Assessing the use of Google Trends to predict credit developments[1]

Edwige Burdeau*
Banque de France, Paris, France – edwige.burdeau@banque-france.fr

Etienne Kintzler
Banque de France, Paris, France – etienne.kintzler@banque-france.fr

A growing part of literature on forecasting is devoted to the usefulness of Google searches data. The frequency of keywords searched on Google over a specific time period is made freely available in near real time through the web facility *Google Trends*. Besides, *Google Correlate*, an application returning the most highly correlated *Google Trends* with user-supplied time series, is gaining attention. For each series of interest, many potential predictors are thus available and can be used for forecasting purposes. In this context, an assessment of the value-added of *Google Trends* and *Google Correlate* to forecast French credit flows for house purchase is carried out. As the relationship between *Google Trends* and the variables of interest is not necessarily linear, the predictive power of such indicators can be improved by using machine learning methods and especially non-linear ones. Hence, different popular machine learning methods are tested: linear methods such as *Least Absolute Shrinkage and Selection Operator* (LASSO) and *Bayesian Model Averaging* (BMA) but also non-linear ones such as Boosting models and *Support Vector Machines* (SVM) with non-linear kernels. The predictive power of each model is assessed using out-of-sample forecasting errors and compared to a benchmark model that does not include *Google Trends* indicators. We find that the usefulness of *Google Trends* for forecasting purposes, several months in advance, is proved and to some extent does not depend on the model chosen. Besides, in some cases, non-linear models show higher predictive power than linear ones.

Keywords: Forecasting; Google Trends; Credit aggregates; Variable selection models.

Monitoring variations in credit flows to the real economy is crucial for a central bank. Indeed, the credit channel is a privileged way for a central bank to finance the economy: a central bank can accommodate its monetary policy by lowering key interest rates. Nevertheless, a lower key interest rate is not immediately passed on credit interest rates, since economic agents need several months to be granted credit such as credit for house purchase. In France, generally three months are needed and given to buyers to find their financing plan after having agreed on their preliminary sale agreement. In this context, a variation in key rates has an impact on credit flows several months after the monetary policy decision. Indicators such as *Google Trends* are made available in quasi real time and may bring early information on expected demand of credit as home buyers used the web to plan their purchase. The main purpose of this article is indeed to assess whether *Google Trends* is a valuable forecasting tool for credit for house purchase. In practice, the use of *Google Trends* raises some issues. In many studies, how the "best" *Google Trends* indicators are chosen is not documented. Furthermore, it is not obvious that one or two terms can be the best predictors of a time series and even if this is the case these terms may no longer be in the future as web users can change their habits. In this study, we propose a more robust way to identify indicators from *Google Trends*, by relying on *Google Correlate*. Besides, we test the predictive power of *Google Trends* with different models and especially variable selection models to evaluate whether our results are model dependent and if considering non-linear effects can improve our results. While variable selection models are not crucial when the set of variables is kept reasonable, these models allow to properly identify a parsimonious set of relevant indicators from a large dataset. Finally, we test the predictive power of our *Google*

[1] This article reflects the opinions of the authors and do not necessarily express the views of the Banque de France.

Trends indicators on both net credit flows and flows of new contracts of credit for house purchase, as repurchase agreements can alter the power of *Google Trends* to forecast correctly net credit flows for house purchase. From these various experiments, it appears that the use of *Google Trends* can help forecast credit flows several months in advance and this property seems not to be model dependent. Finally, our first experiment of a nonlinear model is conclusive, but further efforts should be devoted to this question in the future to identify properly nonlinear patterns in these time series.

1. Google Trends in the literature

Through *Google Trends* and *Google Correlate*, Google gives to anyone the opportunity to analyze trends in web search queries in quasi real time. For most popular terms, Google Trends returns a weekly or monthly time series of the number of queries normalized by the overall number of queries within a geographical area and a time range. Query indexes start in January 2004. Google emphasized in its Google Correlate white paper (2011), while you can easily pick interesting *Google Trends*, choosing the right set of indicators is far from being trivial. In this context, Google developed the application *Google Correlate* which "allows for automated query selection across millions of candidate queries" for any temporal pattern. In practice, Google Correlate can be used in two ways. The user can upload a weekly or monthly time series of her choice; in return the application gives the most highly correlated *Google Trends*. In the other way, the user can identify up to one hundred terms Google queries highly correlated with a pre-specified individual query. In this way, *Google Correlate* gives the opportunity to identify a family of terms not far from those obtain with natural language processing methods.

