**Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends**

Nuno Barreira, Pedro Godinho and Paulo Melo

## Abstract

What?

This work presents a study describing the use of Internet search information to achieve improved nowcasting ability with simple autoregressive models, using data from four countries and two different application domains with social and economic significance: unemployment rate and car sales.

What was found?

The results we obtained differ by country/language and application area. In the case of unemployment, we find that Google Trends data lead to the improvement of nowcasts in three out of the four considered countries: Portugal, France and Italy. However, there are sometimes important differences in the predictive ability of these data when we consider different out-of-sample periods. For car sales, we find that, in some cases, the volume of search queries helps explaining the variance of the car sales data. However, we find little support for the hypothesis that search query data may improve predictions, and we present several possible reasons for these results. Taking all results into account, we conclude that, when Google Trends variables are significantly different from zero in-sample, they tend to lead to improvements in out-of-sample predictive ability.

Why is this important?

The results can have implications for nowcasting, by providing some indications regarding the advantage or not of the use of search data to improve simple models and indirectly by highlighting the sensitivity of the approach to the actual country-specific base, nowcasting period and search data.

**Keywords** Nowcasting; Google Trends; Unemployment; Car Sales

The final publication is available at <https://link.springer.com/article/10.1007/s11066-013-9082-8>:

Barreira, N., Godinho, P., Melo, P. (2013). "Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends", *Netnomics*, Vol. 14, nº 3, pp. 129-165. doi: 10.1007/s11066-013-9082-8

Nuno Barreira

Faculty of Economics of the University of Coimbra, Avenida Dias da Silva, 165, 3004-512 Coimbra, Portugal

email: [nfs.barreira@gmail.com](mailto:nfs.barreira@gmail.com)

Pedro Godinho (corresponding author)

GEMF and Faculty of Economics of the University of Coimbra, Avenida Dias da Silva, 165, 3004-512 Coimbra, Portugal

email: [pgodinho@fe.uc.pt](mailto:pgodinho@fe.uc.pt)

phone number: +351+239790571

Paulo Melo

INESC Coimbra and Faculty of Economics of the University of Coimbra, Avenida Dias da Silva, 165, 3004-512 Coimbra, Portugal

email: [pmelo@fe.uc.pt](mailto:pmelo@fe.uc.pt)

# Introduction

In this paper we present a study that intends to show how the act of searching in the Internet can be used to provide additional information regarding user actions, preferences and intentions, which can in turn be used to better assess particular situations. Using Internet query data from Google Trends, we intend to determine if and how these data can be used to improve forecasting ability using simple models.

Knowing the evolution of key indicators is essential to many decisions. However, often the ability to make such decisions is hampered by the absence of timely information regarding either those indicators or their components. Often, relevant economic indicators are only available with lag, and good forecasts of their current values are essential for making better decisions. This has given rise to the concept of nowcasting: forecasts of current events for which data has not been revealed [4]. We intend to find out whether the search data can be used to improve the ability to perform nowcasting. To do so, limiting ourselves to using simple models, we will analyse whether it is possible to achieve an increase on the predictive accuracy of these models with the addition of search data provided by Google Trends.

We considered two domains (unemployment and car sales) and four different European countries (Portugal, Spain, France, and Italy). Although the four selected countries share some cultural and geographical characteristics (they all belong to the Southern/Western Europe), they use different languages and sport different historical patterns regarding unemployment and automotive preferences on brands and models. We created several simple forecasting models for each country, and verified whether the inclusion of selected Google Trends data improved the nowcasting accuracy.

While the usage of search data to improve forecasting is not new, we consider this work significant since it allowed us to validate previous findings in diverse geographical / linguistic / cultural settings and on different application domains. Doing this we consider to have validated the usage of search data as an additional forecasting data source in some settings, defined the cases in which these data seems to be helpful in increasing the forecasting ability of the models, and in general probed the advantages and limitations of using Google Trends search data in the considered settings. Moreover, we consider the work relevant since it addresses topical social and economic concerns, for which data may not be available in a timely manner (namely unemployment) and for which there is a pressing commercial concern (namely data regarding car sales).

We also considered that it would be interesting to analyse the consistency of the results, and that also led us to choose countries with some points of contact. In fact, most of the available studies about the impact of search query data in the nowcasting performance of econometric models concluded that adding Google Trends variables leads to improved predictive ability (see the literature review below). However, some studies reach different results (some examples can be found in [7, 9]). If we take into account that authors achieving improved forecasting ability with an exciting new set of data are expected to be more inclined to pursue their research and try to publish it, we believe it is of interest to compare the predictive ability of models that include search query data in different in countries for which we did not have reasons to expect radically different results. This analysis of the consistency of results was extended to the definition of different out-of-sample periods, in order to check whether results are consistent through time.

In the interest of simplicity, this work has not tested whether more complex models would better be able to use the search information to improve forecasting, or if instead the use of more complex models would make such information less valuable by improving the forecasting ability of the original data. The usage of simple models can present the advantage of allowing subsequent analysis to discover the mechanisms which explain the benefit of including or not a particular data source. This work, however, does not provide an in-depth analysis of those mechanisms, providing just simple reasoning (when appropriate) to illustrate them.

The main innovation claim of the work is that by performing cross national/cross linguistic study of different realities (which, however, show points of contact) and by analysing the predictive performance in different periods, this work may provide an idea of the consistency of predictive nowcasting improvements and point to the sensitivity of the approach to particular context characteristics. In fact, those characteristics may make the approach better or worse suited to different applications, or even unusable on some of them (as seen in the results presented).

The paper is structured as follows. After this introduction, we present a literature review of related work in Section 2. Section 3 describes the data that was used in this work. Section 4 presents the methodology that we followed in this study. Section 5 presents the results and Section 6 discusses the implications of the results. Finally, Section 7 presents our conclusions and proposes some lines of future work.

# Related work

One of the first applications that suggested that web search data was useful in forecasting economic statistics was presented in [12], which made an analysis of the relation between WorldTracker’s Top500 Keywords Report data and unemployment in the United States (US). The authors concluded that there was a positive, significant association between employment-related searches by Internet users and official unemployment levels disclosed by the US government and they also suggested the viability of using search terms to predict other important macroeconomic data.

More recently, several researchers have analysed the usefulness of data from Google Trends in nowcasting or forecasting the values of several variables. These variables include the incidence of epidemic diseases [15], private consumption [16, 19, 21, 22], retail sales [4, 7], home sales [7], visitor arrivals [7, 9], car sales and unemployment, among others. We will now focus on works that address unemployment and car sales.

Choi and Varian [8] analyse the impact of using the search queries in the "Jobs" and "Welfare & Unemployment" categories of Google Trends in predicting initial US initial claims for unemployment benefits. The authors conclude that adding these variables to a simple AutoRegressive (AR) model improves both the model fit and the out-of-sample Mean Absolute Error (MEA) of the prediction. In a later work that includes more recent data [9], the same authors reach quite different results: adding Google Trends variables increases the adjusted R2 of the model only slightly, and leads to worse out-of-sample predictions. The authors argue that in this case Google Trends data may still be useful for finding turning points, since the model that includes search query data seems to perform particularly well near in the periods surrounding turning points in the data. D’Amauri and Marcucci [11] suggest using a Google Index, based on Google Trends data related to Internet job search, to predict the US unemployment rate, and they show that popular time-series specifications augmented with this indicator improve their out-of-sample forecasting performance for different horizons. The authors consider a very large number of linear and non-linear models, and show that the best forecasts are obtained with models that use the Google Index as the leading indicator. The authors show that the Google Index performs better than the widely accepted Initial Claims indicator as a leading indicator of unemployment. The best model including the Google Index also outperforms the forecasts released in the Survey of Professional Forecasters conducted by the Philadelphia Fed.

Fondeur and Karamé [14] analysed the performance of Google search data in the prediction of youth unemployment in France, concluding that Google search data improve the youth unemployment forecasts. The authors point out that Google search data also improve unemployment forecasts for people belonging to other age ranges, but forecast improvements are not as large as in youth unemployment. Bughin [4] analysed the performance of Google search data in explaining changes in unemployment claims in Belgium. The author considers an AR model that also includes inflation and search query volume as dependent variables, and concludes that the search queries variable is statistically significant, explaining about 15% of the total unemployment spread. Chadwick and Şengül [6] examine the performance of the volume of Google queries in nowcasting the unemployment rate in Turkey. The authors conclude that the models that include the volume of search queries show a better nowcasting performance than a benchmark AR model. Suhoy [18] shows that the volume of search queries can be used to explain the changes in job openings in Israel, which can then be used to nowcast the unemployment rate.

Choi and Varian [7] use Google Trends data in the prediction of sales of different car brands in the US. The authors compare a simple benchmark AR model with two models that include Google Trends variables: a univariate model that assumes that coefficients are independent for all brands and a fixed effects model that assumes that coefficients of lagged sales and of search query data are the same across all brands, and differences in sales volume by brand can be captured by an additive fixed effect. The authors present results for three brands, Chevrolet, Toyota and Ford (for the latter brand, only the univariate model is considered). In the case on Chevrolet, the predictions produced by both models that include Google Trends data have a larger MEA than those of the benchmark model; in the case of Toyota, the predictions achieved with the fixed effects model are better than the ones obtained by the benchmark model, but the univariate model leads to worse predictions; in the case of Ford, the predictions obtained with the univariate model are slightly better than those of the benchmark model. In a later work [9], the same authors use a simple AR model and a model with Google Trends data to nowcast the ‘Motor Vehicles and Parts Dealers’ sales index. The authors conclude that adding Google Trends data leads to an improvement of about 10.5% in the MEA of predictions. Carrière-Swallow & Labbé [5] analysed the performance of Google Trends data in nowcasting car sales in an emerging economy: Chile. The authors conclude that models including Google Trends data always lead to better out-of-sample nowcasts than the benchmark AR or AutoRegressive Moving Average (ARMA) models they consider.

In the most recent years, several different studies address the impact of including Google search data in the prediction accuracy of nowcasts or short-term forecasts of economic variables, including unemployment and car sales. In most cases, the studies conclude that adding Google Trends variable improves the predictive performance of the models. In this study we will similarly analyse whether Google search variables improve the nowcasting accuracy of unemployment and car sales, for four countries of the south-western Europe.

