

Forecasting Spanish unemployment with Google Trends and dimension reduction techniques*

Rodrigo Mulero[†] Alfredo García-Hiernaux[‡]

June 4, 2020

Abstract

This paper presents a method to improve the one-step-ahead forecasts of the Spanish unemployment monthly series. To do so, we use a large number of potential explanatory variables extracted from searches in Google (Google Trends tool). Two different dimension reduction techniques are implemented to decide how to combine the explanatory variables or which ones to use. The results reveal an increase in predictive accuracy of 10-25%, depending on the dimension reduction method employed. A deep robustness analysis confirms this findings, as well as the relevance of using a large amount of Google queries together with a dimension reduction technique, when no prior information on which are the most informative queries is available.

Keywords: unemployment, forecasting, Google Trends, dimensionality reduction, RMSE.

JEL: C32, C52, C53

1 Introduction

Unemployment is an issue currently faced by the vast majority of economies. It is a red hot topic in studies carried out by economists and forecasters. Analyses are often based

*Alfredo García-Hiernaux gratefully acknowledges financial support from UCM-Santander grant ref. PR75/18-21570.

[†]Facultad de Ciencias Económicas. Universidad Complutense de Madrid. Campus de Somosaguas, 28223 Madrid (SPAIN). Email: rmulero@ucm.es

[‡]Corresponding author. Quantitative Economics Department and ICAE. Facultad de Ciencias Económicas. Universidad Complutense de Madrid. Campus de Somosaguas, 28223 Madrid (SPAIN). Email: agarciah@ucm.es, tel: (+34) 91 394 25 11.

on offering explanations, consequences and possible solutions to the problem, by different models that simplify real complexity.

A large number of jobless suffer constrains that generate problems of a macroeconomic nature, such as a decrease in consumption and investment which, eventually, affect GDP. Moreover, unemployment is also related to welfare problems as inequality and social exclusion. At least for these reasons, it is of most importance to correctly predict and evaluate unemployment in order to monitor its evolution, anticipate trend shifts, and design pro-employment policies.

Spain is a country with a high unemployment level compared with its peers peaking, in the recession of year 2013, to 5 million unemployed registered workers. For the purpose of this study, we use the official figures provided by the *Spanish Public Employment Service* (SEPE)¹. Typically, data unemployment is released with certain delay which means that the use of leading, or coincident, indicators will be useful for anticipating its evolution and improving its forecasts (see, e.g., Stock and Watson, 1993, for details on leading indicators).

With this in mind, the aim of this work is to propose some alternatives to univariate models for predicting the Spanish unemployment. We search for models which include additional, free of charge and available-to-everyone up-to-date information. We look for this information on the Internet search engines. These applications contain a large amount of information, available almost instantaneously, and reveals many aspects of the preferences of individuals through their search histories. In this article, without losing generality, we focus on searches in Google and, more specifically, we use one of its tools, known as Google Trends (GT). Our hypothesis is that, using updated search indices obtained from GT there is a large margin to improve the predictions of the Spanish unemployment provided by a univariate model.

However, any forecaster will soon discover that GT is not the panacea. As we will discuss in the next sections, some not trivial decision must be made when trying to optimize the information available on GT. This issue is treated in this paper in an application to the Spanish unemployment forecasting, but the procedures suggested could be applied in other

¹In Spain the main sources of data on unemployment comes from: (i) the Active Workforce Survey (EPA, in Spanish), provided quarterly by the National Statistics Institute, and (ii) the number of registered unemployed workers, provided monthly by the Spanish Public Employment Service (SEPE). We use the latter because of the higher publication frequency. The data has been downloaded from: <https://www.sepe.es/HomeSepe/que-es-el-sepe/estadisticas/empleo.html>

contexts.

Our results show that including GT queries to model Spanish unemployment yields an improvement in terms of forecasting accuracy relative to a univariate benchmark model that ranges 10%-25%. This gain depends on the way the GT information is optimized and is robust to the variables that affect the results of the forecasting exercise.

The paper is organized as follows. Section 2 provides a revision of the literature in the use of GT as explanatory variables, focusing on unemployment applications. Section 3 details the data employed in the analysis, paying particular attention to the GT queries and how those are generated and obtained. Section 4 presents the benchmark model and the proposed alternatives. The latter are based on data reduction methods, which are introduced in Section 5. Section 6 compares the forecasting results of the proposed models relative to the benchmark and Section 7 analyzes the robustness of the previous results. The last section highlights the main findings of the paper.

2 Background and literature

This line of research began in 2004 and has been gaining popularity since then, boosted by the increasing use of the Internet worldwide. Johnson et al. (2004) are the first researchers who exploit this information source. The authors analyze the relationship between access to health related pages and flu symptoms searches with the cases reported by the U.S. Center for Disease Control and Prevention. Also working on Google searches related to the flu, Eysenbach (2006) pioneered to include Google search data in order to improve the forecasts. Similarly, Ginsberg et al. (2009) studied the benefits of using Google searches to estimate outbreaks of influenza in the USA. The result was a tool for estimating and forecasting illnesses, which is known as *Google Flu Trends*. A major contribution of all these studies is the transformation of the benchmark models, with seriously delayed data, to those based on immediately available Google queries results.

The first researchers to look into the economic variables that can be related to these Internet searches are Choi and Varian (2009, 2012). Their hypothesis is that the Internet searches can be related to certain users preferences as, before making a decision (such as buying a car or looking for a job), many consumers carry out a prior Internet search. In their 2012

work, they use different GT categories related to unemployment to build an indicator for estimating the level of unemployment in real time, avoiding the delay incurred in the official figures. Likewise, Askitas and Zimmermann (2009), based on Ginsberg et al. (2009), innovate on the search for GT terms to obtain an indicator to predict unemployment. Coetaneous in time, Francesco D’Amuri has worked intensely in this field. D’Amuri (2009) analyzes how Google forecasts unemployment in Italy. He pays special attention to the potential selection bias in favor of young job seekers, as a consequence of being the greatest consumers of this tool. D’Amuri and Marcucci (2010) show the improvement in unemployment forecasts in the USA, when using an index generated by searches in GT. Finally, D’Amuri and Marcucci (2017) revisit the theory of the previous work, incorporate the effects of the 2008 financial recession and disaggregate the GT searches at a federal level. To sum up, all these works highlight the importance of including GT for estimating unemployment levels. Two very recent works for the USA with similar conclusions are Nagao et al. (2019) and Borup et al. (2019). The latter deserves more attention as it is likely the paper closest to ours. Contrary to most of the literature, the authors work with a large GT queries dataset and use dimension reduction techniques (soft-thresholding) to estimate (with random forest methods) employment models. Our paper differs to theirs in the queries, the samples, the dimension reduction methods applied (PCA and suggested model selection algorithm), the endogenous variable, the benchmark model and the inclusion of a deep robustness exercise.