Since its first release, *Google Trends* indicators have become widely known to have nowcasting and in some cases forecasting properties. The preliminary work of Choi & Varian (2009) emphasized the variety of time series nowcasts which can be improved with *Google Trends* indicators, such as auto sales, retail sales or housing starts. For the housing sector, Wu & Brynjolfsson (2015) illustrated the usefulness of pre-defined Google Trends categories on real estate to nowcast and forecast home sales and house prices in the US. Askitas (2015) also made use of queries classified in the Google Trends category "Real Estate" to nowcast house prices in the US while Chauvet, Gabriel & Lutz (2016) built a Mortgage Default Index Risk to anticipate mortgage delinquency indicators. Finally, Coble & Pincheira (2017) showed that Google Trends can also help to forecast building permits in the US. We contribute to this strand of the literature by forecasting credit for house purchase. *Google Trends* can bring early information on this specific topic: future buyers use web search engine to estimate the cost of credit, find potential houses, or evaluate banks' offers. Besides, the French central bank collects early information from the banks' side aggregated in the *Bank Lending Survey* but few qualitative indicators from the households' side of credit for house purchase are available.

Furthermore, although a large part of the literature has been devoted to simple models reducing the use of *Google Trends* as extra explanatory variables in autoregressive models, some papers has enriched the literature. To nowcast German private consumption, Schmidt & Vosen (2011) reduced the dimension of their initial set of exogeneous variables by taking the principal components of various *Google Trends* categories. Scott & Varian (2012) selected a large set of indicators with *Google Correlate* as an input in a Bayesian variable selection model. Koop & Onorante (2013) emphasized the capacity of *Google Trends* indicators to detect turning points, putting into light that the forecasting property of *Google Trends* is not necessarily linear. In line with this literature, we work on many Google Trends indicators selected by *Google Correlate* and summarized them in principal components. Besides, we use variable selection models to curb the noise of non-relevant indicators and test a non-linear method.

2. Description of the set of models

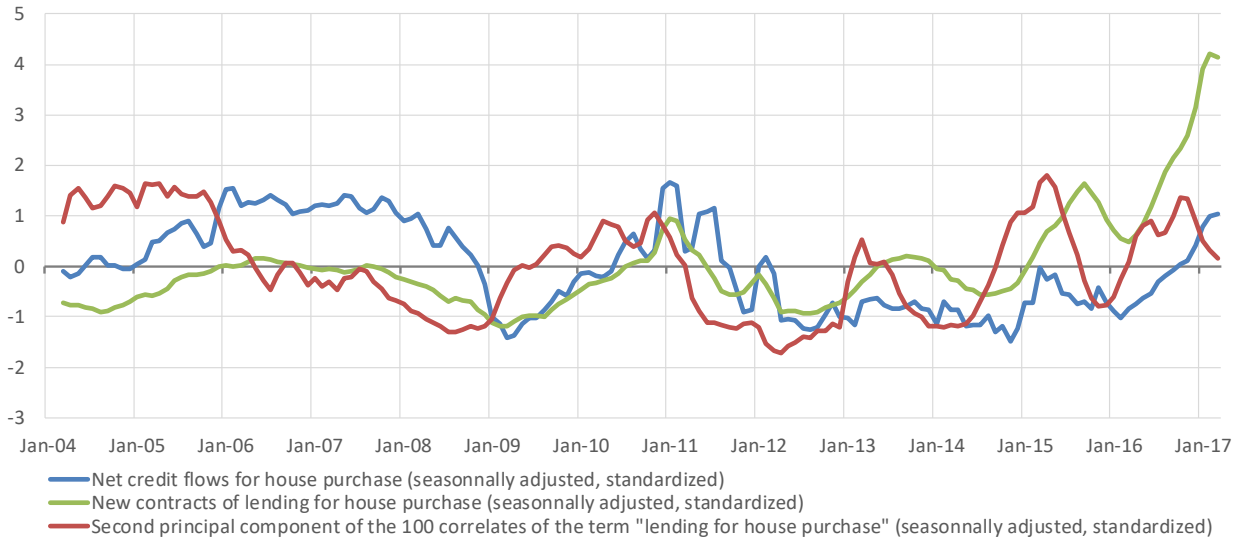
To forecast credit flows for house purchase, we choose to test recent machine learning methods which can deal with a large set of explanatory variables. Indeed, some articles in the forecasting literature, such as Bai & Ng (2008, 2009) et Kim & Swanson (2016), highlighted the value-added of models from the machine learning field.