# Data

## Unemployment

Seasonally adjusted data concerning monthly unemployment rates for Portugal, Spain, France and Italy were collected from the European Central Bank Statistical Data Warehouse website [13]. Monthly data from January 2005 to August 2013 were used, making a total of 104 observations per country. Data were collected on October 14, 2013. The corresponding variables are denoted by *U*C, where C stands for the country: P was used for Portugal, S for Spain, F for France and I for Italy.

For the Google searches, it was assumed that the most relevant searches would concern the unemployment and unemployment benefits. We considered that different countries might have different ways of referring to the relevant information about unemployment benefits, so the simple translation of the term used in a country might fail to capture the relevant searches in another country. So, we started by considering several different terms for each country: *subsidio desemprego* and *desemprego* for Portugal; *subsidio de desempleo*, *desempleo* and *prestacion desempleo* for Spain; *indemnités chomage*, *allocations chomage*, *allocations de chomage*, *chomage* for France; and *disoccupazione ordinaria*, *INPS* [[1]](#footnote-1) *disoccupazione*, *disoccupazione* for Italy. The searches were limited to the country being analysed and no category limitations were defined.

We started by collecting monthly data for the volume indexes of these queries, for the period between January 2005 and December 2012, and used them for some preliminary experiences, in order to get some idea of how the different terms performed in explaining and nowcasting unemployment rates. These preliminary experiences led us to conclude that the terms obtained by simply translating the word unemployment to the native language of each country (*desemprego*, *desempleo*, *chomage* and *disoccupazione*) consistently produced the best results.

Google employs a sampling procedure that introduces measurement error into the series (see, e.g., [5]). Requests for the same query made in different days are based on different samples and, therefore, they will lead to slightly different series. In order to reduce this measurement error we used an API developed externally to Google Trends to collect the data series for all the keywords we considered and for the period between the beginning of 2004 and the end of August 2013. Data were collected each day over a 14-day period (between October 4th 2013 and October 17th 2013), using this API, and the average of each observation over the two weeks was calculated (the procedure is similar to the one used by Carrière-Swallow and Labbé [5], although they used 50 days).

Just to get an idea of the possible impact of measurement error in unemployment-related terms, we calculated average standard deviation for the distribution of weekly observations obtained over the 14 day-period, for the terms *desemprego*, *desempleo*, *chomage* and *disoccupazione*. The average standard deviations ranged between 3.5% for *chomage* and 7.6% for *desemprego*. Using a similar procedure, Carrière-Swallow and Labbé (2011) report a standard deviation of 5.8% for the term "Chevrolet" in queries produced in Chile, near the middle point of the interval we obtained.

Two difficulties arose with this data collecting process:

* Automatic collection of data from Google Trends usually leads to weekly data series. In fact, collected data series have a monthly frequency for the term "*allocations de chomage*" and weekly frequency for all other terms. The data frequency of the collected series depends on the search volume of the different terms and cannot be controlled by the user. In this work we needed monthly series, since we had monthly series for the unemployment.
* Google Trends designates a certain threshold of traffic for search terms, so those with a volume below the threshold are not reported. Data collected by the API were mostly weekly data, and for some search terms the traffic threshold was not reached for some of the weeks (these values showed as zero values).

In order to address the second difficulty, we analysed which observations were missing from the data concerning each term, and we concluded that:

* For those terms that correspond to a direct translation of "unemployment" (*desemprego*, *desempleo*, *chomage* and *disoccupazione*), data were always available for all weeks from the beginning of 2005 to the end of August 2013.
* For all other terms, data were unavailable for some weeks or months of 2005, and in most cases data were also unavailable for some periods in 2006 and/or 2007.

Taking these facts into account, and also considering that the terms that correspond to a direct translation of "unemployment" always produced the best results in the preliminary analysis, we decided to use only the data concerning these terms, disregarding the other ones. We used, for all countries, data from the beginning of 2005 to August 2013. In order to address the need for monthly data, we calculated each monthly volume index as the mean of the daily observations expanded from the weekly data collected from Google Trends (this procedure is identical to the one used in [5]). So, we ended up with 104 monthly observations of a Google Trends query volume index for each country. This series was then scaled by dividing all values by the maximum of the series.

The Google Trends volume indexes thus obtained exhibited some apparent seasonality. We applied the ARIMA-X-12 procedure (see http://www.census.gov/srd/www/x12a/) to these series, in order adjust for seasonality. These seasonally adjusted data series were used in the estimated models, and the corresponding variables are hereafter denoted by *GU*C, where C stands for the country: P was used for Portugal, S for Spain, F for France and I for Italy.

## Car sales

Monthly data for the number of new passenger car registrations in the considered countries, working day and seasonally adjusted, were collected from the European Central Bank Statistical Data Warehouse website [13]. Monthly data from January 2004 to August 2013, collected on October 2, 2013, were used, making a total of 116 observations per country. Data were scaled by dividing each observation by the maximum value of the corresponding series. The corresponding variables are denoted by *S*C, where C stands for the country: P was used for Portugal, S for Spain, F for France and I for Italy.

For the Google searches, it was assumed that potential buyers would directly search for the car brands among which they intended to choose. So, for each country, we searched for the 20 car brands that achieved a largest market share in 2012, and used search queries concerning these brands. In all cases, these 20 brands represented more than 92% of the total car sales. For Portugal, Spain and Italy, data were gathered from associations or companies belonging to the car sales sector. For France, the information was gathered from a site that gathers worldwide information about car sales. The sources for this information were:

1. Portugal: AutoInforma [2].
2. Spain: ANIACAM [1].
3. France: BestSellingCarsBlog [3].
4. Italy: UNRAE [20].

Initially, we considered queries collected in three different contexts: queries included in all categories, queries included in the "Vehicle Brands" category and queries included in the "Vehicle Brands > Vehicle Shopping > Vehicle Specs, Reviews & Comparisons" category. Some preliminary experiences were then conducted, leading us to conclude that the use of queries without category constraints usually led to the best results in explaining and nowcasting car sales. Since, additionally, the API we used to systematically collect the data also prevented us from easily defining Google Trends categories, we ended up using volume indexes without category constraints.

The measurement error problem described before for the unemployment data, also arose in the case of car sales. So, we tried to handle it in a similar way: using an API to collect the data series for the period between the beginning of 2004 and the end of August 2013, during a number of consecutive days and using the average of each observation. Similarly to what was done with the query volume indexes for unemployment, data were collected each day over a 14-day period between October 4th 2013 and October 17th 2013, using the API, and the average of each observation over the two weeks was calculated. In the case of these car brand queries, data was available for all countries, for all weeks from the beginning of 2004 to the end of August 2013. Monthly volume indexes for the months between January 2004 and August 2013 were calculated as the mean of the daily observations expanded from the weekly data collected from Google Trends. Data series were scaled by dividing all observations by the maximum value, and seasonally adjusted by using the ARIMA-X-12 procedure. These seasonally adjusted data series were used in the estimated models, and the corresponding variables are hereafter denoted by *GS*C, where C stands for the country.

# Methodology

We defined two goals for the empirical work:

1. To determine whether the use of the volume of search queries improved the ability of simple models to describe the behaviour of unemployment and car sales.
2. To determine whether the volume of search queries improved the ability of simple models to nowcast unemployment and car sales, in different periods.

The broad lines of the procedure we used to address the first goal were the following:

* Estimation of a simple benchmark model, using the all available data (base model).
* Estimation of a second model including variables related to the quantity of search queries (extended model).
* Comparing the two models according to the adjusted R2 (that is, the adjusted coefficient of determination) and analysing the statistical significance of the variable quantifying search queries, in the second model.

The models were estimated with both the base values of the variables, their logarithms and a logistic transformation of the variables. The models estimated with the base values of the variables will be hereafter referred to as additive models and the models with the logarithmized values will be referred to as multiplicative models. The models that used logistic transformation of the variables always led to conclusions similar to either the additive or the multiplicative models. Therefore, we decided not to include the results obtained with the logistic transformations in the paper.

The stationarity of unemployment, car sales and search queries was checked before defining the models, using the Augmented Dickey–Fuller (ADF) test, with lag length determined according to the Bayesian Information Criterion (BIC). The test was applied both to the values of the variables and to their logarithms, and a 5% p-value was used as a significance threshold. In almost all cases it was concluded that these variables were non-stationary and the variables were differentiated in order to become stationary. In a few cases, it was necessary to differentiate a variable twice in order to achieve stationarity.

For the base model, we considered the use of an ARMA(*p*,*q*) model, and the application of the Hannan and Rissanen (1982) procedure to define the AR and Moving Average (MA) orders of the model. Preliminary experiments showed that the MA terms ended up being removed, both in the unemployment models and in the car sales models. So, it was considered that an AR(*p*) model would be enough, and the MA component was excluded.

For the extended model, the contemporary Google Trends variable was included, and some of its lagged values were also analysed and included if statistically significant. When both the dependent variable and the Google trends variable were integrated of the same order, the Johansen test was performed in order to determine whether the variables were cointegrated. When this was the case, a variable corresponding to the imbalance of the previous period regarding the cointegration relation was included in the model.

Quandt Likelihood Ratio (QLR) tests for a structural break at an unknown point, with 15% trimming at the beginning and end of the sample period, were performed for both the base and extended models. Whenever a structural break was detected, all data up to the moment of the break was removed and the process was repeated – stationarity of the variables was reanalysed for the remaining period and the models were re-estimated.

Several diagnostic tests were performed to the extended model, in order to determine if anything suggested that the model was not properly specified. In particular, we looked for autocorrelation in the residuals (Durbin’s h test and Ljung-Box test with 6 and 12 lags), violations of the normality of the error distribution (Doornik–Hansen test), heteroskedasticity (White test with cross-products and Breusch-Pagan test), Autoregressive Conditional Heteroskedasticity (ARCH) in the residuals (a Lagrange Multiplier (LM) test with 6 and 12 lags) and misspecification (Ramsey RESET test). A 5% significance level was used for the analysis of the results of these tests.

