SERIEs (2021) 12:329–349
https://doi.org/10.1007/s13209-021-00231-x

ORIGINAL ARTICLE

Forecasting Spanish unemployment with Google Trends and dimension reduction techniques

Rodrigo Mulero[1] · **Alfredo García-Hiernaux**[1,2]

Received: 19 November 2020 / Accepted: 2 April 2021 / Published online: 19 April 2021

Abstract
This paper presents a method to improve the one-step-ahead forecasts of the Spanish unemployment monthly series. To do so, we use numerous potential explanatory variables extracted from searches in Google (Google Trends tool). Two different dimension reduction techniques are implemented (PCA and Forward Stepwise Selection) to decide how to combine the explanatory variables or which ones to use. The results of a recursive forecasting exercise reveal a statistically significant increase in predictive accuracy of 10–25%, depending on the dimension reduction method employed. A deep robustness analysis confirms these 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 Classification 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

Alfredo García-Hiernaux gratefully acknowledges financial support from UCM-Santander Grant Ref. PR75/18-21570.

✉ Alfredo García-Hiernaux
agarciah@ucm.es

Rodrigo Mulero
rmulero@ucm.es

[1] Facultad de Ciencias Económicas, Universidad Complutense de Madrid, Campus de Somosaguas, 28223 Madrid, Spain

[2] Quantitative Economics Department, ICAE, Madrid, Spain

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

Numerous 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 2013 recession, to 5 million registered unemployed 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 to anticipate 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 simple 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 reveal many aspects of the individuals' preferences through their search histories. In this paper, without losing generality, we focus on searches in Google. 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 suitable 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 gathered from GT. This issue is treated in the paper in an application to the Spanish unemployment forecasting, although the procedures suggested could be applied in other contexts.

By means of a recursive forecasting exercise, we find that a SARIMA model with additional GT queries, applied to the Spanish unemployment series and relative to a univariate benchmark model, yields a statistically significant improvement in terms of forecasting accuracy that ranges 10–25%. This gain depends on the way the GT information is treated, with Principal Components Analysis (PCA) or Forward Stepwise Selection (FSS), and is robust to the variables that affect the results of the forecasting exercise. In our application, FSS outperforms PCA.

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, the proposed alternatives and their relation with other common methods in the

[1] In 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.

literature. The latter are based on data reduction methods, which are introduced in Sect. 5. Section 6 compares the forecasting results of the proposed models relative to the benchmark and Sect. 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. Coeval 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 (2009) 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 and Schütte (2020). 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 (particularly soft-thresholding) to estimate employment models with random forest methods. Our paper differs to theirs in the queries, the samples, the dimension reduction methods applied (PCA and a suggested FSS), 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 with the GT approach. However, their 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. 2020; 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), Vozlyublennaia (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) and Borup and Schütte (2020). 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 compress and filter this (sometimes huge) amount of information. Although both issues are related, our paper attempts to shed 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 (SEPE 2019). 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

[2] The sample has been increased and modified in Sect. 7 so that a robustness analysis can be performed.

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 Google's 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 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 (Google 2020b).

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:

$$\mathrm{RP}_{(q,l,t)} = \frac{n(q,l,t)}{\Sigma_{q\in Q(l,t)} n(q,l,t)} \times \Pi_{\left(n(q,l,t)>\tau\right)}, \tag{1}$$

where $Q(l, t)$ is the set of all the queries made from l during t and $\Pi_{\left(n(q,l,t)>\tau\right)}$ 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 defined as:

$$\mathrm{IGT}_{(q,l,t)} = \frac{\mathrm{RP}(q,l,t)}{\max\{\mathrm{RP}(q,l,t)_{t\in 1,2,\ldots,T}\}} \times 100. \tag{2}$$

These indexes can be obtained from January 1st 2004 up to 36 h 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 queries between January 2004 and September 2018. The method to choose these terms deserves some explanation. We divide 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 centers, 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); and, finally, (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 2020a), which allows us to download the queries made by the users related to the previous terms.