Within the field, LASSO and Elastic net models introduced by Tibshirani (1996) and Zou & Hastie (2005) are well-known linear models in which a penalized term is added to the cost function to shrink variable parameters, especially those of irrelevant variables. LASSO is a special case in which parameters can be shrank to zero. A shrinkage parameter, reflecting the severity level of the penalization, must be calibrated, generally by cross-validation. While initially used for classification problems, boosting methods and especially component-wise L_2 boosting methods introduced by Buhlmann & Yu (2003) and Buhlmann & Hothorn (2007) can be used for regression problems. These models are estimated recursively, at each step, the estimator is incremented with an elementary function of one variable, called a *base learner*, which minimizes the cost function. In practice, the number of steps needs to be calibrated also by cross-validation. The base learner can be a linear or a spline component. In our case, each type of base learner is tested, but only the linear component give interesting results. Bayesian model averaging or BMA explained in Raftery & al. (1997) estimates linear models of all combinations of the initial set of variables or a sample if the initial number of variables is too high. The final model is the weighted average of all these models. The weight of each model stems from the posterior model probability obtained from the Bayes' theorem. For this purpose, a prior probability for each model, reflecting how *a priori* each model is the good one, needs to be specified. In our case, different models of prior probabilities, uniform, random, or fixed by cross-validation, were tested. Best results were obtained with the uniform one. Finally, we extended this work by testing a support vector machine (SVM) specification. Initially defined by Vapnik et al. (1992) to handle classification problems, support vector machine specifications have been adapted to solve regression problems. Generally considered as a nonparametric technique, SVM specification relies on kernel functions fitted on a finite number of points. In this study, we used the polynomial kernel. Unlike previous models, this model is not a variable selection model but can be calibrated to avoid overfitting. This model is an attempt to see whether nonlinear models should be further investigated.

3. Our dataset

Our time series of interest, net credit flows for house purchase and new contracts of lending for house purchase are extracted from Webstat, the Banque de France statistical website. No further transformations are needed; the monthly time series publicly made available is already seasonally adjusted. For *Google Trends* time series, *Google Correlate* is used in two ways. A first set of indicators is obtained by firstly selecting the 100 terms the most highly correlated with the French translation of the terms "credit for house purchase" and "lending for house purchase". The terms obtained are clearly related to distinctive features of a house purchase, such as terms related to credit interest rates, credit insurance, or names of credit institutions. Then, we extract the 5 first principal components computed from this set of variables and differentiate the non-stationary ones. We tried to include 10 principal components but the supplementary principal components did not show interesting patterns. While the first principal component is trending upward catching a common trend hardly interpretable, the following components, and especially the second one, reflect idiosyncratic trends of the housing sector. This second component seems to anticipate evolutions of credit flows for house purchase, and especially amounts of new contracts (Figure 1).

Figure 1: Evolutions of the net credit flows for house purchase, new contracts of lending for house purchase and the second principal component obtained with the 100 correlates related to lending for house purchase. Each time series was standardized and smoothed over 3 months to ease the reading



Sources : Banque de France, Google Trends

Google Correlate is also used to identify time series correlated with uploaded external time series on credit for house purchase. In this study, only outstanding amounts and annual growth rate of credit for house purchase are correlated enough with some Google Trends to give interesting results. Besides, within the set selected by *Google Correlate*, only few indicators are relevant. We choose to keep 12 *Google Trends* indexes related to terms containing names of French credit institutions. *Google Trends* indexes considered in our models are seasonally adjusted and consider in first differences for non-stationary indicators at a 10% threshold. The final set of exogenous variables is composed of both principal components of the first set and Google Trends indexes of the second set.

