First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.4226

Nowcasting with Google Trends, the more is not always the better

Combes, Stéphanie[a] and Bortoli, Clément[b]
[a]Department of statistical methods, INSEE, France, [b]Department of conjoncture, INSEE, France.

Abstract

National accounts and macroeconomic indicators are usually published with a consequent delay. However, for decision makers, it is crucial to have the most up-to-date information about the current national economic situation. This motivates the recourse to statistical modeling to "predict the present", which is referred to as "nowcasting". Mostly, models incorporate variables from qualitative business tendency surveys available within a month, but forecasters have been looking for alternative sources of data over the last few years. Among them, searches carried out by users on research engines on the Internet – especially Google Trends – have been considered in several economic studies. Most of these exhibit an improvement of the forecasts when including one Google Trends series in an autoregressive model. But one may expect that the quantity and diversity of searches convey far more useful and hidden information. To test this hypothesis, we confronted different modeling techniques, traditionally used in the context of many variables compared to the number of observations, to forecast two French macroeconomic variables. Despite the automatic selection of many Google Trends, it appears that forecasts' accuracy is not significantly improved with these approaches.

Keywords: *nowcasting; Google Trends; macroeconomics; high dimension; machine learning; time series.*

1. Introduction

Official statistics are often published within irreducible delays[1]. But for decision makers, it is crucial to have access to the most up-to-date information about the current national economic situation. This is why developing some efficient forecasting tools and identifying the most relevant data sources are a serious issue in macroeconomics. Real time forecasting (or *nowcasting*) of macroeconomic indicators usually implies to incorporate variables from qualitative business surveys or sometimes financial variables. Over the last few years, forecasters have also been looking into data from the Internet and at trending searches made by Google users in particular.

In 2006, Google launched *Google Trends*, a tool that provides data series free of charge which reflect the interest of Internet users in a query or a set of semantically linked search terms. If this application has been popularized by advertising the most popular searches of the moment, it has also become a well-known source of data for economic studies. Characterized by their high frequency compared to official indicators and their short delay of publication (a week), they have been investigated by numerous economists over the last few years. Indeed, the global evolution of queries made by users about particular products or subjects via the search engine is likely to reflect the potential volume of sales of these products or the predominance of the subject for individuals at the time. These data could therefore be considered as indicators of consumer purchase intention or concerns (for example queries about unemployment benefit may give a hint of the evolution of the unemployment rate). Plus, the soaring penetration rate of equipment of households in computers and Internet connection makes them a credible source of information on individuals (less likely on companies).

The most famous use case of prediction with Google Trends is the Google Flu application developed by Google to forecast the spread of the flu epidemic in real time, based on user queries, in 2008. First launched in the United States, the tool was extended the following year to a dozen European countries, including France. In 2009, the group published an analysis of the benefits of using these series to forecast socio-economic indicators (Choi and Varian, 2009). According to this study, which used American data, forecasting automobile purchases, retail sales and purchases of dwellings could be improved by introducing this type of series into simple models using the dynamics of the series of interest (autoregressive model).

[1] In France, the main quantitative data available on household expenditure for example is the monthly household consumption expenditure on goods, published within one month and its equivalent in services is published within two months. Finally, an initial estimate of quarterly spending on all goods and services is published in the middle of the following quarter.

Askitas and Zimmermann (2009) used the frequency of use of certain search terms to forecast the unemployment rate in Germany; Kulkarni et al. (2009) suggested a link between the frequency of several search terms and housing prices in the United States; Vosen and Schmidt (2013) also used this type of series to forecast household expenditure in the United States.

Using Google Trends data in a variety of fields and incorporating them into more complex econometric models were also tested subsequently. It is along those lines that our study contemplates to contribute. Indeed Google Trends supplies a large pool of series that may convey useful yet hidden information. Automatic variables selection or extraction methods seem to match perfectly this situation where a lot of potential regressors are available, the number of observations is limited, and the expert may not want to constrain the specification of the model too much. In this study, we confronted several approaches used for forecasting in high dimension: variables selection techniques well-known in macroeconomics, variables extraction methods which aim at summarizing a large set of data in a smaller one, averaging methods to take into account the modeling uncertainty, and, eventually, non-parametric methods borrowed from machine learning, which appear to provide accurate predictions in numerous and various fields.

