained by the larger number of variables selected.

Predictive Capacity

While the Google Trends model was, on average, systematically better over the estimation period (based on RMSE) than the retail and SARIMA models, it is not the best for forecasting. Its overall predictive performance is broadly equal to that of the retail and SARIMA models (Table 10). More generally, the results obtained for the various products are mixed. The addition of exogenous data – Google Trends data or quantitative survey indices – does not reduce the forecasting error.

The retail model is the best-performing model for shoes; the Google Trends is slightly better, in terms of RMSFE, than the SARIMA model. For furniture, Google Trends offers the clearest input. “Consumer electronics” also recorded an improvement (by comparison with the SARIMA model) with exogenous data. However, for household appliances and clothing, their input did not improve the results (by comparison with the SARIMA model).

With respect to the combination of data, the results are also mixed. On the one hand, the combination of models offers improved forecasting output (RMSFE) than the global model, except for consumer electronics. On the other hand, the combination of data does not deliver the expected results. Based on RMSFE, the model combination is only better for furniture, the only product for which the Google Trends model outperforms the retail and SARIMA models. The in-sample performances affect the weighting of the models in aggregation. As the Google Trends model produces the best estimations (based on the mean RMSE, see Table 5), its weighting in aggregation is larger (Table 11).

The mean weighting in aggregation is based on the test protocol period, as is the case for the mean RMSE. Clothing is the sole product group for which the Google Trends model weighting is not the largest.

Table 9
Mean RMSE for Models in Estimation of Product Sales Indices

	Google Trends	Retail	SARIMA	Global model	Model combination
Shoes	8.2	10.5	10.9	7.8	7.6
Furniture	6.0	7.3	7.4	5.7	5.5
Household appliances	6.1	6.9	7.2	6.4	5.5
Consumer electronics	5.8	7.4	7.7	5.5	7.2
Clothing	5.3	6.0	6.3	5.9	4.3

Sources: Google Trends, Banque de France DGS SEEC.

Table 10
RMSFE and Standard Deviations related to Google Sampling in the Estimation of Product Sales Indices

	Google Trends		Retail	SARIMA	Global model		Model combination	
Shoes	13.2	0.3	12.7	13.6	13.8	0.4	13.4	0.2
Furniture	11.9	0.5	12.3	12.0	13.2	0.4	11.8	0.5
Household appliances	11.7	0.3	10.4	10.2	12.3	0.3	11.2	0.3
Consumer electronics	15.5	0.3	15.3	16.4	11.5	0.5	13.1	0.3
Clothing	9.8	0.3	10.1	9.2	15.2	0.5	9.7	0.2

Notes: Results included in the main text of the paper for the five products are from the same simulation as those for the total index; RMSFEs are very close to the median values obtained for the thirty simulations.
Reading Note: The RMFSE for the Google Trends model in estimation of the shoe sales index is 13.2; for the thirty simulations carried out to determine the sensitivity of results to Google sampling, the standard deviation is 0.3. The RMSFE of the retail model is 12.7 (and is not impacted by Google sampling).
Sources: Google Trends, Banque de France DGS SEEC.

Table 11
Weighting of Individual Models in the Model Combination

	Google Trends	Retail	SARIMA
Shoes	0.55	0.21	0.19
Furniture	0.71	0.09	0.13
Household appliances	0.47	0.14	0.33
Consumer electronics	0.48	0.20	0.27
Clothing*	0.58	- 0.50	0.80

* For clothing, the values predicted by the three models exhibit strong colinearity, poorly handled by Bayesian aggregation: the contribution of variables in "intermediate" models (see detailed calculation of weightings for the total, in the relevant section) are artificially overvalued; this has a knock-on effect on mean weightings for the model combination.
Sources: Google Trends, Banque de France DGS SEEC.

* *
*

Online retail is rapidly growing. Purchases made online account for a greater proportion of household consumption, and thus the Banque de France monthly retail trends survey. Against this backdrop, the estimation of sales figures released (belatedly) by FEVAD becomes a prominent question.

Up to now, this has been carried out using an autoregressive model. The research set out in this paper looks at the contribution of the exogenous data that are traditional retail indices for physical sales (in monthly retail trend surveys) and Google Trends indices. Each data source provides its own input. The common benefit of such data sources, namely being available before FEVAD releases, is ideal for nowcasting.

However, a new source of data (Google Trends) must be used with caution. Firstly, robustness tests prior to use have been necessary. A system of treating outliers was implemented. Such outliers are sometimes the result of methodological changes introduced by Google and for which little information is made available. The sensitivity of output to the sampling method used by Google prompts multiple simulations to increase the reliability of output. Secondly, it was necessary to reconcile the huge range of possible Google variables with the lack of historical FEVAD time series data (monthly releases date back to 2012). This twin constraint can be overcome by machine learning, using the adaptive lasso process (Zou, 2006). The selection of variables at each iteration, thereby minimising the risks associated with rapid developments in online retail and possible instability of corresponding keywords, as it is possible to backward-extrapolate output with other sets of variables. This way, the model is flexible and offers substantial adaptive capacity, which the ever-changing nature of the modelled phenomenon requires.

The question then arises as to how to exploit the complementarity of the various data sources. In this paper, Bayesian aggregation of single models produces better results in terms of RMSFE, than the global model (adaptive lasso applied to all variables simultaneously). The small size of estimation samples for the models may work against a model with many variables. For example, in the case of the overall index, overfitting is detected for the global model. In addition, aggregation offers clarity in the combination of models, which is useful in production.

In general, the contribution of exogenous data remains mixed. It is clearer for the overall index than the index for products. FEVAD releases are developed from a sample of 70 of its largest respondents (in terms of sales). The number of respondents is therefore lower for product groups; this substantiates the results that are most impactful and thus the most difficult to comprehend. The forecasting error for sales is therefore two to three times greater for products than for the total.

Lastly, one of the possible causes of mixed results lies with model selection. While they meet many of the constraints posed, seasonality is not always fully taken into account. Due to the short time series, the model does not operate on seasonally adjusted series, unlike the standard econometric approach; here, the presence of SARIMA estimation for explanatory variables seeks to capture seasonality.

However, this method overlooks the differences in seasonality between endogenous and exogenous variables.

With longer time series, seasonality in online retail time series should stabilise, offering the opportunity to refine output with other models. As well as the possibility of working on seasonally adjusted time series, combining RegARIMA modelling, for which the residual specification is more applicable, with variable selection methods may prove valuable and is currently not addressed in the existing body of research. □

BIBLIOGRAPHY

Aiofli, M. & Timmerman, A. (2006). Persistence of forecasting performance and combination strategies. *Journal of Econometrics*, 135(1-2), 31–53.
https://doi.org/10.1016/j.jeconom.2005.07.015

Askistas, N. & Zimmerman, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55(2), 107–120.
https://elibrary.duncker-humblot.com/journals/id/22/vol/55/iss/1486/art/5561/

Bates, J. & Granger, C. (1969). The combination of forecasts. *Operational Research Quarterly*, 20(4), 451–468.
https://doi.org/10.1057/jors.1969.103

Bec, F. & Mogliani, M. (2015). Nowcasting French GDP in real-time with surveys and "blocked" regressions: Combining forecasts or pooling information? *International Journal of Forecasting*, 31(4), 1021–1042.
https://doi.org/10.1016/j.ijforecast.2014.11.006

Bortoli, C. & Combes, S. (2015). Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées. Insee, *Note de conjoncture*, mars 2015.
https://www.insee.fr/fr/statistiques/fichier/1408926/mars2015_d2.pdf

Breiman, L. (1996). Stacked Regressions. *Machine Learning*, 24, 49–64.
https://statistics.berkeley.edu/sites/default/files/tech-reports/367.pdf

Choi, H. & Varian, H. (2009). Predicting Initial Claims for Unemployment Benefits. Google, *Technical Report.*
https://static.googleusercontent.com/media/research.google.com/fr//archive/papers/initialclaimsUS.pdf

Choi, H. & Varian, H. (2011). Predicting the Present with Google Trends. Google, *Technical Report.*
http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf

Clements, M. & Galvão, A. (2008). Macroeconomic Forecasting With Mixed-Frequency Data: Forecasting Output Growth in the United States. *Journal of Business & Economic Statistics*, 26(4), 546–554.
https://doi.org/10.1198/073500108000000015

Diebold, F. (1989). Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting*, 5(4), 589–592.
https://doi.org/10.1016/0169-2070(89)90014-9

Diebold, F. & Mariano, R. (1995). Comparative Predictive Accuracy. *Journal of Business & Economic Statistics*, 20(1), 134–144.
https://doi.org/10.1198/073500102753410444

De Gooijer, J. & Hyndman, R. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22(3), 443–473.
https://doi.org/10.1016/j.ijforecast.2006.01.001

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407–499.
https://doi.org/10.1214/009053604000000067

Elliott, G., Rothenberg, T. & Stock, J. (1996). Efficient Tests for an Autoregressive Unit Root. *Econometrica*, 64(4), 813–836.
https://doi.org/10.2307/2171846

Ettredge, M., Gerdes, J. & Karuga, G. (2005). Using Web-based Search Data to Predict Macroeconomic Statistics. *Communications of the ACM*, 48(11), 87–92.
https://www.researchgate.net/publication/200110929_Using_Web-based_search_data_to_predict_macro economic_statistics

Eurostat (2018). *Handbook on Seasonal Adjustment.* Luxembourg: Publications Office of the European Union.
https://ec.europa.eu/eurostat/documents/3859598/8939616/KS-GQ-18-001-EN-N.pdf

FEVAD (2016 and 2017). *Chiffres clés*.
https://www.fevad.com/wp-content/uploads/2016/09/Plaquette-Chiffres-2016_Fevad_205x292_format-final_bd.pdf
https://www.fevad.com/wp-content/uploads/2018/06/Chiffres-Cles-2018.pdf

Granger, C. & Newbold, P. (1974). Spurious Regressions in Econometrics. *Journal of Econometrics*, 2(2), 111–120.
https://doi.org/10.1016/0304-4076(74)90034-7

Hoerl, A. & Kennard, R. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1), 55–67.
http://www.jstor.org/stable/1267351?origin=JSTOR-pdf

Hoeting, J., Madigan, D., Raftery, A. & Volinsky, C. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4), 382–417.
https://www.jstor.org/stable/2676803

Huang, H. & Lee, T. (2010). To Combine Forecasts or to Combine Information? *Econometric Reviews*, 29(5-6), 534–570.
https://doi.org/10.1080/07474938.2010.481553

Hyndman, R. & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. Melbourne, Australia : OTexts.
https://otexts.org/fpp2/

Kozicki, S. & Hoffman, B. (2004). Rounding Error: A Distorting Influence on Index Data. *Journal of Money*, Credit and Banking, 36(3), 319–338.
https://www.jstor.org/stable/3838976

Kuzin, V., Marcellino, M. & Schumacher, C. (2013). Pooling versus model selection for nowcasting GDP with many predictors: Empirical evidence for six industrialized countries. *Journal of Applied Econometric*, 28(3), 392–411.
https://doi.org/10.1002/jae.2279

Marin, J. & Robert, C. (2010). Les bases de la statistique bayésienne. *Rapport des Universités Montpellier II et Dauphine – Crest Insee*.
https://www.ceremade.dauphine.fr/~xian/mr081.pdf

McLaren, N. & Shanbhogue, R. (2011). Using internet search data as economic indicators. Bank of England, *Quarterly Bulletin*, 51(2), 134–140.
https://econpapers.repec.org/RePEc:boe:qbullt:0052

Phillips, P. (1986). Understanding spurious regression in econometrics. *Journal of Econometrics*, 33(3), 311–340.
https://doi.org/10.1016/0304-4076(86)90001-1

Phillips, P. & Perron, P. (1988). Testing for a Unit Root in Time Series Regression. *Biomètrika*, 75(2), 335–346.
http://www.jstor.org/stable/2336182?origin=JSTOR-pdf

Tibshirani, R. (1996). Regression shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.
https://www.jstor.org/stable/2346178

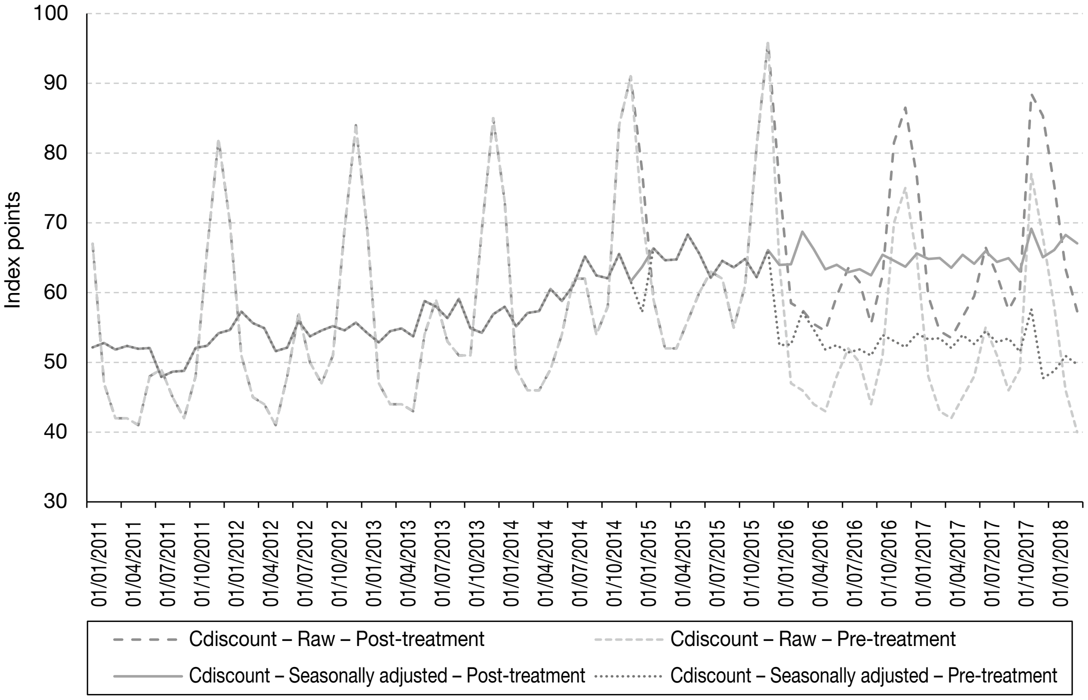
Zeugner, S. (2011). *Bayesian Model Averaging with BMS*.
https://cran.r-project.org/web/packages/BMS/vignettes/bms.pdf

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.
https://doi.org/10.1198/016214506000000735

APPENDIX 1

TREATMENT OF OUTLIERS - THE EXAMPLE OF CDISCOUNT

Figure A1
Treatment of the Break in Series of the Google Trends Index for Cdiscount



Note: The treatment applied to the Cdiscount series is analogous with that for Amazon.
Sources: Google Trends, Banque de France DGS SEEC.

APENDIX 2

LIST OF VARIABLES BY PRODUCT

Table A2
Set of Initial Variables for Each Estimation

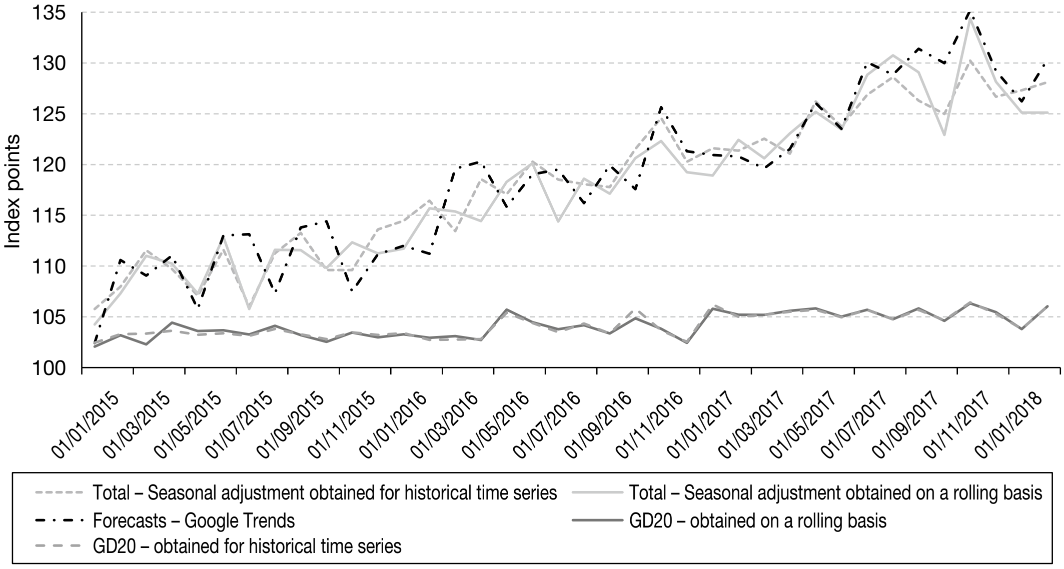
Total	Amazon, eBay, Vente privée, Cdiscount, Fnac, Fnac Darty Group, PriceMinister, Leroy Merlin, UGAP, Castorama, Boulanger, Carrefour, Showroomprive, E. Leclerc, La Redoute, Auchan, Raja, Rue du commerce, 3 Suisses, Promos + Sales Events + Black Friday, Alibaba, Groupon, PhotoBox, Galeries Lafayette, Yves Rocher, Sephora, Decathlon
Clothing	Vertbaudet, Kiabi, H&M, C&A, Jules, home textile, Zara, Suit, Underwear, 3 Suisses, Devred, Robe, Etam, La Redoute, Jeans + Chinos + Trousers, Coat + Jacket, Jacket, Women's clothing, ASOS, Maisons du Monde, Lingerie, Jennyfer, Clothing, Galeries Lafayette, Bonobo, Brandalley, Camaïeu, Showroomprive, Vente Privée, curtain, Blanche Porte, bedsheet, Cushion, Homemaison, Underwear, La Halle, Decathlon
Consumer electronics	iPhone, Apple, Cdiscount, PC Gaming, iPad, Telephone + smartphone , FNAC, Television, Boulanger, Sony, LDLC Pro, Amazon, Phillips, LG Group, Samsung Electronics, Darty, Tablet, Speaker, SLR digital camera, Laptop computer + PC, Bose, JBL, Fnac Darty Group, Soundbar, Camera, Marshall, Samsung Group
Shoes	Shoes, Shoe, Belt, Leather goods, Boots, Sport shoes, Vans, Converse, Zalando, Spartoo, Sarenza, Showroomprive, Prada, Escarpin, Adidas Stan Smith, Women's shoes, Pumps, Men's shoes, Timberland, Football boots, Children's shoes, San Marina, Eram, Dress shoes, J.M Weston, Chaussea, Bexley, Gémo, Handbag, La Halle, Nike shoes
Household appliances	Clubic, Boulanger, Cdiscount, Oven, Fridge, Washing Machine, Darty, Bosch, Electrolux, Conforama, Amazon, Cooker, Électro-dépôt, Brandt, Microwave oven, Fnac Darty Group, Vacuum Cleaner, Whirlpool Corporation, Mistergooddeal, GrosBill, Pulsat, Ubaldi, But
Furniture	But, Legallais, Kitchen, Raja, Staples, Roche Bobois, Castorama, Conforama, Vega, Bureau, Furniture, Leroy Merlin, Ikea, Knives, Cupboard + shelves, Maisons du Monde, Cinna, Wooden furniture, Dresser + cupboard + cabinet, Roset, Table + chair + sofa, Armchair

Reading Note: "+" here denotes a Google Trends index with combined queries.
Sources: Google Trends, Banque de France DGS SEEC.

APPENDIX 3

INSTABILITY OF LATEST OBSERVATIONS IN SEASONAL ADJUSTMENT

Figure A3-I
Stability of Seasonal Adjustment for the Most Recent Observation



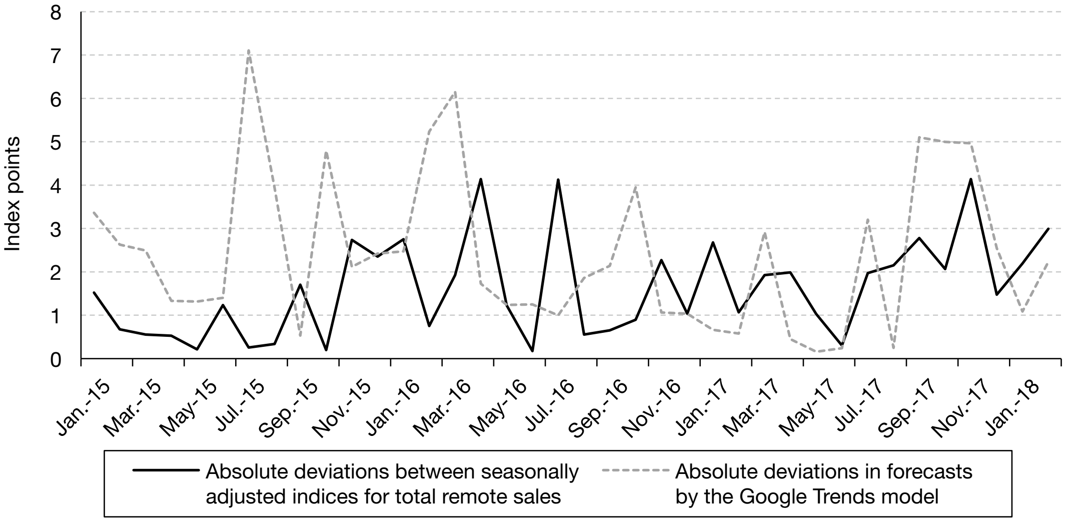
Sources: Google Trends, FEVAD, Banque de France DGS SEEC.

Two groups of series are distinguished. The first applies to remote selling/remote sales and consists of three series: an aggregate remote sales index for which seasonal adjustment takes account of the whole data, a "rolling" aggregate remote sales index, implemented on a rolling basis – i.e. each observation is the last in the seasonally adjusted series obtained using the truncated index on that date, and the Google Trends forecasts index.

The second group is applied to large retailers: the mean absolute deviation between two seasonally adjusted indices for large retailers (obtained using all data versus those available on a rolling basis) is 0.2 (0.05 where the size is related to the amplitude, defined as the largest variation in the reference series – seasonal adjustment obtained for historical data); against 1.6 (0.29 related to amplitude) for both seasonally adjusted indices that make up the total for remote sales.

While revisions to the most recent observations of seasonally adjusted series are not unknown (see Eurostat, 2018), the magnitude observed here presents a challenge: both of these indices vary by proportions similar to the forecasting errors of the models: movements in errors between, on the one hand, the two seasonal adjustments, and on the other hand, forecasts from the Google Trends model and seasonally adjusted series based on historical data, account for this (Figure A3-II).

Figure A3-II
Absolute Deviations with Seasonal Adjustment for the Historical Time Series

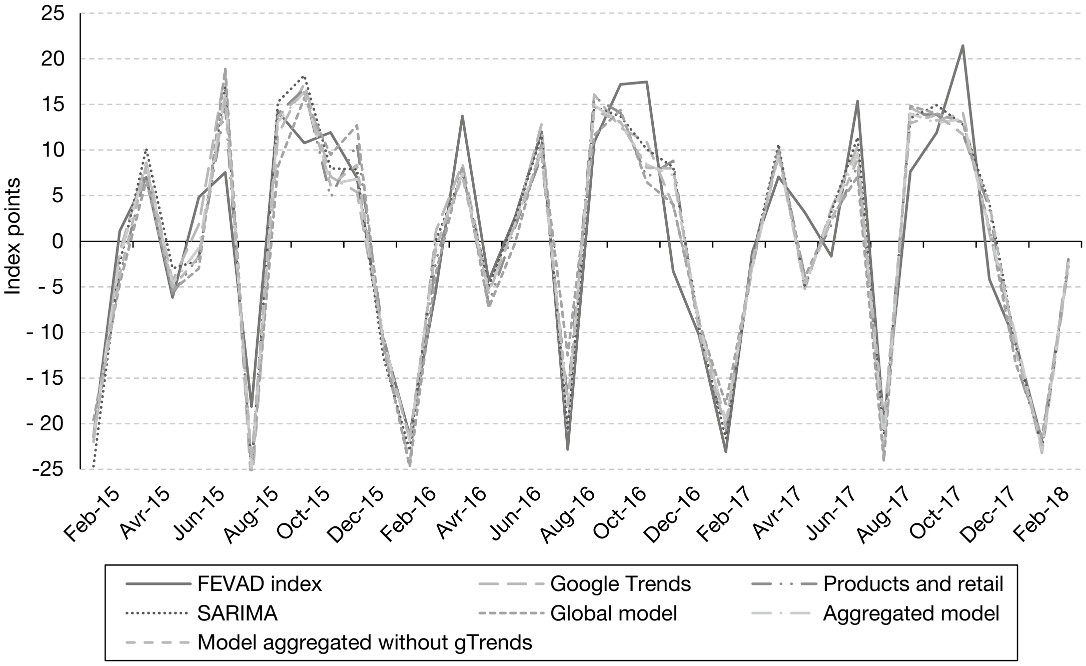


Reading Note: The seasonally adjusted estimation using the data available at present, varies by 4.1 index points from that carried out in quasi-real time. The Google Trends forecast for July 2016 is even closer to the obtained value for the seasonally adjusted index than that obtained using the data available in July 2016.
Sources: Google Trends, FEVAD, Banque de France DGS SEEC.

APPENDIX 4

MODEL FORECASTS FOR THE TOTAL SALES INDEX

Figure A4
Forecasts from the Various Models in Estimation of the Total



Sources: Google Trends, FEVAD, Banque de France DGS SEEC.

APPENDIX 5

DESCRIPTION OF VARIABLE SELECTION IN THE GOOGLE TRENDS MODEL FOR TOTAL SALES ESTIMATIONS

Table A5
Descriptive Statistics for Variables Selected in Estimation of the Total

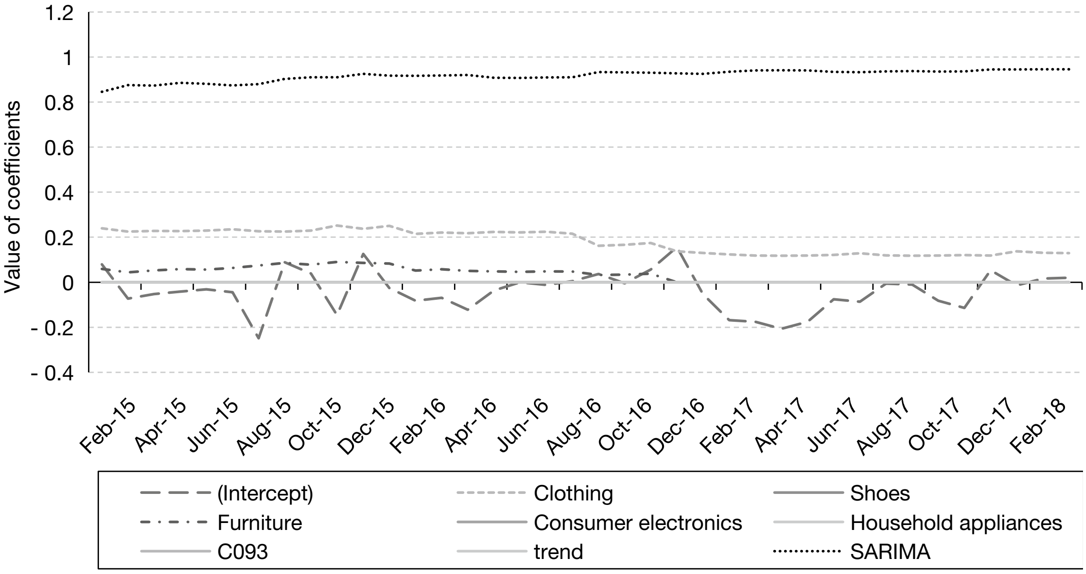
	Mean	Minimum	Maximum	Selections
Amazon	0.07	0.00	0.48	9
eBay	- 0.43	- 0.80	- 0.24	38
Vente-privee.com	- 0.02	- 0.21	0.00	8
Cdiscount	0.01	0.00	0.07	5
FNAC	0.01	0.00	0.18	5
Fnac Darty Group	0.00	0.00	0.00	0
PriceMinister	- 0.21	- 0.37	0.00	37
Leroy.Merlin	0.05	0.00	0.13	32
Central Public Procurement Office	0.01	0.00	0.06	14
Castorama	- 0.02	- 0.09	0.00	13
Boulanger	0.00	0.00	0.00	0
Carrefour	0.01	0.00	0.12	2
Showroomprive.com	0.12	0.00	0.29	35
E.Leclerc	0.00	0.00	0.00	0
La Redoute	0.00	- 0.04	0.00	1
Auchan	0.01	0.00	0.11	5
Raja	0.02	0.00	0.10	13
Rue du Commerce	- 0.05	- 0.38	0.00	8
3 Suisses	0.06	0.00	0.33	15
Promos + Sales Events + Black Friday	0.00	0.00	0.00	0
Alibaba Group	0.00	0.00	0.05	7
Groupon	0.08	0.00	0.14	37
PhotoBox	0.00	0.00	0.00	0
Galeries Lafayette	0.00	- 0.02	0.00	4
Yves Rocher	0.00	0.00	0.00	0
Sephora	0.00	0.00	0.04	2
Decathlon	0.00	- 0.10	0.00	3
Trend	0.00	0.00	0.00	0
SARIMA	0.97	0.93	1.04	38

Sources: Google Trends, FEVAD, Banque de France DGS SEEC.

APPENDIX 6

CHANGE IN COEFFICIENTS IN THE RETAIL TRENDS MODEL IN THE ESTIMATION OF TOTAL SALES

Figure A6
Change in Coefficients for the Retail Model in the Estimation of the Total



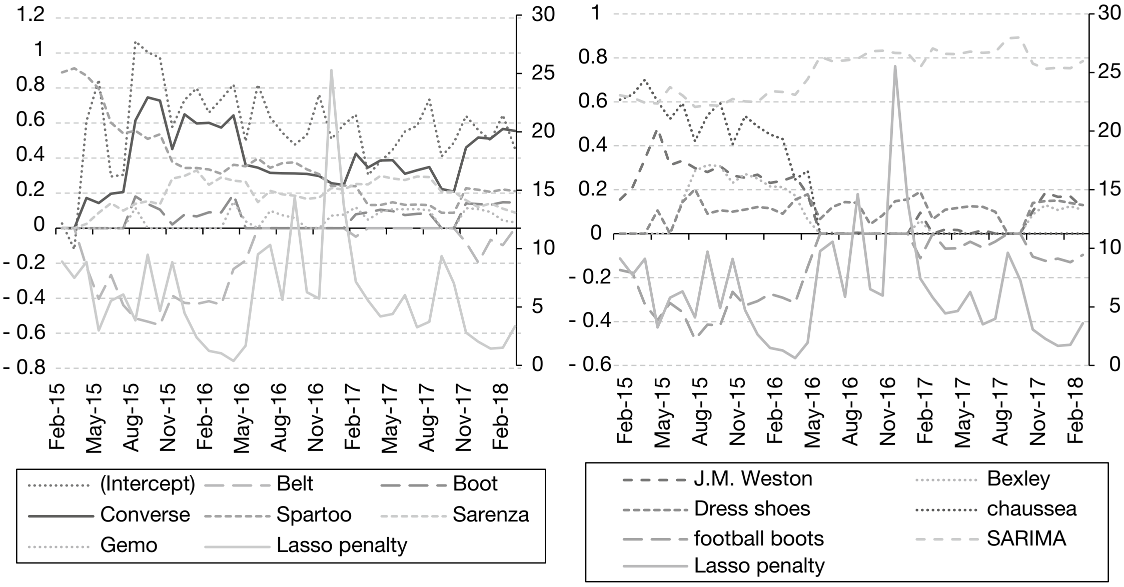
Note: Only selected variables are not shown as a solid line.
Sources: FEVAD, Banque de France DGS SEEC.

APPENDIX 7

STABILITY OF THE GOOGLE TRENDS MODEL IN ESTIMATION OF THE SHOES INDEX

Similar to the figure VIII showing the change in Google Trends model coefficients for the estimation of the total sales index, the lasso penalty is on the secondary axis. Figure A7 below presents only the most often selected variables (at least eight times out of 38 iterations) over the period.

Figure A7
Change in Google Trends Model Estimation, Coefficients of the Index for Shoes



Sources: Google Trends, Banque de France DGS DESS SEEC.

Nowcasting and the Use of Big Data in Short-Term Macroeconomic Forecasting: A Critical Review

Pete Richardson*

Abstract – This paper provides a discussion of the use of Big Data for economic forecasting and a critical review of recent empirical studies drawing on Big Data sources, including those using internet search, social media and financial transactions related data. A broad conclusion is that whilst Big Data sources may provide new and unique insights into high frequency macroeconomic activities, their uses for macroeconomic forecasting are relatively limited and have met with varying degrees of success. Specific issues arise from the limitations of these data sets, the qualitative nature of the information they incorporate and the empirical testing frameworks used. The most successful applications appear to be those which seek to embed this class of information within a coherent economic framework, as opposed to a naïve black box statistical approach. This suggests that future work using Big Data should focus on improving the quality and accessibility of the relevant data sets and in developing more appropriate economic modelling frameworks for their future use.

JEL Classification: C53, E27, E37
Keywords: Big Data, internet search, short-term, macroeconomic forecasting, models, nowcasting

* *Senior Associate with Llewellyn-Consulting, London (pete.w.richardson@gmail.com)*

The author, formerly Head of the Macroeconomic Analysis Division in the Department of Economics, OECD Paris, is grateful to former colleagues including Nigel Pain, David Turner and Christophe André at the OECD and to Robert Kaufmann of Boston University for comments and suggestions on earlier versions of this paper and its presentation to the OECD's Working Group on New Approaches to Economic Challenges (NAEC) in Paris, January 2016 and to the UN Project LINK Conference in New York, October 2015. Thanks also go to the anonymous reviewers for helpful comments and suggestions on an earlier draft. Much of the study was originally carried out as part of the background research for the OECD's study "OECD Forecasts During and After the Financial Crisis: A Post Mortem", as reported in Pain et al. *(2014) and Lewis & Pain (2015).*

Received on 29 September 2017, accepted after revisions on 11 May 2018

To cite this article: Richardson, P. (2018). Nowcasting and the Use of Big Data in Short-Term Macroeconomic Forecasting: A Critical Review. *Economie et Statistique / Economics and Statistics*, 505-506, 65–87. https://doi.org/10.24187/ecostat.2018.505d.1966

Although much has been made of the possible role and uses of so-called Big Data in macroeconomic forecasting, there appear to be relatively few systematic reviews of related empirical work to date.[1] This paper seeks to redress the balance by providing a discussion of the relevance of Big Data for economic forecasting and a critical review of a number of empirical studies published to date, drawing on a number of different sources, including internet search and social media-related information and financial and other transactions-related statistics. It does so primarily from a practical economic forecasting perspective.

As noted by Bok *et al.* (2017), whilst "Big Data" is currently associated with those very large economic data sets derived from internet and electronic transactions sources, many of the related challenges to economists and statisticians existed well before their collection became feasible and pervasive for economics and other disciplines. Such challenges are exemplified by the pioneering work of Burns and Mitchell at the NBER[2] to identify business cycles using a very large range of data sets, the dedicated work of Kuznets and many others in developing consistent frameworks for the measurement of the National Accounts and related statistical concepts culminating in the large range of data collection and analyses currently undertaken. At the same time, developments in econometrics and time-series methods over past decades now permit the construction of consistent methods and suitable platforms for monitoring macroeconomic conditions in almost real time.[3]

The main starting point and motivation for the present review came from an analysis of the OECD's international forecasting record during and after the financial crisis, as described by Pain *et al.* (2014) and Lewis & Pain (2015). In common with many national international institutions, and in line with more recent developments in so-called "nowcasting" techniques, the OECD's near-term macroeconomic assessments routinely take account of forecasts from a suite of statistical models using high frequency economic indicators to provide estimates of near-term GDP growth for the euro area and individual G7 economies for the current and next quarter.[4] These models typically use a Vector Autoregressive "bridge model" approach to combine information from a variety of "soft" indicators, such as business sentiment and consumer surveys, with "hard" indicators, such as industrial production, retail sales, house prices, etc., using different frequencies of data and a variety of estimation techniques. The associated estimation procedures are relatively automated and can be run as new monthly data are released, allowing also for timely updating and model choice according to the available information set.

Empirically, the main gains from using such an approach are typically found to be largest for current-quarter GDP forecasts made at or immediately after the start of the quarter in question, where estimated indicator models appear to outperform simple autoregressive time series models, in terms of both the size of predictive error and directional accuracy. Thus, the largest gains arise once one month of data is available for the quarter being forecast, typically two to three months before the publication of the first official outturn estimate for GDP. For one-quarter-ahead projections, the performance of the estimated indicator models is only noticeably better than simpler time series models once one or two months of information become available for the quarter preceding that being forecast. Modest gains are nonetheless made in terms of directional accuracy from using indicator models.

The general nature of these gains are illustrated in the Figure below which provides a summary of the successive revisions to the OECD's near-term quarterly GDP forecasts for the aggregate G7 economies during the 2008/9 downturn and subsequent recovery period. On this basis, the pre-recession period comparisons show relatively little systematic difference in predictive accuracy as between current- and next-quarter models (shown by the clear and lightly shaded bars respectively). However, from the second half of 2008, through the downturn and subsequent recovery, the current-quarter model predictions are clearly superior to the initial projections, reflecting the relative importance of hard information. The overall conclusion is that GDP indicator models provided a useful

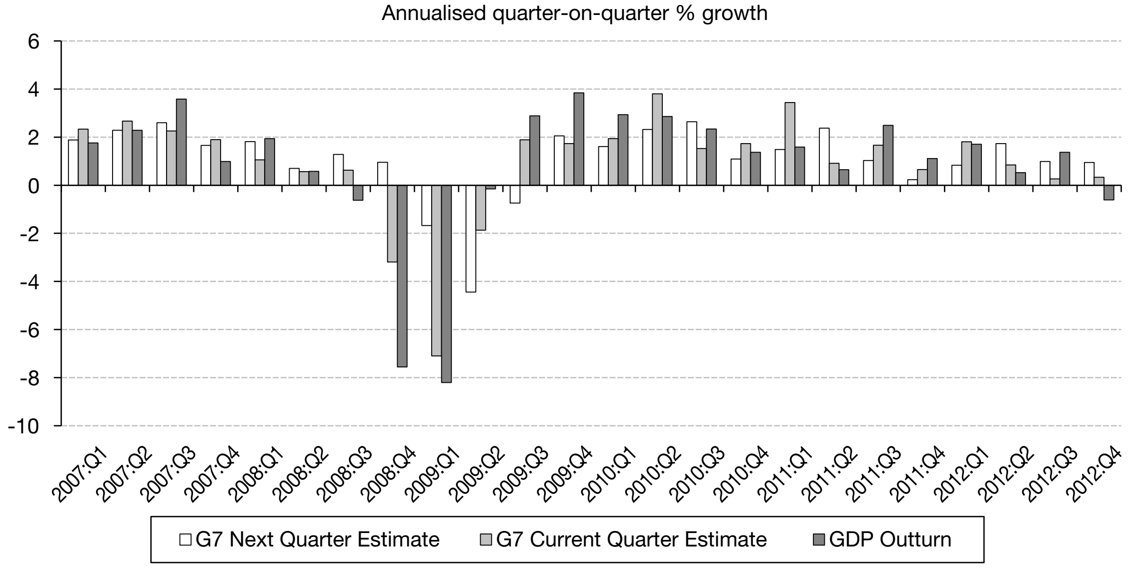
[1] *Useful background to the literature on Big Data sets and their uses in recent empirical studies are also given by Buono* et al. *(2017), Bok* et al. *(2017), Hellerstein & Middeldorp (2012), Hassani & Silva (2015) and Ye & Li (2017).*

[2] *See Burns & Mitchell (1946).*

[3] *In particular see the recent work of Giannone* et al. *(2008) and others, in developing consistent frameworks for near-term statistical analysis and so-called "nowcasting" by combining models for Big Data with modern filtering and estimation techniques.*

[4] *At the OECD these models build on the pioneering work of Sédillot & Pain (2003) and Mourougane (2006) in using short-term economic indicators to predict quarterly movements in GDP by efficiently exploiting available monthly and quarterly information.*

Figure
The OECD's G7 GDP Current and Next Quarter Projections and Outturns Through the Financial crisis



Notes: Current quarter estimates for period 2007-Q1 - 2012-Q4: Mean error = - 0.1; MAE = 1.0; RMSE (actual) = 1.3; RMSE (estimation) = 1.6. For next quarter: Mean error = - 0.2; MAE = 1.6; RMSE (actual) = 2.6; RMSE (estimation) = 2.0.
Figure reports successive OECD forecasts and outturns for quarterly growth in real GDP for the G7 countries, over the period 2007 to 2012, based on the real time OECD short-term indicator models.
Sources: OECD, Pain *et al.* (2014).

basis for assessing current economic conditions through the recession at the point where hard indicator information became available, even though the scale of the global shock was entirely outside of the within-sample experience of the estimated models. Predictive performance was noticeably worse where hard indicator information was absent.

An important limitation in the practical use of indicator and similar nowcast models therefore concerns the lags in availability of hard statistical information, from National Statistical Offices and other statistical and survey agencies. Typically the quarterly goodness-of-fit and out-of-sample predictive performance of such models are found to improve significantly the more information is available for monthly hard indicators during the quarter in question, raising the question of how the availability of more timely from alternative sources might assist the tasks of short-term economic assessment and surveillance.

Big Data, Nowcasting and the Use of Electronic Indicators in Economic Forecasting

Reflecting these concerns, a number of recent academic and institutional studies, mostly post-crisis, have focussed on the possible usefulness of a wider set of data sources than those traditionally provided by the National Statistical authorities and in particular those described as being "Big Data" sets. The term Big Data, has been used in the computing industry field since the early 1990s to describe data sets with sizes which are, or were, typically beyond the ability of commonly used software tools and computer capacities to capture, manage, and process within a tolerable elapsed time, encompassing a wide range of unstructured, semi-structured and structured data sets. However, with the exponential growth of data storage and processing capacities over the recent past, the availability and use of Big Data sets have become increasingly feasible for economists and other analysts.[5]

In this context, a number of recent empirical studies have focussed on the possible usefulness to economic forecasting of three broad sources of such information:

- Internet search statistics, based on the frequency of searches for specific key words or topics;

[5] *In economics, Diebold (2000) was first in describing Big Data as "the explosion in the quantity (and sometimes quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage techniques".*

- Internet-based social media and blog sources, such as Twitter;

- Detailed micro-level transactions data, recorded electronically by rapidly growing and popular financial payments and transactions systems.

The key advantages of using such sources of information lie in the coverage and level of detail they provide (down to individual micro transaction levels) and their timeliness. Being, in principle, available on a near real-time basis they offer a snapshot of current transactions, trends and tendencies well before they become recorded in official statistics. Nonetheless, key challenges remain in their use and development, including interpretation and analysis, as well as traditional concerns about their capture, curation, storage, sharing, visualization, and about privacy.[6]

Against this background, the following sections provide a discussion and critical review of recent studies using data from each of these three main areas for macroeconomic-related analyses and economic forecasting.[7] [8] To complement the review, an annotated summary guide to each of these studies, including their general coverage, the techniques employed and their principal findings and reservations is provided in Appendix.

The Use of Internet Search Information in Macroeconomic Models and Forecasting

Following the pioneering work of Ettredge *et al.* (2005), Choi & Varian (2009a and 2009b) and Wu & Brynjolfsson (2009), a growing body of literature has evolved on the use of internet search statistics in models used for economic forecasting and assessment. Typically such studies involve the construction of weekly, monthly and quarterly time-series indicators related to the "frequency" of internet searches for one or more specific keyword or phrase relevant to a specific topic or category of economic activity for a particular geographic location or country. For example, these might relate to searches for terms such as "welfare and unemployment benefit" or "mortgage foreclosure" or "car loans" etc., for "country A" or "state B". The relevant time-series indicator is then typically added to and tested for significance within a baseline forecasting model on within and out-of-sample bases. The underlying rationale is that internet search has now become a widespread and growing means for economic agents to obtain information relevant to their immediate economic situations, activities and decisions, those which ultimately get reflected in their behaviour and the wider set of economic statistics and accounts for a particular sector, concept or activity. Hence, the value of such an indicator for forecasting lies in its embodying relevant additional information available quickly, at high frequency and with a significant lead time over the transaction being recorded in official statistics.

Whilst earlier studies used raw internet search statistics from diverse search engines, Google Labs have since developed fairly refined tools within the Google Trends/Google Insights for Search website which enable individual researchers to recover tailor-made sample statistics on the frequency of searches for specific keywords and phrases by location and on a near real-time basis, starting from 2004. The relatively restricted historical samples available pose some limitations on their general usefulness for macroeconomic modelling, as does the sampling method which inevitably varies over time, as discussed below. Nonetheless a wide range of studies have emerged, originally focussing mainly on labour market indicators, but then widening to include housing, tourism, retail sales and consumption, housing markets, inflation expectations and financial markets, and for a range of countries.

Labour Market Studies

The earliest and most numerous set of studies using internet search indicators for economic forecasting are those related to labour markets and unemployment. The pioneering study by Ettredge *et al.* (2005), predating the use of Google Trends and other Big Data sources by several years, looks at US monthly unemployment over the period 2001-2004, using an internet search indicator of job-searches from a variety of internet sources. Using a relatively simple autoregressive forecasting model, it finds a significant relationship between search variables and published US unemployment data for adult males, one superior to the alternative use of official weekly claims data. Broadly similar results are reported for monthly total unemployment for Germany by Askitas & Zimmermann

[6] *For a thorough and up-to-date description of the various Big Data sets available and their uses see also Buono* et al. *(2017).*

[7] *In this respect, this review represents a snapshot of published studies available at or around Spring 2018.*

[8] *This review does not include more recent works presented in this issue, which were not available at the time of writing.*

(2009) using Google Search statistics for the period 2004-2008, followed by Choi & Varian (2009b) for the United States, D'Amuri (2009) for Italy, D'Amuri & Marcuccio (2009) and Tuhkuri (2015) for the United States at aggregate and state levels, Suhoy (2009) for Israel, Anvik & Gjelstad (2010) for Norway and McLaren & Shanbhogue (2011) for the United Kingdom.

Most of these studies use a similar method of adding an internet-search indicator to relatively naïve time-series autoregressive models, in level or first-differenced terms. In some cases, most notably D'Amuri & Marcuccio (2009), more sophisticated models including other economic variables and leading indicators relevant to unemployment are used. Though sensitivity to the choice of baseline model and search keywords is often noted, most of these studies find the relevant internet-search indicator to be statistically significant and to provide superior out-of-sample performance compared with naïve baseline models and in some cases other relevant indicators, for example the US Survey of Professional Forecasters.

The more recent US study by Tuhkuri (2015) is generally more thorough in the choice and sophistication of data, statistical models and estimation techniques. The overall finding is that improvements in predictive accuracy from using Google search data appear robust across different model specifications and search terms, but are generally modest compared with previous studies and limited to short-term predictions, and that the informational value of internet search data tends to be somewhat time specific.

Consumption Studies

Studies of consumption, retail sales and car sales using internet search indicators include those by Choi & Varian (2009a, 2011), Kholodilin *et al*. (2010) and Schmidt & Vosen (2011) for the United States, Chamberlin (2010) for the United Kingdom, Bortoli & Combes (2015) for France, Toth & Hajdu (2012) for Hungary, and Carriére-Swallow & Labbé (2010) for Chile. Both the methods used and the results obtained vary considerably across these studies.

Some studies follow modelling strategies similar to those used for predicting unemployment by adding relevant internet search indicators to relatively naïve baseline time series forecasting models, whilst others include search indicators in combination with other measures of consumer sentiment or broad macroeconomic activity. For the United States, Schmidt & Vosen (2011) use more fully specified reduced form economic models of consumption which include lagged income, interest rates and stock market price variables. In most cases, internet-search variables are found to be significant either in their own right or in combination with other variables, though sometimes the gains are found to be relatively small. For car sales in Chile, Carrière-Swallow & Labbé (2010) find the introduction of car brand search indicators to significantly improve goodness-of-fit and predictive performance of baseline autoregressive models and also outperform broader measures of economic activity.

The results of Schmidt & Vosen (2011) in particular tend to show the individual significance of such variables to be greatest with simple AR(1) models, as might be expected (as discussed in a later section). Using more semi-structural consumption function specifications, they are found to perform as well as or in combination with the Conference Board Indicator, and the best one-month-ahead nowcasts are given by models including the Google Indicator. An interesting by-product of this study is the finding that the Michigan Consumer Sentiment indicator appears to have no additional predictive value.

Also of interest is the later study of consumption and new car sales by Schmidt & Vosen (2012), where Google-based indicators are found to be generally useful in modelling and predicting the effects of changes in motor vehicle scrapping schemes (so called "cash for clunkers" schemes) for the United States, France, Germany and Italy over the period 2002-2009. Such a finding suggests the possibly useful role in detecting and predicting the effects of special events or structural change at times when other timely information are not available. However, the authors note that the major challenges in such circumstances often lie in the identification of significant irregular events and constructing an appropriate measure from available search data.

The more recent Insee paper, by Bortoli & Combes (2015), reviews the usefulness of Google indicators for modelling French consumption at different levels of aggregation. The overall results are mixed, and suggest that search statistics improve monthly expenditure forecasts in only a limited way and for a narrow set of goods and services (clothing, food, household durables and transport).

Other Personal Sector Studies

Other largely personal sector-specific studies have involved housing market variables, tourism and inflation expectations. For housing markets, Webb (2009) finds high correlations between searches for "foreclosure" and recorded US home foreclosures, whilst Wu & Brynjolfsson find an internet search-based housing indicator significant and strongly predictive for US house sales and prices, as well as the sales of home appliances. Hellerstein & Middeldorp (2012) find similar improvements for predicting US mortgage refinancing, though the gains are found to be insignificant beyond a lead time of one week. McLaren & Shanbhogue (2011) report relatively strong results for UK house prices, with an internet search indicator out-performing other indicators over the period 2004-2011.

With regard to tourism and travel, Choi & Varian (2011) report significant results for Hong Kong tourism. Artola & Galen (2012) find similar results when adding Google based indicators to ARIMA models of the UK demand for holidays in Spain, although they also report considerable sensitivity to the choice of both baseline model and search keywords, particularly when used in different languages. Examining a range of inflation expectations indicators Guzmán (2011) finds that higher frequency Google-based indicators to generally outperform lower frequency traditional measures in use.

Financial Markets

A considerable number of studies have examined the relevance for search-based indicators for financial markets, though not specifically in a forecasting context. For example, Andrade *et al.* (2009) use such measures in identifying market volatility bubbles in the run up to the 2007 Chinese stock market bubble, Vlastakis & Markellos (2010) show strong correlations between search volume data by company name and trading volumes and excess stock returns for the 30 largest companies traded on the New York Stock Exchange.

Da *et al.* (2010, 2011) find similar correlations between product search variables and revenue surprises and investor attention for 3000 US companies, whilst Preis *et al.* (2012) find strong correlations between name searches and transactions volumes for the S&P 500 companies. Dimpf & Jank (2012) also report strong co-movements between Google company name searches (as a measure of investor attention) and US stock market movements and volatility, with Google search indicators providing better out-of-sample forecasts than ARIMA models. Hellerstein & Middeldorp (2012) find a Google search indicator to be significant in modelling movements in certain dollar-Renminbi forward market variables, but with low predictive power.

Overall, the lack of firm evidence or forecasting applications in the financial market area is perhaps of less importance given the wider availability of high frequency statistics for financial market variables.[9]

Wider Macroeconomic Studies

In contrast to the previous studies, where internet search-based indicators are included directly as explanatory variables in regression models for individual economic variables, Koop & Onorante (2013) use a different approach by introducing Google search-based probability measures into a dynamic model switching (DMS) nowcasting system, one in which current outcomes are regressed on lagged values of the set of dependent variables and Google indicators. That is, instead of using internet search volumes as simple regressors, they also allow them to determine the weights given to alternative nowcasting equation estimates over time. The intuition here is that internet search information may provide the researcher with useful information about which macroeconomic variables are most important to economic agents concerns and expectations at given points in time. This would make sense, for example, in a context where the underlying economic structure is not constant, and are therefore particularly suited to deal with sudden unexpected events like financial crises or recession.

Applying this method to models for monthly US data for a selection of monthly macroeconomic variables (including inflation, industrial production, unemployment, oil prices, money supply and other financial indicators), they find dynamic switching models to be generally superior to others, regardless of whether these models involve search-based probabilities

[9] *This contrasts with studies of financial markets based on social media indicators, as described in a later section, where high frequency forecasting is a very specific focus of interest.*

or not. Firstly, the inclusion of search data is found in many cases to give improvements in nowcast performance, complementing the existing literature by showing that internet search variables are not only useful when dealing with specific disaggregate variables, but can be used to improve nowcasting of broad macroeconomic aggregates. Secondly, they also find that information from search variables is often best included in the form of model probabilities as opposed to simple regressors. The general results are however somewhat mixed across variables, being most positive for inflation, wage, price and financial variables, inconclusive for industrial production and strongly inferior for unemployment.

Limitations of the Use of Internet Search-Based Indicators

Although broadly supportive of the general usefulness of internet search-based measures for short term assessment and nowcasting for a variety of economic variables, many of the above studies note that results tend to be mixed across topics and subject to a number of specific limitations and possible biases, reflecting both the qualitative nature of the data sets and the modelling frameworks in use.

The Data Sets

Firstly, it should be noted that the various measures do not specifically correspond to the absolute number of searches but rather the proportion of searches carried out on a particular subsample using specified keywords or topics over a particular time period, suitably scaled. For this reason the data sets used often need to be "cleaned" subjectively for specific outliers, one-off events or aberrant search terms which might otherwise swamp the data.[10] At the same time, by their very nature, high frequency internet search-based indicators draw on a variable and non-stratified sample, one which changes continuously over time. Both of these factors are likely to add noise to the underlying measures and make them more qualitative than they seem at first sight. Indeed in many cases the qualitative nature of internet search statistics begs the question of the general nature of the underlying relationship e.g. with regard to the scale, linearity or even the sign.[11] Secondly, the shortness of the available samples for internet search information dating from the mid-2000s limits the scope for the stability and testing within a range of existing models, both statistical and structural.[12] Most studies therefore rely on relatively short samples of high-frequency data which are also sometimes subject to strong seasonality, which risks swamping the underlying relationships. At least visually, this seems to be the case for a number of early studies claiming to illustrate close historical relationships between the search indicator and variable in question.

A number of studies also note the sensitivity of results to the choice of keywords and baseline models.[13] The former is necessarily a handicap which implies the need for care in the construction of an indicator targeted for a specific use. Much is left to the individual researcher, to design/construct their own indicators – which has considerable advantages for use in specialist areas – but to date there appear to be no standardised published measures available for specific purposes such as general macroeconomic surveillance or analysis at national or international levels.

The Modelling Frameworks

Concerning sensitivity to the choice of baseline models, it is worth noting that, with few exceptions[14], studies reporting high significance or superior out-of-sample forecasts often do so by comparison with relatively naïve low-order univariate autoregressive time-series baseline models. These results are probably not surprising then, to the extent that without additional information such models are seldom able to provide more than smooth short-term projections adjusting recent out-turns to longer term trends and hence fail to pick up erratic short-term movements or turning points.

Relatively few studies appear to have been done to systematically test or embed internet search-based variables within existing indicator model frameworks used to forecast

[10] For example, the death of Michael Jackson in June 2009 resulted in a huge surge in internet search activity, with a major negative effect on the relative shares of searches for all other topics in that period.

[11] For example, the intensity of search activity for a range of economic variables might be associated with both positive and negative movements in the variable in question and may be time or episode specific.

[12] For example, see the comments of Chamberlin (2010), Schmidt & Vosen (2012) and Bortoli & Combes (2015).

[13] For example, see the comments of Artola & Galen (2012), Askitas & Zimmermann(2015), Chamberlin (2010) and Tkacz (2013).

[14] Notable exceptions here include D'Amuri (2009), D'Amuri & Marcuccio (2009), Schmidt & Vosen (2011).

near-term movements or turning points in key GDP or trade aggregates, or to augment or predict other high-frequency indicators significantly ahead of publication. Important exceptions here are found in the work of Koop & Onorante (2013) in combining search information with dynamic probability switching models, and Galbraith & Tkacz (2015) in the testing and use of internet search variables within more extensive indicator systems.

A relatively small subset of studies does however successfully use search-based indicators to augment and improve more conventional economic and/or indicator-based models or to allow for special factors in specific relationships at macro and sectoral levels. Although much of the literature also aims to improve the detection of turning points, very little seems to have been done to systematically test or embed internet search-based variables within existing indicator and bridge-model frameworks used to forecast near-term movements in key GDP or trade aggregates, or to augment/predict other high frequency indicators significantly ahead of publication. Further work in all the above areas would seem necessary to exploit the key advantages of internet search-based indicators over other indicators, as the relevant data sets are extended and improved over time.