We used the models thus defined not only to address the first goal of this work, but also as a starting point to the analysis of the predictive ability of the search query variables – that is, as a starting point for the second goal of this work. The broad lines of the procedure used for addressing this second goal were:

* Splitting the data series into two periods: the first one used for model estimation (in-sample period) and the second one for testing the predictive ability of the estimated models (out-of sample period).
* Estimation of a benchmark model with the same regressors as the benchmark model defined for the whole period, using only in-sample data.
* Estimation of an extended model with the same regressors as the extended model defined for the whole period, using only in-sample data.
* Comparing the out-of-sample predictive power of the two models, using static forecasts (i.e., forecasts that use the actual value for the lagged variable).

Three different data splits between in-sample and out-of-sample periods were used in order to analyse how consistent and stable would be the possible improvements of predictive accuracy. The in-sample period started at the beginning of the sample, if no structural break was detected, or at the first observation after the structural break, whenever the QLR test detected a structural break. Data split 1 considered an in-sample period ending in August 2010 and an out-of-sample period of September 2010 – August 2013 (36 data points). Data split 2 considered an in-sample period ending in August 2011 and an out-of-sample period of September 2011 – August 2013 (24 data points). Data split 3 considered an in-sample period ending in August 2012 and an out-of-sample period of September 2012 – August 2013 (12 data points). In order to get a better idea of the stability and consistency of nowcast improvements, in each data split we analysed not only the predictions for the whole out-of-sample period but also for a sub-period consisting of the first 12 months of the out-of-sample period.

Predictive accuracy was compared based on the Root Mean Square Forecast Error (RMSFE), and two statistical tests were performed. The first test is a one-sided F test of equal mean square error based on McCracken [17]. The second test is a forecast encompassing test to find out to whether or not the forecasts produced by the base model encapsulate all the useful predictive information contained in the forecasts of the extended model, based on Clark and McCracken [10]. Both these tests are defined for nested models, and so they are applicable to the models we are comparing, in which the base model is a restricted version of the extended model. These tests were applied using a procedure for RATS made available by Clark and Nakata (<http://www.estima.com/procs_perl/clarkforetest.src>).

## Unemployment

The ADF test led us to conclude that both the unemployment variables, , and their logarithms, , were integrated of order one for Portugal, Spain and France. In the case of Italy, we concluded that both the unemployment variable and its logarithm were integrated of order two. The ADF test was also applied to the Google Trends query volume variables and to their logarithms. We concluded that these variables were integrated of order one for Portugal, France and Italy and stationary for Spain.

In order to define the benchmark model, we estimated an AR(6) model for each country, and iteratively removed lagged values of the unemployment variable with coefficients that were not significantly different from zero, and always keeping the first lag in the model. Whenever diagnostic tests indicated it might be useful to add other lags to the model, they were also tested and kept if significantly different from zero. This was the case of the model defined for Italy, in which the Ljung-Box test with 12 lags led us to think that higher-order lags might be useful and, in fact, we ended up keeping the 11th lag of the unemployment variable. The base model can be defined as

, (1)

where *Lags-UC* is the set containing the orders of significant lags and the number one (the first lag was kept even if it was not significantly different from zero),  is the unemployment variable differentiated  times,  being the number of differentiations required in order to achieve stationarity (one for Portugal, Spain and France and two for Italy),  is the series of error terms and  are coefficients.

Another model was estimated, adding Google Trends-related variables to the base model: the extended model. In this model, we always introduced the contemporary Google Trends variable and also tested the first two lags of this variable, with the intention of keeping them in the model if significantly different from zero. Only in the case of the model defined for Spain was one of these lags significantly different from zero (the second lag). For Portugal and France we used to Johansen test to find out whether unemployment and the Google Trends variables were cointegrated (in the case of Spain and Italy the variables were not integrated of the same order). For both Portugal and France, we concluded that the variables were cointegrated and added to the model a variable corresponding to the imbalance of the previous period regarding the cointegration relation. The extended model is defined as

 (2)

*Lags-GUC* is the set containing both the orders of significant lags of the Google Trends variable and the number zero (the contemporary value of the Google Trends variable is always included in the model),  is the Google Trends variable differentiated  times,  being the number of differentiations required in order to achieve stationarity (one for Portugal, France and Italy and zero for Spain), *EC* is the imbalance regarding the cointegration relation (the term  was included only when significant evidence of cointegration was found, that is, only for Portugal and France),  is the series of error terms and  and  are coefficients. The other variables and coefficients are identical to the ones defined for the base model.

## Car sales

The ADF test led us to conclude that both the car sales variables, , and their logarithms, , were integrated of order one for all countries, so these variables were differentiated once. The Google Trends variables and their logarithms were found to be integrated of order one for Portugal, France and Italy and integrated of order two for Spain.

The initial models that were estimated showed very significant deviations from the linear regression assumptions, for all countries. For example, significant deviations from the normality of the residuals, autocorrelation of the residuals and heteroskedasticity were found in most models. An analysis of the data series led us to hypothesize that such deviations would be due to some sharp rises and declines arising in the car sales series, often a sharp rise immediately followed by a sharp decline or vice versa. Such sharp rises and declines are usually related to specific issues related to the country being analysed, which make it attractive to bring forward or postpone buying a car. Such issues may be related to specific promotions or tax-related (for example, related to changes in car taxation) and often occur at the year end. Using Portugal as an example, in Figure 5 we can see very sharp rises in December 2008 and December 2010, followed by sharp declines in the following month. These movements can possibly be explained by changes in the car sales taxation: in the beginning of 2009 there was an increase in the taxes on sales of diesel cars, and in the beginning of 2011 there was a general increase in the taxes on car sales.

In order to handle this problem, we added dummy variables for the periods when we detected very large changes in car sales, and kept them in the models when their coefficients were significantly different from zero. Since there were strong indications that car sales in January might vary in the opposite direction of car sales in December, for tax-related reasons, we considered the inclusion of a variable that might capture this effect. This variable, denoted by , was defined as the product of a dummy variable, which takes the value one in the January of every year, by the car sales variable corresponding to the previous month of December. If our analysis was correct, this variable should have a negative coefficient in the estimations. As we can see in the next Section, this was always the case.

Similarly to the procedure followed for unemployment, to define the benchmark model, we estimated an AR(6) model for each country, and iteratively removed non-significant lagged values of the car sales variable, always keeping the first lag in the model. The base model is defined as

, (3)

where *Lags-SC* is the set containing the orders of significant lags and the number one (even if the first lag is not significantly different from zero),  is the car sales variable differentiated once, *TS* is the set of time periods corresponding to observations with sharp changes in the volume of car sales, is a dummy variable that only takes the value one at time *i*,  is the series of error terms and ,  and  are coefficients. Variable  was described before, and it can be defined as

, (4)

where  is a dummy variable that takes the value one only if the observation corresponds to the month of January. The term  was only kept in (3) when  was significantly different from zero.

In order to define the extended model, we added Google Trends-related variables to the base model. In this model we always introduced the contemporary Google Trends variable. Lagged values of the Google Trends variable were never significantly different from zero, so they were not included in the extended model. The Johansen test showed that car sales and the Google Trends variable were cointegrated for France but not for Portugal nor for Italy (in the case of Spain the variables were not integrated of the same order). So, in the models corresponding to France, a variable corresponding to the imbalance of the previous period regarding the cointegration relation was added. The extended model is defined as:

 (5)

 is the Google Trends variable differentiated  times,  being the number of differentiations required in order to achieve stationarity (one for Portugal, France and Italy and two for Spain), *EC* is the imbalance regarding the cointegration relation (included only for France),  is the series of error terms and  and  are coefficients. The other variables and coefficients are identical to the ones defined for the base model.

# Results

In this section, we present the main results obtained using the methodology described before. The results are summarized in Tables 1-16, and will now be briefly analysed. For each country and for both unemployment and car sales, two tables are presented. The first one shows the most important results concerning the models estimated for explaining the behaviour of the variable being considered. The second table shows summarized data concerning the nowcasting ability of the models estimated for the three data splits.

## Unemployment

Tables 1 and 2 present the results concerning Portugal. The base data is presented in Fig. 1 in the Appendix. The volume of search queries and unemployment were found to be cointegrated and therefore the imbalance of the previous period was included as a variable in the extended model. The diagnostic tests suggested there were no specification problems. Table 1 shows that, for the models estimated for the whole period, the coefficients of the Google Trends variables are significantly different from zero at the 1% level and the coefficient of the imbalance of the previous period is always somewhat significant. The inclusion of these variables clearly improves the adjusted R2.

Table 1: Main results concerning the models estimated for the unemployment rate in Portugal, for the whole period (January 2005 – August 2013).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dependent variable: Δ*U*P(*t*) | | |  | Dependent variable: Δ*Log-U*P(*t*) | | |
|  | Base model | Base model  + Δ*GU*P, *EC* |  |  | Base model | Base model  + Δ*GU*P, *EC* |
| Constant | 0.000207 | -0.000712\* |  | Constant | 0.00173 | -0.00986 |
| Δ*U*P(*t*-1) | 0.728\*\*\* | 0.634\*\*\* |  | Δ*Log*-*U*P(*t*-1) | 0.723\*\*\* | 0.663\*\*\* |
| Δ*GU*P(*t*) | - | 0.00737\*\*\* |  | Δ*Log*-*GU*P(*t*) | - | 0.0398\*\*\* |
| *EC*(*t*-1) | - | -0.0105\*\* |  | *EC*(*t*-1) | - | -0.00874\* |
| Adjusted R2 | 0.523 | 0.608 |  | Adjusted R2 | 0.521 | 0.595 |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3.

Table 2 shows that the volume of search queries is always significant at the 1% level in the in-sample period of each data split, but the imbalance variable is no longer significant at the 10% level. The extended model leads to a significant decrease the RMSFE whenever nowcasts are extended until August 2013, and also in the predictions concerning the first out-of-sample year of data split 2. The only case in which there is no improvement in predictive ability is when we try to nowcast the first out-of-sample year of data split 1. However, even in this case the encompassing test indicates that the extended model adds useful predictive information.

