Empirical Economics
https://doi.org/10.1007/s00181-022-02347-w

Forecasting unemployment with Google Trends: age, gender and digital divide

Rodrigo Mulero[1] · **Alfredo Garcia-Hiernaux**[2]

Received: 21 March 2022 / Accepted: 6 December 2022

Abstract
This paper uses time series of job search queries from Google Trends to predict the unemployment in Spain. Within this framework, we study the effect of the so-called digital divide, by age and gender, from the predictions obtained with the Google Trends tool. Regarding males, our results evidence a digital divide effect in favor of the youngest unemployed. Conversely, the forecasts obtained for female and total unemployment clearly reject such effect. More interestingly, Google Trends queries turn out to be much better predictors for female than male unemployment, being this result robust to age groups. Additionally, the number of good predictors identified from the job search queries is also higher for women, suggesting that they are more likely to expand their job search through different queries.

Keywords Digital divide · Forecasting · Gender · Google Trends · Unemployment

JEL Classification C32 · C52 · C53

1 Introduction

The world has undergone a dramatic change with the rise of the Internet in the twenty-first century. Particularly, job seeking has been strongly influenced, being carried out

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

✉ Alfredo Garcia-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 and ICAE, Facultad de Ciencias Económicas, Universidad Complutense de Madrid, Campus de Somosaguas, 28223 Madrid, Spain

Published online: 30 December 2022

increasingly by online resources. In fact, this job searching method has become the most common one, used by employed and unemployed people, as it is expected to improve the chances of finding a job while considerably reducing searching costs. For instance, in 2009 more than 70% of US young unemployed searched for a job online (Khun and Mansour 2014). In Spain, a recent survey among unemployed people revealed that 98% of them use Internet searches and job search websites (Adecco 2016).

Simultaneously, in parallel with the increasing use of the Internet, the digital revolution makes available for researchers a huge amount of data that can be exploited to produce more accurate predictions of countless variables, including unemployment figures. In the last decade, numerous studies employ data mined from the Internet searches to improve the predictions obtained with most common models. Pioneer papers applying this approach arise in the field of medicine see, e.g., Johnson et al. (2004), who analyze the relationship between Internet searches for flu symptoms and the number of cases reported in the USA. In the last decade, this idea has produced a fruitful literature with, in particular, many papers focused on predicting the evolution of the labor market. Some examples are the studies by Choi and Varian (2009, 2012); Pavlicek and Kristoufek (2015); Niesert et al. (2020) and Caperna et al. (2020). Other papers, closer to ours, will be discussed in the next section.

The queries in the search engines, in our case Google Trends (GT), are free and easily obtained. They also offer broader and more up-to-date data than commonly used surveys, which are released with some delay. Yet the data mined from Internet searches are far from being the panacea. In this regard, Cebrián and Domenech (2022) evidence non-negligible issues related to its measurement accuracy. Similarly, Naccarato et al. (2018) point out that the unemployment data cover a known population with an estimated and reliable error distribution. In contrast, data downloaded from GT are not a probabilistic sample of the population and so, its error distribution is unknown. Researchers should keep in mind that its representativeness is closely linked to the consumption patterns and Internet penetration rate. In this sense, only with perfectly global spread of the Internet and same usage patterns by age and gender, data mined from search engines would be thoroughly representative.

Concerning this last point, numerous studies have addressed the issue of the so-called *digital divide*, regarding the differences in the uptake of Internet access and usage patterns by age, race, gender and socioeconomic status (see, e.g., Novak and Hoffman 1998; Enoch and Soker 2006; Abbey and Hyde 2009; Hidalgo et al. 2020). The reader is referred to van Dijk (2020) for a complete and up-to-date survey on this topic. Particularly, Gómez (2019) presents the asymmetries in terms of access and usage to the network in Spain. As expected, usage is higher the higher is the level of education and the better is the economic situation, although these gaps have been considerably reduced in the last decade (see, also, Cañón Rodríguez et al. 2016). To monitor this, the Spanish National Institute of Statistics releases a report on the digital gap by groups of age, education levels and gender (see INE 2020).