The Use of "Social Media" and Twitter Based Information in Macroeconomic Modelling

In many respects social media data sets, such as those embodied in Twitter and other user-based blogs, are potentially richer and therefore have important advantages over indicators based on internet search frequencies:

- Sample sizes are often considerably larger and available on a virtually continuous basis;

- The data are more varied in scope, with greater general and specific detail of posts;

- They permit a more stratified approach, by analysing information coming from selected representative samples or well-defined user groups;

- The absence of pre-preparation/filtering by data proprietors, as with Google Trends, may be an advantage or disadvantage.

Social media blog entries and Tweets can be about any topic, being totally up to the user what they choose to broadcast. For the most part they are publicly available either directly in raw form or indirectly through social media Application Programming Interfaces (API's). This makes them an increasingly accessible and popular source of information for researchers to construct general and specific mood or intentions indicators at a given place and time, and for particular topics of interest.

Financial Markets

To date, the large majority of published empirical studies using social media data as input to economic models and forecasting,[15] are relatively near-term and in the area of stock market prices and finance. Gilbert & Karahalios (2010) for example use a dataset of over 20 million LiveJournal posts, to construct a measure of public anxiety (the Anxiety Index). This is based on a panel of 13 thousand LiveJournal contributors, chosen by linguistic classifiers on the basis of entries for 2004, as a sub-sample known to frequently express varying degrees of general anxiety (not specifically economic events). This sub-sample is then used to construct the Anxiety Index based on their daily blog posts through 2008 and tests carried out for its possible "influence" on the S&P500 stock market index, using a baseline statistical relationship involving lagged index values and the lagged levels and changes in the volume of transactions.

Using a combination of regression and Granger causality tests, the broad conclusion is that the Anxiety Index contains statistically significant information not apparent from the market data. The authors note however that this result is somewhat weakened by further testing for the inclusion of the existing Chicago Board Options Exchange VIX index[16], which in some models tends to dominate the Anxiety Index. Even so, general collinearity with the VIX is seen as a possible validation of the usefulness of the more broadly based Anxiety Index as a measure of stock market uncertainty. The authors nonetheless note that more work

[15] *Previous applications based on social media and so-called mood indicators cover a fairly wide range of topics including ; book sales (Gruhl* et al.*, 2005); cinema box office receipts (Mishne & Glance, 2005 and Liu* et al.*, 2007); influenza pandemics (Ritterman* et al.*, 2009); TV ratings (Wakamiya* et al.*, 2011); and election results (O'Connor* et al.*, 2010; Tumasjan* et al.*, 2010).*

[16] *The VIX index is a popular measure of the stock market's expectation of volatility implied by S&P 500 index options, calculated and published by the Chicago Board Options Exchange (CBOE), colloquially referred to as the* fear index *or the* fear gauge*, see Brenner & Galai (1989).*

needs to be done in overcoming the inherent difficulties in interpreting blog-based information and potential ambiguities, as well as potential index volatility associated with non-economic external events and, importantly, that the sample year 2008 was exceptional in many respects.

A number of parallel studies have looked only at correlations between the social media-based mood indicators and relevant economic variables, rather than in forecasting models. For example, Zhang *et al.* (2010) examine a very large sample of daily Twitter entries between March and September 2009 to estimate a variety of measures of differing degrees of positive and negative moods, ranging from fear to hope. These are then correlated against corresponding values of the Dow Jones, NASDAQ and S&P500 indices, as well as the VIX index. Statistically significant correlations are found, consistent with negative impacts of lagged mood indicators on current stock market prices and the VIX. However, the authors note that this holds for both positive and negative mood indicators, indicative of the relative importance of emotional outbursts as opposed to specific mood directions over the sample period.

Along similar but more formal lines, Bollen *et al.* (2011), examine the relationship between mood indicators derived from large-scale Twitter feeds and changes in the Dow Jones index over time.[17] Specifically the text content of daily Twitter feeds are analysed using two mood tracking tools from March to December 19, 2008. The first tool, OpinionFinder, analyses the text content of tweets to give a daily time series of the positive vs. negative balance of the public mood. The second tool, the Google-Profile of Mood States (GPOMS), analyses text content to provide a more detailed view of changes in public sentiment using six different mood states (Calm, Alert, Sure, Vital, Kind, and Happy). The resulting indicators are then correlated against the Dow Jones index on a daily basis, using a general autoregressive model and Granger causality testing framework. The authors conclude that results support the view that the accuracy of stock market prediction models is significantly improved (by around 6%) when some but not all mood dimensions are included.[18] In particular, variations along the public mood dimensions of "Calm" and "Happy" as measured by GPOMS appear to have some predictive value, but not the overall balance of optimism and pessimism as measured by OpinionFinder.

Following up on this work, Mao *et al.* (2012), focus more closely on the relevance of finance-specific Twitter information as opposed to general positive and negative mood expressions. Specifically they examine the relationship between the daily number of tweets that mention S&P500 stocks and associated stock prices and traded volumes at the aggregate level, for each of 10 industry sectors and at the individual company level, for Apple Inc. This is done through correlations between daily stock market measures over an approximate 3 month period (February to May 2012) and the Twitter volume indicators. The analysis is then extended using simple linear autoregressive regression models, to predict the stock market indicators with the Twitter data indicator as an exogenous input. The overall results are fairly mixed and vary between different levels of aggregation.

Significant correlations are found at the aggregate level between the Twitter indicator and both levels and changes in prices, though not trading volumes. For 8 out of 10 industry and corporate sectors (notable exclusions being consumer discretionary and stable categories), statistically significant correlations are found for the levels of traded volumes but not prices. For the financial sector and Apple Inc. (the most highly tweeted categories) correlations are statistically significant for both volumes and prices. These results are broadly mirrored in the tests for predictive accuracy with the Twitter indicator improving forecasts for volumes and prices at the aggregate and financial sectors but for volumes only for Apple Inc. Even so, the predictions of directional changes in the sample period are at best 68% accurate for the aggregate and financial sectors and only 52% for Apple Inc. i.e. close to a random walk. The authors conclude that the relevant correlations are statistically significant and help predict some stock market movements on a daily basis, although more work is required to refine the choice of search words, to screen for spurious tweets, to collecting longer-term data, and to combine indicators for the number, relevance and sentiments of individual tweets.

[17] *For a similar but more micro and higher frequency approach see Wolfram (2010).*

[18] *As discussed further below, this "landmark" result is hotly disputed by other authors, see Lachanski & Pav (2017).*

Noting the scope for measurement and classification errors associated with computational machine learning-based processing of blog-related data sources, the further work of Mao *et al.* (2014) focus instead on a simpler set of indicators, based on the frequency of use of terms related to market "bullishness" or "bearishness" in both Twitter posts and Google search queries.[19] These are calculated on daily (for Twitter) over the period 2010 to 2012, and weekly (for Google Trends) bases over the period 2007 to 2012, and then compared with other investor sentiment indicators. Relative predictive powers are then analysed in the context of small dynamic models of the US, UK Canadian and Chinese stock market prices and returns.[20] Comparing between measures and adjusting for frequencies, Twitter-based measures of market bullishness are found to lead and "predict" changes in corresponding Google-based measures, whilst both measures are found to be positively correlated with, and lead established investor sentiment surveys for the United States.[21]

Using a fairly detailed dynamic VAR modelling framework for the United States (one also including trading volumes and other sentiments indicators as explanatory variables), the Twitter-based indicator is found to be both statistically significant and provide better predictions of stock returns on a daily basis. An additional feature is that high levels of Twitter bullishness are found to be associated with changes in daily stock returns over the following days, with there being a reversion to normal levels within the next two to five days. The corresponding Google-based indicator is also found to be statistically significant but with lower predictive power, attributed to its low frequency and lack of relevant dynamics. Similar correlation results are also found for the UK, Canada and China, (within simpler bi-variate models), but with lower predictive power for China. The Google indicator is also found to be significantly correlated with all four stock market prices but with lower predictive power. The authors note that the overall results are promising in terms of predictive correlation but are less clear with regard to causality, which remains a challenging research problem for Big Data analysis and the development of appropriate experimental design methods and machine learning algorithms.

Other notable contributions to the finance literature using Twitter-based indices include Arias *et al.* (2012), which applies complex decision tree computer algorithms to Twitter based information to analyse movie box office sales and stock market prices, and Ranco *et al.* (2015), which examines the impact of Twitter-based measures of so-called "event study" effects on the stock returns of 30 leading companies within the Dow Jones index between 2013 and 2014. On a more detailed basis, Bartov *et al.* (2015), covering 300 companies over the period 2009 to 2012, examine whether aggregate opinion in individual tweets about a firm can help predict the firm's earnings and stock returns around earnings announcements, and whether the ability to predict abnormal returns is greater for firms in weaker information environments.

The Twitter-based literature, emanating initially from computational information science and artificial intelligence studies, is not without critics within the economics and finance world. Indeed, a recent review by Lachanski & Pav (2017) strongly criticises both the general approach and the results of Bollen *et al.* (2011), which they consider incompatible with both information theory and the investor-sentiment based text mining. Attempting to replicate similar mood indicators and models, they find some in-sample but almost no out-of-sample correlation with the Dow Jones index. Whilst this might be attributable to minor differences in data coverage and the selection of the time period studied, they conclude that Bollen's results are very much an outlier and that there is little or no credible evidence that Twitter-based measures of general collective moods can be used to forecast index activity on a daily basis. Overall, they argue that the Bollen *et al.* (2011) study is fundamentally flawed and has contributed to a "growing deadweight loss to the finance literature".

Labour Markets

To date, there appear to be relatively few published Twitter-related economic studies outside the area of financial markets. An important exception has been the work on labour market's by Antenucci *et al.* (2014) at the University of Michigan, in developing measures of labour market flows from social media data. Specifically

[19] *A specific advantage of such an approach is that these terms are used fairly unambiguously and in a focussed way to refer to financial market conditions.*

[20] *Specifically for the US, they examine a number of market indicators including the Dow-Jones and S&P indices; for the UK, the FTSE100; for Canada the S&P/TSX; and for China the SSE index.*

[21] *These include the Daily Sentiment Index and the US Advisors' Sentiment Report of Investors Intelligence.*

large sample Twitter-based data were used to produce indexes of job loss, job search and job posting as a means of analyzing high frequency weekly estimates of job flows from July 2011 to early November 2013. Measures are first derived from the frequency of use of job loss and search-related phrases in the sample of Tweets, which are then combined into composite measures using their principal components to track initial claims for unemployment insurance at medium and high frequencies. The resulting index is found to have a greater signal to noise ratio than initial claims data which might be of value to policy makers needing high-frequency, real-time indicators. Over the sample period, the indicator is found to account for 15 to 20 percent of the variance of the prediction error of the consensus forecast for initial claims. The index was also considered useful in providing realtime indicators of events such as Hurricane Sandy and the 2013 government shutdown, although this body of work is currently said to be under revision since the original model began to deviate in its estimates around mid-2014.

The Limitations in the Use of Social Media in Forecasting Studies to Date

Overall, the challenges in the extraction and use of social media based data sets are considerable and perhaps greater than those involved with internet search material. Typically the researcher has to devise methods of searching across large sets of blog entries to identify within a given sample and timeframe the frequency of the use of specific phrases or keywords by those posting on blogs. For example, this might include looking for phrases indicating concepts like job security and job loss, company and consumer product names or those used more generally to indicate degrees of anxiety or confidence with respect to life in general or more specifically economic and financial conditions. Thus whilst being richer in content, they are also, arguably, more exposed to differences in linguistics, interpretation and nuances in the use of language, than for internet search related data.

For these reasons, and given the huge volume of data being processed, much of the work in this area builds on developments in the informatics, machine learning and artificial intelligence domain, for the design and application of sophisticated automated filters to mine the information content of simple text blog entries. Indeed it is notable that much of the original literature originates in the study of computational, linguistic and machine learning methods as opposed to economics and finance. For this reason, these studies are not always embedded in the sound and familiar theoretical and empirical frameworks more commonly used in other areas of economic research and econometrics. Although these studies may often embody “state of the art” computational machine learning techniques, there is relatively little evidence of testing of one measure or method against another to see whether all the “bells and whistles” are superior to simpler frequency balance measures.

Similarly, there is certain sense of searching for a “Holy Grail” financial market indicator which is both broadly based and able to explain, predict or, at best, correlate with chosen financial variables. Possibly because of the enormous sample sizes involved with raw high frequency data, the chosen time samples often seem to be idiosyncratic and restrictively short, as noted by Lachanski & Pav (2017). In this context, the more recent work of Mao *et al.* (2015) focussing on simpler variables that are more narrowly defined to be of relevance to financial markets over a longer sample period and comparing between measures may be more rewarding. Even so, there is often an excessive focus on very (daily) near-term predictive power and in having more detailed and workable model for US stock prices, as opposed to those for other economies. Both factors clearly limit their overall relevance for macroeconomic analysis as opposed to profit-driven trading applications.

Similar to the body of research based on internet search variables, the models used in many social media-based studies are almost exclusively statistical and, in the absence of other explanatory variables, may be too simple to tell much about the underlying dynamics or relative predictive values of the different indicators being analysed. In addition, a surprising and perhaps important omission in these studies is the fact that financial markets are inherently international and therefore linked to each other and influenced by other global phenomena.

The Structure and Uses of Other Big Data: Electronic Transactions and Confidence Indicators

To the extent that a large and growing share of global financial and commercial transactions are supported by electronic payments and transactions systems, there has also been growing interest in the use of high frequency statistics from a number of such sources as indicators within informal and formal forecasting and assessment frameworks. Typically, these systems cover a range of different detail and frequencies, down to the individual transaction level. As a result, confidentiality and proprietorial ownership rights pose important limitations on their uses beyond the privileged few.

SWIFT Trade Transactions Indicators

In this context, recent ICC Global Surveys of Trade and Finance and recent EBRD blog reports draw particular attention to the use of SWIFT indicators in tracking trade credit and the volume trade transactions.[22] [23] Whilst making a number of important caveats about the form and coverage of this type of data, both reports provide useful illustrations of the sharp year-on-year decline in SWIFT trade-related messages (accounting for a significant share of trade letters of credit) from end-2008 to end-2009 and later in early 2011 and their relationship with global and regional trends in trade over the same periods.

On a more country-specific basis, a recent Australian Reserve Bank study[24] examines the possible use of various electronic indicators of wholesale and retail payments from commercial banks in forecasting a range of macroeconomic aggregates including consumption, domestic demand and GDP. The overall results are mixed in finding that a SWIFT payments indicator used in combination with the principal components of other more conventional short-term macroeconomic indicators, significantly improves short-term predictive performance relative to naïve autoregressive baseline models, whilst other retail payments indicators including credit card transactions do less well.

Following the same general idea, SWIFT, in collaboration with CORE Louvain, has constructed a number of global and regional indicators for use in specific nowcasting applications.[25] In particular, SWIFT (2012) reports the use of an OECD aggregate index of filtered transactions in a suite of GDP bridge models, finding the most significant results using a dynamic mixed frequency forecasting model for quarterly movements in OECD real GDP for the period 2000 to 2011. As with most of the internet search indicator-based studies, the underlying baseline model is a relatively simple statistical ARMA model, taking account of no other relevant information.

An important caveat to these studies is that SWIFT indicators generally relate to the volumes of messages as opposed to the levels or values of transactions and therefore need to be carefully filtered for message versus transactions content and for their coverage, as between trade, financial and other activity-related transactions. Nonetheless, the broad results to date are generally supportive of the broad approach, and having the advantage of being available for a longer sample period, merit further investigation within a wider range of indicators and economic aggregates.

Payments Transactions Statistics

The recent Bank of Canada study by Galbraith & Tkacz (2015) reports an interesting approach combining a range of financial and transactions indicators within a set of mixed frequency GDP indicator models. These models combine measures of the growth in values and volumes of monthly and quarterly Canadian debit, credit and cheque transactions cleared through the Canadian Payments Association (CPA) on a daily basis, with composite leading indicators for the US and Canada, monthly unemployment rates and lagged GDP growth. A key finding is an improvement in the accuracy of the earliest GDP predictions through the inclusion of debit card payments observed for the first two months of the prediction period, although such improvements are not detectable once the previous quarter's GDP value is observed. Overall, this supports the possible value of combining electronic transactions with other data measurable on a daily basis. An obvious limitation to the use of this class of information is its confidentiality and inaccessibility to the general public for research uses, even in processed form.

[22] *In particular, see the Global and Regional trends sections of the ICC Reports "Global Survey of Trade and Finance: Rethinking Trade Finance", for 2010, 2011 and 2012, and the EBRD blogs "Trade Finance on the Way to Recovery in the EBRD Region", January 2011 and "Rising uncertainty for trade finance as IFI additionality increases", February 2012.*

[23] *The Society of Worldwide Interbank Financial Telecommunication (SWIFT) network covers the financial transactions of over 10,000 financial institutions and businesses worldwide (210 countries).*

[24] *See Gill* et al. *(2012).*

[25] *In particularly see "The SWIFT Index: Technical Description", SWIFT, February 2012.*

ADP Employment Indicators

A further example of the use of real-time transactions systems data is given by the work reported in the ADP National Employment Report (2012) for the United States.[26] In this study monthly and bi-weekly payroll data processed by the ADP's system (responsible for the payrolls of establishments covering approximately 20% of U.S. private sector workers) are filtered and classified by size and industry to provide pair-wise matches with the sample used in producing BLS monthly employment data. A set of adjusted sectoral ADP indicators are then used, in conjunction with the Philadelphia Federal Reserve ADS Business Conditions Index[27], to estimate a system of VAR equations to predict monthly changes in BLS private employment data by sector, since April 2001. Although the significance of individual variables is not reported and it is unclear what restrictions are placed across individual parameters/sectoral contributions, the overall in-sample correlations appear to be relatively high (0.83 to 0.95) and the models appear to track overall monthly movements in BLS employment for the total private sector and 5 broad sectors fairly closely.

The Ceridian-UCLA Pulse of Commerce Index

Another "Big Data" indicator of interest for high frequency analysis for United States activity is the Ceridian-UCLA Anderson Pulse of Commerce Index (PCI). Essentially this index is based on Ceridian electronic card payment services for US diesel sales to freight haulage companies. In principle, the transactions data can be tracked and analysed on a yearly, monthly, weekly and daily basis by location and volumes of fuel purchases to provide a detailed high-frequency picture of US road trucking activities including interstate highways and cities, shipping ports, manufacturing centres and border crossings with Canada and Mexico. The PCI's main advantage over other economic indicators is its basis on real-time, actual fuel consumption data in advance of published monthly statistics. Its main disadvantage for economic research is that it is not freely available within the public domain. To date no published analytical studies appear to be available, although UCLA Anderson produce a monthly newsletter 4 to 5 days in advance of the publication of monthly industrial production data and reports that back-testing to 1999 shows the index to closely match growth in real GDP and changes in Industrial Production.

* *
*

The forecasting experience of many national and international forecasters during the 2008-2009 recession and beyond have been similar in that existing models, methods and analyses were not particularly well suited to predict or analyse the scale of the crisis. This very much reflected the nature of the underlying situation, the lack of systematic evidence of the scale of the financial shock, the nature of the international linkages involved and the mechanisms by which financial shocks translated into shocks to the real economy. By contrast, short-term indicator and nowcasting models for world trade and GDP for the G7 economies have sometimes proved to be extremely useful and more accurate, within a current quarter or nowcasting situation. Even so, such models appear to have been limited or poor in going much beyond the current quarter and detecting possible turning points, reflecting the limitations of "soft" survey-based activity indicators and a lack of "hard" information. Overall, such a situation suggests a research priority in improving the timeliness and availability of relevant information.

On the basis of a review of the recent academic literature, a broad conclusion is that internet search and social media based indicators and other Big Data sources provide a novel and possibly useful means of measuring various aspects of consumer and business behaviour in an almost a real time basis. At the same time they may embody information which cannot be captured by other economic indicators or be available on such a timely basis. For this reason, they warrant further development and monitoring in parallel with other macroeconomic indicators.

The range of available empirical studies reviewed provides interesting insights and evidence of significant correlations and predictive performance across a range of topics. Overall, however, the results are generally quite mixed, reflecting both the relative simplicity of the

[26] *See "The ADP National Employment Report", Automatic Data Processing Inc. and Moody's Analytics, October 2012.*

[27] *The ADS business conditions index is based on the framework developed in Aruoba* et al. *(2009). The index takes on board a combination of high and low frequency indicators, including weekly initial jobless claims; monthly payroll employment, industrial production, personal income less transfer payments, manufacturing and trade sales, and quarterly real GDP.*

models used and important limitations in terms of quality, form, sample sizes and their "qualitative" nature. In these respects more needs to be done to:

- Refine and improve the quality standards for available Big Data sets and their accessibility;

- Develop better methods for extracting relevant economic information relevant to specific fields of economic research;

- Improve the means of comparing and testing between alternative measures;

- Further adapt and improve relevant testing and modelling frameworks, to be more useful to the task of incorporating near-term information in short-term macroeconomic forecasts.

Nonetheless, there are some clear examples where such indicators could usefully augment existing nowcasting and other indicator-based approaches as part of the general selection of variables to be analysed. Now that the first wave of such studies seems to have become less prominent in the literature it would be useful to take advantage of some of these lessons in the design of ongoing future work, rather than see the general topic dismissed as a fad or dead end.

Big Data sources including transactions based and other financial indicators have been more thinly used, until very recently. The results obtained to date also seem rather mixed, although trade finance indicators do appear to have given the right signals prior to and during the financial crisis. They also show some promising features but are equally limited in terms of information content and transparency. As for internet search and social media-based studies, they warrant further investigation in the context of statistical and semi-structural economic and indicator frameworks. In contrast to internet based indicators, this class of data is, in the main, subject to wider concerns about their confidentiality and is therefore available, so far, to a relatively small audience, mostly central bankers and statisticians. Important priorities in this area are therefore to develop suitable quality standards and to improve their accessibility for statisticians and economic researchers in a suitably relevant and condensed form.

The overall message is that Big Data sets provide new and useful sources of information for economic analysis, but also warrant further refinement, development and monitoring in parallel with other macroeconomic indicators and forecasting techniques. As such they are a welcome addition to the economist's and statistician's toolkit for short-term analysis. □

BIBLIOGRAPHY

ADP & Moody's Analytics Enhance (2012). *ADP National Employment Report.*
http://mediacenter.adp.com/news-releases/news-release-details/adp-and-moodys-analytics-enhance-adp-national-employment-report/

Andrade, S. C., Bian, J. & Burch, T. R. (2009). Does information dissemination mitigate bubbles? The role of analyst coverage in China. *University of Miami Working Paper.*

Andrade, S. C., Bian, J. & Burch, T. R. (forthcoming). Analyst Coverage, Information, and Bubbles. *The Journal of Finance and Quantitative Analysis*, 48(5), 1573–1605.
https://doi.org/10.1017/S0022109013000562

Antenucci, D., Cafarella, M., Levenstein, C., Ré, C. & Shapiro, M. (2014). Using Social Media to Measure Labour Market Flows. University of Michigan, *NBER Working paper* N° 20010.
https://doi.org/10.3386/w20010

Anvik, C. & Gjelstad, K. (2010). "Just Google It!"; Forecasting Norwegian unemployment figures with web queries. CREAM Publication N° 11.
http://hdl.handle.net/11250/95460

Arias, M., Arratia, A. & Xuriguera, R. (2014). Forecasting with Twitter Data. In: *ACM Transactions on Intelligent Systems and Technology*, 5(1), 1–24.
https://doi.org/10.1145/2542182.2542190

Armah, N. (2013). Big Data Analysis: The Next Frontier. *Bank of Canada Review*, Summer 2013, 32–39.
https://www.bankofcanada.ca/wp-content/uploads/2013/08/boc-review-summer13-armah.pdf

Artola, C. & Galen, E. (2012). Tracking the Future on the Web: Construction of leading indicators using Internet searches. *Bank of Spain Occasional Paper* N° 1203.
https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/12/Fich/do1203e.pdf

Aruoba, S. B., Diebold, F. X. & Scotti, C. (2009). Real-Time Measurement of Business Conditions. *Journal of Business and Economic Statistics,* 27(4), 417–427.
https://doi.org/10.1198/jbes.2009.07205

Askitas, N. & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly,* 55(2), 107–120.
https://doi.org/10.3790/aeq.55.2.107

Bartov, E., Faurel, L. & Mohanram, P. (2015). Can Twitter Help Predict Firm-Level Earnings and Stock Returns? Rotman School of Mnagaemnet, *Working Paper* N° 2631421, July 2015.
https://dx.doi.org/10.2139/ssrn.2782236

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. & Tambalotti, A. (2017). Macroeconomic Nowcasting and Forecasting with Big Data. New York Federal Reserve *Staff Report* N° 830, November 2017.
https://www.newyorkfed.org/research/staff_reports/sr830

Bollen, J., Mao, H. & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
https://doi.org/10.1016/j.jocs.2010.12.007

Bortoli, C. & Combes, S. (2015). Contribution from Google Trends for forecasting the short-term economic outlook in France: limited avenues. Insee, *Conjoncture de la France*.
https://www.insee.fr/en/statistiques/1408911?sommaire=1408916

Brenner, M. & Galai, D. (1989). New Financial Instruments for Hedging Changes in Volatility. *Financial Analysts Journal*, 45(4), 65–71.
https://www.jstor.org/stable/4479241

Buono D., Mazzi, G. L., Kapetanios, G., Marcellino, M. & Papailas, F. (2017). Big data types for macroeconomic nowcasting. *Eurostat Review on National Accounts and Macroeconomic Indicators*, 1/2017, 93–145.
https://ec.europa.eu/eurostat/cros/system/files/euronaissue1-2017-art4.pdf

Burns, A. F. & Mitchell, W. C. (1946). Measuring Business Cycles. NBER Book Series, *Studies in Business Cycles* N° 2.
https://www.nber.org/books/burn46-1

Carrière-Swallow, Y. & Labbé, J. (2010). Nowcasting with Google Trends in an Emerging Market. *Bank of Chile Working Paper* N° 588. Reprinted (2013) in: *Journal of Forecasting*, 32(4), 289–298.
https://doi.org/10.1002/for.1252

Chamberlin, G. (2010). Googling the present. *Economic and Labour Market Review*, 4(12), 59–95.
https://doi.org/10.1057/elmr.2010.166

Choi, H. & Varian, H. (2009a). Predicting the present with Google Trends. Google, *Technical report*, April 2009.
http://dx.doi.org/10.2139/ssrn.1659302

Choi, H. (2009b). Predicting Initial Claims for Unemployment Benefits. Google, *Technical report*, July 2009.
https://ssrn.com/abstract=1659307

Choi, H. & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88, 2–9.
http://dx.doi.org/10.1111/j.1475-4932.2012.00809.x

Cousin, G. & Hillaireau, F. (2018). En attente du titre. *Economie et Statistique / Economics and Statistics* (this issue)

Da, Z., Engelberg, J. & Gao, P. (2010). In Search of Earnings Predictability. University of Notre Dame and University of North Carolina at Chapel Hill, *Working Paper*.
https://pdfs.semanticscholar.org/b68e/aeeac8e5fcd42cff698c7c96dee5e357623a.pdf

Da, Z., Engelberg, J. & Ga, P. (2011). In Search of Attention. *Journal of Economic Finance*, 66(5), 1461–1499.
https://econpapers.repec.org/RePEc:bla:jfinan:v:66:y:2011:i:5:p:1461-1499

D'Amuri, F. (2009). Predicting unemployment in short samples with internet job search query data. *MPRA Paper* N° 18403.
https://econpapers.repec.org/RePEc:pra:mprapa:18403

D'Amuri, F. & Marcucci, J. (2009). "Google It!" Forecasting the US Unemployment Rate with a Google Job Search Index. *ISER Working Paper Series* N° 2009-32.
https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2009-32

Della Penna, N. & Huang, H. (2009). Constructing Consumer Sentiment Index for U.S. Using Internet Search Patterns. University of Alberta, *Working Paper* N° 2009-26.
https://ideas.repec.org/p/ris/albaec/2009_026.html

Diebold, F. X. (2000). "Big Data" Dynamic Factor Models for Macroeconomic Measurement and

Forecasting: A Discussion of the Papers by Lucrezia Reichlin and by Mark W. Watson. In: Dewatripont, M., Hansen, L. P. & Turnovsky, S. (Eds.), *Advances in Economics and Econometrics*, Eighth World Congress of the Econometric Society, pp. 115–122. Cambridge: Cambridge University Press.
https://www.sas.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF

Dimpfl, T. & Jank, S. (2012). *Can internet search queries help to predict stock market volatility?* New York: Social Science Research Network.

EBRD (2011). Trade Finance on the Way to Recovery in the EBRD Region. EBRD blog, January 2011.

EBRD (2012). Rising uncertainty for trade finance as IFI additionality increases. EBRD blog February 2012.

Ettredge, M., Gerdes J. & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the Association of Computing Machinery*, 48(11), 87–92.
https://doi.org/10.1145/1096000.1096010

Galbraith, J. W. & Tkacz, G. (2015). Nowcasting GDP with electronic payments data. ECB *Statistics Paper Series* N° 10.
https://econpapers.repec.org/RePEc:ecb:ecbsps:201510

Giannone, D., Reichlin, L. & Small, D. (2008). Nowcasting: The realtime informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4), 665–676.
https://econpapers.repec.org/RePEc:eee:moneco:v:55:y:2008:i:4:p:665-676

Gilbert, E. & Karahalios, K. (2010). Widespread Worry and the Stock Market. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.
https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1513

Gill, T., Perera, D. & Sunner, D. (2011). Electronic Indicators of Economic Activity. *Reserve Bank of Australia Bulletin,* June 2012.
https://www.rba.gov.au/publications/bulletin/2012/jun/1.html

Gruhl, D., Guha, R., Kumar, R., Novak, J. & Tomkins, A. (2005). The predictive power of online chatter. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005.
https://doi.org/10.1145/1081870.1081883

Guzmán, G. C. (2011). Internet Search Behaviour as an Economic Forecasting Tool: The Case of Inflation Expectations. *The Journal of Economic and Social Measurement,* 36(3), 119–167.
https://ssrn.com/abstract=2004598

Hassani, H. & Silva, E. (2015). Forecasting with Big Data: A Review. *Annals of Data Science*, 2(1), 5–19.
https://doi.org/10.1007/s40745-015-0029-9

Hellerstein, R. & Middeldorp, M. (2012). Forecasting with Internet Search Data. Federal Reserve Bank of New York, *Liberty Street Economics,* January 4, 2012.
https://libertystreeteconomics.newyorkfed.org/2012/01/forecasting-with-internet-search-data.html

ICC (2010). *Rethinking Trade Finance 2010: An ICC Global Survey*. Paris: International Chamber of Commerce.
https://iccwbo.org/publication/icc-global-report-on-trade-finance-2012/

International Institute of Forecasters' Workshop (2014). *Using Big Data for Forecasting and Statistics*. Summary of proceedings of the 11th IIF workshop, April 2014, hosted by the ECB.
https://forecasters.org/wp-content/uploads/11th-IIF-Workshop_BigData.pdf

Jansen, B. J., Ciamacca, C. C. & Spink, A. (2008). An analysis of travel information searching on the web. *Information Technology & Tourism*, 10(2), 101–108.
https://doi.org/10.3727/109830508784913121

Kholodilin, K. A., Podstawski, M. & Siliverstovs, B. (2010). Do Google Searches Help in Nowcasting Private Consumption? Real-Time Evidence for the US. DIW Berlin *Discussion Paper* N° 997.
https://dx.doi.org/10.2139/ssrn.1615453

Koop, G. & Onorante, L. (2013). *Macroeconomic Nowcasting Using Google Probabilities.*
https://www.ecb.europa.eu/events/pdf/conferences/140407/OnoranteKoop_Macroeconomic-NowcastingUsingGoogleProbabilities.pdf

Lachanski, M. & Pav, S. (2017). Shy of the Character Limit: "Twitter Mood Predicts the Stock Market" Revisited". *Econ Journal Watch*, 14(3), 302–345.
https://ideas.repec.org/a/ejw/journl/v14y2017i3p302-345.html

Lewis, C. & Pain, N. (2015). Lessons from OECD forecasts during and after the financial crisis. *OECD Journal: Economic Studies*, 5(1), 9–39.
https://doi.org/10.1787/19952856

Liu, Y., Huang, X., An, A., & Yu, X. (2007). *ARSA: a sentiment-aware model for predicting sales performance using blogs*. New York: ACM.
http://doi.org/10.1145/1277741.1277845

Mao Y., Wei, W., Wang, B. & Liu, B. (2012). Correlating S&P 500 stocks with Twitter data. *Proceedings of the 1st ACM Intl. Workshop on Hot Topics on Interdisciplinary Social Networks Research*, 69–72.
http://doi.org/10.1145/2392622.2392634

Mao, H., Counts, S & Bollen, J. (2014). Quantifying the effects of online bullishness on international financial markets. European Central Bank, *Statistics Papers Series* N° 9.
https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp9.en.pdf?177000b829d4450b007f3d3a612cab18

McLaren, N. & Shanbhogue, R. (2011). Using internet search data as economic Indicators. *Bank of England Quarterly Bulletin,* 51(2), 134–140.
https://econpapers.repec.org/RePEc:boe:qbullt:0052

Mishne, G. & Glance, N. (2005). Predicting Movie Sales from Blogger Sentiment. *Proceedings of the AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*.
https://www.microsoft.com/en-us/research/publication/predicting-movie-sales-from-blogger-sentiment/

Mourougane, A. (2006). Forecasting Monthly GDP for Canada. OECD Economics Department, *Working Papers* N° 515.
https://doi.org/10.1787/421416670553

O'Connor, B., Balasubramanyan, R., Routledge, B. & Smith, N. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceeding of the International AAAI Conference on Weblogs and Social Media*.
https://www.researchgate.net/publication/221297841_From_Tweets_to_Polls_Linking_Text_Sentiment_to_Public_Opinion_Time_Series

Pain, N., Lewis, C., Dang, T., Jin, Y. & Richardson, P. (2014). OECD Forecasts During and After the Financial Crisis A Post Mortem. OECD Economics Department, *Working Papers* N° 1107.
https://doi.org/10.1787/5jz73l1qw1s1-en

Preis, T., Reith, D. & Stanley, H. E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society* 368(1933), 5707–5719.
https://doi.org/10.1098/rsta.2010.0284

Ranco G., Aleksovski, D., Caldarell,i G., Grcar, M. & Mozeti, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns, *PLoS ONE*, 10(9), 1–21.
https://doi.org/10.1371/journal.pone.0138441

Ritterman, J., Osborne, M. & Klein, E. (2009). Using prediction markets and Twitter to predict a swine flu pandemic. *Proceedings of the 1st International Workshop on Mining Social Media*, pp. 9–17.
https://www.research.ed.ac.uk/portal/en/publications/using-prediction-markets-and-twitter-to-predict-a-swine-flu-pandemic(dcc11feb-77be-44c1-b07a-47da57aba7b8).html

Schmidt, T. & Vosen, S. (2010). Forecasting Private Consumption: Survey-based Indicators vs. Google Trends. *Ruhr Economic Papers* N°155. Also in: *Journal of Forecasting* (2011), 30(6), 565–578.
https://dx.doi.org/10.2139/ssrn.1514369

Schmidt, T. & Vosen, S. (2012). Using Internet Data to Account for Special Events in Economic Forecasting. *Ruhr Economic Papers* N° 382.

Sédillot, F. & Pain, N. (2003). Indicator Models of Real GDP Growth in Selected OECD Countries. OECD Economics Department *Working Papers* N° 364.
http://dx.doi.org/10.1787/275257320252

Suhoy, T. (2009). Query Indices and a 2008 Downturn: Israeli Data. Bank of Israel *Discussion Paper* N° 2009/06.
https://www.boi.org.il/deptdata/mehkar/papers/dp0906e.pdf

SWIFT (2012). The SWIFT index: Technical Description. Society for Worldwide Interbank Financial Telecommunication Inc.

Tkacz, G. (2013). Predicting Recessions in Real-Time: Mining Google Trends and Electronic Payments Data for Clues. *C.D. HOWE Institute commentary* N° 387.
https://ssrn.com/abstract=2321794

Toth, J, & Hajdu, M. (2012). Google as a tool for nowcasting household consumption: estimations on Hungarian data. Institute for Economic and Enterprise Research. Central European University *Research Working Paper*.
https://gvi.hu/files/researches/47/google_2012_paper_120522.pdf

Tuhkuri, J. (2015). *Big Data: Do Google Searches Predict Unemployment?* Masters thesis, University of Helsinki, 2015.
http://urn.fi/URN:NBN:fi:hulib-201703273213

Tumasjan, A., Sprenger, T., Sandner, P. & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*?, pp. 178–185.
https://www.researchgate.net/publication/215776042_Predicting_Elections_with_Twitter_What_140_Characters_Reveal_about_Political_Sentiment

Vlastakis, N., & Markellos, R. N. (2012). Information Demand and Stock Market Volatility, *Journal of Banking and Finance,* 36(6), 1808–1821.
https://econpapers.repec.org/RePEc:eee:jbfina:v:36:y:2012:i:6:p:1808-1821

Varian, H. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.
https://econpapers.repec.org/RePEc:aea:jecper:v:28:y:2014:i:2:p:3-28

Wakamiya, S., Lee, R. & Sumiya, K. (2011). Crowd-Powered TV Viewing Rates: Measuring Relevancy between Tweets and TV Programs. In: Xu, J., Yu, G., Zhou, S. & Unland, R. (Eds.), *Database Systems for Adanced Applications. DASFAA 2011. Lecture Notes in Computer Science,* vol. 6637, pp. 390–401. Berlin, Heidelberg: Springer.
https://doi.org/10.1007/978-3-642-20244-5_37

Webb, G. K. (2009). Internet Search Statistics as a Source of Business Intelligence: Searches on Foreclosure as an Estimate of Actual Home Foreclosures. *Issues in Information Systems*, X(2), 82–87.
https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1014&context=mis_pub

Wolfram, M. S. A. (2010). *Modelling the stock market using Twitter*. M.S. thesis, School of Informatics, University of Edinburgh, 2010.
http://homepages.inf.ed.ac.uk/miles/msc-projects/wolfram.pdf

Wu, L. & Brynjolfsson, E. (2009 and 2013). The future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. *SSRN papers*.
https://dx.doi.org/10.2139/ssrn.2022293

Ye, M. & Li, G. (2017). Internet big data and capital markets: a literature review. *Financial Innovation,* 3(6).
https://doi.org/10.1186/s40854-017-0056-y

Zhang, X., Fuehres, H. & Gloor, P. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear. *Procedia – Social and Behavioral Sciences*, 26, 55–62.
https://doi.org/10.1016/j.sbspro.2011.10.562

APPENDIX

SUMMARY OF RECENT STUDIES USING INTERNET SEARCH AND SOCIAL MEDIA-RELATED INDICATORS FOR MACROECONOMIC FORECASTING AND "NOWCASTING"

Authors	Sector/Topic/Country	Methods and data	Key results	Notes/comments
Andrade *et al.* (2009 and forthcoming)	Analysis of the role of analysts and information dissemination in the run up to the 2007 Chinese stock market bubble.	Correlates different measures of bubble intensity against analyst coverage as measure of information dissemination. Uses a Google search index as check on their timing and intensity.	Significant negative relation found between bubble intensity and analyst coverage. Notes strong positive correlation between the Google search index and volume of new accounts.	This study is mostly tangential to forecasting issues.
Antenucci *et al.* (2014)	University of Michigan study of Twitter based labour market indicators for the period July 2011 to early November 2013.	Estimates indices of job loss, job search and job postings using large sample Twitter as a means of analyzing high frequency weekly estimates. Combines individual measures into composite measures using their principal components to track initial claims for unemployment insurance at medium and high frequencies.	Indicator is found to account for 15 to 20 percent of the variance of the prediction error of the consensus forecast for initial claims. The index also considered useful in providing realtime indicators of events such as Hurricane Sandy and the 2013 government shutdown.	This work is currently under revision since the original model began to deviate in its estimates around mid-2014.
Anvik & Gjelstad (2010)	Forecasting monthly changes in Norwegian unemployment.	Uses Google search indicators related to job search and welfare criteria in simple ARIMA forecasting models of monthly unemployment.	Significant improvements in RMSE found by adding Google search indicators in basic models, and superior to other leading indicators.	Limited to non-economic ARIMA models. Good discussion of the practical limitations of internet search data.
Artola & Galen (2012)	Bank of Spain study of British tourism inflows to Spain.	Uses Google search indicators related to UK search for Spanish holiday destinations in simple ARIMA model of British tourist inflows.	Google search indicator is found to be significant, with improvements in predictive value sensitive to choice of baseline model.	Notes limitations to Google search indicators and sensitivity to choice of language and search criteria.
Askitas & Zimmermann (2009)	Forecasting monthly changes in German unemployment.	Uses Google search indicators in univariate error correction models.	Strong correlations found with models predicting trends and turning points.	Notes limitations in existing data sets and scope for wider use.
Bortoli & Combes (2015)	French Insee study of the use of internet search indicators in predicting consumer's expenditures at the aggregate and detailed disaggregate levels.	Introduces Google search indicators for a wide range of aggregate and disaggregate consumption items into a multivariate indicator model framework.	Finds that search indicators do not improve the forecasting of monthly aggregate household consumption. Results for certain goods (clothing, household durables, and food) and some services (transport) are more positive but generally mixed.	Includes excellent review of the strengths and limitations of internet search variables and their uses. Notes in particular concerns about the continuity and structural stability of internet search based measures.
Bollen *et al.* (2011)	Uses OpinionFinder and Google POMS over the period March to December 19, 2008 to identify 6 Twitter-based measures of mood states. Examines the relationship between mood indicators and changes in the Dow Jones index from March to December 2008.	Correlates mood indicators against the Dow Jones index within a general autoregressive model and Granger causality testing framework on a daily basis.	The broad results suggest that the prediction accuracy of the standard stock market prediction models is significantly improved when some but not all mood dimensions are included.	Notes that variations in measures of Calm and Happiness as measured by GPOMS appear to have some predictive value, but not for general happiness as measured by the OpinionFinder tool.
Carrière-Swallow & Labbé (2010)	Bank of Chile study of the sales of automotive products.	Adds a Google search indicator for the most popular car brands in Chile to simple and high order autoregressive models for year on year car sales in combination with a general indicator of economic activity.	Models including Google search indicator found to significantly out-perform both simple and more complex baseline models within and outside the sample period.	
Chamberlin (2010)	UK NSO study modelling a range of monthly UK statistics, including retail sales, home purchases, car registrations and foreign travel.	Adds Google search indicators to simple monthly first difference autoregressive models.	Results are mixed: significant for detailed expenditures and mortgage approvals but poor for total retail sales, car purchase and travel.	No out-of-sample tests done. Notes sensitivity to search query choice and seasonality.

Authors	Sector/Topic/Country	Methods and data	Key results	Notes/comments
Choi & Varian (2009a)	Google Research team study, modelling a range of different monthly US demand variables.	Adds Google search indicators to simple AR models.	Models including Google search indicators found to generally outperform baseline models, results are mixed with little or no gains for motor vehicles and housing sector.	Innovative and original study. Notes that sampling method may add noise but anticipates improvements over time.
Choi & Varian (2009b)	Forecasting US monthly unemployment benefits claims.	Adds Google search indicators to simple autoregressive models of unemployment claims.	Significant improvements in forecasting accuracy found relative to baseline model.	Results in line with other countries. Models are strictly non-economic.
Choi & Varian (2012)	Consolidates earlier studies and extends methods to include Hong Kong tourism and Australian consumer sentiment.	Adds Google search indicators to simple autoregressive models and tests for out of sample forecasting accuracy.	Significant improvements in forecasting accuracy found.	
Da *et al.* (2010)	Study of US cross-company performance and revenue surprises.	Uses Google search indicators for individual firm products to predict revenue surprises within a time series panel..	Finds significant relationship between search volumes and earnings surprises and company performance.	
Da *et al.* (2011)	Study of a large sample of US company stock performance.	Uses search frequency indicators as measure of investor attention for 3000 US companies.	Finds strong correlations though different from existing proxies for company attention.	
D'Amuri (2009)	Analysis of quarterly Italian unemployment.	Adds Google search indicators of job search enquiries to quarterly multivariate ARIMA models including industrial production and employment expectations variables.	Google search indicators found to be significant and superior to established leading indicators. Small sample results found to be better than for longer samples.	
D'Amuri & Marcuccio (2009)	Analysis of monthly US unemployment in aggregate and at state levels.	Tests the addition of Google search indicator in ARIMA models over a wide range of model forms and specifications, including other relevant US leading indicators.	Combined with Initial Claims indicators, models including Google search indicators found to outperform other models across a wide range of specifications at aggregate and most state levels.	Models including Google search indicators found to be superior to those using the Survey of Professional Forecasters.
Dimpfl & Jank (2012)	Study of daily US stock market volatility.	Uses Google search indicator by company name as measure of investor attention. Tests relationship with stock market prices and volatility within an ARIMA model framework.	Finds strong co-movements between Google searches and market movements and volatility, with search queries providing more precise in and sample prediction.	
Ettredge *et al.* (2005)	Earliest study of US monthly unemployment (2001-2004) using a range of internet search data predating Google Trends.	Constructs and correlates an Internet search based measure of job search within a simple forecasting model.	Finds significant correlation between job-search and unemployment data with significant trade-off between explanatory power and lead time. Index found to be superior to weekly initial claims data. Relationship is only significant for males.	Author strongly promotes future use of internet search statistics as a means of predicting a wider range of macroeconomic data, and proposes related study of consumer confidence.
Galbraith & Tkacz (2015)	Bank of Canada study combining a range of financial and transactions indicators within a set of mixed frequency GDP indicator models.	Models combine measures of the growth in values and volumes of monthly and quarterly Canadian debit, daily credit with composite leading indicators for the US and Canada, monthly unemployment rates and lagged GDP growth.	A key finding is the improvement in accuracy for the earliest GDP nowcasts through the inclusion of debit card payments observed for the first two months of the nowcast period, although such improvements are not detectable once the previous quarter's GDP value is observed (in month 3).	Provides overall support for the need for combining electronic transactions with other data, measured with some accuracy at a daily frequency.
Gilbert & Karahalios (2010)	Twitter based study constructing a broad Anxiety Index based on LiveJournal blog entries. Tests through 2008 data for possible influence of the index on daily changes in the Standard and Poor index (the S&P500).	Estimates a baseline statistical relationship between S&P, its lagged values, and levels and changes in the volume of transactions taking place and the VIX Fear index	Uses a combination of regression and Granger causality tests and finds a statistically significant relationship between the Anxiety Index and future stock market prices.	Notes that result is weakened by inclusion of the VIX index which tends to dominate. Notes difficulties in interpreting blog-based linguistic expressions, index volatility due to external factors and the exceptional nature of 2008.

Authors	Sector/Topic/Country	Methods and data	Key results	Notes/comments
Gill *et al.* (2011)	Reserve Bank of Australia review of the use of various electronic indicators as a means of improving information and forecasts of main Australian macro aggregates.	Use a range indicators of retail and wholesale bank transfer (SWIFT) and card transactions in AR(1) and principal components models for retail sales, consumption, domestic demand and GDP.	Results are mixed. SWIFT payments indicators are found significant in some AR models, but best in principal components models in combination with other measures. Results using retail payments indicators are less significant.	The authors suggest wider use of electronic indicators to improve the real-time measurement of economic aggregates. Suggest that such data are likely to become more useful as payments behaviour and internet use become more stable over time.
Guzmán (2011)	Study of Google search indicators as measure of real-time US CPI inflation expectations.	Tests forecasting performance for search indicators relative to 36 other indicators of inflation expectations and TIPS spreads.	Results suggest higher frequency measures outperform lower frequency measures in use, in terms of accuracy, predictive power. Out-of-sample forecasts using the Google search indicator have lowest forecast errors across the range of indicators used.	
Hellerstein & Middeldorp (2012)	New York Fed blog review of current literature on use of internet search counts in a range of modelling areas includes new work on US financial markets.	Adds Google search indicator for home and mortgage refinancing to a small dynamic model of the refinancing index, also including the influence of market yields. Adds Google search indicator to models of the Renminbi-dollar forward market variables.	Results are mixed. Google search indicator significantly improves forecast performance for mortgage refinancing, but gains are limited by insignificance of lead times. Search indicator found significant in Renminbi forward market analysis although predictive power is low.	Concludes that improvements in predictive power are not universal and do not provide explanatory power beyond more traditional methods, but nonetheless a useful addition to the economist's toolkit.
Kholodilin *et al.* (2010)	Examines usefulness of Google search indicators in nowcasting year-on-year growth in monthly US private consumption (2007-2010).	The Google search indicator-based forecasts are compared to benchmark AR(1) model and others including the consumer surveys and financial indicators.	Google search based forecasts found more accurate than for benchmark model. Similar results found for models including consumer survey and financial variables.	
Koop & Onorante (2013)	Examines the use of Google search probability variables in monthly US dynamic switching models for nine US macroeconomic variables (inflation, industrial production, unemployment, oil prices, money supply and other financial indicators).	Introduces Google search-based probability measures into a dynamic model switching (DMS) nowcasting system in which current outcomes are regressed on lagged values of the set of dependent variables and Google indicators.	Inclusion of internet search data gives improvements in many cases, but best included as model switching probabilities rather than simple regressors. General results are mixed; positive for inflation, wage, price and financial variables, less so for industrial production and inferior for unemployment.	Innovative approach combining search information with a sophisticated DMS nowcasting system.
Lachanski & Pav (2017)	Attempt to replicate Bollen *et al.* (2011) using similar Twitter based data sets methods.	Correlates mood indicators against the Dow Jones index within a general autoregressive model and Granger causality testing framework on a daily basis.	Finds some in-sample but almost no out-of-sample evidence that such a measure contains information relevant to the Dow Jones index.	Concludes that Bollen *et al.* results are an outlier and that there is little/no credible evidence that the collective mood content of raw Twitter text data from the universe of tweets can be used to forecast index activity at the daily time scale.
Mao *et al.* (2012)	Examines the relationship between Tweets mentioning the S&P 500 index and stock prices and traded volume between February and May 2012. Analysis done at the aggregate level, for each of 10 industry sectors and at the company level, for Apple Inc.	Uses simple linear autoregressive regression models, to predict the stock market indicators with the Twitter data an exogenous input.	Generally mixed results. Significant correlations at aggregate level with levels and changes in prices but not trading volumes. Significant correlations for 8 out of 10 sectors with traded volumes but not prices. Significant correlations for both volumes and prices for the financial sector and Apple Inc. Results are broadly mirrored in the tests for predictive accuracy.	Predictions of directional changes in the sample period are at best 68% accurate for the aggregate and financial sectors and only 52% for Apple Inc., close to a random walk.

Authors	Sector/Topic/Country	Methods and data	Key results	Notes/comments
Mao *et al.* (2014)	Analysis of Twitter and Google search-based indicators of "bullishness" or "bearishness" calculated on daily basis (Twitter) over the period 2010 to 2012, and weekly (for Google Trends) over the period 2007 to 2012.	Makes cross comparisons and with other investor sentiment indicators and analyses relative predictive powers in small dynamic models of for US, UK Canadian and Chinese stock market prices and returns. US model is notably more complete. Twitter-based measures found to lead changes in Google-based measures, both are positively correlated with other measures of US investor sentiment.	Twitter-based indicator statistically significant and provides better predictions of stock returns for the US. Google-based indicator also significant but with lower predictive power. Similar correlations for the UK, Canada and China, within simpler bi-variate model, but with lower predictive power for China. Google indicators are significantly correlated for stock market prices but with lower predictive power.	Notes lack of evidence with regard to causality. Notes need to develop appropriate experimental design methods and machine learning algorithms for processing Tweets and for testing causality.
McLaren & Shanbhogue (2011)	Bank of England paper examining use of Internet search data for UK labour and housing markets.	Adds Job Seekers search variable to first-differenced AR models of unemployment including other indicators and house prices, over the period 2004-2011.	Mixed results. Job Seekers indicator significant but outperformed out of sample by Claimant Counts. Stronger results for house prices, Internet search variable in AR(1) model outperforming ones based on other indicators over the period 2004-2011.	Notes limitations in the approach but concludes that search data provide additional insights not covered by business surveys. Bank to monitor search data within range of indicators in reviewing UK economic prospects.
Preis *et al.* (2010)	Examines weekly Google search data looking for possible links between search volume data and weekly US financial market fluctuations.	Complex correlation analysis of company name search and transactions volumes for S&P 500.	Finds evidence of strong correlations. Recurring patterns found using new method for quantifying complex correlations.	
Schmidt & Vosen (2010)	Examines predictive performance of Google search indicator for US private consumption.	Performance assessed relative to Michigan Consumer Sentiment and Conference Board Confidence Index in simple AR models and more conventional consumption functions including lagged income, interest rate and stock market price variables.	Google search indicator outperforms survey based indicators in simple AR models. With an extended consumption function, both Google and Conference Board indicators offer improvements, with the former useful for one month ahead predictions.	Michigan index found to have no additional value.
Schmidt & Vosen (2012)	Examines use of internet search data to predict special events when timely information is not available. Specifically it looks at car scrapping programs in four countries (France, Germany Italy and the United States).	Uses small quarterly dynamic models of changes in consumption over the period 2002 to 2009, including income and a Google search indicator, effectively entering as a shift variable during the relevant programs.	Finds the inclusion of search query data into statistical forecasting models improves the forecasting performance in almost all cases.	Notes that major challenge is to identify irregular events and finding the appropriate time series from Google search statistics.
Suhoy (2009)	Examines use of Google search indicators across a range of sectors and variables for Israel, using query categories including human resources, home appliances, travel, real estate, food and drink and beauty and personal care.	Applies Granger causality tests, first differenced linear and two-state Bayesian models to test for co-movement in indicators and growth cycles.	Labour market indicator found most predictive, improving monthly projection of changes in unemployment rates. Finds weekly frequency useful in real-time monthly monitoring, with query indices preceding official data by up to two months. Co-movements in search queries found useful in assessing economic slowdowns.	
Tkacz (2013)	Canadian study examining the use of Google search indicator for predicting recent turning points and recessions in key macroeconomic indicators.	Examines internet recession-related search indicators alongside other financial and payments variables within probit models to predict turning points in GDP and unemployment.	Finds that Google searches for "recession" and "jobs" would have predicted the 2008 recession up to three months in advance of its onset. Shortness of sample prevents analysis of other turning points.	Provides good review of the nature and limitations of search related variables, noting both advantages in their timeliness but also their qualitative nature and sensitivity to specific choices.

Authors	Sector/Topic/Country	Methods and data	Key results	Notes/comments
Toth & Hajdu (2012)	Examines use of Google search indicators to predict household consumption, retail sales and car sales in Hungary.	Constructs and tests search indicators for retail sales and car sales in simple autoregressive baseline model using monthly data for the period 2004-2011.	A combination of Google variables are found to be significant when used in combination with autoregressive terms. Similar results are obtained for quarterly consumption, though with smaller sample size.	Highly seasonal data set.
Tuhkuri (2015)	PhD thesis study of the use of internet search data in predicting US unemployment at economy wide and state levels.	Introduces internet search frequency information on unemployment benefits into a range of AR benchmark and panel data models at the state level.	Improvements in predictive accuracy using Google data appear robust to different model specifications and search terms, but are generally modest and limited to short-term predictions. The informational value of internet search data also tends to be time specific.	Provides an excellent review of the literature and thorough insights into a variety of tests including causality and stability.
Vlastakis & Markellos (2012)	Study of information demand at market and firm level using data for the largest 30 stocks traded on the NYSE.	Proxies demand by weekly internet Google Trends search volume data by company name.	Results suggest significant relationship with individual stock trading volumes and the conditional variance of excess stock returns. Significance of search indicators diminishes using implied rather than historical measures of volatility at the firm and market levels.	Study confirms theoretical proposition that information demand is positively related to risk aversion.
Webb (2009)	Examines the relationship between Google searches on the keyword “foreclosure” and actual U.S. home foreclosures over the period 2004-2008.	Uses bi-variate correlation and regression analysis.	Finds a high correlation between the two variables providing a reasonably accurate estimate of trends in actual U.S. home foreclosures.	
Wolfram (2010)	Applies Natural Programming Language sampling methods to very high frequency Twitter feeds over a 10 day period in 2010, to predict hourly and daily movements in the individual stock prices for Apple, Google, Intel and other selected stock prices.	Uses automated Support Vector Regression (SVR) methods to model and simulate stock price movements over the very near-term.	Model based predictions were found to be close to baseline for Apple and Google stocks over a very short (15 minute) period, but become unstable as the forecast distance increases (to 30 minutes). Concludes that that relevant information can be extracted to give small but significant advantages in predicting market prices.	Notes the need to improve sampling by more clearly identifying influential users and creating rules specific to the Twitter dataset for focussing more specifically on the topic of financial markets.
Wu & Brynjolfsson (2009 and 2013)	Seminal paper examining the use of Internet search data to predict US housing market trends and sales of house appliances in 2008-2009.	Relevant search indicators are constructed and then introduced into quarterly dynamic joint autoregressive models for house purchases and prices at the state level, including fixed effects variables.	Housing search index found to be significant and strongly predictive of both future housing market sales and prices compared with an underlying baseline model. Out-of-sample predictions and mean absolute errors significantly smaller than baseline model. Similar results found for home appliance sales.	
Zhang *et al.* (2011)	Examine large sample of daily Twitter entries between March and September 2009. Estimates a variety of measures of differing degrees of positive and negative moods, ranging from fear to hope.	Correlates these against corresponding values of the Dow Jones, NASDAQ and S&P500 indices, as well as the VIX index.	Finds statistically significant correlations, consistent with negative impacts of lagged mood indicators on current stock market prices and the VIX.	Notes that the result holds for positive and negative mood indicators, suggesting the relative importance of emotional outbursts as opposed to the specific mood indicator during the sample period.

