Big Data and Macroeconomic Nowcasting: From Data Access to Modelling

Dario Buono[1], George Kapetanios[2], Stephan Krische[3],

Massimiliano Marcellino[4], Gian Luigi Mazzi[5], Fotis Papailias[6] [7]

Abstract

Parallel advances in IT and in the social use of Internet-related applications, provide the general public with access to a vast amount of information. The associated Big Data are potentially very useful for a variety of applications, ranging from marketing to tapering fiscal evasion.

From the point of view of official statistics, the main question is whether and to what extent Big Data are a field worth investing to expand, check and improve the data production process and which types of partnerships will have to be formed for this purpose. Nowcasting of macroeconomic indicators represents a well-identified field where Big Data has the potential to play a decisive role in the future.

In this paper we present the results and main recommendations from the Eurostat-funded project “Big Data and macroeconomic nowcasting”, implemented by GOPA Consultants, which benefits from the cooperation and work of the Eurostat task force on Big Data and a few external academic experts.

Keywords: Nowcasting, Big Data, Machine Learning, Shrinkage, Forecast Combination

[1] European Commission, Eurostat, dario.buono@ec.europa.eu
[2] King’s College, george.kapetanios@kcl.ac.uk
[3] GOPA Consultants, stephan.krische@gopa.de
[4] Bocconi University, massimiliano.marcellino@unibocconi.it
[5] European Commission, Eurostat, gianluigi.mazzi@ec.europa.eu
[6] Queen's University Management School, f.papailias@qub.ac.uk
[7] The views expressed are the author’s alone and do not necessarily correspond to those of the corresponding organisations of affiliation.

1. Introduction

The recent global crisis has emphasized the importance for policy-makers and economic agents of a real-time assessment of the current state of the economy and its expected developments, when a large but incomplete and noisy information set is available. The main obstacle is the delay with which key macroeconomic indicators such as GDP and its components, but also fiscal variables, regional/sectoral indicators and disaggregate data, are released. The project focuses on the particular case of using Big Data for macroeconomic nowcasting, thus possibly enhancing the timely availability and precision of early estimates of key macroeconomic variables, and potentially providing new Big Data based coincident and leading indicators of economic activity.

In a nowcasting context, Big Data provides potentially relevant complementary information with respect to standard data, being based on rather different information sets. Moreover, it is timely available and, generally, not subject to subsequent revisions – all relevant features for indicators providing information about the current state of an economy. Furthermore, it can provide a more granular perspective on the indicator of interest, both in the temporal and in the cross-sectional dimensions.

While it is clear that there is a potential for an important contribution in this context, Big Data raises a number of old and new issues, first of all related to the availability of relevant data, as well as the continuity and quality of its provision. In this context, the establishment of reliable partnerships with data providers, both in the public and the private sector, as well as with the research community, could represent a critical factor of success.

Hence, we discuss the main challenges raised by data preparation, cleaning, filtering, and evaluation, all the more relevant in the context of official statistics, in view of providing a set of recommendations associated with all phases of an operational step-by-step approach for using Big Data in a nowcasting exercise. Additional details can be found, e.g., in Marcellino (2016).

2. Big Data search

The starting point for an assessment of the potential benefits and costs of the use of Big Data for macroeconomic nowcasting is the identification of their source. A first main provider is represented by Social Networks (human-sourced information), broadly defined to include proper social networks but also blogs and comments, pictures, videos, Internet searches, mobile data content, e-mails. The associated data is, typically, loosely structured and often ungoverned.

A second main source of Big Data are Traditional Business Systems (process-mediated data). These processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. Traditional business data is the vast majority of what IT managed and processed, in both operational and BI systems. Usually structured and stored in relational database systems, including also "Administrative data", it can be further grouped into data produced by Public Agencies (medical records, social insurance, etc.) and data produced by businesses (commercial transactions, banking/stock records, e-commerce, credit cards, etc.).

A third, fast expanding, provider of Big Data is the so-called Internet of Things (machine-generated data). This data is derived from sensors and machines used to measure and

record events and developments in the physical world. The well-structured nature of machine-generated data makes it suitable for computer processing, but its size necessitates the use of new statistical approaches.

From an economic nowcasting point of view, all the three types of big data are potentially relevant. For example, selected internet searches (Social Networks), credit card transactions (Traditional Business Systems), or number of navigating commercial vessels in a certain area (Internet of Things) could all provide useful leading indicators for GDP growth of a country. Hence, a crucial step for a proper use of Big Data for nowcasting is a careful search and classification of existing data, having clearly in mind the specificities and the characteristics of the target variable.

3. Assessment of big-data accessibility and quality

Unfortunately, not all existing Big Data are also available. In fact, most data pass through private providers and are related to personal aspects. Hence, once a potentially useful Big Data source is identified for nowcasting a specific indicator of interest, it should be evaluated whether and at what cost the information is actually available. Besides availability, an additional issue is continuity of data provision, which could not be guaranteed and is particularly relevant for the use of internet data in official statistics. The availability of a good quality documentation or of a regularly updated metadata associated to Big Data is another important characteristic when analysing and selecting available Big Data.