From 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]

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

4 Benchmark model and proposed alternatives

We follow Box and Jenkins (1976) ARIMA methodology to obtain our benchmark model. The univariate monthly time series model considered is:

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

where $\phi_p(B) = 1-\phi_1 B-\cdots-\phi_p B^p, \theta_q(B) = 1-\theta_1 B-\cdots-\theta_q B^q$ are polynomials in B of degrees p and q, respectively, while $\Phi_P(B) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{sP}$ and $\Theta_Q(B) = 1 - \Theta_1 B^s - \cdots - \Theta_Q B^{sQ}$ 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 uncorrelated Gaussian variates with mean zero and variance σ_a^2. To meet the traditional Box and Jenkins' 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 (graphics, autocorrelation and partial autocorrelation functions, and unit root 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 beginning of the global financial crisis, which hardly hit 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 Eqs. (4a–4b), whose residuals do not evidence any sign of misspecification and are now compatible with the statistical assumptions on a_t:

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

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

We will use this model as benchmark in the forecasting exercises in Sects. 6 and 7.[5] Although it is not the purpose of the analysis, some theoretical implications can be drawn from the empirical identification of this model. As the nonstationary tests do not reject the unit root hypothesis, the hysteresis theory (see Blanchard and Summers 1987), which indicates that shock effects on the unemployment will persist because of the rigidity of the labor market, cannot be rejected either. This results in line with the analysis, also for Spain, performed by Romero-Avila and Usabiaga (2007), and Cheng et al. (2014).

The alternative models are build on top of the benchmark. We propose to include additional explanatory series in Eq. (4a) and keep the ARMA noise structure, in

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

[5] Note that the model introduces a step in levels in Eq. (4a), which corresponds to an impulse in its differenced version. The 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.

Eq. (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 function:

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

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

where exogenous variables x_{it}, $i = 1, 2, 3, \ldots, I$ 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 level of 79.770 people. The estimates of the ARMA parameters are also presented in Table 2, along with those of the alternative models.[6]

5 Data reduction

There are basically two groups of methods to overcome the dimensionality curse arisen from the use of many 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 explanatory ones. This is often called variable (or feature) 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 a Forward Stepwise Selection algorithm 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 its projection onto the principal component.

[6] All the models have been estimated by maximum likelihood through their state-space equivalent form (see, Casals et al. 2016) and using adequate initial conditions (Garcia-Hiernaux et al. 2009).

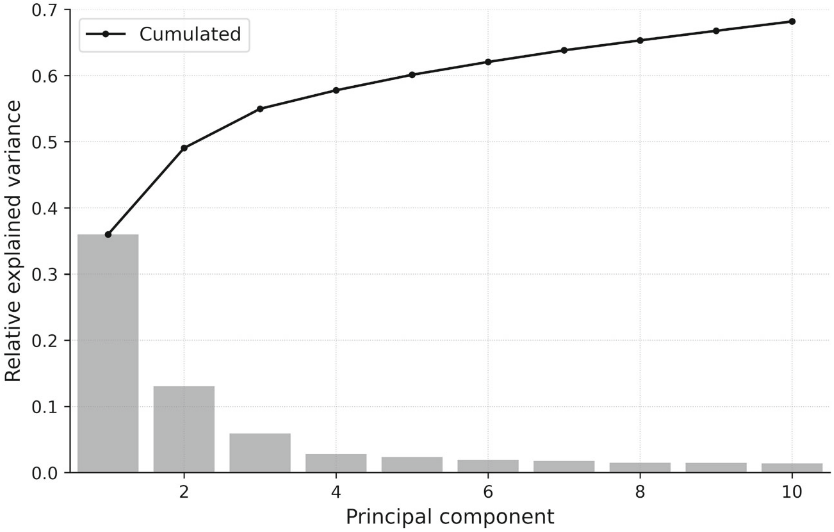

Fig. 1 PCA analysis of the 163 GT queries

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 Fig. 1.