The next section describes more precisely our data and the treatments that were operated on them. Then, we remind our reader quickly with the concepts behind the different techniques we used, and eventually, we present and discuss the main results.

2. Data

The main attraction of the Google Trends data for the economic outlook lies in the fact that they can be mobilized quickly and at a higher frequency than most traditional economic series. Indeed, data related to one given week are published at the end of the very week. Data can also be filtered by geographic origin: we could therefore restrict our study to searches carried out in France. Available data are pretreated which means that raw series corresponding to the real frequency of use of a search term are not made public. Applied treatments are not very well documented but series are supposedly corrected accordingly to a trend resulting from an increase in popularity of the search engine itself. They are normalized too so that their maximum always equals 100, which means that they might be revised between one extraction at a certain date and another one later on and that direct comparison between two distinct series is not possible.

Google provides categories grouping queries by topics. More of one hundred of them are available organized in a three levels hierarchy. Normalization of categories differs from keyword's one: the frequency of the category in the first week of 2004 is used as a reference, the following points in the series are expressed as deviations from this level. Since the

meaning of a search term can evolve over time, it seems preferable to work on categories or concepts rather than on specific terms. Plus the strategy of choice of keywords would be very subject to subjectivity in addition to consequent manual task. For example, the "Sports" category aggregates all search terms linked with the field of sport. French Google users have shown an increased interest in this topic in the summers of even years (figure 1). Indeed searches related to sport showed a marked increase during the football World Cup 2006, 2010, 2014, the European football championships and the Olympic Games in the summers of 2004, 2008 and 2012. Purchases of televisions usually increase significantly at times of major sports events, so using the "Sports" category seems to be a natural choice to measure the degree of interest that a sports event can generate among French consumers.

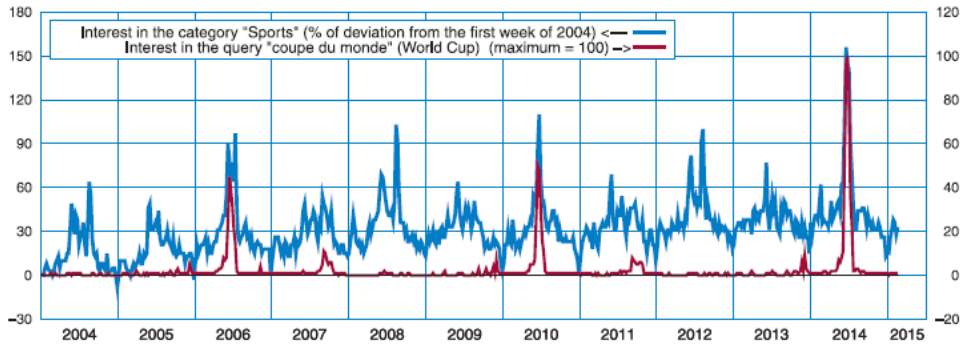

Figure 1. Examples of Google Trends chronicles for two chosen keywords. Source: Google Trends (2015)

In the context of our study, we selected a pool of 50 categories which may be correlated with the macroeconomic situation in one way or the other. The Google Trends categories were first transformed into a monthly format, weeks overlapping a month were distributed accordingly to the number of days in each month[2]. Series were then seasonnally adjusted; their monthly growth rates were computed to produce the explanatory variables, as well as their first time lag (i.e. the value of this growth rate in the previous month).

For this study, two targets have been considered: the household consumption in goods and the manufacturing production index. Representing more than half of GDP, household consumption is the largest item in final domestic demand, its estimate gives therefore a good outline of the whole activity. The first available data is the monthly household consumption expenditure on goods, published within one month. The publication of the manufacturing production index takes more time (two months), but its variation explains most of the quarterly GDP's evolution, it is then crucial to be able to produce accurate advanced estimates. In order to forecast these indices in real time or before they are published, usual

[2] This aspect makes it difficult to use techniques mixing data with different frequencies such as MIDAS (Mixed-data sampling), the advantage of the higher frequency was then not exploited here.

models incorporate variables from qualitative business tendency surveys available within a month. It seems reasonable to believe that the volume of queries made by users about particular products via the search engine could also reflect the potential volume of sales of these products, and, to a lesser extent, about production.