On the other hand, the papers by Fondeur and Karamé (2013) and Naccarato et al. (2018) also analyze the unemployment by means of GT queries, but they focus, particularly, on youth unemployment in France and Italy, respectively. As far as we know, only Vicente et al. (2015) deal with the Spanish unemployment. However, the paper models and predicts the unemployment with only two GT queries plus a confident indicator. As a result, they do not cope with the dimension reduction problem. Additionally, their forecasting horizon is only of 12 periods and they do not vary the sample, which could make their conclusions sample-dependent.

Moreover, the use of GT queries and Internet searches, in general, as tools for modeling and forecasting has extended to distinct economic fields as: tourism (Pavlicek and Kristoufek, 2015; Siliverstovs and Wochner, 2018), inflation and GDP (Woo and Owen, 2019; Niesert et al., 2019; Poza and Monge, 2020), or even oil consumption (Yu et al., 2019).

Recently, two opposite mainstreams show up in the way this source of information should be used. While most of the authors stand up for the use of a few queries to reduce the noise in the analysis, see D’Amuri (2009), Fondeur and Karamé (2013), Vozlyublennaiia (2014), D’Amuri and Marcucci (2017), Naccarato et al. (2018) or Yu et al. (2019); some others favor the use of more queries, see Pan et al. (2012), Li et al. (2017) or Borup et al. (2019). From our viewpoint, the use of GT information to improve models and their forecasts has currently two problems to be solved: 1) what are the suitable queries to extract the most informative series? and, 2) how to comprime and filter this (sometimes huge) amount of information? Although both issues are related, our paper attempts to shade some light on the second one by applying two data reduction methods to a significant amount of GT queries results.

3 Data

This section details both, the unemployment data used as endogenous variable and the GT queries employed as potential explanatory variables.

3.1 Unemployment data

The unemployment series used in the paper is provided by the Spanish State Employment Service. It is released monthly during the first week of the next month and represents the number of people declaring to look for a job at a public employment office. The sample extends from January 2004 to September 2018, so that it covers business cycle expansions and recessions, with a total of 177 monthly observations.²

3.2 Google Trends (GT)

Google browser is the most used search engine on the planet. According to NetMarketShare (2019), the Google browser had in December 2018 a 77,1% and an 85,8% share in desktop computers and mobile devices, respectively. For this reason, GT represents a reliable estimation of all the searches made on the Internet.

GT is a search trends feature that shows how frequently a given search term is entered into Googles search engine, relative to the site’s total search volume over a given period of time. Google launched this tool in May 2006 and released an extension called *Google Search*

²The sample has been increased and modified in Section 7 to perform a robustness analysis.

Insight in August 2008. In 2012, both tools were merged to create the current version of GT, which is the one employed in this paper.

Mathematically, being $n(q, l, t)$ the number of searches for the query q , in the location l during the period t , the relative popularity (RP) of the query is expressed as:

$$RP_{(q,l,t)} = \frac{n(q, l, t)}{\sum_{q \in Q(l,t)} n(q, l, t)} \times \Pi_{(n(q,l,t) > \tau)}, \quad (1)$$

where $Q(l, t)$ is the set of all the queries made from l during t and $\Pi_{(n(q,l,t) > \tau)}$ is a dummy variable whose value is 1 when the query is sufficiently popular (the absolute number of search queries $n(q, l, t)$ exceeds τ) and 0 otherwise. The resulting numbers are then scaled on a range of 0-100 depending on the proportion of a topic with respect to the total number of all the search topics. So, the index of GT is expressed as in the following equation:

$$IGT_{(q,l,t)} = \frac{RP_{(q,l,t)}}{\max\{RP_{(q,l,t)}_{t \in \{1,2,..,T\}}\}} \times 100. \quad (2)$$

These indexes can be obtained from January 1st 2004 up to 36 hours prior to the search. GT excludes search data conducted by very few users and shows the topics of popular searches, assigning a zero in terms with a low search volume. In addition, searches performed repeatedly from the same machine in a short time period are removed. Finally queries containing apostrophes and other special characters are filtered.

We have conducted a search of 200 job query terms between January 2004 and September 2018. The method to choose these terms deserves some explanation. We have divided the terms of the searches in four groups: 1) series representing the queries related to leading job search applications (*e.g.*, *Infojobs*, *Jobday*, *LinkedIn*, etc); 2) searches related to Spanish unemployment centres, whether online, physical, public or private (*e.g.*, *Employment office*, *SEPE*, *Randstad*, etc); 3) queries related to standard job searching terms (*e.g.*, *Job offers*, *How to Find a Job*, *How to Find Work*, etc); 4) searches directly related to those companies that generate most employment in Spain (*e.g.*, *work in Inditex*, *Carrefour work*, *Santander job*). In order to complement these queries we also use the available GT tool called ‘related searches’ (see, Google, 2020), which allows us to download the queries made by the users related to the previous terms.

