First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.4243

Some guidance for the use of Big Data in macroeconomic nowcasting

Mazzi, Gian Luigi[a1]
[a]Eurostat, European Commission, Luxembourg.

Abstract

This paper develops an operational step by step approach aiming to facilitate the use of Big Data in nowcasting exercises. Each step includes a description of the problem and a set of recommendations addressing the most relevant available solution. The approach includes nine steps starting from the theoretical availability of Big Data until the publication of new nowcasting including also Big Data. In designing this operational step by step approach, the preliminary results of an ongoing Eurostat project on Big Data and macroeconomic nowcasting have been used as a starting point. Further elaboration has been carried out in order to make the operational step by step approach more concrete and prescriptive. Its aim is to provide a concrete help for experts involved in the construction of nowcasting especially in the judgment about the usefulness of the presence of Big Data in their models. It also provides guidance related to the dissemination of new nowcasting based also on Big Data.

Keywords: *Big Data; Nowcasting.*

[1] The information and views set out in this paper are those of the author and do not necessarily reflect the official opinion of the European Commission.

1. Introduction

The availability of Big Data is opening new challenging ways of producing statistics. Big Data can be particularly relevant to increase the timeliness of macroeconomic indicators by means of new types of nowcasting. At present, Big Data should be viewed realistically more as a complement of traditional information to produce nowcasting instead of an alternative. The presence of Big Data can substantially change the traditional ways of building up nowcasting. This paper, also based on the preliminary results of an ongoing project on Big Data and macroeconomic nowcasting, is proposing a step by step approach for the utilization of Big Data in a nowcasting exercise.

2. A step by step approach for the use of Big Data in a nowcasting exercise

2.1. The step by step approach

In this section we are proposing some guidance, expressed in a form of a step by step approach, for using Big Data when building up macroeconomic nowcasting. Each step will be accompanied by a detailed description as well as one or more recommendations, suggested to the compilers of macroeconomic nowcasting. Table 1 summarises the various steps together with their aim while, in the following subsection, each step will be further detailed.

Table 1.

Steps	**Title**	**Aim**
Step 0	Big Data usefulness within a nowcasting exercise	Checking for the existence of adequate Big Data sources
Step 1	Big Data search	Identification of the appropriate Big Data
Step 2	Availability and quality	Verification of Big Data availability and its quality
Step 3	Accounting for Big Data specific features	Move from unstructured Big Data to a structured dataset
Step 4	Big Data pre-treatment	Removing deterministic and periodic undesirable effects
Step 5	Presence of bias	Checking Big Data for the presence of bias

Step 6	Big Data modelling	Identifying the best modelling strategy
Step 7	Results evaluation of Big Data based nowcasting	Checking for the effective contribution of Big Data
Step 8	Implementing Big Data based nowcasting	Timing and scheduling for the new nowcasting

2.2. Big Data usefulness within a nowcasting exercise

2.2.1. Description

This first step should investigate the potential usefulness of Big Data for a specific indicator of interest, such as GDP growth, inflation or unemployment rate. Big Data sources should be considered for their ability of improving the overall quality of existing nowcasting or of producing timelier estimates. The theoretical soundness of the relationships between existing Big Data sources and the target variables should also be investigated.

2.2.2. Recommendations

- Suggest the use of Big Data only when there are well founded expectations of their usefulness either for fixing problems in existing nowcasting or to improve the timeliness.
- Do not consider Big Data sources with doubtful or even spurious correlations with the target variable.

2.3. Big Data search

2.3.1. Description

Once Big Data passes the "need check" in the previous step, the next action of the Big Data based nowcasting exercise is a careful search for the specific Big Data to be collected. There are many potential providers such as social networks, traditional business systems, the Internet of things, etc. It is very difficult to give general guidelines on a preferred data source because the choice is heavily dependent on the target indicator of the nowcasting exercise.

2.3.2. Recommendations

- Searching in the wider possible set of Big Data having clearly in mind the specificities and the characteristics of the target variable as well as what we want to nowcast.
- Checking for the adherence of available Big Data to what the target variable is really measuring.

2.4. Availability and quality

2.4.1. Description

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

2.4.2. Recommendations

- Privileging data providers which are able to give sufficient guarantee of the continuity of the data process and of the availability of a good and regularly updated metadata associated to the Big Data
- Privileging Big Data sources which ensure sufficient time coverage to properly building up a nowcasting exercise.

2.5. Accounting for Big Data specific features

2.5.1. Description

The third step analyzes specific features of the collected Big Data. A first issue concerns the amount of the required storage space and the associated need of specific hardware and software for storing and handling the Big Data. The second issue is the type of the Big Data, as it is often unstructured and may require a transformation into cross-sectional or time series observations.

2.5.2. Recommendations

- Creating a Big Data specific IT environment where the original data are collected and stored with associated routines to automatically convert them into structured, either cross-sectional or time-series datasets.
- Ensure the availability of an exhaustive documentation of the Big Data conversion process.

2.6. Big Data pre-treatment

2.6.1. Description

Even when already available in numerical format or after their transformation into numerical form as in the previous step, pre-treatment of the Big Data is often needed to remove deterministic patterns such as outliers and calendar effects and deal with data irregularities, like missing observations. Furthermore, seasonal and non-seasonal short-term movements (i.e. infra-monthly ones) should be removed accordingly to the characteristic of the target variable. 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 when the data are available at high frequency. The size of the datasets suggests resorting to robust and computationally simple univariate approaches.

2.6.2. Recommendations

- Whenever possible, all the data treatment described in this step should be done within a unique framework in order to avoid inconsistencies between different parts of the process.
- The filtering of Big Data should be consistent to the one used for the target variables: for example if the target variable is not seasonally adjusted, there is no reason to remove the seasonal component from Big Data and vice-versa.