In this sense, our paper exploits GT time-series data from 2004 to 2018, on a collection of more than 170 search-related items, to predict unemployment figures. The official Spanish unemployment series, disaggregated by age groups and gender, are applied to yield one-step-ahead out-of-sample forecasts. This disaggregation allows

us to study, not only the effect of the *digital divide* by groups of age when forecasting the unemployment with Internet searches, but also by gender, and its interaction with the previous age groups. In spite of the flourishing literature on the predictive power of GT, this is unprecedented to the best of our knowledge.

The paper is organized as follows. Section 2 provides a revision of the literature in the use of GT as predictors, focusing on unemployment applications and digital divide. Section 3 details the data employed in the analysis, paying particular attention to the GT queries and how those are generated. The benchmark model and the proposed alternatives are presented in Sect. 4. Section 5 compares the forecasting results of the proposed models relative to the benchmarks for all the combinations of gender and age group. A discussion and some concluding remarks close the paper.

2 Literature review

The available (quasi) real-time GT data allow nowcasting models, which provide more accurate estimates than those generated by conventional indicators. This has been shown by many authors for unemployment figures, and other variables related to the labor market, in countries such as the USA, Germany, Romania, the UK, Spain, France, Italy or Canada.

Focusing on the USA, Choi and Varian (2009, 2012) and Nagao et al. (2019) demonstrate that introducing an indicator of the number of searches in Internet improves the results of conventional models when predicting the unemployment. D'Amuri and Marcucci (2009) provide similar results, in this case by applying an index generated by searches in GT. Later, the same authors revisit the theory of their previous work, disaggregate the GT searches at a federal level and incorporate the effects of the 2008 Great Recession in D'Amuri and Marcucci (2017). More recently, Borup and Schütte (2022) analyze the impact in the forecast when using a large amount of GT-query variables. They conclude that GT variables do not seem to be better predictors for the unemployment than the classical macroeconomic and financial series. However, combining many GT series, preferably with nonlinear procedures, increases the forecasting power, significantly overtaking the above-mentioned classical indicators.

For Germany, Askitas and Zimmermann (2009) compute the improvement of the unemployment forecasts using three groups of keywords (*unemployment agency*, *unemployment rate*, *staff consultant*) and a set of queries linked by the Boolean operator 'OR,' captured from the job vacancies websites *Monster* and *Jobboerse*.

The effect of GT searches at a regional level has been studied by Simionescu (2020). This paper's main contribution is to analyze the results of this methodology when forecasting the unemployment rate in Romanian counties, which are heterogeneous in terms of economic and social development. In a subsequent paper, Simionescu et al. (2020) estimate the impact of the Brexit on the unemployment for the UK also with GT predictors.

Regarding Spain, several studies have arisen recently. Vicente et al. (2015) and González-Fernández and González-Velasco (2018) obtain better predictive accuracy when forecasting the Spanish unemployment with two and one GT searching terms, respectively, than with univariate and multivariate models that do not include this

information. Similarly, Mulero and Garcia-Hiernaux (2021) find a remarkable precision gain when forecasting monthly unemployment with numerous GT-query searches and dimensionality reduction techniques. Finally, Simionescu and Cifuentes-Faura (2022) also demonstrate the capacity of GT data to predict Spanish and Portuguese unemployment (at a regional level), in this case by means of dynamic panel data models.