Table 2: Main results concerning the forecasting models estimated for the unemployment rate in Portugal and corresponding forecasts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Base model | | | |  | Extended model | | | | |
| Dependent  Variable | Data split | Adjusted  R2 | Out-of-sample  RMSFE |  | Additional  variables | Coefficient  added var. | Adjusted  R2 | Out-of-sample  RMSFE | Percentage improvement |
| Δ*U*P | 1 | 0.493 | 3 years: 0.00141 |  | Δ*GU*P(*t*) | 0.00560\*\*\* | 0.554 | 3 years: 0.00126 | 10.26%\*\*\*,+++ |
| 1 year: 0.000766 |  | *EC*(*t*-1) | -0.00573 | 1 year: 0.000794 | -3.62%++ |
| 2 | 0.470 | 2 years: 0.00164 |  | Δ*GU*P(*t*) | 0.00518\*\*\* | 0.527 | 2 years: 0.00145 | 12.13%\*\*\*,+++ |
| 1 year: 0.00166 |  | *EC*(*t*-1) | -0.00449 | 1 year: 0.00140 | 15.75%\*\*\*,+++ |
| 3 | 0.521 | 1 year: 0.00166 |  | Δ*GU*P(*t*) | 0.00631\*\*\* | 0.592 | 1 year: 0.00143 | 13.72%\*\*\*,+++ |
|  | *EC*(*t*-1) | -0.00903\* |
| Δ*Log-U*P | 1 | 0.509 | 3 years: 0.00928 |  | Δ*Log*-*GU*P(*t*) | 0.0362\*\*\* | 0.568 | 3 years: 0.00823 | 11.28%\*\*\*,+++ |
| 1 year: 0.00615 |  | *EC*(*t*-1) | -0.00574 | 1 year: 0.00665 | -8.06%+++ |
| 2 | 0.498 | 2 years: 0.0105 |  | Δ*Log*-*GU*P(*t*) | 0.0328\*\*\* | 0.552 | 2 years: 0.00906 | 13.90%\*\*\*,+++ |
| 1 year: 0.0113 |  | *EC*(*t*-1) | -0.00451 | 1 year: 0.00938 | 16.63%\*\*\*,+++ |
| 3 | 0.514 | 1 year: 0.00991 |  | Δ*Log*-*GU*P(*t*) | 0.0366\*\*\* | 0.580 | 1 year: 0.00828 | 16.46%\*\*\*,+++ |
|  | *EC*(*t*-1) | -0.00814 |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3 and data splits are defined in Section 4. In the case of the improvement, the stars refer to the one-sized F test of forecast improvement of McCracken [17]. The ‘+’ signs refer to the one-sized F test of forecasting encompassing of Clark and McCracken [10].

The results for Spain, presented in Tables 3 and 4 (base data can be found in Fig. 2 in the Appendix), are very different from the results obtained for Portugal. For Spain, the QLR test detected a structural break in February 2009 for both the additive and the multiplicative models, so models were estimated using just data from March 2009 onwards. After restricting the data, diagnostic tests suggested no specification problems.

Table 3: Main results concerning the models estimated for the unemployment rate in Spain, for the period after the detected structural break (March 2009 – August 2013).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dependent variable: Δ*U*S(*t*) | | |  | Dependent variable: Δ*Log-U*S(*t*) | | |
|  | Base model | Base model  + *GU*S |  |  | Base model | Base model  + *GU*S |
| Constant | 0.000411\*\* | -0.000113 |  | Constant | 0.00221\*\*\* | -0.0000738 |
| Δ*U*S(*t*-1) | 0.701\*\*\* | 0.692\*\*\* |  | Δ*Log*-*U*S(*t*-1) | 0.656\*\*\* | 0.724\*\*\* |
| *GU*S(*t*) |  | 0.000964 |  | *Log*-*GU*S(*t*) |  | 0.00511 |
|  |  |  |  | *Log*-*GU*S(*t*-2) |  | -0.00830\*\* |
| Adjusted R2 | 0.646 | 0.642 |  | Adjusted R2 | 0.688 | 0.704 |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3.

In Table 3 we can see that the coefficient of the contemporary volume of search queries is not significantly different from zero. Although the second order lag of the volume of search queries is significant at the 5% level in the multiplicative model, we cannot rule out this being a statistical glitch: the coefficient of the variable has a negative sign, contrary to what was expected and, looking at Table 4, we can see that its coefficient is not significantly different from zero in two of the three in-sample periods.

The addition of Google Trends variables usually leads to worse unemployment predictions for Spain. When the additive model is used, RMSFE is always larger, although in two cases there is some evidence of the variables adding useful predictive information, according to the encompassing test. In the case of the multiplicative model, significant RMSFE improvements are achieved in the one-year nowcasts performed in data splits 1 and 3, but surprisingly, in these cases, the Google Trends variables are not significantly different from zero in the model estimated in-sample. In the multiplicative model, the encompassing test finds significant evidence that the extended model adds useful predictive information, even when the RMSFE worsens.

We must also stress that, for Spain, results may be somehow affected by a small number of observations in the in-sample periods. Due to the structural break found in the data, we were left with 17, 29 and 31 in-sample observations for data splits 1, 2 and 3, respectively. This may be insufficient to obtain robust models, particularly in the case of the first data split.

Table 4: Main results concerning the forecasting models estimated for the unemployment rate in Spain and corresponding forecasts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Base model | | | |  | Extended model | | | | |
| Dependent  Variable | Data split | Adjusted  R2 | Out-of-sample  RMSFE |  | Additional  variables | Coefficient  added var. | Adjusted  R2 | Out-of-sample  RMSFE | Percentage improvement |
| Δ*U*S | 1 | 0.661 | 3 years:0.000950 |  | *GU*S(*t*) | 0.00893 | 0.683 | 3 years: 0.00122 | -28.43%+ |
| 1 year: 0.000981 |  | 1 year: 0.00103 | -5.02% |
| 2 | 0.620 | 2 years: 0.000903 |  | *GU*S(*t*) | 0.00633 | 0.632 | 2 years: 0.00117 | -30.00% |
| 1 year: 0.000894 |  | 1 year: 0.00110 | -23.14%+++ |
| 3 | 0.599 | 1 year: 0.000973 |  | *GU*S(*t*) | 0.00201 | 0.605 | 1 year: 0.00105 | -7.67% |
| Δ*Log-U*S | 1 | 0.726 | 3 years: 0.00411 |  | *Log*-*GU*S(*t*) | 0.0231 | 0.736 | 3 years: 0.00476 | -15.84%++ |
| 1 year: 0.00460 |  | *Log*-*GU*S(*t*-2) | -0.00821 | 1 year: 0.00422 | 8.16%\*\*,++ |
| 2 | 0.683 | 2 years: 0.00375 |  | *Log*-*GU*S(*t*) | 0.0202\* | 0.732 | 2 years: 0.00468 | -29.58%++ |
| 1 year: 0.00355 |  | *Log*-*GU*S(*t*-2) | -0.0158\*\* | 1 year: 0.00535 | -50.92%+ |
| 3 | 0.647 | 1 year: 0.00428 |  | *Log*-*GU*S(*t*) | 0.00651 | 0.661 | 1 year: 0.00379 | 11.38%\*\*\*,+++ |
|  | *Log*-*GU*S(*t*-2) | -0.00983 |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3 and data splits are defined in Section 4. In the case of the improvement, the stars refer to the one-sized F test of forecast improvement of McCracken [17]. The ‘+’ signs refer to the one-sized F test of forecasting encompassing of Clark and McCracken [10].

The results concerning France are shown in Tables 5 and 6 (again, base data can be found in the Appendix, in Fig. 3). The QLR test detected a structural break in March 2008 in the multiplicative model (significant at the 5% level), but not in the additive model. So, data from April 2008 onwards was used in the multiplicative model. The volume of search queries and unemployment were found to be cointegrated and therefore the imbalance of the previous period was included in the extended model. The diagnostic tests suggested there were no specification problems in the models.

Table 5 shows that, for the models estimated for the whole period, both the coefficients of Google Trends variables and those of the imbalance of the previous period are significantly different from zero at the 1% level, and the inclusion of these variables clearly improves the adjusted R2.

Table 5: Main results concerning the models estimated for the unemployment rate in France, for the whole period (January 2005 – August 2013) in the case of the additive model, and for the period after the detected structural break (April 2008 – August 2013) in the case of the multiplicative model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dependent variable: Δ*U*F(*t*) | | |  | Dependent variable: Δ*Log-U*F(*t*) | | |
|  | Base model | Base model  + Δ*GU*F, *EC* |  |  | Base model | Base model  + Δ*GU*F, *EC* |
| Constant | 0.0000317 | 0.00111\*\*\* |  | Constant | 0.00139 | -0.147\*\*\* |
| Δ*U*F(*t*-1) | 0.946\*\*\* | 0.696\*\*\* |  | Δ*Log*-*U*F(*t*-1) | 0.833\*\*\* | 0.385\*\*\* |
| Δ*U*F(*t*-2) | -0.212\*\* | -0.224\*\* |  |  |  |  |
| Δ*U*F(*t*-4) | -0.263\*\* | -0.315\*\*\* |  | Δ*Log*-*U*F(*t*-4) | -0.391\*\*\* | -0.388\*\*\* |
| Δ*U*F(*t*-5) | 0.344\*\*\* | 0.245\*\*\* |  | Δ*Log*-*U*F(*t*-5) | 0.307\*\*\* | 0.111 |
| Δ*GU*F(*t*) |  | 0.00464\*\*\* |  | Δ*Log*-*GU*F(*t*) |  | 0.0451\*\*\* |
| *EC*(*t*-1) |  | -0.0446\*\*\* |  | *EC*(*t*-1) |  | -0.0715\*\*\* |
| Adjusted R2 | 0.646 | 0.726 |  | Adjusted R2 | 0.596 | 0.739 |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3.