Of the 200 queries initially raised, we finally obtained data for 163 of them, as certain searches do not meet the conditions laid out by the GT index.³

4 Benchmark model and proposed alternatives

We follow Box et al. (2015) ARIMA methodology to obtain our benchmark model. The univariate monthly time series model considered is:

$$\Phi_P(B^s)\phi_p(B)\nabla^d\nabla_s^D u_t = \mu + \Theta_Q(B^s)\theta_q(B)a_t, \quad (3)$$

where $\phi_p(B) = 1 - \phi_1 B - \dots - \phi_p B^p$, $\theta_q(B) = 1 - \theta_1 B - \dots - \theta_q B^q$ are polynomials in B of degrees p and q , respectively, while $\Phi_P(B) = 1 - \Phi_1 B^s - \dots - \Phi_P B^{sP}$ and $\Theta_Q(B) = 1 - \Theta_1 B^s - \dots - \Theta_Q B^{sQ}$ are polynomials in B^s of degrees P and Q , respectively, and s is the seasonal frequency ($s = 12$ in our case). Moreover, μ is a constant, B is the lag operator so that $Bu_t = u_{t-1}$, $\nabla = (1 - B)$ is the difference operator and a_t is a sequence of independent Gaussian variates with mean zero and variance σ_a^2 . To meet the traditional Box et al. (2015) modelling requirements of stationarity and invertibility, we assume that all the zeros of the polynomials in B and B^s are outside the unit circle and have no common factors. This is often called as the seasonal autoregressive integrated moving average (SARIMA) form of the stochastic process u_t .

The identification using common tools (autocorrelation and partial autocorrelation functions and non-stationarity tests) leads us to a SARIMA(2, 1, 1) \times (0, 1, 1)₁₂ model. However, the residuals do not seem to represent a Gaussian white noise process due to an influential outlier in 2008. This is not surprising as this date corresponds to the global financial crisis, which hardly hit the Spanish unemployment.⁴ In order to model this outlier we include a step dummy variable defined as: $\xi^{08/03} = 1$, when $t < 2008/03$ and $\xi^{08/03} = 0$, otherwise.

The final model is presented in Equation (4a-4b), whose residuals do not evidence any

³All the information about the queries, GT data and multiple estimates are available from the authors upon request.

⁴Between March 2008 and January 2009 the number of unemployed increased by 44.6% in Spain.

sign of misspecification and are now compatible with the statistical assumptions on a_t .

$$u_t = \omega_0 \xi^{08/03} + \eta_t; \quad (4a)$$

$$(1 - \phi_1 B - \phi_2 B^2) \nabla \nabla_{12} \eta_t = (1 - \Theta_1 B^{12}) a_t. \quad (4b)$$

We will use this model as benchmark in the forecasting exercises in Sections 6 and 7.⁵

The alternative models are build on top of the benchmark. We propose to include additional explanatory series in Equation (4a) and keep the ARMA noise structure, in Equation (4b), as long as the statistical diagnosis does not reveal any sign of misspecification. Therefore, the proposed alternative models can be represented as the transfer funtion:

$$u_t = \omega_0 \xi^{08/03} + \sum_{i=0}^I \beta_i x_{it} + \eta_t; \quad (5a)$$

$$(1 - \phi_1 B - \phi_2 B^2) \nabla \nabla_{12} \eta_t = (1 - \Theta_1 B^{12}) a_t, \quad (5b)$$

where exogenous variables x_{it} , $i = 1, 2, 3, \dots, I$ will depend on the two different methods proposed to summarize the huge amount of information downloaded from GT. These two alternatives are detailed in the next section. The estimates for the benchmark model can be found in Table 2, for $I = 0$. As expected, the value for $\hat{\omega}_0$ is negative and highly significant, which implies that the financial crisis yielded a permanent increase in the Spanish unemployment of 79.770 people. The estimates of the ARMA parameters are also presented in Table 2, along with those of the alternative models.

5 Data reduction

There are basically two groups of methods to overcome the dimensionality curse arisen from the use of a large number of GT queries results. The first one exploits the redundant information of the data and creates a smaller set of new variables, each being a combination of the original ones, which replicates most of the information contained originally. These techniques are usually known as dimensionality reduction methods; see Van Der Maaten et al. (2009) for a complete survey. The second one encompasses the procedures that drop the less relevant variables from the original dataset by keeping the most explicative ones. This is often called

⁵The same model was identified if we use $\log(u_t)$ instead of u_t as the endogenous variable. The results of the paper do not change significantly when the log transformation is applied.

feature (or model) selection (see, e.g., Guyon and Elisseeff, 2003).

This section presents two methods (one of each of the previous groups) used to compare the forecasting performance of the Spanish unemployment, by reducing the amount of information obtained via GT. First, we briefly describe the Principal Component Analysis (PCA), one of the most widely used dimensionality reduction methods. Second, we propose an algorithm of feature selection adapted to our problem.

5.1 Principal Component Analysis

PCA is one of the most popular algorithms for dimensionality reduction. The reader unfamiliar with this procedure may consult Jolliffe (2002).

Broadly speaking, given the set of GT queries results (which is 163-dimension), PCA is the standard technique for finding the best -from a least-squares error sense- subspace of a lower dimension, I . The first principal component is the one that minimizes the distance between the data and their projection onto the principal component. The second principal component is chosen in the same way, but must be uncorrelated with the first one (or perpendicular to its direction), and so on.

In our case we compute the first 10 principal components, which accumulate around 70% of total variance of the GT result series. Interestingly, the two first components explain close to 50% of total variance. We stop at component 10 in an attempt to capture more information even if from the third one onward the marginal contribution to total variance is quite low; see Figure 1.

Figure 1 should be around here

The first alternative to the benchmark model consists of including the previous principal components as the explanatory variables x_{it} in Equation (5a). This means that x_{it} will be the i th-principal component, $i = 1, 2, \dots, I$ and $I = 1, 2, \dots, 10$, calculated from the set of variables obtained from GT ($N = 163$).⁶

⁶Additional information on the computation of the PCA, the weights and the correlation of the principal components with the original GT variables can be obtained from the authors upon request.

5.2 Model selection

Now we propose an alternative model based on a feature selection method. As before, we start with the original set of 163 queries. The process consists of estimating Model (5a-5b) with a potential explanatory variable, without lags, in Equation (5a). We do this for each variable in our set of 163 series. Therefore, a model is estimated for each variable. Once the estimation loop is finished, we sort the models by the lowest AIC criterion.⁷ This allows us to choose the best model out of all the estimates, obviously under the previous criterion. Next, we compute the one-step-ahead out-of-sample forecasts in the evaluation sample (2015/12 to 2018/09 in our case) based on the estimates of the selected model. We save these forecasts and calculate its corresponding Root Mean Squared Error (RMSE).⁸ If the RMSE is lower than the one obtained with the benchmark model, we repeat this process again, by adding a new explanatory variable to the previous model. For this, we rerun the model selection loop and choose the next variable whose model minimizes the information criterion. We repeat this process until the inclusion of a variable, whose model yields the lowest information criterion, does not provide a lower RMSE than that obtained with the benchmark model. Notice that the RMSE is only used to make the algorithm stop. Figure 2 depicts a diagram that illustrates the algorithm.⁹

The resulting models of this procedure to be compared against the PCA-based method and the benchmark is again defined as the transfer function (5a-5b), but in this case x_{it} is the variable chosen by the proposed feature selection method, with $i = 1, 2, 3, \dots, I$ and $I = 0, 1, 2, \dots$ until the algorithm stops.