Can Mobile Phone Data Improve the Measurement of International Tourism in France?

Guillaume Cousin* and Fabrice Hillaireau**

Abstract – Since July 2015, the *Banque de France* and the French Ministry for the Economy and Finance have been experimenting with the use of mobile phone data to estimate the number and overnight stays of foreign visitors in France. The purpose of the experiment is to assess the ability of such data to eventually replace, in part or in whole, the traffic data by mode of transport currently used to establish the representativeness of foreign visitor surveys (*Enquête auprès des visiteurs venant de l'étrangers* or EVE). Mobile phone data have yet to be incorporated into the method used to count tourists. However, estimates based on mobile phone data have a number of benefits in terms of the time required to obtain data, the level of temporal and geographical detail and short-term trend monitoring. This ongoing trial illustrates the difficulty of exploiting original Big Data and demonstrates the importance of drawing on traditional survey data to improve the quality of estimates.

JEL Classification : Z3, Z39
Keywords: Big Data, tourism statistics, mobile phone data

** Banque de France (guillaume.cousin@banque-france.fr)*
*** French Ministry for the Economy and Finance (fabrice.hillaireau@finances.gouv.fr)*

We sincerely thank François Mouriaux, Bertrand Collès, François-Pierre Gitton and Yann Wicky for their valuable contribution to the design and drafting of this article, as well as the Orange team with whom we conducted the experiment, in particular Sylvain Bourgeois, Jean-Michel Contet and Kahina Mokrani.

Received on 9 August 2017, accepted after revisions on 14 January 2019
Translated from the original version: "Les données de téléphonie mobile peuvent-elles améliorer la mesure du tourisme international en France ?"

Pour citer cet article : Cousin, G. & Hillaireau, F. (2018). Can Mobile Phone Data Improve the Measurement of International Tourism in France? *Economie et Statistique / Economics and Statistics*, 505-506, 89–107.
https://doi.org/10.24187/ecostat.2018.505d.1967

Since July 2015, the Directorate-General of Statistics of the Banque de France and the Directorate-General for Enterprise (DGE) of the French Ministry for the Economy have been experimenting with the use of mobile phone data to estimate the number of foreign visitors in France and their overnight stays as part of a survey partnership aimed at developing the tourism satellite account and determining the balance of payments.[1]

Counting foreign visitors and their overnight stays is necessary for the purpose of generating tourism statistics and estimating France's "travel services" trade balance. These statistics are of particular importance because of the weight of tourism in the French economy. In 2015, tourism consumption accounted for 7.27%[2] of GDP, 32% of which came from non-resident visitors. In the same year, France welcomed 84.5 million foreign tourists (DGE, 2016a), generating 52.6 billion euros of revenue from travel services recorded in the balance of payments.[3]

At present, visitor numbers are measured based on traffic data by mode of transport combined with counting sessions and surveys. Such sessions provide a detailed breakdown of border-crossing based on the country of origin of visitors, while traffic data by mode of transport serve as a basis for calculating extrapolation coefficients. Current estimates of visitor numbers require improvements which can, however, prove complicated when seeking to take into account rapidly-occurring changes such as the conversion of major airports into hubs – thereby multiplying the number of nationalities present on the same flight – and the difficulties involved in using road counts at borders in countries with multiple entry points (for example, France and Belgium share approximately three hundred border-crossing points). The purpose of experimenting with the use of mobile phone data is to evaluate the capacity of such data, in time, to replace traffic data by mode of transport for measuring foreign visitor numbers in France.

Mobile phone data have already been used to monitor tourism in Estonia and by a number of departmental and regional tourism committees, such as Bouches-du-Rhône Tourisme,[4] to measure visitor numbers and flows. Therefore, mobile phone data appeared to represent a potential solution for overcoming the new limitations affecting, or likely to affect, current data sources. This is because they are a rich source of information about the location and mobility of individuals. Gonzales *et al.* (2008) were among the first to use this data source to build a model of human mobility based on a sample of 100,000 mobile phones monitored over a 6-month period. This type of data has since been used to identify mobility patterns (Calabrese *et al.*, 2011, 2013), notably commuting (Aguilera *et al.*, 2014). Widhalm *et al.* (2015) sought to build a typology of urban activity patterns based on travel time, frequency and location. Many other uses have been explored in a wide range of areas (ONS, 2016).

Official statistical agencies have identified the potential benefits of Big Data for conducting population censuses (Vanhoof *et al.*, 2018; Givord *et al.*, this issue), but also for measuring tourism. Estonia has been a pioneer in this area, with an experiment reported in two main articles: Ahas *et al.* (2008), who found a strong correlation between mobile phone data and accommodation statistics, and Kroon (2012), who presented an experiment conducted by the Bank of Estonia involving the use of mobile phone data as a potential data source for estimating trade in travel services. Eurostat (2014) has since produced a feasibility study on such data for the purpose of monitoring tourism.

However, mobile phone data analysis is rarely used to measure tourism, and our experiment has the advantage of focusing on a relatively large country (twelve times larger than Estonia) welcoming a significant number of tourists (28 times more than Estonia). This paper also has the peculiarity of being written from the point of view of official statistics compilers and provides a different perspective to most other papers, aiming as it does to promote a regular operational use of Big Data for the development of statistical indicators. Given this background, it is important to test the quality of the indicators by comparing them

[1] *The partnership concerns the surveys known in French as SDT (*Suivi de la Demande Touristique*, or Tourism Demand Survey) and EVE (*Enquête auprès des visiteurs étrangers*, or Foreign Visitor Survey). The first survey involves collecting data on tourism demand among French nationals and is based on a representative sample of French households. The second data collection relates to tourism demand among non-residents visiting France and focuses on flows as evidenced by methods such as "border surveys". The partnership enables the integrated production of reference data for the official statistics for which each institution is responsible.*

[2] *Cf. DGE,* Compte satellite du tourisme *(base 2010); Insee,* Comptes nationaux *(base 2010).*

[3] *Cf. Webstat, Banque de France.*

[4] *Cf. https://www.myprovence.pro/bouches-du-rhone/projets-majeurs/projet-flux-vision-tourisme.*

to alternative data – in this instance, the reference survey on international tourism in France (EVE) combined with card payment data.

Mobile phone data have yet to be incorporated into the method used to estimate tourist arrivals, which has been limited to traffic data, counts and surveys. The test period highlighted many specificities in the use of Big Data: access to data, anonymisation and technical constraints, and quality of the data and of the indicators constructed on the basis of such data. It also meant that processing methods could be developed to render them more usable.

The Initial Requirement: Consolidation of the Current System Used to Estimate Foreign Tourist Arrivals

The Estimation of Foreign Tourist Arrivals is Based on Counts and a Survey Conducted at Border-Crossing Points

The current system used to estimate foreign tourist arrivals was developed based on the legacy of the border survey carried out between 1963 and 2001 with a view to bringing it into line with the wider context of the free movement of capital and the creation of the eurozone and of an area of free movement of persons (Schengen zone). The system is known in French as the *Enquête auprès des visiteurs venant de l'étranger* (EVE) and combines traffic data, counting sessions and a survey (Banque de France, 2015). The reason why monitoring international tourism presents such a challenge is that there is no sampling frame from which to carry out a traditional survey (such as is the case with "outgoing tourism", for example[5]). The EVE system is thus based, first, on a traffic census at the country's exit points (ports, airports, train stations offering international routes, road borders). Data on air, sea and rail passenger flows are collected from the various transporters and carriers, while road traffic flows are estimated by Cerema[6] using fixed or mobile automata distributed along all the borders (more than one hundred and fifty counting points in total). The second stage involves detailing the total flow, i.e. breaking it down into resident and non-resident flows. This requires counting operations to be conducted by researchers located at different points throughout the country. In airports, non-residents are counted in boarding lounges, which provides a basis for estimating the split between residents and non-residents based on a sample of flights and for extrapolating therefrom. In the case of travel by road, counting is conducted at border crossing points. The distribution of outgoing traffic by country or area of residence is then further specified using the responses to the EVE survey questionnaires. The EVE survey is thus based on a combination of traffic data (external data), specific counting operations (over a million vehicles counted at the borders, over 120,000 air passengers) and the survey itself, administered to over 80,000 visitors in both 2015 and 2016, the questionnaire being available in twelve languages.

The EVE Method Faces an Increased Need to Adjust the Traffic and Count Data

For counts (or tallies), the main challenge facing the EVE method in terms of statistical adjustment concerns road travel. For this particular mode of transport, the difficulties arise from both external traffic data, which are based on fewer measurement points, and counting and survey sessions, some of which are relatively unproductive. In order to obtain spontaneous responses at the end of their stay in France, travellers are interviewed in motorway rest areas near the border, where visitor numbers can vary unpredictably depending, for example, on the travel routes of coach operators. It follows that the extrapolated results relating to the distribution of outgoing visitor flows by geographical area can fluctuate erratically for some areas or countries or origin, meaning that specific corrective measures may need to be considered. Similarly, but to a lesser degree, splitting outgoing air traffic according to the country of origin of visitors needs to take into account the importance of transit in Paris airports and the role of major airports as hubs. The EVE survey is carried out with specific objectives in terms of questionnaires by area of origin. In the case of airports, the survey design is based on a sample of flights sampled in such a way as

[5] *French people travelling abroad. For these, a representative sample was obtained as part of the SDT (*Suivi de la demande touristique, *or Tourism Demand Monitoring) survey.*

[6] *Cerema (*Centre d'étude et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement*) is an administrative public establishment under the joint authority of the French urban development and sustainable development ministries.*

to collect questionnaires from visitors from a range of countries of origin. However, the link between flight destination and the country of origin of visitors is made tenuous by the fact that some passengers transit through an intermediate destination before returning to their country of origin. For example, Asian tourists are likely to fly from Paris to Frankfurt when leaving France, thereby complicating the targeting of the survey.

Mobile Telephony is a Potential Source for Counting Visitor Numbers

The decision to experiment with the use of mobile phone data was largely based on the potential of such data to address the issue of monitoring international tourism, as highlighted by several experiments. In France, some local authorities use such data to measure tourist arrivals. In Europe, Estonia has used mobile phone data as a main source for measuring incoming and outgoing international tourism since 2008-2010, while other countries have also shown an interest in using data of this kind, the potential of which has been highlighted in an in-depth study (Eurostat, 2014). However, the experiment conducted by the Banque de France and the DGE is unique by virtue of its scale, with the population of interest consisting of international tourists in France representing 85 million people a year. In addition, the territory where the count was conducted (metropolitan France) covers an area of 552 thousand square kilometres. By comparison, the number of tourist arrivals in Estonia is approximately three million a year in a country with a surface area twelve times smaller.

From a technical point of view, the use of mobile phone data is made possible by the fact that operators have access to the list of connections between cell towers and mobile phones, whether for the signals emitted passively or for mobile phone activity (calls, messages, reception of data via the Internet, etc.). The country of residence of the operator issuing the SIM[7] card of mobile phones connecting to the French network is also known to operators based in France and provides a basis for building a mass database of the signals emitted by roaming mobile phones. Mobile phone data therefore include variables of interest for the production of tourism statistics.

Experimental Framework and Procedures

The Arrangement of a Service Contract Corresponds to the Financing of a Cooperative Research and Development Initiative

While experimenting with data obtained from mobile phones for the purpose of counting the number of foreign tourist arrivals in France evidently amounts to a Big Data approach, the context of the experiment cannot be said, however, to reflect an open data approach. This is because the data are held by the various operators with access to a mobile network within metropolitan France. To gain access to these data in the form of statistics, the Banque de France and the DGE issued an invitation to tender in the spring of 2015 and received two proposals, reflecting, on the one hand, the interest of operators in collaborating with official statistical agencies with a view to gauging a new field of application for their Big Data and, on the other, the need for public funding as part of an arrangement to share research and development costs and the costs specific to the provision of information within the context of the experiment. When interviewing the candidates, expectations in terms of the transparency of data collection and processing methods were a key discussion point. The experiment was based on a distinction between the detail of the aggregation and anonymisation algorithms, which relate to the protection of the service provider's intellectual property, the market shares of the operator selected for the different populations and territories examined being a matter of commercial confidentiality, and the determinant variables for assessing the statistical quality of the data and the adjustment choices, all of the methodological options having to be gauged by the Banque de France and the DGE. A more detailed explanation of the distinction is provided below.

At the end of the tender process, the contract was awarded to Orange Business Services. The experiment focuses on data dating from early July 2015 to late June 2017. The last available

[7] *A SIM (Subscriber Identity Module) card is a chip used in mobile phones to store information that is specific to a mobile network subscriber, in particular for GSM, UMTS and LTE networks. It also allows the data and applications of the user, of the user's operator or, in some cases, of third parties to be stored. A SIM card contains an IMSI number, made up of a mobile country code (MCC), a mobile network code (MNC) and a mobile subscription identification number (MSIN).*

data point at the time of writing this paper was March 2017.

The selected bid has three characteristics: a pre-existing Big Data formatting module already available on the market, a balance between respect for intellectual property and methodological transparency, and a "phased" process governing the methodology used to develop the indicators.

The service provider had already developed a Big Data processing module adapted, in particular, to the needs of users looking for spatio-temporal data on visitor numbers (quantifying groupings of individuals in a given location, such as a cultural or sporting event). However, the proposed method had never been used for the purpose of observing international tourism over the entire national territory of France.

The service provider undertook to provide the two partners with a sufficient level of information to ensure that the method used would enable them to comply with the statistical quality standards established, in particular, for international and European institutions and be intelligible to the various audiences with an interest in tourism statistics. A knowledge and understanding of the methodology used serves to guarantee the independent ability to interpret results as well as the ability, where relevant, to make necessary revisions. At the same time, it was important to provide the service provider with an assurance of confidentiality in relation to the algorithm used to move from the basic datum of the signal transmitted by a SIM card to one or several towers to a raw datum representing a proxy of the anonymous physical person. The level of detail of the shareable methodological information therefore varies at stages 2, 3 or 4 and 5, as detailed below.

The mobile network operator's pre-existing module does not record the detail of the movement of SIM cards prior to processing their movements by aggregating the data following the requirement of the study. The processing method validated by the *Commission nationale de l'Informatique et des libertés* (French National Commission on Informatics and Liberty, CNIL) requires that the behaviours studied be predefined. Predefined behaviours are the sole target of real-time incremental counts of connections to the provider's networks, without personal data being stored. The counting method includes five stages (Diagram):

- Stage 1: the counting criteria are defined in advance with the Banque de France and the DGE and correspond to the tourist behaviours of interest (arrival, overnight stay);

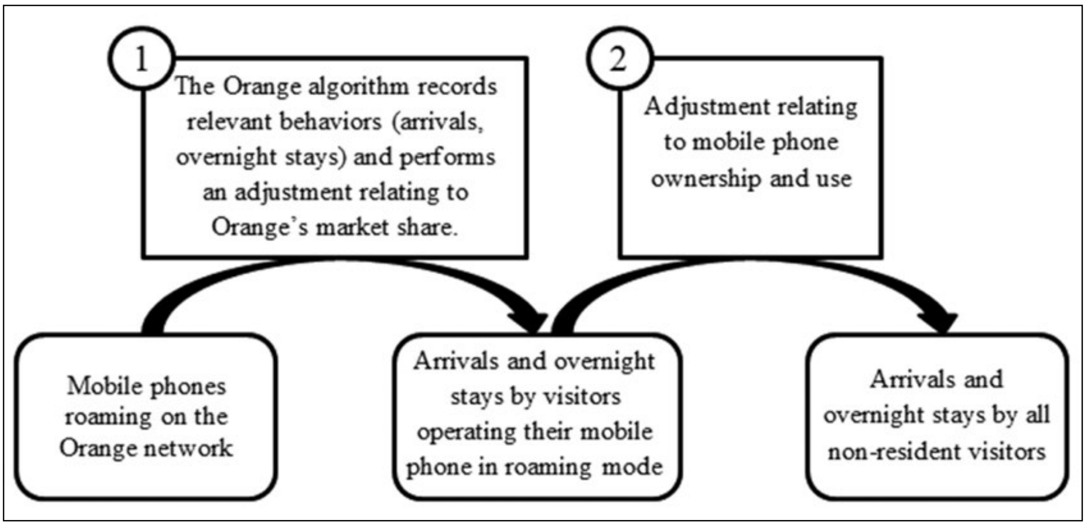
- Stage 2: mobile phone connection data are fed into an algorithm in real time. These involve signalling data, which include all of the communication between mobile phones and cell towers. The recorded signals are those emitted passively by mobile phones to connect to a cell tower based on their position as well as the data transiting via cell towers when mobile phones are being used (calls, SMS, use of mobile applications). These data originate solely from mobile phones equipped with SIM cards issued by a non-resident operator and roaming on the Orange network. They are not stored. The algorithm processes and anonymises the data as and when they are collected, rather like a meter. The algorithm constructs estimates of visitor numbers in terms of the number of mobile phones, by area of origin;

- Stage 3: the service provider carries out a "spatio-temporal" adjustment by aggregating the data on the different networks ("2G", "3G " "4G") and correcting the effects associated with the constant changes being made to these networks (putting into service of new cell towers, temporary or long-term unavailability);

- Stage 4: the transition from connection data to an estimate of the number of foreign mobile phones present within metropolitan France requires an adjustment to the service provider's market share of roaming customers, by country of origin and operator of origin in combination. Market share by country-operator is measured based on the distribution between the different operators of SMS sent from roaming mobiles, which is available in real time to the service provider;

- Stage 5: lastly, the number of visitors by area of origin is estimated based on a traditional statistical adjustment relating to mobile phone ownership and usage rates. Since it is carried out ex post on the results aggregated by country (or larger area) of assumed residence, this last adjustment lends itself to the execution of several scenarios.

Diagram
Simplified Methodology Used to Construct the Indicators


1
The Orange algorithm records relevant behaviors (arrivals, overnight stays) and performs an adjustment relating to Orange's market share.
2
Adjustment relating to mobile phone ownership and use
Mobile phones roaming on the Orange network
Arrivals and overnight stays by visitors operating their mobile phone in roaming mode
Arrivals and overnight stays by all non-resident visitors

A Setup Stage is Required Prior to the Observation Period

The setup stage involves defining, in conjunction with the service provider, the different behaviours of interest in order that they can be measured. In the case of the present experiment, the definitions of tourist arrivals and overnight stays had to be translated into criteria relating to the presence of mobile phones on a network. An overnight stay was thus defined as the presence of a mobile phone between midnight and six o'clock in the morning. An arrival is counted when the first overnight stay is recorded following an absence the previous night. Such hypotheses allow mobile phone data to be interpreted in terms of behaviours. Some studies have used similar hypotheses to examine mobility patterns based, for example, on presence at home between midnight and eight o'clock (Akin & Sisiopiku, 2002). The initial criteria were gradually completed in order to correct a number of measurement deviations.

The setup stage is also an opportunity to define the territory of interest. This requires selecting the cell towers involved in the counts. A decision was made not to take into account the flows received by towers located on French soil but close to the borders, a factor deemed necessary when analysing the initial results generated from the operator's pre-existing module. The reach of cell towers being unaffected by administrative boundaries, it is important not to count foreign residents outside France. For the data aggregated from the entire territory of metropolitan France, which are the primary target of the Banque de France and the DGE, we may therefore expect a slight underestimation of foreign tourist arrivals. Their numbers cannot be estimated since, over the experimental period, the overall noise does not allow bias measurements of such accuracy to be carried out. In the case of data broken down at a regional level, the problem of cell tower selection also arises at each administrative border, with the map of towers and their reach not being aligned with the map of regions and departments. This requires a specific focus on allocating cell towers on the basis of the spatial groupings sought.

The Series Obtained and their Usefulness

The Series Received

The operator provides the Banque de France and the DGE with estimates of the number of international tourist arrivals and overnight stays. These estimates are provided monthly within a theoretical period of one month. The indicators are provided at a daily frequency. Estimates are provided for 29 geographical areas of visitor origin. The data received are in CSV format. One file contains the arrivals while another records the overnight stays. Each file contains three columns: area of origin, date and number of overnight stays or arrivals for the corresponding origin × date intersection. The files received thus contain approximately 900 lines (29 geographical areas plus a total line times approximately 30 days). For a given day and country of origin, it is thus possible to determine the number of tourists who arrived and the number of tourists who spent the night in France on that day.

It is important to note that mobile phone data were initially chosen to calibrate the measurement of tourist flows travelling by road. The service provider was therefore required to include a distribution of tourist arrivals by border and mode of transport. The subtlety of the collection process and the operator's expertise were expected to identify the different means of transport, with rail travel being characterised, for example, by a high number of SIM cards operating at the same speed and over an identified route. In actual fact, the significant differences found between the data of the operator's pre-existing module and the contextual data (EVE survey) were such that the identification of the mode of transport was quickly abandoned. The approach ultimately adopted therefore focuses on the primary requirement to count the number of foreign visitors arrivals, aggregated for all modes of transport and borders.

While Initially Disappointing, the Quality of the Data Improved

Comparison of the indicators derived from mobile phone data and survey data provides a basis for assessing their reliability and examining potential sources of variance. Such comparisons have been carried out by several studies devoted to the analysis of mobility and the construction of origin-destination matrices. The results obtained using mobile phone data are, in some cases, close to the survey results, but are of a higher level (Calabrese *et al.*, 2013). In the field of mobility, a more recent study (Bonnet *et al.*, 2015) compared the results of the global transport survey with estimates derived from passive mobile phone data provided by Orange. The authors found a strong correlation between the two types of estimates and reached similar estimates in terms of the total number of trips in the Île-de-France region (9% difference). However, their study covered a short period (twelve days).

In the context of the experiment conducted by the Banque de France and the DGE, the first deliveries of estimates in the third quarter of 2015 showed significant differences with the estimates of the EVE survey, the only available source for an estimate of tourist arrivals and overnight stays in metropolitan France. The indicator obtained from mobile phones pointed to over one hundred million tourist arrivals in the third quarter alone, whereas the approximate number of tourist arrivals is 85 million a year.

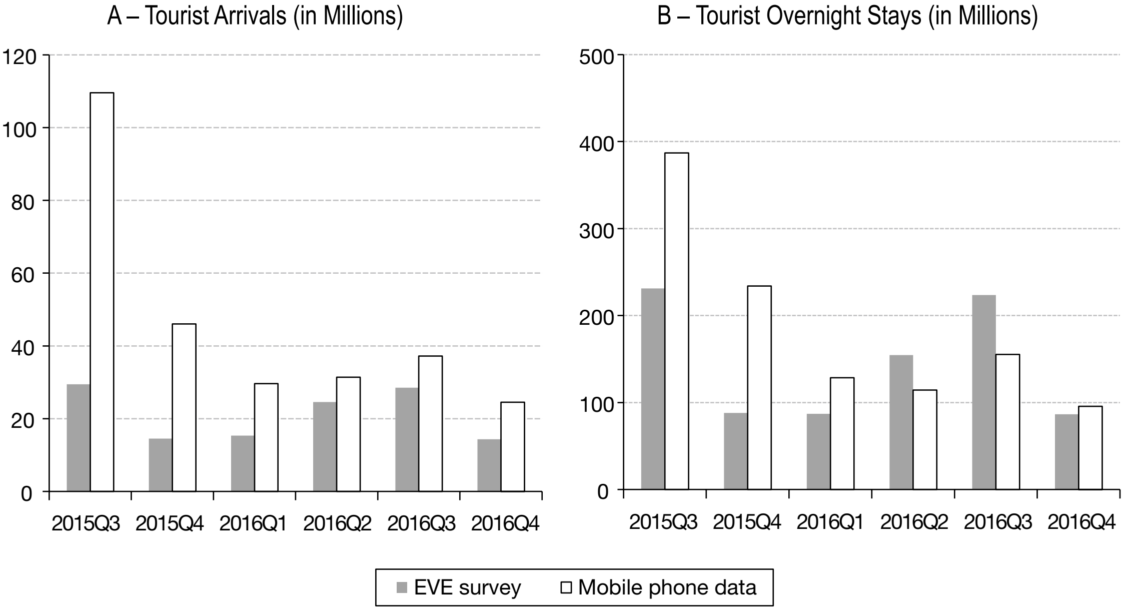
It soon became apparent that the indicators developed based on a single collection are unusable for various reasons set out below. Arrivals of tourists and day visitors (i.e. visitors who do not spend the night on French soil) in France were thus found to be of insufficient quality. Further work is therefore needed on "consolidated" indicators,[8] such as the number of overnight stays, each overnight stay being defined as presence confirmed several times in the same location over a defined time frame. Overnight stays are therefore not counted in the case of arrivals previously defined by the measurement system. Rather, the number of arrivals is deduced from the number of observed overnight stays. This is entirely consistent with the letter and spirit of the international definition of a tourist: a tourist is a visitor whose visit includes at least one night in a territory which is not his or her habitual residence.[9] Lastly, the goal of counting the number of one-day visitors (i.e. visits which do not include an overnight stay), made difficult by the exclusion of border areas where cell towers are likely to cover a portion of foreign territory, was soon abandoned, it being impossible to define it either directly or as a balance. The improvement work therefore focused mainly on an analysis of overnight stays.

To improve quality, various corrections were made throughout the experimental period; these are described below. The result was to bring the estimates of arrivals and overnight stays based on mobile phone data closer to trustworthy levels compared to the EVE survey estimates (Figure I). The difference between the estimates of the total number of overnight stays decreased from 67% in the first quarter delivered (third quarter 2015) to 10% for the final quarter currently available (fourth quarter 2016).

[8] *More generally, monitoring observations over time helps to overcome many of the defects of the measurement system, although the CNIL's requirements restrict such monitoring to 3 consecutive months.*

[9] *See the website of the World Tourism Organization: http://media.unwto.org/fr/content/comprende-le-tourisme-glossaire-de-base*

Figure I
Comparison of Measurements by the EVE Survey and Mobile Phone Data



Coverage: Quarterly arrivals of non-resident tourists.
Sources: Banque de France, DGE.

However, the quality of the estimates remains insufficient for three reasons. First, significant differences remain between the two sources in relation to countries or areas of origin. Some nearby areas appear to be overestimated whereas remote areas are underestimated. In the 4th quarter of 2016, for tourist overnight stays, the total difference of 10% between mobile phone data and the estimate obtained from the EVE survey arises from the compensation of very significant differences at country level. The estimates of overnight stays based on mobile phone data are 78% higher than those of the EVE survey for Germany, but approximately 80% lower for the United States, Canada and Brazil, for example. These differences stem in all likelihood from the values of the mobile phone usage rates used to adjust the data; for tourists originating from remote countries, the adjustment factors poorly reflect visitor behaviours. This limitation is inherent to mobile phone data: the quality of the estimates depends on knowledge of the operator's penetration rate by population segment. This limitation has been noted in several other studies, including when the population of interest is the resident population (Bonnet *et al.*, 2015). However, unlike the definitions of tourist behaviours, the adjustment factors relating to the use of mobile phones can be modified *a posteriori*. Therefore, the quality of the data can be improved by furthering our understanding of behaviours.

Second, estimates of tourist arrivals are less robust than estimates of overnight stays. This is due to the problem of interrupted stays (see next section), which has a comparatively greater effect on arrivals than on overnight stays. Lastly, for some countries or areas of origin, there is a seasonality in the estimates obtained from phone data that differs significantly from the survey and which appears to be of limited credibility. For example, in the case of Spain, the EVE survey indicates that tourist overnight stays increase by 80% between the second and third quarters, reflecting the seasonality characteristic of the data from the professions in question (air traffic to tourist destinations, hotel occupancy, etc.), whereas the mobile phone data indicate a much lower increase of 13%. It follows that estimates obtained from mobile phone data are not yet of sufficient quality to replace the traffic data currently used.

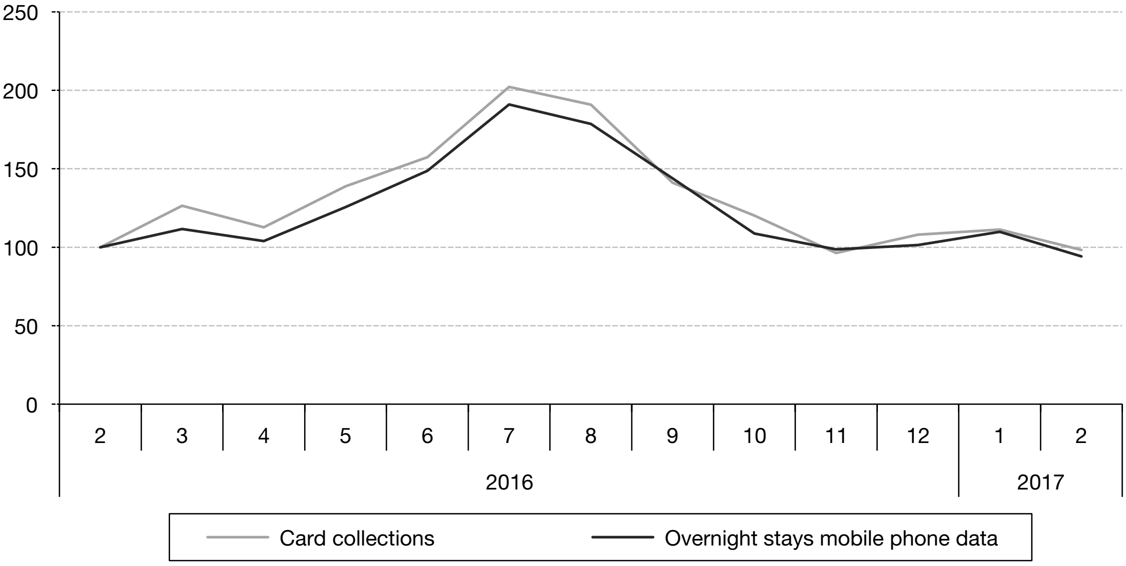
The Data are Potentially Adapted to Short-Term Trend Monitoring

As noted above, mobile phone data have yet to be made more reliable at different levels. This is largely because of the adjustments relating to mobile phone usage among visitors travelling from remote areas and artificially interrupted stays.

An analysis of these evolving data is of greater value, all the more so since such data are generally available before traditional survey data and provide a greater degree of chronological detail since daily estimates are available. The potential of mobile phone data for monitoring short-term trends can be seen by comparing them to the payment card data collected monthly by the Banque de France, which relate to cash payments and withdrawals made in France using a non-resident card and which are aggregated by country. Such spending is not an exact reflection of travel credits (use by residents of foreign cards, spending in currencies withdrawn in the country of origin or prepaid by simple bank transfer) but overlaps with it to a great extent, particularly in the case of specific nationals from particular countries among whom the use of card payments is the primary or even exclusive payment method. Payment card data also have the advantage of being available on a monthly basis with a detailed geographical breakdown, thereby allowing comparison with mobile phone data. Over the period February 2016 to February 2017,[10] the revenue derived from payment cards and tourist overnight stays as measured using mobile phone data were highly correlated: the correlation coefficient between the two series was 0.986 (Figure II). Furthermore, the high correlation between revenue derived from the collection of bank card data and tourist overnight stays estimated using mobile phone data was also found to be true for the different countries observed, albeit with some exceptions. Thus, while estimates of the number of overnight stays in levels show significant differences with the results of the EVE survey for certain countries (for example, the United States, Canada and Brazil), the correlation between overnight stays and payment card spending is high for all countries except Brazil (very limited use of payment cards), Morocco and Luxembourg (on account of the high proportion of cards issued in Luxembourg but used by residents of other countries). For the other countries, the correlation coefficient between overnight stays estimated based on mobile phone data and revenue obtained from card payments ranges between 0.66 and 0.97. Furthermore, some countries where the number of overnight stays is significantly underestimated using mobile phone data present fairly accurately estimated short-term trends

[10] *The decision to start in February 2016 is based on the fact that the last significant methodological change caused a break in series between January and February 2016.*

Figure II
Tourist Overnight Stays Estimated Using Mobile Phone Data and Revenue from Payment Card Transactions (Base 100 in February 2016)



Coverage: Expenditure in France using non-resident cards (excluding online) and overnight stays of non-resident tourists.
Sources: Banque de France, DGE.

(Canada, United States). Mobile phones therefore provide reliable trend estimates and their use for short-term estimates is conceivable pending calibration of the estimates in levels.

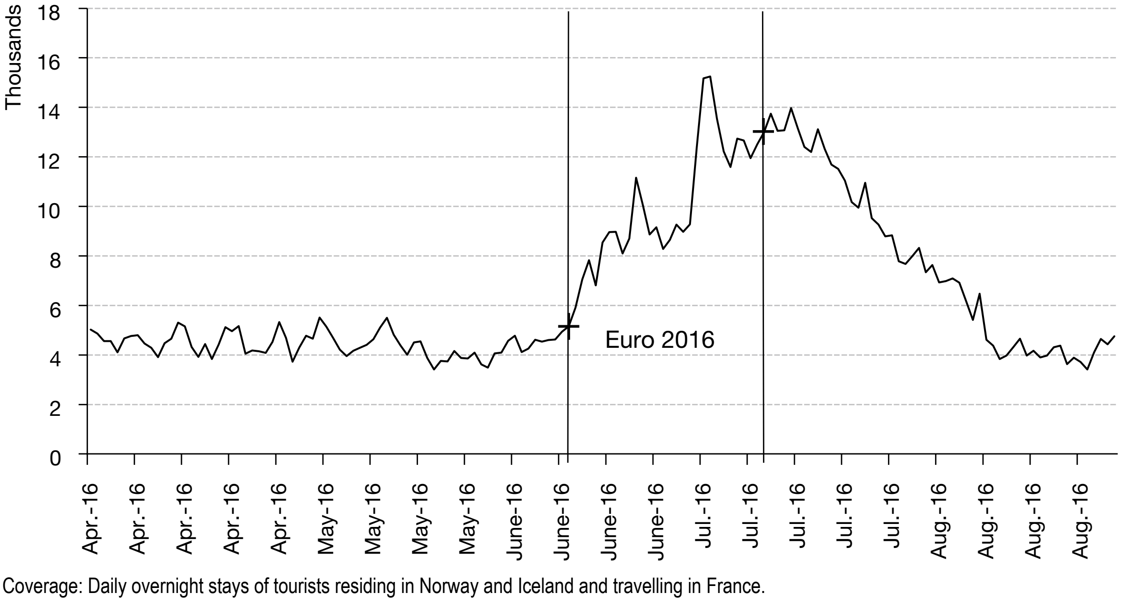
Mobile phone data also provide an insight into shock events affecting tourist arrivals. These are not so readily identifiable with the EVE survey, which produces quarterly results. Mobile phone data are also better at covering some relatively rare areas of origin. The example of the Euro 2016 football competition provides an illustration of the measurement accuracy of mobile phone data: the success enjoyed by the Icelandic team is reflected in the gradual increase in the number of visitors from that country (Figure III).

Sources of Bias in the Estimation of Arrivals and Overnight Stays

A Typical Measurement Problem: Sporadic Connections

The first cause of overestimation relates to "sporadic connections". The term is used to refer to mobile phone connections which are found to be roaming on the Orange network but which do not use Orange as a preferred network. The presence of such connections on the network does not correspond to an adjustment hypothesis used in the service provider's pre-existing algorithm. The operator first observes the roaming mobile phones present on its network, whether they be active or not, and then proceeds with an adjustment relating to its market share in order to infer the total number of mobile phones roaming in France, regardless of network. The key element of this adjustment is market share as measured by the number of SMS sent from roaming mobile phones, by country-operator. The adjustment is pertinent if the distribution of mobile phones in terms of presence on the network is equal to the distribution of SMS messages. This may not necessarily be the case, however, in particular because of the preferential agreements which national operators may have entered into with foreign operators. The mobile phone of a foreign tourist (i.e. someone who holds a contract in country P1 with operator E1P1) in France may thus be received first by the cell tower of French operator F1 with preferential agreements with E1P1, if the state of the network permits it. However, it may also be received by the cell tower of another French operator (F2) if the state of the preferred network is inadequate. Sporadic connections are connections made by a non-resident tourist (in all likelihood the holder of a contract with an operator who has entered into a preferential agreement with another French operator, F1) received sporadically on network F2, for example when travelling through a "black spot" of network F1. If the tourist in question

Figure III
Daily Overnight Stays of Icelandic and Norwegian Tourists



Coverage: Daily overnight stays of tourists residing in Norway and Iceland and travelling in France.
Sources: Banque de France, DGE.

does not use his or her mobile phone actively over that short period, she/he will not be represented in the numbers used to calculate market shares. The adjustment key is therefore underestimated and the extrapolation is carried out with an excessively high coefficient, implying an excessive estimate for the number of SIM cards of the relevant country over the considered period and territory. Because of this problem, it was necessary to make several changes to the measurement criteria. These changes are detailed in the next section.

Reception Interruption: A Cause of Underestimation of the Average Length of Stays and of Overestimation of Arrivals

The second measurement problem is in part related to the first and concerns artificial interruptions of stays. The counts available in the spring of 2016 indicate significantly excessive arrivals and overnight stays of an order of magnitude compatible with the contextual data, implying an excessively short average length of stay. To improve the measurement of the length of stays and arrival numbers, the Banque de France and the DGE requested an examination of the following intuition: stays may be artificially shortened by reception interruptions. Such interruptions can be caused by many factors: a mobile phone which has run out of battery or been deliberately turned off, travel to an area not covered by the Orange network, etc. The result is an automatic overestimation of the number of tourist arrivals: when reception resumes, if the interruption included an overnight stay, the SIM card is treated as a new arrival. The phenomenon also results in an underestimation of the number of overnight stays. For example, during a traditional one-week holiday stay on French soil, a SIM card signal may be detected regularly over three days, not be detected for two days and again be detected over two days prior to leaving the country. The operator's pre-existing system will therefore assume that two tourist arrivals have taken place, with the first arrival corresponding to a three-night stay and the second to a two-night stay.

To evaluate this hypothesis, a first test was performed in early 2016 over a limited area and period conducive to analysis: a hill station. The advantage of the chosen area is that we have a good understanding of tourist behaviours in such settings: foreign customers, a significant proportion of stays starting on Saturdays and ending the following Saturday, and a clearly defined reception area on account of natural barriers. From this, it was found that the proportion of tourist stays concerned by artificial interruptions may be too high. This result is further supported by another finding at a national level over two observation periods: March 2016 and September-November 2016. For these periods, the number of arrivals was studied based on a requirement for absence prior to arrival. While this requirement is commonly set at one day, the aim of the observation was precisely to determine whether the arrivals recorded by the algorithm are genuine arrivals or artificial arrivals of individuals already present within the studied territory. Over the course of March, the application of a more stringent absence requirement (two days of absence prior to an arrival) resulted in the number of total arrivals decreasing by 13%. If the requirement is extended to six days of absence, the number of arrivals decreases by 37%.

The underestimation is therefore confirmed without it being possible to correct it on account of the inability to distinguish between artificially interrupted stays and genuine regular short stays, this distinction being necessary to establish the number of overnight stays. The application of a more stringent absence factor is not self-evident and would imply a departure from the statistical definition of tourism. Beyond the impact on the aggregates, which is significant if we exclude, for example, all arrivals prior to three days before the previous arrival, a more fundamental problem arises: should we rely on probabilistic reasoning, which would involve correcting an unsatisfactory measurement by adapting and altering definitions? A direct intervention on the source data, a physical identification of anomalies and correction prior to determining the aggregates appear preferable but are not feasible at present. While in terms of behaviour the identification of problematic cases appears unequivocal (sudden exit from the network, generally at a distance from a border, or over a route incompatible with an actual exit from the territory in question, and an equally sudden return), such an identification is not compatible with the operator's pre-existing measurement system. The question of interrupted stays is referred to in the study conducted by Estonia on international travel monitoring (Kroon, 2012). The authors mitigate the difficulty by adopting a number of hypotheses. They consider a mobile phone to be present if the phone has remained inactive for less than 7 days and that the traveller has left if the period of inactivity is greater than 7 days. Such a solution is not ideal in the case of France, a major transit point.

The Country of Issue of the SIM Card and the Country of Residence May Differ

A third measurement difficulty concerns behaviours which weaken the hypothesis that the country where the SIM card was issued is the same as the country of residence of the owner of the mobile phone. Having been identified at the outset of the experiment by analogy with research conducted on French tourists, this difficulty has been in part corrected.

Converting the Number of SIM Cards into Tourist Arrivals is not Straightforward

Lastly, the final difficulty concerns the ultimate adjustment phase, which occurs *a posteriori*, independently of the operator's pre-existing system, in order to distribute arrivals and overnight stays according to the issuing area of the tourists. As noted in the section relating to data quality, the adjustment factors used to extrapolate tourist arrivals from the number of mobile phones are not always appropriate. The data relating to ownership rates in different countries come from GSM Alliance. However, there are significant limitations to using data on ownership rates among populations due, on the one hand, to the difference in representativeness between the total population of a country and the percentage of that population visiting France and, on the other, to the specific behaviours exhibited when travelling abroad. Ownership rates can vary according to the tariffs charged by operators in a given country and according to sociocultural factors (such as populations being more or less sensitive to security matters and with more less intense connection habits).

In the absence of external data on ownership rates, the service provider combines the data relating to ownership rates with several sets of coefficients determined based on broad groups of countries defined according to their remoteness.

This adjustment problem, while it may not prevent a trend analysis[11] of indicators on a country-by-country basis, nevertheless complicates the use of aggregate indicators. By positing that the trends in tourist arrivals for residents of country A are measured adequately and that the same applies to country B, the trends in overnight stays for residents of both countries taken together cannot be determined without drawing on external data relating to the respective contributions of both countries to tourist arrivals in the area considered. Thus, besides the numbers by country which are, in some cases, significantly underestimated, the evolving data are affected when considering a range of nationalities.

Corrections Made, Resulting Benefits and Limitations

The first correction made involved introducing a requirement for mobile phone loyalty to the operator's network in order to reduce the noise caused by sporadic connections by mobile phones connecting to the network only in the event of connection to their preferred network being lost. As indicated above, taking these phones into account results in an overestimation of the number of overnight stays and arrivals on account of the adjustment made by the operator in relation to its market share of roaming customers. In order to distinguish between regular and occasional users, the first criterion to be introduced concerns the total amount of time spent on the network, which must be higher than 9 hours over a 21-hour period. This first criterion was added for the delivery of the data from November 2015 and resulted in a decrease in the estimated number of overnight stays of around 30% (see Figure IV). It was further strengthened by adding a new loyalty requirement for the delivery of the data from February 2016. This requirement, applied systematically, requires that at least three network events be performed over the course of the 24-hour period prior to or following the recorded tourist overnight stay. Despite these successive improvements, sporadic mobile connections continued to introduce noise into the measurement. Current studies are tending to focus on selecting operators of the country of origin of visitors based on their loyalty to the network when roaming in France.

[11] *Over short periods at the very least: long-term analysis presupposes stability in mobile phone usage behaviour.*

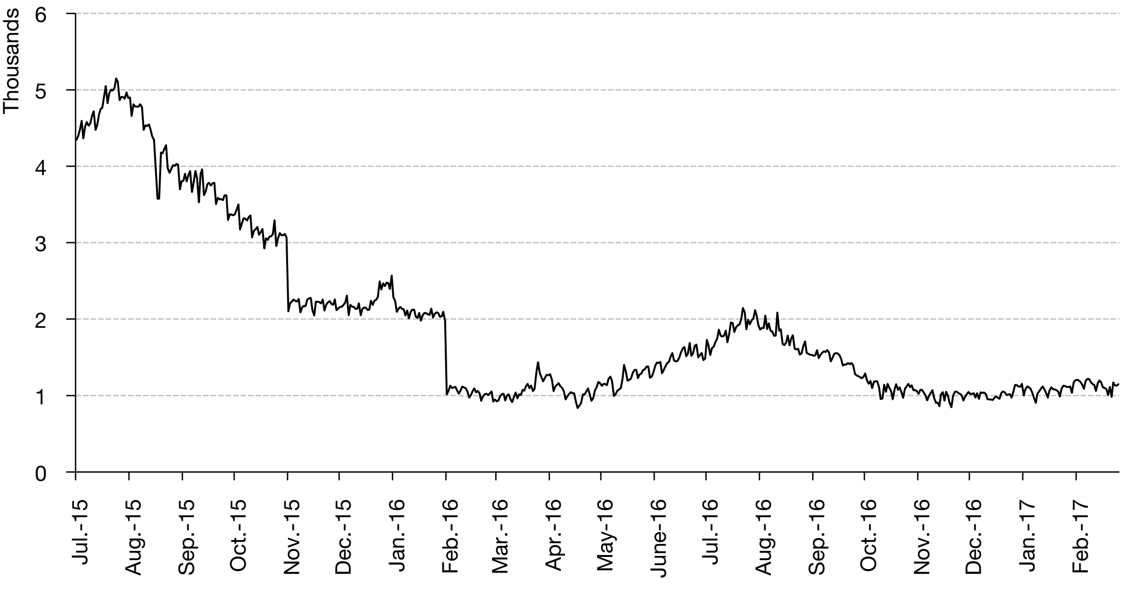
The second important measurement correction concerns French residents using a mobile phone with a foreign SIM card, which may be the case, for example, of cross-border workers (i.e. French residents working abroad) who have taken out a phone contract with a foreign company. Such behaviour tends to somewhat limit the validity of the hypothesis according to which the user's country of residence is the same as the country which issued the user's SIM card, resulting in an overestimation of the number of tourist overnight stays. The correction made drew on the significant contribution of the working group led by Tourisme & Territoires[12] in relation to the segmentation of the observed population. This is unavoidable for resident population data since the frequency and duration of trips make it possible to distribute individuals into different categories. The number of nights spent in a given territory means that someone carrying a mobile phone may be deemed to be a resident of that territory, regardless of the characteristics of his or her SIM card. Applied more simply to SIM cards with a foreign code, this method of segmentation serves to eliminate individuals who spend more than half of their nights in France over a two-month period. Therefore, the service provider added a non-residence requirement to the mobile phone data processing algorithm. Individuals who had spent more than one month on French soil over the previous two months were treated as residents and were therefore excluded from the measurement of international tourism, thereby addressing the case of cross-border workers.

Because of the necessary learning process, the first data delivery to take account of this correction was the delivery made in February 2016. Combined with the increased requirement for loyalty to the mobile network, this correction resulted in a reduction of the total estimate of overnight stays of approximately 50%. The impact on overnight stays is significantly greater than on arrivals. The average length of a holiday stay in France is 6.8 days for all international customers (DGE, 2017), while residents spent almost all overnight stays in France. However, the correction remains imperfect since it implies excluding from the measurement tourists who stay in France for a period of more than one month, whereas the statistical definition includes stays of up to one year.

Insofar as the anonymity requirement prevents individual data relating to connections between mobile phones and cell towers from being stored, the alteration of the definitional

[12] See http://www.tourisme-territoires.net/zoom-sur-le-projet-flux-vision-tourisme/

Figure IV
Tourist Overnight Stays Estimated Based on Mobile Phone Data



Coverage: Daily overnight stays of non-resident tourists travelling in France.
Sources: Banque de France, DGE.

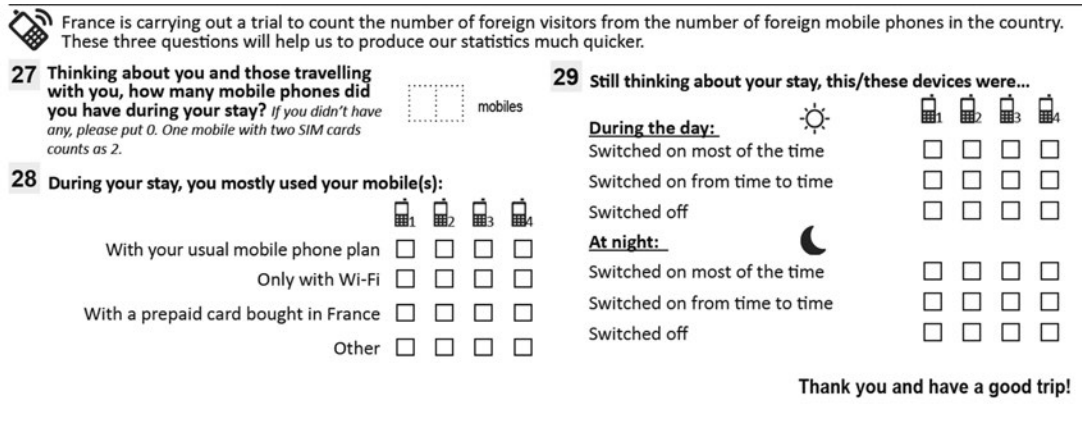
criteria used to define relevant behaviours for monitoring tourism cannot be projected backwards. The effect of the corrections described is in evidence in Figure IV in the form of the breaks in the series observed between November 2015 and February 2016.

The Specific Problem of Behavioural Adjustment Requires an Exogenous Collection

Since an understanding of mobile phone usage behaviour among visitors was not enough to satisfactorily adjust the counting of SIM cards, the Banque de France and the DGE made the decision to incorporate a set of questions relating to the use of mobile phones into the EVE survey questionnaire. These questions (Box) were added in January 2017 and the first results will be available for analysis at the end of the year. It will then be possible to determine a coefficient for each of the main countries of residence and thereby to improve the adjustment by country of residence. However, while usage behaviours have been found to vary widely both temporally and spatially, the use of mobile phone data will prove more costly because of the requirement for a collection dedicated to adjustment. This will have an impact on the cost-benefit assessment to be carried out at the end of the experiment in order to decide whether or not to incorporate these data in the current production process.

Among the difficulties related to ownership and use rates are the family composition of tourist groups, which is likely to play a significant role: it is not simply a matter of estimating the number of individuals carrying mobile phones, the number of accompanying people also having to be measured. The issue of the impact of group composition is not specific to the measurement of foreign tourist traffic. Contextual data are available for French tourists: according to a study based on the permanent tourist demand monitoring (SDT) scheme carried out by the DGE and the Banque de France in the spring of 2015, while the mobile phone ownership rate is relatively consistent among residents aged 15 years and over, it varies very significantly up to the age of fifteen. The number of tourists aged under fifteen years accompanying a tourist aged over fifteen also depends heavily on the period (school holidays or term time), the type of accommodation (hotel/campsite/rental) and, therefore, the area. Inclusion of this impact will also be an important stage in improving the measurement of French tourist arrivals. The tourists-to-SIM cards ratio is inevitably higher in a family-friendly camping area at the height of the summer season than outside school holidays in an area dominated by work-related tourism and its representatives equipped with several mobile phones, or even with mobile phones equipped with several SIM cards; in such cases, the difficulty is to have suitable and sufficiently refined adjustment factors.

Box – **Questions on Mobile Phone Usage, EVE 2017 Collection**



France is carrying out a trial to count the number of foreign visitors from the number of foreign mobile phones in the country.
These three questions will help us to produce our statistics much quicker.

27 Thinking about you and those travelling with you, how many mobile phones did you have during your stay? *If you didn't have any, please put 0. One mobile with two SIM cards counts as 2.*

□□ mobiles

28 During your stay, you mostly used your mobile(s):

	1	2	3	4
With your usual mobile phone plan	□	□	□	□
Only with Wi-Fi	□	□	□	□
With a prepaid card bought in France	□	□	□	□
Other	□	□	□	□

29 Still thinking about your stay, this/these devices were...

	1	2	3	4
During the day:				
Switched on most of the time	□	□	□	□
Switched on from time to time	□	□	□	□
Switched off	□	□	□	□
At night:				
Switched on most of the time	□	□	□	□
Switched on from time to time	□	□	□	□
Switched off	□	□	□	□

Thank you and have a good trip!

Future Developments and Uncertainties

The Abolition of Roaming Charges in the European Union

The abolition of roaming charges within the European Union has been in force since June 2017. Operators began to limit such charges to certain destinations a number of years ago. Consequently, tourist behaviours within the European Union are expected to become more standardised, the assumption being that everyone within the EU will eventually come to use their mobile phones as if they were in their country of habitual residence. This would help to improve the accuracy of estimates for European countries, although the transition phase between current habits and the anticipated standardisation may result in noise in the estimates, the expected increase in the use of roaming mobile phones abroad being likely to lead to an artificial increase in the number of visitors recorded.

However, this risk should be put into perspective since the system is not only based on actual mobile phone usage but also on passive connections to the network, which are not charged to users; therefore, the impact will not necessarily be significant. Furthermore, inasmuch as some operators already no longer charge roaming fees to their subscribers on internal destinations within the European Union, the transition phase has been underway for many months and its impact will be spread out over an extended period. The data collected among tourists as part of the EVE survey should enable such behavioural changes to be measured.

While European customers account for approximately 79% of foreign tourist arrivals in France (DGE, 2016b), customers from further afield represent a not insignificant and indeed growing proportion of total visitor numbers. For these customers, roaming charges are likely to remain high and to deter a large proportion of tourists from using their original SIM cards.

WiFi Connection

Specific connection practices may limit the scope of the measurement method based on mobile phone networks. Thus, some tourists, such as North American and Chinese travellers, are already opting for WiFi hotspots and connection to an Internet network in order to use specific voice communication applications without having to connect to a mobile phone network. The use of such applications, popular among young travellers and technophiles, has so far eluded measurement. In addition, the availability of a WiFi connection is identified by tourist sites as a key pull factor and technical solutions are flourishing, for example in coastal areas. The dedicated questions added to the EVE survey in early 2017 should help to better gauge the scale of the phenomenon.

Supranational Contracts

Another development affecting the accuracy of the system (rather than its exhaustiveness) is the rise of supranational contracts, which is likely to weaken the link between the country of the detected SIM card and the country of residence of its user. This development may be further encouraged by the abolition of roaming charges within the European Union (see above), although the ban comes with some restrictions. These restrictions are designed to dissuade extreme uses (characterised by minority usage in the country of issuance of the contract). The difficulty of inferring the user's place of residence from the nationality of the SIM card is already established, as evidenced by the very high counts of overnight stays seemingly by visitors from Luxembourg in France: this finding – one of the first of the experiment carried out by the Banque de France and the DGE – has proved resistant to the various improvements made to the method. Therefore, the only explanation lies in the fact that Luxembourg companies sell contracts used by residents of other countries, which is no doubt connected to the number of cross-border workers in Luxembourg, who may reside in France but also in Belgium or Germany.

Assessment of the Experiment and Avenues for Further Improvement

The Type of Partnership Put in Place and the Research Approach Appear Suited to the Specific Features of These Big Data

The experiment suggests a number of key success factors for a partnership between a private company holding Big Data and the institutions responsible for producing official statistics. In the case of this ongoing experiment, the research was based on a tried and tested data formatting module used to serve different needs to those set out here (analysis of events within a limited area – city, department, etc. – as opposed to a level-based assessment of visitor flows and length of stay across metropolitan France as a whole). The advantage is that datasets were available from the very outset of the partnership, which was conducive to an empirical approach and to comparison with the reference data held by the Banque de France and the DGE. The drawback lies in the low significance of the initial results in a context in which the upstream processing stage (phases 2 to 4) depended entirely on the supplier's expertise. This particular obstacle was overcome by putting in place a co-development approach. This requires that the parties commit proportionate and balanced resources, hence the importance of a partnership management structure that actively involves the experts and draws on the appropriate level of decision-making. In this context, the ability of the two parties to make adjustments rapidly (agile method) was found to be essential. For example, the Banque de France and the DGE decided to include a module in the 2017 survey questionnaire that allowed for variables to be collected on mobile phone usage behaviour which, if robust, will improve the potential for exploiting mobile phone data for statistical purposes.

The co-development approach and the agile method also appear to be adapted to the specific features of Big Data: the observation of very frequent events across the entire territory of metropolitan France requires greater computational capabilities than those required for current processing, hence the importance of high reactivity to adapt computational capabilities. Some definitional adjustments require a period of examination of variants, which is inherent to this type of experiment. Definitional stability, which is necessary to the construction of series and their comparison with existing sources, is not immediately achievable, the implication being that researchers need to be willing to interpret successive results that incorporate changes in method, which requires a high level of interaction between them. A test by sampling or by restricting the experiment to a limited territory would have served to mitigate this problem, but would not have achieved the goal of exhaustive measurement across the entire territory, which is one of the main contributions which these data are expected to make.

The Advances Achieved During the Experiment Open Up New Possibilities for Short-Term Trend Monitoring and the Regionalisation of Tourism Statistics

The setup of data processing provides results adapted to the short-term monitoring of tourist arrivals over the short-term and to the measurement of shocks in the case of one-off events (sporting event, festival, attack, comparison of populations in high season and low season). Excluding the setup period, the speed at which statistics can be generated from mobile telephone data (less than thirty days before the end of the observed month) is a definite advantage compared to survey-based collection methods while also being comparable to bank card data collection methods. Such uses require certain precautions, including the use of trend rather than volume indicators. As a second area of interest, the statistical processing derived from the experiment should provide, for each of the principal countries of residence, a satisfactory distribution of overnight stays across the thirteen metropolitan regions.