Table 6: Main results concerning the forecasting models estimated for the unemployment rate in France and corresponding forecasts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Base model | | | |  | Extended model | | | | |
| Dependent  Variable | Data split | Adjusted  R2 | Out-of-sample  RMSFE |  | Additional  variables | Coefficient  added var. | Adjusted  R2 | Out-of-sample  RMSFE | Percentage improvement |
| Δ*U*F | 1 | 0.686 | 3 years: 0.000501 |  | Δ*GU*F(*t*) | 0.00675\*\*\* | 0.755 | 3 years: 0.000509 | -1.67%+++ |
| 1 year: 0.000551 |  | *EC*(*t*-1) | -0.0258\*\*\* | 1 year: 0.000534 | 3.06%+++ |
| 2 | 0.663 | 2 years: 0.000471 |  | Δ*GU*F(*t*) | 0.00545\*\*\* | 0.744 | 2 years: 0.000447 | 4.97%\*\*,+++ |
| 1 year: 0.000482 |  | *EC*(*t*-1) | -0.0426\*\*\* | 1 year: 0.000417 | 13.43%\*\*\*,+++ |
| 3 | 0.650 | 1 year: 0.000456 |  | Δ*GU*F(*t*) | 0.00533\*\*\* | 0.735 | 1 year: 0.000478 | -4.81%++ |
|  | *EC*(*t*-1) | -0.0458\*\*\* |
| Δ*Log-U*F | 1 | 0.646 | 3 years: 0.00554 |  | Δ*Log*-*GU*F(*t*) | 0.0657\*\*\* | 0.745 | 3 years: 0.00551 | 0.62%+++ |
| 1 year: 0.00671 |  | *EC*(*t*-1) | -0.0383\*\* | 1 year: 0.00637 | 4.95%\*,++ |
| 2 | 0.640 | 2 years: 0.00463 |  | Δ*Log*-*GU*F(*t*) | 0.0333\* | 0.801 | 2 years:0.00588 | -27.00%+++ |
| 1 year: 0.00475 |  | *EC*(*t*-1) | -0.103\*\*\* | 1 year: 0.00432 | 9.13%\*\*,+++ |
| 3 | 0.612 | 1 year: 0.00446 |  | Δ*Log*-*GU*F(*t*) | 0.0401\*\* | 0.777 | 1 year: 0.00612 | -37.10%+++ |
|  | *EC*(*t*-1) | -0.0978\*\*\* |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3 and data splits are defined in Section 4. In the case of the improvement, the stars refer to the one-sized F test of forecast improvement of McCracken [17]. The ‘+’ signs refer to the one-sized F test of forecasting encompassing of Clark and McCracken [10].

Table 6 shows that there seem to be some prediction improvements achieved by adding Google Trends-related variables, but at the first sight they do not seem to be consistent. If we look at the one-year nowcasts, we can see that in the first year of the out-of-sample period of data split 1 (September 2010-August 2011), both the additive and the multiplicative extended models seem to lead to modest prediction improvements; in the first year of the out-of-sample period of data split 2 (September 2011-August 2012), both extended models lead to significant improvements; in the out-of-sample period of data split 3 (September 2012-August 2013), both extended models lead to worse predictions than the corresponding base models. In all cases the encompassing test finds significant evidence that the extended model adds useful predictive information. So, it can be the case that Google Trends variables add predictive ability but, sometimes, the also add a level of noise that outweighs such predictive ability.

Tables 7 and 8 present the results concerning Italy (as before, base data can be found in the Appendix, in Fig. 4). Diagnostic tests showed that some autocorrelation seems to remain in the multiplicative model, although mild (only when 12 lags are considered, and not significant at the 1% level). The added Google Trends variable is significant at the 5% level in the additive model and at the 10% level in the multiplicative one, and in both cases they lead to a small improvement in the adjusted R2.

Table 7: Main results concerning the models estimated for the unemployment rate in Italy, for the whole period (January 2005 – August 2013).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dependent variable: Δ2*U*I(*t*) | | |  | Dependent variable: Δ2*Log-U*I(*t*) | | |
|  | Base model | Base model  + Δ*GU*I |  |  | Base model | Base model  + Δ*GU*I |
| Constant | 0.0000359 | -0.0000045 |  | Constant | 0.000386 | 0.0000244 |
| Δ2*U*I(*t*-1) | -1.04\*\*\* | -1.00\*\*\* |  | Δ2*Log-U*I(*t*-1) | -0.955\*\*\* | -0.932\*\*\* |
| Δ2*U*I(*t*-2) | -0.590\*\*\* | -0.569\*\*\* |  | Δ2*Log-U*I(*t*-2) | -0.329\*\*\* | -0.322\*\*\* |
| Δ2*U*I(*t*-3) | -0.199\*\* | -0.200\*\* |  |  |  |  |
| Δ2*U*I(*t*-11) | 0.254\*\*\* | 0.243\*\*\* |  | Δ2*Log-U*I(*t*-11) | 0.246\*\*\* | 0.234\*\*\* |
| Δ*GU*I(*t*) |  | 0.00895\*\* |  | Δ*Log*-*GU*I(*t*) |  | 0.0564\* |
| Adjusted R2 | 0.683 | 0.700 |  | Adjusted R2 | 0.691 | 0.698 |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3.

Table 8 shows that the addition of Google Trends variables always leads to a reduction in the RMSFE. The significance of the prediction improvement varies, but they are always significant at the 1% or 5% levels when nowcasts are extended until August 2013.

Table 8: Main results concerning the forecasting models estimated for the unemployment rate in Italy and corresponding forecasts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Base model | | | |  | Extended model | | | | |
| Dependent  Variable | Data split | Adjusted  R2 | Out-of-sample  RMSFE |  | Additional  variables | Coefficient  added var. | Adjusted  R2 | Out-of-sample  RMSFE | Percentage improvement |
| Δ2*U*I | 1 | 0.692 | 3 years: 0.00192 |  | Δ*GU*I(*t*) | 0.00474 | 0.693 | 3 years: 0.00182 | 5.21%\*\*\*,+++ |
| 1 year: 0.00200 |  | 1 year: 0.00194 | 2.88%\* |
| 2 | 0.681 | 2 years: 0.00198 |  | Δ*GU*I(*t*) | 0.00644 | 0.687 | 2 years: 0.00183 | 7.62%\*\*\*,+++ |
| 1 year: 0.00213 |  | 1 year: 0.00203 | 4.64%\*\*,++ |
| 3 | 0.690 | 1 year: 0.00167 |  | Δ*GU*I(*t*) | 0.00798\* | 0.700 | 1 year: 0.00145 | 13.27%\*\*\*,+++ |
| Δ2*Log-U*I | 1 | 0.701 | 3 years: 0.0207 |  | Δ*Log*-*GU*I(*t*) | 0.0256 | 0.697 | 3 years: 0.0201 | 3.10%\*\*,++ |
| 1 year: 0.0230 |  | 1 year: 0.0225 | 2.22% |
| 2 | 0.697 | 2 years: 0.0198 |  | Δ*Log*-*GU*I(*t*) | 0.0387 | 0.697 | 2 years: 0.0188 | 5.22%\*\*\*,++ |
| 1 year: 0.0220 |  | 1 year: 0.0213 | 3.15%\*,+ |
| 3 | 0.700 | 1 year: 0.0168 |  | Δ*Log*-*GU*I(*t*) | 0.0477 | 0.702 | 1 year: 0.0152 | 9.79%\*\*\*,+++ |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3 and data splits are defined in Section 4. In the case of the improvement, the stars refer to the one-sized F test of forecast improvement of McCracken [17]. The ‘+’ signs refer to the one-sized F test of forecasting encompassing of Clark and McCracken [10].

## Car sales

Tables 9 and 10 present the results concerning car sales for Portugal. The base data is presented in Fig. 5 in the Appendix. The QLR test detected a structural break in August 2005 in the additive model, but no structural break was found in the multiplicative model. So, the additive model was estimated using just data from September 2005 onwards. The Johansen test showed no cointegration between car sales and Google Trends variables, so no imbalance variable was added to the extended model. The diagnostic tests suggested no specification problems in the multiplicative model, but some autocorrelation in the additive model (significant at the 5% level but not at the 1% level, and only in the Ljung-Box test). The coefficients of Google Trends variables are always significantly different from zero at the 1% level and they always lead to an increase in the adjusted R2.

Table 9: Main results concerning the models estimated for car sales in Portugal, for the period after the detected structural break (September 2005 – August 2013) in the case of the additive model, and for the whole period (January 2004 – August 2013) in the case of the multiplicative model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dependent variable: Δ*S*P(*t*) | | |  | Dependent variable: Δ*Log-S*P(*t*) | | |
|  | Base model | Base model  + Δ*GS*P |  |  | Base model | Base model  + Δ*GS*P |
| Constant | 0.000294 | 0.00228 |  | Constant | -0.000429 | 0.00156 |
| Δ*S*P(*t*-1) | -0.0663 | -0.0358 |  | Δ*Log*-*S*P(*t*-1) | -0.0886 | -0.0762 |
| *D*2008:12 | 0.170\*\*\* | 0.171\*\*\* |  | *D*2008:12 | 0.248\*\*\* | 0.252\*\*\* |
| *D*2009:01 | -0.148\*\*\* | -0.168\*\*\* |  | *D*2009:01 | -0.490\*\*\* | -0.507\*\*\* |
| *D*2010:12 | 0.278\*\*\* | 0.273\*\*\* |  | *D*2010:12 | 0.330\*\*\* | 0.326\*\*\* |
|  |  |  |  | *D*2011:01 | -0.255\*\* | -0.231\*\* |
| *D*2012:01 | -0.0679\* | -0.0712\* |  | *D*2012:01 | -0.256\*\*\* | -0.258\*\*\* |
| *A*Dec/Jan | -1.499\*\*\* | -1.440\*\*\* |  | *A*Dec/Jan | -0.943\*\*\* | -0.916\*\*\* |
| Δ*GS*P(*t*) |  | 0.612\*\*\* |  | Δ*Log*-*GS*P(*t*) |  | 0.702\*\*\* |
| Adjusted R2 | 0.771 | 0.797 |  | Adjusted R2 | 0.639 | 0.660 |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3.