4. Our main results

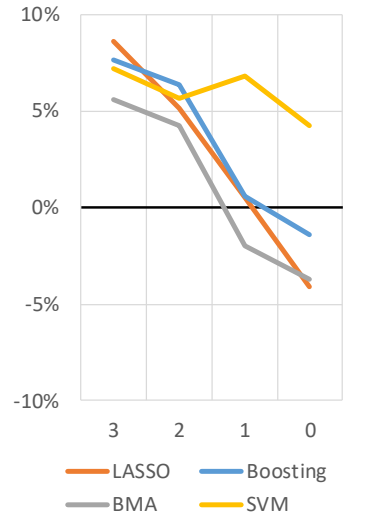
The objective of this study is to evaluate the forecasting power of *Google Trends*. For this purpose, we estimate, for the two variables of interest, the different models mentioned previously, the LASSO specification, the Elastic Net one, the L_2 Boosting approach, BMA approach and SVM model with a polynomial kernel. For each monthly time series of interest, we run 4 sets of estimations, depending on the number of months h in advance, from 3 to 0, of the forecast. For each set of estimations, we predict the variable of interest in month m with information available h months earlier, the observed value of the dependent variable in m-h-1, m-h-2, m-h-3, and monthly *Google Trends* indicators for the month m-h. To assess the predictive power of our models, out-of-sample monthly forecasts from January 2012 to December 2016 are computed for all models. These out-of sample forecasts are obtained recursively, for each month within this time range, we re-estimate and re-calibrate each model by considering a dataset from February 2004 to the last observation date and forecast the next observation. From these forecasts, the predictive power of each model is estimated using the root mean squared error (RMSE) on the forecasting period. To benchmark these results, we also estimate out-of-sample forecasts on the same period range with only lagged values of the variable of interest, available at the time of the prevision; this model is designated by the abbreviated term “AR”.

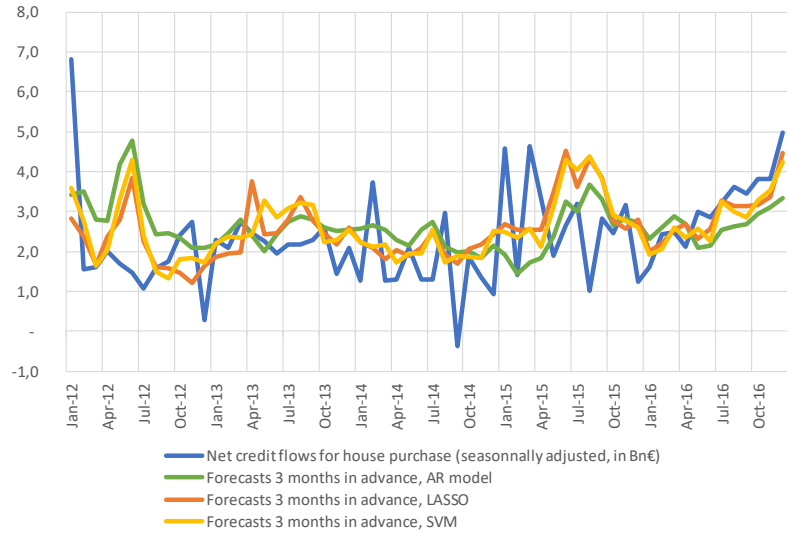
Results for net credit flows for house purchase and new contracts of lending for house purchase are shown in figures 2 and 3. In this study, new contracts are not taken in first differences even though we cannot reject the unit root hypothesis when considering credit evolutions in 2016. Nevertheless, we

decided to show this result as this can illustrate the higher predictive power of *Google T*rends for new contracts, especially in 2013, rather than for net credit flows. Indeed, net credit flows do not include renegotiated loans, while some Google queries are clearly related to this type of contracts. From these results, it appears that Google Trends have limited nowcasting power for both variables of interest; gains in RMSE are of small magnitude, generally below 5% (Figures 2a and 3a). But, *Google Trends* indicators appear to have forecasting power for both time series, especially several months in advance. In this case, gains in RMSE are quite substantial for every model and every time series, from 5% to more than 20% in one case (Figures 2a and 3a). Besides, it appears that *Google Trends* indicators bring substantial information to better catch cycle upturns and downturns, especially in 2012, during the first semester of 2015 and in 2016 (Figures 2b and 3b). The best predictor from *Google Trends* identified by all models is indeed the second component of the 100 correlates displayed in figure 1, the parameter associated with this variable is positive for all models and of the same order of magnitude than the autoregressive term. Finally, considering nonlinear features with the SVM can significantly improve forecasting results.

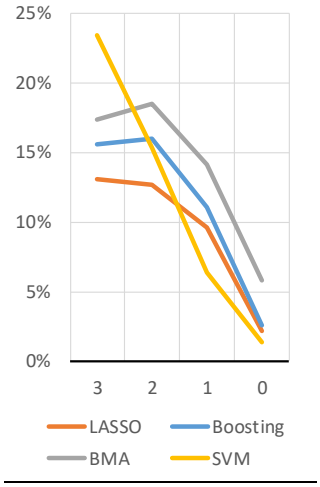
Indeed, our experiments confirm that using both *Google Trends* and *Google Correlate* for forecasting purposes of credit flows of loans for house purchase can significantly improve our forecasts on a medium term. Thanks to a simple approach using *Google Correlate* to identify a family of terms and a principal component analysis to summarize information, we obtained a robust indicator of future evolutions of credit for house purchase. From these results, it appears that some aspects need further investigations. Firstly, *Google Trends* may be a valuable tool to catch medium frequency cycles better than high frequency ones, analyzing *Google Trends* within a frequency approach could open new horizons. Secondly, as *Google Trends* indexes appear to be more interesting in level than in first differences, models coping with non-stationarity should be privileged. Finally, our first experiment with a nonlinear specification gave satisfactory results, which advocate for the use of other nonlinear methods.