However, Long-Term Use Requires Other Improvements

To enable dissemination by different users, future research should enable significant improvements in two areas. The first such improvement concerns the reduction of the overall noise of the measurement. This relates to the very heart of the operator's pre-existing system, which served as a starting point for the experiment and implies changes to the basic algorithms. Bypassing insufficient quality by adapting the definitions of predefined behaviours cannot be deemed to be an acceptable solution. The second area of improvement relates to our understanding of mobile phone usage behaviour. The creation of an external database on mobile phone use rates among foreign visitors in France, segmented based on the principal countries of origin, is necessary. Given this, the cost of collecting high-quality external data represents one of the key factors in assessing the value of using mobile phone data. Since such data are intended to reduce our reliance upon or even replace the collection of external data relating to overall traffic by means of transport,[13] deploying a large-scale collection method to adjust the collected data to the data which they are designed to replace would hardly be appropriate.

In the short term, and in order to align with the adopted approach aimed at achieving visible progress according to relatively close milestones, the stakeholders in the experiment undertook to test a method designed to combine greater control of raw data quality and the preservation of basic algorithms: in order to limit the noise related to the sporadic connections and excessively short stays, it is necessary to measure a sporadicity rate for each of the foreign operators with a view to retaining the operators most loyal to the Orange network. One strength of this choice is that it will not be defined *a priori* on the basis of preferential roaming agreements, but will instead be measured on the ground. It will also be evolving, with the list of operators included in the counts having to be updated on a regular basis. The question of the representativeness of the different operators will arise, with the level of distinctiveness of customer profiles varying according to whether the operator is low cost, long-standing, targeted at technophiles, etc. While attractive in principle, this new version has yet to be evaluated.

Fulfilling the Initial Objectives of the Experiment Leads to Developing Processing that Stretches the Connection between Big Data and the Statistical Series Produced

At present, mobile phone data do not provide a basis for consolidating the data relating to traffic leaving the territory of metropolitan France. They cannot be substituted for traffic data and the EVE method therefore needs to be maintained in its current architecture.

The proposed solutions to improve the quality of estimates focus on sampling strategies. The selection of foreign operators with the fewest sporadic connections falls under this heading. Another possible solution is to monitor volunteers in order to achieve a better understanding of behaviours. Mobile phone operators are adept at monitoring samples of voluntary users and sell such studies more often than studies which focus on entire populations. Such studies have the benefit of avoiding some of the drawbacks faced when attempting to perform exhaustive measurements, such as the sheer volume of data and the inflexibility of anonymisation algorithms. In the case of tourism, a method such as this would generate detailed results on mobility behaviours, including, in particular, travel frequency, duration and destination. For operators with access to a network in one or several border countries, monitoring on both sides of the border may be worth considering. Beyond the statistical dimensions (extrapolation of behaviours observed in a sample of volunteers to the entire population), adopting such a method requires a legal framework suited to the processing of data relating to natural persons.

In one sense, the idea of considering a sampling-based approach is a paradoxical outcome of the experiment, the initial motivation being to exploit a source of exhaustive data in a straightforward manner. To go in this direction presupposes evaluating the sustainability, and indeed the transparency, of such an approach, given the speed at which technologies and the associated behaviours tend to change. This could generate significant costs in maintaining the sampling frame, in a context in which official statistics are dependent on their users for clear information about their methods and any changes thereto.

To conclude that it is necessary to implement a strategy which involves the sampling-based processing of Big Data implies renouncing one of their assumed strengths: namely, the rapid generation, based on raw and exhaustive sources, of highly representative and easily interpretable results. In some sense, this amounts to altering them in order to transform them into traditional data or, put differently, data whose use goes hand-in-hand with acquisition and reprocessing costs.

* *
*

The experiment conducted by the Banque de France and the DGE therefore suggests viewing mobile phone data as an additional source of information and not as a source capable of replacing currently existing data collections. The same conclusion was reached by Eurostat

[13] These data, and, in particular, the data generated by the Cerema survey for road travel, provide a point of reference for determining the survey design and serve to calibrate the statistical model ensuring the representativeness of the questionnaire data. It should be noted that visitor questionnaires remain necessary for understanding tourist behaviours: expenditure by type, type of accommodation and activities.

in its 2014 report on mobile phone data. The most pertinent uses of such data in the context of international tourism in France are short-term trend analysis and the regionalisation of data from the EVE survey. However, the monitoring of international tourism in France represents a very specific research context on account, first, of the size of the population of interest and its diversity (means of transport, country of origin, mobile phone behaviour) and, second, of the specific characteristics of the territory (borders, surface area, transit and cross-border work phenomena). Using mobile phone data to produce tourism statistics in levels remains conceivable, subject to improving the algorithms and furthering our understanding of visitor behaviour pertaining to mobile phone usage. □

BIBLIOGRAPHY

Aguiléra, V., Allio, S., Benezech, V., Combes, F. & Million, C. (2014). Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transportation Research Part C.: Emerging Technologies,* 43(2), 198–211.
http://dx.doi.org/10.1016%2Fj.trc.2013.11.007

Ahas, R., Aasa, A., Roose, A., Mark, Ü. & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3), 469–486.
https://doi.org/10.1016/j.tourman.2007.05.014

Akin, D. & Sisiopiku, V. P. (2002). *Estimating Origin-Destination Matrices Using Location Information from Cellular Phones.* Puerto Rico, USA: Proc. NARSC RSAI.
https://s3.amazonaws.com/academia.edu.documents/7109318/PuertoRicopaper_finall.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1549286310&Signature=h7qw7eNszCA7KWGLEd4lvSaLzWw=&response-content-disposition=inline;filename=Estimating_origin_destination_matrices_u.pdf

Banque de France (2015). *Méthodologie – La balance des paiements et la position extérieure de la France.*
https://www.banque-france.fr/sites/default/files/media/2016/11/16/bdp-methodologie_072015.pdf

Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M. & Smoreda, Z. (2015). Passive Mobile Phone Dataset to Construct Origin-Destination Matrix: Potentials and Limitations. *Transportation Research Procedia*, 11, 381–398.
https://doi.org/10.1016/j.trpro.2015.12.032

Calabrese, M., Di Lorenzo, Liu, L. & Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing,* 10(4), 36-44.
http://dx.doi.org/10.1109/mprv.2011.41

Calabrese, M., Di Lorenzo, G., Ferreira Jr., J. & Ratti, C. (2013). Understanding Individual Mobility Patterns From Urban Sensing Data: A Mobile Phone Trace Example. *Transportation Research Part C: Emerging Technologies*, 26, 301–313.
https://doi.org/10.1016/j.trc.2012.09.009

DGE (2016a). *Chiffres clés du tourisme.* Édition 2016.
https://www.entreprises.gouv.fr/files/files/directions_services/etudes-et-statistiques/stats-tourisme/chiffres-cles/2016-Chiffres-cles-tourisme-FR.pdf

DGE (2016b). *Le 4 pages de la DGE*, N° 60.
https://www.entreprises.gouv.fr/etudes-et-statistiques/4-pages-60-touristes-etrangers-france-2015

DGE (2017). *Le 4 pages de la DGE*, N° 71.
https://www.entreprises.gouv.fr/etudes-et-statistiques/4-pages-71-touristes-etrangers-france-2016

Eurostat (2014). Feasibility Study on the Use of Mobile Phone Positioning Data for Tourism Statistics. *Consolidated Report Eurostat Contract* N° 30501.2012.001-2012.452
http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf

Gonzales, M. C., Hidalgo, C. A. & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature,* 453(7196), 779–782.
https://doi.org/10.1038/nature06958

Kroon, J. (2012). Mobile Positioning as a Possible Data Source for International Travel Service Statistics. United Nations, Economic Commission for Europe, Geneva, Switzerland, 31 October-2 November 2012, *Seminar on New Frontiers for Statistical Data Collection*. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP6.pdf

ONS (2016). Statistical uses for mobile phone data: literature review. *Methodology working paper series* N° 8. https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesno8statisticalusesformobilephonedataliteraturereview

Organisation mondiale du tourisme. *Comprendre le tourisme : glossaire de base*. http://media.unwto.org/fr/content/comprende-le-tourisme-glossaire-de-base

Tourisme & Territoires. *Zoom sur le projet Flux Vision Tourisme*. http://www.tourisme-territoires.net/zoom-sur-le-projet-flux-vision-tourisme/

Widhalm, P., Yang, Y., Ulm, M. Athavale, S. & Gonzales, M. C. (2015). Discovering Urban Activity Patterns in Cell Phone Data. *Transportation*, 42(4), 597–623. https://doi.org/10.1007/s11116-015-9598-x

Estimating the Residential Population from Mobile Phone Data, an Initial Exploration

Benjamin Sakarovitch*, Marie-Pierre de Bellefon*, Pauline Givord, and Maarten Vanhoof*****

Abstract – Many studies are focused on using data derived from mobile phones to construct statistical indicators. Mobile phone data have the advantage of providing information with both high spatial resolution and at high frequency, allowing applications such as measurements of the spatial or temporal details of population presence. Nonetheless, using mobile phone data to construct statistical indicators raises difficulties: data from a single operator are not representative of the whole population and they often lack socio-demographic detail, which limits their quality for many applications. This article is based on a database of mobile phone records from subscribers collected by a large French operator. It aims to offer a view on the potential, but also the problems posed by mobile phone data, specifically by illustrating how indicators of residential populations can or can not be estimated from them.

JEL Classification: C55, C81, R23
Keywords: mobile phones, call detail records (CDR), present population

Reminder:

The opinions and analyses in this article are those of the author(s) and do not necessarily reflect their institution's or Insee's views.

This research was carried out within the framework of a tripartite agreement between Orange, Eurostat and the Insee's Directorate of Methodology and Statistical and International Coordination. The authors would particularly like to thank Zbigniew Zmoreda and Cezary Ziemlicki for their welcome to the Sense laboratory and their support with the use of mobile phone data, the participants of the Eurostat Big Data Task Force and especially Michail Skaliotis and Fernando Reis, as well as Elise Coudin and Vincent Loonis for their valuable advices. The authors remain solely responsible for any errors or omissions that may remain in the article.

** Insee (marie-pierre.de-bellefon@insee.fr ; benjamin.sakarovitch@insee.fr)*
*** Insee, Crest (pauline.givord@oecd.org)*
**** Open Lab, Newcastle University / Orange Labs (m.vanhoof1@ncl.ac.uk)*

Received on 20 February 2018, accepted after revisions on 5 October 2018
Translated from the original version: "Estimer la population résidente à partir de données de téléphonie mobile, une première exploration"

To cite this article: Sakarovitch, B., Bellefon, M. (de), Givord, P. & Vanhoof, M. (2018). Estimating the Residential Population from Mobile Phone Data, an Initial Exploration. *Economie et Statistique / Economics and Statistics*, 505-506, 109–132. https://doi.org/10.24187/ecostat.2018.505d.1968

The use of Big Data, linked to rapid advances in the capability to store and analyse huge volumes of data, has expanded significantly over the last decade. Big Data, created by the digital trails generated by the activities of individuals or companies, are often studied from the viewpoint of predictive analysis or to support decision-making. Another use is that they can also serve as source of observations useful to the construction of statistical indicators, which explains the interest shown in these data by official statistics institutes.[1] The expected opportunities for the use of Big Data in official statistics would be to reduce publication times by taking advantage of the very rapid access to useful information (e.g. in the field of economic analysis), but also to produce more detailed statistics (in particular, geographically) than the ones currently based on survey data, and finally to reduce the workload of collecting information from people and companies. As an example, automatic price gathering (from e-commerce sites or from invoicing data of major retailers) is used by several statistical institutes to construct consumer price indexes.[2] The use of alternative or additional sources to "conventional" data is subject to multiple studies, although the idea is certainly not new. Notwithstanding that official statistics have been complementing statistical surveys with government sources for decades now (e.g. for a long time Insee has used its statistical tracking of salaries on employers' social security returns), the integration of Big Data sources raises new, specific questions such as technical issues regarding data in large volumes or unstructured formats.

Data from mobile phones form part of the sources identified as particularly promising to supplement statistical information. Such data consist of regular records for the location (or at least the location of the cell tower the phone is connected with) date and time of phones belonging to the subscribers of a mobile phone operator. As such, mobile phone data can provide information on population presences at specific locations over specific time periods, and this at a very fine levels of geographical and temporal precision. Although official statistics produce information about residential population (especially by means of the census), access to the fine detail of mobile phone data would make it possible to detect the number of people who are at a given moment (which depends, for example, on tourist visits, business behaviour, etc., see Terrier, 2009), as well as the movements of people between several points. Regularly locating subscribers thus enables the mapping of population presence and the way it changes (Deville *et al*., 2014; Debusschere *et al*., 2016; Ricciato *et al*., 2015). For example, these data can be used to measure the variability in visitor numbers to certain places during the day or during the year, to improve precise knowledge of travel times using different means of transport (in particular for "small" daily journeys) and to draw up detailed mobility matrices (see Aguiléra *et al*., 2014, for evaluating performance of the Île-de-France transport network or Demissie *et al*., 2014, for Senegal). The visitor profiles of an area at different moments in time can assist the analysis of regional dynamics. Since we can expect presence (or activity) profiles to change during the day depending on the type of place (home, workplace or travel hub), Toole *et al*. (2012) were able to distinguish the main activity of areas, depending on the daily presence profiles observed mobile phone data (e.g. shops, residential, industrial or car park) across the Boston area. For France, Vanhoof *et al*. (2017) applied a similar approach at municipality scale, and revealed a correlation between aggregated activity profiles of mobile phone cell towers and the type of communities they are located in, as defined by the French statistical office's (Insee) zoning of urban areas. Ultimately, the information from mobile phone data can also be used to enhance the analysis of interpersonal networks, for example by analysing the strength of communications between subscribers or regions (Grauwin *et al*., 2017).

Nevertheless, using mobile phone data raises several questions. Firstly, it is necessary to guarantee respect for subscribers' privacy. Being able to reconstruct individual journeys using the trails left by subscribers creates a risk of "re-identification". Even by deleting all direct mentions about their identity, from a certain point onwards it is possible to attribute an observed journey to a single person with high probability (Montjoye *et al*, 2013). This requires that mobile phone data be aggregated at an adequate level to prevent individual identification, or that privacy will be protected by procedures that do not allow for practitioners to have direct access to sensitive data. The former solution has the disadvantage that it reduces information and relevance of the data, whereas the latter requires new platforms and procedures to be implemented with regard to most present-day

[1] e.g. see the Scheveningen Memorandum (2013)

[2] In France, the "checkout data" project is based on price records taken from invoicing data from several large retailers (see Leonard et al., 2017, *and* Economie et Statistique / Economics and Statistics *N° 509 forthcoming).*

situations.[3] Secondly, in technical terms, mobile phone data for millions of subscribers represent huge volumes that require suitable storage and computation infrastructures.

Notwithstanding the questions raised, statistical offices are interested in the potential of this type of Big Data. For example, a Eurostat report (2014) studied the potential of mobile phone data to improve the accuracy of current tourism indicators. Additionally, several national statistical institutes have launched initial experiments in using different Big Data sources and a coordination programme was launched in 2016 to share knowledge on this subject.[4] One central element is access procedures for official statistics institutes that cover both subscribers' privacy and business confidentiality for the companies involved. For France, a CNIS report offers guidelines on reusing company data in official statistics (2016), specifically highlighting the case of mobile phone data.[5] Simultaneously, other European official statistics institutes have begun negotiations with national operators and are now engaged in experimental projects (Debusschere *et al.*, 2016, for Belgium). Such experiments are needed to define what information at what level of aggregation is needed to construct relevant statistical indicators (Vanhoof *et al.*, 2018).

In the case of mobile phone data, experiments have raised multiple questions. Firstly, using mobile phone data from one operator raises questions of representativeness. Access to an operator's data will only supply information about its subscribers, who only makes up part of the population. Understanding this bias requires additional information, such as the local coverage of these operators, which is necessary to obtain more detailed spatial statistics. Additionally, the level of mobile phone ownership can vary depending on population characteristics: Some people may not have a mobile phone – e.g. Wesolowski (2013) highlighted problems of the unequal distribution of telephones in different social groups in Kenya for the use of this type of data, while others may have several mobile phones.

A second difficulty in using mobile phone data relates to the grid of cell towers, which in principle does not match normal geographical grids (e.g. administrative subdivisions). Cell towers are not distributed uniformly – there are more in densely populated areas and fewer in rural areas. To use them across more traditional territorial units, translations from the cell tower grid need to be made, which introduces approximations (Ricciato *et al.*, 2015).

Finally, it is essential to clarify what can be measured from mobile phone data. These data are produced "naturally" (sometimes called "organic data", as opposed to "designed data", supplied using surveys constructed with the aim of measuring the study object[6]), they simply reflect the trails left by subscribers on the mobile phone network. For a statistical indicator to have a meaning everyone can understand, it is essential first to agree a definition of what we want to measure. For example, a tourist is generally defined as a person registered "outside their usual environment". Tourist visit measurements for a place therefore require distinguishing, among people present in this place, those who do not live there but also those who do not work there regularly. To measure this information from records of subscribers' journeys require being able to identify a person's home or even their "usual" workplace (Janzen *et al.*, 2018). Several studies on this question have been conducted based on mobile phone data. For example, Ahas *et al.* (2010) showed that it is possible, using trails left by an individual on the network, to reconstruct their "anchor places", i.e. places important to them, where they go repeatedly – their home and workplace being the most obvious of them (Ahas *et al.*, 2010). As also emphasised by Song *et al.* (2010), the time spend by each person is generally concentrated on a limited number of places. Several algorithms have been suggested to identify a subscriber's likely home from observed journey profiles (Vanhoof *et al.*, 2017; Bojic *et al.*, 2015; Isaacman *et al.*, 2011). This point is essential as it is a prerequisite to many other analyses (Blondel *et al.*, 2015) that go beyond the simple scope of tourism.

This study is offered to, based on a practical example, illustrate the empirical questions raised by the use of mobile phone data. The study will use mobile phone data from subscribers to a French telephone operator over the course of five months in 2007. It will try to

[3] *For example, the Opal project (http://www.opalproject.org/about-us/) offers providing researchers a platform to run algorithms on mobile phone data to which the researcher does not have direct access: we are talking about Open Algorithm rather than Open data.*

[4] *https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP5_Mobile_phone_data.*

[5] *See Cnis-Insee report "Reuse of Company Data by the Public Statistical System".*

[6] *In particular, this distinction has been suggested by the Census Bureau, see https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html.*

estimate resident population figures from this data, compare them with estimates from official statistics taken as reference data, and analyse sources of discrepancy. This methodology makes it possible to test the relevance of several home-detection algorithms and several data aggregation techniques: two essential questions when you want to use mobile phone data.

The rest of this article is organised as follows. The first section describes the different types of mobile phone records and details how to convert them to a localised population count. A second section discusses the coverage of cell towers and their translation to administrative divisions. The next section presents the different methods used to estimate resident populations. It gives details about the representativeness questions and suggests solutions that can be used to resolve them, as well as comparisons with reference sources. Finally, the last section suggests some other ideas to use mobile phone data to characterise population presence dynamics.

Data Records

A mobile phone network enables communication by transmitting radio waves between devices, repeater towers and the operator's centralised switches that direct the connection to other repeater towers for the person being called. These networks have a cellular structure, i.e. each cell tower covers a certain area and a telephone can change cell without the communication being cut off.

Principle of Mobile Phone Records

The data used here are records by "repeater towers" of the cellular network, which report the presence of subscribers' cellular telephones near these cell towers. They are mounted on towers with known coordinates. In principle it is therefore possible to construct indicators about visitor numbers to certain places, or very finely detailed geographical and temporal mobility behaviour. The frequency and regularity of these records, and therefore the level of detail (granularity) at which we will be able to construct these indicators, depends on the data type. There are several data types.

CDR (Call Detailed Records) relate to making or receiving a call or an SMS, i.e. a deliberate action by the subscriber. We therefore call them active data. These data as generally used for invoicing and operators therefore recording them "by default". In France, operators have to retain these data for six months. Besides indicating the location of subscribers, these data can be used, for example, for studies on user behaviour (call frequency, preference for text messages, etc.).

Signalling data, what we will call passive data, are generated from telecommunication and internet networks (2G, 3G, 4G), using the fact that all mobile telephones connect regularly to the nearest cell towers (with variable frequency that can range from three hours to ten minutes) without necessarily arising from the user's action on the mobile. They therefore provide more complete information than CDR, for example if you want to measure the number of visits to a place at a given moment or track people's movements. However, processing these data is more expensive. By default, these "events" are not recorded by operators: to do so requires very large storage capacities.

In terms of population coverage, the data recorded by an operator, whether active or passive, relate only to their subscribers. However, there may be "roaming" agreements that enable one operator's subscribers to use its competitors' networks when they are outside the area covered by their own operator. In France, there are few roaming agreements between the national operators, and this "roaming" situation essentially relates to foreign subscribers. In particular, this means that it is possible to identify people only passing through France, as long as they are using their telephones (for CDR data) or they at least leave them switched on (for signalling data). The SIM card makes it possible to identify the telephone operator's home country, from which the telephone subscriber's probable nationality can be inferred.[7]

Approach to convert records to population counts

A series of processing operations is needed to derive information useful for official statistics from data recorded by the mobile network (Diagram).

[7] Before June 2017, these overseas roaming costs were invoiced by the operators. Since this date, the European Commission required such invoicing to end. It is possible that this will ultimately lead to creating a more competitive European market, as the nationals from one country are more easily able to use a foreign operator and it will therefore be more difficult to identify these journeys.

Box 1 – **Description of Mobile Phone Data Used**

The study relates to using an anonymised file of CDR data (Call Detailed Records) containing the complete record of subscribers' activities for the operator Orange across mainland France over a five month period, from mid-May to mid-October 2007[(a)]. The records cover about 18 million SIM cards and more than 20 billion observations. These data contain no direct information about the subscriber's name or address. However, for the study, it was possible to supplement them with certain information taken from a Customer Relationship Management (CRM) file, designated in the article as "customer file". For 12.4 million SIM cards also included in the CDR (i.e. about two-thirds), this customer file indicates the *départements* in which subscribers have stated they live. The subscriber (as identified in the customer file) is not necessarily the telephone user. Imagine for example the case of parents funding the subscription for mobile phones used by their children. Furthermore, information from the customer file can "expire", such as when moving home, as the information is not always updated.

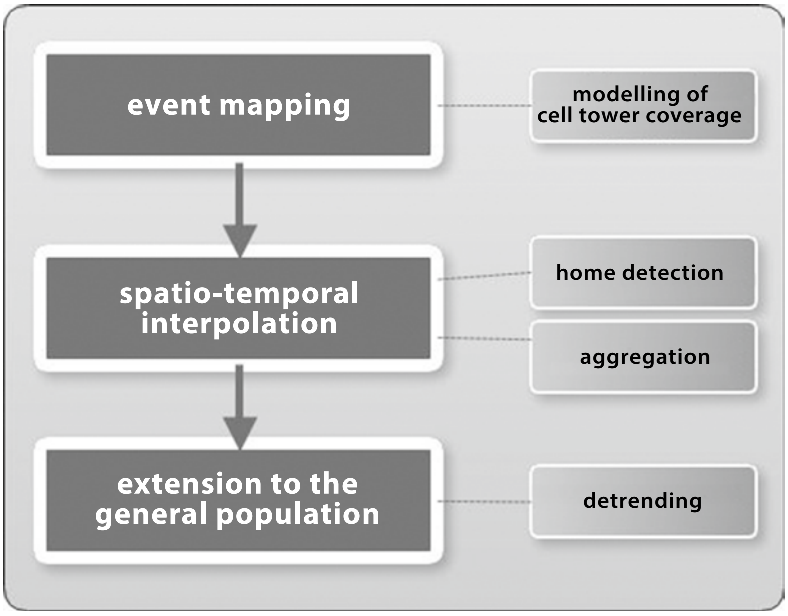
(a) These anonymised data are available to the Orange Labs SENSE laboratory, for purposes of research projects.

Table A
Structure of Call Detail Records and Essential Variables

Sending SIM card	Receiving SIM card	Event type	Sending cell tower	Receiving cell tower	Timestamp	Duration
SIM-1	SIM-2	Call	A-1	A-2	13/06/2007 -14:26:03	7m32s
SIM-1	SIM-3	SMS	A-3	-	25/08/2007 -12:04:58	-

Note: For SMS messages, we don't know the cell tower through which the message is received.

Diagram
Diagram to Show Processing of Mobile Phone Data for Oficial Statistics


event mapping
modelling of cell tower coverage
spatio-temporal interpolation
home detection
aggregation
extension to the general population
detrending

The first step is mapping events recorded on the network (calls or SMS). The location of the event is inferred from information available on the location of the cell towers. To practically perform the mapping, you have to define a spatial grid on which you want to locate the different events, and secondly model the area of cell tower coverage (in particular, based on the technical characteristics of the cell towers, if they are available, see Ricciato *et al.*, 2017). As detailed in the section "A Very Non-Uniform Grid", we use the simplest way of modelling cell tower cover areas by using the Voronoï tesselation of the cell towers (see Box 2). Based on this coverage model, the event will then be located on the chosen spatial grid.

The second step is to perform a spatio-temporal aggregation to convert the record of events into aggregated data matching a predetermined definition. This consists of defining aggregation units (both temporal and spatial) to produce statistical indicators. For example, we may want to construct indicators of populations present in places based on traditional administrative subdivision (by IRIS, municipality, etc.) at specific moments of the day, or at least over fixed time periods. The grid used to convert from cell towers to places, related to their technical characteristics, does not naturally match the conventional territorial subdivision. It is therefore necessary to perform a spatial interpolation. In some case, this spatial interpolation must be coupled with a temporal interpolation, as the records from SIM cards have neither defined nor regular frequency: for example, based on call activities, we may have the location of the same telephone at 7:47 am then at 8:12 am, however the location of this same telephone at 8 am is not directly known. If the aim is to measure the population over specific times, it will be necessary to reconstruct the probable location at 8am from these available data. Finally, to estimate the resident population indicators, we must try to infer the probable home location, based on the times and locations of available in the data. The home detection algorithms that perform this step are described in a next section.

A final step seeks to obtain estimates for the reference population (the entire population of France), based on the aggregated data subscriber counts from moble phone data. This aggregation is supported by external sources (e.g. operators' market share). Several possible estimates are presented for the reference population, depending on the depth of additional information available, stressing the underlying hypotheses. These results are compared to reference statistics (resident populations, such as measured by taxation sources processed by Insee).

Approximation of Cell Tower Coverage: Simulation from Taxation Data

A Very Non-Uniform Grid of the Country

Spatial coverage across the country is uneven. For each operator, the repeater towers that supply the main information about location are sited unevenly across the country. As shown in Figure II, in 2007 the cell towers of the operator Orange were very densely distributed in urban areas but much less densely in rural areas. Furthermore, mobile infrastructures can boost the network locally to prevent it becoming saturated during events leading to large crowds – sporting events, concerts, demonstrations. In more structural terms, the development of technologies (successive releases of 2G, 3G, 4G, etc.) leads to renewing the network and therefore changes to the location of cell towers.

In practice, we can infer the probable position of a telephone from the cell towers to which it is connected. The simplest solution is to assume that it is connected to the nearest cell tower.[8] We can define subdivision of the country using a Voronoï tessellation (Box 2), which matches each cell tower to all the points in space that are nearest to it. This model of coverage is an approximation of the actual coverage of cell towers. It does not take into account that in the real world coverage areas overlap and that the load of telephones present in a given area is split among the various cell towers covering it. Still, in our simplification we consider the Voronoï polygons of all cell towers as our spatial unit of observation. Due to the unequal distribution of cell towers across the country, the areas of these polygons are very variable in size (Figure I). Figure II shows the distribution of their areas. We can see that while many Voronoï cells have quite areas small area (a few hectares), the range of areas covered is very broad and goes up to more than ten thousand hectares for some cells. These large areas do not correspond to the effective coverage of the cell towers but arise from the Voronoï tessellation in regions where the cell towers are a long way from each other and can even actually include "white" areas where no signal is received.

As confirmed by Figure VI-A, the smallest Voronoï cells are located in the most densely-populated areas.

[8] *This is an approximation based on the assumption that cell towers all transmit with the same power and in all directions. In reality, one mast can hold several aerials transmitting in transmission directions (all over the place) and with different ranges. Scholus (2015) or Tennekes (2015) constructed an inference model for the position of the mobile based on the detailed observation of the properties of cell towers, as well as knowledge of the distance between the telephone and the cell tower that retransmitted the signal. However, this information (properties of cell towers, distance to the telephone) is not always available in the data. Furthermore, having very frequent information can permit triangulations that make it possible to identify the position of a mobile precisely. In the ideal case, where the distances to several cell towers (at least 3) are reported, it is possible to use triangulation to deduce the exact position of the telephone.*

Translating Voronoï Cells to Another Grid

The purely technical geometric partitioning of the space by Voronoï polygons obviously does not coincide with the subdivisions of the country used for circulating regional statistical data. Indeed, there is no reason for cell tower coverage to correspond to administrative boundaries of municipalities or *départements*, nor should they be contained within the finer grids used by official statistics, such as IRIS (the building blocks for circulation of infra-municipal information, interlinked within the community geography and forming uniformly-sized units in terms of population[9]). As a consequence, it requires translation to a

[9] https://www.insee.fr/fr/metadonnees/definition/c1523

Box 2 – **Partitioning the Space, the Voronoï Tessellation**

The Voronoï tessellation is a partitioning of the space based on a set of given points: the seeds. Each point on the plane is allocated to the seed to which it is closest. The boundaries between the different areas of the plane form the sides of polygons containing exactly one seed.

This subdivision of the plane is useful for processing mobile data when you only know the locations of the various cell towers (which therefore form these seeds). We then assume that a call is transmitted using the nearest cell tower, which therefore means that the telephone is located in the Voronoï polygon associated with this cell tower.

Figure A
Example of a Tessellation Using Voronoï Polygons Derived from 7 Points

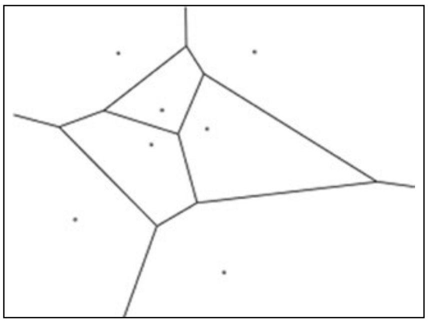
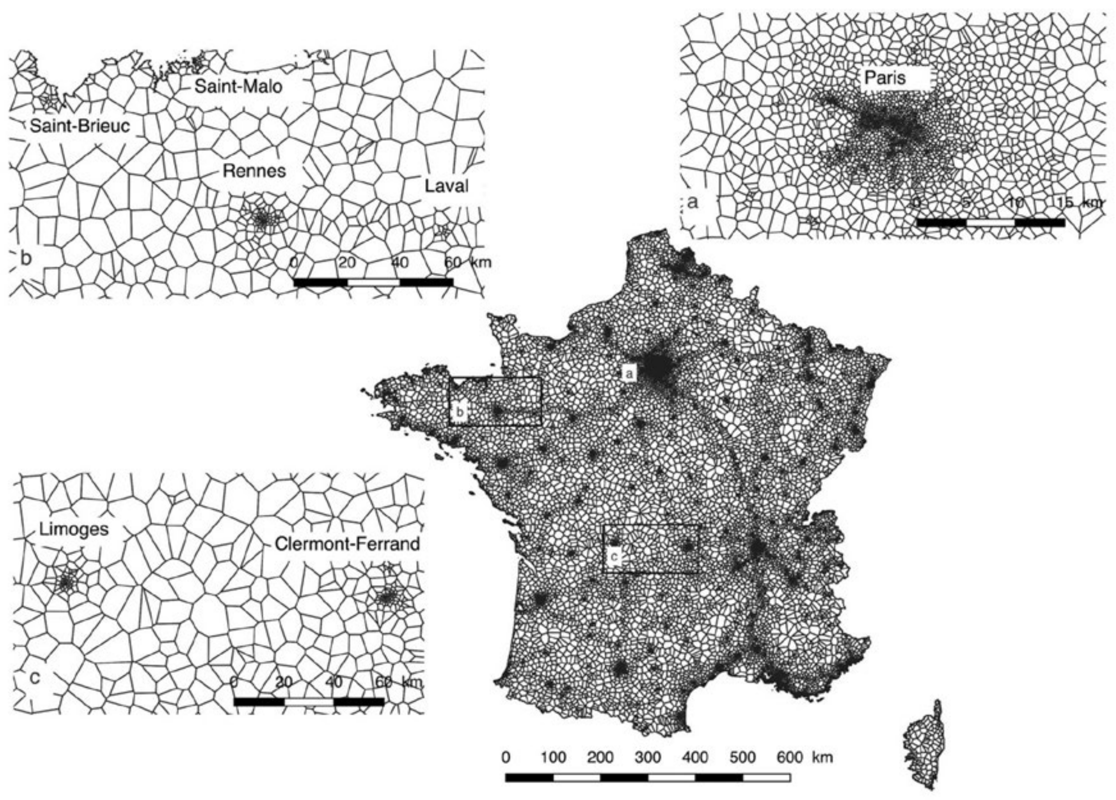
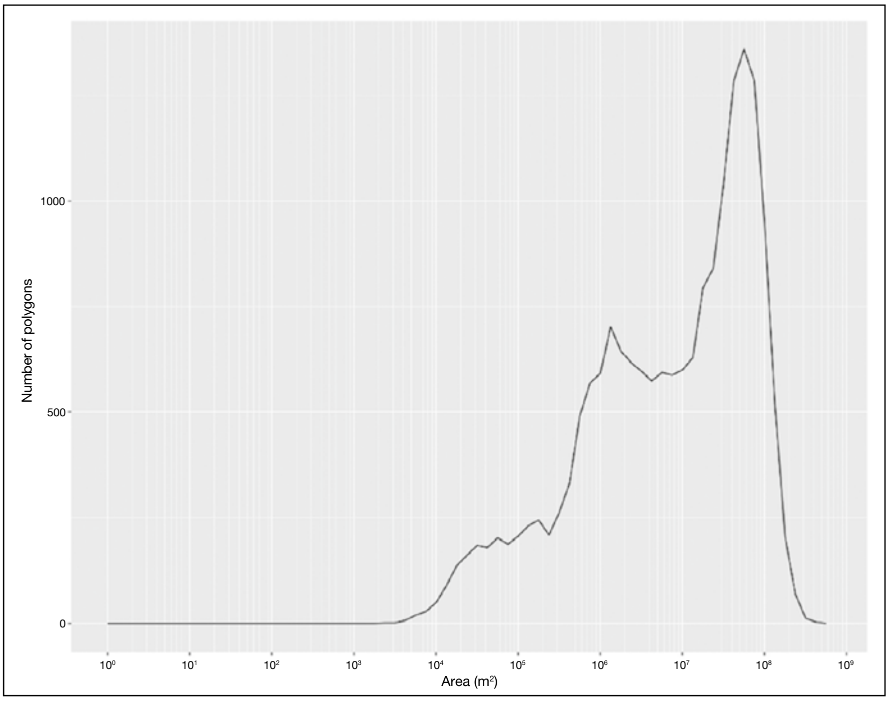


Figure I
Voronoï Polygons Associated with Cell Towers for the Operator Orange, Metropolitan France



Sources: Orange CDR.

Figure II
Area Distribution (in m²) of Voronoï Polygons Associated with Cell Towers for the Operator Orange



Reading note: The modes of the distribution are at 10^6 and 10^8 m² (the graph is plotted on a logarithmic scale); there are no polygons with area less than 10^3 m².

conventional administrative grid to estimate regional statistics from mobile telephone data and compare them with information provided at the scale of this administrative grid.

In what follows, and due to the lack of better information, this translation will be done simply weighting the areas of the polygons as they are situated in the administrative grid. The base administrative grid chosen is the municipal grid, divided into *arrondissements* (districts) for Paris, Lyon and Marseille. Subscribers' counts for the administrative grid will correspond to the sum of the estimated subscribers in the Voronoï polygons that are entirely enclosed within a unit of the administrative grid and the number of estimated subscribers weighted by the proportion of the areas of the Voronoï polygon covering an administrative unit in the case that Voronoï polygons overlap several administrative units (see also equation 1).[10]

$$N_c = \sum\nolimits_{V_j} \frac{A_{V_j \cap C}}{A_{V_j}} N_{V_j} \qquad (1)$$

Where N_C represents the estimated number of subscribers in administrative unit C, N_{V_j} the number of subscribers detected in Voronoï polygon V_j, A_{V_j} the area of this Voronoï polygon, and $A_{V_j \cap C}$ the area of the intersection between the administrative unit and the Voronoï polygon.

Within equation 1, we base on the assumption that the population density present is uniform over the whole polygon. This assumption is obviously debatable, in particular in rural areas where dwellings are typically more concentrated. In the next section we evaluate the impact of the translation from Voronoï grid to the administrative grid by reproducing it for a conventional official statistics dataset, namely Tax files. Since tax files are complete (available for the entire population) and geolocated, they serve us well to investigate the effect of translating between both grids.

Simulating the Approach on Tax Data to Evaluate the Scale of the Approximation

Insee has complete information about the resident population at regional scale. The "Localised Social and Tax File" (Filosofi), which replaces and supplements the "Localised Tax Revenue File" (RFL), is made up from complete files of physical persons' tax returns

[10] *https://www.insee.fr/fr/metadonnees/definition/c1523*

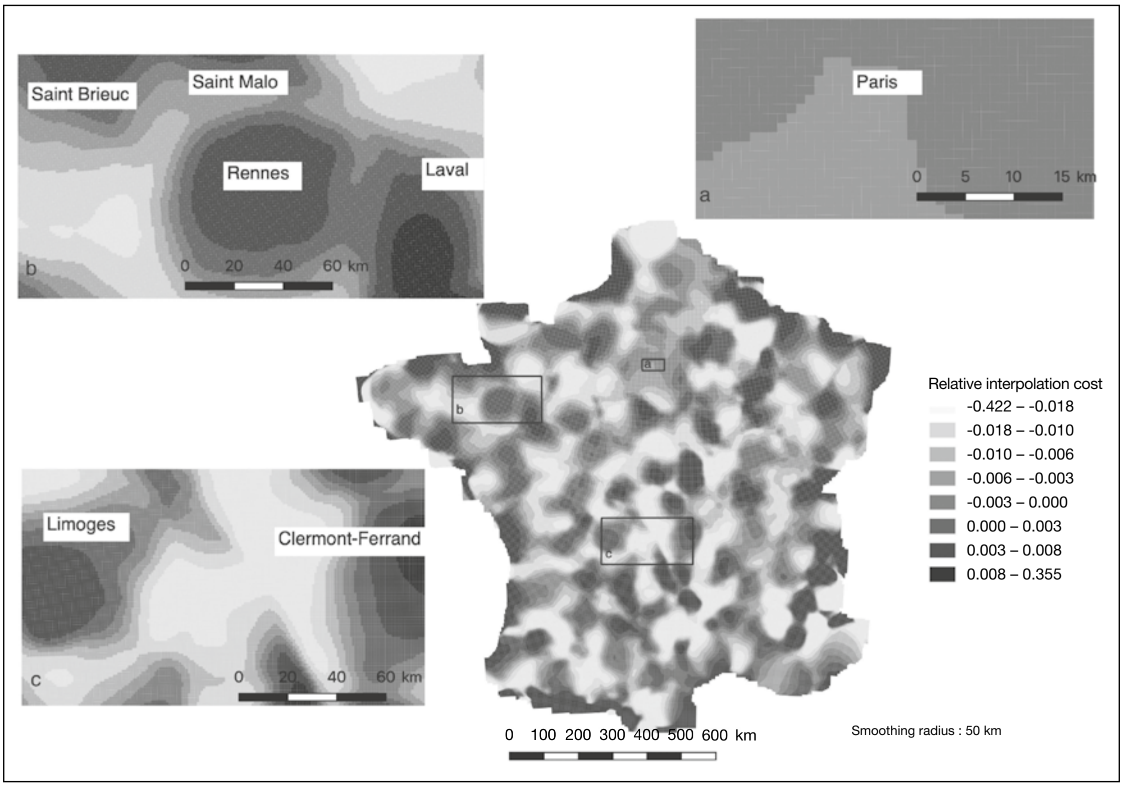
and local housing tax. This information is available in even greater detail than the mobile telephone data, since it is geolocated.[11] However, temporal accuracy is much less since this information is produced annually. Furthermore, these tax files only provide information about where people live and not about their actual presence in certain places (which can vary during the day). Nonetheless, they may constitute an interesting source of comparison to evaluate the suitability of mobile telephone data to reconstruct conventional statistical indicators, such as population density.

From the geolocated tax data, we estimate the spatial distribution of the number of inhabitants for municipalities and Voronoï polygons. Similar to previous section, we translate the information at the Voronoï polygon grid to the administrative grid and compare the outcome with the direct estimate at administrative level (municipality, in this case). The term "interpolation cost" describes the measurement error contributed by the translation between both grids. In practice, this is the difference between the number of inhabitants measured directly for the administrative unit and the estimate obtained from the translation (normalised to the reference).

Figure III illustrates the distribution across the country of the interpolation cost. For municipalities located in rural areas, this interpolation generally leads to overestimating the municipal population. The grid translation is actually based on the assumption that the population density is uniform over an entire polygon. The grid of cell towers is looser in less densely-populated areas. The corresponding

[11] https://www.insee.fr/fr/statistiques/fichier/2520034/donnee-carroyees-documentation-generale.pdf

Figure III
Relative Difference between the Municipal Population and the Population Estimated by Grid Translation (Interpolation Cost)



Reading note: The interpolation cost is the difference between the municipal population obtained directly in the tax source and that estimated from spatial interpolation (using equation (1)). A negative interpolation cost corresponds to an overestimate of the municipal population, a positive interpolation cost is an underestimate).
Sources: Filosofi 2011; authors' computations.

polygons therefore cover a larger surface area, even if dwellings are more spread out – this makes the underlying assumption all the less plausible.[12] We find these differences when we estimate the effects by size of municipality. For municipalities smaller than 10,000 inhabitants, the relative difference related to spatial interpolation is an average overestimate of 53% (see Figure C1 in the Online complements[13]). Conversely, for municipalities larger than 10,000 inhabitants, spatial interpolation rather tends to underestimate the actual municipal population – nonetheless the relative differences are smaller (without ever being negligible): they are 10% on average.

The results suggest that using a grid that does not superimpose directly over the "conventional" subdivisions is a significant factor for the quality of estimates produced from these data. One solution would be to set aside the administrative subdivision by considering Voronoï polygons as the base unit, but it has the disadvantage of being based on a grid – that for the cell towers – that is neither stable over time nor uniform in space. This partitioning of the space is also based on an approximation of the true cell tower coverage, which probably affects the quality of the results obtained. In reality, antennas on cell towers are directional and only provide coverage up to a certain distance. This also explains the presence of "white areas" mapped by ARCEP since 2017.[14] In addition, their areas of coverage are very often super-imposed, unlike a tessellation. Having this information about the technical capabilities of cell towers would make it possible to refine the actual partitioning of the corresponding space. For example, exploratory work by the Dutch Central Bureau of Statistics (CBS) proposes using a Bayesian inference procedure to allocate and point in space to one or other of the nearly cell towers, based on their power and orientation.[15] Future work will be able to reveal the benefit gained in terms of accuracy and the cost in terms of complexity. But the data we are using does not contain the technical information needed for this exploration. Furthermore, as discussed below, other problems are raised by using mobile phones, which result both from the definition of a concept (how to convert the record of a telephone call in the management data to a statistical indicator?)[16] and from that of their statistical treatment (how to obtain representative estimates of the whole population from the subscribers to a single operator?).

Constructing Statistical Indicators from Data

Home Detection

The data we have correspond to the trails left by subscribers during their journeys. In principle, these recurrent journeys indicate the use of places specific to the subscriber and so it seems possible to infer subscribers' likely home, or workplace, or other places important to them. Such information is useful or even essential to construct certain statistical indicators, such as home/work journey times or tourist numbers in certain regions. Regarding tourists, for example, they are defined according to the "statistical" definition established by the United Nations World Tourism Organization Statistics Department, stating that tourism is "the activities conducted by people during their travels and stays in places located outside their usual environment for a consecutive period not exceeding one year, for leisure, business or other reasons". While the usual environment can be interpreted to vary in size, it includes at least the home and workplace. This information is rarely available in the anonymised files to which researchers or statisticians have access and thus several home detection algorithms have been proposed that try to infer them from mobile phone data.

The general principle of home detection is to define the home from criteria based on the frequency and/or times (generally the night) users are present in a place. Vanhoof *et al.* (2018) offers a review of several home detection methods. We extract five methods to be used here:

– "Maximum activity": Home is the place where most events (making and receiving calls or SMS) take place during the period under study;

[12] Finally, it should be noted that this consists of differences related to the number of inhabitants in the municipality – numerical differences can be amplified for very small municipalities.

[13] See the link to Online complements at the end of the article.

[14] The site https://www.monreseaumobile.fr/ can be used to see the white areas by network and operator.

[15] This work is accessible from the available mobloc R package R, the address: https://github.com/MobilePhoneESSnetBigData/mobloc. It is also described in Dutch here: https://circabc.europa.eu/webdav/CircaBC/ESTAT/ESTP/Library/2017 ESTP PROGRAMME/46. Advanced Big Data Sources - Mobile phone and other sensors, 6 – 9 November 2017 - Organiser_ EXPERTISE FRANCE/Mobile_Phone2.pdf.

[16] A statistical indicator here means the quantification of a social reality (e.g. the population present), based on a convention to be defined (for Desrosières (2008), "to quantify means to agree then to measure").

– "Number of distinct days": Home is the place where activities were recorded on the largest number of distinct days during the period under study;

– "Time constraint": Home is the place where most activities between 7 pm and 9 am were recorded during the period under study;

– "Spatial constraint": Home is the place where most activities were recorded within a 1 km radius around the cell tower during the period under study;

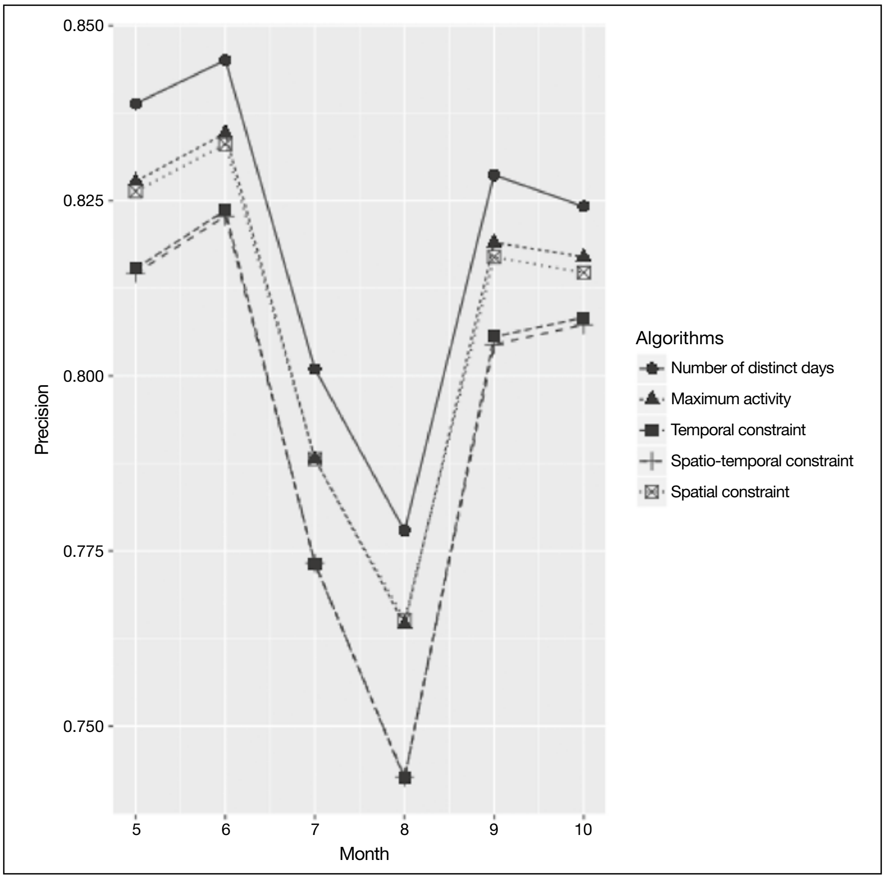
– "Spatio-temporal constraint", a combination of the previous two.

These different algorithms are all reasonable guesses. Nonetheless, they all have their limits and, for each, we can also easily think of situations where the identification of home will be incorrect. For a given subscriber, different methods can identify different places as the likely home.

To evaluate the performance of the different home detection algorithms, we have additional information supplied by the customer file, which contains the post code of the subscriber's home. This information is only available for two-thirds of the subscribers, but nonetheless it is possible to compare this post code with the estimate of home supplied by the different algorithms. Furthermore, we also have data on the resident population supplied by the localised tax files.

Figure IV shows the accuracy of the five proposed home detection algorithms, by comparing their residential population estimations with information from the customer file. The estimates are made for each month and by *département*. The accuracy corresponds to the proportion of subscribers included in the customer file for whom we have correctly identified the *département* in which they live (in the sense of it matching that in the customer file). Over the entire study period, the algorithm

Figure IV
Accuracy at *Département* Level of Home Detection Algorithms Based on the Customer File



Note: Accuracy correspond to the proportion of subscribers included in the customer file for whom the localisation algorithm determines the same *département* as the customer file.

using the number of distinct days (i.e. the place where activities were recorded on the largest number of distinct days) performed the best. Even at this aggregated regional level,[17] we see that the difference between the home *département* as identified by the algorithms and as stated in the customer file remains large (never less than 15%). The discrepancies can partly be explained by the inadequacy of the heuristic home detection methods we used. For example, accuracy falls significantly in summer and can very probably be explained by the fact that a significant proportion of the population is on holiday at that time and do not reside in their usual *département* for the whole month. This difference may also be linked to a quality problem with the customer file. Even ignoring the summer months, we see reduced accuracy over the whole period for all the algorithms (the differences observed in September-October are greater than those observed in May-June), which may be due partly to an effect of the customer file ageing (e.g. updates not made when subscribers move house). In addition, the data only contain records for the end of May (18 days) and the beginning of October (14 days), which may also explain the poorer performance than in June and September, respectively.

Additionally, it is worth noting that a user is considered as having a home in a *département* if the cell tower allocated to them by the home detection algorithm is within the *département*. There can be marginally edge effects for cell towers with Voronoï cells that overlap several *départements*. The Online complements contain maps representing the geographical distribution of this accuracy for June and August.

Adjusting Data to Obtain Estimators of Resident Population

The mobile phone data available to us relate to a single operator's subscribers only. To estimate statistics on the entire French population it is therefore necessary to perform detrending. This detrending should make it possible to convert from the subscriber population to the total population, which may differ for two reasons. The first is that the operator only covers a proportion of mobile phone subscribers. This operator's market share indicates the order of magnitude of the relative difference that we expect to find between the actual population and "raw" estimates obtained with mobile phone data. According to the *Autorité de régulation des communications électroniques et des postes* (Arcep – French electronic and postal communications regulator), the national market share of the operator Orange in 2007 was 46.7%.[18]

The second reason is that there is no simple correlation between the population of physical people and that of SIM cards. All physical people do not own a telephone (such as very young children) and conversely some have several (especially for business reasons). So we have to consider the penetration level, i.e. the ratio of the number of telephones over the reference population (the population at 1st January of year $N-1$ published by Insee). For example, in 2007, the number of portable telephones per inhabitant estimated by Arcep was 85.6% across all of mainland France. It was 81.6% for the Rhône-Alpes region but only 66.0% in Franche-Comté. In two regions, Île-de-France and the PACA region, these levels were even higher than 100% (122.3% and 104.3%, respectively).[19] Part of these differences may be linked to the characteristics of the populations. For example, the CREDOC digital barometer shows large disparities based on age in 2007: Nearly all of 18-24 year olds were equipped with a telephone while this was only true for a third of the over-70s.[20]

Formally, converting the number of subscribers N_{HD_i} identified as residing in a given spatial unit i (from the home detection algorithm – HD – corresponding to the number of separate days, the most effective according to the results above) to the resident population in this unit is supplied by the following accounting operation:

$$\widehat{N_i} == \tau_i^{-1} \cdot \alpha_i^{-1} \cdot N_{HD_i} \qquad (2)$$

where α is the local market share of the operator Orange, and τ the penetration level. These two parameters are likely to vary throughout the country, because of the Orange market share but also because of the composition of

[17] In principle, the more aggregated level is less interesting, as the interest aroused by the sources derived from mobile phone data is precisely in obtaining estimators with finer spatial granularity.

[18] See Ruling 07-0706 from Arcep dated 6 September 2007, https://www.arcep.fr/uploads/tx_gsavis/07-0706.pdf

[19] Arcep, Le Suivi des Indicateurs Mobiles – *Figures at 31 December 2007. https://www.arcep.fr/index.php?id=9545 "Geographical distribution of mainland customers".*

[20] 2015 digital barometer, available at https://www.arcep.fr/uploads/tx_gspublication/CREDOC-Rapport-enquete-diffusion-TIC-France_CGE-ARCEP_nov2015.pdf Table 2 – Proportion of individuals having a mobile telephone, p. 24.

the resident population. To obtain local estimates of residential population from mobile phone data, we therefore want to have accurate information about the variables corresponding to the detrending (at least the operator's share and penetration level) at fine geographical levels. The difficulty here is that this information is generally available at an aggregated level (national or regional). Using it uniformly over the whole country creates the risk of not being able to distinguish between actual differences in population and different market shares for different administrative units.

To investigate (and quantify) the importance of different effects on our residential population estimates, we perform detrending while increasingly adding additional information to the equation. Found estimates can then be compared to those observed in the tax source in order to understand the contribution of different effects such as market share. A first, "raw" estimate simply consists of correcting for a size effect. We simply multiply the obtained subscriber counts from mobile phone data with the ratio of the number of subscribers available in the data to the size of the residential population in mainland France (i.e. 18 million out of a total mainland population of about 62 million in 2007). A second estimate can be based on incorporating open source information on the market share and penetration rate, as was done in the previous paragraph.

A third source of information that can be added is not open-source but was available to us. It consists of the customer file, which provides an estimate of the regional distribution of subscribers. We therefore use this file to construct a detrending by *département*. This geographical level appears both sufficiently broad to reduce the problems of spatial approximation raised from using the grid supplied by Voronoï polygons, and sufficiently fine to be able to ignore the spatial heterogeneity of market shares and the penetration level mobile phones amongst the population. The number of subscribers residing in *département* k is estimated from addresses available in the customer file. As these addresses are only available for part of the file of SIM cards we have, we adjust by the size of the customer file (which comes back to supposing that the customer file's lack of coverage is uniform over the whole country). So the *département* market share simply corresponds to the ratio of this estimate of the number of subscribers residing in the *département* over the total number of inhabitants in this *département* supplied by tax sources.

$$\alpha_k \tau_k = \frac{Tot_{HD}}{Tot_{CRM}} \cdot \frac{N_{CRM_k}}{N_{Insee_k}} \qquad (3)$$

Where *k* represents the index for the *département.*

A fourth and final source of information is based on *Deville et al.* (2014) who suggest estimating the municipal population densities from equivalent mobile data and a model taking account of the "superlinear effect of densely-populated areas on human activities". We therefore use this method only as a comparison with the different ways of detrending that we are suggesting.[21]

The population is then estimated using the model:

$$N_{Insee_C} = \alpha \cdot N_{HD_C^{\beta}} \qquad (4)$$

where the parameters α and β are themselves estimated by generalised linear regression N_{Insee_C} is the number of residents in the municipality according to the tax source and N_{HD_C} is the number of people identified as resident in the municipality with mobile data.

Performing this detrending on the subscriber counts obtained from mobile phone data, we can compare the obtained residential population estimates with the aggregated tax data at different spatial scales, investigating differences in magnitude and regional distribution.

The correlation between both is measured by means of two indicators: the cosine similarity and empirical correlation coefficient (Box 3). These indicators are both independent of the size of the population involved. They therefore amount to confirming whether the estimates from mobile phone data result in residential population densities consistent with those given by the tax source.