As it is often the case for applications with many variables (i.e., with high dimension), it is difficult to interpret the principal components obtained, as they are often a linear combination of many variables. However, we will try to give some insight on the three first principal components by looking at their correlations with the original queries. The first principal component is positively and highly (linearly) related to job-search apps (e.g., *LinkedIn, Indeed, Milanuncios*), unemployment centers (e.g., *Randstad, SEPE*) and some, but not many, queries related to companies (as *work for Decathlon* or *Lidl*). The second principal component is mostly related to queries seeking jobs in particular firms (*as hiring in Carrefour, Eulen, CaixaBank, Mercadona*, etc). Finally, the third principal component mostly represents the queries related to standard job searching terms (e.g., *employment, public employment, work, job vacancies*, etc).[7]

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

[7] Additional information on the computation of the PCA, their weights, the correlation of the principal components with the original GT queries and its visualization (heatmap) are available from the authors upon request.

this method is similar to the Principal Component Regression (PCR). In PCR, the principal components of the explanatory variables are used as regressors. Particularly, as we do here, one often uses a few principal components for regression, making PCR a shrinkage procedure. The main advantage of our proposal with respect to original PCR is that we additionally incorporate a model for the noise. Certainly, Eq. (5a) can be considered as a PCR when x_{it} for $i = 1, \ldots, I$ is a subset of the principal components previously calculated. However, our proposal also includes Eq. (5b) in the model, so that ω_0 and the βs can jointly be estimated with ϕ_1, ϕ_2 and Θ_1, which capture the remaining autocorrelation of the residuals.

5.2 Forward stepwise selection

Now we propose an alternative model based on a FSS 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 Eq. (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.[8] 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).[9] 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 procedure.[10]

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

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

[8] Akaike's Information Criterion is computed as AIC $= E\big[-2L(\beta)\big] = T \log \hat{\sigma}^2_{\text{ML}} + 2k$, where T is the sample size, $\hat{\sigma}^2_{\text{MV}}$ 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.

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

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

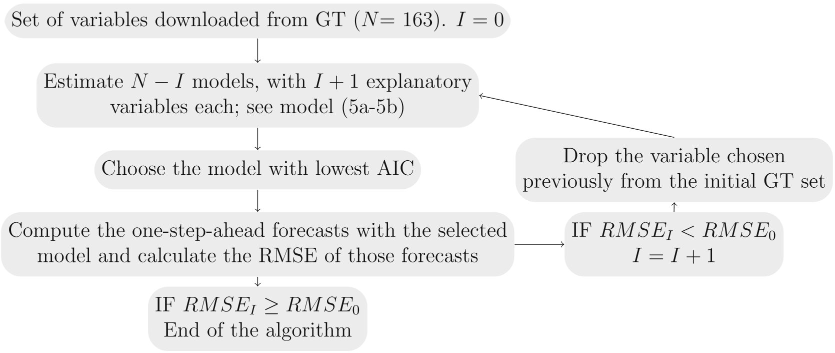
Set of variables downloaded from GT (N= 163). $I = 0$
Estimate $N - I$ models, with $I + 1$ explanatory variables each; see model (5a-5b)
Choose the model with lowest AIC
Compute the one-step-ahead forecasts with the selected model and calculate the RMSE of those forecasts
IF $RMSE_I \geq RMSE_0$ End of the algorithm
IF $RMSE_I < RMSE_0$ $I = I + 1$
Drop the variable chosen previously from the initial GT set

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

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}.[11] 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.

