2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8364

Big data and official data: a cointegration analysis based on Google Trends and economic indicators

Crosato, Lisa; Mariani, Paolo; Marletta, Andrea and Zavanella, Biancamaria
Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy

Abstract

In this paper the relationship between the Industrial Production Index (IPI), the confidence index for the manufacturing sector and Google searches for several words linked to the economic situation is explored. In particular, time series referred to the period January 2004 - September 2016 on Italian data. An analysis of significant correlations between the selected indicators is achieved to explore the probable comovements of same. Adding one observation at a time since the first forewarning signs of the 2008 crisis, we find that a few Google searches and the IPI cointegrate, particularly during the strong downward trend leading to January 2009, while there is no cointegration between confidence indicators and the IPI. These results suggest that searches in google and the IPI or the confidence indexes are influenced by common circumstances. Finally forecasts of the IPI obtained through VECM models suggest that the evolution of the IPI can be well represented using the real time Gtrends selected variables.

Keywords: *Big Data; Google trends; Confidence indicators, Cointegration*

1. Introduction

The Industrial Production Index (henceworth, IPI) is one of the main monthly indicator attesting the current health of a country's economy. Accordingly, several contributions in the literature proposed to forecast it usually imputing hard data as regressors, from macroeconomic variables to business-specific indicators (Bodo and Signorini, 1987; Bruno and Lupi, 2004; Hassani et al., 2013). Soft data, such as text analysis in media and other sentiment indicators were introduced instead by Ulbricht et al. (2016) to predict the German IPI with more than 17,000 models.

The degree of novelty of the paper consists in the combined use of data coming from hard and soft data. The basic idea is to analyse the comovements of time series related to the general and personal economic situation achieved by different sources. The main goal of the paper is to understand whether web based soft index numbers together with confidence indicators may help in predicting the hard IPI. In their work, Ulbricht et al. (2016) showed that when it comes to the forecast of industrial production models using media data clearly outperform models without media data. Here the aim is to understand whether even models considering Gtrends data performed better than those taking into account only hard data.

The empirical strategy of the paper is to proceed by subsequent selection of variables. Firstly, the selection is obtained by simple visual inspection on the range of variability; secondly it is realized a correlation analysis with the IPI. If the correlations between the IPI and the soft indicators is significant, it is possible to represent this relationship through timeseries modeling. Last selection step of indicators is to exclude the stationary ones in order to proceed to the final cointegration analysis. Finally the presence of more than one cointegration relationship among time serie is tested, to end up with VECM based short term forecasts of the IPI.

The paper is organized as follows. After a brief introduction, data are shown in Section 2, in Section 3 methodology about time series and the choice of selected indicators is presented; finally main results and conclusions are discussed in the last part of the paper.

2. Data

This paper makes use of three data sources, two of which official and a third one non-official. The first is the Industrial Production Index, monthly released by ISTAT (Italian Institute of Statistics) with two months of delay with the reference period. The IPI is a 2010 fixed base Laspeyres index and is the main conjunctural indicator measuring real output for all facilities located in Italy.

The second data source is the Italian confidence index for manufacturing, monthly released by ISTAT with about 15 days of delay with respect to the interviews. In particular, here data refer to opinions on current level of orders, current economic situation, future level of orders and future economic situation.

The third data source we use is Google Trends, a free tool by Google showing the interest of some keyword during time. It allows to monitoring tendencies about a topic detecting the search frequencies on the web. Typing the keyword (or the topic), it is possible to extract frequencies and trends.

The economic literature has been using Google trends since its appearance in 2004 (see Hassani and Silva (2015) for a recent review on forecasting using Big Data). Google trends data are released as monthly frequencies of searches starting from January 2004, therefore this is the initial date for all our time series. Since the interest of this paper regards in understanding whether Google searches can be considered and used as proxies of the IPI, the searched words in Google Trends are related to the economic situation.

The words we have searched for in Google Trends are economic crisis, recovery, GDP, gross domestic product, public debt, spread, recession, unemployment, employment, job. We also construct naive composite Gtrends indicators by summing up frequencies associated to related words so obtaining four more variables: Total cycle = economic crisis + recession + recovery, Total occupation = unemployment + employment + job, Total Debt = public debt + spread plus a mixed-up variable three words = economic crisis + unemployment + public debt.

The official statistics we use in the paper are all expressed as index numbers in base 2010, so in order to have a fair comparison, also the Gtrendsdata have been indexed to 2010. To this end, the single and composite words monthly frequencies were divided by the mean of 2010 respective frequencies.