3. Methods

In this study, we confronted several approaches usually used for forecasting in high dimension: variables selection techniques well-known in time series, model averaging, and some machine learning techniques involving regression trees.

In dimension reduction problem, we can adopt two standard approaches: either we suspect that some variables are more important than the others, then the emphasis is put on identifying them; either we believe that some latent – unobserved – variables explain most of the comovements of the series considered altogether. In the first case, it is common to use iterative algorithms which add (respectively remove) a certain number of variables from an initial empty (respectively full) linear model, on the basis of a significance criterion. The main risk is to select a model which is far from the best possible model or to overfit, *i.e.* to adjust perfectly on observations used for estimation, which is generally associated with bad performance in forecasting new ones. More efficient approaches in terms of optimization have been developed such as penalized regressions - like LASSO (Tibshirani, 1996) or Elastic Net (Efron et al. 2004) - incorporating a penalty term in the objective function to favor parsimonious solutions. The main idea is to trade-off between the quality of the adjustment and some metric calculated on coefficients which prevents overfitting. In case the hypothesis of sparsity is challenged, variable extraction methods like principal component regression or partial least squares may reveal to be more efficient since they summarize the large set of data in a smaller set supposed to approximate the latent yet important variables.

When an estimator is the result of a model search amongst a collection of models in which multiple estimators are computed, one can potentially obtain an even better predictor by averaging these estimators for some selected models. In this respect, bayesian model averaging approach (Raftery et al., 1997) – BMA – combines multiple bayesian regressions weighted with their likelihood given the data (adjusted for complexity in a BIC criterion fashion) and a prior distribution on models (choosing here a binomial distribution with a probability lower than 0.5 for each variable to be included in order to search among parsimonious models).

Eventually, given the momentum of machine learning techniques in the context of a growing interest for “Big Data”, and their aknowledged performances in multiple and various fields, we also tried regression trees aggregation techniques like bagging and random forests.

Bagging (Breiman, 1996) consists in aggregating regression trees built on bootstrap samples (block bootstrap samples were used here to account for autocorrelation of time series). Any modeling technique can actually be bagged. Random forests (Breiman, 2001) introduce more randomness by sampling a set of regressors from the initial set of variables at each separation step of each trees. It entails more diversity in the aggregated trees since trees in bagging are usually very close since some variables get systematically selected. Boosting (Schapire et al., 1998) is quite different, it is an additive adaptive procedure which takes into account the biggest forecasting errors at one iteration when calibrating at the next iteration. This is done by actualizing some observations' weights.

To evaluate and compare the different approaches, we computed Root Mean Square Error (RMSE) in a pseudo real time fashion. After we fixed a first window - from 2004 to mid 2011; we proceeded for each month from end 2011 to end 2015 as follows: we extended the window by one month, estimated and calibrated the model, produced the one step ahead forecast and computed the forecast errors. The series of forecasts errors were eventually used to compute RMSE for each variable and methods.

4. Results

For households consumption in goods, a simple autoregressive model with one lag gets a RMSE of 0.56, we can see in Table 1 that this is hard to beat. For manufacturing production, given the publication delay, it makes no sense to compare to an AR(1) model, therefore we used the historical mean as benchmark (with a RMSE about 0.99). Again, the improvements are very limited. But it would be fairer to compare with models founded on business tendency surveys.

Table 1. Best performances (in RMSE/ out RMSE) obtained for different methods.

Variable	Stepwise	E. Net	BMA	Bagged E Net	Bagging	RF
Households consumption	0.48/0.52	0.53/0.53	0.49/0.52	0.58/0.53	0.61/0.56	0.54/0.53
Manufacturing production	0.88/0.99	0.90/0.99	0.88/0.97	0.91/0.98	0.82/0.92	0.64/0.98

Source: Bortoli & Combes (2016).

In each case, it seems not possible to expect more than 5-10% improvement. For households consumption forecasts, there is no clear winner. Selected Google Trends by stepwise selection or Elastic Net are multiple and diverse but the most contributing variable remains the lagged target. For manufacturing production, best performances are obtained for bagging of regression trees, other approaches don’t really compete with the historical mean.