Table 10: Main results concerning the forecasting models estimated for car sales in Portugal and corresponding forecasts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Base model | | | |  | Extended model | | | | |
| Dependent  Variable | Data split | Adjusted  R2 | Out-of-sample  RMSFE |  | Additional  variables | Coefficient  added var. | Adjusted  R2 | Out-of-sample  RMSFE | Percentage improvement |
| Δ*S*P | 1 | 0.663 | 3 years: 0.0712 |  | Δ*GU*P(*t*) | 0.598\*\*\* | 0.703 | 3 years: 0.0715 | -0.32% |
| 1 year: 0.114 |  | 1 year: 0.115 | -0.15% |
| 2 | 0.783 | 2 years: 0.0282 |  | Δ*GU*P(*t*) | 0.632\*\*\* | 0.810 | 2 years: 0.0285 | -1.01%+ |
| 1 year: 0.0327 |  | 1 year: 0.0335 | -2.52% |
| 3 | 0.775 | 1 year: 0.0228 |  | Δ*GU*P(*t*) | 0.621\*\*\* | 0.801 | 1 year: 0.0222 | 2.78%\*,+++ |
| Δ*Log-S*P | 1 | 0.590 | 3 years: 0.124 |  | Δ*Log*-*GU*P(*t*) | 0.614\*\* | 0.612 | 3 years: 0.121 | 2.84%\*\*,++ |
| 1 year: 0.159 |  | 1 year: 0.152 | 3.92%\*\*,+ |
| 2 | 0.668 | 2 years: 0.0996 |  | Δ*Log*-*GU*P(*t*) | 0.672\*\* | 0.690 | 2 years: 0.0988 | 0.75% |
| 1 year: 0.113 |  | 1 year: 0.113 | -0.01% |
| 3 | 0.656 | 1 year: 0.0837 |  | Δ*Log*-*GU*P(*t*) | 0.683\*\*\* | 0.677 | 1 year: 0.0822 | 1.98%+ |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3 and data splits are defined in Section 4. In the case of the improvement, the stars refer to the one-sized F test of forecast improvement of McCracken [17]. The ‘+’ signs refer to the one-sized F test of forecasting encompassing of Clark and McCracken [10].

Looking at Table 10, we notice that the Google Trends variables are still very significant in the models estimated using in-sample data, for all data splits. However, we can find only mild evidence of improved nowcasting ability by adding Google Trends variables. The additive model only seems to lead to improved predictions in data split 3, and the improvement is significant only at the 10% level. In the multiplicative case, the extended model seems to lead to significant improvement in the predictive ability when data split 1 is considered but, in the other data splits, only non-significant improvements are found.

Tables 11 and 12 present the results concerning car sales in Spain. The base data is presented in Fig. 6 in the Appendix. Diagnostic tests seem to point out to some ARCH in the residuals: the LM test with 6 lags leads to values significant at the 5% level in both models; however, the results are not significant at the 1% level, and the LM test with 12 lags does not lead to significant results at the 5% level.

Table 11 shows that the coefficients of Google Trends variables are not significantly different from zero at the 5% level and their sign is negative, contrary to what was to be expected. So, it is no surprise to find, in Table 12, that the extended model usually produces non-significant prediction improvements or no improvements at all.

Table 11: Main results concerning the models estimated for car sales in Spain, for the whole period (January 2004 – August 2013).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dependent variable: Δ*S*S(*t*) | | |  | Dependent variable: Δ*Log-S*S(*t*) | | |
|  | Base model | Base model  + Δ*GS*S |  |  | Base model | Base model  + Δ*GS*S |
| Constant | -0.000677 | -0.000739 |  | Constant | -0.00188 | -0.00183 |
| Δ*S*S(*t*-1) | -0.0743 | -0.0765 |  | Δ*Log*-*S*S(*t*-1) | -0.0407 | -0.0581 |
| *D*2007:04 | -0.0944\*\* | -0.0937\*\* |  |  |  |  |
| *D*2007:12 | 0.115\*\*\* | 0.115\*\*\* |  | *D*2007:12 | 0.124\*\* | 0.108\* |
| *D*2008:01 | -0.196\*\*\* | -0.197\*\*\* |  | *D*2008:01 | -0.222\*\*\* | -0.225\*\*\* |
| *D*2010:07 | -0.213\*\*\* | -0.212\*\*\* |  | *D*2010:07 | -0.408\*\*\* | -0.387\*\*\* |
| *D*2012:08 | 0.0983\*\*\* | 0.0977\*\*\* |  | *D*2012:08 | 0.258\*\*\* | 0.246\*\*\* |
| *D*2012:09 | -0.126\*\*\* | -0.127\*\*\* |  | *D*2012:09 | -0.349\*\*\* | -0.352\*\*\* |
| Δ*GS*S(*t*) |  | -0.0597 |  | Δ2*Log*-*GS*S(*t*) |  | -0.218\* |
| Adjusted R2 | 0.480 | 0.476 |  | Adjusted R2 | 0.508 | 0.517 |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3.

Table 12: Main results concerning the forecasting models estimated for car sales in Spain and corresponding forecasts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Base model | | | |  | Extended model | | | | |
| Dependent  Variable | Data split | Adjusted  R2 | Out-of-sample  RMSFE |  | Additional  variables | Coefficient  added var. | Adjusted  R2 | Out-of-sample  RMSFE | Percentage improvement |
| Δ*S*S | 1 | 0.460 | 3 years: 0.0360 |  | Δ*GU*S(*t*) | -0.0285 | 0.453 | 3 years: 0.0360 | -0.06% |
| 1 year: 0.0214 |  | 1 year: 0.0210 | 2.02% |
| 2 | 0.453 | 2 years: 0.0412 |  | Δ*GU*S(*t*) | -0.0604 | 0.448 | 2 years: 0.0415 | -0.73% |
| 1 year: 0.0408 |  | 1 year: 0.0405 | 0.53% |
| 3 | 0.455 | 1 year: 0.0411 |  | Δ*GU*S(*t*) | -0.0731 | 0.451 | 1 year: 0.0420 | -2.12% |
| Δ*Log-S*S | 1 | 0.451 | 3 years: 0.0957 |  | Δ2*Log*-*GU*S(*t*) | -0.170 | 0.453 | 3 years: 0.0946 | 1.10% |
| 1 year: 0.0506 |  | 1 year: 0.0480 | 5.24%\*\*,++ |
| 2 | 0.428 | 2 years: 0.111 |  | Δ2*Log*-*GU*S(*t*) | -0.194 | 0.435 | 2 years: 0.111 | 0.12% |
| 1 year: 0.106 |  | 1 year: 0.103 | 2.56%\* |
| 3 | 0.442 | 1 year: 0.113 |  | Δ2*Log*-*GU*S(*t*) | -0.231 | 0.452 | 1 year: 0.115 | -1.65% |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3 and data splits are defined in Section 4. In the case of the improvement, the stars refer to the one-sized F test of forecast improvement of McCracken [17]. The ‘+’ signs refer to the one-sized F test of forecasting encompassing of Clark and McCracken [10].

The results obtained for car sales in France are shown in Tables 13 and 14 (as before, base data is presented in the Appendix in Fig. 7). The volume of search queries and car sales were found to be cointegrated and, therefore, we included the imbalance of the previous period as a variable in the extended model. The diagnostic tests suggested there were no specification problems in the models.

Similarly to the case of Spain, coefficients of Google Trends variables are not significantly different from zero and their sign is negative, contrary to what was to be expected. Also, the coefficients of the imbalance of the previous period are not significantly different from zero. Table 12 shows that extended model only improves the RMSFE in the one-year nowcasts of data sample 1, and in all other cases it increases the forecast error.

Table 13: Main results concerning the models estimated for car sales in France, for the whole period (January 2004 – August 2013).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dependent variable: Δ*S*F(*t*) | | |  | Dependent variable: Δ*Log-S*F(*t*) | | |
|  | Base model | Base model  + Δ*GS*F |  |  | Base model | Base model  + Δ*GS*F |
| Constant | -0.00286 | 0.0989 |  | Constant | -0.00356 | -0.0285 |
| Δ*S*F(*t*-1) | -0.402\*\*\* | -0.369\*\*\* |  | Δ*Log*-*S*F(*t*-1) | -0.431\*\*\* | -0.404\*\*\* |
| *D*2007:12 | 0.105\*\*\* | 0.106\*\*\* |  | *D*2007:12 | 0.128\*\*\* | 0.129\*\*\* |
| *D*2009:05 | 0.199\*\*\* | 0.199\*\*\* |  | *D*2009:05 | 0.226\*\*\* | 0.227\*\*\* |
| *D*2010:11 | 0.0953\*\* | 0.0967\*\* |  | *D*2010:11 | 0.112\*\* | 0.115\*\* |
| *D*2011:04 | -0.178\*\*\* | -0.169\*\*\* |  | *D*2011:04 | -0.211\*\*\* | -0.200\*\*\* |
| *EC*(*t*-1) |  | -0.0691 |  | *EC*(*t*-1) |  | -0.0611 |
| Δ*GS*F(*t*) |  | -0.114 |  | Δ*Log*-*GS*F(*t*) |  | -0.143 |
| Adjusted R2 | 0.450 | 0.448 |  | Adjusted R2 | 0.431 | 0.428 |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3.