We have measured the differences at several scales. Clearly, we will use the Voronoï polygons, which is the finest spatial scale available with mobile phone data. Because Voronoï

[21] *The model proposed by* Deville et al. *relates to population densities. The model is estimated by least squares weighted by the population of municipalities over the logarithms of densities. The interest of official statistics focuses more towards population counts. We therefore favour a model better suited to counts and we estimate the parameters by generalised linear regression based on a Poisson family (equation 4), on which a logarithmic link function is applied.*

Box 3 – **Cosine Similarity and Empirical Correlation Coefficient**

For each geographical level, we can define the vectors of observations using data taken from the tax source ($\vec{x}$) and mobile phone data ($\vec{y}$). We therefore call the empirical correlation coefficient:

$$cor\left(\vec{x},\vec{y}\right)=\frac{\overrightarrow{(x-\bar{x})\cdot(y-\bar{y})}}{\left\|\overrightarrow{(x-\bar{x})}\right\|\cdot\left\|\overrightarrow{(y-\bar{y})}\right\|} \tag{5}$$

Where $\bar{x}$ and $\bar{y}$ are the empirical means of the sample. It is also standard to use the cosine similarity, making it possible to measure if these two vectors are similar. Formally, this is the scalar product normalised to the product of the norms of the two vectors.

$$cosim\left(\vec{x},\vec{y}\right)=\cos(\theta)=\frac{\vec{x}\cdot\vec{y}}{\left\|\vec{x}\right\|\cdot\left\|\vec{y}\right\|} \tag{6}$$

This measurement is therefore independent of the norm of each vector. In principle it is more indicated for measuring densities, while the correlation coefficient provides information about the divergences on a like-for-like basis.

polygons do not naturally superimpose on statistical or administrative subdivisions, we will also investigate subdivisions by IRIS (first intra-community level), by community, by employment area and by *département*. Figure V represents the correlation and cosine similarity between the population estimate and the population derived from geo-referenced tax data, for each level of granularity.

We note that there are at least two reasons for finding differences between the results provided by the two sources. Firstly, the measuring concepts for the home are not the same (in one case, the information is derived directly from the tax residency declaration, in the other it is only obtained very indirectly from the subscriber's call behaviour). Secondly, one of the sources is complete while the other requires detrending with only a limited amount of additional information available to enable this detrending.

The results bring out significant divergences in the estimates obtained at very fine levels. The biggest divergences are observed in IRIS, the empirical correlation is 0.61. Using Voronoï polygons, the observations are closer compared to the IRIS grid, probably because the fact that it does not require resorting to a translation between grids, which removes one source of deviation.

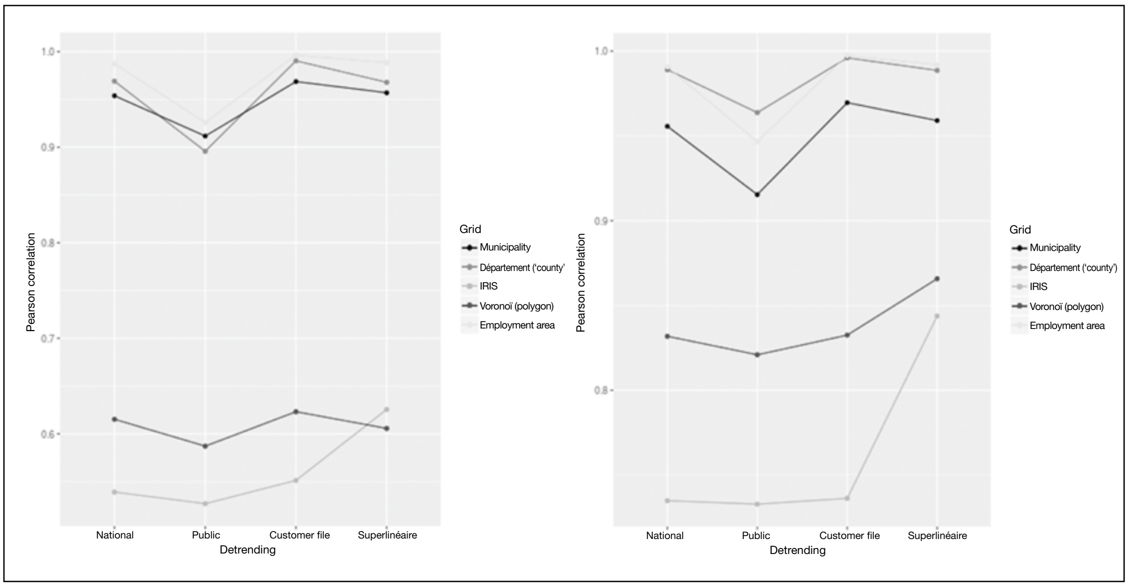
The difference is smallest at the most aggregated levels. In practice, it corresponds to the accuracy of the home-detection algorithm, which can vary over the different *départements* (particularly because the cell towers are not distributed uniformly across the country). This detrending is founded exactly on the data supplied by the tax source by *département*, and it is not surprising that the estimates obtained are very similar. However, it is surprising to observe that the "loss" of accuracy by municipality is low compared to by *département*.

We have also tested the quality of our estimates for a statistical zoning that might be more relevant to mobile phone data: the zoning by employment areas. An employment area is a geographical space within which most of the working age people live and work (Aliaga, 2015). This zoning is constructed iteratively, with the aim of maximising the number of working age people who live and work in an area. In 2010, France had 322 employment areas that formed a complete partitioning of the country in similar surface areas, that hold the middle between municipalities and *départements*. Compared to other zonings, employment areas offer the best correlations between mobile phone data estimates and residential populations derived from tax data and irrespective of the detrending method. One probable explanation is that employment areas are suitable for studying the local labour market meaning that we assume most working age people that live in an employment area will also place the majority of caal in that area, at least during working days. While there is inaccuracy in the precise location of an individual's home, there is therefore a high chance that the home detection algorithms will place the individual's home in the right employment area. Our results suggest that employment areas are an appropriate geographical scale for analysing population estimates made using mobile phone data.

Figures V-A and V-B also allow comparison of the differences obtained depending on the available additional information: simple ratio of the number of subscribers, use of "public data" (the operator's national market share and regional penetration levels), use of

Figure V
Empirical Correlation and Cosine Similarity between Resident Population Estimates and the Tax Source Based on the Detrending Method and Aggregation Grid

A – Empirical Correlation B – Cosine Similarity



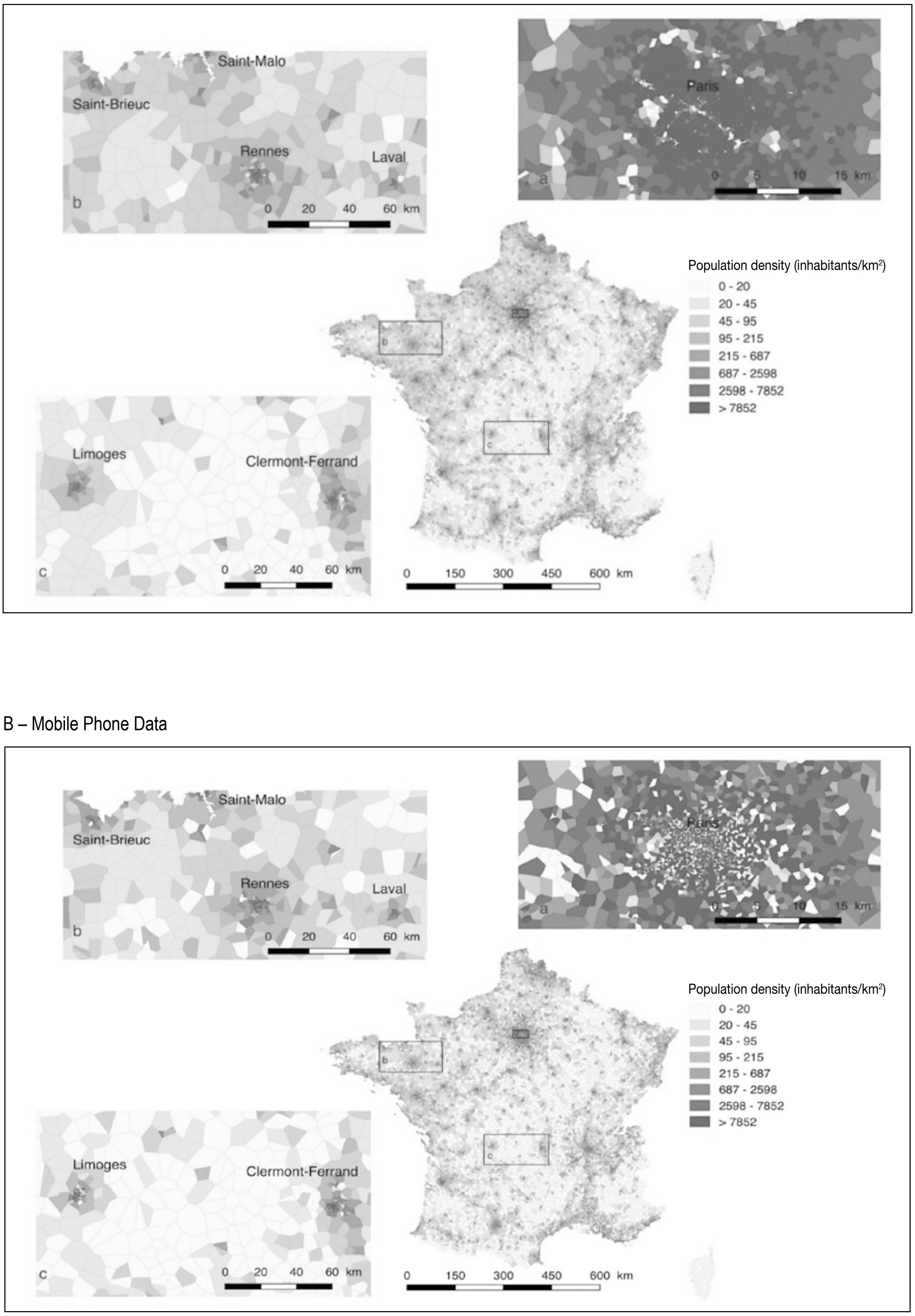
Reading note: For employment areas, by detrending the estimates using the customer file, we find 0.99 correlation between the population estimated from mobile data and the tax resident population.
Sources: CDR, customer file for "f.client" detrending, Arcep 2007 data for "public" adjustment and Filosofi 2011; authors' calculations.

the customer file to detrend the population observed at *département* level and by estimation from the superlinear model proposed by Deville *et al*. The best estimates are obtained from customer file information. However, using additional information such as penetration levels rather tends to degrade estimates by comparison with a simple rule of three on the volume of subscribers normalised to the French mainland population. Using regional penetration levels, which can mask non-uniform intra-municipal behaviour, seems to contribute more noise instead of improving the accuracy of the estimate. Furthermore, the superlinear model estimated at national level does not yield better results in terms of the empirical correlation or cosine similarity than detrending by *départemental* market shares. It is by taking into account information about the representativeness of the operator's customers at an intermediate geographical level (the *département*) that we obtain the best results, even without considering potential non-linear effects but with simple local detrending.

Figures VI and VII provide a cartographic representation – beyond nationally-aggregated indicators – to compare the differences between population densities estimated using tax and mobile data (with *départemental* detrending by market shares). In addition, comparing these two sets of maps illustrates how many estimates based on municipality are closer to the reference than estimates at the scale of Voronoï polygons. In the close-up areas, especially around Paris, it is clear that the change of grid and aggregation by municipality or *arrondissement* provides information closer to the available references.

The map in Figure VIII shows the relative differences between the predictions made by municipality and detrended by *département* using the customer file, with municipal populations obtained from the tax file. The areas where the difference is highest roughly match the parts of the country where the spatial interpolation procedure creates the best outcomes (as illustrated in Figure III). We therefore remain dependent on the grid represented by Voronoï cells to produce a municipal estimate. The inaccuracy is remains highest in places where the assumption of uniform population distribution in Voronoï polygons has less chance of being confirmed (such as in areas with unevenly distributed dwellings over the region of the municipality). Sometimes the differences between estimate and reference are very large.

Figure VI
Population Density by Voronoï Polygon Calculated Using Tax Data and Mobile Phone Data



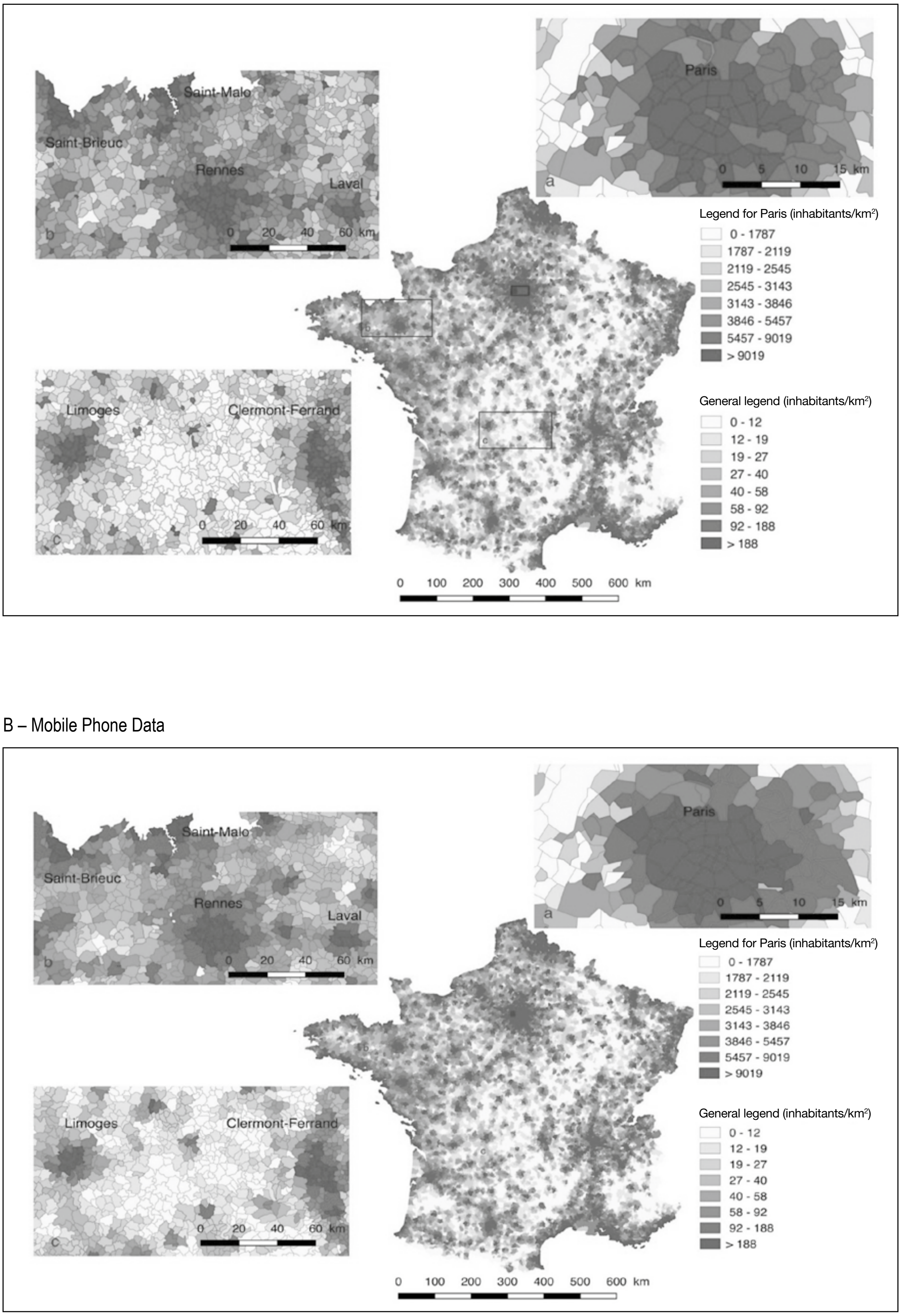
A – Tax Data

B – Mobile Phone Data

Note: The estimates are detrended by département using the customer file.
Sources: A, Filosofi; B, CDR, customer file and Filosofi; authors' calculations.

Figure VII
Population Density by Municipality Calculated Using Tax Data and Mobile Phone Data

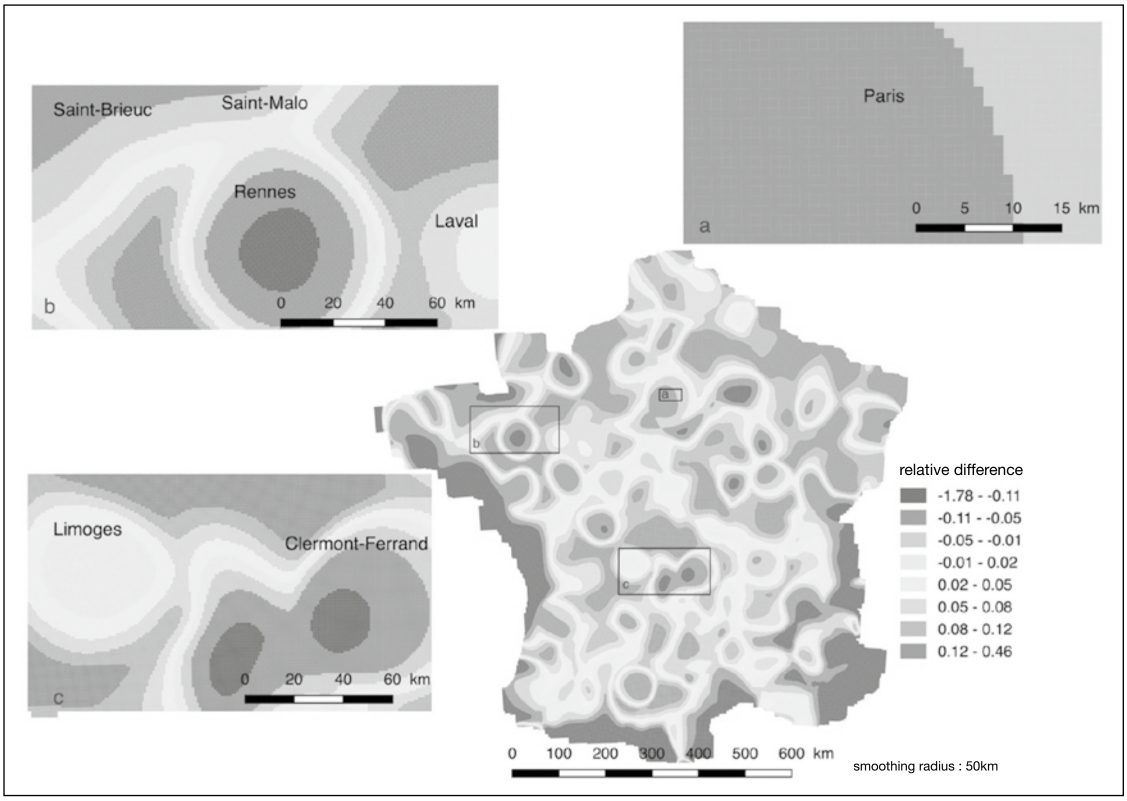
A – Tax Data



B – Mobile Phone Data

Note: The estimates are detrended by *département* using the customer file.
Sources: A, Filosofi; B, CDR, customer file and Filosofi ; authors' calculations.

Figure VIII
Map of the Relative Difference by Municipality between the Resident Population Estimate Detrended Using the Customer File and the Tax Source



Note: The differences are smoothed spatially for the representation.
Reading note: In the light grey areas the estimated populations are overestimated by a factor between 0.11 and 1.78; in the dark grey areas, it is underestimated by a factor between 0.12 and 0.46.
Sources: Orange 2007 Call Detail Records and customer file and Filosofi 2011; authors' calculations.

In some areas, the population of the municipality is underestimated by nearly half the municipal population, while in others it is overestimated by more than double (Figure VIII). These figures cover the estimates in section 2.3 on the interpolation cost in the tax source. This result is also confirmed by a more systematic analysis of the errors using statistical analysis (see Online complement C4).

Indicators such as the correlation coefficient or the cosine similarity do not take into account the spatial organisation of the points measured. However, it is plausible that the differences between the observed and predicted variables are spatially correlated, as illustrated by Figures III and VIII. For example, we may assume compensation phenomena between nearby municipalities, which are partly covered by the same cell towers and therefore by the same Voronoï polygons. Population estimates using Voronoï polygons will be distributed between these municipalities, which will create a correlation between the estimated values for these municipalities. Furthermore, as the error linked to using spatial interpolation is correlated to population density, it is likely that the differences will be similar for neighbouring municipalities. Spatial autocorrelation indicators such as Moran's *I* (Box 4) are an additional means of illustrating these phenomena.

We have calculated the value of Moran's *I* for four variables: the gross interpolation cost, the relative interpolation cost (compared to the number on inhabitants in the municipality), the gross difference and the relative difference. The four indices are significant, which confirms that these variables are not distributed randomly over the country and that there is indeed a spatial phenomenon involved.

Moran's spatial autocorrelation index for the gross interpolation cost is negative (and not significant). This is explained by the fact that when the subdivision into Voronoï polygons

Box 4 – **Moran's *I***

Spatial autocorrelation indices measure the spatial dependence between values of the same variable in different places in the space. The more the observation values are influenced by observation values that are geographically close to them, the greater the spatial correlation.

- Spatial autocorrelation is positive when similar values of the variable to be studied are grouped geographically.

- Spatial autocorrelation is negative when the dissimilar values of the variable to be studied come together geographically: nearby locations are more different than remote locations.

- In the absence of spatial autocorrelation, it can be considered that the spatial allocation of the observations is random.

The Moran index compares the way neighbouring observations co-vary with the covariance of all observations. The concept of neighbourhood is introduced using weights w_{ij} that take a value of 1 of observations y_i and y_j are similar, and 0 if not. The null hypothesis is a lack of spatial autocorrelation.

$$I_w = \frac{n}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij} \left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right)}{\sum_i \left(y_i - \bar{y}\right)^2}$$

$I_w > 0$ if there is positive spatial autocorrelation

leads to overestimating the population of a municipality, the population of neighbouring municipalities is underestimated, since the total population is constant. However, when the interpolation cost is normalised for the number of inhabitants, this index becomes positive – and very small, although it is significant (Table 2). Dividing by the size of the estimated population actually smooths the differences since the overestimates areas have their weight reduced relative to the underestimated areas. The gross differences and relative differences are positively correlated in space, a sign that some areas significantly concentrate municipalities having differences larger or smaller than the mean.

Table 1
Spatial Autocorrelation of Differences and the Interpolation Cost

Variable	Value of Moran's *I*
Gross difference	0.14***
Relative difference	0.13***
Gross interpolation cost	-0.11
Relative interpolation cost	0.009***

Note: *, **, *** indicate the significance at limits of 10%, 5% and 1%.

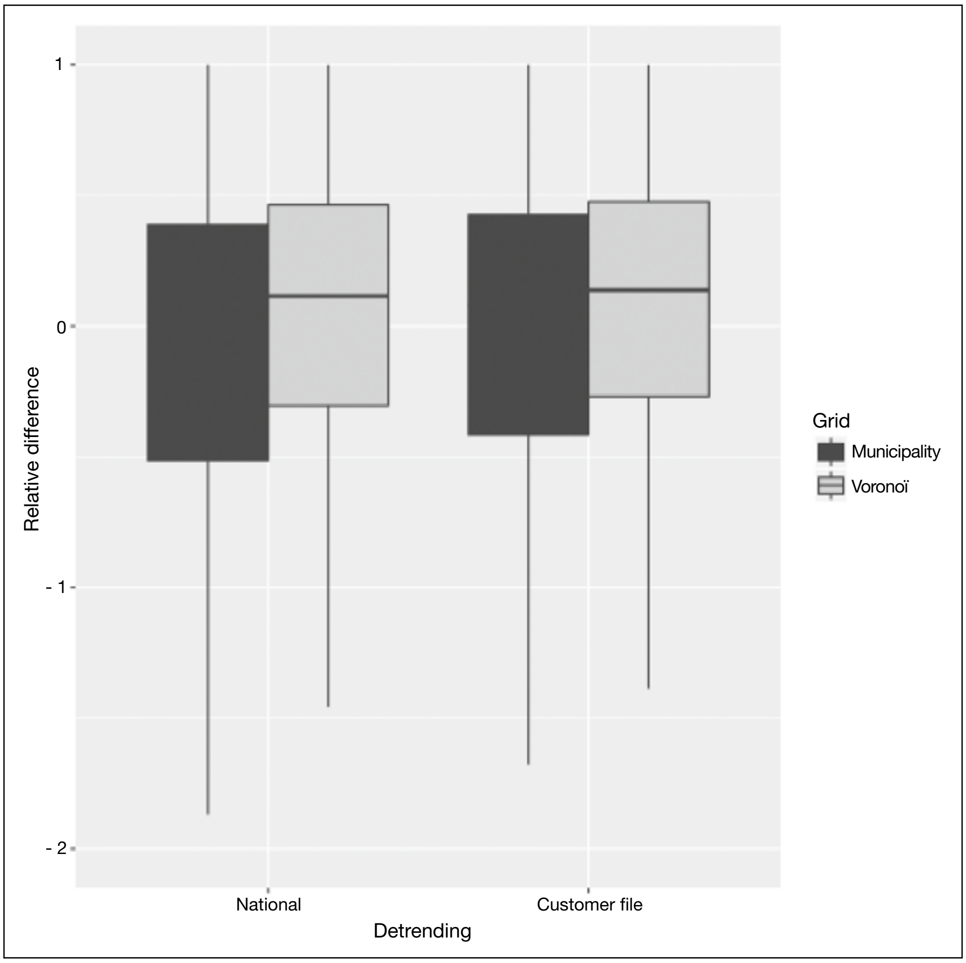
Finally, the distribution of municipal population differences, as represented in Figure IX, is narrower when detrending using the customer file by *département*. However, the median of this relative difference remains small, at the level of both the Voronoï cell and the municipality when detrending.

Using Temporal Granularity: Estimating Seasonal Variations

An important advantage of mobile phone data, other than spatial accuracy, is that it provides frequently captured data. In fact, in mobile phone data we have semi-continuous records about the presence of people, this is, when they use the network. This dimension was used indirectly in the previous estimates to identify subscribers' probable homes, but it was then used to estimate static values (the population). Using the dynamic aspects more directly can provide interesting information about the dynamic of the regions, such as variations in seasonal visitors. Such information could supplement the conventional indicators from official statistics, which only provide information about long-term changes in populations (supplied by censuses), with finer temporal information about tourist visitor numbers. Mobile phone records can be used to identify areas in which we observe large differences during the year, with great geographical precision. Looking at variations rather than at the absolute numbers of residential population estimates partly remedies the weaknesses highlighted by the previous analyses. In particular, knowing the local variability in market shares of the operator whose data we are using is less essential for investigating temporal trends then it is creating entire population estimates.

By way of illustration, we can focus on the summer months and for each month calculate the number of distinct subscribers identified in the mobile phone data for an area during

Figure IX
Distribution of Differences between Estimated Population and the Tax Source

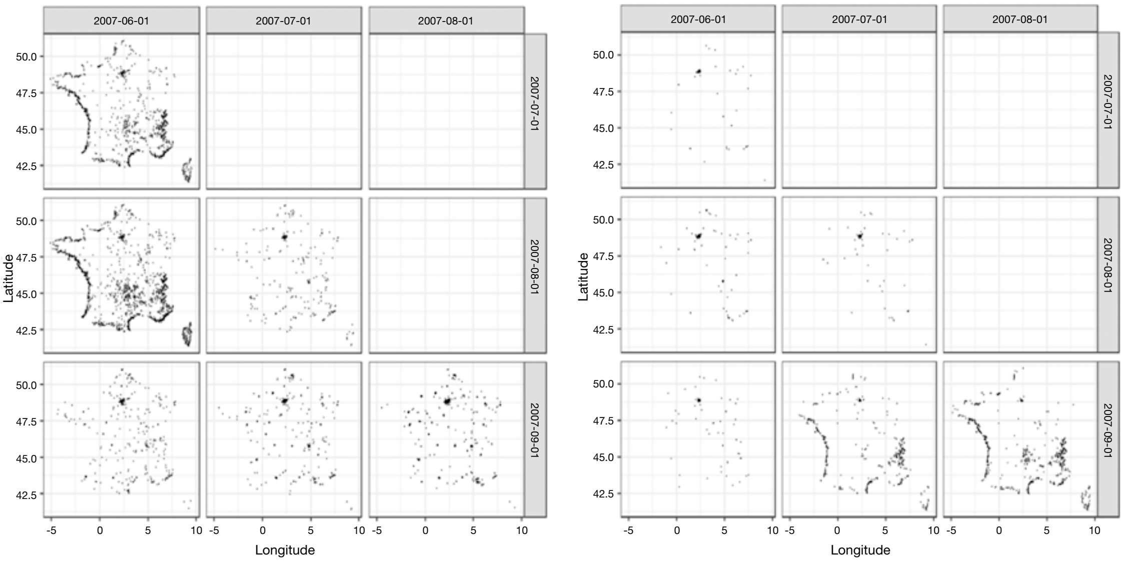


Note: For a clearer representation, outlying points are not shown. However, they represent a non-zero part of the population, for approximately 250 Voronoï cells where no tax resident is considered to live a total of nearly 60,000 users are estimated to live there.

a given month, normalised by the number of residents in previous months (here also using the most effective algorithm associated with the number of distinct days presence over a month). We work directly on the grid supplied using Voronoï polygons, to overcome difficulties linked to transposing the administrative subdivision described above. For each Voronoï cell we therefore have 6 variables corresponding to the ratios for July, August and September, normalised for the estimate of residents for June, July and August. Over all the Voronoï polygons, these variables are distributed to correspond approximately to a log-normal law centred around 1 – corresponding to a situation where the people present in a given month are identical to those identified as resident in the previous month. However, these differences can be very large, which results in a very thick tail to the distribution. To highlight the geographical distribution of these differences, Figure X shows the logarithm of these variables for the different months. To better bring out the large variations, we use different maps to show the areas where the changes are most marked. Figure X-A indicates locations where populations have increased by more than 50% between two months, and Figure X-B shows a map of the locations where they have decreased by more than 50%. The changes match the guess: in touristic areas (particularly coastal or mountain areas), we observe large population increases between June and July and between July and August, which disappear in September to return to a situation similar to that before the two holiday months.

In the rest of the country, changes are less pronounced. Still, we can also remark seasonal changes. For example, increases in populations outside the large urban centres can be observed during summer months and are reversed in September.

Figure X
Variation of the Population Present by Month



A – Increases in the Detected Population

B – Decreases in the Detected Population

Reading note: Between June and August, the population detected as inhabitants around the dark cell towers more than doubled, essentially on the coast and in the mountains (see part A). In Online complement C4, the light blue points show cell towers around which the population fell by less than half.
Sources: CDR; authors' calculations.

* *
*

These initial analyses suggest that it would be difficult using mobile phone sources to reproduce accurate statistics for residential population counts, as produced by official statistics. This result is not surprising in itself, given the differences between the two sources (declared tax residency versus residency reconstructed by the mobile phone analyses). We may also mention the limits inherent to the “active” nature of the data used, the locations are frequent on average but not always very regular. The signalling data, which supply information about the location at a systematic frequency, may make it possible to identify homes better, for example. Even by limiting oneself to the CDR data, widespread use of unlimited text messaging packages (still not widespread in 2007) has increased their use – and therefore also the possibilities for locating subscribers more regularly. Furthermore, the availability of para-data on the coverage of cell towers seems crucial insofar as a major part of the differences found seems to come from the approximation made by modelling coverage areas using a Voronoï tessellation.

This rapid change of mobile phone usage raises a major issue for the use of this type of data by official statistics. The statistical indicators that it produces are based on clear and shared concepts – a measurement convention on the value we want to measure. To use the indicators over time, in principle it is necessary for the data (and what they relate to) to be consistent over time. A constant change of content, and the methods needed to deal with them, could complicate interpretation of the results. It therefore seems premature to target the publication of standardised indicators using mobile phone data. Furthermore, using data from a single operator raises important questions about the possibility of accessing the information needed for detrending, in particular regarding local market shares, a necessary condition for detrending at a fine level. Finally, unequal coverage of the country raises difficulties in reproducing precise analyses on grids that have meaning.

Despite these limits, records taken from mobile phones supply a rich raw material for structural studies, as they illuminate regional phenomena, by giving information about the behaviour of individuals or other variables beneficial to regional development. Thus Pucci *et al.* (2015) present an illustration of using this type of data to describe the practices and uses of urban space (in which the grid of cell towers is sufficiently small to enable accurate analyses), and Aguilera *et al.* (2014) use them on performance measurements for urban transport networks (journey times, occupancy of trains, etc.). We can assume that these variables are less sensitive to the choice of operator and therefore that the detrending issues are raised less intensely. Galiana *et al.* (2018) were concerned with studying social and spatial segregation in urban units of Paris, Lyon and Marseille. By identifying a subscriber's probable home, and by characterising the district in which they live based on socioeconomic characteristics supplied by Insee, we can calculate social segregation indicators, quantifying the tendency of people only to communicate with people living in a similar district to their own in terms of income level, and to assess if this behaviour is more or less marked depending on whether or not they live in a better-off district. This study also proposes to measure segregation in space and its change, which corresponds to the fact of meeting, during the day or the week, people coming from various districts, or conversely the fact of remaining confined to a circle similar to their own. □

Link to Online Complements: https://www.insee.fr/en/statistiques/fichier/3706217?-sommaire=3706269/505-506_Sakarovitch-de-Bellefon-Givord-Vanhoof_complement.pdf

BIBLIOGRAPHY

Aguiléra, V., Allio, S., Benezech, V., Combes, F. & Milion, C. (2014). Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transportation Research Part C: Emerging Technologies*, 43(2), 198–211.
https://doi.org/10.1016/j.trc.2013.11.007

Ahas, R., Silm, S., Järv, O., Saluveer, E. & Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1), 3–27.
https://doi.org/10.1080/10630731003597306

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J.-L, Nurmi, O., Potier, F., Schmücker, D., Sonntag, U. & Tiru, M. (2014). *Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Consolidated Report.* Luxembourg: Publications Office of the European Union.
https://doi.org/10.2785/55051

Aliaga, C. (2015). Les zonages d'étude de l'Insee: une histoire des zonages supracommunaux définis à des fins statistiques. *Insee Méthodes*, 129.
https://www.insee.fr/fr/information/2571258

ARCEP (2008). Le Suivi des Indicateurs Mobiles – les chiffres au 31 décembre 2007.
https://archives.arcep.fr/index.php?id=9545&L=1

Blondel, V. D., Decuyper, A. & Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1), 1–55.
https://doi.org/10.1140/epjds/s13688-015-0046-0

Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S. & Ratti, C. (2015). Choosing the Right Home Location Definition Method for the given Dataset. In: Liu, T.-Y., Scollon C., Zhu W. (Eds.) *Social Informatics. 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings*, pp. 194–208. Springer International Publishing.
https://doi.org/10.1007/978-3-319-27433-1_14

Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., De Meersman, F., Seynaeve, G., Wirthmann, A., Demunter, C., Reis, F. & Reuter, H. I. (2016). Big data et statistiques : un recensement tous les quarts d'heure…, *Carrefour de l'Economie*, 2016/10.
https://economie.fgov.be/fr/file/801/download?token=Juj2pHbV

Debusschere, M., Sonck, J. & Skaliotis, M. (2016). Official statistics and mobile network operator partner up in Belgium, *The OECD Statistics Newsletter* N° 65, 11–14.
https://issuu.com/oecd-stat-newsletter/docs/oecd-statistics-newsletter-11-2016?e=19272659/40981228

Demissie, M. G., Phithakkitnukoon, S., Sukhvibul, T., Antunes, F., Gomes, R. & Bento, C. (2016). Inferring Passenger Travel Demand to Improve Urban Mobility in Developing Countries Using Cell Phone Data: A Case Study of Senegal. *IEEE Transactions on Intelligent Transportation Systems*, 17(9), 2466–2478.
https://doi.org/10.1109/TITS.2016.2521830

Deville, P., Linard, C., Martine, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D. & Tatem, A. J. (2014). Dynamic population mapping using mobile phone data, 111(45), 15888–15893. https://doi.org/10.1073/pnas.1408439111

DGINS (2013). Scheveningen Memorandum on Big Data and Official Statistics. https://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13

Desrosières, A. (2008). *Pour une sociologie historique de la quantification : L'Argument statistique I.* Paris : Presses des Mines. https://doi.org/10.4000/books.pressesmines.901

Galiana, L., Sakarovitch, B. & Smoreda, Z. (2018). *Ségrégation urbaine un éclairage par les données de téléphonie mobile*. Journées de méthodologie statistique de l'Insee, 12-14 juin 2018. http://jms-insee.fr/wp-content/uploads/S25_2_ACTEv2_GALIANA_JMS2018.pdf

Grauwin, S., Szell, M., Sobolevsky, S., Hövel, P., Simini, F., Vanhoof, M., Smoreda, Z., Barabási, A.-L. & Ratti, C. (2017). Identifying and modeling the structural discontinuities of human interactions. *Scientific Reports*, 7. https://doi.org/10.1038/srep46677

Grégoir, S., & Dupont, F. (2016). La réutilisation par le système statistique public des informations des entreprises. *Rapport du groupe de travail Insee-Cnis*. https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2016_143_reutilisation_syst_stat_information_ets.pdf

Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Martonosi, M., Rowland, J. & Varshavsky, A. (2011). Identifying Important Places in People's Lives from Cellular Network Data. In: Lyons, K., Hightower, J. & Huang, E. M. (Eds.), *Pervasive Computing*, vol. 6696, pp. 133–151. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21726-5_9

Janzen, M., Vanhoof, M., Smoreda, Z. & Axhausen, K. W. (2018). Closer to the Total? Long-Distance Travel of French Mobile Phone Users. *Travel Behaviour and Society*, 11, 31–42. https://doi.org/10.1016/j.tbs.2017.12.001

Léonard, I., Sillard, P., Varlet, G. & Zoyem, J.-P. (2017). Données de caisses et ajustements qualité. Insee, *Document de travail* N° F1704. https://www.insee.fr/fr/statistiques/fichier/2912650/F1704.pdf

Montjoye, Y. A. (de), Hidalgo, C.A., Verleysen, M. & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Science Report*, 3. https://doi.org/10.1038/srep01376

Pucci, P., Manfredini, F. & Tagliolato, P. (2015). Mobile Phone Data to Describe Urban Practices: An Overview in the Literature. In: *Mapping Urban Practices Through Mobile Phone Data,* pp. 13–35. Springer, Cham. https://doi.org/10.1007/978-3-319-14833-5_2

Ricciato, F., Widhalm, P., Craglia, M. & Pantisano, F. (2015). *Estimating Population Density Distribution from Network-based Mobile Phone Data*. Luxembourg: Publications Office. https://doi.org/10.2788/863905

Ricciato, F., Widhalm, P., Pantisano, F. & Craglia, F. (2017). Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing*, 35, pp. 65–82. https://doi.org/10.1016/j.pmcj.2016.04.009

Scholtus, S. (2015). Aantekeningen over het toewijzingsalgoritme voor Daytime Population. Statistics Netherlands, *Internal CBS note.*

Song, C., Qu, Z., Blumm, N. & Barabasi, A.-L., (2010). Limits of Predictability in Human Mobility. *Science* 327(5968), 1018–1021. https://doi.org/10.1126/science.1177170

Tennekes, M. (2015). Uitvoering toewijzings algoritme. Statistics Netherlands, *Internal CBS note.*

Tennekes, M. (2019). *R package for mobile location algorithms and tools: MobilePhoneESSnetBigData/mobloc*. R, Mobile Phone ESSnet Big Data. https://github.com/MobilePhoneESSnetBigData/mobloc (Original work published 2018)

Terrier, C. (2009). Distinguer la population présente de la population résidente. Insee, *Courrier des Statistiques* N° 128, 63–70. https://www.epsilon.insee.fr/jspui/bitstream/1/8564/1/cs128k.pdf

Toole, J. L., Ulm, M., González, M. C. & Bauer, D. (2012). *Inferring land use from mobile phone activity*. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12* (p. 1). Beijing, China: ACM Press. https://doi.org/10.1145/2346496.2346498

Vanhoof, M., Combes, S., & de Bellefon, M.-P. (2017). Mining mobile phone data to detect urban areas. In: *Proceedings of the Conference of the Italian Statistical Society*. Florence, Italy: Firenze University Press. https://eprint.ncl.ac.uk/file_store/production/241585/32829DBE-235C-4902-A175-0A8A0BD-CAFD4.pdf

Vanhoof, M., Plotz, T. & Smoreda, Z. (2017). Geographical veracity of indicators derived from mobile phone data. In: *Netmob 2017 Book of abstracts*. https://arxiv.org/abs/1809.09912

Vanhoof, M., Reis, F., Ploetz, T., & Smoreda, Z. (2018). Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics. *Journal of Official Statistics*, 34(4), 935–960. https://doi.org/10.2478/jos-2018-0046

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., & Buckee, C. O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of The Royal Society Interface*, 10(81), 20120986–20120986. https://doi.org/10.1098/rsif.2012.0986

Big Data and Audience Measurement: A Marriage of Convenience?

Lorie Dudoignon*, Fabienne Le Sager* et Aurélie Vanheuverzwyn*

Abstract – Digital convergence has gradually altered both the data and media worlds. The lines that separated media have become blurred, a phenomenon that is being amplified daily by the spread of new devices and new usages. At the same time, digital convergence has highlighted the power of big data, which is defined in terms of two connected parameters: volume and the frequency of acquisition. Big Data can be as voluminous as exhaustive and its acquisition can be as frequent as to occur in real time. Even though Big Data may be seen as risking a return to the paradigm of census that prevailed until the end of the 19th century – whereas the 20th century belonged to sampling and surveys. Médiamétrie has chosen to consider this digital revolution as a tremendous opportunity for progression in its audience measurement systems.

JEL Classification : C18, C32, C33, C55, C80
Keywords: hybrid methods, Big Data, sample surveys, hidden Markov model

Reminder:

The opinions and analyses in this article are those of the author(s) and do not necessarily reflect their institution's or Insee's views.

* *Médiamétrie (ldudoignon@mediametrie.fr; flesager@mediametrie.fr; avanheuverzwyn@mediametrie.fr)*

Receveid on 10 July 2017, accepted on 10 February 2019
Translation by the authors from their original version: "Big Data et mesure d'audience : un mariage de raison ?"

To cite this article: Dudoignon, L., Le Sager, F. & Vanheuverzwyn, A. (2018). Big Data and Audience Measurement: A Marriage of Convenience? *Economie et Statistique / Economics and Statistics,* 505-506, 133–146. https://doi.org/10.24187/ecostat.2018.505d.1969

During the 20th century, census has gradually declined in favor of sample surveys. The founding act can be considered to be Anders N. Kiaer's paper at the Congress of the International Statistical Institute in 1895 entitled *Observations et expériences concernant des dénombrements représentatifs*. In 1934, Jerzy Neyman published the reference article in sampling theory *"On the two different aspects of representative methods, the method of stratified sampling and the method of purposive selection"*. The growth of telephone equipment then encouraged the use of sample surveys in many fields (public statistics, politics, health, marketing, audience measurement, etc.). The end of the 20th century saw a new paradigm shift with the emergence of Big Data: a return to the census. As a major player in this digital revolution, the media sector has seen its measurement systems multiply and sometimes, inevitably, contradict itself. Médiamétrie, a benchmark institute for media audience measurement in France, has had to change its methods to take advantage of the best of each source.

The first part of the article deals with the relative advantages and limits of survey data and Big Data, with an emphasis on the notion of quality in its various dimensions. This will allow to explain why Médiamétrie has chosen to see survey data and Big Data as complementary rather than in competition with one another. Indeed, we will look at how hybrid approaches: "the mix of two data sources that differ in both nature and level to create a third, richer or more detailed one" have become the natural approach (Médiamétrie, 2010). The second part will illustrate these approaches through two operational implementations in media audience measurement. We shall begin by introducing the hybrid method used to measure internet audiences as part of the French market standard since 2012 – an example of a so-called panel-up approach (Dudoignon *et al.*, 2012). We shall finish by illustrating the so-called log-up approach used to measure the audience of special interest channels (Dudoignon *et al.*, 2014). In both cases, for Big Data to have any meaning or value, we must first understand how it is acquired, which often includes technical aspects performed to "clean" the data and process it in such a way as to create a potentially happy marriage with survey data.

Preamble: Data available in Audience Measurement

Both survey data and Big Data exist for television and especially internet media. In both cases, audience measurement is based on a panel and a semi-automatic system of measurement. In this introduction, our aim is to briefly describe the existing audience measurement systems for television and internet applied by *Médiamétrie* in France.

Internet

Internet audience measurement relies on two types of system: user-centric measurement is dedicated to tracking internet site and app audience behaviour for individuals across all of their devices. These systems are based on panels of individuals whose connections are measured using meter software installed on their computers, mobile phones or tablets that feeds data back to Médiamétrie's servers. The second type of system is called site-centric. This kind of measurement relies on the insertion of tags (Box 1) into the websites and apps of clients subscribing to the measurement and produces a total counting of the number of visits, page views and connection times.

Internet Audience Measurement on Computers

Since the home computer is often a shared device, the panel consists of a cluster sample of all individuals aged 2 years and over within a household. Therefore, the primary unit in the panel is the household and the secondary unit is the individual. Primary sampling units are recruited in accordance with the empirical quota method. Once the meter has been installed on all household computers, a pop-up screen or window will appear each time there is a connection. The secondary sampling units (individual household members) then have to identify themselves by ticking the box that corresponds to them. In September 2018, the panel comprised approximately 6,200 households with internet access *via* a computer, i.e. more than 14,000 individuals.

The scope of measurement is not limited to connections to the internet at home. In fact, for the population in employment, a significant proportion of their connections to the internet occurs in the workplace. Nevertheless, the effort required by individuals to take part in

Box 1 – **Description of Measurement Technologies**

What is a Tag?

In web analytics, a tag is an element that is inserted into each content to be measured, so as to count the number of content views. The content can be a page, an app, a podcast or even audio or video content. A code is inserted into the source code of the content. This generates a log on the third-party measurement system server each and every time a content is viewed. This then makes a total counting of connections to the tagged content possible.

What is Audio Watermarking?

A technology used for television audience measurement, audio watermarking consists of the insertion into the broadcast being measured of a mark (similar to a tattoo) that is inaudible to the human ear. This digital tattoo is inserted by a professional embedder validated by Médiamétrie. The principle is to modify the signal broadcasting the program with some additional information, without affecting sequence audibility. At the other end, the watermark is read by the TV meters connected to the TV sets owned by panelists. The mark inserted by the embedder contains identification information for the channel broadcasting the program, as well as regular markers of the broadcast time. In this way, we can differentiate between the audience watching a live broadcast, the audience watching a pre-recording and the audience watching via a catch-up TV platform.

measurement – also known as the "response burden" – prevents us from insisting that all secondary sampling units on the panel are additionally measured at their place of work (if they have a computer with internet access at work). Such insistence would likely lead to very low response rates. Therefore, the system is supplemented by an independent panel of individuals who have internet access on a work computer. In September 2018, there were 2,000 individuals on this panel, and it is linked to the preceding panel by statistical matching (Fisher, 2004).

Internet Audience Measurement on Tablets

The principle of internet audience measurement on tablets is very similar to that for computer measurement. Given that tablets are still hardly used in businesses, the scope for measurement is for the moment restricted to households. The panel consists of a cluster sample of individuals from within the recruited households. The latter must install a measurement app on all tablets used in their home and must change the settings to ensure their connections are sent to Médiamétrie's servers. As soon as the app is launched, the user can be identified. In September 2018, the panel consisted of 2,000 households, or 5,200 individuals aged 2 and older.

Internet Audience Measurement on Mobile Phones

Unlike computers and tablets, mobile phones are devices that are primarily for personal use. Consequently, the panel is made up of individuals recruited by quota sampling. The minimum age for measurement participants is set at 11 years old, and in accordance with the constraints imposed by France's Data Protection Act of 6th January 1978, participation by minors is subject to the consent of an adult with parental authority. Like the system for measuring connections *via* tablets, the panelist must install an app on its mobile phone. This app routes the connections to Médiamétrie's servers. All internet traffic on the phone is attributed to the main user of the phone. Any use of the mobile phone by a secondary user is therefore, by convention, assigned to the main user. In September 2018, the panel consisted of 11,000 individuals aged 11 and older.

Measurement of Secure Connections

Participation in user-centric measurement systems begins with the signature of an agreement between Médiamétrie and its panelists. This agreement details the respective commitments of Médiamétrie and the panelists. In particular, Médiamétrie undertakes to collect panelist user data for purely statistical purposes. Furthermore, Médiamétrie undertakes to never disclose the identity of its panelists to any third party for advertising or commercial purposes. Finally, it undertakes to take all necessary precautions to preserve the security of the data collected and, in particular, to prevent any distortion, corruption or unauthorized third-party access to this data. In return, the panelists undertake to keep their participation in the survey and the means of their participation confidential, in order to avoid any attempt at influence by stakeholders, publishers or operators with an interest in audience measurement. They also undertake to install the measurement software, to log on where appropriate, to inform Médiamétrie of any change in their situation, and to agree to be contacted by Médiamétrie.

Once the agreement has been signed, the panelists authorise Médiamétrie to have full access to their internet usage data, including their HTTPS connections and their IP address. However, for technical reasons, data collected from secure connections is in some cases less detailed than data gathered from HTTP connections. For example, for the measurement of connections *via* tablets, only the domain name will be available in the logs returned to Médiamétrie's servers in the case of an HTTPS connection, whereas the full URL will be collected for an HTTP connection.

Television

Médiamétrie's Médiamat panel is the reference in television audience measurement in metropolitan France. This measurement is based on a panel of individuals consisting of a cluster of some 5,000 households that own at least one television set. All active television sets are included in the measurement scope, i.e. those that are used at least once a month to watch television. Each of these TVs is connected to a TV meter that uses audio watermarking technology (cf. Box 1) to detect the channel being watched on the TV at any time. Individuals in the household must participate in the measurement by stating that they are in front of the TV using a remote control connected to the meter. Médiamétrie's servers continuously collect the data recorded by the TV meters. Although the panelists are instructed to state the presence of all household members in front of the TV screen, only the audience results for individuals aged 4 and older are fed back.

The TV return path (Box 2) is technically possible in two scenarios: ADSL, cable and satellite set-top boxes when they are connected to the internet, and smart TVs. We should note that although most television sets on the market today are smart TVs, in reality it is still quite rare for them to be connected to the internet. In these two scenarios only, return path data are available from the operator distributing the broadcast and they indicate which channel or service the set-top box is turned onto. No measurement is taken for any usage of the television without the set-top box. For example, if the television is connected to several modes of reception – *via* DTTV and an ADSL set-top box – any programs watched *via* DTTV will not be measured.

Quality of Survey Data and Big Data

Although there is no single definition of survey data quality (Dussaix, 2008), this is even more true of data quality in general. We can, however, keep in mind that quality is a real concern for most statistical agencies and that most of these would agree that it is a multidimensional concept that is difficult to assess (Lyberg, 2012). For our discussion, we have chosen to retain the six dimensions of quality used by Statistics Canada and the Australian Bureau of Statistics. They are: relevance, accuracy, time-to-market, accessibility, interpretability and consistency (Brackstone, 1999; Institut de Statistique du Québec, 2006). We should also note that the OECD adds two additional dimensions: credibility and cost-effectiveness in their assessment of the quality of statistical output (OECD, 2011). It is not a question here of discussing the definition of the dimensions of the quality of the surveys but of proposing a comparative analysis of "survey data" vs "Big Data" on each of these dimensions.

Relevance

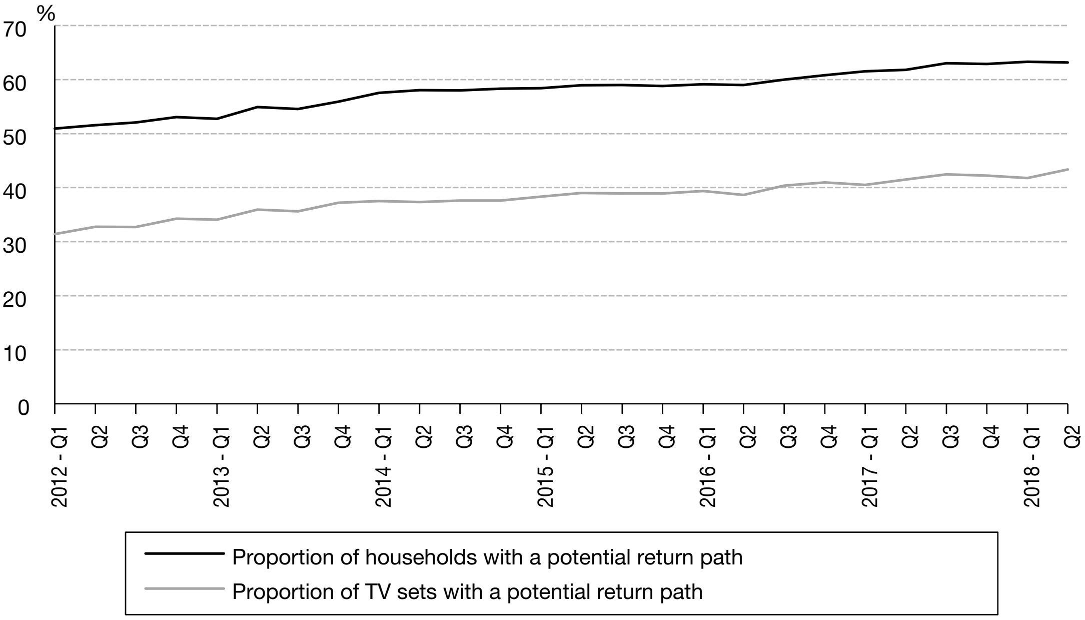
The relevance of a study or a measurement corresponds to its utility and its ability to meet the needs of users or customers. This criterion is obviously the first choice when assessing quality. The relevance of panel audience measurement is not generally called into question insofar as these systems have been designed in close collaboration with their users. In fact, for each media, a committee composed of members representing broadcasters and users, advertisers and media agencies, publishers and operators, and Médiamétrie, has been created on a parity basis. The objective of each committee is to define, orientate and validate measurements and surveys that serve as a reference for each of the media types concerned.

However, panel audience measurement cannot fully meet every need, in particular, when it comes to measuring very confidential or very fragmented usages, given that these would necessarily be poorly represented – or even not represented at all – within a sample. Increasing the sample size is clearly not a pertinent answer because the relevance of a study includes the budgetary constraints of its users. Conversely, Box 2 – **What is the Potential of Return Path Data in Television?**

A TV return path is the possibility for a broadcaster to collect some digital information back from users, about their TV consumption. The return path is technically possible for all set-top boxes that are connected to the internet, and for smart TVs. In concrete terms, this type of data collection is implemented by telecoms operators and satellite operators such as CanalSat (one of the major French suppliers of cable programs).

It is estimated that this return path is currently possible for a little over 60% of French households with at least one television set, but for barely more than 40% of television sets. In fact, the set-top box is very often connected to just one main television set and is not linked up to additional sets. This represents a potential, since not all set-top boxes that can be connected to the internet are necessarily connected.

Figure A
Evolution of the Potential of Return Path in Television



Coverage: Metropolitan France.
Sources: Médiamétrie – Home Devices.

Big Data does not fully meet the needs of users since it can identify machine usage but not individual usage. It is therefore essential to pre-process this kind of data to clean it up and transform it into meaningful information. Below are some real examples of this kind of pre-processing. They provide some valuable information on emerging and niche usages that cannot be measured by samples because of their volume. On the first criterion, relevance, the complementarity between survey data and Big Data for the purposes of audience measurement is clear to see.

Accuracy

In our context, accuracy means correctly describing the media behaviour of French people. Although it is generally acknowledged that results from surveys are flawed because of sampling errors and the problem of non-response there is a tendency to think on the contrary that Big Data is accurate because it covers the entire scope of measurement. It is nothing of the sort. Actually, as we noted above, Big Data brings in information about machines and not about individuals, which is an obvious source of error. Furthermore, if the technologies used to measure are not properly controlled, they can lead to implementation or interpretation errors. This brings us back to the pre-processing phase that should partially clean up these interpretation errors. As far as implementation errors are concerned (for example, wrong implementation of an internet tag), the best way to proceed is to install a monitoring system to detect these flaws as early as possible and to correct them before too large a volume of data becomes affected. It should be noted that this type of monitoring is also necessary for panel measurement since it uses content marking technology (web tag or audio watermarking for television) for the purposes of audience measurement.

Time-to-Market (or Speed of Delivery)

Time-to-market refers to the lag between the analysis reference period and the delivery of results. In the context of media audience measurement, this is a very important criterion. Any excessive delay in delivering results would render these results obsolete and of very limited interest to users. For internet, results are generally made available monthly and must be published in the month following the analysis period. For television, the delays are much shorter. The first audience results for daily programs are published from 9am the following morning. These results are then consolidated eight days later with the inclusion of time-shifting viewing in the seven days following the original broadcast.

For survey data, site-centric data or return path data, when automatic measurement technologies are used, raw data can in theory be acquired almost in real time. The freshness of the results can therefore be ensured as soon as the pre-processing and processing operations of these data are performed in limited time. In both cases, this involves the implementation of very strict, automated and industrial production processes.

Accessibility

Audience measurement results are accessed *via* reporting interfaces that are available to all subscribers. This kind of interface in particular can manage various user permissions, and thus grant access to less or more information depending on their subscription. From the user point of view, accessibility will be considered as satisfactory if the results consultation tool is both ergonomic and efficient in terms of computational and display times. Internally, our teams tasked with producing results and performing additional analysis have ready access to all of the data. Nevertheless, even in-house, this access is limited to anonymous data. Only the management and panel coordination teams have access to personal information that can be used to contact panelists.