5.3 Alternative procedures

Other tools might be used to select the forecasting explanatory variables and models. Some related alternatives in the literature are the Lasso regression (Tibshirani 1996), the Model Confident Set (MCS, Hansen et al. 2011) and the Bayesian Model Averaging (BSA, Hoeting et al. 1999), although others can also be found. Each of these procedures has its pros and cons with respect to our methods, which were chosen mainly because of its conceptual simplicity (SARIMAX models, PCA and AIC criterion are well-known for most forecasters) at an affordable computational cost. In this sense, Lasso regression could be a good alternative as it works well for a big number of potential explanatory variables and its computational cost is low. However, it has two drawbacks in this application: (i) as far as we know, there is currently no procedure to estimate the Lasso regression in a transfer function model like (5a–5b) that include MA terms,[12] and (ii) as we do not have a large sample size ($T = 177$) and the number of potential explanatory variables is big ($N = 163$), we cannot directly perform a Lasso regression as we will not have enough degrees of freedom to estimate. Therefore, some kind of dimension reduction technique should be previously used anyway. Regarding the MCS, although it is a powerful tool for model comparison, it does not fit our problem as well as our procedures do. That is because of the slight difference between model selection and feature selection. MCS is a model selection procedure. In this application

[11] Notice 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.

[12] Indeed, our models are estimated by maximum likelihood. Obviously, an alternative ADL model could be used but then the number of lags will be high if one wants to correctly capture the seasonal effect captured by the MA factor.

there would be a huge number of models to compare (even restricting to only two GT queries in model (5a–5b) will yield $\binom{163}{2} = 13{,}203$ models to compare!). For that reason, it seems logical to first deal with the feature selection problem, and then chose the best model. Obviously, the MCS (or BMA) can be used to select the best model among those given in Table 2, but for the sake of simplicity we decide to use the most common out-of-sample RMSE comparison for this purpose. In turn, in order to apply BMA, a prior distribution over the considered models must be specified, which is usually non-trivial. Similarly to MCS, in our application the number of models under consideration is huge and the computational cost of BMA will become enormous.

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. To this aim we use a recursive (expanding) forecasting scheme. In the exercise, all the estimations converge adequately and no model shows 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 Sect. 5.1, and (ii) the results for specific GT queries chosen by the features' selection algorithm of Sect. 5.2. The main statistics shown 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 explanatory 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 2 presents the estimates of the SARIMA 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 the AR2 is only significant for the models that include just one explanatory variable, either the first principal component or the *LinkedIn* query. In turn, the estimated seasonal moving average ($\hat{\Theta}_1$) is always highly significant and negative. All these values show the stability and robustness of the models, whose coefficients and statistics do not vary significantly when additional explanatory variables are sequentially incorporated.

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

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

I	$\hat{\beta}_I$	Normality	No autocorrelation	Homoskedasticity
	Principal components-based models			
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)
	Feature selection-based models			
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 parentheses

and Fig. 3 show the RMSE improvement of the compared methodologies against the benchmark.

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 Fig. 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.[13] 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.

[13] Table 4 in "Appendix" presents the estimates of the coefficients associated to each variable and model.

Table 2 Estimates of the coefficients in Eq. (5b)

I	x_{It}	Estimates				AIC	RMSE		DM
		$\hat{\omega}_0$	$\hat{\Theta}_1$	$\hat{\phi}_1$	$\hat{\phi}_2$		Value	%	p value
	Benchmark model								
0	–	−7.977 (2.086)	−.255 (.092)	.604 (.080)	.207 (.086)	657.751	2.432	100	–
	Principal components-based models								
1	Comp.1	−8.594 (1.884)	−.291 (.098)	.656 (.080)	.155 (.086)	654.235	2.350	96.66	.189
2	Comp.2	−8.148 (2.048)	−.279 (.097)	.654 (.080)	.160 (.084)	655.119	2.295	94.40	.071
3	Comp.3	−8.186 (2.064)	−.251 (.099)	.678 (.077)	.135 (.082)	656.210	2.219	91.24	.036
4	Comp.4	−7.959 (2.094)	−.232 (.099)	.682 (.075)	.134 (.081)	655.548	2.248	92.45	.084
5	Comp.5	−8.125 (2.136)	−.231 (.097)	.682 (.076)	.135 (.082)	658.332	2.294	94.32	.142
6	Comp.6	−8.353 (1.989)	−.223 (.102)	.726 (.077)	.091 (.084)*	655.937	2.387	98.18	.379
7	Comp.7	−8.451 (2.012)	−.216 (.101)	.719 (.078)	.099 (.085) *	659.543	2.385	98.07	.368
8	Comp.8	−8.525 (2.060)	−.213 (.102)	.723 (.078)	.095 (.086)*	659.639	2.371	97.51	.336
9	Comp.9	−8.519 (2.074)	−.211 (.102)	.724 (.078)	.094 (.086)*	659.686	2.374	97.65	.348
10	Comp.10	−8.517 (2.075)	−.211 (.102)	.723 (.080)	.095 (.086)*	658.672	2.377	97.76	.364