Table 14: Main results concerning the forecasting models estimated for car sales in France and corresponding forecasts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Base model | | | |  | Extended model | | | | |
| Dependent  Variable | Data split | Adjusted  R2 | Out-of-sample  RMSFE |  | Additional  variables | Coefficient  added var. | Adjusted  R2 | Out-of-sample  RMSFE | Percentage improvement |
| Δ*S*F | 1 | 0.413 | 3 years: 0.0465 |  | Δ*GU*F(*t*) | -0.178 | 0.463 | 3 years: 0.0475 | -2.06%+++ |
| 1 year: 0.0672 |  | *EC*(*t*-1) | -0.242\*\*\* | 1 year: 0.0620 | 7.78%\*\*,+++ |
| 2 | 0.466 | 2 years: 0.0318 |  | Δ*GU*F(*t*) | -0.188 | 0.484 | 2 years: 0.0392 | -23.07% |
| 1 year: 0.0289 |  | *EC*(*t*-1) | -0.179\*\* | 1 year: 0.0345 | -19.43% |
| 3 | 0.449 | 1 year: 0.0346 |  | Δ*GU*F(*t*) | -0.121 | 0.458 | 1 year: 0.0386 | -11.42% |
|  | *EC*(*t*-1) | -0.134\* |
| Δ*Log-S*F | 1 | 0.390 | 3 years: 0.0585 |  | Δ*Log*-*GU*F(*t*) | -0.231 | 0.441 | 3 years: 0.0613 | -4.67%++ |
| 1 year: 0.0792 |  | *EC*(*t*-1) | -0.253\*\*\* | 1 year: 0.0727 | 8.18%\*\*\*,+++ |
| 2 | 0.444 | 2 years: 0.0454 |  | Δ*Log*-*GU*F(*t*) | -0.235 | 0.465 | 2 years: 0.0552 | -21.57% |
| 1 year: 0.0388 |  | *EC*(*t*-1) | -0.192\*\* | 1 year: 0.0463 | -19.12% |
| 3 | 0.425 | 1 year: 0.0514 |  | Δ*Log*-*GU*F(*t*) | -0.154 | 0.434 | 1 year: 0.0564 | -9.70% |
|  | *EC*(*t*-1) | -0.140\* |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3 and data splits are defined in Section 4. In the case of the improvement, the stars refer to the one-sized F test of forecast improvement of McCracken [17]. The ‘+’ signs refer to the one-sized F test of forecasting encompassing of Clark and McCracken [10].

The results for Italy are presented in Tables 15 and 16. Base data is presented in Fig. 8 in the Appendix. The Johansen test showed no cointegration between car sales and Google Trends variables, so no imbalance variable was added to the extended model. The diagnostic tests suggested no specification problems in the models. The coefficients of Google Trends variables are not significantly different from zero and the adjusted R2 is almost identical to the one of the base model.

Table 15: Main results concerning the models estimated for car sales in Italy, for the whole period (January 2004 – August 2013).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dependent variable: Δ*S*I(*t*) | | |  | Dependent variable: Δ*Log-S*I(*t*) | | |
|  | Base model | Base model  + Δ*GS*I |  |  | Base model | Base model  + Δ*GS*I |
| Constant | -0.00190 | -0.00189 |  | Constant | -0.00334 | -0.00329 |
| Δ*S*I(*t*-1) | -0.235\*\*\* | -0.234\*\*\* |  | Δ*Log*-*S*I(*t*-1) | -0.231\*\*\* | -0.231\*\*\* |
| Δ*S*I(*t*-2) | -0.173\*\*\* | -0.175\*\*\* |  | Δ*Log*-*S*I(*t*-2) | -0.174\*\*\* | -0.180\*\*\* |
| *D*2004:08 | -0.0903\*\*\* | -0.0899\*\*\* |  | *D*2004:08 | -0.107\*\* | -0.105\*\* |
| *D*2005:05 | -0.258\*\*\* | -0.258\*\*\* |  | *D*2005:05 | -0.355\*\*\* | -0.354\*\*\* |
| *D*2005:06 | 0.337\*\*\* | 0.337\*\*\* |  | *D*2005:06 | 0.426\*\*\* | 0.425\*\*\* |
| *D*2008:12 | 0.0981\*\*\* | 0.0984\*\*\* |  | *D*2008:12 | 0.130\*\*\* | 0.132\*\*\* |
| *D*2010:04 | -0.258\*\*\* | -0.258\*\*\* |  | *D*2010:04 | -0.321\*\*\* | -0.321\*\*\* |
| *A*Dec/Jan | -1.08\*\*\* | -0.0899\*\*\* |  | *A*Dec/Jan | -1.06\*\*\* | -1.08\*\*\* |
| Δ*GS*I(*t*) |  | 0.0186 |  | Δ*Log*-*GS*I(*t*) |  | 0.0594 |
| Adjusted R2 | 0.766 | 0.764 |  | Adjusted R2 | 0.743 | 0.742 |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3.

In Table 16 it can be seen that, with the exception of data split 1 in the additive model, the extended model always leads to slightly smaller values of the RMSFE. However, improvements are very slight and never significant, even at the 10% level. In fact, the encompassing test never finds any evidence that the extended model may add any useful predictive ability in relation to the base model.

Table 16: Main results concerning the forecasting models estimated for car sales in Italy and corresponding forecasts.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Base model | | | |  | Extended model | | | | |
| Dependent  Variable | Data split | Adjusted  R2 | Out-of-sample  RMSFE |  | Additional  variables | Coefficient  added var. | Adjusted  R2 | Out-of-sample  RMSFE | Percentage improvement |
| Δ*S*I | 1 | 0.809 | 3 years: 0.0277 |  | Δ*GU*I(*t*) | -0.0236 | 0.806 | 3 years: 0.0320 | -0.41% |
| 1 year: 0.0248 |  | 1 year: 0.0251 | -1.16% |
| 2 | 0.799 | 2 years: 0.0290 |  | Δ*GU*I(*t*) | 0.00875 | 0.796 | 2 years: 0.0290 | 0.05% |
| 1 year: 0.0336 |  | 1 year: 0.0336 | 0.01% |
| 3 | 0.775 | 1 year: 0.0233 |  | Δ*GU*I(*t*) | 0.0100 | 0.773 | 1 year: 0.0232 | 0.17% |
| Δ*Log-S*I | 1 | 0.829 | 3 years: 0.0485 |  | Δ*Log*-*GU*I(*t*) | 0.00389 | 0.827 | 3 years: 0.0484 | 0.04% |
| 1 year: 0.0367 |  | 1 year: 0.0367 | 0.17% |
| 2 | 0.813 | 2 years: 0.0533 |  | Δ*Log*-*GU*I(*t*) | 0.0472 | 0.811 | 2 years: 0.0533 | 0.15% |
| 1 year: 0.0590 |  | 1 year: 0.0590 | 0.01% |
| 3 | 0.768 | 1 year: 0.0467 |  | Δ*Log*-*GU*I(*t*) | 0.0479 | 0.766 | 1 year: 0.0465 | 0.44% |

\*: significant at the 10% level; \*\*: significant at the 5% level; \*\*\*: significant at the 1% level. Variables are defined in Section 3 and data splits are defined in Section 4. In the case of the improvement, the stars refer to the one-sized F test of forecast improvement of McCracken [17]. The ‘+’ signs refer to the one-sized F test of forecasting encompassing of Clark and McCracken [10].

# Discussion

The global picture we get from the results is somewhat less rosy than the picture gathered from other studies, and it is very different for the two variables we tried to nowcast: while our results indicate that Google Trends data improves the predictive ability of the unemployment models in most of the countries we considered, we get little evidence that search query data may lead to better predictions concerning car sales, at least in the countries we examine.

Regarding unemployment, it seems to be possible to use Google Trends variables to improve nowcasts in three out of the four considered countries: Portugal, France and Italy. In all these three cases, the coefficients of the search query volume are significantly different from zero when we consider the whole period (or the period after structural breaks, when they were detected), although more so for Portugal and France than for Italy. When we consider one-year out-of-sample sub-periods, the addition of search query data improves predictions in two out of three sub-periods for Portugal and France, with the improvements concerning Portugal being more significant than in the case of France. In the case of Portugal and France, the encompassing test indicates that Google Trends data add useful predictive information even in the sub-periods in which they lead to larger forecast errors than the base model. For Italy, there are nowcasting improvements in all out-of-sample sub-periods that were considered. For Spain, the coefficient of the contemporaneous volume of search queries is not significant when the model is estimated for the period ending in August 2013, and in most cases adding search query data leads to worse predictions.

These results lead us to conclude that adding Google Trends data usually leads to better unemployment nowcasts, but not always so. We can find two possible explanations for the difference in results for different countries and for different periods. The first one is related to the existence of different levels of noise in the search query data. In fact, Google Trends data are a measure of the relative proportion of searches that concern a particular query. So, they may have a high level of noise due not only to the characteristics of the specific use of the search query and due to the sampling procedure used by Google (we tried to reduce this noise but averaging the values over a 2-week period), but also due to possible changes in the total volume of searches unrelated to the query we are considering. This level of noise may be different for different countries and for different periods, leading to different predictive ability of the Google Trends data. The second possible explanation is that user behaviour is continually changing, and that it is quite different in different countries, leading to changes in the predictive content of Google Trends data across time periods and across geographical locations. The fact that, in many cases, the encompassing test indicates that search query data add useful predictive information, even when the RMSFE worsens, leads us to favour the former explanation, but further studies are required to establish the causes for these differences.

At the first sight, our results are not as good as the results obtained in other studies concerning the use of search query volume in nowcasting unemployment. Choi and Varian [8] and D'Amauri and Marcucci [11] conclude that Google Trends data improve predictions of initial claims of unemployment benefits and unemployment rate, respectively. Since a later work of Choi and Varian [9], including more recent data, leads to the conclusion that adding Google Trends data leads to worse prediction of the initial claims, we must hypothesize that prediction improvements concerning variables related to US unemployment may also change according to the considered out-of-sample period. Fondeur and Karamé [14] analyse a country also considered in this paper: France. Although they focus on the prediction of youth unemployment, they find that the inclusion of Google Trends data improves the prediction of unemployment for all age ranges (although the improvement is greater for youth unemployment). The out-of-sample forecasting period of the authors goes from January 2009 to July 2011, so it is previous to most of our out-of-sample periods. In our out-of-sample period that shows a bigger intersection with that of those authors (the first out-of-sample year of data split 1), we also achieve an improvement in the prediction of unemployment by using Google Trends data (although a modest improvement). It is probable that, if we were to consider the prediction of youth unemployment (instead of the general unemployment rate), we might have achieved a bigger improvement in the predictive ability, since younger people probably rely more on the Internet than older people.