Technical difficulties of access to Big Data are increasingly rare these days and are no longer a priority issue for development. By contrast, legal constraints oblige us to limit access to this type of data and even to reduce the quantity of information gathered. Although in the past, digital data could sometimes be collected without the knowledge of the individuals, this kind of practice is no longer possible, in Europe at least. Most stakeholders who are currently gathering this kind of (site-centric or return path) data have had to put in a lot of effort to become compliant with the European General Data Protection Regulation (Box 3).

Intelligibility (or Interpretability)

Whether for panel data or Big Data, the intelligibility of the data mainly relates to technology. It is possible to think of the raw data generated by tagging technology for television and internet (what we call logs) as hardly intelligible at all. Only after pre-processing will this data be translated into an interpretable format. The statistician obviously cannot work alone. This type of data necessitates close collaboration between the technical teams who develop the tagging solutions, the I.T. teams who collect and process the data, the statistical teams who devise the analysis, and the customer liaison teams who install the tagging solution into their websites and channels.

Although they may appear complex, media content tagging solutions can, after translation, provide intelligible data that is also easy to enrich with metadata describing the content in detail (e.g. for online video content, the ability to specify if it was a series, which season, which episode, and when the original TV broadcast was, etc.). Automatic measurement solutions that do not use tags are generally much less intelligible. Take, for instance, internet audience measures based on the capture of network traffic for a device. Over 90% of the collected data is irrelevant, since it cannot describe the behaviour of the individual using the device. The data collected actually includes all of the technical information flows, e.g. updates to software and applications, which are totally transparent for the user.

Rendering this kind of data intelligible is a real challenge, since any mistake in filtering the data usually leads to an interpretation error. With tagging solutions, it is possible to only collect

Box 3 – **European General Data Protection Regulation: The Changes Affecting Professionals**

The new European regulation which came into force on 25th May 2018 introduces or strengthens the following principles.

• Strengthening of the rights of persons: Users must be informed of the collection and use of their data. At all times, they must be able to give their consent, or object if necessary. Users have new rights: in particular, the right to restriction of processing; the right to data portability; the right to erase data.

• Responsibility of agents (data controller and processor): The regulation reduces the obligations of prior formalities at the CNIL (the French authority for data protection). On the other hand, the new regulation introduces the principle of demonstrability: the ability to prove compliance with the regulation at all times through detailed documentation of all personal data processing activity. In concrete terms, the data controller undertakes to: keep up-to-date detailed registers of personal data processing activity; to systematically carry out impact assessments before each processing activity that presents a high risk to the rights and freedoms of natural persons; to ensure the compliance of any data processors. The regulation also strengthens the sanctions to be applied against the data controller in the event of a non-compliance: up to 20 million euro or 4% of global turnover.

• Privacy by Design: The company must take into account the notion of respect for private life, beginning at the design phase of a product or application. The data controller must implement all technical and organizational measures that are necessary to comply with the protection of personal data, from the design phase and by default.

• Creation of a Data Protection Officer role (DPO): This new expert will identify and coordinate the actions to be taken within the company or organization that pertain to the protection of personal data: from internal communications to checks on regulatory compliance, as well as being the point of contact with the supervisory authority.

data that is useful, which therefore makes these solutions a lot easier to interpret.

Consistency

Without the hybrid approaches, a stakeholder could end up with several figures representing the performance of an identical content. For example, the average number of viewers for the video content over a period, and the number of set-top boxes tuned into that video content for at least one minute. These two indicators are based on different units and are not comparable, but they may alarm unaware users who see both of them published. Médiamétrie should therefore provide the necessary consistency. Firstly, by clearly explaining the concepts and indicators, as well as how to interpret them. Next, by offering solutions to reconcile these different-natured data so as to produce a consistent measure. Panel data and Big Data consistency is then the very essence of Médiamétrie's hybrid measures.

In addition to the six dimensions described above, another one that must be considered regarding Big Data is confidence (or, to use the OECD term, credibility). Some media stakeholders have installed site-centric or return path systems of measurement. As is the case, for example, of the biggest players on the web – GAFA[1] and the telecoms operators. Such players use these to offer measurement services to publishers who also use their distribution platform. As it is generally very hard to be the judge of one's own case, even if one possesses the utmost discipline and honesty, other market players will always call their credibility into question. In such a context, "proprietary" Big Data often requires certification by a trusted third party to be recognized and shared by the market. This is the role of ACPM[2] in France which certifies the number of newspapers and magazines distributed.

Some Examples of Hybrid Approaches to Media Audience Measurement

Two approaches to hybrid measurement are theoretically possible. The approach chosen depends on the user's expressed need. In the first approach, which we call panel-up, Big Data enriches the information gathered in the media survey, which is usually a panel, as described in the preceding section. In this approach, Big Data will be considered as auxiliary information that is taken into account in order to improve the precision of the survey results. The second approach, which we call log-up, involves the enrichment of Big Data. We construct a model based on the survey data, thereby allowing us to

[1] *Google, Apple, Facebook and Amazon, the four American giants that dominate the digital market.*

[2] *Press and Media Statistics Alliance.*

estimate the consumer profile for this media. We will now illustrate each one of these approaches.

Hybrid Internet Audience Measurement on Computers

Coexistence of Two Complementary Measures

In the context of internet audience measurement on computers, two types of complementary measures have coexisted for a number of years now. As detailed in the first part of this article, user-centric measurement is provided by Médiamétrie//NetRatings. It is based on a panel of 16,000 individuals that can estimate the audience and usage of all websites in France. For their part, site-centric measurement tools can provide comprehensive results for website and app consumption in terms of page views, visits and duration. Subscribers to site-centric measurement systems can only access their own results and may not see their position compared to competitors. We call this proprietary measurement. They must then refer to the Médiamétrie// NetRatings panel to find their position.

Launch of a Hybrid Measure in October 2012

Médiamétrie wanted to release a hybrid measurement system onto the market that could take advantage of both measures while still respecting a number of constraints:

- All websites should be able to benefit from the accuracy gain delivered by site-centric measurement, not just those that have subscribed to that measurement;

- The site-centric data used should be consistent with the panel measurement scope;

- The resulting hybrid data should be compatible with media planning tools which require individual data to input into their calculation engine.

In consideration of the three aforementioned constraints, we decided to go for a panel-up approach. Site-centric results are seen as counts known for the total population. The fundamental theoretical principle is this: “whenever we possess auxiliary information, we must seek to use it” (Ardilly, 2006). The idea, therefore, is to use this information by introducing additional auxiliary variables when weighting the sample (Dudoignon *et al.*, 2012). Site-centric data for around 400 entities was then sent to Médiamétrie. By data, we mean all of the connection logs collected by the site-centric measurement tools.

Consistency between Site-Centric and Panel Data

Site-centric data is not inherently comparable with panel data for the same entity. In particular, they differ in two aspects: geographical coverage and the terminals measured. Indeed, site-centric measurement counts connections across all devices (computers, mobile phones, tablets, games consoles, etc.) and regardless of the country where the connections occur. In order to introduce site-centric results as weighted auxiliary variables in the panel calibration, the two scopes must be exactly comparable. Consequently, we developed a pre-processing step for site-centric data in order to ensure this consistency. Firstly, the site-centric data is filtered on the device being measured, in this case the computer. Connections from abroad are then dismissed. Other more technical filters are also applied which can notably exclude logs that contain connections performed by robots.

The final step consists of aggregating URLs consistently between the two measures. The objective of this last step is to ensure that these auxiliary variables are consistent between panel and population. The only way to ensure this consistency is to tag all of the URLs of the various entities.

Problems Encountered

The problems encountered were first and foremost related to the representativeness of the entities introduced in the panel calibration. Unfortunately, no site-centric results are available for all web content. Some stakeholders are opposed to subscribing to a site-centric system of measurement. Others have proprietary measurement systems that have not been certified by a trusted third party.

Moreover, it was hard to envisage how we could introduce these 400 entities as weighted auxiliary variables in the panel calibration. Therefore, we decided to make a carefully judged selection of entities. The first rule used was to only include entities for which the number of visitors in the panel was greater than 100, and to minimize the correlation between the entities we introduced. The final selection of entities had to respect the following constraints:

- Consistently cover all population targets in terms of gender, age and socio-professional category;

- Be varied in terms of content (news, travel, cars, etc.);

- Be of limited size in order to allow convergence of the calibration algorithm, without discriminating the calibrated weights distribution, as this would limit the gain in precision.

In the end, a little over 150 entities were chosen to be included in the basis for panel calibration. The introduction of these additional auxiliary variables in the weighting process directly impacts on the quality of the calibrated weights. The ratio between the maximum weight and the minimum weight is higher and we observe that calibrated weights accumulate towards the limits, which lead to a loss of accuracy and to greater instability of the results (Roy *et al.*, 2001).

Currently, the CALMAR macro program is used for the calibration (Sautory, 1993). Tests are conducted with new algorithms to summarize the auxiliary information – calibration to the principal components (Goga *et al.*, 2011) – or to relax the benchmark constraints on some auxiliary variables – ridge regression calibration (Alleaume *et al.*, 2013) –, these algorithms allowing either to improve the quality of the calibrated weights or to introduce a larger number of entities.

Extension of the Method to Global Internet Measurement

Since October 2017, the French market standard for internet audience measurement has been the Global Internet measurement, i.e. on the three screens (computers, mobile phones and tablets). The Global Internet measurement is based on the three panels described above, which have a common part. Indeed, some panelists belong to several panels and are measured on several types of devices. In September 2018, the number of panelists measured on several of their devices is about 6,000 individuals.

The three internet panels are combined by statistical matching to produce audience results on three screens, taking into account the duplication between devices. The site-centric measurement described in the previous section allows the identification of the device used by the user to connect, but without distinction between mobile phone and tablet. A hybrid method by calibration similar to that carried out on the computer internet audience measurement is performed on the sample resulting from a first statistical matching between the panels on mobile phones and tablets. A second statistical matching under constraint of weights conservation is then performed with the computer panel to create the hybrid measurement of the Global Internet.

Hybrid Measurement for Television

As indicated above, panel audience measurement does not always allow the most detailed measurement of very fragmented usages. This is true for Médiamat whose 5,000 households are insufficient to offer a daily service to thematic channels that are exclusively received *via* satellite (CanalSat), ADSL, fibre optic or cable.

In response to the need to assess the value of special interest channels, we chose the log-up approach because it can provide these channels with additional information at little cost, which is always an important consideration and especially so for this category of stakeholders whose marketing research budgets are limited. We are only dealing here with TV data for television channels (i.e. broadcast and not video on demand – VOD). Advertising distribution models are very different between broadcast and VOD or digital platforms, at least for the moment, in France.

To clearly understand the solution developed by Médiamétrie for the hybrid measurement of special interest channels, we must understand firstly, the differences between set-top box usage and individual viewing. To begin with, we notice deviations between set-top box usage and the usage of the television that the set-top box is linked up to. For example: the set-top box can send backlogs that do not correspond to human activity, such as automatic reboots. Furthermore, the set-top box may be switched on and the television switched off: this is very often the case overnight.

In addition, deviations were observed between TV usage and watching TV alone, since the TV remains primarily a family media and a significant part of viewing time is spent watching (the same television) together. Around 40% of the time that individuals aged 4 and over spent in front of the television involves multiple simultaneous viewers, and this figure peaked at 60% for certain weekend time slots (Source: Médiamétrie//Médiamat).

We therefore use a two-step method. The first step is to shift from set-top box to television set. We begin by pre-processing the raw logs, so as to clean up any technical log data and to establish the audience tickets. For each channel viewing, we obtain data of the type: start time, finish time, channel identifier. Next, we proceed to truncate the set-top box usage for those times when the television is probably off. To do this, we shorten the longest audience tickets. The parameters of the truncating function can be estimated from the observed audience tickets durations in the Médiamat panel for the same universe (Figure I).

The second step is to individualize the audience tickets obtained in the 1st step at television set level. This second step presents the most difficulties.

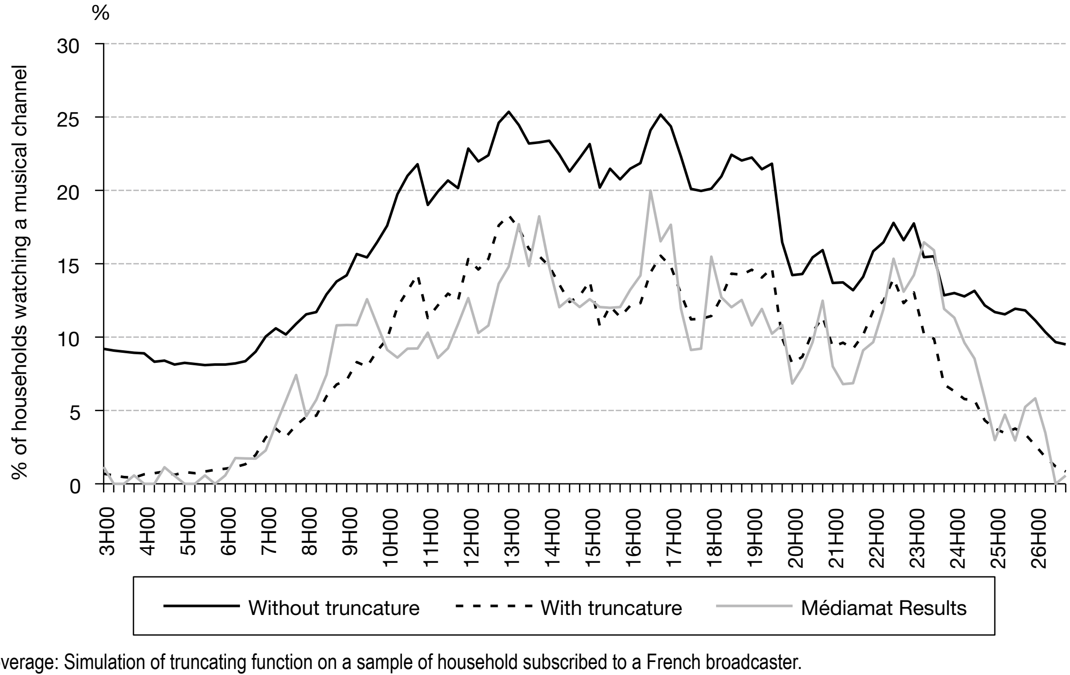
We decided on a modeling approach based on knowledge of the sociodemographic profile of the set-top boxes to be individualized (number of people in household, gender, age, SPG and relationship between individuals). Since the individuals in the household are known, we then only need to determine who is watching the TV when it is turned on. With this approach, we do not therefore use the comprehensiveness of return path data collected by the operators, but only the data from a sample of subscribers who agree to state the nature of their household and who authorize the operator and Médiamétrie to have access to the TV usage data on their set-top box. All of the data is made completely anonymous. Even though the comprehensiveness of the data is not used, the low cost of recruiting a panelist allows us to obtain a large sample size for minimal outlay. This then meets the needs of the thematic channels. The individualization of television set audience tickets without this additional information on household's composition would be hard to envisage.

The individualization of the audience is based on hidden Markov models that can be represented schematically as shown in the Diagram below (Rabiner, 1989; Rabiner *et al.*, 1993).

In our case, the time could be cut into 5-minute steps (but we can choose a longer or shorter time). We then have:

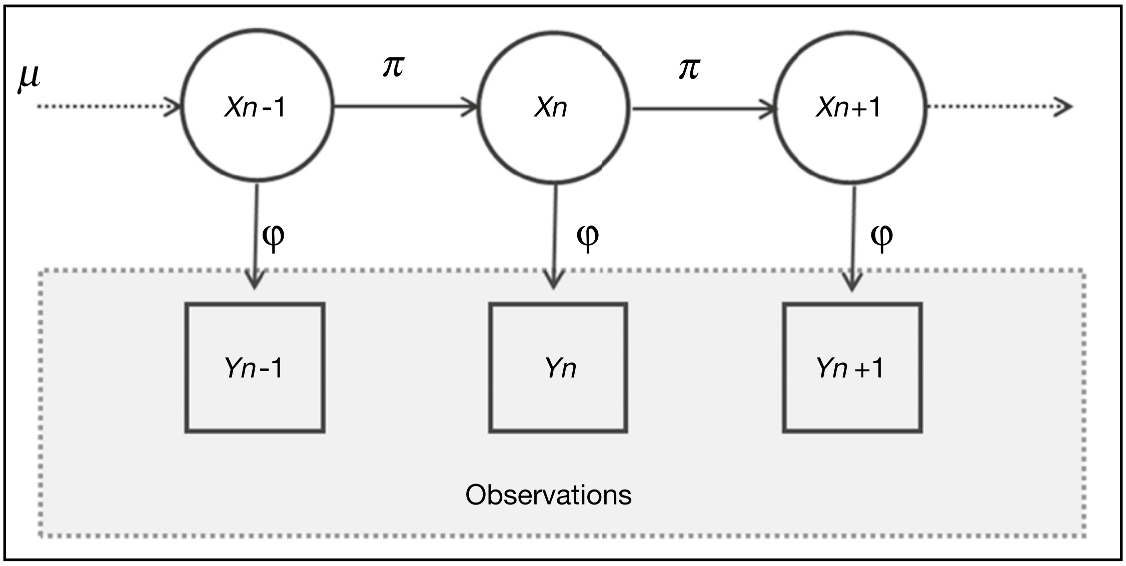
- Observations Y which correspond to the television channels watched, which we group by theme, e.g.: youth, sport, cinema, etc. Y_n is the major theme during the n^{th} time step;

Figure I
Effects of Truncating Function on a Musical Channel



Coverage: Simulation of truncating function on a sample of household subscribed to a French broadcaster.
Sources: Return path data of this broadcaster.

Diagram
Schematic Representation of a Hidden Markov Model



Note: The Markov chain {*Xn*} is not directly observed. Observations {*Yn*} are generated through a memoryless channel, which means that each *Yn* depends only on the state *Xn* at the same moment.

- A hidden phenomenon X, which stands for the individuals in front of the television. X_n describes all individuals in the household watching television at time n, which enables the correlations between individuals of the same household to be preserved, and therefore the overall levels of watching TV together.

We chose hidden Markov models because their characteristic properties perfectly describe the phenomenon to be modeled, namely:

- A short memory process: to know who is watching the TV at time $n + 1$, we only have to look at who was watching it at time n. We do not need to know the full history of who was in front of the TV;

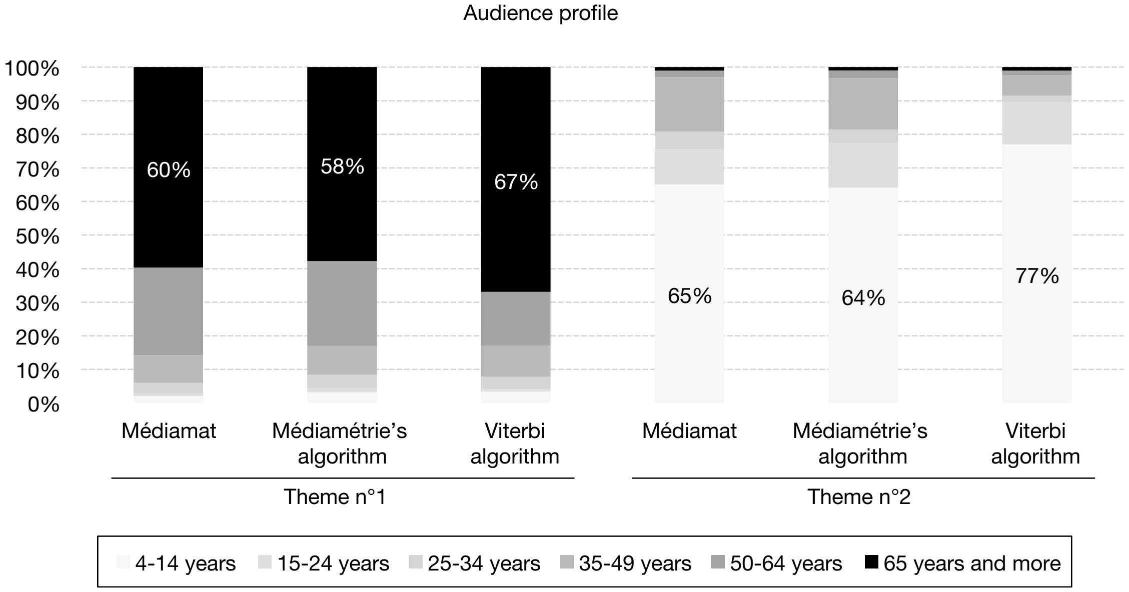
- Observations through a memoryless channel: the TV channel being watched at time n only depends on the individuals who are in front of the television at the same time.

The possible states for X depend on the size and composition of the household. For a single person household, modeling is pointless (the one individual in that household is watching the TV). For a two-person household, for example a couple, there are three possible states: the reference person alone, the partner alone, or the couple. For a three-person household, for example a couple with one child, there are seven possible states: the reference person alone, the partner alone, the child alone, the reference person with the child, the partner with the child, the couple or the couple and the child.

It can be easily demonstrated that for a household of size k, the number of possible states is $2^k – 1$. We have deployed a household typology that describes all household compositions to consider: one person in the household, two persons in the household (couple), two persons in the household (single parent and a child), three persons in the household (couple and child), three persons in the household (single parent and two children), three persons in the household (three adults), etc. For each household type, there is a corresponding sub-model characterized by a set of parameters $M = (\mu,\pi,\varphi)$ where μ is the initial state, π the transition matrix and φ the probabilities of observation. All parameters can be simply estimated using Médiamat panel data, which here serves as a sample for learning.

Once the model parameters are known, we only have to estimate how many people are in front of each television set. Most often, people want to estimate the most likely sequence $\{Xn\}$ by using the Viterbi algorithm (dynamic programming), which allows to do it without calculate all the possibilities. But considering the most likely solution leads to caricatured behaviour estimates (only children in front of youth channels, etc.) and does not reproduce behavioural diversity. We prefer then to use an algorithm with a random component.

Figure II
Comparison of Algorithms – Example of Results on Two Themes with Very Marked Profiles



Reading note: 60% of theme 1 audience is 65 years old and over, in Médiamat panel. With Viterbi algorithm, this is increasing to 67%, that means an over-estimation of older people. While, with Médiamétrie's algorithm, the result is closer to the panel reality with 58%.
Coverage: Audience profile on two themes.
Sources: Tests of individualization on Médiamat panel.

The Médiamat panel is also used as a test sample for the choice of algorithm. Using the panel data, the presences are estimated with the individualization algorithm, then we compare the obtained results with those from Médiamat. The comparisons are not made on a unitary basis (household by household) because the published results are averages and so this could lead to compensations. Instead, the main audience indicators by theme and by channel are compared and we choose the algorithm that minimises the deviations. Figure II gives an illustration of the comparisons that have been made to build the algorithm.

* *
*

The emergence of Big Data – the new Oil – and the development of capacities to store and process this data have raised the prospect of the end of audience measurement in favour of more accurate, more reliable and less expensive measurement systems (Vanheuverzwyn, 2016).

In the first part of this article, we demonstrated that issues surrounding quality were of equal concern for Big Data and survey data. The two examples shown of hybrid approaches clearly show that quality also lies in the processing and modeling that can be applied. Some perfectly good data could lead to incoherent or irrelevant results, especially if we lose sight of the users' needs.

Rather than marking the end, we are observing today an evolution, or even a revolution, in audience measurement towards hybrid measures. There is no question that we must leverage the advantages of different observation systems in order to create others that are more complex and richer. With this outlook, new application fields will open up in research and development. Starting with the theory and practice of surveys. In fact, the utilization of Big Data could be considered as a response to the increasing prevalence of non-response in surveys. The question of the trade-off between bias and variance, estimation bias and calibrated weights variance, has been raised and is worth pursuing. It could lead to the development of more effective calibration algorithms capable of taking many more weighted auxiliary variables into account. It could also result in the development of new hybrid methods based on statistical matching or imputation techniques. Research in machine learning also offers interesting prospects for enriching Big Data and it cannot be ignored in the context of audience measurement.

However, the responses that we put forward to address the needs of observing individual behaviour must be, as they have always been, part of a framework that respects privacy and the legal restrictions associated with the processing of personal data. This is not so much a legal question as an ethical one (Tassi, 2014). The entry into force of the European General Regulations on Data Protection and the public debates that took place upstream, made it possible to highlight the drifts in the measurement of internet usages. Surveys, for which the consent of the individual is inherent, therefore regain a central role. □

BIBLIOGRAPHY

Alleaume, F. & Dudoignon, L. (2013). Calage sur information auxiliaire incertaine : proposition d'algorithme de redressement ridge. *Actes des 45e Journées de Statistique de la SFdS,* Toulouse, 2013.
http://papersjds13.sfds.asso.fr/submission_189.pdf

Ardilly, P. (2006). *Les Techniques de sondage.* Paris : Éditions Technip.
http://www.editionstechnip.com/en/catalogue-detail/113/techniques-de-sondage-les.html

Brackstone, G. (1999). La gestion de la qualité des données dans un bureau de statistique. *Techniques d'enquête,* 25(2), 159–171.
https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1999002/article/4877-fra.pdf?st=FSaA6d3F

Brackstone, G. (2006). Le rôle des méthodologistes dans la gestion de la qualité des données. In : Lavallée, P. & Rivest, L.-P.*, Méthodes d'enquêtes et sondages.* Paris : Dunod.
https://www.dunod.com/sciences-techniques/methodes-d-enquetes-et-sondages-pratiques-europeenne-et-nord-americaine

Deville, J.-C. & Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418), 376–382.
https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1992.10475217#.XGbmljNKiiM

Dudoignon, L. & Logeart, J. (2014). Mesure hybride de l'audience TV. *Actes des 46e Journées de Statistique de la SFdS*, Rennes, 2014.
http://papersjds14.sfds.asso.fr/submission_128.pdf

Dudoignon, L. & Zydorczak, L. (2012). Enquête et données exhaustives : un nouveau défi pour les mesures d'audience. *Actes en ligne du 7e Colloque Francophone sur les Sondages*, Rennes, 2012.
http://sondages2012.ensai.fr/wp-content/uploads/2011/01/Dudoignon-Zydorczak-Mesures-Hybrides-Médiamétrie-2012-Article.pdf

Dussaix, A-M. (2008). La qualité dans les enquêtes. *MODULAD,* 39, 137–171.
https://www.rocq.inria.fr/axis/modulad/archives/numero-39/Tutoriel-Dussaix/Dussaix-39.pdf

EUROSTAT (2007). *Handbook on Data Quality Assessment Methods and Tools.*
https://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK ON DATA QUALITY ASSESSMENT METHODS AND TOOLS I.pdf

Fischer, N. (2004). Fusion statistique de fichiers de données. *Thèse de doctorat.* Paris : Conservatoire National des Arts et Métiers.
https://cedric.cnam.fr/fichiers/RC899.pdf

Goga, C., Shehzad, M.-A. & Vanheuverzwyn, A. (2011). Régression en composantes principales versus ridge régression en sondages. Application aux données Médiamétrie. *Actes des 43e Journées de Statistique de la SFdS*, Tunis, 2011.
https://www.researchgate.net/publication/292133976_Regression_en_composantes_principales_versus_ridge_regression_en_sondages_Application_aux_donnees_Mediametrie

Institut de la Statistique du Québec (2006). *Le cadre intégré de la gestion de la qualité de l'Institut de la statistique du Québec.*
http://www.stat.gouv.qc.ca/institut/CadreGestion_qual.pdf

Kiaer, A. N. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9(2).
https://gallica.bnf.fr/ark:/12148/bpt6k61560p?rk=42918;4

Lyberg, L. (2012). La qualité des enquêtes. *Techniques d'enquête,* 38(2), 115–142.
https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012002/article/11751-fra.pdf?st=NfC31Ekj

Médiamétrie & Médiamétrie//NetRatings (2010). Les mesures hybrides – Synergies et rapprochement entre les mesures de l'Internet. *Le Livre Blanc*. https://www.mediametrie.fr/fr/les-publications-scientifiques-de-2008-2016

Neyman, J. (1934). On the Two Different Aspects of Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4), 558–625. https://www.jstor.org/stable/2342192

OCDE (2011). *Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities*. http://www.oecd.org/sdd/21687665.pdf

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286. https://ieeexplore.ieee.org/document/18626

Rabiner, L. R. & Juand, B.-H. (1993). *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall. https://dl.acm.org/citation.cfm?id=153687

Roy, G. & Vanheuverzwyn, A. (2001). Redressement par la macro CALMAR : applications et pistes d'amélioration. In: Lejeune, M. (Ed.), *Traitement des fichiers d'enquêtes*. Grenoble : Presses Universitaires de Grenoble. https://www.pug.fr/produit/314/9782706110295/traitements-des-fichiers-d-enquetes

Sautory, O. (1993). La macro CALMAR : redressement d'un échantillon par calage sur marges. Insee, *Méthodes*. https://www.insee.fr/fr/information/2021902

Tassi, P. (2014). La data est-elle éthique-compatible et quelques questions posées par les données. *8^{e} Colloque Francophone sur les Sondages*, Dijon, 2014. https://www.mediametrie.fr/fr/les-publications-scientifiques-de-2008-2016

Vanheuverzwyn, A. (2016). Mesure d'audience et données massives : mythes et réalités. *9^{e} Colloque Francophone sur les Sondages*, Gatineau, 2016. http://paperssondages16.sfds.asso.fr/submission_104.pdf

Econometrics and Machine Learning

Arthur Charpentier*, Emmanuel Flachaire and Antoine Ly*****

Abstract – On the face of it, econometrics and machine learning share a common goal: to build a predictive model, for a variable of interest, using explanatory variables (or features). However, the two fields have developed in parallel, thus creating two different cultures. Econometrics set out to build probabilistic models designed to describe economic phenomena, while machine learning uses algorithms capable of learning from their mistakes, generally for classification purposes (sounds, images, etc.). Yet in recent years, learning models have been found to be more effective than traditional econometric methods (the price to pay being lower explanatory power) and are, above all, capable of handling much larger datasets. Given this, econometricians need to understand what the two cultures are, what differentiates them and, above all, what they have in common in order to draw on tools developed by the statistical learning community with a view to incorporating them into econometric models.

Codes JEL / JEL Classification : C18, C52, C55
Keywords: learning, Big Data, econometrics, modelling, least squares

Reminder:

The opinions and analyses in this article are those of the author(s) and do not necessarily reflect their institution's or Insee's views.

* *University of Rennes 1 & CREM (arthur.charpentier@univ-rennes1.fr)*
** *Aix-Marseille University, AMSE, CNRS & EHESS (emmanuel.flachaire@univ-amu.fr)*
*** *University of Paris-Est (antoine.ly.pro@gmail.com)*

Received on 2 September 2017, accepted after revision on 29 May 2018
Translated from the original version: "Économétrie et *Machine Learning*"

Pour citer cet article: Charpentier, A., Flachaire, E. & Ly, A. (2018). Econometrics and Machine Learning. *Economie et Statistique / Economics and Statistics*, 505-506, pp. 147–169.
https://doi.org/10.24187/ecostat.2018.505d.1970

The earliest use of quantitative techniques in economics probably dates back to the sixteenth century (Morgan, 1990). However, it was not until the twentieth century that the term "econometrics" was first used, giving birth to the Econometric Society in 1933. Machine learning techniques are more recent. It was Arthur Samuel, widely regarded as the father of the first self-learning programme, who first coined the term "machine learning", which he defined as "a field of study that gives a computer the ability without being explicitly programmed" (Samuel, 1959). Among the earliest techniques are Hebb's cell assembly theory (Hebb, 1949) (which later gave birth to the "perceptron" in the 1950s, and then to neural networks), with Widrow and Hoff (1960) demonstrating, around fifteen years later, the links with least-squares methods, the SVM (support vector machine) and, more recently, boosting methods. While the two communities have developed in parallel, big data require links to be built between the two approaches by bridging the "two cultures" referred to by Breiman (2001a), contrasting mathematical statistics, which may be likened to traditional econometrics (Aldrich, 2010), with computational statistics and machine learning more generally.

Econometrics and supervised statistical learning techniques are similar, while also being very different. To start with, the two appear similar, with both using a database (or data table), i.e. observations $\{(y_i, x_i)\}$, with $i = 1, \cdots, n$, $x_i \in \mathcal{X} \subset \mathbb{R}^p$ and $y_i \in \mathcal{Y}$. If y_i is qualitative, we speak of a classification problem,[1] and, otherwise, of a regression problem. The two approaches also share common ground at the other end since, in both cases, the aim is to build a "model", i.e. a function $m : \mathcal{X} \mapsto \mathcal{Y}$ which will be interpreted as a prediction.

However, there are significant differences in between. Historically, econometric models have been based on economic theory, generally with parametric models. Traditional statistical inference methods (such as maximum likelihood and the method of moments) are thus used to estimate the values of a vector of parameters θ, in a parametric model $m_\theta(\cdot)$, by a value $\widehat{\theta}$. As in statistics, unbiased estimators are important since a lower bound on the variance can be obtained (Cramér-Rao bound). Asymptotic theory plays an important role (Taylor expansions, law of large numbers and central limit theorem). In statistical learning, by contrast, nonparametric models are often built based almost exclusively on data (i.e. no distribution hypothesis), and the meta-parameters used (tree-depth, penalty parameter, etc.) are optimised by cross-validation.[1]

Beyond the foundations, while the (often asymptotic) properties of $\widehat{\theta}$ (viewed as a random variable, thanks to the underlying stochastic representation) have been extensively studied in econometrics, statistical learning focuses to a greater extent on the properties of the optimal $m^\star(\cdot)$ based on a criterion that remains to be defined, or even simply $m^\star(x_i)$ for observations i deemed to be of interest for example in a test population. The problem of the choice of model is also viewed from a somewhat different perspective. Following Goodhart's law ("When a measure becomes a target, it ceases to be a good measure"), the goodness-of-fit of a model is penalised after the fact in econometrics by its complexity in the validation or selection phase, while in statistical learning it is the objective function which takes account of the penalty.

From High Dimension to Big Data

In this paper, a variable will be a vector of $\mathbb{R}^n$, such that by concatenating the variables, they can be stored in a matrix X, of size $n \times p$, with n and p being potentially large.[2] The fact that n is large is not a problem in itself. Many theorems in econometrics and statistics are obtained when $n \to \infty$. By contrast, the fact that p is large is problematic, particularly if $p > n$.

Portnoy (1988) showed that the maximum likelihood estimator retains the asymptotic normality property if p remains small in relation to n ($p^2/n \to 0$ where $n, p \to \infty$). Indeed, it is not uncommon to speak of high dimension when $p > \sqrt{n}$. Another important concept is the idea of "sparsity", which is based not on the dimension p but on the actual dimension, in other words the number of truly

[1] *The term "classification" will be used when $\mathcal{Y}$ is a set of classes, typically a binary classification, $\mathcal{Y} = \{0,1\}$, corresponding to the outcome of an indicator variable. The term is less dated than "discrimination" and more general than the determination of a "score" (often an intermediate step). It should not be confused with unsupervised classification (such as "ascending hierarchical classification"), which involves the creation of homogeneous classes based on a similarity measure (in this case, the term "creation or construction of classes" or "clusters" is sometimes used).*

[2] *Extensions are possible with MRI-type images as predictive variables, or climate data with maps as predictive variables. It is possible fall back on the typical case of data in the form of vectors by using the Tucker decomposition (Kolda & Bader, 2009).*

important variables. It is thus possible to have $p > n$ while having convergent estimators.

The high dimension can be frightening because of the curse of dimensionality (Bellman, 1957). The volume of the unit sphere, in dimension p, tends towards 0 when $p \to \infty$. In such cases, the space is said to be "parsimonious" – i.e. the likelihood of finding a point close to another becomes increasingly small (we may even speak of a "sparse" space). While the idea of reducing the dimension by using a principal component analysis may seem attractive, the analysis suffers from a number of flaws in high dimension. The solution often revolves around the selection of variables (which raises the problem of multiple tests or computational time).

To use the terminology of Bühlmann & van de Geer (2011), the problems highlighted here correspond to those encountered in high dimension, an essentially statistical problem. From a computational perspective, we may go a little further, with truly Big Data. In the foregoing, the data were stored in a matrix X, of size $n \times p$. There can be issues with storing such a matrix or even with using a matrix widely used in econometrics, $X^T X$ ($n \times n$). The first-order condition of the linear model is associated with the solution to $X^T\left(X\beta - y\right) = 0$. In reasonable dimension, the Gram-Schmidt decomposition is used. In high dimension, the numerical descent and gradient methods may be used, where the gradient is approximated by subsampling (Zinkevich *et al.,* 2010). This computational dimension is often overlooked, despite the fact that it has been the basis of a significant number of methodological advances in econometrics.

Nonparametric and Computational Statistics

The purpose of this paper is to explain the major differences between econometrics and statistical learning, which correspond to the cultures alluded to by Breiman (2001a) in referring, in the context of statistical modelling, to the data modelling culture (based on a stochastic model, such as logistic regression or a Cox model) and the algorithmic modelling culture (based on the implementation of an algorithm, such as random forests or support vector machines; for a complete list, see Shalev-Shwartz & Ben-David, 2014). However, the boundary between the two is blurred. At the intersection, we find, for example, nonparametric econometrics, which is based on a probabilistic model (like econometrics) while focusing to a greater extent on algorithms and their performance rather than on asymptotic theorems.

Some Machine Learning Tools

Neural Networks

Neural networks are semiparametric models. Nevertheless, this family of models can be approached in the same way as nonparametric models: the structure of neural networks (presented below) can be modified to extend the class of functions used to approximate a variable of interest. More specifically, Cybenko (1989) showed that the set of neural functions is dense in the space of continuous functions on a compact space. In other words, we have a theoretical framework which guarantees a form of universal approximation. It also requires defining a neuron and emphasises the existence of a sufficient number of neurons to approximate any continuous function on a compact domain. Thus, a continuous phenomenon can be approximated by a sequence of neurons: this sequence is referred to as a "single-layer neural network". While the universal approximation theorem was demonstrated in 1989, the first functional artificial neuron was introduced by Franck Rosenblatt in the mid-twentieth century in Rosenblatt (1958). Referred to now as "basic neuron", this neuron is known as "Perceptron". In its earliest uses, it was used to determine the gender of an individual presented in a photo. It introduced the first mathematical representation of a biological neuron:

- The synapses transmitting the information to the cell are represented by a real vector. The dimension of the input vector of the neuron (which is none other than a function) corresponds biologically to the number of synaptic connections;

- Each signal transmitted by a synapse is then analysed by the cell. Mathematically, the schema is expressed by weighting the different components of the input vector;

- Depending on the information acquired, the neuron decides whether or not to resend a signal. The phenomenon is replicated by introducing an activation function. The output signal is modelled by a real number computed as an image by the activation function of the weighted input vector.

Thus, an artificial neuron is a semiparametric model. The choice of activation function is left to the user. A basic neuron may then be formally defined by:

1. An input space $\mathcal{X}$, generally $\mathbb{R}^k$ with $k \in \mathbb{N}^*$;

2. An output space $\mathcal{Y}$, generally $\mathbb{R}$ or a finite set (typically $\{0,1\}$, although here we prefer $\{-1,+1\}$);

3. A vector of parameters $w \in \mathbb{R}^p$;

4. An activation function $\phi : \mathbb{R} \to \mathbb{R}$. Ideally, this function should be monotonic, derivable and bounded (here, "saturating") to ensure certain convergence properties.

This last function ϕ is reminiscent of logistic or probit transformations, which are popular in econometrics (which are cumulative distribution functions, of value in $[0,1]$, ideal when $\mathcal{Y}$ is the set $\{0,1\}$). For neural networks, preference is given to the hyperbolic tangent, the arctangent function or the sigmoid functions for classification problems on $\mathcal{Y} = \{-1,+1\}$ (the latter evoke the logistic transformation performed by econometricians). The term neuron is used to refer to any application f_w of $\mathcal{X}$ in $\mathcal{Y}$ defined by:

$$y = f_w(x) = \phi(w^T x), \quad \forall x \in \mathcal{X}$$

For the perceptron, introduced by Rosenblatt (1958), a basic neuron is assimilated to the function:

$$y = f_w(x) = signe(w^T x), \quad \forall x \in \mathcal{X}$$

According to this formalisation, many statistical models, such as logistic regressions, may be viewed as neurons. Any GLM (Generalised Linear Model) could be interpreted as an artificial neuron where the activation function ϕ is none other than the inverse of the canonical link function. If g denotes the link function of the GLM, w the vector of parameters, y the variable to be explained and x the vector of explanatory variables of the same dimension as w:

$$g(\mathbb{E}[Y \mid X = x]) = w^T x$$

We return to neural modelling by taking $\phi = g^{-1}$. However, the chief difference between GLMs and the neural model is that the latter requires no distribution hypothesis on $Y \mid X$ (here there is no need to introduce a probabilistic model). Furthermore, when the number of neurons per layer increases, convergence is not necessarily guaranteed if the activation function does not verify certain properties (which is not the case for the majority of the canonical link functions of GLMs). However, neural network theory imposes additional mathematical constraints on the function g (detailed in Cybenko, 1989). Thus, for example, a logistic regression may be viewed as a neuron, whereas generalised linear regressions do not verify all the necessary hypotheses.

To extend the analogy with the functioning of the nervous system, it is then possible to connect different neurons. We speak of a layered neural network structure. Each layer of neurons receives the same observation vector every time. To revert to a more econometric analogy, we might imagine an intermediate step, for example by not performing a regression on the raw variables x but a smaller set of orthogonal variables obtained based on a principal component analysis. Consider A as the matrix associated with this linear transformation, with A of size $k \times p$ if we wish to use the p first components. Take z as the transformation of x, where $z = A^T x$ ($z_j = A_j{}^T x$). One generalisation of the above model may be to posit:

$$y = f(x) = \phi(w^T z) = \phi(w^T A^T x), \quad \forall \mathrm{x} \in \mathcal{X}$$

where $w \in \mathbb{R}^p$. Here we have a linear transformation (by considering a principal component analysis), although we can imagine a generalisation with nonlinear transformations:

$$y = f(x) = \phi(w^T F_A(x)), \quad \forall x \in \mathcal{X}$$

where F is a function $\mathbb{R}^k \to \mathbb{R}^p$. It is the two-layer neural network. More generally, in order to formalise the construction, the following notations are introduced:

- $K \in \mathbb{N}^*$: number of layers;

- $\forall k \in \{1, \cdots K\}$, p_k represents the number of neurons in the layer k;

- $\forall k \in \{1, \cdots K\}$, W_k denotes the matrix of the parameters associated with the layer k. More specifically, W_k is a matrix $p_k \times p_{k-1}$ and for any $\in \{1, \cdots p_k\}$, $w_{k,l} \in \mathbb{R}^{p_{k-1}}$ denotes the weight vector associated with the basic neuron l of the layer k;

- $W = \{W_1, .., W_K\}$ denotes the set of parameters associated with the neural network;

- $F^k_{W_k} : \mathbb{R}^{p_{k-1}} \to \mathbb{R}^{p_k}$ denotes the transfer function associated with the layer k. For the purpose of simplification, we may also write F^k;

- $\widehat{y}_k \in \mathbb{R}^{p_k}$ will represent the image vector of the layer $k \in \{1, \cdots, K\}$;

- $F = F_W = F^K \circ \cdots \circ F^1$ will denote the transfer function associated with the global network. In this respect, if $x \in \mathcal{X}$, we may note $\widehat{y} = F_W(x)$.

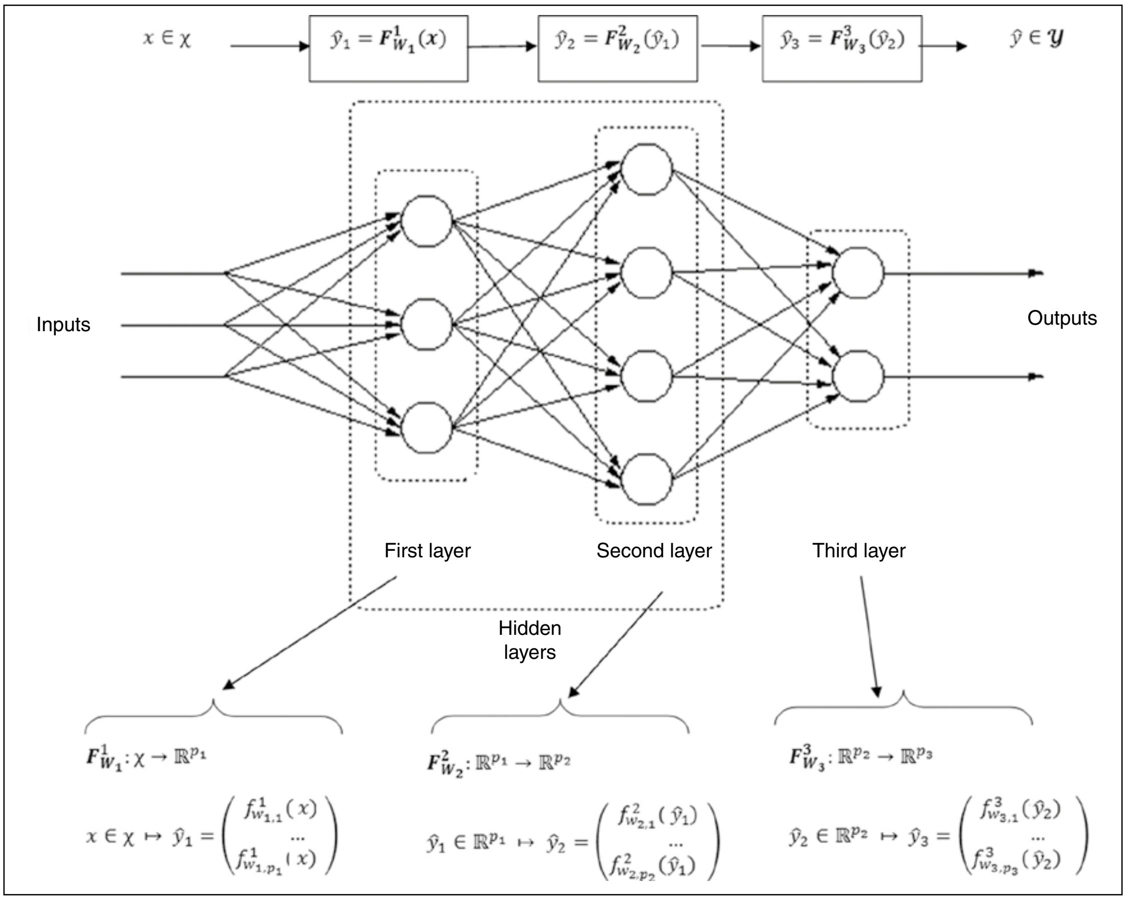
Diagram 1 provides an illustration of the notations presented here.[3] Each circle represents a basic neuron. Each rectangle encompassing several circles represents a layer. The first layer taking as "input" the observations $x \in \mathcal{X}$, is referred to as the input layer, while the output layer denotes the layer providing as "output" the prediction $\widehat{y} \in \mathcal{Y}$. The other layers are commonly known as hidden layers. A multilayer neural network is, therefore, a semiparametric model whose parameters are the set of components of the matrices W_k for any integer k of $\{1, \cdots, K\}$. Each activation function associated with each neuron (each circle of Diagram I) is to be determined by the user.

Once the model parameters to be calibrated have been identified (here, the reals forming the matrices W_k for each layer $k \in \{1, \cdots, K\}$), it is necessary to define a loss function ℓ. Indeed, it is worth recalling that the aim of supervised learning on a learning basis of $n \in \mathbb{N}^*$ couples $(y_i, x_i) \in \mathcal{Y} \times \mathcal{X}$ is to minimise the empirical risk (see Online complements – see the link at the end of the article):

$$\widehat{\mathcal{R}}_n(F_W) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, F_W(x_i))$$

[3] See: http://intelligenceartificielle.org.

Diagram 1
Example of Notations Associated with the Multilayer Neural Networks


$x \in \chi$
$\hat{y}_1 = F^1_{W_1}(x)$
$\hat{y}_2 = F^2_{W_2}(\hat{y}_1)$
$\hat{y}_3 = F^3_{W_3}(\hat{y}_2)$
$\hat{y} \in \mathcal{Y}$
Inputs
Outputs
First layer
Second layer
Third layer
Hidden layers
$F^1_{W_1}: \chi \to \mathbb{R}^{p_1}$
$x \in \chi \mapsto \hat{y}_1 = \begin{pmatrix} f^1_{W_{1,1}}(x) \\ \cdots \\ f^1_{W_{1,p_1}}(x) \end{pmatrix}$
$F^2_{W_2}: \mathbb{R}^{p_1} \to \mathbb{R}^{p_2}$
$\hat{y}_1 \in \mathbb{R}^{p_1} \mapsto \hat{y}_2 = \begin{pmatrix} f^2_{W_{2,1}}(\hat{y}_1) \\ \cdots \\ f^2_{W_{2,p_2}}(\hat{y}_1) \end{pmatrix}$
$F^3_{W_3}: \mathbb{R}^{p_2} \to \mathbb{R}^{p_3}$
$\hat{y}_2 \in \mathbb{R}^{p_2} \mapsto \hat{y}_3 = \begin{pmatrix} f^3_{W_{3,1}}(\hat{y}_2) \\ \cdots \\ f^3_{W_{3,p_3}}(\hat{y}_2) \end{pmatrix}$

To illustrate this point, let us consider the following example, which will also serve to illustrate the approach adopted. Let us assume that we are observing a phenomenon through observations $y_i \in [-1,1]$. The aim is to explain this phenomenon based on the independant variables x which are assumed to have actual values. The "universal approximation theorem" tells us that a single-layer neural network should enable the phenomenon to be modelled (subject to it being continuous). However, the theorem provides no indication of convergence speed. The user retains control of the choice of structure, which may be a simple neuron whose activation function is the hyperbolic tangent function:

$$y_1 = \tanh\left(w_0 + w_1 x\right)$$

where the parameters w_0, w_1 are to be optimised in order to minimise the empirical risk over the learning data.

Based on the universal approximation theorem, by adding several neurons, the error is expected to reduce. However, since the function to be estimated is not known, it can only be observed through the sample. Mechanically, learning-based error decreases when parameters are added. Error analysis by means of a test enables our ability to generalise to be assessed (Box 1).

A second model, which uses several neurons, may thus be considered. For example:

$$\begin{aligned} y_2 = {} & w_a \tanh\left(w_0 + w_1 x\right) + w_b \tanh\left(w_2 + w_3 x\right) \\ & + w_c \tanh\left(w_4 + w_5 x\right) \end{aligned}$$

where the parameters $w_0, .., w_5$ and w_a, w_b, w_c are the parameters to be optimised. Calibrating a neural network thus amounts to reiterating these structural modification steps until the risk is minimised on a test basis.

For a fixed neural network structure (i.e. fixed number of layers, number of neurons per layer and activation functions), the programme therefore amounts to determining the set of parameters $W^* = \left(W_1, ..., W_K\right)$ in such a way that:

$$W^* \in \underset{W=(W_1,...,W_K)}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, F_W\left(x_i\right)\right) \right\}.$$

This formula underlines the importance of the choice of function ℓ. This loss function quantifies the average error of our model F_W based on learning. *A priori*, ℓ can be chosen arbitrarily. However, from the point of view of working out an optimisation programme, sub-differentiable and convex cost functions are preferable for guaranteeing the convergence of the optimisation algorithms. In addition to the quadratic loss function $\ell_2(y, \widehat{y}) = (y - \widehat{y})^2$, traditional loss functions include the hinge function $-\ell(y, \widehat{y}) = max(0, 1 - y\widehat{y})$ – and the logistic function $-\ell(y, \widehat{y}) = \log(1 - e^{-y\widehat{y}})$.

Neural networks were used very early on in economics and finance, notably on corporate defaults (Tam & Kiang, 1992; Altman *et al.,* 1994) and, more recently, credit rating (Blanco *et al.,* 2013; Khashman, 2011). However, structures such as those presented above are generally limited. Deep learning is more particularly characteristic of more complex neural networks (sometimes more than ten layers with hundreds of neurons per layer). Today, these

Box 1 – **Learning and Test Samples**

In the literature on learning, assessing the quality of a model based on the data used to build it says nothing about how the model will behave with new data. This is what is known as the "generalisation" problem. The traditional approach thus involves splitting the sample (of size n) in two: one part to train the model (the learning base, in-sample, of size m) and another to test it (the test base, out-of-sample, of size $n - m$). The latter allows for the measurement of a real predictive risk Let us suppose that the data are generated by a linear model $y_i = x_i^T \beta_0 + \varepsilon_i$ where the ε_i are realisations of independent centred distributions. The in-sample empirical quadratic risk is:

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left(\left[x_i^T \widehat{\beta} - x_i^T \beta_0\right]^2\right) = \mathbb{E}\left(\left[x_i^T \widehat{\beta} - x_i^T \beta_0\right]^2\right)$$

for any observation i. If the residuals ε are Gaussian, this risk equals $\sigma^2 p / m$, where p is the size of the vectors x_i. By contrast, the out-of-sample empirical quadratic risk is:

$$\mathbb{E}\left(\left[\mathrm{x}^{\mathrm{T}} \widehat{\beta} - x^T \beta_0\right]^2\right)$$

Where x is a new observation, which is independent of the others. We may note that:

$$\mathbb{E}\left(\left[x^T \widehat{\beta} - x^T \beta_0\right]^2 \mid x\right) = \sigma^2 x^T (X^T X)^{-1} x$$

and by integrating in relation to x:

$$\begin{aligned} \mathbb{E}\left(\left[x^T \widehat{\beta} - x^T \beta_0\right]^2\right) &= \mathbb{E}\left(\mathbb{E}\left(\left[x^T \widehat{\beta} - x^T \beta_0\right]^2 \mid x\right)\right) \\ &= \sigma^2 \operatorname{trace}\left(\mathbb{E}[xx^T]\,\mathbb{E}[X^T X]^{-1}\right) \end{aligned}$$

➔

Box 1 (contd.)

The expression is then different from that obtained in-sample, and by drawing on Groves & Rothenberg (1969), we can show that:

$$\mathbb{E}\left(\left[x^T\widehat{\beta}-x^T\beta_0\right]^2\right)\geq\sigma^2\frac{p}{m}$$

Except for certain simple cases, there is no simple formula. We may note, however, that if $x\sim\mathcal{N}\left(0,\sigma^2\mathbb{I}\right)$, then x^Tx follows a Wishart distribution, and:

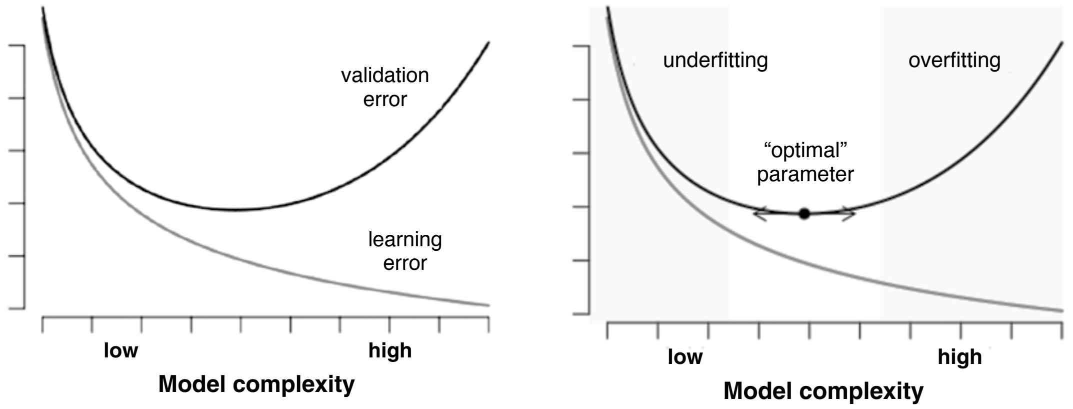
$$\mathbb{E}\left(\left[x^T\widehat{\beta}-x^T\beta_0\right]^2\right)=\sigma^2\frac{p}{m-p-1}$$

Let us now consider the empirical version: if $\widehat{\beta}$ is estimated on the m first observations,

$$\widehat{\mathcal{R}}^{\text{IS}}=\sum_{i=m+1}^{m}[y_i-x_i^T\widehat{\beta}]^2 \text{ and } \widehat{\mathcal{R}}^{\text{OS}}=\sum_{i=m+1}^{n}[y_i-x_i^T\widehat{\beta}]^2$$

and as noted by Leeb (2008), $\widehat{\mathcal{R}}^{\text{IS}}-\widehat{\mathcal{R}}^{\text{OS}}\approx 2\cdot\nu$ where ν represents the number of degrees of freedom. Figure A shows the respective evolution of $\widehat{\mathcal{R}}^{\text{IS}}$ and $\widehat{\mathcal{R}}^{\text{OS}}$ according to the complexity of the model (number of degrees in a polynomial regression, number of nodes in splines, etc.). $\widehat{\mathcal{R}}^{\text{IS}}$ always decreases with complexity (light curve). However, $\widehat{\mathcal{R}}^{\text{IS}}$ is non-monotonic (dark curve). If the model is too simple, it is a poor predictor, but if it is too complex, "over-learning" arises: it starts to model noise.