Table 2 continued

I	x_{It}	Estimates				AIC	RMSE		DM
		$\hat{\omega}_0$	$\hat{\Theta}_1$	$\hat{\phi}_1$	$\hat{\phi}_2$		Value	%	*p* value
	Feature selection-based models								
1	LinkedIn	−7.966 (1.932)	−.337 (.095)	.629 (.085)	.195 (.084)	647.824	1.889	77.67	.012
2	Carrefour	−8.086 (1.929)	−.336 (.097)	.703 (.082)	.120 (.084)*	644.086	1.900	78.13	.010
3	Ikea	−7.815 (1.901)	−.361 (.097)	.724 (.084)	.103 (.085)*	641.162	1.841	75.70	.008
4	HFE**	−7.820 (2.012)	−.335 (.100)	.746 (.083)	.084 (.087)*	636.719	**1.828**	**75.16**	.009
5	HFJ**	−7.683 (1.863)	−.312 (.097)	.777 (.078)	.060 (.084)*	629.927	1.855	76.27	.012
6	Milanuncios	−7.791 (1.787)	−.297 (.100)	.765 (.087)	.081 (.091)*	626.675	2.076	85.36	.063
7	Telefonica	−8.556 (1.865)	−.311 (.101)	.786 (.085)	.063 (.089)*	623.215	2.140	87.99	.097
8	Lidl	−8.213 (1.637)	−.278 (.096)	.802 (.084)	.054 (.087)*	619.549	2.328	95.72	.336
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	–

Standard errors are in parentheses. 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

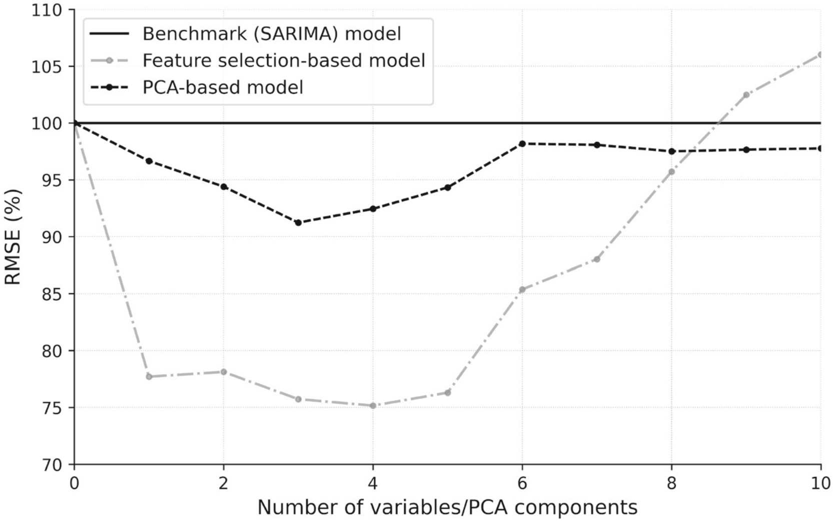

Fig. 3 Forecasting accuracy of the models: RMSEs comparison relative to the benchmark model forecasts

That is why our algorithm (see Fig. 2) stops here, when $I = 9$ as $\text{RMSE}_0 < \text{RMSE}_9$. We just include $I = 10$ for comparison purpose.