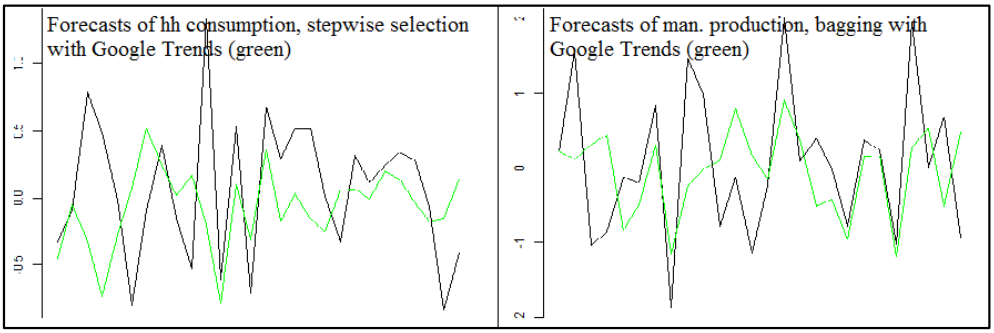

Figure 2. Examples of forecasts for households consumption and manufacturing prod.. Source: Bortoli & Combes

5. Discussion

Several reasons may explain these mitigated results. First, Google Trends series are very short, the comparison and evaluation protocol may suffer from the small number of observations and conclusions discussed here must be considered with caution. Plus, targets used in this study are not exempted of flaws. Household consumption in goods, for example, exhibits almost the properties of a white noise. Whatever the model or inputs, this series will remain hard to forecast. On the other hand, results for the purchase of certain goods (especially clothing and household durables) are more positive and some Google Trends categories seem to be probable explanatory variables (Bortoli and Combes, 2015). But expecting to get better forecasts for aggregated variables with a wide range of series thanks to automatic methods without any human intervention may be too naive. This may well be one limit to the attractiveness of "Big Data".

As far as Google Trends is concerned, it is worth noting than if they were to be used in the context of recurrent and official forecasts, we would not be in a position to judge the way these categories are built. Indeed their composition is unknown and we couldn't guarantee that the volume of searches is always enough to make statistical assumptions. Their composition may also change over time, especially when a popular new query appears at a given date. In addition, the series provided are the result of random sampling and can therefore differ from one data extraction to another. Repetitive forecasts founded on different extractions of the data will impose to re-estimate models systematically. The lack of transparency about treatments processed or sampling is one of the serious weaknesses of this tool, even if it proves to be more effective in another application than ours. From the official statistics' point of view, the sustainability use of the tool is also questionable. Indeed, Google Trends application is, by design, dependent on the technological developments in the search engine itself, continuously adapted to meet the needs of its users: the performance of the search engine and the underlying algorithms may evolve and lead to a change in the way in

which users use it (Lazer et al., 2014). Plus, since it was first created, the tool and the range of series available have changed substantially. Likewise, the free of charge access is based on the current marketing strategy of the company. Finally the behavior of users are continuously evolving, yet the quality of the data depends greatly on the individuals' habits of looking for information through the research engine. The growing share of smartphone applications could eventually lead to a reduction in the part played by search engines: the ability of trending searches to capture their behavior may decrease.

References

Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55(2), 107–120.

Bortoli, C., & Combes, S. (2015). Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées. *Note de Conjoncture, INSEE.*

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. *Wadsworth.*

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24 (2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, 5.

Choi, H., & Varian, H. (2009). Predicting the Present with Google Trends. *Technical report, Google*.

Claeskens G. (2012). Focused estimation and model averaging for high-dimensional data, an overview. *Statistica neerlandica*, 66(3), 272–287.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics,* 32(2), 407–499.

Kulkarni R., Haynes, K., Stough, R., & Paelinck, J. (2009). Forecasting housing prices with Google econometrics. *Research Paper, George Mason University School of Public Policy,* 457(10), 1012–1014.

Kunsch, H. R. (1989). The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics,* 17(3), 1217-1241.

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

Raftery, A., Madigan, D., & Hoeting, J. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association,* 437, 179– 191.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.*, 58, 267–288.

Vosen, S. & Schmidt, T. (2011). Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends. *Journal of Forecasting*, 30(6),565–578.

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*,26,1651-86