3. Time series methodology and choice of the selected indicators

The time series from the three data sources here shown differ at least in two aspects. First, the IPI and the confidence indicators (when needed) are published already deseasonalized, while the Gtrends variables must be treated for seasonality. Therefore, the R-interface to X13ARIMA-SEATS method by the United States Census Bureau is applied. Second , they are released with different lags with respect to the date of the information they are referred to. Morevoer the IPI of two months earlier is available at the end of each month, while confidence indicators and Gtrends variables refer to the current month. Accordingly, the data matrix is shaped anticipating all confidence and Gtrends indicators by two months. All the time series thus obtained are represented in figure 1. A quick glance to the series reveals different degrees of variability among the time series, highlighting the structural difference among the indicators.

The flatter series is for sure the IPI, followed by the confidence indicators and the Gtrends variables. Gtrends variables are clearly more volatile and subject to sudden jumps in correspondance of particular events (for instance, see in figure 1 the spikes in economic crisis from spring 2008 onwards and of three words at the end of the Berlusconi Government in summer-fall 2011).

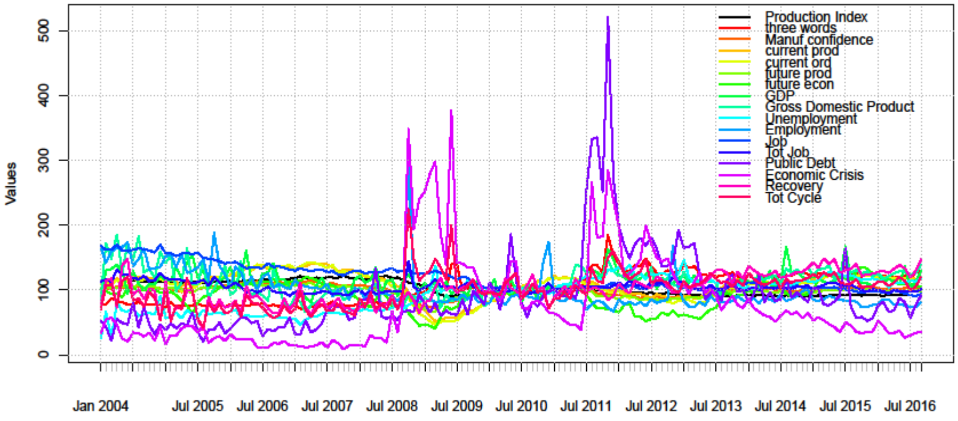

Figure 1. Time series of the selected indicators. Sources: ISTAT official statistics and our own elaborations on google trends data.

If the official sources (IPI and confidence index for manufacturing) and the non-official one (Google Trends) present common movements, it is more simpler to analyse the evolution of the phenomena involved. This could help in terms of prediction with the aim to move up the description of the economic conjuncture in comparison with official data. The cointegration analysis get back the idea according to two or more economic variables, although characterised by a different behaviour in the brief period, the could have some co-movements and tendency in the longrun period.

The final aim of the paper is to explore whether Gtrends variables and confidence indicators may show some predictive power on IPI. For this purpose it is used a multivariate time series model (VAR or VECM if any cointegration relationship appears). In particular, a forward approach is adopted adding one observation at a time from April 2008 onwards to monitor changes in the cointegration relationship during the observed period.

The selection process of the initial variables could be divided into three steps:

- removal of indicators showing too wide a range of variation (spread, recession and total debt);
- restriction to variables which correlate with the IPI and elimination of all variables showing no significant correlation relationship with the IPI and, among the remaining, those presenting a correlation coefficient lower than 0.3 (GDP, Gross domestic product, total job, public debt);
- test for the presence of Unit root in the series, as a preliminary information for the cointegration analysis. using a Phillips Perron test for unit root on the whole set of 100X12 series.

According to the selection process here described, the only variables not discarded are the threewords index (economic crisis + unemployment + public debt), job, economic crisis and all the confidence indicators.

4. Results

After the selection process of indicators correlated to the IPI, the next step consists in the cointegration analysis of the IPI index with one of the remaining variables in turn. Results of the Engle and Granger test for cointegration, reported in figure 2, point to no cointegration neither between the IPI and the confidence indexes, nor between the IPI and job. On the contrary, the IPI and three words do cointegrate and so do IPI and economic crisis, although there are some spikes when the turbolence in the two Gtrends variables is higher. The cointegration analysis between confidence indicators and threewords reveals a similar outcome.

This could be viewed as a first result of the paper contributing to define a selection strategy for Gtrends variables to increase forecasting models, although at present restricted to this particular case. If variables cointegrate when influenced by a common factor or by a combination of common factors, it might tentatively say that a few of the Gtrends variables and the IPI share some pattern drivers.