As mentioned, D'Amuri (2009) applies the same idea to beat the official unemployment forecast in the USA. Interestingly, this author is the first to focus on a potential selection bias, finding a *digital divide* effect in favor of younger people, arguing that they are the greatest consumers of Internet. In contrast with the considerable amount of papers dealing with the unemployment prediction gain when using GT predictors, only a few focus on this selection bias related to *digital divide*. Some of them are Fondeur and Karamé (2013) and Naccarato et al. (2018), who analyze the unemployment in France and Italy, respectively. The former studies the predictive capability of GT variables when forecasting the French unemployment in three age ranges 15–24, 25–49 and over 50, finding statistically significant results only for the youngest. The latter, supported by the selection bias argument, only focuses on the predictive capacity in youth unemployment. Additionally, Dilmaghani (2018, 2019) also investigates the effects of the *digital divide* on this methodology. Her first work analyzes the forecast improvements when introducing GT searches to predict the unemployment rate in the USA only for youths between 16 and 24 of age, distinguishing among Whites, Hispanics or African-Americans, and males or females. Her second paper shows an improvement in the prediction of the Canadian unemployment rate for the age grouped between 25 and 44 years old.

3 Data

This section introduces the data employed in the analysis. We first describe the disaggregated (in gender and age) unemployment series. Second, we detail the GT queries that will be used as predictors devised to improve the forecast of the unemployment figures.

3.1 Unemployment data

This research analyzes the unadjusted and disaggregated unemployment series supplied by the Spanish Public Employment Service (SEPE). Each observation is released the first week of the next month and reports the number of people (by age group and gender) declaring to look for a job at a public employment office. Figure 1 shows the availability of the unemployment data and why (quasi) real-time predictors as GT variables may improve its forecasts.

Our sample covers the period from January 2004 to September 2018, for a total of 177 monthly observations, including business cycle expansions and recessions. The data are disaggregated by age and gender groups, as presented in Fig. 2. This disaggregation is supported by the following arguments.

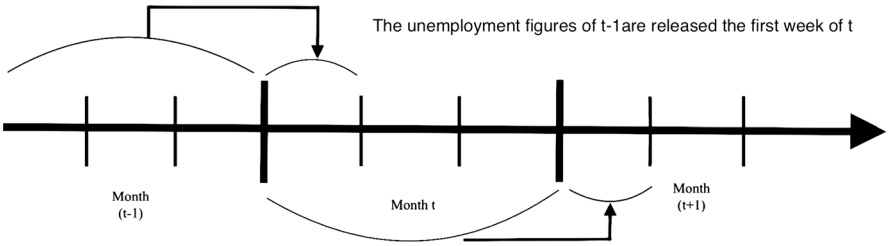

Fig. 1 Spanish monthly unemployment data availability

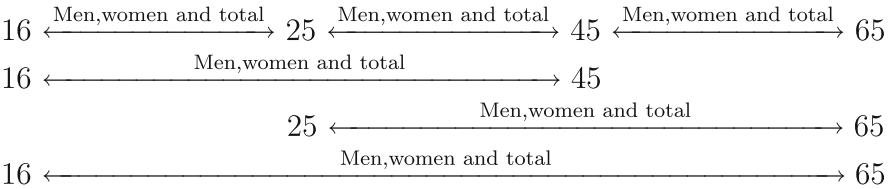

Fig. 2 Description of the unemployment series used as endogenous variables. The numbers correspond to years of age

Less than 25 years of age. This is the age range most commonly used by the literature. We will use these series (female, male and total unemployment) to contrast the potential selection bias previously reported by D'Amuri (2009) and Fondeur and Karamé (2013).

Between 25 and 45 years of age. These series contain most of the working age population in Spain. They will be used to compare the results against the youths.

Older than 45 years of age. This group is characterized by a lower Internet use, and therefore, we expect the lowest—if any—gain in the unemployment forecast in both genders, when adding GT queries.

Finally, to better investigate the effect of different groups of age and make them comparable with most of the literature, we build two additional groups: older than 25 and younger than 45 years of age. We then end up with six age (three non-overlapped and three overlapped) groups and three samples (females, males and totals unemployed) for each of them. This amounts to eighteen endogenous variables. The disaggregation will permit us to study the overall gender effect and make gender and age comparisons of the potential *digital divide* effects. Figure 3 depicts all the series introduced above.