Figure 2 should be around here

The first repetition of the loop defined in Figure 2 provides a ranking sorted by increasing AIC, of the explanatory variables obtained in the GT queries (see Appendix, Table 5). The

⁷Akaike's Information Criterion is computed as $AIC = E(-2L(\beta)) = T \log \hat{\sigma}_{MV}^2 + 2k$, where T is the sample size, $\hat{\sigma}_{MV}^2$ the maximum likelihood estimate of the innovations variance and k is the number of parameters to be estimated in the model, Akaike (1974). We perform the same exercise by using the Bayesian Information Criterion (BIC) and the final results do not vary.

⁸Let $\hat{a}_{l+1|l}$ with $l = 1, 2, \dots, L$ be a sequence of L one-step-ahead forecast errors, we compute the RMSE as $\left(\frac{1}{L} \sum_{l=1}^L \hat{a}_{l+1|l}^2\right)^{1/2}$.

⁹The code for the feature selection algorithm, the PCA as well as the forecasting analysis in Sections 6 and 7 (written in Python 3.6) is available from the authors upon request.

variable that provides the lowest AIC is the query for the term *LinkedIn*. The professional social network had three million users in Spain in 2012 (Jiménez, 2012). The inclusion of this variable considerably improves the model in terms of different information criteria and residual statistics. When repeating the exercise keeping *LinkedIn* in the model, as x_{1t} , the procedure leads to the selection of the query for the term *Carrefour job*, denoted by x_{2t} .¹⁰ Carrefour is a distribution company with 1,088 stores in Spain in 2019 (Osorio, 2019). The rest of the explanatory variables chosen and their order of selection are presented in the next section, Table 2.

6 Prediction evaluation

This section investigates the accuracy of the methods exposed previously when forecasting the Spanish unemployment in an out-of-sample validation of 34 periods. All the estimated models converge adequately and show no evidence of poor specification.

Table 1 presents the most common residual statistics for Model (5a-5b) by including cumulatively and sequentially: (i) the principal components given in Section 5.1, and (ii) the results for specific GT queries chosen by the features selection algorithm of Section 5.2. The main statistics are: Normality test (Jarque-Bera test), absence of autocorrelation (Ljung-Box test) and of heteroskedasticity (Goldfeld-Quandt test). Residuals do not evidence non-normality nor autocorrelation, although a few of them (when adding the principal components as explicative variables particularly) may be heteroskedastic. For the PCA-based models, p-values of the coefficients show poor explanatory power from the second principal component onward (except maybe the 6th one). Conversely, all the feature selection-based models have significant estimated coefficients (see Table 1, parameter $\hat{\beta}_I$).

Table 1 should be around here

Table 2 presents the estimates of the ARIMA parameters and the step-dummy variable, the AIC and the RMSE both, absolute and relative to the benchmark's. The coefficients $\hat{\omega}_0$ measuring the effect on the unemployment of the 2008 financial crisis shows a stable negative and significant value in all the models. When looking at the autoregressive polynomial coefficients ($\hat{\phi}_1$ and $\hat{\phi}_2$), the AR1 always provides a significant and positive coefficient while

¹⁰Notice that this is not the second variable found in the first iteration of the feature selection algorithm, see Table 5, but the first variable found in the second iteration.

the AR2 is only significant for the models that include only one explanatory variable, either the first principal component or the *LinkedIn* query. In turn, the estimated seasonal moving average ($\hat{\Theta}_1$) is always significant and negative. All these figures show the stability and robustness of the models, whose coefficients and statistics do not vary significantly when additional explanatory variables are sequentially incorporated.

Table 2 should be around here

Akaike’s criterion is considerably lower for the feature selection-based models (relative to PCA-based and benchmark models) and it decreases with each additional explanatory GT query. This was expected as a result of the design of the feature selection algorithm.

Regarding the forecasting accuracy, the RMSE of each of the models for the out-of-sample forecast period 2015/12 – 2018/09 is evaluated. In other words, a comparison of this error measure is made over a total of 33 one-step-ahead forecasts. Table 2 and Figure 3 show the RMSE improvement against the benchmark of the compared methodologies.

Figure 3 should be around here

The major advantage for the PCA-based models appears when $I = 3$, a gain close to 9% of predictive accuracy relative to benchmark’s. This result is compatible with the fact that from the third principal component, the relative explained variance of each additional component is marginal (see Figure 1). Regarding the feature selection-based model, the best improvement occurs with $I = 4$, *i.e.*, when the model incorporates GT queries for the terms *LinkedIn*, *Carrefour job*, *Ikea employment* and *How to Find a Job (HFJ)*. In such a case, the gain in terms of RMSE relative to benchmark’s is around 25%. Interestingly, the higher leap in forecast accuracy comes with the introduction of the GT search *LinkedIn*, which, individually, represents an improvement in predictive accuracy of 22.3%. The rest of the variables, instead, add a relative minor advance.¹¹ Furthermore, from the inclusion of the fifth variable, the forecasting precision begins to decrease almost linearly and when $I = 9$ it becomes even worse than the benchmark’s. That is why our algorithm (see Figure 2) stops here, when $I = 9$ as $RMSE_0 < RMSE_9$. We just include $I = 10$ with a comparison purpose.

¹¹Table 4 in the Appendix presents the estimates of the coefficients associated to each variable and model.