A small sample size (it reaches 176 observations for the last forecast) could imply that an important part of the RMSE is noise due to estimation error. Moreover, the out-of-sample size is not very large either (33 forecasts), which adds more uncertainty to the RMSE. To incorporate the latter source of uncertainty in our forecast evaluation, Table 2 offers in its last column the *p* value of the Diebold and Mariano (1995) test. The null hypothesis is that the two predictions, those obtained from the benchmark and the corresponding alternative model, have the same accuracy. Accordingly, a small *p* value evidences that the suggested model predicts better than the benchmark with a particular significance level. Thus, the PCA-based model including the first three components outperforms the benchmark with a 3.6% significance level, while the FSS-based model with the first three (or four) variables beats the benchmark with a 0.9% significance level.

Notice that the Diebold and Mariano test is appropriate in this application even if it does not account for parameter estimation error (see, e.g., West 2006; Escanciano and Olmo 2010). Although here we apply this test to forecast provided by estimated (not known) models and, therefore, they are subject to parameter uncertainty, in all the cases treated the out-of-sample size is small relative to the in-sample size. This makes the extra term related to parameter estimation error, which is not accounted by the limiting variance derived by Diebold and Mariano (1995), to vanish asymptotically (see, West 2006). Thus, assuming there are no estimation effects is expected to be a good approximation in our forecasting evaluation exercise.

7 Robustness analysis

As the analysis in the previous section 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 the forecasting evaluation: (i) the specification sample, (ii) the forecasting sample, (iii) the number of forecasting periods, and (iv) the date of the data extraction (as explained in Sect. 3.2, 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 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%. In terms of Diebold and Mariano's test, 7 models beat the benchmark with a 5% significance level. 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 with the series extracted in 2019/09; and (4) the lower is the number of forecasting periods, the higher is the forecasting accuracy.

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 (this is, indeed, a recursive forecasting scheme), 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 better specification) increases. For instance, our FSS algorithm in Fig. 2 chooses *LinkedIn* and *Ikea* as the first two best queries to be included in the model. The recursive forecasting scheme implies to update each observation, re-estimate and produce a new forecast with that model. So, in this exercise we do not rerun our FSS algorithm with each update. Our hypothesis is that, including a re-specification step when adding a new observation (i.e., rerunning the FSS algorithm to search for the best model with each update) will yield even better forecasting results. This will be, obviously, in exchange for a non-negligible increase in the computational cost, and remains as 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 questions arise: (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

Table 3 Robustness analysis

Exercise number	Specification sample		Forecast number	End of forecast	Data downloaded	First variable found	Best variable[a]	Best RMSE			DM
	Start	End						Var. num	Variable	%[b]	*p* value
1	2004/01	2015/12	33	2018/09	2018/09	LinkedIn	LinkedIn	4	HFJ	75.2	.013
2	2005/01	2016/12	33	2019/09	2019/09	LinkedIn	LinkedIn	1	LinkedIn	89.5	.209
3	2006/01	2015/12	33	2018/09	2018/09	LinkedIn	LinkedIn	1	LinkedIn	77.5	.009
4	2006/01	2016/12	33	2019/09	2019/09	LinkedIn	LinkedIn	1	LinkedIn	89.5	.212
5	2008/01	2015/12	33	2018/09	2018/09	LinkedIn	LinkedIn	1	LinkedIn	75.1	.017
6	2008/01	2016/12	33	2019/09	2019/09	LinkedIn	LinkedIn	1	LinkedIn	93.6	.300
7	2010/01	2015/12	33	2018/09	2018/09	Job offers	LinkedIn	5	MediaMarkt job	76.6	.003
8	2010/01	2016/12	33	2019/09	2019/09	LinkedIn	LinkedIn	1	LinkedIn	90.5	.059
9	2004/01	2013/12	33	2016/09	2018/09	Carrefour job	–	0	–	–	–
10	2004/01	2013/12	48	2017/12	2019/09	LinkedIn	LinkedIn	1	LinkedIn	85.6	.009
11	2004/01	2014/12	48	2018/12	2019/09	Cabify job	–	0	–	–	–
12	2004/01	2015/12	12	2016/12	2018/09	LinkedIn	LinkedIn	1	LinkedIn	**68.7**	.003
13	2005/01	2016/12	12	2017/12	2018/09	LinkedIn	LinkedIn	5	LIDL job	82.2	.135
14	2006/01	2017/12	12	2018/12	2019/09	LinkedIn	–	0	–	–	–