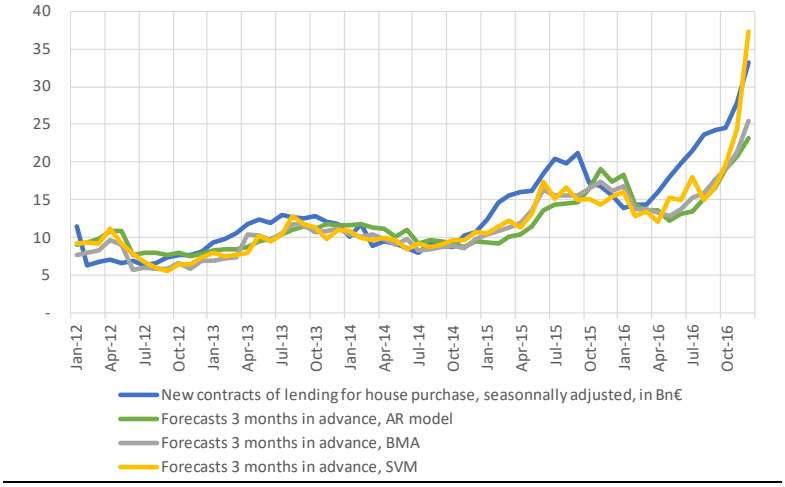
Figures 2a and 2b: Net credit flows of lending for house purchase – Gain in RMSE compared to the AR model, in function of the number of months in advance forecasts are estimated, for each type model (only the best models of each type are presented) and Comparisons between the variable of interest and out-of-sample forecasts 3-months in advance for different type of models (r.h.s.)





Figures 3a and 3b: New contracts of lending for house purchase – Gain in RMSE compared to the AR model, in function of the number of months in advance forecasts are estimated, for each type model (only the best models of each type are presented) and Comparisons between the variable of interest and out-of-sample forecasts 3-months in advance for different type of models (r.h.s.)





References

Askitas, N. (2015). Trend-Spotting in the Housing Market. IZA Discussion Paper No. 9427.

Bai, J., &Ng, S. (2008). Forecasting economic time series using targeted predictors. Journal of Econometrics, Elsevier, vol. 146(2), p. 304-317.

Bai, J., &Ng, S. (2009). Boosting diffusion indices. Journal of Applied Econometrics, John Wiley & Sons, Ltd., vol. 24(4), p. 607-629.

Boser, B. E., Guyon I. M., &Vapnik V. N. (1992). A training algorithm for optimal margin classifiers. Proc. 5th Annu. Workshop on Comput. Learning Theory, ACM Press, p. 144-152.

Buhlmann, P., &Yu, B. (2003). Boosting with the L2 Loss: Regression and Classification, Journal of the American Statistical Association, 98, issue, p. 324-339.

Buhlmann, P., &Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting. Statistical Science, Vol. 22, No. 4, p. 477-505.

Chauvet, M., Gabriel, S., &Lutz, C. (2016). Mortgage default risk: New evidence from internet search queries. Journal of Urban Economics, 96, November, p. 91–111.

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

Coble, D., &Pincheira, P. (2017). Nowcasting Building Permits with Google Trends. MPRA Paper.

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H. & Kumar, S (2011). Google Correlate Whitepaper

Kim, H. H., &Swanson, N. R. (2016). Mining Big Data Using Parsimonious Factor, Machine Learning, Variable Selection and Shrinkage Methods. International Journal of Forecasting.

Koop, G., &Onorante, L. (2013). Macroeconomic nowcasting using Google probabilities. Mimeo

Raftery, A. E., Madigan, D. &Hoeting J. A. (1997). Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association, Vol. 92, n° 437, p. 179-191.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society Series B, 73, issue 3, p. 273-282.

Scott, S., &Varian H. (2012). Predicting the present with Bayesian structural time series. Tech. Google.

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

Wu, L., &Brynjolfsson, E. (2015). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. Economic Analysis of the Digital Economy, p. 89–118.

Zou, H., &Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B, 67, issue 2, p. 301-320.