Figure A
Generalisation and Over-Learning



Reading note: The light curve represents the in-sample empirical risk on the learning sample, while the dark curve represents the out-of-sample risk on the test sample.

structures are very popular in signal analysis (image, text, sound) because they are capable, based on a very large quantity of observations, of extracting information which humans are incapable of perceiving and to deal with non-linear problems (LeCun *et al.,* 2015).

Information extraction can, for example, be performed through convolution. As an unsupervised procedure, it has produced excellent results in image analysis. In technical terms, this may be seen as a kernel-based transformation (as used in SVM techniques; see next section). While an image may be viewed as a matrix, with each coordinate representing a pixel, a convolution amounts to applying a transformation to a point (or area) of this matrix, thereby producing a new datum. The procedure can thus be repeated by applying different transformations (hence the notion of convolutional layers). The final vector obtained can then be fed into a neural model. More generally, a convolutional layer may be seen as a filter allowing the initial datum to be transformed.

One intuitive explanation for deep learning, and particularly deep networks, being so powerful for describing complex relationships in data is their construction around a simple functional approximation and the use of a form of hierarchy (Lin *et al.,* 2016). Nevertheless, deep learning models are more difficult to use since they require a significant degree of empirical judgement. While open-source libraries (Keras, Torch, etc.) currently allow more readily for parallel computations by using, for example, GPUs (Graphical Processor Units), the user is still required to determine the structure of the most appropriate neural network.

Support Vector Machines

As noted above, in machine learning classification problems (as in signal processing), observations in the set $\{-1,+1\}$ are preferable (rather than $\{0,1\}$ in econometrics). With this notation, Cortes & Vapnik (1995) laid the theoretical foundations of support vector machine (SVM) models, an alternative to the then very popular neural networks. The initial idea of SVM methods involves finding a separating hyperplane dividing space into two sets of points as homogeneously as possible (i.e. containing identical labels). In dimension two, the algorithm involves determining a line separating the space into two areas that are as homogeneous as possible. Since it is a problem which may sometimes have an infinite number of solutions (there may be an infinity of lines separating the space into two distinct and homogeneous areas), an additional constraint is generally added: the separating hyperplane must be located as far as possible from the two homogeneous subsets which it generates (Diagram 2). In such cases, we speak of margin. The algorithm thus described is a soft- or hard-margin linear SVM.

If a plane can be entirely characterised by a directional vector *w* orthogonal to the latter and a constant *b*, applying an SVM algorithm to a set of $n \in \mathbb{N}^*$ points x_i of $\mathbb{R}^p$ labelled by $y_i \in \{-1,1\}$ thus amounts to solving a constrained optimisation programme similar to a lasso problem (quadratic deviation under linear constraint; see Online complements – link at the end of the article). In particular, we are led to solving the following:

$$\left(w^\star, b^\star\right) = \underset{w,b}{\operatorname{argmin}}\left\{\| w \|^2\right\} = \underset{w,b}{\operatorname{argmin}}\left\{w^T w\right\}$$

under constraints

$$\forall i \in \{1,\cdots,n\}, \begin{cases}\omega^T x_i + b \geq +1 \text{ when } y_i = +1\\ \omega^T x_i + b \leq -1 \text{ when } y_i = -1\end{cases}$$

The constraint can be loosened by allowing a point in a subset not to have the same label as the majority of the points in the subset provided it is not too far from the boundary. These are known as soft-margin linear SVMs. Heuristically, and indeed in practice, we cannot have $y_i\left(w^T x_i + b\right) - 1 \geq 0$ for any $i \in \{1,\cdots,n\}$; we loosen by introducing positive variables ξ such that:

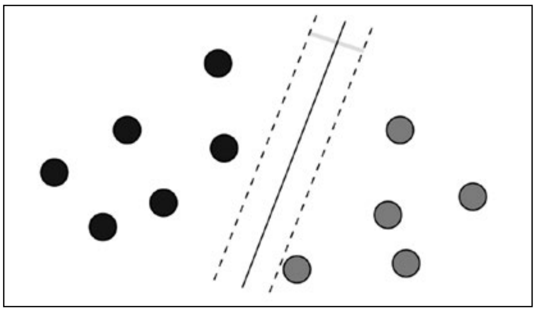
$$\begin{cases}\omega^T x_i + b \geq +1 - \xi_i \text{ lorsque } y_i = +1\\ \text{É}^{\mathrm{T}}\mathrm{x}_{\mathrm{i}} + b \leq -1 + \xi_i \text{ lorsque } y_i = -1\end{cases} \qquad (1)$$

with $\xi_i \geq 0$. A misclassification occurs if $\xi_i > 1$, and a penalty is then applied as a price to pay for each error. The aim then is to solve a quadratic problem:

$$\min\left\{\frac{1}{2}\omega^T \omega + C1^T 1_{\xi>1}\right\}$$

under constraint (1), which can be efficiently solved numerically by coordinate descent.

Diagram 2
Illustration of a Margin SVM



Sources: Vert (2017).

If the points cannot be separated, another possibility is to transfer them into a higher dimension in such a way that the data become linearly separable. Finding the right transformation separating the data is, however, very difficult. One mathematical trick for elegantly solving this problem involves defining the transformations $T(\cdot)$ and the scalar products using a kernel $K(x_1,x_2) = \langle T(x_1), T(x_2)\rangle$. One of the most common choices for a kernel function is the radial basis function (Gaussian kernel) $K(x_1,x_2) = \exp\left(-\|x_1 - x_2\|^2\right)$. However, no rules have so far been devised for choosing the "best" kernel. This technique is based on distance minimisation and does not predict the probability of being positive or negative, although a probabilistic interpretation is nonetheless possible (Grandvalet *et al.*, 2005).

Trees, Bagging and Random Forests

Classification trees were introduced by Breiman *et al.* (1984) and then by Quinlan (1986). We speak of CART, or Classification and Regression Tree. The idea is to divide (based on the notion of branching) the input data consecutively until an allocation criterion (in relation to the target variable) is reached, based on a pre-defined rule.

The intuition: entropy $H(x)$ is associated with the amount of disorder in the data x in relation to the modalities of the classification variable y, and each partition aims to reduce this disorder. The probabilistic interpretation is to create groups that are as homogeneous as possible by reducing the variance of each group (intra-group variance), or in an equivalent manner by creating two groups that are as different as possible by increasing the variance between the groups (inter-group variance). At each stage, the partition providing the most significant reduction of disorder (or of variance) is chosen. The complete decision tree is developed by repeating this procedure across all the sub-groups, where each step results in a new partition into 2 branches, which subdivides the dataset into 2. Lastly, a decision about when to put an end to the creation of new branches is made by carrying out the final allocations (leaf nodes). There are several options. One option is to build a tree until all leaves are pure, i.e. composed of a single observation. Another option is to define a stopping rule linked to the size or decomposition of the leaves. Examples of stopping rules can be of minimum size (at least 5 elements per leaf) or minimum entropy. We speak of the pruning of the tree: the tree is allowed to grow, and then certain branches are cut *a posteriori* (which is different from introducing a stopping criterion *a priori* to the growth process of the tree – for example by imposing a minimum size on the leaves, or other criteria discussed in Breiman *et al.*, 1984).

At a given node, formed of n_0 observations (x_i, y_i) with $i \in \mathcal{I}_0$, we cut into two branches (one on the left and one on the right), thus partitioning $\mathcal{I}_0$ into $\mathcal{I}_g$ and $\mathcal{I}_d$. Let I be the criterion of interest, such as the entropy of the node:

$$I(y_0) = -n_0 p_0 \log p_0 \text{ where } p_0 = \frac{1}{n_0}\sum_{i\in\mathcal{I}_0} y_i$$

or the variance of the node:

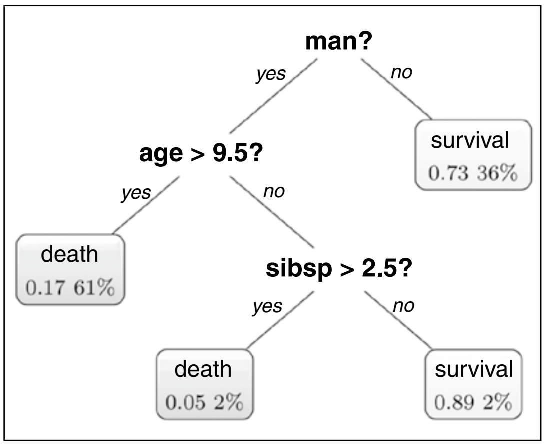
$$I(y_0) = n_0 p_0 (1 - p_0) \text{ where } p_0 = \frac{1}{n_0}\sum_{i\in\mathcal{I}_0} y_i$$

the latter also being the Gini impurity index.

The left and right branches are partitioned if the gain $I(y_0) - \left[I(y_{\mathrm{g}}) + I(y_d)\right]$ is sufficiently significant. In the construction of the trees, the aim is to determine the partition that provides the greatest possible gain. This combinatorial problem being complex, Breiman *et al.* (1984) proposed a partition according to one of the variables, with $\mathcal{I}_{\mathrm{g}} = \{i \in \mathcal{I}_0 : x_{k,i} < s\}$ and $\mathcal{I}_{\mathrm{d}} = \{i \in \mathcal{I}_0 : x_{k,i} > s\}$, for a variable k and a threshold s (if the variable is continuous; otherwise, groupings of modalities are considered for qualitative variables).

The decision trees thus described are simple to obtain and easy to interpret (as shown by Diagram 3 on the data of the Titanic[4]), although they are not robust, and their predictive power is often very limited, particularly if the tree is very deep. One obvious idea is to develop a set of more or less independent tree models which, together, predict better than a single-tree model. The bootstrap method will be used, by sampling (with replacement) n observations among $\{(x_i, y_i)\}$. Each sample thus generated can be used to estimate a new classification tree, thus forming a forest of trees. It is the aggregation of all these trees that gives the prediction. The overall result is less sensitive than the initial sample and often gives better prediction results. These techniques, known as bagging (short for bootstrap aggregating), are similar to bootstrap techniques in regression (for example to construct confidence tubes in a functional regression).

Diagram 3
Illustration of a Decision Tree Used to Predict the Survival Rate of a Passenger on the Titanic



Reading note: A woman (man: no) had a 73% chance of survival, with women representing 36% of the population.

[4] *This dataset, which contains information on all the passengers and crew members on the Titanic, with the variable indicating whether the person survived, has been widely used to illustrate classification techniques, see https://www.kaggle.com/c/titanic/data.*

Bagging involves generating random samples by sampling with replacement from the original sample, as with the bootstrap method. Random forests are based on the same principle as bagging, but during the construction of a classification tree, at each branch, a subset of m covariates is drawn randomly. In other words, each branch of a tree is not based on the same set of covariates. This helps to increase the variability between the trees and, ultimately, to obtain a forest composed of less correlated trees.

Choice of Classification Model

Given a model $m(\cdot)$ approximating $\mathbb{E}[Y \mid X = x]$, and a threshold $s \in [0,1]$, let us posit:

$$\hat{y}^{(s)} = 1[m(x) > s] = \begin{cases} 1 \text{ si } m(x) > s \\ 0 \text{ si } m(x) \leq s \end{cases}$$

The confusion matrix is then the contingency table associated with the countings $N = [N_{u,v}]$ with:

$$N_{u,v}^{(s)} = \sum_{i=1}^{n} 1\left(\hat{y}^{(s)} = u, y_j = v\right)$$

for $(u,v) \in \{0,1\}$. Table 1 presents such a matrix, with the name of each of the components: TP for true positives, corresponding to the 1 predicted in 1, TN for true negatives, corresponding to the 0 predicted in 0, FP for false positives, corresponding to 0 predicted in 1, and FN for false negatives, corresponding to 1 predicted in 0.

Several quantities are derived from this table. Sensitivity is the probability of predicting 1 in the population of 1, or the true positive rate. Specificity is the probability of predicting 0 in the population of 0 or the true negative rate. However, the true negative rate will be of greater interest, i.e. the probability of predicting 1 in the population of 0. The representation of these two values when s varies gives the ROC curve (receiver operating characteristic):

$$ROC_s = \left(\frac{FP_s}{FP_s + VN_s}, \frac{VP_s}{VP_s + FN_s}\right)$$
$$= (sensitivity_s, 1 - specificity_s) \quad \text{pour } s \in [0,1]$$

Table 1
Confusion Matrix, or Contingency Table for a Given Threshold *s*

	$y = 0$	$y = 1$	
$\hat{y}_s = 0$	VN_s	FN_s	$VN_s + FN_s$
$\hat{y}_s = 1$	FP_s	VP_s	$FP_s + VP_s$
	$VN_s + FP_s$	$FN_s + VP_s$	n

This curve is presented in the next section, based on real data. The two values widely used in machine learning are the index κ, which compares observed and expected accuracy using a random model (Landis & Koch, 1977), and the AUC, corresponding to the area under the ROC curve. For the first index, once s is chosen, let $N^{\perp}$ be the contingency table corresponding to independent cases (defined based on N in the chi-square independence test. We then posit:

$$total\ precision = \frac{TP + TN}{n}$$

whereas:

$$random\ precision = \frac{[TN + FP] \cdot [TP + FN] + [TP + FP] \cdot [TN + FN]}{n^2}$$

We may then define:

$$\kappa = \frac{total\ precision - random\ precision}{1 - random\ precision}$$

Traditionally, s will be set at 0.5, as in naive Bayesian classification, although other values may be retained, in particular if the two errors are not symmetrical. There are compromises between simple and complex models measured by the number of parameters (or degrees of freedom more generally) in terms of performance and cost. Simple models are generally easier to compute, but can also lead to poorer goodness-of-fit (with high bias, for example). By contrast, complex models can provide a more accurate goodness-of-fit, but also risk being more costly in terms of computation. Furthermore, they go beyond the data or have greater variance and, just as with overly simple models, present significant test errors. As noted above, in machine learning, the optimal model complexity is determined using the bias-variance compromise.

From Classification to Regression

Historically, machine learning methods have focused on classification problems (with possibly more than 2 modalities[5]), with relatively little interest being shown in cases

[5] *For example, in the case of letter or number recognition.*

where the variable of interest y is continuous. Nevertheless, a number of techniques can be adapted, such as trees and random forests, boosting and neural networks.

In the case of regression trees, Morgan & Sonquist (1963) proposed the AID method, based on the variance decomposition formula with an algorithm similar to the algorithm of the CART method described above. In a classification context, we would calculate, at each node (in the case of the Gini impurity index by adding on the left leaf $\{x_{k,i} < s\}$ and the right leaf $\{x_{k,i} > s\}$:

$$I = \sum_{i:x_{k,i}<s} \bar{y}_g\left(1-\bar{y}_g\right) + \sum_{i:x_{k,i}>s} \bar{y}_d\left(1-\bar{y}_d\right)$$

where $\bar{y}_g$ and $\bar{y}_d$ denote the frequencies of 1 in the left and right leaf, respectively. In the case of a regression tree, we use:

$$I = \sum_{i:x_{k,i}<s} (y_i - \bar{y}_g)^2 + \sum_{i:x_{k,i}>s} (y_i - \bar{y}_d)^2$$

corresponding to the (weighted) sum of intra-group variance. The optimal distribution is the distribution with the highest intra-group variance (the aim is for the leaves to be as homogeneous as possible).

In the context of random forests, a majority criterion is often used in classification (the predicted class is the majority class in a leaf), whereas for regression the predictions across all the trees are averaged. In a regression context (y continuous variable), the idea is to create a succession of models based on the boosting method (Box 2), which, in this case, takes the form:

$$m^{(k)}(x) = m^{(k-1)}(x) + \alpha_k \underset{h\in\mathcal{H}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - m^{(k-1)}(x) + h(x))^2 \right\}$$

where α_k is a shrinkage parameter and where the second term corresponds to a regression tree, on the residuals, $y_i - m^{(k-1)}(x_i)$. However, there are other techniques which allow for sequential learning. In an additive model (GAM), the aim is look for a notation in the form:

$$m(x) = \sum_{j=1}^{p} m_j\left(x_j\right) = m_1\left(x_1\right) + \cdots + m_p\left(x_p\right)$$

The idea of projection pursuit is based on a decomposition of the linear combinations and not of the explanatory variables. Let us consider a model:

$$m(x) = \sum_{j=1}^{k} g_j\left(\omega_j^T x\right) = g_1\left(\omega_1^T x\right) + \cdots + g_k\left(\omega_k^T x\right)$$

As with additive models, the functions $g_1, \cdots, g_k$ are to be estimated, as are the directions $\omega_1, \cdots, \omega_k$. This notation is relatively general and allows for interactions and cross effects to be considered (which is something that could not be done with additive models, which do not take into account nonlinearities). For example, in dimension 2, a multiplicative effect $m\left(x_1, x_2\right) = x_1 \cdot x_2$ is expressed as follows:

$$m\left(x_1, x_2\right) = x_1 \cdot x_2 = \frac{(x_1 + x_2)^2}{4} - \frac{(x_1 - x_2)^2}{4}$$

in other words $g_1(x) = x^2/4$, $g_2(x) = -x^2/4$, $\omega_1 = (1,1)^T$ and $\omega_2 = (1,-1)^T$. In the simple version, with $k = 1$, with a quadratic loss function, we may use a Taylor expansion to approximate $[y_i - g\left(\omega^T x_i\right)]^2$, and construct an

Box 2 – **Slow Learning by Boosting**

The idea of boosting, introduced by Shapire & Freund (2012), is to learn slowly from the errors of the model, in an iterative manner. In the first stage, a model m_1 is estimated for y, based on X, giving error ε_1. In the second stage, a model m_2 is estimated for ε_1, based on X, giving error ε_2, etc. After k iterations, the model is then selected:

$$\begin{aligned} m^{(k)}(\cdot) &= \underbrace{m_1(\cdot)}_{\sim y} + \underbrace{m_2(\cdot)}_{\sim \varepsilon_1} + \underbrace{m_3(\cdot)}_{\sim \varepsilon_2} + \cdots + \underbrace{m_k(\cdot)}_{\sim \varepsilon_{k-1}} \qquad (2) \\ &= m^{(k-1)}(\cdot) + m_k(\cdot) \end{aligned}$$

Here, the error ε is seen as the difference between y and model $m(x)$, but it may also be seen as the gradient associated with the quadratic loss function.

Equation (2) may be seen as a gradient descent, but expressed dualistically. The problem will then be recast as an optimisation problem:

$$m^{(k)} = m^{(k-1)} + \underset{h\in\mathcal{H}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell\left(y_i - m^{(k-1)}(x_i), h(x_i)\right) \right\} (3)$$

where the space $\mathcal{H}$ is relatively simple (in such cases we speak of a weak learner). Traditionally, the functions $\mathcal{H}$ are staircase functions (found in classification and regression trees) known as stumps. To ensure that learning is slow, it is not uncommon for a shrinkage parameter to be used, and rather than positing, for example, $\varepsilon_1 = y - m_1(x)$, $\varepsilon_1 = y - \alpha \cdot m_1(x)$ is posited, with $\alpha \in [0,1]$.

iterative algorithm in the standard way. If we have an initial value ω_0, let us note that:

$$\sum_{i=1}^{n}[y_i - g\left(\omega^T x_i\right)]^2 \approx \sum_{i=1}^{n} g'(\omega_0{}^T x_i)^2 \left[\omega^T x_i + \frac{y_i - g\left(\omega_0{}^T x_i\right)}{g'\left(\omega_0{}^T x_i\right)} - \omega^T x_i\right]^2$$

corresponding to approximation in the generalised linear models on the function $g(\cdot)$ which was the link function (assumed to be known). We recognise a weighted least squares problem. The difficulty here is that the functions $g_j(\cdot)$ are unknown.

Applications

Big Data have required the development of estimation techniques capable of overcoming the limitations of parametric models, which are seen as too restrictive, and of traditional nonparametric models, whose estimation can be difficult in the presence of a large number of variables. Statistical learning, or machine learning, provides new nonparametric estimation methods, which perform well in a general context and in the presence of a large number of variables.[6] However, greater flexibility comes at the cost of a sometimes significant lack of interpretation.

In practice, one important issue is to determine the best model. The answer to this question depends on the underlying problem. If the relationship between the variables is approximated by a linear model, a correctly specified parametric model should perform well. By contrast, if the parametric model is not correctly specified, since the relationship is highly nonlinear and/or involves significant cross effects, then the statistical methods derived from machine learning should perform better.

The correct specification of a regression model is a common hypothesis, but one that is seldom verified and justified. In the following applications, we show how statistical methods derived from machine learning can be used to justify the correct specification of a parametric regression model or to detect a misspecification.

Sales of Child Car Seats (Classification)

Here, we will be drawing on an example used in James *et al.* (2013). The dataset contains the sales of child car seats at 400 stores (*sales*), as well as several variables, including the quality of the shelving location (*shelveloc*, equal to "poor", "average" and "good") and price (*price*).[7] A binary dependent variable is artificially created to describe high or low sales (*high* = "yes" if $sales > 8$ and "no" if not). In this application, the aim is to identify the determinants of a good volume of sales. We begin by considering a latent linear regression model:

$$y^\star = \gamma + x^T\beta + \varepsilon, \quad \varepsilon \sim G(0,1), \tag{4}$$

where x is composed of k explanatory variables, β is a vector of k unknown parameters and ε is an *i.i.d.* error term with a distribution function G with zero expectation and unit variance. The dependent variable y^* is not observed, with only y, with:

$$y = \begin{cases} 1 & \text{si}\, y^\star > \xi \\ 0 & \text{si}\, y^\star \leq \xi \end{cases} \tag{5}$$

The probability of y being equal to 1 may then be expressed as follows:

$$\mathbb{P}\left(Y = 1\right) = G\left(\beta_0 + x^T\beta\right) \tag{6}$$

where $\beta_0 = \gamma - \xi$.[8] This model is estimated by maximum likelihood by selecting a parametric distribution G. If it is assumed that G is the normal distribution, it is a probit model; if it is assumed that G is the logistic distribution, it is a logit model. In a logit/probit model, there are two possible sources of misspecification:

- The linear relationship $\beta_0 + x^T\beta$ is misspecified;

- The parametric distribution used G is incorrect.

In the event of misspecification, of whatever kind, the estimation is no longer valid. The most flexible model is the following:

$$\mathbb{P}\left[Y = 1 | X = x\right] = G\left(h\left(x\right)\right) \tag{7}$$

where h is an unknown function and G an unknown distribution function. The bagging, random forest and boosting methods can be

[6] *See, among others, Hastie* et al. *(2009) and James* et al. *(2013).*

[7] *It is the Carseats dataset from the ISLR library.*

[8] $\mathbb{P}[Y=1]=\mathbb{P}[Y^\star > \xi]=\mathbb{P}[\gamma + x^T\beta + \varepsilon > \xi]=\mathbb{P}[\varepsilon > \xi - \gamma - x^T\beta]$ *which can ultimately be written as* $\mathbb{P}[\varepsilon < \gamma - \xi + x^T\beta]$*. Given* $\gamma - \xi = \beta_0$*, we obtain* $\mathbb{P}[Y=1] = G(\beta_0 + x^T\beta)$*. In general, it is assumed that the variance of the error term is equal to* σ^2*, in which case the parameters of model (6) are* β_0 / σ *and* β / σ*, which means that the parameters of latent model (4) are not identifiable and are estimated to within one scale parameter.*

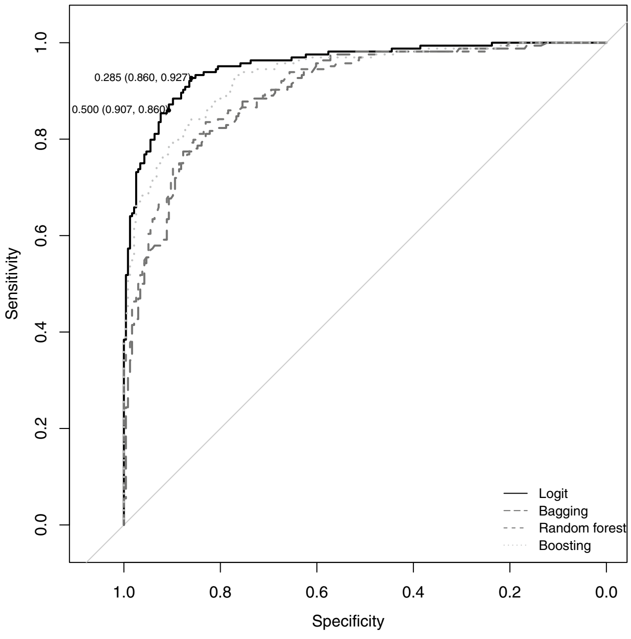
used to estimate this general model without making a preliminary choice about the function *h* and the distribution *G*. The estimation of the logit/probit model nevertheless performs better if *h* and *G* are correctly specified.

Model (6) is estimated using the logistic distribution for *G*, while model (7) is estimated with the bagging, random forest and boosting methods. A 10-fold cross-validation analysis is performed (Box 3). The individual probabilities of the out-of-sample data, i.e. of each of the folds not used for the estimation, are used to assess the quality of the classification.

Figure I shows the ROC curve and the area under the curve (AUC) for the logit, bagging, random forest and boosting estimations. The ROC curve is a graph that simultaneously represents the quality of the prediction in the two classes, for different values of the threshold used to classify the individuals (the term is "cutoff"). One obvious way of classifying individuals is to assign them to the class for which they have the highest estimated probability. In the case of a binary variable, this amounts to predicting the class for which the estimated probability is higher than 0.5. However, a different threshold could be used. For example, in Figure I, a point on the ROC curve of the logit model indicates that by using a threshold of 0.5, the correct prediction rate for the answer "no" is 90.7% (specificity), while the correct prediction rate for the answer "yes" is 86% (sensitivity). Another point indicates that by using 0.285, the correct prediction rate for the answer "no" is 86% (specificity), while the correct prediction rate for the answer "yes" is 92.7% (sensitivity). As described above, an ideal classification model would have an ROC curve of the form Γ. The best model is the model whose curve is above the others. One criterion commonly used to select the best model is the criterion with the largest area under the ROC curve (AUC). The advantage of such a criterion is that it is easy to compare and does not depend on the choice of classification threshold.

In our example, the ROC curve of the logit model is above the other curves and has the largest area under the curve (AUC = 0.9544). These results indicate that this model provides the best classification predictions. Since it no other model performs better, this finding suggests that the linear logit model is correctly specified and that there is no need to use a more general and more complex model.

Figure I
Sales of Car Seats: ROC Curves and Areas Under the Curve (AUC)



AUC	Logit	Bagging	Random Forest	Boosting
	0.9544	0.8973	0.9050	0.9313

Sources: Simulated data on 400 points of sale of baby car seats with the data set "Carseats" from James *et al.* (2013), https://CRAN.R-project.org/package=ISLR

Box 3 – **K-Fold Cross Validation**

Cross-validation is based on the idea of building an estimator by removing an observation. Since the aim is to build a predictive model, the prediction obtained from the estimated model will be compared with the missing observation:

$$\widehat{\mathcal{R}}^{\text{CV}} = \frac{1}{n}\sum_{i=1}^{n}\ell\left(y_i,\widehat{m}_{(i)}\left(x_i\right)\right)$$

The main problem of this method (known as the leave-one-out method) is that it requires calibrating n models, which can be problematic in high dimension. An alternative method is k-fold cross-validation, which involves using a partition of $\{1,\cdots,n\}$ in k groups (or folds) of the same size, $\mathcal{I}_1,\cdots,\mathcal{I}_k$ (let $\mathcal{I}_{\bar{j}} = \{1,\cdots,n\}\setminus\mathcal{I}_j$). With $\widehat{m}_{(j)}$ built on the sample $\mathcal{I}_{\bar{j}}$, we then posit:

$$\widehat{\mathcal{R}}^{k-\text{CV}} = \frac{1}{k}\sum_{j=1}^{k}\mathcal{R}_j \text{ or } \mathcal{R}_j = \frac{k}{n}\sum_{i\in\mathcal{I}_j}\ell\left(y_i,\widehat{m}_{(j)}\left(x_i\right)\right)$$

Using $k = 5,10$ presents two advantages compared to $k = n$ (corresponding to the leave-one-out method): (1) the number of estimations to be performed is far too low, i.e. 5 or 10 rather than n; (2) the samples used for the estimation are less similar and, therefore, less correlated with each other, which tends to avoid excessive variance (James *et al.*, 2013).

Purchase of Caravan Insurance (Classification)

Here, we will be drawing on an example used in James *et al.* (2013). The dataset contains 85 variables on the demographic characteristics of 5,822 individuals.[9] The dependent variable (*purchase*) indicates whether the individual has purchased caravan insurance; it is a binary variable, corresponding to "yes" or "no". In the dataset, only 6% of the individuals took out such insurance. The classes are therefore highly imbalanced.

Model (6) is estimated using the logistic distribution function, while model (7) is estimated by the bagging, random forest and boosting methods (the tuning parameters are those used by James *et al.* (2013), n.trees = 1,000 and shrinkage = 0.01). A 10-fold cross-validation analysis is performed. The individual probabilities of the out-of-sample data, i.e. of each of the pieces not used for the estimation, are used to assess the quality of the classification.

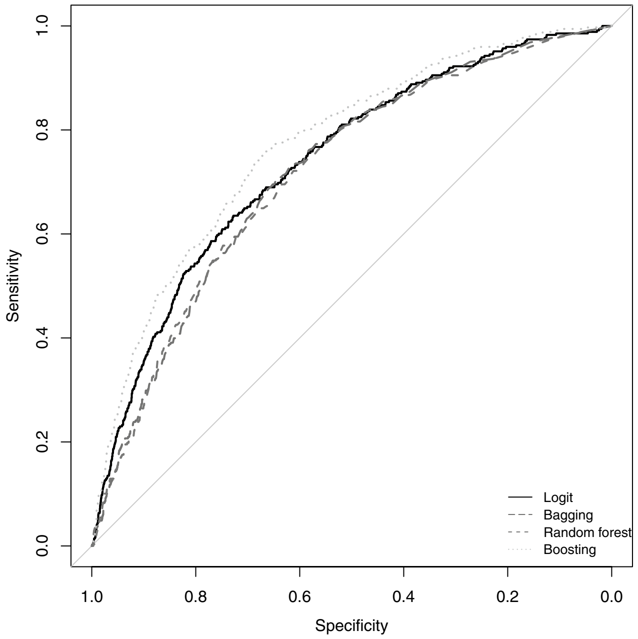
Figure II shows the ROC curve and the area under the curve (AUC) for the logit, bagging, random forest and boosting estimations. The curve of the boosting model is above the other curves and has the largest area under the curve (AUC = 0.7691). These results indicate that boosting provides the best classification predictions. Compared to the previous example, the curves are relatively far from the L shape, which suggests that the classification will not be as good.

It is important to consider the results of a standard classification, i.e. with a classification threshold of 0.5, which is often used by default in software (the prediction of the answer of individual i is "no" if the estimated probability of the individual answering "no" is higher than 0.5; if not, it is "yes"). The left side of Table 2 shows the correct classifications with this threshold (threshold of 0.5) for the different methods. With the best model and the standard threshold (boosting and threshold of 0.5), the "no" answers are 99.87% correct while the "yes" answers are all wrong. This equates to using a model which predicts that no one buys caravan insurance. For analysts, choosing such a model is absurd since their main focus is precisely the 6% of individuals who purchased such insurance. This result is explained by the presence of highly imbalanced classes. Indeed, by predicting that no one buys insurance, the error rate is "only" 6%. However, these are errors which result in not explaining anything.

Several methods can be used to overcome this problem, linked to highly imbalanced classes (Kuhn & Johnson, 2013, Chapter 16). One simple solution is to use a different classification threshold. The ROC curve presents the results according to several classification thresholds, where the perfect classification is illustrated by the couple (specificity, sensitivity) = (1,1), i.e. by the upper-left corner of the graph. The classification threshold corresponding to the point on the ROC curve closest to this corner is selected as the optimal classification threshold. The right side of Table 2 shows the correct classification rates with the optimal thresholds for the different methods (the optimal thresholds

[9] *It is the Caravan dataset from the ISLR library under R.*

Figure II
Purchase of Insurance: ROC Curves and Areas Under the Curve (AUC)



AUC	Logit	Bagging	Random Forest	Boosting
	0.7372	0.7198	0.7154	0.7691

Sources: Experimental dataset "Caravan" on the consumption of caravan insurance, James *et al.* (2013). https://CRAN.R-project.org/package=ISLR

of the logit, bagging, random forest and boosting methods are 0.0655, 0.0365, 0.0395 and 0.0596, respectively). With boosting and an optimal threshold, the "no" answers are 68.6% correct, while the "yes" answers are 73.85% correct. The aim of the analysis being to correctly predict the individuals likely to buy caravan insurance ("yes" class) and to distinguish them sufficiently from the others ("no" class), the optimal threshold performs far better than the standard threshold (0.5). With a logit model and an optimal threshold, the correct classification rate for the "no" class is 72.78%, while the rate for the "yes" class is 63.51%. Compared to boosting, the logit model is slightly better at predicting the "no" class, but is significantly worse at predicting the "yes" class.

Personal Loan Defaults (Classification)

Consider the German database of personal loans, used in Nisbet *et al*. (2001) and Tufféry (2001), with 1,000 observations and 19 explanatory variables, including 12 qualitative variables: by disjuncting them (by creating an indicator variable for each modality), we obtain 48 potential explanatory variables. A recurring question in modelling is to determine which variables merit being used. The

Table 2
Purchase of Insurance: Sensitivity to the Choice of Classification Threshold

	Threshold of 0.5		Optimal Thresholds	
	Specificity	Sensitivity	Specificity	Sensitivity
Logit	0.9967	0.0057	0.7278	0.6351
Bagging	0.9779	0.0661	0.6443	0.7069
Random Forest	0.9892	0.0316	0.6345	0.6954
Boosting	0.9987	0.0000	0.6860	0.7385

Sources: Experimental dataset "Caravan" on the consumption of caravan insurance, James *et al.* (2013). https://CRAN.R-project.org/package=ISLR most obvious solution for an econometrician may be a stepwise method (with running through all possible combinations of variables being, on the face of it, too complex in high, forward or backward dimension). The set of variables in a backward approach is shown in the first column of Table 3 (see Box 4 for the principles governing penalisation and the choice of explanatory variables). The table provides a comparison with two other approaches: first, the lasso method, by suitably penalising the norm ℓ_1 of the vector of parameters β (last column). We note that the first two variables considered as null (for a sufficiently large λ) are the first two to emerge from a backward procedure. One last method has been proposed by Breiman (2001b), using all of the trees created when building a random tree: the importance of the variable x_k in a forest of T trees is given by:

$$Importance\left(x_k\right)=\frac{1}{T}\sum_{t=1}^{n}\sum_{j\in N_{t,k}} p_t\left(j\right)\Delta\mathcal{I}\left(j\right)$$

where $N_{t,k}$ denotes the set of nodes of the tree t using the variable x_k as a separation variable, $p_t\left(j\right)$ denotes the proportion of observations in a node j, and $\Delta\left(j\right)$ is the index variation at the node j (between the preceding node, the left leaf and the right leaf). The central column of Table 3 shows the variables by decreasing order of importance when the index used is the Gini impurity index.

With the stepwise approach and the lasso method, we remain with linear logistic models. In the case of random forests (and trees), interactions between variables can be taken into account when 2 variables are present. For example, the variable *residence_since* ranks very high among the predictive variables (third most important variable).

Wage Determinants (Regression)

The Mincer wage equation (Mincer, 1974; Lemieux, 2006) has traditionally been used

Table 3
Credit: Choice of Variables, Sequential Sorting, Based on a Stepwise Approach, by Importance Function in a Random Forest and by Lasso

Stepwise	AIC	Random Forest	Gini	Lasso
checking_statusA14	1112.1730	checking_statusA14	30.818197	checking_statusA14
credit_amount(4e+03,Inf]	1090.3467	installment_rate	20.786313	credit_amount(4e+03,Inf]
credit_historyA34	1071.8062	residence_since	19.853029	credit_historyA34
installment_rate	1056.3428	duration(15,36]	11.377471	duration(36,Inf]
purposeA41	1044.1580	credit_historyA34	10.966407	credit_historyA31
savingsA65	1033.7521	credit_amount	10.964186	savingsA65
purposeA43	1023.4673	existing_credits	10.482961	housingA152
housingA152	1015.3619	other_payment_plansA143	10.469886	duration(15,36]
other_payment_plansA143	1008.8532	telephoneA192	10.217750	purposeA41
personal_statusA93	1001.6574	Age	10.071736	installment_rate
savingsA64	996.0108	savingsA65	9.547362	property_magnitudeA124
other_partiesA103	991.0377	checking_statusA12	9.502445	age(25,Inf]
checking_statusA13	985.9720	housingA152	8.757095	checking_statusA13
checking_statusA12	982.9530	jobA173	8.734460	purposeA43
employmentA74	980.2228	personal_statusA93	8.715932	other_partiesA103
age(25,Inf]	977.9145	property_magnitudeA123	8.634527	employmentA72
purposeA42	975.2365	personal_statusA92	8.438480	savingsA64
duration(15,36]	972.5094	purposeA43	8.362432	employmentA74
duration(36,Inf]	966.7004	employmentA73	8.225416	purposeA46
purposeA49	965.1470	employmentA75	8.089682	personal_statusA93
purposeA410	963.2713	duration(36,Inf]	8.029945	personal_statusA92
credit_historyA31	962.1370	purposeA42	8.025749	savingsA63
purposeA48	961.1567	property_magnitudeA122	7.908813	telephoneA192

Sources: Dataset “Credit” of the casdataset library of R, loans to households in Germany (Nisbet *et al.*, 2001; Tufféry, 2001). http://cas.uqam.ca/

Box 4 – **Penalisation and Methods for the Choice of Explanatory Variables**

To select relevant explanatory variables in econometrics, we may use criteria *ex post* relating to the quality of the model penalising the complexity, in practice the number of explanatory variables (such as R^2 adjusted or the Akaike criterion – AIC – see the online complement). In the forward method, we start with a regression on the constant before adding one variable at a time, retaining the variable that most improves the model according to the chosen criterion, until adding a variable reduces the quality of the model. In the backward method, we start with a regression on all the variables before adding one variable at a time, removing the variable that most improves the quality of the model, until removing a variable reduces the quality of the model. Stepwise methods introduce ensemble methods to limit the number of tests.

The machine learning strategy involves penalising *ex-ante* in the objective function, even at the risk of constructing a biased estimator. Typically, the following is built:

$$\left(\hat{\beta}_{0,\lambda},\hat{\beta}_{\lambda}\right)=\text{argmin}\left\{\sum_{i=1}^{n}\ell\left(y_i,\beta_0+x^T\beta\right)+\lambda\,\textit{pénalisation}\left(\beta\right)\right\} \quad (8)$$

where the penalisation function will often be a norm $\|\cdot\|$ chosen *a priori*, and a penalisation parameter λ.

to explain (individual) wages according to the individual's education, experience and gender:

$$\log\left(\text{wage}\right)=\beta_0+\beta_1\,ed+\beta_2\,exp+\beta_3\,exp^2+\beta_4\,fe+\varepsilon \quad (9)$$

where *ed* is the level of education, *exp* is the level of professional experience and *fe* is a dummy variable equal to 1 if the individual is a woman. According to human capital theory, the expected wage increases with experience, at an increasingly slow rate, until it reaches a threshold before decreasing. The introduction of the square of *exp* enables such a relationship to be taken into account. The presence of variable *fe* allows for any wage gap between men and women to be measured.

Model (9) establishes a linear relationship between wage and level of education and a quadratic relationship between wage and professional experience. These relationships may seem too restrictive. Several studies have shown, in particular, that wages do not fall after a certain age and that a quadratic relationship or a higher-degree polynomial is more appropriate (Murphy & Welch, 1990; Bazen & Charni, 2017).

Model (9) also establishes that the wage gap between men and women is independent of the level of education and experience. It is too restrictive if, for example, the average wage gap between men and women is low for unskilled jobs and high for skilled jobs, or low among early-career workers and high among late career workers (interaction effects). The most flexible model is the fully nonparametric model:

$$\log\left(wage\right)=m\left(ed,exp,fe\right)+\varepsilon \quad (10)$$

where $m(\cdot)$ is a random function. It has the advantage of being able to take into account any nonlinear relationships and complex interactions between the variables. However, its significant flexibility is at the cost of a more difficult interpretation of the model. Indeed, a 4-dimensional graph would be needed to represent the function m. One solution is to represent the function m in 3 dimensions by fixing the value of one of the variables, although the represented function may differ significantly with a different fixed value.

We will use data from a survey by the US Census Bureau carried out in May 1985 drawn from Berndt (1990) and available under R.[10] The two models are estimated and a 10-fold cross-validation analysis is used to select the best approach. Parametric model (9) is estimated by ordinary least squares (OLS). Fully nonparametric model (10) is estimated by the method of splines since it includes few variables and also by the bagging, random forest and boosting methods.

The results of the 10-fold cross-validation are presented in Table 4. The best model is the model that minimises the criterion $\widehat{\mathcal{R}}^{10-CV}$. The results show that model (9) is at least as effective as model (10), which suggests that parametric model (9) is correctly specified.

Determinants of House Prices in Boston (Regression)

Here, we will be drawing on one of the examples used in James *et al.* (2013), whose data

[10] *It is the CPS1985 dataset from the AER library.*

are available under R. The dataset contains the median values of house prices (*medv*) in $n = 506$ neighbourhoods around Boston along with 13 other variables, including the average number of rooms per house (*rm*), the average age of houses (*age*) and the percentage of households with a low economic status (*lstat*).[11]

Consider the following linear regression model:

$$medv = \alpha + x^T \beta + \varepsilon \qquad (11)$$

where x = [chas,nox,age,tax,indus,rad,dis,lstat, crim,black,rm,zn,ptratio] is a vector in dimension 13 and β is a vector of 13 parameters. This model specifies a linear relationship between the value of houses and each of the explanatory variables. The most flexible model is the fully nonparametric model:

$$medv = m(x) + \varepsilon. \qquad (12)$$

The estimation of this model by the Kernel method or the method of splines can be problematic since the number of variables is relatively high (there are 13 variables here) or, at least, too high to consider estimating a surface in dimension 13. We estimate the two models and use a 10-fold cross-validation analysis to select the best approach. Parametric model (11) is estimated by ordinary least squares (OLS) and fully nonparametric model (12) is estimated using three different methods: bagging, random forest and boosting (here we use the default values used in James *et al.*, 2013, pp. 328–331).

Table 5 shows the results of the 10-fold cross-validation. Based on the in-sample results (on the learning data), the bagging and random forest methods are found to be vastly more effective than the OLS estimation of linear model (11), the criterion $\widehat{\mathcal{R}}^{10-CV}$ going from 21.782 to 1.867 and 1.849. The out-of-sample results (on data other than those used to estimate the model) tend in the same direction, although the difference is less significant, with the criterion $\widehat{\mathcal{R}}^{10-CV}$ going from 24.082 to 9.59 and 9.407. These results illustrate a common phenomenon with nonlinear methods such as bagging and random forest, which can be highly effective in predicting the data used in the estimation, but less effective at predicting out-of-sample data. This explains why the selection of the best estimation is typically based on an out-of-sample analysis.

The difference between the estimation of models (11) and (12) is significant. Such a difference suggests that the linear model is misspecified and that nonlinear relationships

[11] It is the Boston dataset from the MASS library. For a complete description of the data, see: https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Boston.html.

Table 4
Wage: Fold Cross Validation Analysis (*K* = 10): Performance of the Estimation of Linear Model (9) and Fully Nonparametric Model (10)

$\widehat{\mathcal{R}}^{10-CV}$	Model (9)	Model (10)			
	OLS	Splines	Bagging	Random forests	Boosting
Out-of-sample	0.2006	0.2004	0.2762	0.2160	0.2173

Source: Population census, USA, 1985, Berndt (1990). Dataset CPS1985 from AER Library. https://rdrr.io/cran/AER/man/CPS1985.html

Table 5
House Prices in Boston - Fold Cross Validation Analysis (*K* = 10): Performance of the Estimation of Linear Model (11) and Fully Nonparametric Model (12)

$\widehat{\mathcal{R}}^{10-CV}$	Model (11)	Model (12)		
	OLS	Splines	Random forests	Boosting
In-sample	21.782	1.867	1.849	7.012
Out-of-sample	24.082	9.590	9.407	11.789

Coverage: Districts of the Boston metropolitan area.
Sources: James *et al.* (2013), Boston data set from the MASS library. https://stat.ethz.ch/R-manual/Rdevel/library/MASS/html/Boston.html

and/or interaction effects are present in the relationship between house prices and the explanatory variables x. This requires looking for a better parametric specification. Based on parametric model (11), and in order to take into account any potential nonlinearities, the following generalised additive model (GAM) may be considered:

$$medv = m_1(x_1) + m_2(x_2) + \ldots + m_{13}(x_{13}) + \varepsilon \quad (13)$$

where $m_1, m_2, \ldots m_{13}$ are unknown functions. The advantage of this model is that it allows for any nonlinear relationship between the dependent variable and each of the explanatory variables to be considered. Furthermore, it does not suffer from the curse of dimensionality problem since each of the functions is of dimension 1 and it is easily interpretable. However, it does not take into account any potential interaction effects. The estimation of generalised additive model (13) by the method of splines, as part of a 10-fold cross-validation analysis, gives value $\widehat{\mathcal{R}}^{10-CV} = 13.643$. Compared to parametric model (11), there is a significant gain (13.643 vs. 24.082). However, the difference with the fully nonparametric model (12) remains substantial (13.643 vs 9.590, 9.407, 11.789). Such a difference suggests that taking into account individual relationships which may be highly nonlinear is not sufficient and that interaction effects between variables are present. The simplest interaction variables among the pairs of variables $(x_i \times x_j)$ could be included in the model, but that would imply adding a significant number of variables to the original model (78 in this case), which would have an impact on the quality of the estimation of the model. In any case, as things stand, what can be said is that the linear model is misspecified and that there are potentially significant interaction effects in the relationship between *medv* and X, the identification of such effects remaining a delicate matter.

To go further, the tools developed in statistical learning may be of great use. For example, the random forest estimation technique involves measures of the significance of each of the variables in the estimation of the model. Table 6 shows these measurements in relation to model (12), estimated on the whole sample. The results suggest that the variables *rm* and *lstat* are the most significant variables to explain house price variations *medv*. This finding suggests enriching the initial relationship by adding the interaction effects linked to these two variables only, which are the most significant.

The generalised additive model including the interaction variables is estimated on the whole sample:

$$medv = m_1(x_1) + \cdots + m_{13}(x_{13}) + (rm : x)\gamma$$
$$+ (lstat : x)\delta + \varepsilon \quad (14)$$

where $(rm : x)$ represents the interaction variables of *rm* with all the other variables of x and $(lstat : x)$ represents the lstat interaction variables with all the other variables of x.[12] Analysis of the results of this estimation suggests that functions $\widehat{m_i}$ are linear for all variables except for the DIS variable, whose estimated relationship is shown in Figure III. This variable measures the average distance from five employment centres within the region. The effect appears to decrease more rapidly with distance when the latter is not very significant. Beyond a certain distance (beyond 2, in log), the effect is reduced and continues to decrease, albeit at a slower rate. This nonlinear relationship can be approximated by a piecewise linear regression by considering a node.

Table 6
House Prices: Measures of the Importance of Each Variable in the Random Forest Estimation of Model (12), by Considering the Whole Sample

	% IncMSE	IncNodePurity
rm	61.35	18 345.41
istat	36.20	15 618.22
dis	29.37	2601.72
nox	24.91	1034.71
age	17.86	554.50
ptratio	17.43	626.58
tax	16.60	611.37
crim	16.26	1701.73
indus	9.45	237.35
black	8.72	457.58
rad	4.53	166.72
zn	3.10	35.73
chas	0.87	39.05

Coverage: Districts of the Boston metropolitan area.
Sources: James *et al.* (2013), Boston data set from the MASS library. https://stat.ethz.ch/R-manual/Rdevel/library/MASS/html/Boston.html

[12] We have *(rm:x)=[rm × chas, rm × nox, rm × age, rm × tax, rm × indus, rm × rad, rm × dis, rm × lstat, rm × crim, rm × black, rm × zn, rm × ptratio]* and *(lstat:x)=[lstat × chas, lstat × nox, lstat × age, lstat × tax, lstat × indus, lstat × rad, lstat × dis, lstat × crim, lstat × black, lstat × zn, lstat × ptratio].*

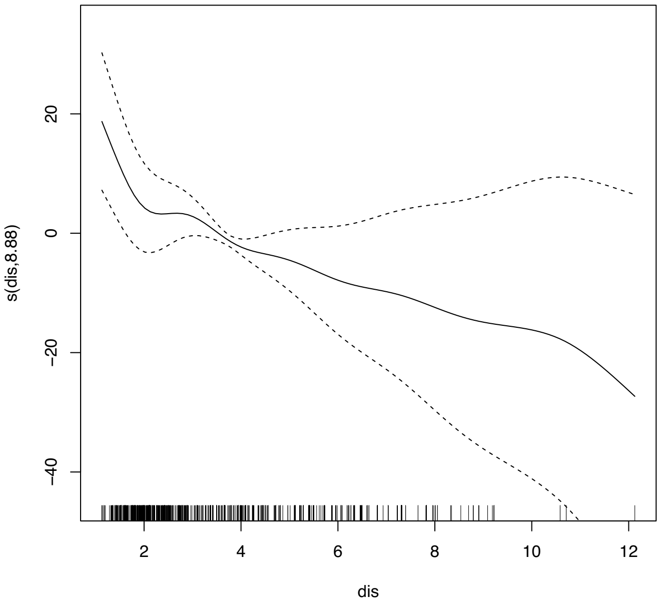
Lastly, the above analysis suggests considering the following linear model:

$$medv = \alpha + x^T \beta + (dis - 2) + \theta + (rm : x)\gamma \\ + (lstat : x)\delta + \varepsilon \qquad (15)$$

where $(dis - 2)$ is equal to the value of its argument if the latter is positive, and to 0 if it is not. Compared to the original linear model, this model includes a piecewise linear relationship with the DIS variable, as well as interaction effects between *rm*, *lstat* and each of the other variables of x.

Table 7 shows the results of the 10-fold cross validation of the estimation of parametric models (11) and (15), estimated by ordinary least squares (OLS), and of the generalised additive model (14) estimated by splines. It shows that the addition of interaction variables and of the piecewise linear relationship in model (15) produces far better results than the initial model (11): the criterion $\widehat{\mathcal{R}}^{10-CV}$ is divided by more than two, going from 24.082 to 11.759. By comparing these results with the results of Table 5, we also find that parametric model (15), estimated by OLS, is as effective as general model (12) estimated by boosting ($\widehat{\mathcal{R}}^{10-CV} = 11.789$). The difference with the bagging and random forest methods is not very significant ($\widehat{\mathcal{R}}^{10-CV} = 9.59, 9.407$). Lastly, the bagging, random forest and boosting methods served to highlight the misspecification of the original parametric model and then to find a far more effective parametric model by taking into

Figure III
Estimation of the Relationship m_7 (x_7) in the Generalised Additive Model (14), where x_7 = dis.



Note: Estimation of the m7 (x7) relationship for the dis variable in the generalized additive model; the dotted lines correspond to the 95% confidence intervals.
Coverage: Districts of the Boston metropolitan area.
Sources: James *et al.* (2013), Boston data set from the MASS library. https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Boston.html

Table 7
House Prices in Boston - Fold Cross Validation Analysis (*K* = 10): Performance of the Estimation of Linear Models (11) and (15) and of Model (14) Including the Interaction Effects and With a Piecewise Nonlinearity

$\widehat{\mathcal{R}}^{10-CV}$	Model (11)	Model (14)	Model (15)
	OLS	Splines	OLS
Out-of-sample	24.082	13.643	11.759

Coverage: Districts of the Boston metropolitan area.
Sources: James *et al.* (2013), Boston data set from the MASS library. https://stat.ethz.ch/R-manual/Rdevel/library/MASS/html/Boston.html

account the effects of appropriate nonlinearities and interactions.

* *
*

While the two cultures (or two communities) of econometrics and machine learning have developed in parallel, the number of links between the two is constantly increasing. Whereas Varian (2014) outlined the significant contributions of econometrics to the machine learning community, our aim here was to present concepts and tools developed over time by that very community and which may be of use to econometricians, in a context of ever increasing data volumes. The probabilistic foundations of econometrics are without doubt its key asset, allowing not only for model interpretability, but also for the quantification of uncertainty. Nevertheless, the predictive performance of machine learning models is of value insofar as they allow for the identification of a misspecified econometric model. In the same way that nonparametric techniques provide a point of reference for assessing the relevance of a parametric model, machine learning tools help to improve an econometric model by detecting a nonlinear effect or an overlooked cross effect.

An illustration of the potential interactions between the two communities can be found, for example, in Belloni *et al.* (2010, 2012), in the context of the choice of instrument in a regression. Using the data produced by Angrist & Krueger (1991) relating to an academic achievement problem, they show how to effectively implement instrumental econometric techniques when 1,530 instruments are available (a recurring problem with the increase in the volume of data). As we have seen throughout this paper, although the approaches adopted may differ fundamentally in the two communities, econometricians have much to gain from using many of the tools developed by the machine learning community. □

Link to the Online complements: https://www.insee.fr/en/statistiques/fichier/3706234?sommaire=3706269/505-506_Charpentier-Flachaire-Ly_complement_EN.pdf

BIBLIOGRAPHY

Aldrich, J. (2010). The Econometricians' Statisticians, 1895-1945. *History of Political Economy*, 42(1), 111–154.
https://doi.org/10.1215/00182702-2009-064

Altman, E., Marco, G. & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*, 18(3), 505–529.
https://doi.org/10.1016/0378-4266(94)90007-8

Angrist, J. D. & Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, 106(4), 979–1014.
https://doi.org/10.2307/2937954

Bazen, S. & Charni, K. (2017). Do earnings really decline for older workers? *International Journal of Manpower*, 38(1), 4–24.
https://doi.org/10.1108/IJM-02-2016-0043

Bellman, R. E. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Belloni,A.,Chernozhukov,V.& Hansen,C.(2010). Inference for High-Dimensional Sparse Econometric Models. *Advances in Economics and Econometrics*, *10th World Congress of Econometric Society*, 245–295
https://doi.org/10.1017/CBO9781139060035.008

Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica*, 80(6), 2369–2429.
https://doi.org/10.3982/ECTA9626

Blanco, A. Pino-Mejias, M., Lara, J. & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications*, 40(1), 356–364.
https://doi.org/10.1016/j.eswa.2012.07.051

Breiman, L. Fiedman, J., Olshen, R. A. & Stone, C. J. (1984). *Classification And Regression Trees*. Chapman & Hall/CRC Press Online.
https://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf

Breiman, L. (2001a). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3), 199–231.
https://doi.org/10.1214/ss/1009213726

Breiman, L. (2001b). Random forests. *Machine learning*, 45(1), 5–32.
https://doi.org/10.1023/A:1010933404324

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High Dimensional Data: Methods, Theory and Applications*. Berlin: Springer Verlag.
https://doi.org/10.1007/978-3-642-20192-9

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273–297.
https://doi.org/10.1023/A:1022627411411

Grandvalet, Y., Mariéthoz, J., & Bengio, S. (2005). Interpretation of SVMs with an Application to Unbalanced Classification. *Advances in Neural Information Processing Systems* N° 18.
https://papers.nips.cc/paper/2763-a-probabilistic-interpretation-of-svms-with-an-application-to-unbalanced-classification.pdf

Groves, T. & Rothenberg, T. (1969). A note on the expected value of an inverse matrix. *Biometrika*, 56(3), 690–691.
https://doi.org/10.1093/biomet/56.3.690

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer Verlag.
https://doi.org/10.1007/978-0-387-84858-7

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.
https://doi.org/10.1002/1097-4679(195007)6:3<307::AID-JCLP2270060338>3.0.CO;2-K

James, G., D. Witten, T. Hastie, & Tibshirani, R. (2013). An Introduction to Statistical Learning. *Springer Texts in Statistics* 103.
https://doi.org/10.1007/978-1-4614-7138-7

Khashman, A. (2011). Credit risk evaluation using neural networks: Emotional versus conventional models. *Applied Soft Computing*, 11(8), 5477–5484.
https://doi.org/10.1016/j.asoc.2011.05.011

Kolda, T. G. & Bader, B. W. (2009). Tensor Decompositions and Applications. *SIAM Review*, 51(3), 455–500.
https://doi.org/10.1137/07070111X

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer Verlag.
https://doi.org/10.1007/978-1-4614-6849-3

Landis, J. R. & Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174.
https://doi.org/10.2307/2529310

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, **521,** 436–444.
https://doi.org/10.1038/nature14539

Leeb, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli*, 14(3), 661–690.
https://doi.org/10.3150/08-BEJ127

Lemieux, T. (2006). The "Mincer Equation" Thirty Years After Schooling, Experience, and Earnings. In: Grossbard, S. (Ed.), *Jacob Mincer: A Pioneer of Modern Labor Economics*, pp. 127–145. Boston, MA: Springer Verlag.
https://doi.org/10.1007/0-387-29175-X_11

Lin, H. W., Tegmark, M. & Rolnick, D. (2016). Why does deep and cheap learning work so well?
https://arxiv.org/abs/1608.08225

Mincer, J. (1974). Schooling, Experience and Earnings. New York: NBER.
https://www.nber.org/books/minc74-1

Morgan, J. N. & Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58(302), 415–434.
https://doi.org/10.1080/01621459.1963.10500855

Morgan, M. S. (1990). *The history of econometric ideas*. Cambridge, UK: Cambridge University Press.