In the case of car sales, we find that the volume of search queries helps explaining the variance of the car sales data in Portugal and, to a lesser extent, in Spain. However, in none of the considered countries were we able to achieve a consistent improvement in car sales prediction accuracy due to the inclusion of search query data. In fact, the only country for which the coefficients of the search query volume are significantly different from zero is Portugal and, even for Portugal, we do not find consistent out-of-sample prediction improvements (although we do find some improvements). Moreover, in most cases, the encompassing test finds no useful predictive information in search query data. Several reasons may explain the lack of predictive ability of Google Trends data in this context:

* Information search regarding car purchases often takes place long time before the car is bought, but how long before changes from person to person. So, for cars sold in a given month, the corresponding search queries may be diluted over a large number of previous months, reducing the predictive content of search query data.
* A large number of searches concerning car brands can be related to the search for repair shops and parts sellers, instead of reflecting a purchase intention.
* Car purchases by companies (especially fleet renovations) represent a significant volume of sales that will usually not be accompanied by a correspondingly high volume of information searches.
* Specific circumstances, like changes in car taxation or promotions, often have a decisive effect in the timing of car sales. Such circumstances may contribute to change the relation between search patterns and car sales (given tax change or promotion announcements, people may look for car information sooner or later than they usually would do).

Our results seem to be roughly in line with those of Choi and Varian [7], who are unable to achieve consistent improvements in predicting the sales of car brands by using search volume data. Differently from us, in [9], the same authors conclude that the volume of search queries "suvs" and "insurance" leads to a significant improvement in the "Motor Vehicles and Parts Dealers" index, and Carrière-Swallow and Labbé [5] find that an index of search queries for nine of the most popular car brands in Chile leads to improvements in car sales prediction. Differences between our results and those of these authors may be due to cultural differences between countries, methodological differences (Choi and Varian are not just considering car sales, since they consider an index that also includes car parts and accessories, and Carrière-Swallow and Labbé define the search query index differently from us), or to the use of different out-of-sample periods.

So, the results we achieved are very different for the two variables we considered and for different countries. When Google Trends variables are significantly different from zero in-sample, they tend to lead to improvements of the predictive ability, but this is not always the case. Our results also stress that some care must be taken with the analysis of the results achieved by adding search query data to prediction models: in fact, we find that sub-periods with very significant improvements in out-of-sample predictive accuracy may be followed by sub-periods in which these variables seem to hamper the predictive ability of the models.

# Final remarks and future research

In this paper we analysed if and how the use of Internet query data from Google Trends could be used to improve nowcasting ability using simple models. We have shown how search data can sometimes be used to improve the forecasting of the current value of indicators usually delivered with lag (nowcasting). We considered forecasts of two variables, unemployment and car sales, in four countries from the same geographical region, Portugal, Spain, France and Italy. The analysis focused on four countries in a limited geographical area, which inherently share some common factors. Although expanding the analysis to more countries could prove desirable, it was not done (yet) because of time and resource limits. Nevertheless, the authors feel that the variation on results makes the sample size enough to demonstrate the unevenness of the approach.

We concluded that adding Google Trends data usually leads to improvements in unemployment predictions in three of the considered countries (Portugal, France and Italy) and that these data helps explaining the variance of the car sales data in Portugal and, to a lesser extent, in Spain. However, our results give little support to the hypothesis that these data may improve car sales forecasts.

In our analysis, we assumed that data concerning the variables we are nowcasting would be released at most one month after the period to which they pertain. If this lag is larger (e.g., Suhoy [18] states that unemployment rates in Israel are released with quarterly frequency, and preliminary unemployment trend data is released with a two-month delay), the usefulness of Google Trends data is expected to increase. The higher frequency with which Google Trends data are available may also lead to increased usefulness in other situations not directly analysed in this paper.

An interesting question, not addressed by us, concerns the impact of using more sophisticated benchmark models. Such analysis would be particularly useful in the case of unemployment, both because in this case we achieve some prediction improvements when we add Google Trends data and because there are some more sophisticated models available in the literature (e.g., the Phillips curve). Following this line, Bughin [4] finds that contemporaneous inflation has a negative, statistically significant coefficient in a model for Belgian unemployment that also includes the volume for "unemployment" search queries, but he reports no predictive results. With respect to the impact of more sophisticated baseline models in the predictive ability of Google Trends, we may speculate that Google Trends data is unlikely to be a good proxy to other interesting variables that may explain unemployment, like inflation. So, we believe that the predictive ability of Google Trends variables comes from the intrinsic meaning of these variables, and not from them being good proxies to other excluded economic factors. Therefore, in the cases in which Google Trends data is currently leading to better predictions, we believe that nowcasting improvements would be achieved by using Google Trends data even with better benchmark models. However, this is a hypothesis that should be tested in future works.

In this work, we defined that in each extended model we would only include one Google Trends variable, whose values were directly gathered from Google Trends. We think that this approach will reduce the risk of achieving spurious results, but we also think it will not maximize the probability of achieving good forecasts. In the future we intend to try different approaches. These approaches include gathering data for a lot of search queries related to the variables being studied and aggregating it into a small number of indices. Such indices may be obtained by using principal component analysis (similarly to what was done by Vosen and Schmidt [18]) or by fitting the Google Trends series to a linear model (similarly to what was done by Carrière-Swallow and Labbé [5]).

Finally, in the future it will also be important to examine the reasons for the differences between the results we reached and the results presented by other authors. Methodological differences may explain some differences, but they are unlikely to explain everything. Cultural differences between countries and the use of different out-of-sample periods may also have a role in these differences.

## Acknowledgements

The work of Paulo Melo has been partially supported by Fundação para a Ciência e a Tecnologia (FCT) under project grant PEst-C/EEI/UI0308/2011.

# References

[1] ANIACAM. Asociación Nacional de Importadores de Automóviles, Camiones, Autobuses y Motocicletas. <http://www.aniacam.com/>. Accessed 3 March 2013.

[2] AutoInforma. Homologações, Estatísticas, Estudos e Serviços do Setor Automóvel. <http://www.autoinforma.pt/>. Accessed 14 October 2013.

[3] BestSellingCarsBlog. <http://bestsellingcarsblog.com/>. Accessed 3 March 2013.

[4] Bughin, J.R. (2011) 'Nowcasting' the Belgian Economy. SSRN working paper. <http://ssrn.com/abstract=1903791>. Accessed 7 July 2013.

[5] Carrière-Swallow, Y., & Labbé, F. (2011). Nowcasting with Google Trends in an Emerging Market. Journal of Forecasting, 32(4), 289-298

[6] Chadwick, M., & Şengül, G. (2012). Nowcasting Unemployment Rate in Turkey: Let's Ask Google. CBRT Working Paper, 12/18. <http://www.tcmb.gov.tr/research/discus/2011/WP1113.pdf>. Accessed 7 July 2013.

[7] Choi, H., & Varian, H. (2009). Predicting the Present with Google Trends. Technical Report, Google. <http://www.google.com/googleblogs/pdfs/google_predicting_the_present.pdf>‎. Accessed 7 July 2013.

[8] Choi, H., & Varian, H. (2009). Predicting Initial Claims for Unemployment Benefits. Technical Report, Google. [http://research.google.com/archive/papers/initialclaimsUS.pdf. Accessed 7 July 2013](http://research.google.com/archive/papers/initialclaimsUS.pdf.%20Accessed%207%20July%202013).

[9] Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. Economic Record, 88(S1), 2-9.

[10] Clark, T.E., & McCracken, M.W. (2001). Tests of Equal Forecast Accuracy and Encompassing for Nested Models. Journal of Econometrics, 105(1), 85-110.

[11] D'Amauri, F., & Marcucci, J. (2009). ‘Google it!’ Forecasting the US Unemployment Rate with a Google Job Search Index. MPRA Paper No. 18248. <http://mpra.ub.uni-muenchen.de/18248/1/MPRA_paper_18248.pdf>. Accessed 7 July 2013.

[12] Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomics statistics. Communications of the ACM, 48(11), 87-92.

[13] European Central Bank Statistical Data Warehouse. <http://sdw.ecb.europa.eu/>. Accessed 14 October 2013.

[14] Fondeur, Y., & Karamé, F. (2013). Can Google Data Help Predict French Youth Unemployment?. Economic Modelling, 30, 117-125.

[15] Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457, 1012-1014.

[16] Kholodilin, K., Podstawski, M., & Siliverstovs, B. (2010). Do Google Searches Help in Nowcasting Private Consumption? A Real-Time Evidence for the US. KOF Working Papers, 256. <http://kofportal.kof.ethz.ch/publications/download/1735/wp_256_1.pdf>. Accessed 7 July 2013.

[17] McCracken, M.W. (2007). Asymptotics for out of sample tests of Granger causality. Journal of Econometrics, 140(2), 719-752.

[18] Suhoy, T. (2009). Query indexes on a 2008 downturn: Israeli data. Bank of Israel Discussion Paper No. 2009.06. <http://www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf>. Accessed 7 July 2013.

[19] Toth, I., & Hajdu, M. (2012). Google as a tool for nowcasting household consumption: estimation on Hungarian data. 31th CIRET Conference, Vienna, Austria. <https://www.ciret.org/conferences/vienna_2012/papers/upload/p_218-934263.pdf>. Accessed 7 July 2013.

[20] UNRAE. Unione Nazionale Rappressentanti Autoveicoli Esteri. <http://www.unrae.it/>. Accessed 3 March 2013.

[21] Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google Trends. Journal of Forecasting, 30(6), 565-578.

[22] Vosen, S., & Schmidt, T. (2012). A monthly consumption indicator for Germany based on internet search query data. Applied Economics Letters, 19(7), 683-687.

# Appendix

## Google Trends Data versus Indicator evolution

In the figures within this appendix Google Trends data are compared with the evolution of the indicator variable.

### Unemployment

Fig. 1 Unemployment (Portugal) vs. Google Trends Data

Fig. 2 Unemployment (Spain) vs. Google Trends Data

Fig. 3 Unemployment (France) vs. Google Trends Data

Fig. 4 Unemployment (Italy) vs. Google Trends Data

### Car Sales

Fig. 5 Car Sales (Portugal) vs. Google Trends Data

Fig. 6 Car Sales (Spain) vs. Google Trends Data

Fig. 7 Car Sales (France) vs. Google Trends Data

Fig. 8 Car Sales (Italy) vs. Google Trends Data

1. INSP is the acronym for the Istituto Nazionale della Previdenza Sociale, the Italian National Institute for Social Security. [↑](#footnote-ref-1)