3.2 Google Trends

We use Google data because the queries introduced in this browser are a reliable estimation of all the searches made on the Internet. We download the data from a tool named Google Trends (GT). 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. The index can be collected from January 1, 2004, up to 36 hours prior to the search.

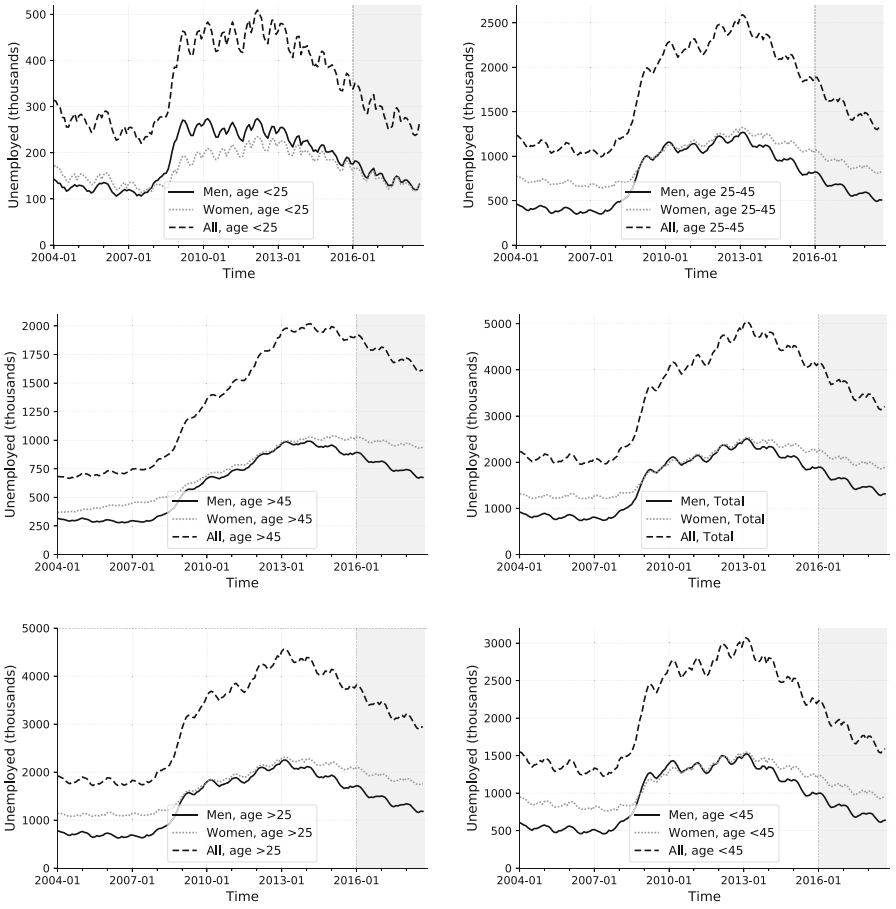

Fig. 3 Unemployment series by gender and age groups. The shaded area corresponds to the validation period

Google makes some data cleaning in its trends. For instance, searches performed repeatedly from the same machine in a short period of time are removed and just counted once. For more details about GT and on how its index is created, see Mulero and Garcia-Hiernaux (2021). Because of the huge amount of information gathered, as pointed out by Blazquez and Domenech (2018), GT has become a useful tool for studies related to large-scale data.