RMSE and other indicators for various models

[a] Best variable is the variable with the highest impact on RMSE reduction

[b] Relative to benchmark SARIMA model specified in the corresponding sample. The best RMSE overall is in bold font

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 PCA, and (ii) a proposed algorithm for FSS. The gains in RMSE relative to the benchmark are around 10% for the PCA-based model and 25% for the FSS-based model. The improvement of the FSS-based model 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 larger amount of GT data used and the application of dimension reduction techniques.

Besides the gain in predictive accuracy found to forecast the Spanish unemployment, the paper also casts some light to the discussion in the literature about using more or less explanatory variables. Our results on the robustness exercise shows that it seems better to introduce only a few 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. Our FSS method demonstrates its potential to find it. As to the discussion about less of more queries, we show that the larger is the number of GT queries, the higher is the probability of finding one or more excellent indicators. At least, when no prior information on which are the most informative queries is available.

Data availability The data that support the findings of this study are available in the Spanish State Employment Service at https://www.sepe.es/HomeSepe/que-es-el-sepe/estadisticas/datos-avance/datos.html and Google Trends searches at https://trends.google.es/trends/. We conducted a search of 200 queries related to job seeking, divided into queries related to: (1) leading job search applications; (2) Spanish unemployment centers (online, physical, public or private); (3) standard job searching terms; and, (4) those companies that generate most employment in Spain. The column *Name* in Table 5 gives some examples of the queries used. Section 3 gives more information about the data.

Appendix

See Tables 4 and 5.

Table 4 Estimates of the β_i coefficients in Eq. (5a)

i	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\sigma}^2_{\hat{a}_t}$
0	–	—	–	–	–	—	–	–	–	–	8.116
1	.205 (.060)	–	–	–	–	–	–	–	–	–	7.373
2	.189 (.055)	.025 (.012)	–	–	–	–	–	–	–	–	7.055
3	.197 (.054)	.024 (.012)	−.019 (.011)	–	–	—	–	–	–	–	6.781
4	.200 (.050)	.034 (.012)	−.022 (.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

β_i for $i = 1, 2, \ldots, I$ are the coefficients corresponding to variables x_{it} for $i = 1, 2, \ldots, I$. Standard errors are in parentheses. One asterisk (*) denotes non-significant values at 10%

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

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

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Askitas N, Zimmermann KF (2009) Google econometrics and unemployment forecasting. Appl Econ Q 55(2):107

Blanchard OJ, Summers LH (1987) Hysteresis in unemployment. Eur Econ Rev 31(1):288–295

Borup D, Schütte ECM (2020) In search of a job: forecasting employment growth using Google Trends. J Bus Econ Stat (**in press**)

Box GEP, Jenkins G (1976) Time series analysis: forecasting and control. Holden-Day, San Francisco

Casals J, Garcia-Hiernaux A, Jerez M, Sotoca S, Trindade A (2016) State-space methods for time series analysis: theory, applications and software. Chapman and Hall/CRC, Boca Raton

Cheng S, Wu T, Lee K, Chang T (2014) Flexible Fourier unit root test of unemployment for PIIGS countries. Econ Model 36:142–148

Choi H, Varian H (2009) Predicting initial claims for unemployment benefits. Google Inc., pp 1–5

Choi H, Varian H (2012) Predicting the present with Google Trends. Econ Rec 88:2–9

D'Amuri F (2009) Predicting unemployment in short samples with internet job search query data. MPRA Paper 18403, University Library of Munich, Germany

D'Amuri F, Marcucci J (2009) "Google it!" Forecasting the US unemployment rate with a Google job search index. MPRA Paper 18248, University Library of Munich, Germany

D'Amuri F, Marcucci J (2017) The predictive power of Google searches in forecasting US unemployment. Int J Forecast 33(4):801–816

Diebold F, Mariano R (1995) Comparing predictive accuracy. J Bus Econ Stat 13:253–263