A Big Data feature specifically relevant for nowcasting applications is the overall number of temporal observations in the frequency of the target economic indicator (typically, months/quarters). Unfortunately, this is generally low, even if in high frequency or cross-sectionally there can be thousands of observations, as Big Data generation and collection has started only recently. A short temporal sample is problematic as the Big Data based indicators need to be related to the target low frequency macroeconomic indicators and, without a long enough sample, the parameter estimators can be noisy and the ex-post evaluation sample for the nowcasting performance too short.

A final issue and recommendation, specifically relevant for Big Data nowcasting, is the control of the stability of the relationship with the target variable. This is a common problem also with standard indicators, as the type and size of economic shocks that hit the economy vary over time. In the case of Big Data an additional potential source of instability is the fact that both their size and quality keeps changing over time, in general much faster than for standard data collection.

4. Big data preparation

Even when available, there are often substantial costs to make the Big Data suitable for nowcasting exercises. Actually, big data is often unstructured, so that a proper mapping into regularly spaced cross-sectional or time series observations is required. This topic is not considered formally in the econometric or statistical literature on Big Data, and there is no unique way to transform unstructured into structured data, as the specific transformation depends on the type of Big Data. However, most transformations can be treated in a unified analytical context, where they are considered as functions that map the Big Data into time series space.

This requires to "clean" the variables prior to econometric modelling, replacing outliers and missing observations with reasonable estimates, removing other deterministic effects (such as calendar ones) and filtering for seasonal and other short-term periodic movements such as intra-monthly or intra-weekly ones. When the number of variables is really large and/or the adjustment has to be implemented many times, as in the case of recursive forecasting exercises, it is convenient to work on a series by series basis. Since not all seasonal and calendar adjustment methods can be applied when data are available at high frequency, appropriate adjustment techniques need to be identified or developed when the data are available at high frequency. The size of the datasets suggests resorting to robust and computationally simple univariate approaches.

5. Designing a Big Data modelling strategy

Once temporally structured, properly cleaned, Big Data is available, we can proceed to identify and implement one or several proper econometric methods to match the target indicator with the Big Data based explanatory variables, and conduct a careful in-sample and pseudo-out-of-sample evaluation (cross-validation) of the alternative methodologies. We discuss these aspects in the next two steps.

Big Data prevents the use of standard econometric methods. For example, when the number of regressors is larger than that of observations (N>>T, as in FAT datasets), OLS estimation clearly cannot be used, as well as OLS based statistics, such as t-tests and F-tests to check the significance of regressors. Moreover, selecting regressors by means of information criteria also becomes not doable, as 2^N models should be compared, a number larger than one million already for N=20 regressors. Furthermore, standard statistical theory to prove econometric properties such as unbiased and consistency of the estimators typically relies on fixed N and T diverging asymptotics (suited for TALL datasets, where T⋙N). Instead, with big (potentially HUGE) data both N and T diverging asymptotics is needed, which is much more complex.

A common approach is to either aggregate the data or to impose strong a priori assumptions on the econometric models for the disaggregate data. Clearly, in general these assumptions are invalid, and data aggregation leads to a loss of information. Hence, proper Big Data econometrics is needed.

Big data econometrics has received a big boost in the recent past, more in terms of estimation and testing than forecasting. There are many approaches available, which can be categorised into five main classes that we now briefly summarize, providing guidelines on when to use each of them.

Machine Learning methods. In, machine learning methods, which are particularly suited for FAT datasets, OLS estimation is regularised to become feasible when N is large. This is typically achieved by adding a set of (nonlinear) constraints on the model parameters, which are thus shrunk towards pre-specified values, preferably towards zero in order to achieve a more parsimonious specification. This class includes methods such as Ridge Regression, the seminar work on LASSO Regression by Tibshirani (1996), Adaptive LASSO, Elastic Net, SICA, Hard Thresholding, Boosting and Multiple Testing. Unfortunately, few applications of these methods are available in the context of macroeconomic nowcasting and forecasting.

Heuristic Optimisation. The rationale of Heuristic Optimisation is to use algorithms to apply, in a Big Data context, information criteria that reach a good balance between model fit and

parsimony by assigning a penalty dependent on the number of model parameters (which is equal to that of regressors in the linear context). Within this class of algorithms, it is worth mentioning Simulated Annealing, Genetic Algorithms, and MC³. As the methods are iterative, and sometimes simulation based, they can become computationally very demanding when N is really large, say already about 1000. As they should be applied recursively in a macroeconomic forecasting context, not only for forecast evaluation but also for cross-validation, the computational aspect can become quickly prohibitive. However, in other contexts, e.g. for the construction of coincident and leading composite indexes, these methods could be quite useful at the stage of indicator selection.

Dimensionality reduction techniques. A third class of econometric methods to properly handle big data is based on the idea of reducing the dimension of the dataset by producing a much smaller set of generated regressors, which can then be used in a second step in standard econometric models to produce nowcasts and forecasts in common ways. There are naturally many ways to carry out dimensionality reduction, the most common are Principal Component Analysis and Partial Least Squares, which can handle TALL datasets, and Sparse Principal Component Analysis, which is also suited for FAT and HUGE datasets.