However, the information provided by GT has evident limitations when applied as a potential predictor. Dilmaghani (2019) lists four issues. First, the data sources of the search engines are not probabilistic samples of the population (see Naccarato et al. 2018). As these data reflect the part of the population that used the Internet, it can potentially suffer from selection bias (e.g., job search queries are possibly driven by young people from larger urban centers). Second, GT does not distinguish between those looking for job opportunities when they are unemployed or just contemplating a job switch. This difference can be crucial, as job queries for the unemployed often show

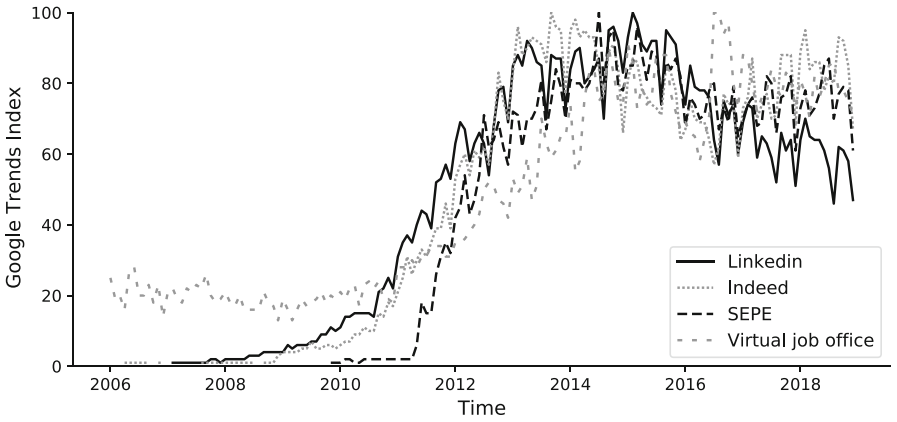

Fig. 4 GT index for the queries *LinkedIn*, *Indeed*, *SEPE* and *Virtual job office*. Selection from the 163 queries used

a counter-cyclical behavior, but searches addressed by employees are usually assumed to be procyclical. Third, GT does not provide information on users' sociodemographic characteristics. Fourth, GT index is calculated with a sampling method. Therefore, GT series may change if some new observations are added, which could yield some bias in the estimates (see, Vicente et al. 2015; Cebrián and Domenech 2022).

On top of the above drawbacks, inaccuracies can appear from an unsuitable keywords selection or data processing. A summary of the procedure applied in the paper, which follows these lines, could be useful to researchers working with GT. We conduct a search of more than 170 job query terms between January 2004 and September 2018. We group the search terms into four sets, based on what they are representative of. Specifically, Group 1 includes series representing queries related to leading job search applications, e.g., *Infojobs, Indeed, Monster*; Group 2 is made up of searches related to Spanish unemployment centers, either online, physical, public or private, e.g., *Employment office, SEPE, Randstad*; Group 3 contains queries related to standard job searching terms, e.g., *Job offers, How to Find a Job, Job vacancy*; and finally, Group 4 consists of searches related to the companies that generate most employment in Spain, e.g., *work in Inditex, Orange work, Santander job*. Additionally, we incorporate the information provided by the 'related searches' GT tool, which allows us to capture other queries related to the terms above.[1] As illustration, Fig. 4 shows the GT index for four selected queries. Notice that the time evolution of the indexes is similar to those of the unemployment series, but the correlations likely vary across them.

[1] Specific information about all the queries used and GT downloaded data is available from the authors upon request. A report on the 'related searches' can be found in: https://support.google.com/trends/answer/4355000.

4 Models

This section presents the models applied in the paper. First, we introduce the univariate models that will generate the benchmark predictions for each unemployment series defined by age group and gender. Second, we describe the alternative models that incorporate the information of the GT queries, which will potentially improve the benchmark forecasts.