2.7. Presence of bias

2.7.1. Description

This step requires assessing the presence of a possible bias in the answers provided by the Big Data, due to the so-called “digital divide” or the tendency of individuals and businesses not to report truthfully their experiences, assessments and opinions. Another relevant and partially related problem, particularly relevant for nowcasting, is the possible instability of the relationship with the target variable. This is a common problem also with standard indicators and traditional nowcasting exercises. Both issues can be however tackled at the modelling and evaluation stages.

2.7.2. Recommendations

- If a bias in the Big Data answers is observed, provided that it has been reasonably stable in the last few years, a bias correction can be included in the nowcasting strategy.
- If a bias in the Big Data answers is very unstable, then the Big Data should be considered not reliable enough to be used in a nowcasting exercise.
- In order to deal with a possible instability of the relationships between the Big Data and the target variables, nowcasting models should be re-specified on a regular basis (e.g. yearly) and occasionally in presence of unexpected events.

2.8. Big Data modelling

2.8.1. Description

This step requires the identification of the most appropriate econometric technique when building up a nowcasting exercise with Big Data. It is important to be systematic about the correspondence between the nature and the size of the selected Big Data and the method that is used. There are a number of dimensions along which we wish to differentiate.

In the first one we address the choice between the use of methods suited for large but not huge datasets, and therefore applied to summaries of the Big Data (such as Google Trends, commonly used in nowcasting applications), or of techniques specifically designed for Big Data. For example, nowcasting with large datasets can be based on factor models, large BVARs, or shrinkage regressions.

Huge datasets can be handled by sparse principal components, linear models combined with heuristic optimization, or a variety of machine learning methods, such as LASSO and LARS regression which, though, are generally developed assuming i.i.d. variables. It is difficult to provide an a priori ranking of all these techniques and there are few empirical comparisons and even fewer in a nowcasting context, so that it may be appropriate to apply and compare a few of them for nowcasting the specific indicator of interest. In absence of a multifrequency problem, those techniques can work for variable selection or data reduction as well as for the estimation of the nowcasting of the target variable.

In the second dimension we address the problem of the frequency of the available data. If this frequency is mixed, then specific techniques for mixed frequency data become relevant after having selected the variables or having reduced the dimension of the variables space accordingly with the techniques discussed above. Among mixed frequency models, UMIDAS stands out but also Bridge models can deserve a certain attention. UMIDAS provides a very flexible framework of analysis and can be adapted to work together with most if not all Big Data methods be they machine learning or econometric.

2.8.2. Recommendations

- In absence of any a priori information on the relative performance of various techniques, as many methods as possible should be evaluated and compared in a nowcasting context in order to select the best performing one.
- Alternative modelling strategies should be compared also by looking at the balance between their complexity in computational terms and their empirical performance.
- In case of mixed frequency data, linear methods such as UMIDAS and, as a second best, Bridge, should be privileged.
- Forecast combination and model averaging techniques, also when the mixed frequency aspect is present, can be used as an alternative to a large-scale comparison among competing techniques.

2.9. Results evaluation of Big Data based nowcasting

2.9.1. Description

The final step consists of a critical and comprehensive assessment of the contribution of Big Data for nowcasting the indicator of interest. This should be carried out within a real-time or a pseudo-real time exercise. In order to avoid, or at least reduce the extent of, data and model snooping, a cross-validation approach should be followed, whereby various models and indicators, with and without Big Data, are estimated over a first sample and they are selected and/or pooled according to their performance, but then the performance of the preferred approaches is re-evaluated over a second sample.

This procedure provides a reliable assessment of the gains in terms of enhanced nowcasting performance from the use of Big Data. For some critics about the usefulness of Big Data see Hartford (2014) and Lazer et al. (2014).

2.9.2. Recommendations

- Conducting an in-depth real-time or pseudo real-time simulation of competing models in order to evaluate their relative performance in nowcasting the variable of interest.
- Models including Big Data should be preferred when they significantly lead to an improvement of the reliability and accuracy of the nowcasting at the same point in time.
- Models including Big Data should also be preferred when they allow for timelier nowcasting without any significant loss in terms of reliability and accuracy.

2.10. Implementing Big Data based nowcasting

2.10.1. Description

In case the in-depth comparative analysis carried out in the previous steps suggests that the use of Big Data can improve the nowcasting for a given variable of interest, they can be then implemented. At this stage, the institution in charge of producing nowcasting should take several relevant decisions related to the number of the nowcasting to be implemented and their scheduling. For example, it is possible to decide to publish just one nowcasting (e.g. at the very end of the reference period or at the very beginning of the following one), to produce two nowcastings (e.g. on in the middle of the reference period and one at the very end), or to produce a sequence of nowcasts scheduled at weekly or even daily frequency. Such decisions should take into account, among other, the trade-off between timeliness and reliability, the user needs as well as some more institutional considerations.

2.10.2. Recommendations

- Implementing and publishing the most reliable nowcasts available either at the end of the reference period or at the beginning of the following one.
- Moving towards a daily or weekly update on nowcasting already during the reference period, only after detailed pros and cons analysis and a consultation of the most relevant stakeholders.
- The new Big Data based nowcasting should be accompanied by clear metadata and widely available reference and methodological papers.

3. Conclusions

In this paper, we have proposed a new operational step by step approach for using Big Data in a nowcasting exercise. It aims to facilitate the activity of experts involved in the construction of nowcasting by providing a set of recommendations associated to various operational steps.

References

Hartford, T. (2014, April). Big data: Are we making a big mistake? Financial Times. Retrieved from 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.