Shrinkage Estimators and Bayesian Methods. Shrinkage estimators typically regularize OLS estimation, making it feasible also when N is very large and larger than T, by adding a set of a priori constraints on the model parameters. The constraints induce a bias in the estimator but also reduce its variance. As a consequence, some bias is also introduced in the forecasts, but their efficiency can be improved. In a classical context, shrinkage can be obtained in various ways, including by means of stochastic constraints on the model parameters. However, Bayesian methods are more commonly used, where prior distributions on the parameters permit to achieve the desired level of shrinkage (besides allowing the user to incorporate a priori opinions on the relevance of each explanatory variable).

Nowcast pooling. Forecast pooling (or combination) has a long tradition of empirical success, and nowcast pooling is promising as well. Possible reasons for the good performance of forecast pooling may be model misspecification, model uncertainty and parameter non-constancy, which are attenuated by weighting. As these features are likely present when modelling with big data, forecast combination could be helpful also in this context. Hence, an alternative procedure in the presence of a big set of potentially useful leading indicators for the target variable of interest is to use a (possibly very large) set of small econometric models to produce nowcasts, one model for each of the N available indicator or small subset of them, and then to combine the resulting many nowcasts or forecasts into a single prediction. An additional benefit of this approach is an evaluation of the forecasting ability of each of the many available indicators, which is of interest by itself.

6. Methodological findings and recommendations

As Hartford (2014) put it: "Big data has arrived, but big insights have not. The challenge now is to solve new problems and gain new answers - without making the same old statistical mistakes on a grander scale than ever." In fact, assessing in a critical and comprehensive way the contribution of Big Data for nowcasting an indicator of interest is of key importance and could be done by answering a set of questions, that we discuss next.

Is there a "Big Data hubris"? "Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis, see Lazer et al. (2014). In a nowcasting context, this is the case for studies that are purely based on Big Data indicators, and it is not surprising that Big Data indicators are useful when used this way, but their usefulness could be spurious. To attenuate the Big Data hubris we should think, as mentioned, of Big Data based indicators as complements to existing soft and hard data-based indicators, include all of them in econometric models, and assess the marginal contribution of each type of indicator. This evaluation should be conducted in an out-of-sample context, as in-sample Big Data can lead to overfitting. Various specifications including alternative combinations of traditional and Big Data based indicators should be used for nowcasting over a training sample, and their associated forecast errors evaluated and compared, for example in terms of mean squared and mean absolute forecast errors.

Is there a risk of "False positives"? This question can be reformulated as: can we get some Big Data based indicators that nowcast well just due to data snooping? Similarly, can we get positive results because of model snooping, since we have seen that various econometric approaches are available? The risk of false positives is always present in empirical analysis and is magnified in our case by the size of data and number of models. Only a careful and honest statistical analysis can attenuate this risk. In particular, we recommend coomparing alternative indicators and methods over a training sample, select the preferred approach or combine a few of them, and then test if they remain valid in a previously unused sample.

Are correlations mistaken for causes when interpreting the results? Again, this is a common problem in empirical analysis that is exacerbated in a Big Data context. For example, a large number of internet searches for "filing for unemployment" can predict future unemployment without, naturally, causing it. Hence, we should abstain from a causal interpretation of the results, unless it can be sustained by economic theory and/or institutional considerations.

Is there instability in the nowcasting ability of specific Big Data based indicators? Instability is a common source of errors in nowcasting exercise also with standard indicators. It can be due to a variety of reasons, such as recurrent crisis, more general institutional changes, discontinuity in data provision, etc. In the case of Big Data, there are some additional specific reasons for instability, such as the increasing use of internet and what Lazer et al. (2014) labelled "Algorithm Dynamics", namely the continuous changes made by engineers to improve the commercial service and by consumers in using that service. Instability is indeed often ignored in the current big data literature. Unfortunately, detecting and curing instability is complex, even more so in a big data context. However, some fixes can be tried, borrowing from the recent econometric literature on handling structural breaks.

7. Conclusions

Overall, we are very confident that Big Data are valuable in a nowcasting context, not only to reduce the errors but also to improve the timeliness, frequency of release and extent of data revision. The approach we have developed in this presentation naturally requires further extensions and adjustments, for example to consider more systematically the transformation from unstructured to structured Big Data, filtering issues, more advanced Bayesian estimation and forecasting methods, more careful real time evaluations, and alternative ways to present the results. Nonetheless, we hope that this introductory presentation will be nonetheless useful for many users.

References

Harford, T. (2014, April). Big data: Are we making a big mistake? Financial Times. Available at http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html #ixzz2xcdlP1zZ

Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014). "The Parable of Google Flu: Traps in Big Data Analysis", Science, 143, 1203-1205.

Marcellino, M. (2016), “Nowcasting with Big Data”, Keynote Speech at the 33rd CIRET conference.

Tibshirani, R. (1996). “Regression Shrinkage and Selection via the Lasso”, Journal of the Royal Statistical Society B, 58, 267-288.