Escanciano JC, Olmo J (2010) Backtesting parametric value-at-risk with estimation risk. J Bus Econ Stat 28(1):36–51

Eysenbach G (2006) Citation advantage of open access articles. PLoS Biol 4(5):e157

Fondeur Y, Karamé F (2013) Can Google data help predict French youth unemployment? Econ Model 30:117–125

Garcia-Hiernaux A, Casals J, Jerez M (2009) Fast estimation methods for time series models in state space form. J Stat Comput Simul 79(2):121–134

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014

Google (2020a) Find related searches. https://support.google.com/trends/answer/4355000

Google (2020b) Google Trends Data. https://trends.google.es/trends/?geo=ES

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hansen PR, Lunde A, Nason JM (2011) The model confidence set. Econometrica 79(2):453–497

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14(4):382–401

Jiménez R (2012) Linkedin sets up in Spain 9 years later. https://elpais.com/tecnologia/2012/03/27/actualidad/1332838659_013202.html

Johnson HA, Wagner MM, Hogan WR, Chapman WW, Olszewski RT, Dowling JN, Barnas G et al (2004) Analysis of web access logs for surveillance of influenza. In: Medinfo, pp 1202–1206

Jolliffe IT (2002) Principal component analysis. Springer series in statistics. Springer, New York

Li X, Pan B, Law R, Huang X (2017) Forecasting tourism demand with composite search index. Tour Manag 59:57–66

Naccarato A, Falorsi S, Loriga S, Pierini A (2018) Combining official and Google Trends data to forecast the Italian youth unemployment rate. Technol Forecast Soc Chang 130:114–122

Nagao S, Takeda F, Tanaka R (2019) Nowcasting of the US unemployment rate using Google Trends. Finance Res Lett 30:103–109

NetMarketShare (2019) Browser market share. https://netmarketshare.com/?options=

Niesert RF, Oorschot JA, Veldhuisen CP, Brons K, Lange R-J (2020) Can Google search data help predict macroeconomic series? Int J Forecast 36(3):1163–1172

Osorio VM (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 DC, Song H (2012) Forecasting hotel room demand using search engine data. J Hosp Tour Technol 3(3):196–210

Pavlicek J, Kristoufek L (2015) Nowcasting unemployment rates with Google searches: evidence from the Visegrad group countries. PLoS ONE 10(5):e0127084

Poza C, 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. Int Econ 163:163–175

Romero-Avila D, Usabiaga C (2007) Unit root tests and persistence of unemployment: Spain vs. the United States. Appl Econ Lett 14(6):457–461

SEPE (2019) Employment statistics. https://www.sepe.es/HomeSepe/que-es-el-sepe/estadisticas/empleo.html

Siliverstovs B, Wochner DS (2018) Google Trends and reality: do the proportions match? Appraising the informational value of online search behavior: evidence from Swiss tourism regions. J Econ Behav Org 145:1–23

Stock JH, Watson MW (1993) A procedure for predicting recessions with leading indicators: econometric issues and recent experience. In: Business cycles, indicators and forecasting. University of Chicago Press, pp 95–156

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc B 58(1):267–288

Van Der Maaten L, Postma E, Van den Herik J (2009) Dimensionality reduction: a comparative review. J Mach Learn Res 10:66–71

Vicente MR, López-Menéndez AJ, Pérez R (2015) Forecasting unemployment with internet search data: does it help to improve predictions when job destruction is skyrocketing? Technol Forecast Soc Chang 92:132–139

Vozlyublennaia N (2014) Investor attention, index performance, and return predictability. Journal of Banking & Finance 41:17–35

West KD (2006) Forecast evaluation. Handbook of economic forecasting, vol 1. Elsevier, pp 99–134

Woo J, Owen AL (2019) Forecasting private consumption with Google Trends data. J Forecast 38(2):81–91

Yu L, Zhao Y, Tang L, Yang Z (2019) Online big data-driven oil consumption forecasting with Google Trends. Int J Forecast 35(1):213–223