7 Robustness analysis

As the analysis in the previous section clearly demonstrates a much better forecasting performance of the feature selection-based model, we carry out a robustness analysis only for this methodology. We do so by varying all the variables that may have some influence in the result of our previous forecasting evaluation: (i) the estimation sample, (ii) the forecasting sample, (iii) the number of forecasting periods, and (iv) the date of the data extraction (as explained in Section 3.2 the GT index may differ for different download dates). Although with a few exceptions, the results shown in Table 3 are pretty unambiguous: the use of GT queries along with the proposed feature selection-based model clearly improves the forecasting accuracy in terms of RMSE relative to benchmark's. The best RMSE implies a gain of 31.3%, we found better forecasting results in 11 out of 14 models and the average benefit (of the 14 models) is close to 15%. Besides this main finding, some additional interesting facts can be withdrawn from this robustness check: (1) *LinkedIn* is definitively the key explanatory variable (when this term is not the best variable there is no predictive improvement); (2) the best RMSEs are usually obtained when adding extra explanatory variables to *LinkedIn*; (3) more explanatory variables (and better forecasting results) are found with the data downloaded in 2018/09 than in the series extracted in 2019/09; and (4) the lower is the number of forecasting periods, the higher is the forecasting accuracy.

Table 3 should be around here

While points (1) and (2) of the previous observed facts are related to the high impact of the *LinkedIn* GT search result on the forecasting of the Spanish unemployment, points (3) and (4) are likely related to the design of the exercise. Regarding the latter, in our paper the models are specified with the information given in the *Specification sample* (see Table 3) and although they are re-estimated with the observations added in each period, they are not re-specified. Thus, when the forecasting sample increases, the probability of finding a different model that better fits the new sample (*i.e.*, a new specification) increases. Our hypothesis is that, including a re-specification step when adding a new observation will yield even better forecasting results although, obviously, in exchange for a non-negligible increase in the computational cost. This is an open question for future research.

8 Final remarks

This paper studies whether additional information, collected in form of time series from queries applied to GT, improves in some extent the forecast accuracy of the Spanish unemployment, obtained with a univariate model. When conducting this analysis, two drawbacks show up: 1) what are the best queries one can introduce in GT, and 2) how to deal with the huge amount of information one can download from it. The first question is not the scope of this work but could be a subject of future research. In contrast, we compare two different ways to deal with close to 200 series downloaded: (i) the use of the standard techniques of principal components analysis, and (ii) a proposed algorithm of feature selection. The gains in RMSE relative to benchmark's are around 10% for the PCA-based model and 25% for the feature selection-based model. The improvement of the feature selection-based model proposed is confirmed in a robustness analysis. Compared to the literature, our gain is greater than the 15% obtained by Vicente et al. (2015) for the same endogenous variable (but different period) and greater than the common 10-19% range find by, *e.g.*, D'Amuri and Marcucci (2017) and Fondeur and Karamé (2013). The reason of this could be the large amount of GT data used and the dimension reduction techniques.

Besides the gain in predictive accuracy found to forecast the Spanish unemployment, the paper also shades some light to the discussion in the literature about using more or less explanatory variables. Our results on the robustness exercise shows that it is a good idea to introduce a small number of GT explanatory variables in the model. In our case, the best RMSE varies from 0 to 5 exogenous variables, depending on the sample and other parameters of the exercise. It certainly does on the endogenous variable to be analyzed as well.

Finally, in our application, the variable *LinkedIn* clearly arises as the best leading indicator among close to 200 series: it is the black cat in the dark room. Our feature selection-based model demonstrates its potential discovering the black cat. So, another relevant finding is that the larger is the dark room the higher is the probability of finding one or more black cats. At least, when no prior information is available on which are the most informative queries.

References

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

- Askitas, N. and Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2):107.
- Borup, D., Schütte, E. C. M., et al. (2019). In search of a job: Forecasting employment growth using google trends. Technical report, Department of Economics and Business Economics, Aarhus University.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Choi, H. and Varian, H. (2009). Predicting initial claims for unemployment benefits. *Google Inc*, pages 1–5.
- Choi, H. and Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88:2–9.
- D’Amuri, F. (2009). Predicting unemployment in short samples with internet job search query data. Technical report.
- D’Amuri, F. and Marcucci, J. (2010). ‘google it!’ forecasting the us unemployment rate with a google job search index. *FEEM working paper*.
- D’Amuri, F. and Marcucci, J. (2017). The predictive power of google searches in forecasting us unemployment. *International Journal of Forecasting*, 33(4):801–816.
- Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS biology*, 4(5):e157.
- Fondeur, Y. and Karamé, F. (2013). Can google data help predict french youth unemployment? *Economic Modelling*, 30:117–125.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012.
- Google (2020). Find related searches. <https://support.google.com/trends/answer/4355000>.
- Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.
- Jiménez, R. (2012). LinkedIn sets up in spain 9 years later. https://elpais.com/tecnologia/2012/03/27/actualidad/1332838659_13202.html.

- Johnson, H. A., Wagner, M. M., Hogan, W. R., Chapman, W. W., Olszewski, R. T., Dowling, J. N., Barnas, G., et al. (2004). Analysis of web access logs for surveillance of influenza. In *Medinfo*, pages 1202–1206.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York.
- Li, X., Pan, B., Law, R., and Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism management*, 59:57–66.
- Naccarato, A., Falorsi, S., Loriga, S., and Pierini, A. (2018). Combining official and google trends data to forecast the italian youth unemployment rate. *Technological Forecasting and Social Change*, 130:114–122.
- Nagao, S., Takeda, F., and Tanaka, R. (2019). Nowcasting of the us unemployment rate using google trends. *Finance Research Letters*, 30:103–109.
- NetMarketShare (2019). Browser market share. <https://netmarketshare.com/?options=>.
- Niesert, R. F., Oorschot, J. A., Veldhuisen, C. P., Brons, K., and Lange, R.-J. (2019). Can google search data help predict macroeconomic series? *International Journal of Forecasting*.
- Osorio, V. M. (2019). Carrefour multiplies by two the number of shops in spain in 5 years. <http://www.expansion.com/empresas/distribucion/2019/04/11/5cae569f268e3edb348b465c.html>.
- Pan, B., Wu, D. C., and Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*.
- Pavlicek, J. and Kristoufek, L. (2015). Nowcasting unemployment rates with google searches: Evidence from the visegrad group countries. *PloS one*, 10(5):e0127084.
- Poza, C. and Monge, M. (2020). A real time leading economic indicator based on text mining for the spanish economy. fractional cointegration var and continuous wavelet transform analysis. *International Economics*, In press.
- Silverstovs, B. and Wochner, D. S. (2018). Google trends and reality: Do the proportions match?: Appraising the informational value of online search behavior: Evidence from swiss tourism regions. *Journal of Economic Behavior & Organization*, 145:1–23.