Murphy, K. M. & Welch, F. (1990). Empirical Age-Earnings Profiles. *Journal of Labor Economics*, 8(2), 202–229.
https://doi.org/10.1086/298220

Nisbet, R., Elder, J. & Miner, G. (2011). *Handbook of Statistical Analysis and Data Mining Applications*. New York: Academic Press.

Portnoy, S. (1988). Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity. *Annals of Statistics*, 16(1), 356–366.
https://doi.org/10.1214/aos/1176350710

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.
https://doi.org/10.1007/BF00116251

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
https://doi.org/10.1037/h0042519

Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229.
https://doi.org/10.1147/rd.33.0210

Shalev-Shwartz, S. & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, UK: Cambridge University Press.

Shapire, R. E. & Freund, Y. (2012). *Boosting. Fondations and Algorithms*. Cambridge, A MIT Press.

Tam, K. Y. & Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, 38, 926–947.
https://doi.org/10.1287/mnsc.38.7.926

Tufféry, S. (2001). *Data Mining and Statistics for Decision Making*. Hoboken, NJ: Wiley.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.
https://doi.org/10.1257/jep.28.2.3

Vert, J. P. (2017). Machine learning in computational biology. Cours à l'Ensae ParisTech.
http://members.cbio.mines-paristech.fr/~jvert/teaching/

Widrow, B. & Hoff, M. E. Jr. (1960). Adaptive Switching Circuits. *IRE WESCON Convention Record*, 4, 96–104.
https://apps.dtic.mil/dtic/tr/fulltext/u2/241531.pdf

Zinkevich M. A., Weimer, M., Smola, A. & Li, L. (2010). Parallelized Stochastic Gradient Descent. *Advances in neural information processing systems* 23, 2595–2603.
https://papers.nips.cc/paper/4006-parallelized-stochastic-gradient-descent.pdf

Citizen Data and Trust in Official Statistics

Evelyn Ruppert*, Francisca Grommé,* Funda Ustek-Spilda**
and Baki Cakici***

Abstract – From smartphones, meters, fridges and cars to internet platforms, the data of digital technologies are the data of citizens. In addition to raising political and ethical issues of privacy, confidentiality and data protection, this calls for rethinking relations to citizens in the production of data for statistics if they are to be trusted by citizens. We outline an approach that involves co-producing data with citizens as partners of statistical production, from the design of a data production platform to the interpretation and analysis of data. While raising issues such as data quality and reliability, we argue co-production can potentially mitigate problems associated with the re-purposing of Big Data. We argue that in a time of "alternative facts", what constitutes legitimate knowledge and expertise are major political sites of contention and struggle and require going beyond defending existing practices towards inventing new ones. In this context, we contend that the future of official statistics not only depends on inventing new data sources and methods but also mobilising the possibilities of digital technologies to establish new relations with citizens.

JEL Classification: A14, C93, O35, O38
Keywords: citizen science, co-production, experimentalism, privacy-by-design, smart statistics

* *Department of Sociology, Goldsmiths University of London (E.Ruppert@gold.ac.uk; F.Gromme@gold.ac.uk)*
** *Department of Media and Communications, London School of Economics (f.ustek-spilda@lse.ac.uk)*
*** *Technologies in Practice, IT University of Copenhagen (bakc@itu.dk)*

We would like to thank two anonymous reviewers for their comments and suggestions. We are also grateful for the support and involvement of numerous national and international statisticians who made this research possible at the UK Office for National Statistics, Statistics Netherlands, Statistics Estonia, Turkish Statistical Institute, Statistics Finland, Eurostat and UNECE. Funding from a European Research Council (ERC, Consolidator Grant – Agreement n° 615588) supported the writing of this article. Principal Investigator, Evelyn Ruppert, Goldsmiths, University of London.

To cite this article: Ruppert, E., Grommé, F., Ustek-Spilda, F. & Cakici, B. (2018). Citizen Data and Trust in Official Statistics. *Economie et Statistique / Economics and Statistics*, 505-506, 179–193.
https://doi.org/10.24187/ecostat.2018.505d.1971

National Statistical Institutes (NSI) experiments concerning the potential of Big Data generated by various digital technologies as a new source for the making of official statistics have now been underway for about five years. These have led to the identification of several concerns such as data access, data ownership, privacy and ethics, data representativeness, data quality and so on. Amongst other things, these concerns are understood as potential risks to the reputation and public image of NSIs working with Big Data sources, as identified in a report of the UNECE Big Data Privacy task team (UNECE, 2014). That report summarised a number of strategies to mitigate such risks including the enforcement of ethical principles through instruments of accountability and informed consent; establishing strong compliance controls; developing monitoring systems to track reputational threats; ensuring transparency and understanding through clear communication with stakeholders about the use of data and the organisation of dialogues with the public; and creating a crisis communication plan. The report also argued, as have others produced by international bodies such as Eurostat's Big Data Task Force, that repurposing Big Data sources not only presents technical challenges but potentially could undermine citizens' trust in how NSIs generate data and produce official statistics. Similar challenges are encountered when NSIs seek to repurpose administrative data generated by other government departments, which has introduced not only technical challenges but for some NSIs also raised concerns about how data is shared, joined up and used for purposes other than for what they were originally generated.

Of course, questions of citizen trust in official statistics are not new. While trust is also a concern in relation to other stakeholders including ministries, government agencies, media, universities, and other public or private research organisations that rely on official statistics, it is trust in relation to citizens that concerns us here. The history of established methods of generating social and population statistics, such as census questionnaires, surveys and time diaries demonstrates that elaborate practices have been required to secure citizens' trust in how data is generated and used for official statistics. Through practices such as focus groups, the pilot testing of questions, and consultations with civic organisations about issues of consent, data protection, privacy, impartiality and professional standards, NSIs have sought to secure the trust of citizens (Struijs *et al.*, 2014, p. 2). Understood in this way, trust is not the result of one but myriad practices through which the trustworthiness of official statistics is accomplished.

Big Data, because it is generated not by governments but private corporations such as platform owners, if used for official statistics could undermine these practices and the trust they have relatively well performed. As some statisticians have noted, "[of] critical importance is the implication of any use of Big Data for the public perception of a NSI as this has a direct impact on trust in official statistics" (Struijs *et al.*, 2014, p. 3). While Struijs *et al.* argue that such risks can be mitigated by other practices such as "being transparent about what and how Big Data sources are used", we suggest that while necessary this would be insufficient due to another significant issue: the repurposing of Big Data for official statistics constitutes a break and detachment in the relation between NSIs and citizens. While not without problems, established methods such as those noted above have involved more-or-less direct relations between NSIs and citizens to secure data as a collective accomplishment and social good. These relations enable citizens to be relatively active in their identification such as how they translate their knowledge and experiences into responses to questions and, we suggest, in turn contribute to accomplishing trust in and the legitimacy of official statistics.

This proposition was initially put forward in the "Socialising Big Data" project, which involved collaborative workshops with national and international statisticians and led to a proposal for a social framework for Big Data (Ruppert *et al.*, 2015). The framework posited models of social ownership that stress sharing, collaborative, and co-operative possibilities and that imagine Big Data as a social and collective rather than private resource. The approach that we develop in this article builds on this aspiration to develop the concept of "citizen data" as a form of "re-attachment" and social ownership that establishes new relations with citizens as co-producers of data for official statistics rather than as ever more distant subjects whose impressions and confidence need to be managed.

We contend that this understanding of new relations is critical in two ways. First, unlike some uses of the term that define citizen data as data about citizens, our conception recognises that Big Data and citizens are inseparable: the data of digital technologies is the data *of* citizens. Second, relations that involve more direct engagements with citizens are necessary to address another consequence of detachment when data such as that generated by social media, mobile phones and browsers is repurposed: the risk of a widening gap between citizens' actions, identifications and experiences and how they are categorised, included and excluded in statistics, the interpretation of that data, and citizens' identifications with the resulting statistics.[1] We refer to this risk as a widening gap because these consequences are not entirely new or limited to Big Data.[2] Former Eurostat Director General Walter Radermacher expressed this more generally as a gap between citizen experiences and official statistics which in turn calls for "subjective statistics".[3] In saying so he stressed the need for a more democratic debate between citizens and data producers and owners to achieve a "more subjective, differentiated understanding of our world", instead of "technocrats and politicians sitting together and confronting citizens in the end".[4] For our concept of citizen data this requires processes of co-production that involve direct relations with citizens in the production of data for making official statistics.

Our argument draws on several years of fieldwork conducted at NSIs and international statistical organisations (see Box; see also the working paper by Grommé *et al.*, 2017). This research led to the identification of four principles for citizen data that started from key "matters of concern" statisticians have expressed about the future of official statistics which we encountered in our fieldwork. We consider these as matters of concern for two reasons. First, to recognise them as normativities that influence and guide statisticians' actions and development of practical solutions (Boltanski & Chiapello, 2007). Second, to engage in a form of critique that does not dismiss the concepts of our research subjects but first engages with how they conceive and define concepts to then consider how concepts can be reconceived (Latour, 2004). That is, taking up the concerns statisticians have expressed does not mean to agree with them and their assumptions but to engage with and then reconceive those concerns. The four matters of concern we identified as significant to our concept of citizen data are experimentalism, citizen science, smart statistics and privacy-by-design. In the next part of this article we introduce each concern and then draw on a range of literature in the social sciences to reconceive each and then express them as principles of citizen data. Central to our

[1] *For example, experiments with mobile phone data to model mobility encounter problems when attempting to interpret the meaning of travel patterns.*

[2] *We are aware that issues of representation also affect established statistical methods. GDP, Gross Domestic Product, for instance, is one such highly debated official statistic. Columbia University economist Joseph Stiglitz draws attention to how GDP has come to be "fetishised" as "the" indicator of how well a national economy is doing, despite various shortcomings (Stiglitz* et al.*, 2009). Consequently, Fleurbaey (2009) suggests moving "beyond GDP" and draws attention to other approaches, including recent developments in the analysis of sustainability, happiness and the theory of social choice and fair allocation to the studies of social welfare. Similar arguments have also been raised for employment indicators, especially with respect to people working in non-regular employment arrangements (see Hussmanns, 2004).*

[3] Fieldwork notes*, Eurostat conference "Towards More Agile Social Statistics", Luxembourg, 28-30 November 2016.*

[4] Idem.

Box – **The research project**

Our concept of citizen data comes from several years of ethnographic fieldwork that we conducted at five NSIs and two international statistical organisations, which involved observing conferences and meetings, following and analysing publications, and conducting interviews and engaging in conversations with statisticians. More precisely, this article builds on and summarises key points in an ARITHMUS working paper by Grommé *et al.* (2017). ARITHMUS (Peopling Europe: How data make a people), an ERC funded project, began in 2014 with a team of six researchers: Evelyn Ruppert (Principal Investigator), Baki Cakici, Francisca Grommé, Stephan Scheel, and Funda Ustek-Spilda (Postdoctoral Researchers), and Ville Takala (Doctoral Researcher). We followed working practices at five NSIs (UK Office for National Statistics, Statistics Netherlands, Statistics Estonia, Turkish Statistical Institute, and Statistics Finland) and two international organisations (Eurostat and UNECE). Amongst other things, we followed statisticians' debates about and experiments with digital technologies and big data and their implications for official statistics. Based on this fieldwork we conducted two workshops with a project advisory group of statisticians to discuss some of our analyses such as the changing relations between NSIs and citizens as a consequence of new digital technologies and big data sources. This led to a working paper that summarised some of the arguments outlined in this article and introduced the concept of citizen data, which was reviewed by the advisory group (Grommé *et al.*, 2017). That review led to a collaborative workshop with the advisory group and a broader group of statisticians, academic researchers, information designers and facilitators on the development of design principles for the co-production of an app for citizen data. Rather than summarising empirical material from our ethnography and the workshops, our objective here is to outline the conception of citizen data that we have developed as a result of this research.

re-conception is that the future of official statistics not only depends on working with new digital technologies, data sources and inventing methods, but on establishing new relations to citizens (Ruppert, 2018).

We have intended this discussion of a concept of citizen data principally for statisticians but also for social science researchers for three key reasons. One is that we have brought concepts and understandings advanced in the social sciences to bear on matters of concern expressed by statisticians. In this way, we contribute more generally to social science research methods. Another reason is that the principles and concept of citizen data also apply to debates within the social sciences concerning research methods that engage with digital technologies and Big Data sources. That is, while the issues and objectives of social science research are different, relations to citizens in the production of knowledge are a shared concern. Third, as reflected in our research method which involved workshops with statisticians, a concept of citizen data calls for experimental engagements not only with citizens but also between social scientists and statisticians.

Experimentalism

The first matter of concern that we have come across in our fieldwork is experimentalism. Government agencies and corporations have embraced experimentation as a necessary part of innovation. Official statistics is a good example as attested by the development of innovation laboratories, sandboxes, hackathons and exploratory research projects.[5] For statisticians, experiments with new digital technologies and Big Data are methods to develop new ways of thinking, techniques, and skills in the production of official statistics. There are also various strands within the social sciences that engage with experimentalism. Relatively new, however, is the adoption of experimenting as a method to open scientific and technological expertise to different actors to generate new ways of thinking. In areas as diverse as wheelchair design, Big Data and synthetic biology, social scientists have adopted experimentalism to generate new spaces of problem formulation, engage with different actors and consider different possibilities.[6] That is, a key premise is that experimental modes of *collaboration* can generate new ways of thinking.

Broadly speaking, we can distinguish two models through which collaborative experiments may seek to achieve this. The first is through various forms of participation intended to achieve a degree of democratisation by opening up scientific and technical debates and processes to publics (Marres, 2012). The second is to experiment collaboratively to develop and explore new problem formulations, transcend ingrained styles of reasoning, disrupt existing hierarchies and critically examine how knowledge is created (Rabinow & Bennett, 2012). This is the model of a "collaborator" (or, co-laboratory) where participants engage in the common exploration of a topic. The Socialising Big Data project previously mentioned engaged with this model by conducting workshops and discussions with national statisticians, genomic scientists and waste management engineers to define and develop shared concepts for understanding Big Data (Ruppert *et al.*, 2015). Another form of collaboration involves the co-production of a "thing" – a tangible end-product – through which collaborators *practically* explore and develop shared concepts and issues. Working on a common product makes "issues experimentally available to such an extent that "the possible" becomes tangible, formable, and within reach" (Binder *et al.*, 2015, p. 12). As a method, it forces participants to make future modes of working explicit (Muniesa & Linhardt, 2011). Generally, from the social studies of science we learn that such collaborative experiments also require reshaping relations between participants, technologies and knowledge. This is also a principle of what is called in the social sciences and humanities, practice-based research, which involves an engagement between participants and the skills, materials, small tasks and everyday labour, in addition to texts and spoken word, that are enrolled in making things (Jungnickel, 2017). Making things, as opposed to unravelling or deconstructing them, involves a close entanglement with different participants and can increase understanding of the skills, relations and infrastructures that are part of an end-product (*ibid.*).

Experimentalism is especially recognised as a necessary approach to uncertainty and change. For example, in an article on a collaboration between academics, farmers and environmentalists, Waterton and Tsouvalis (2015, p. 477) ask how "the politics of nature can be envisioned for an age conscious of the complexity,

[5] *See for example, experimental statistics produced by Eurostat: http://ec.europa.eu/eurostat/web/experimental-statistics/.*

[6] *For these three examples, see: https://entornoalasilla.wordpress.com/english/; Ruppert* et al.*, 2015; and http://www.anthropos-lab.net/about.*

contingency, and relationality of the world?" They investigate a collaboration between themselves as social scientists with environmental experts and farmers to improve water quality. In their experience, a shared inquiry opened up questions of how to understand water pollution: in terms of isolated causes or wider socio-technical relations and histories. They thus adopted an agenda of experimentation that understands the generation of knowledge as involving "hybrid forums" (Callon *et al.*, 2011) or "new collectives" (Latour, 2006) in which participants reflexively engage in reconstructing the relations, histories and stakeholders involved in an issue. Uncertainty is not something to be solved, instead it needs to be acknowledged and worked with in an ongoing collective process of knowledge production. In practice this entails a "care-full" approach (Grommé, 2015) which entails the exercise of responsibilities for monitoring and documenting who and what are (unavoidably) included and excluded; avoiding ambiguity about the terms of evaluation by making explicit how outcomes are assessed; recognising that failure is likely caused by myriad factors; and, understanding that values are inseparable from facts. "Care-full" therefore does not only refer to a cautious approach, but also active acknowledgement that experiments continually reshape relations and redistribute effects in sometimes unexpected ways.

As a principle of citizen data, experimentalism thus involves not only experimenting but collaborating to make ways of thinking and generating knowledge "open" to the influence and insights of others and in doing so imagining and speculating on alternatives and possibilities (Stengers, 2010). It requires being accountable to and accounting for the procedures and practices of experiments. Finally, it means being open to how relations between different participants in the making of knowledge might be organised differently. Taking up our point on new relations between citizens and NSIs, experimentalism thus involves active and open forms of participation and influence. We develop this further through a second principle, that of citizen science, to explore how relations between NSIs and citizens in the making of data and official statistics might further be reconceived.

Citizen Science

Some statistical organisations have started experimenting with models of citizen engagement in the production of data. Such models often draw on existing conceptions of citizen science, which we will briefly discuss here to explore how we might reconceive them. Different models of citizen science conceive of citizens as not only research subjects, but as actively involved in the production of data as opposed to traditional methods where they are usually understood as respondents. There are many definitions and interpretations of citizen science and the terms of citizen engagement in the making of data. The European Commission (EC), for example, defines it as the "production of knowledge beyond the scope of professional science, often referred to as lay, local and traditional knowledge" (European Commission, 2013, p. 5). Goodchild (2007) uses the term to describe communities or networks of citizens who act as observers in some domain of science. This is the most commonly accepted definition especially evident in the significant momentum citizen science has gained in the natural sciences in recent years (Kullenberg & Kasperowski, 2016, p. 2). However, the practice of engaging people in collecting and submitting data for scientific purposes goes back at least to the 1960s, though the term itself was not used until the 1990s (*ibid.*).[7]

A second version involves citizens not as only observers but co-producers or producers of scientific studies and data to reflect their own concerns, needs and questions. This version includes local and activist-oriented approaches referred to as "community based auditing", "civic science", "community environmental policing", "street science", "popular epidemiology", "crowd science", and "Do It Yourself Science" (Kullenberg & Kasperowski 2016, p. 2). These versions range from citizens seeking close alliances with scientific and knowledge institutions to citizens engaging in the production of independent knowledge together with scientists.

Citizens' objectives for engaging in scientific data production are multiple, ranging from documenting concerns about environmental issues, to creating online archival maps of local historical sites or transcribing Shakespearean contemporaries.[8] Goodchild (2007, p. 219) suggests that people who generally participate and share information on the internet are more likely to volunteer geographic information and

[7] *For some researchers, it includes the National Audubon Society's Annual Christmas Bird Count in early 1900s, where citizens participated in the observation and enumeration of bird species.*

[8] *Some of these examples are documented at www.zooniverse.org.*

contribute to data collection initiatives such as OpenStreetMap (OSM). On this basis he argues that two kinds of people are likely to participate: people who seek self-promotion and volunteer personal information on the internet to make it "available to friends and relations, irrespective of the fact that it becomes available to all"; and, people who seek personal satisfaction derived from contributing anonymous information and seeing it appear as part of a developing "patchwork" of collective contributions (*ibid.*, p. 219).

Jasanoff (2003) notes that models of citizen science can facilitate meaningful interaction among policymakers, scientific experts, corporate producers and publics (pp. 235–236). She argues that the pressure for accountability in expert decision-making is manifest in the demand for greater transparency and wider participation. However, participatory opportunities cannot alone ensure the representative and democratic governance of science and technology. Jasanoff underscores that the attention of modern states has focused on refining "technologies of hubris" that are designed to facilitate management and control by bracketing off uncertainty, political objections and the unforeseen complexities of everyday life (p. 238). What is lacking is not just knowledge, but ways to bring uncertain, unknown processes and methods into the dynamics of democratic debate (pp. 239–240). For this reason Jasanoff suggests citizen science as a possible model of democratic interaction between different stakeholders in the production of science. In this way citizen science models can be thought of as "technologies of humility", that is, *social* technologies that involve relations between governments, decision-makers, experts, and citizens in the management of technology for "assessing the unknown and the uncertain, 'modest assessments'" that engage citizens as active agents of knowledge, insight, and memory (p. 243; italics in the original).

One concern with the role of non-scientists in the production of science are the implications for established scientific principles.[9] However, as Goodchild (2007) demonstrates, while strictly speaking citizen science might not fulfil scientific criteria per se, it can potentially open up new ways of thinking and approaching data. This is especially relevant for practices of democratisation, which call for different forms of reasoning, as captured in Herbert Simon's (1947) conception of "satisficing" rather than "optimizing" or "maximizing" in decision-making. In opposition to abstractions such as utility theory he advanced an understanding based on how people reason in practice. Practical reasoning, he argued, involves juggling numerous criteria and arriving at a "good enough" solution rather than engaging in an infinite search for all possible ones, evaluating them and then arriving at the best one. Gabrys & Pritchard (2015) take a similar approach to suggest that the adequacy of an answer depends on how practical questions are posed. Instead, they define "just good enough data" to counter the reliance on measurement accuracy as the only objective and criterion for evaluating environmental data gathered through citizen sensing practices. Measurements of environmental phenomena meet different objectives or questions, which are often not known in advance. For instance, a "rough" measurement to identify a pollution event when it is happening or when it has happened might be sufficient and "good-enough". What Gabrys & Pritchard draw attention to is that the potential uses or value of data are often not known in advance and that there is value in organising data production and interpretation as practices of searching for potential rather than reiterating and replicating already known objectives or questions through previously established methods.

Recent experiments by statistical organisations with models of citizen engagement include a pilot project by Statistics Canada using OSM for crowdsourcing citizen work to help fill in data gaps on geolocations (Statistics Canada, 2016).[10] OSM is a collaborative initiative designed to create a free and editable map of the world. The application for Statistics Canada allows users to select a geolocation and edit, for instance, the name of a street. Another example is from the European Commission's Joint Research Centre on Citizen Science and Open Data which has explored possible models of citizen engagement for monitoring the spread of invasive alien plant species (IAS) (Cardoso *et al.*, 2017). That report argued that the implementation of the IAS Regulation could benefit from the contributions of citizens in providing "accurate, detailed, and timely information on IAS occurrences and distribution for efficient prevention, early detection, rapid response, and to allow for evaluation of management measures" (p. 5). Additionally, this form of citizen engagement could raise awareness and increase public support for the regulation as

[9] *Also see Gabrys* et al. *(2016) for discussions about data quality and credibility.*

[10] *The pilot was organized by Statistics Canada in collaboration with OpenNorth, MapBox, City of Ottawa and OSM Canada. OpenNorth is a non-profit organization developing digital tools for civic engagement.*

well as supporting citizens in acquiring skills and better understanding of scientific work (Socientize Consortium, 2014). The United Nations has also identified citizen science data production on environmental issues as necessary to the measurement and monitoring of sustainable development goals (SDGs) (United Nations, 2016). Modes of citizen engagement are recognised as key to ensuring that the 2030 Agenda for Sustainable Development is country-owned and context specific and with goals linked to national values and priorities. While these initiatives conceive of citizen engagement in varying ways, they generally limit it to tasks such as data production, verification and classification. This has led to criticisms of these forms of citizen science as exploitative of citizens as free public labour (DataShift, n.d.; Piovesan, 2017; Paul, 2018). What they point to is that tasks related to data cleaning, coding or analysis as well as design, architecture or interpretation are reserved for experts while citizens are limited to being no more than research subjects or assistants.

We reconceive of citizen science in a way that is more closely aligned with what Jasanoff expresses as the inclusive generation of knowledge. But, following from our argument about detachment, we suggest that inclusivity involves the right to make claims and articulate concerns about how environmental, economic and social issues should be categorised and known.[11] Arguably, this is the claim citizen scientists make when they engage in the independent production of data to challenge or supplement official and scientific knowledge. However, our conception of citizen data envisages citizens not as independent but as co-producers. In this way, we conceive of citizen data as involving new relations between citizens and NSIs in ways that combine statistical science and citizen science. Such a conception could involve citizen engagement in statistical production and lead to statistics that are more representative and inclusive of citizens' concerns, needs and experiences, as well as their own identifications. As such, it would necessitate an approach that is flexible and experimental in its criteria (Paul, 2018) so that it can adapt to the shifting needs and requirements of not only citizens, but also what matters to them. As we suggest below, this includes broadening the understanding of ethics beyond consent, fairness, and data protection to what is arguably at the core of the rise of citizen science: citizens as active in the making and shaping of the data through which official statistics and knowledge are generated. In the next section, we explore what this understanding of ethics might mean in relation to another matter of concern: proposals for "smart statistics".

Smart Statistics

Propositions by Eurostat for the development of "smart statistics" build on conceptions of "smart cities", usually understood as the use of Big Data, urban sensors, Internet of Things (IoT) and other forms of data production and data integration to streamline municipal governance and transportation infrastructures, rejuvenate local economies, transform the urban environment to make it more sustainable, liveable, and socially inclusive (see for instance Henriquez, 2016). While smart cities have been defined in various ways, the concept generally refers to on the one hand how "cities are increasingly composed of and monitored by pervasive and ubiquitous computing and, on the other, whose economy and governance is being driven by innovation, creativity and entrepreneurship, enacted by smart people" (Kitchin, 2014, p. 1). In this view, Big Data offers the possibility of real-time analysis of city life, new modes of urban governance, and envisioning and making more efficient, sustainable, competitive, productive, open and transparent cities.

Leveraging "smart systems" such as smart energy, smart meters, smart transport, and so on is an objective of proposals for "smart statistics" put forward by Eurostat's Big Data Task Force. The proposals seek to engage with the potential of the proliferation of digital devices and sensors connected to the internet and how the data they generate might be embedded in statistical production systems such that statistics could be produced in "real-time" and "automatically".[12] In this view, data capturing, analysis and processing are envisioned as embedded in activities that generate and simultaneously analyse data. The adoption of such an approach could dramatically transform the production system for official statistics and calls for rethinking business processes and architectures, laws and regulations, ethics, methodologies, and so on.

[11] This is an understanding advanced in the field of critical citizenship studies and summarised in Isin & Ruppert (2015) and Isin & Saward (2013). Being a citizen is understood as a political subjectivity that includes not only the possession of rights but the right to make rights claims such as the right to shape how data is made about them and the populations of which they are being constituted as a part (Ruppert, 2018).

[12] Eurostat Big Data Task Force (2016) "Smart Statistics". Draft document. October.

Two approaches for generating smart statistics understood in this way have been proposed: using third party systems that exist for other purposes than statistics but from which statistical information can be extracted (e.g., mobile phones); or developing entirely new data production practices such as sensors and digital devices exclusively for generating statistical information.[13] The third-party approach engenders many of the concerns we previously identified such as data access and ownership, privacy and ethics, data representativeness, quality, and trust as well as greater detachment between citizens and NSIs. However, the latter approach of designing new devices of data production, provides an opportunity to mitigate these issues. That is, we reconceive of smart statistics as not only requiring that NSIs rethink the technical and organisational aspects of statistical production systems, but also their relations to citizens. As noted in the discussion of citizen science, this could involve models of co-production that engage citizens in the production of smart statistics.

It would, however, mean being care-full in the ways we previously outlined including a broader understanding of ethics that extends throughout the production of official statistics. Ethics of course have long been central principles of official statistics, which address the values of utility, professional standards and ethics, scientific principles, transparency, quality, timeliness, costs, respondent burden, and confidentiality (UN, 2014).[14] These principles constitute what we would call an ethic of care for data, such as care for the quality, accessibility and clarity of data, but also for relations and accountabilities to citizens through practices such as data protection, confidentiality, consent, and trust. While the origins of these principles are a mix of legal, governmental, political and professional rationales and requirements, they tend to operate as part of everyday working values and commitments. This is evident in claims made by statisticians such as "just because you can, doesn't mean you should" use Big Data sources.

The fundamental principles of official statistics thus express a broad conception of ethics that includes relations to citizens that social science research calls procedural ethics (Guillemin & Gillam, 2004). Procedural ethics are understood as an estimation of the ethical issues that might be involved when research and data production are undertaken. However, Guillemin and Gillam note a second dimension of ethics in research, which they term "ethics in practice" (id., p. 261). It concerns the recurrent, iterative, and uncertain ethical moments that happen during research and which may be odds with that covered in a procedural ethics review. This latter understanding is relevant to practices involved in the co-production of smart statistics, which, by definition, involve uncertainty, adaptation and responsiveness to the interactions, interests and demands of different stakeholders. As such, co-production demands an ethic of care that recognises and is responsive to the dependence on relations to citizens and their labours to "create, hold together and sustain" data (Puig de la Bellacasa, 2012, p. 198).

The concept of citizen data we propose thus reconceives of smart statistics as involving new relations to citizens as co-producers of data production platforms. It is a conception that calls for a care-full approach that enlarges the understanding of ethics to include the demands, interests and contributions of citizens at different stages of the development of new devices of data production rather than at the backend as an afterthought or correction. As such, it is a model that builds on the premises of another matter of concern, privacy-by-design, which addresses issues of privacy and consent at the frontend of software design, which we address next.

Privacy-by-Design

Big Data and new data sources come with new questions concerning privacy, consent and confidentiality that are not always fully addressed by existing regulatory frameworks. As such, privacy-by-design has become as matter of concern for NSIs. Privacy-by-design is understood as the embedding of privacy protection at the software design stage of data production platforms, devices or applications. It entails designing privacy protection with citizens in mind at the outset and the implementation of these designs in a

[13] Ibid. *One example is Statistics Netherlands collection of data for statistics about road traffic intensities which are produced purely on the basis of road sensors. See: https://www.cbs.nl/en-gb/our-services/innovation/nieuwsberichten/recente-berichten/new-steps-in-big-data-for-traffic-and-transport-statistics.*

[14] *Six principles are that: official statistics must meet the test of practical utility; be developed according to strictly professional considerations, scientific principles and professional ethics; present information on the scientific standards of their sources, methods and procedures; may be generated from all types of sources such as surveys or administrative records and the source chosen with regard to quality, timeliness, costs and the burden on respondents; are to be strictly confidential and used exclusively for statistical purposes; and the laws, regulations and measures governing them should be public.*

transparent manner. As such, privacy-by-design is a response to the problem of privacy, consent, and confidentiality through software and which can be used in tandem with other tools, such as privacy impact assessments. By employing privacy-by-design, privacy issues are addressed at the beginning of the design process, in contrast to other approaches that aim at solving privacy issues after software development is complete or leave privacy considerations to legal or regulatory frameworks.

Cavoukian *et al.* (2010) define privacy-by-design through seven foundational principles: proactive not reactive and preventative not reactive; privacy as the default; privacy embedded into design; full functionality that leads to positive sum, not zero-sum outcomes; end-to-end lifecycle protection; visibility and transparency; and respect for user privacy. These principles require designs to be committed to privacy from the beginning and to limit data production to ways that are respectful of citizens' expectations. The principles also require that data production software addresses the likelihood that data may exist after the software stops functioning. The authors also emphasise that the lifecycle of software must be considered when deciding on how to best protect privacy, including making plans for deleting data once the software reaches the end of its lifecycle. Finally, the principles compel organisations dealing with personal data to be transparent in their goals and to remain accountable to citizens.

However, the production and processing of personal data present many other challenges for privacy in addition to individual privacy. Nissenbaum (2004) argues that privacy norms need to be tied to specific contexts. She describes three principles that have dominated debates around privacy throughout the 20th century, namely, limiting surveillance of citizens by governments, restricting access to private information, and curtailing intrusions into private places. She suggests a new term, "contextual integrity", to deal with the new challenges introduced by digital technologies. Contextual integrity demands that information gathering is kept appropriate to the context and obeys the governing norms of distribution within it. The key insight is that norms of distribution vary across cultures, historical periods, locales, and other factors. Additionally, contextual integrity requires awareness of not only the specific site of data production but also the relevance of related social institutions (Nissenbaum, 2009).

Approaches that aim to protect individual privacy may still lead to undesired outcomes in large-scale data production efforts. When individually anonymised data are joined to create profiles, individuals who fit the profile could still experience effects even when they are not identified individually. For example, Graham (2005) discusses how software can be used to assign different categories to different parts of a city based on school performance, house prices, crime rates, etc., which might potentially orchestrate inequalities and discriminate inhabitants, even when they are not personally identified. Similarly, Zwitter (2014) has identified and problematised the potential discriminatory "group effects" of anonymised data such as in practices of profiling.

The use of Big Data also introduces additional privacy challenges. Barocas and Nissenbaum (2014) argue that anonymity and consent are often fundamentally undermined in Big Data applications, and that other approaches are needed to protect integrity, such as policies based on moral and political principles that serve specific contextual goals and values. Instead of focusing on anonymity in Big Data applications, they instead emphasise securing informed consent, not only as a choice for subjects to waive consent or not, but a requirement that data collectors justify their actions in relation to norms, standards, and expectations. To an extent this is addressed in the recently implemented General Data Protection Regulation (GDPR) in member states across the European Union, which is based on a broad understanding of personal data and privacy and will end practices of general consent by default for the production of personal data.[15] It introduces the requirement to think "what is personal data" for all private and public stakeholders which demand, hold or archive personal data, as well as what are the ethical practices required to deal with personal data, given the complexity and connectedness of data systems and proven non-neutrality of algorithms. In sum, privacy is not a single thing but depends on the context of production, accountability for group effects, and mechanisms of informed consent.

Recently, scholars working to address the technical challenges of privacy in relation to Big Data have proposed a method of privacy protection by taking advantage of blockchain technology (Montjoye *et al.*, 2014; Zyskind

[15] *The General Data Protection Regulation came into force in May 2018. See: https://www.eugdpr.org/.*

et al., 2015). Blockchain is a distributed computing method where many devices communicate with one another over a shared network, without requiring a central server to authorise the participation of each member or to keep a list of currently connected members. By applying blockchain technology to privacy, it becomes possible to encrypt and distribute private data over a large network without requiring a trusted central server.

Blockchain privacy methods are intended to solve underlying privacy challenges using a technical framework during software development. However, as we have indicated above, they do not stand on their own as the sole solution to ensuring privacy, but rather supplement legal and policy-oriented considerations such as contextual integrity, group effects and modes of consent through software design. We thus reconceive of privacy-by-design beyond software to include citizen privacy as a right that should be built into not only the frontend of software design but through relations with citizens as co-producers in the production of official statistics. That is, like ethics, privacy is processual and cannot be settled through the one-time granting of consent or software design alone or independent of specific contexts.

* *
*

In sum, we have taken up matters of concern expressed by statisticians and reconceived of them as principles of citizen data. Through the discussion of the four principles of experimentalism, citizen science, smart statistics and privacy-by-design, we have explored how citizen data can create new attachments and relations between citizens and NSIs, and between citizens' actions, identifications and experiences and how they are categorised, included and excluded in statistics. In this regard, we argue it has the potential to produce new statistical variables desired and identified by citizens, increase their identification with official statistics and possibly advance their role as also users of statistics. Indeed the latter may well be a collateral effect of co-producing statistics with citizens in ways that are more in accordance with their experiences and knowledge.

We place the significance of our concept of citizen data within the current proliferation of data production platforms that enable myriad data generators (e.g., platform owners) and analysts (e.g., researchers, governments, media) to produce statistics and knowledge of societies (Ruppert *et al.*, 2013). Indeed, many topics of interest to NSIs such as price levels, the economy, consumer sentiment or tourism can be measured using Big Data generated by browsers, social media or devices such as mobile phones that can be accessed and analysed by different actors. Some would claim that this represents a "democratisation" of knowledge and the erosion of validated knowledge and expertise about societies. However, as Ruppert *et al.* (2013) contend, this widening distribution of data and analysis means that knowledge of societies does not cohere in single authoritative accounts to the same extent that it perhaps did in the recent past. Instead, what constitutes legitimate knowledge and expertise have become major sites of political contention and struggle as revealed in current debates about "alternative facts".

Proposals that NSIs need to thus defend the quality and legitimacy of official statistics through gatekeeping practices such as demonstrating their trustworthiness by making their statistical practices transparent and thus assessable, fact checking competing statistics, and "calling out bad numbers" certainly have a role to play. However, they potentially play into the premise that what is at stake is winning a competition of "facts". They ignore that what constitutes "public facts" should be open to democratic contestation and deliberation because they inevitably involve normative judgements about social meaning and choices about which experiential realities matter (Jasanoff & Simmit, 2017). We thus suggest NSIs have a role to play in fostering official statistics as social and collective accomplishments where their legitimacy is derived from conditions of co-production that address data subjects as citizens with rights to be active participants. Such an approach understands data and official statistics as social technologies that require new forms of engagement and relations between experts, decision-makers, and citizens for addressing collective problems (Jasanoff, 2003) and as matters of democratic deliberation where citizens are active in the making and shaping of knowledge about societies of which they are a part.

We recognise that the concept of citizen data raises many practical and political questions. For one, we are not suggesting that existing methods and their relations to citizens will become obsolete. However, methods such as surveys and questionnaires will likely change as digital technologies are increasingly adopted and a concept of citizen data can possibly inform those changes. That is, beyond Big Data sources, how data is produced by NSIs using various methods can be reconceived along the lines of what we call citizen data. While online or digital surveys and censuses, for example, are being adopted they do not imagine the possibilities of co-production. Different modes of co-production could be adopted that utilise the affordances of digital technologies and potentially produce data that more closely aligns with the experiences and knowledge of citizens.

Throughout our discussion we have defined co-production as involving citizens in the statistical production process. What this would mean practically is of course a major question and extends to issues of representativeness and inclusion in that process. This is a matter of concern for all methods especially taking into account the heterogeneity of citizens. For methods that mobilise digital technologies such as online censuses and surveys this is potentially exacerbated by what has come to be called the "digital divide". These are only some of the possible practical and political issues that arise from citizen data, which we also addressed in the collaborative workshop with statisticians noted previously. While we have not reported on the outcomes of that workshop in this article, one outcome was imagining alternative "roadmaps" for engaging with citizens at different stages of the statistical production process, from the co-design of prototypes for data generation platforms and apps to the establishment of co-operative forms of data ownership. In other words, citizen data does call for rethinking statistical production processes and some of their fundamental premises.

For example, aspects of statistical production that would need to be rethought are those of data standards and quality. However, as noted, the principle of experimentalism calls for being open to such questions and not settling them in advance including what may or could constitute quality. Interestingly, this is also recognised in NSI experiments with Big Data generated by third party systems where concerns about quality as well as others such as the representativeness of data have been raised. One solution statisticians propose is that statistics that repurpose Big Data could be adopted not as replacements but auxiliary, complementary or supplementary to existing data sources. While possibly relegating such data to a different status and role, this response provides an opportunity to rethink how statistics are made "official". That is, it suggests that there is not one mode of production or set of standards through which data can be made official. We suggest that this also applies to existing methods that produce data for official statistics but which involve myriad standards and where quality is not singularly defined or measurable. However, the concept of citizen data that we have developed introduces a critical difference that goes beyond issues of standards and quality. It proposes that the authority and expertise to make statistics official are not founded in a single institution, but in processes of co-production and direct relations to citizens. In that regard, citizen data approaches claims of "alternative facts" as not matters of accuracy and standards but of the relations to citizens through which data and in turn statistics are made official. □

BIBLIOGRAPHY

Barocas, S. & Nissenbaum, H. (2014). Big Data's End Run Around Anonymity and Consent. In Lane, J., Stodden, V. Bender, S. & Nissenbaum, H. (Eds.), *Privacy, Big Data, and the Public Good*, pp. 44–75. Cambridge, MA: Cambridge University Press.

Binder, T., Brandt, E., Ehn, P. & Halse, J. (2015). Democratic Design Experiments: Between Parliament and Laboratory. *CoDesign*, 11(3-4), 152–165. https://doi.org/10.1080/15710882.2015.1081248

Boltanski, L. & Chiapello, E. (2007). *The New Spirit of Capitalism*. London: Verso.

Cardoso, A. C., Tsiamis, K., Gervasini, E. ***et al.*** **(2017).** Citizen Science and Open Data: a model for Invasive Alien Species in Europe. Joint Research Centre (JRC) and the European Cooperation in Science and Technology (COST Association), *Workshop Report*. Brussels, BE. https://doi.org/10.3897/rio.3.e14811

Callon, M., Burchell, G., Lascoumes, P. & Barthe, Y. (2011). *Acting in an Uncertain World: An Essay on Technical Democracy*. Cambridge, MA: MIT Press.

Cavoukian, A., Taylor, S. & Abrams, M. E. (2010). *Privacy by Design*: Essential for Organizational Accountability and Strong Business Practices. *Identity in the Information Society*, 3(2), 405–413.
https://doi.org/10.1007/s12394-010-0053-z

DataShift (n.d.). Global Goals for Local Impact: Using Citizen-Generated Data to Help Achieve Gender Equality.
http://civicus.org/thedatashift/wp-content/uploads/2017/01/LanetUmojaProcessandApproach.pdf (accessed 22 February 2018)

European Commission (2013). Environmental Citizen Science. *Science for Environment Policy Indepth Report* N° 9. Bristol: University of the West of England, Science Communication Unit.
http://ec.europa.eu/environment/integration/research/newsalert/pdf/IR9_en.pdf (accessed 22 February 2018)

Fleurbaey, M. (2009). Beyond GDP: The Quest for a Measure of Social Welfare. *Journal of Economic literature*, 47(4), 1029–1075.
https://doi.org/10.1257/jel.47.4.1029

Gabrys, J., Pritchard, H. & Barratt, B. (2016). Just Good Enough Data: Figuring Data Citizenships Through Air Pollution Sensing and Data Stories. *Big Data & Society*, 3(2), 1–14.
https://doi.org/10.1177/2053951716679677

Gabrys, J. & Pritchard, H. (2015). Just Good Enough Data and Environmental Sensing: Moving Beyond Regulatory Benchmarks toward Citizen Action. In *Infrastructures and Platforms for Environmental Crowd Sensing and Big Data*. Barcelona: European Citizen Science Association.
https://ecsa.citizen-science.net/sites/default/files/envip-2015-draft-binder.pdf (accessed 22 February 2018)

Goodchild, M. F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69(4), 211–221.
https://doi.org/10.1007/s10708-007-9111-y

Graham, S. (2005). Software-Sorted Geographies. *Progress in Human Geography*, 29(5), 562–580.
https://doi.org/10.1191/0309132505ph568oa

Grommé, F., Ustek-Spilda, F., Ruppert, E. & Cakici, B. (2017). Citizen Data and Official Statistics: Background Document to a Collaborative Workshop. ARITHMUS *Working Paper* N° 2.
http://arithmus.eu/wp-content/uploads/2015/02/ARITHMUS-collaborative-workshop-wp_final-version-060717-1.pdf

Grommé, F. (2015). *Governance by Pilot Projects: Experimenting with Surveillance in Dutch Crime Control* (Doctoral thesis). Amsterdam: University of Amsterdam.
http://hdl.handle.net/11245/1.486712 (accessed 6 February 2019)

Guillemin, M. & Gillam, L. (2004). Ethics, Reflexivity, and "Ethically Important Moments" in Research. *Qualitative Inquiry*, 10(2), 261–280.
https://doi.org/10.1177/1077800403262360

Henriquez, L. (2016). *Amsterdam Smart Citizens Lab: Towards Community Driven Data Collection*. Amsterdam: De Waag Society and AMS Institute.
https://waag.org/sites/waag/files/media/publicaties/amsterdam-smart-citizen-lab-publicatie.pdf (accessed 2 April 2017)

Hussmanns, R. (2004). Measuring the Informal Economy: From Employment in the Informal Sector to Informal Employment. *Working Paper* N° 53.
http://www.ilo.org/wcmsp5/groups/public/---dgreports/---integration/documents/publication/wcms_079142.pdf (accessed 30 April 2018)

Isin, E. & Ruppert, E. (2015). *Being Digital Citizens*. London: Rowman & Littlefield International.

Isin, E. & Saward, M. (2013). *Enacting European Citizenship*. Cambridge: Cambridge University Press.

Jasanoff, S. (2003). Technologies of Humility: Citizen Participation in Governing Science. *Minerva*, 41(3), 223–244.
https://doi.org/10.1023/A:1025557512320

Jasanoff, S. & Simmet, H. R. (2017). No Funeral Bells: Public Reason in a "post-Truth" Age. *Social Studies of Science*, 47(5), 751–770.
https://doi.org/10.1177/0306312717731936

Jungnickel, K. (2017). Making Things to Make Sense of Things: DIY as Research Subject and Practice. In: Sayers, J. (Ed.), *The Routledge Companion to Media Studies and Digital Humanities*, pp. 492–502. Oxon: Routledge.

Kitchin, R. (2014). The Real-Time City? Big Data and Smart Urbanism. *GeoJournal*, 79(1), 1–14.
https://doi.org/10.1007/s10708-013-9516-8

Kullenberg, C. & Kasperowski, D. (2016). What Is Citizen Science? – A Scientometric Meta-Analysis. *PLOS ONE*, 11(1), e0147152.
https://doi.org/10.1371/journal.pone.0147152 (accessed 2 April 2017)

Latour, B. (2004). Why Has Critique Run Out of Steam? From Matters of Fact to Matters of Concern. *Critical Inquiry*, 30(2), 225–248.
https://doi.org/10.1086/421123

Latour, B. (2006). Which Protocol for the New Collective Experiments? *Boletín CF+S*, (32/33).
http://habitat.aq.upm.es/boletin/n32/ablat.en.html (accessed 2 April 2017)

Marres, N. (2012). *Material Participation: Technology, the Environment and Everyday Publics*. Basingstoke: Palgrave Macmillan.

Montjoye, Y.-A. (de), Shmueli, E., Wang, S. S. & Pentland, A. S. (2014). OpenPDS: Protecting the Privacy of Metadata through SafeAnswers. *PLOS ONE*, 9(7).
https://doi.org/10.1371/journal.pone.0098790

Muniesa, F. & Linhardt, D. (2011). Trials of Explicitness in the Implementation of Public Management reform. *Critical Perspectives on Accounting*, 22(6), 550–566.
https://doi.org/10.1016/j.cpa.2011.06.003

Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review*, 79(1), 119–158.
https://nyuscholars.nyu.edu/en/publications/privacy-as-contextual-integrity (accessed 2 April 2017)

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford: Stanford University Press.

Paul, K. T. (2018). Collective Organization of Discourse Expertise Using Information Technology – CODE IT! *Information Technology*, 60(1), 21–27.
https://doi.org/10.1515/itit-2017-0022

Piovesan, F. (2017). *Statistical Perspectives on Citizen-Generated Data*. [Online].
http://civicus.org/thedatashift/wp-content/uploads/2015/07/statistical-perspectives-on-cgd_web_single-page.pdf (accessed 22 February 2018)

Puig de la Bellacasa, M. (2012). "Nothing Comes Without Its World": Thinking with Care. *The Sociological Review*, 60(2), 197–216.
https://doi.org/10.1111/j.1467-954X.2012.02070.x

Rabinow, P. & Bennett, G. (2012). *Designing Human Practices: An Experiment with Synthetic Biology*. Chicago: University of Chicago Press.

Ruppert, E. (2018). *Sociotechnical Imaginaries of Different Data Futures: An Experiment in Citizen Data*. 3e Van Doornlezing. Rotterdam, NL: Erasmus School of Behavioural and Social Sciences.
https://www.eur.nl/sites/corporate/files/2018-06/3e%20van%20doornlezing%20evelyn%20ruppert.pdf (accessed 21 Jan 2019)

Ruppert, E., Law, J. & Savage, M. (2013). Reassembling Social Science Methods: The Challenge of Digital Devices. *Theory, Culture & Society, Special Issue on "The Social Life of Methods"*, *30*(4), 22–46.
https://doi.org/10.1177/0263276413484941

Ruppert, E., Harvey, P., Lury, C., Mackenzie, A., McNally, R., Baker, S. A., Kallianos, Y. & Lewis, C. (2015). A Social Framework for Big Data. Project Report. CRESC, The University of Manchester and The Open University.
http://research.gold.ac.uk/13483/ (accessed 2 April 2017)

Simon, H. (1947). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. New York: Macmillan.

Socientize Consortium (2014). Green Paper on Citizen Science. Citizen Science for Europe: Towards a Better Society of Empowered Citizens and Enhanced Research. European Commission Digital Science Unit.
https://ec.europa.eu/digital-single-market/en/news/green-paper-citizen-science-europe-towards-society-empowered-citizens-and-enhanced-research (accessed 22 February 2018)

Statistics Canada (2016). *Open Building Data: An Exploratory Initiative*.
http://www.statcan.gc.ca/eng/crowdsourcing (accessed 18 February 2018)

Stengers, I. (2010). *Cosmopolitics*. Vol. 1–2. Minneapolis: University of Minnesota Press.

Stiglitz, J. E., Sen, A. & Fitoussi, J.-P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Paris: CMESP.
http://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report (accessed 30 April 2018)

Struijs, P., Braaksma, B. & Daas, P. J. H. (2014). Official statistics and Big Data. *Big Data & Society*, 1(1), 1–6.
https://doi.org/10.1177/2053951714538417

UNECE (2014). The Role of Big Data in the Modernisation of Statistical Production Project. Report of the Big Data Privacy Task Team.
http://bit.ly/2eTHDOe (accessed 2 April 2017)

United Nations (2014). “Fundamental Principles of Official Statistics”. *Resolution adopted by the General Assembly on 29 January 2014. A /RES/68/261*.
http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf (accessed 21 Jan 2019)

United Nations (2016). Make Sustainable Development Goals Relevant to Citizens. New York: Economic and Social Council.
https://www.un.org/press/en/2016/ecosoc6782.doc.htm (accessed 2 April 2017)

Waterton, C. & Tsouvalis, J. (2015). On the Political Nature of Cyanobacteria: Intra-Active Collective Politics in Loweswater, the English Lake District. *Environment and Planning D: Society and Space*, 33(3), 477–493.
https://doi.org/10.1177/0263775815594305

Zwitter, A. (2014). Big Data Ethics. *Big Data & Society*. 1(2) 1–6.
https://doi.org/10.1177/2053951714559253

Zyskind, G., Nathan, O., & Pentland, A. (2015). Decentralizing Privacy: Using Blockchain to Protect Personal Data. *2015 IEEE Security and Privacy Workshops,* pp. 180–184.
https://doi.org/10.1109/SPW.2015.27

N °503-504 – 2018

MÉLANGES / *VARIA*

INCITATIONS À L'EMPLOI / WORK INCENTIVES IN FRANCE

NOUVEAUX IMPACTS DE LA GLOBALISATION / *NEW IMPACTS OF GLOBALIZATION*

N °500-501-502 – 2018

LOGEMENT ET MARCHÉS DU LOGEMENT / *HOUSING AND HOUSING MARKETS*

SYSTÈMES DE LOGEMENT DANS L'OCDE / *HOUSING SYSTEMS IN THE OCDE*

OFFRE, DEMANDE ET PRIX SUR DIFFÉRENTS MARCHÉS DU LOGEMENT / *SUPPLY, DEMAND AND PRICES ON HOUSING MARKETS*

ACCESSION À LA PROPRIÉTÉ, MOBILITÉ RÉSIDENTIELLE - DYNAMIQUES A MOYEN TERME / *ACCESS TO HOME OWNERSHIP, RESIDENTIAL MOBILITY: DYNAMICS IN THE MEDIUM-TERM*

ÉVALUATIONS D'IMPACT ET MÉTHODES / *EVALUATIONS OF IMPACT AND METHODS*

- L'impact de la hausse des droits de mutation immobiliers de 2014 sur le marché du logement français / *The impact of the 2014 increase in the real estate transfer taxes on the French housing market*
- L'information aux acheteurs affecte-t-elle le prix de vente des logements ? L'obligation d'information et le modèle de prix hédoniques – un test sur données françaises / *Does information to buyers affect the sales price of a property? Mandatory disclosure and the hedonic price model – A test on French data*
- Évaluation des méthodes utilisées par les pays européens pour le calcul de l'indice officiel des prix des logements / *An evaluation of the methods used by European countries to compute their official house price indices*

Economie et Statistique / Economics and Statistics

Objectifs généraux de la revue

Economie et Statistique / Economics and Statistics publie des articles traitant de tous les phénomènes économiques et sociaux, au niveau micro ou macro, s'appuyant sur les données de la statistique publique ou d'autres origines. Une attention particulière est portée à la qualité de la démarche statistique et à la rigueur des concepts mobilisés dans l'analyse. Pour répondre aux objectifs de la revue, les principaux messages des articles et leurs limites éventuelles doivent être formulés dans des termes accessibles à un public qui n'est pas nécessairement spécialiste du sujet de l'article.

Soumissions

Les propositions d'articles, en français ou en anglais, doivent être adressées à la rédaction de la revue (redaction-ecostat@insee.fr), en format MS-Word. Il doit s'agir de travaux originaux, qui ne sont pas soumis en parallèle à une autre revue. Un article standard fait environ 11 000 mots (y compris encadrés, tableaux, figures, annexes et bibliographie, non compris éventuels compléments en ligne). Aucune proposition initiale de plus de 12 500 mots ne sera examinée.

La soumission doit comporter deux fichiers distincts :

- Un fichier d'une page indiquant : le titre de l'article ; le prénom et nom, les affiliations (maximum deux), l'adresse e-mail et postale de chaque auteur ; un résumé de 160 mots maximum (soit environ 1 050 signes espaces compris) qui doit présenter très brièvement la problématique, indiquer la source et donner les principaux axes et conclusions de la recherche ; les codes JEL et quelques mots-clés ; d'éventuels remerciements.
- Un fichier anonymisé du manuscrit complet (texte, illustrations, bibliographie, éventuelles annexes) indiquant en première page uniquement le titre, le résumé, les codes JEL et les mots-clés.

Les propositions retenues sont évaluées par deux à trois rapporteurs (procédure en « double-aveugle »). Les articles acceptés pour publication devront être mis en forme suivant les consignes aux auteurs (accessibles sur https://www.insee.fr/fr/information/2410168). Ils pourront faire l'objet d'un travail éditorial visant à améliorer leur lisibilité et leur présentation formelle.

Publication

Les articles sont publiés en français dans l'édition papier et simultanément en français et en anglais dans l'édition électronique. Celle-ci est disponible, en accès libre, sur le site de l'Insee, le jour même de la publication ; cette mise en ligne immédiate et gratuite donne aux articles une grande visibilité. La revue est par ailleurs accessible sur le portail francophone Persée, et référencée sur le site international Repec et dans la base EconLit.

Main objectives of the journal

Economie et Statistique / Economics and Statistics publishes articles covering any micro- or macro- economic or sociological topic, either using data from public statistics or other sources. Particular attention is paid to rigor in the statistical approach and clarity in the concepts and analyses. In order to meet the journal aims, the main conclusions of the articles, as well as possible limitations, should be written to be accessible to an audience not necessarily specialist of the topic.

Submissions

Manuscripts can be submitted either in French or in English; they should be sent to the editorial team (redaction-ecostat@insee.fr), in MS-Word format. The manuscript must be original work and not submitted at the same time to any other journal. The standard length of an article is of about 11,000 words (including boxes if needed, tables and figures, appendices, list of references, but not counting online complements if any). Manuscripts of more than 12,500 words will not be considered.

Submissions must include two separate files:

- A one-page file providing: the title of the article; the first name, name, affiliation-s (at most two), e-mail et postal addresses of each author; an abstract of maximum 160 words (about 1050 characters including spaces), briefly presenting the question(s), data and methodology, and the main conclusions; JEL codes and a few keywords; acknowledgements.
- An anonymised manuscript (including the main text, illustrations, bibliography and appendices if any), mentioning only the title, abstract, JEL codes and keywords on the front page.

Proposals that meet the journal objectives are reviewed by two to three referees ("double-blind" review). The articles accepted for publication will have to be presented according to the guidelines for authors (available at https://www.insee.fr/en/information/2591257). They may be subject to editorial work aimed at improving their readability and formal presentation.

Publication

The articles are published in French in the printed edition, and simultaneously in French and in English in the online edition. The online issue is available, in open access, on the Insee website the day of its publication; this immediate and free online availability gives the articles a high visibility. The journal is also available online on the French portal Persée, and indexed in Repec and EconLit.

N° 505-506 - 2018

Economie ET Statistique

Economics AND Statistics

Au sommaire
du prochain numéro :
Mélanges

Forthcoming:
Varia

ISBN 978-2-11-151217-7 - ISSN 0336-1454 - ECO 505-506
Parution avril 2019 - PRIX : 17,20 €