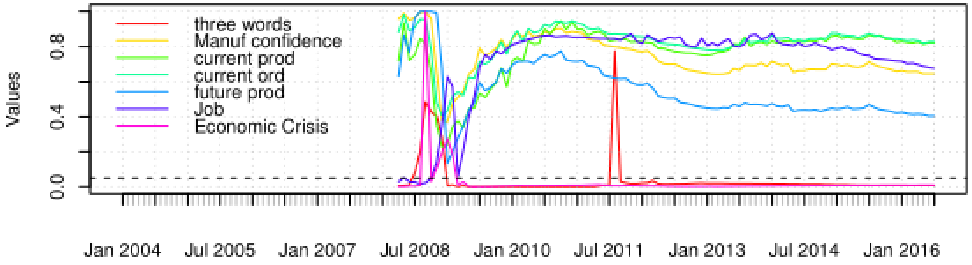

Figure 2. Engle and Granger cointegration test (p-values of the Phillips Perron test on residuals). Both tests are applied adding one observation at a time from April 2008 onwards. Sources: ISTAT official statistics and our own elaborations.

Another important result of this work can be obtained through a simple prediction based on a VECM model estimated on the IPI, the threewords index and one confidence indicator in turn. This is a way to measure the possible contribution to prediction of IPI by one or more confidence indicators and to better exploit pieces of information shared by Gtrends variables and the confidence indicators, although none of the latter cointegrate with the IPI. Again the VECM model was estimated 100 times on the month by month augmented time series, resulting in 100 forecasted values of the IPI, from May 2008 to September 2016 (see figure 3). The preliminary Johansen test confirms one rank of cointegration almost always for the confidence indicator on future orders and to a minor extent for the composite manufacturing confidence index, while the introduction of current orders or current production indicators seems to weaken the cointegration relationship between threewords and the IPI. Therefore, the 100 IPI forecasted values in figure 3 are obtained through VECM models based on IPI, threewords and the manufacturing confidence index or the confidence in future production. As can be seen predictions closely follow the actual values of the production index, in downward as well as in upward changes. The median percentage absolute error is smaller for the confidence in future production (0.9%) with respect to the manufacturing composite confidence (1.1%) mainly due to the protracted fall in the forecast for April 2009, when the IPI had already turned up. Note that these predictions are available two months earlier than the official IPI. For this reason, it is possible to claim that Gtrends data could be useful in prediction models to disclose movements of IPI when they are used in combination with hard data.

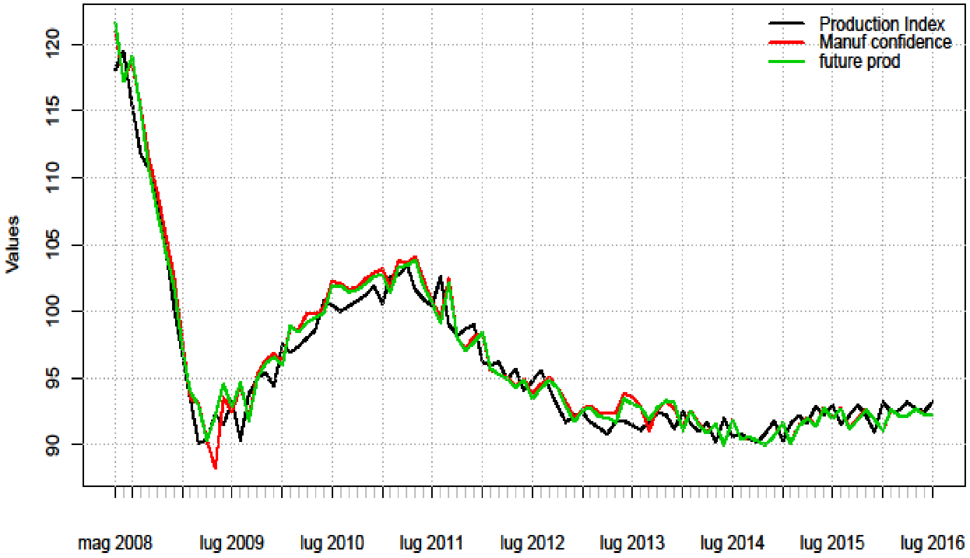

Figure 3. Recursive forecast of the IPI by VECM models using one of the listed variables together with the IPI and the three words Gtrends variable. Sources: ISTAT official statistics and our own elaborations.

References

Bodo, G., & Signorini L. F. (1987). Short-term forecasting of the industrial production index. International Journal of Forecasting 3(2), 245–259.

Bruno, G., & Lupi C. (2004). Forecasting industrial production and the early detection of turning points. Empirical economics 29(3), 647–671.

Hassani, H., Heravi, S., & Zhigljavsky A. (2013). Forecasting UK industrial production with multivariate singular spectrum analysis. Journal of Forecasting 32(5), 395–408.

Hassani, H., & Silva E. S. (2015). Forecasting with big data: A review. Annals of Data Science 2(1), 5–19.

Ulbricht, D., Kholodilin K. A., & Thomas T. (2016). Do media data help to predict german industrial production? Journal of Forecasting.