- Stock, J. H. and Watson, M. W. (1993). A procedure for predicting recessions with leading indicators: Econometric issues and recent experience. In *Business cycles, indicators and forecasting*, pages 95–156. University of Chicago Press.
- Van Der Maaten, L., Postma, E., and Van den Herik, J. (2009). Dimensionality reduction: a comparative review. *Journal of Machine Learning Research*, 10:66–71.
- Vicente, M. R., López-Menéndez, A. J., and Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change*, 92:132–139.
- Vozlyublennaia, N. (2014). Investor attention, index performance, and return predictability. *Journal of Banking & Finance*, 41:17–35.
- Woo, J. and Owen, A. L. (2019). Forecasting private consumption with google trends data. *Journal of Forecasting*, 38(2):81–91.
- Yu, L., Zhao, Y., Tang, L., and Yang, Z. (2019). Online big data-driven oil consumption forecasting with google trends. *International Journal of Forecasting*, 35(1):213–223.

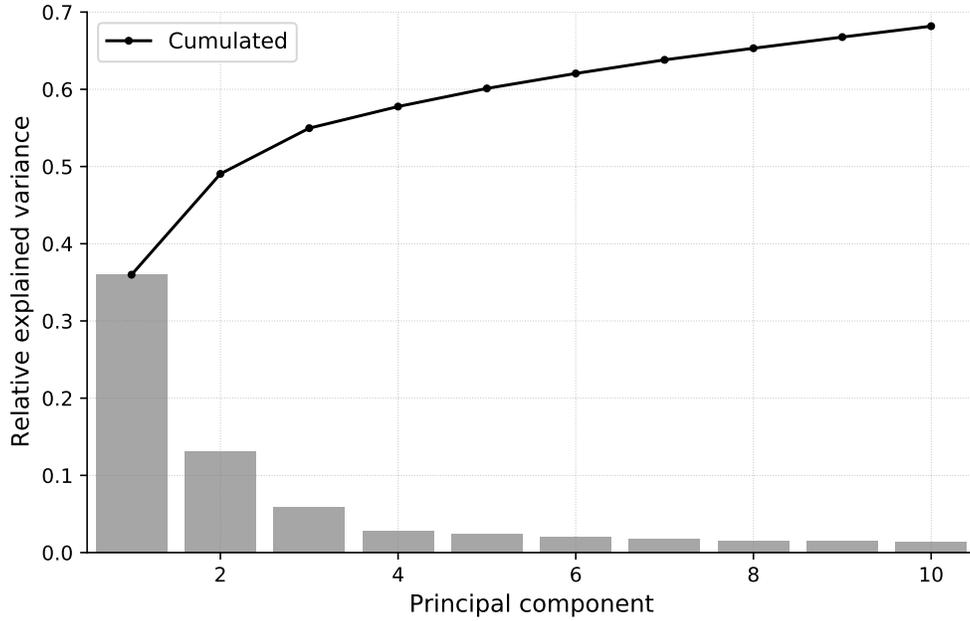


Figure 1: PCA analysis of the 163 GT queries.

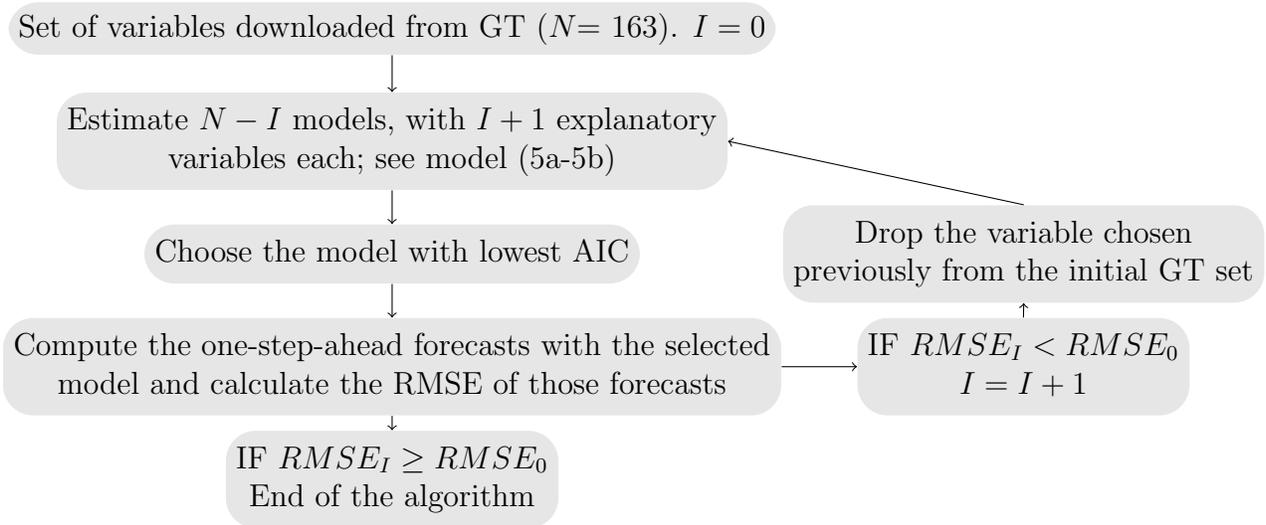


Figure 2: Feature selection algorithm. $I = 0$ corresponds to the benchmark univariate model.

Table 1: Estimates of the $\hat{\beta}_I$ coefficients and common residual tests for model I .

I	$\hat{\beta}_I$	Normality	No autocorrelation	Homoskedasticity
<i>Principal components-based models</i>				
1	.082 (.037)	.09 (.96)	37.21 (.60)	1.52 (.17)
2	.037 (.232)	.32 (.85)	37.32 (.59)	1.74 (.07)
3	.042 (.122)	.50 (.78)	38.13 (.55)	1.70 (.08)
4	-.046 (.116)	.78 (.68)	38.96 (.52)	1.93 (.03)
5	.028 (.457)	.68 (.71)	40.49 (.45)	1.70 (.08)
6	.026 (.065)	.44 (.80)	42.48 (.36)	1.75 (.07)
7	.010 (.684)	.45 (.80)	42.43 (.37)	1.91 (.03)
8	-.009 (.743)	.44 (.80)	42.02 (.38)	1.80 (.06)
9	.002 (.914)	.47 (.79)	41.87 (.39)	1.82 (.05)
10	-.001 (.972)	.46 (.79)	41.95 (.39)	1.82 (.05)
<i>Feature selection-based models</i>				
1	.205 (.001)	.94 (.62)	40.95 (.43)	1.61 (.12)
2	.025 (.041)	.93 (.63)	33.15 (.77)	1.71 (.08)
3	-.019 (.079)	1.62 (.45)	37.16 (.60)	1.80 (.06)
4	.014 (.019)	.62 (.73)	32.39 (.60)	1.78 (.06)
5	-.037 (.027)	1.12 (.57)	34.40 (.72)	1.80 (.06)
6	-.084 (.039)	2.40 (.30)	31.55 (.83)	1.70 (.08)
7	-.024 (.021)	.50 (.78)	32.18 (.81)	1.48 (.20)
8	.020 (.041)	.70 (.71)	30.92 (.85)	1.42 (.25)
9	.055 (.017)	1.49 (.48)	31.15 (.84)	1.21 (.54)
10	.014 (.060)	2.78 (.25)	36.04 (.65)	1.40 (.27)

The null hypothesis of the residual tests are: Normality, absence of autocorrelation and homoskedasticity. P-values are in parenthesis.

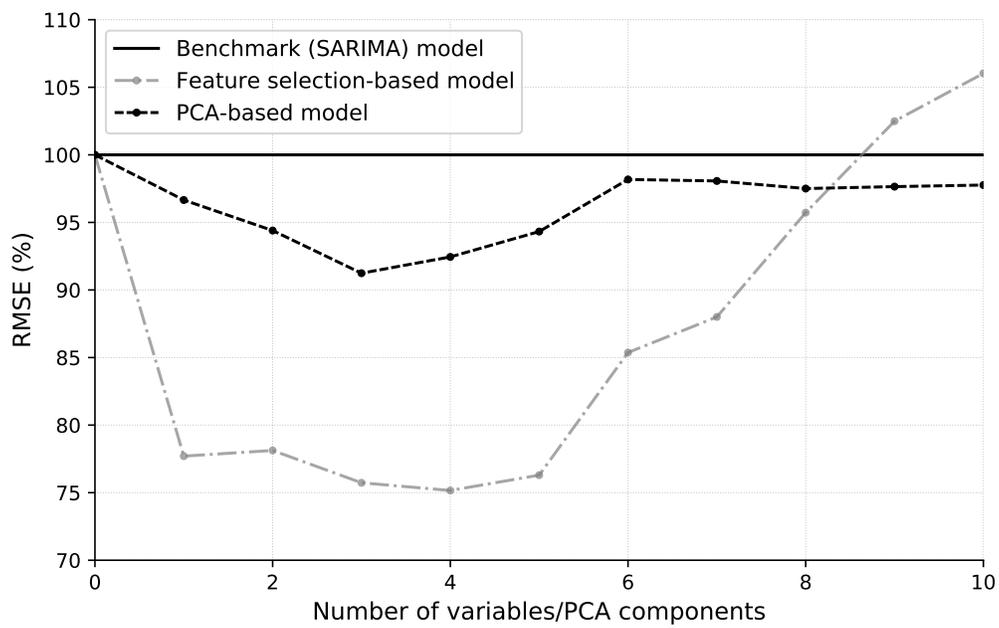


Figure 3: Forecasting accuracy of the models: RMSEs comparison relative to the benchmark model forecasts.

I	x_{It}	$\hat{\omega}_0$	$\hat{\Theta}_1$	$\hat{\phi}_1$	$\hat{\phi}_2$	AIC	RMSE	RMSE (%)
0	-	-7.977 (2.086)	-255 (.092)	.604 (.080)	.207 (.086)	657.751	2.432	100%
<i>Benchmark model</i>								
<i>Principal components-based models</i>								
1	Comp.1	-8.594 (1.884)	-.291 (.098)	.656 (.080)	.155 (.086)	654.235	2.350	96.66%
2	Comp.2	-8.148 (2.048)	-.279 (.097)	.654 (.080)	.160 (.084)	655.119	2.295	94.40%
3	Comp.3	-8.186 (2.064)	-.251 (.099)	.678 (.077)	.135 (.082)	656.210	2.219	<u>91.24%</u>
4	Comp.4	-7.959 (2.094)	-.232 (.099)	.682 (.075)	.134 (.081)	655.548	2.248	92.45%
5	Comp.5	-8.125 (2.136)	-.231 (.097)	.682 (.076)	.135 (.082)	658.332	2.294	94.32%
6	Comp.6	-8.353 (1.989)	-.223 (.102)	.726 (.077)	.091 (.084)*	655.937	2.387	98.18%
7	Comp.7	-8.451 (2.012)	-.216 (.101)	.719 (.078)	.099 (.085)*	659.543	2.385	98.07%
8	Comp.8	-8.525 (2.060)	-.213 (.102)	.723 (.078)	.095 (.086)*	659.639	2.371	97.51%
9	Comp.9	-8.519 (2.074)	-.211 (.102)	.724 (.078)	.094 (.086)*	659.686	2.374	97.65%
10	Comp.10	-8.517 (2.075)	-.211 (.102)	.723 (.080)	.095 (.086)*	658.672	2.377	97.76%
<i>Feature selection-based models</i>								
1	LinkedIn	-7.966 (1.932)	-.337 (.095)	.629 (.085)	.195 (.084)	647.824	1.889	77.67%
2	Carrefour	-8.086 (1.929)	-.336 (.097)	.703 (.082)	.120 (.084)*	644.086	1.900	78.13%
3	Ikea	-7.815 (1.901)	-.361 (.097)	.724 (.084)	.103 (.085)*	641.162	1.841	75.70%
4	HFE**	-7.820 (2.012)	-.335 (.100)	.746 (.083)	.084 (.087)*	636.719	1.828	75.16%
5	HFJ**	-7.683 (1.863)	-.312 (.097)	.777 (.078)	.060 (.084)*	629.927	1.855	76.27%
6	Milanuncios	-7.791 (1.787)	-.297 (.100)	.765 (.087)	.081 (.091)*	626.675	2.076	85.36%
7	Telefonica	-8.556 (1.865)	-.311 (.101)	.786 (.085)	.063 (.089)*	623.215	2.140	87.99%
8	Lidl	-8.213 (1.637)	-.278 (.096)	.802 (.084)	.054 (.087)*	619.549	2.328	95.72%
9	Mercadona	-8.792 (1.562)	-.228 (.099)	.824 (.084)	.035 (.089)*	615.147	2.492	102.47%
10	Volkswagen	-9.129 (1.476)	-.196 (.106)	.886 (.085)	-.014 (.091)*	610.602	2.578	106.00%

Table 2: Estimates of the coefficients in Equation (5b). Standard errors are in parenthesis. One asterisk (*) denotes non significant values at 10%. Two asterisks (**) denote acronyms: HFE and HFJ stand for *How to Find Employment* and *How to Find a Job*, respectively. The best RMSE for each model is underlined. The best RMSE overall is in bold font.

Exercise number	Specification sample		Number of forecasts	End of forecast	Data downloaded	First variable found	Best variable ^a	Variables num	Best RMSE	
	Start	End							Variable	Relative ^b (%)
1	2004/01	2015/12	33	2018/09	2018/09	LinkedIn	LinkedIn	4	HFJ	75.16
2	2005/01	2016/12	33	2019/09	2019/09	LinkedIn	LinkedIn	1	LinkedIn	89.50
3	2006/01	2015/12	33	2018/09	2018/09	LinkedIn	LinkedIn	1	LinkedIn	77.55
4	2006/01	2016/12	33	2019/09	2019/09	LinkedIn	LinkedIn	1	LinkedIn	89.55
5	2008/01	2015/12	33	2018/09	2018/09	LinkedIn	LinkedIn	1	LinkedIn	75.08
6	2008/01	2016/12	33	2019/09	2019/09	LinkedIn	LinkedIn	1	LinkedIn	93.60
7	2010/01	2015/12	33	2018/09	2018/09	Job offers	LinkedIn	5	MediaMarkt job	76.57
8	2010/01	2016/12	33	2019/09	2019/09	Carrefour job	LinkedIn	1	LinkedIn	90.50
9	2004/01	2013/12	33	2016/09	2018/09	Carrefour job	-	0	-	100.0
10	2004/01	2013/12	48	2017/12	2019/09	LinkedIn	LinkedIn	1	LinkedIn	85.65
11	2004/01	2014/12	48	2018/12	2019/09	Cabify job	-	0	-	100.0
12	2004/01	2015/12	12	2016/12	2018/09	LinkedIn	LinkedIn	1	LinkedIn	68.67
13	2005/01	2016/12	12	2017/12	2018/09	LinkedIn	LinkedIn	5	LIDL job	82.23
14	2006/01	2017/12	12	2018/12	2019/09	LinkedIn	-	0	-	100.0

^a Best variable is the variable with highest impact on RMSE reduction. ^b Relative to benchmark SARIMA model specified in the corresponding sample. The best RMSE overall is in bold font.

Table 3: Robustness analysis. RMSE and other indicators for various models.

Appendix

i	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\sigma}_{\hat{\alpha}_i}^2$
0	-	-	-	-	-	-	-	-	-	-	8.116
1	.205 (.060)	-	-	-	-	-	-	-	-	-	7.373
2	.189 (.055)	.025 (.012)	-	-	-	-	-	-	-	-	7.055
3	.197 (.054)	.024 (.012)	-	-	-	-	-	-	-	-	6.781
4	.200 (.050)	.034 (.012)	-.019 (.011)	.014 (.006)	-	-	-	-	-	-	6.467
5	.199 (.048)	.038 (.010)	-.023 (.010)	.018 (.005)	-.037 (.017)	-	-	-	-	-	6.053
6	.219 (.055)	.039 (.013)	-.023 (.010)	.020 (.005)	-.040 (.016)	-.084 (.041)	-	-	-	-	5.821
7	.227 (.054)	.043 (.012)	-.024 (.010)	.020 (.005)	-.047 (.015)	-.094 (.040)	-.024 (.010)	-	-	-	5.585
8	.208 (.054)	.040 (.011)	-.024 (.009)	.024 (.005)	-.048 (.015)	-.092 (.036)	-.025 (.010)	.020 (.010)	-	-	5.321
9	.180 (.054)	.044 (.011)	-.023 (.008)	.025 (.005)	-.052 (.014)	-.100 (.035)	-.025 (.010)	.022 (.010)	.055 (.023)	-	5.096
10	.173 (.050)	.045 (.010)	-.023 (.008)	.029 (.005)	-.056 (.013)	-.103 (.034)	-.027 (.010)	.019 (.009)	.074 (.024)	.014 (.007)	4.934

Table 4: Estimates of the β_i coefficients in Equation (5a). β_i for $i = 1, 2, \dots, I$ are the coefficients corresponding to variables x_{it} for $i = 1, 2, \dots, I$. Standard errors are in parenthesis. One asterisk (*) denotes non significant values at 10%.

Position	Name	AIC	p-value for $\hat{\beta}_1$	$\hat{\sigma}_{a_t}^2$
1	<i>LinkedIn</i>	647.82	.001	7.37
2	Job offers	653.47	.004	7.73
3	<i>Carrefour</i> work	653.53	.039	7.73
4	SEPE	653.72	.010	7.72
5	<i>Nortempo</i> employment	654.15	.042	7.78
\vdots	\vdots	\vdots	\vdots	\vdots
161	Work in <i>Carrefour</i>	659.75	.985	8.11
162	<i>La Caixa</i> work	659.75	.986	8.11
163	Work in <i>Telefonica</i>	659.75	.994	8.11

Table 5: Ranking of some variables after the first round of the algorithm ($I = 0$) for feature selection.