4.1 Univariate models

The ARIMA representation (Box and Jenkins 1976) is chosen to obtain our benchmark models. The general univariate monthly time series representation considered here is:

$$\Phi_P(B^{12})\phi_p(B)\nabla^d\nabla_{12}^D u_t = \mu + \Theta_Q(B^{12})\theta_q(B)a_t, \tag{1}$$

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 $\Phi_P(B^{12}) = 1-\Phi_1 B^{12}-\cdots-\Phi_P B^{12P}$ and $\Theta_Q(B^{12}) = 1-\Theta_1 B^{12}-\cdots-\Theta_Q B^{12Q}$ are polynomials in B^{12} of degrees P and Q, respectively, and 12 is the seasonal frequency. In addition, B is the lag operator so that $Bu_t = u_{t-1}$, $\nabla = (1-B)$ is the difference operator, μ is a constant and a_t is a sequence of uncorrelated Gaussian variates with zero-mean and constant variance, σ_a^2. As it is common in time series, we assume that all the zeros of the polynomials in B and B^{12} are outside the unit circle (stationarity and invertibility requirements) and have no common factors. Model (1) is sometimes known as the Seasonal AutoRegressive Integrated Moving Average (SARIMA) form of the stochastic process u_t. Actually, u_t should be written u_{it} for each unemployment stochastic process, where i denotes the corresponding gender-age group; see Fig. 2. However, we avoid sub-index i in all the elements of Eq. (1) for the sake of simplicity.

In order to identify an appropriate univariate model for each endogenous variable, we apply the methodology proposed by Garcia-Hiernaux et al. (2022). Essentially, the procedure first detects the number of unit roots at the zero and seasonal frequencies and suggests the corresponding transformation that induces stationarity; in our case $\nabla\nabla_{12}$. Next, the autoregressive and moving average orders are selected, first for the regular and then for the seasonal part, by estimating a sequence of models. After pruning some non-significant parameters, this leads us to the model presented in Eq. (2). In all the cases, we add a step dummy variable to capture the effect of the 2008 global financial crisis, which hardly hit the Spanish unemployment level. The series corrected from this outlier are denoted by u_{it}^*. The residuals of the final models are tested with the algorithm NID (see, Garcia-Hiernaux et al. 2012), which shows no evidence of autocorrelation.

This identification process returns two very similar models: first, a SARIMA$(1, 1, 1) \times (0, 1, 1)_{12}$ for men unemployed between 25 and 45, over 25 and total; total unemployment between 25 and 45; and women unemployed under 45, and second a SARIMA$(2, 1, 1) \times (0, 1, 1)_{12}$ model for the remaining series. As the series depicted in Fig. 3 are not very different from each other, similar SARIMA structures were

expected. However, the estimated values of the parameters substantially vary across models. The final base model is represented by Eq. (2), whose residuals do not evidence any sign of misspecification and are compatible with the statistical assumptions made on a_t[2]:

$$(1 - \phi_{i1}B - \phi_{i2}B^2)\nabla\nabla_{12}u_{it}^* = (1 - \Theta_{i1}B^{12})a_{it}, \tag{2}$$

where u_{it}^* is the corresponding unemployment series. As mentioned above, when $i = \{$men between 25 and 45, men over 25, all men, women under 45, total between 25 and 45$\}$ then $\phi_{i2} = 0$. We will use these models as benchmarks in the forecasting exercises of Sect. 5.[3]

4.2 Models including GT searches

As alternative models, we apply the simple idea of including additional explanatory variables for u_{it}^* and keep the ARMA noise structure for the residuals, as long as the statistical diagnosis does not reveal any sign of misspecification. Hence, these models are represented by the following transfer function:

$$u_{it}^* = \sum_{j=1}^{J} \beta_{ij}x_{ijt} + \eta_{it}; \tag{3a}$$

$$(1 - \phi_{i1}B - \phi_{i2}B^2)\nabla\nabla_{12}\eta_{it} = (1 - \Theta_{i1}B^{12})a_{it}, \tag{3b}$$

where the indicators x_{ijt}, $j = 1, 2, 3, \ldots, J$ for each u_{it}^* will be selected from all the series mined from GT.

Now, we briefly explain the selection feature methodology to choose the indicators x_{ijt} used in Eq. (3a), which was proposed by Mulero and Garcia-Hiernaux (2021). The process consists in a relatively simple AIC-based forward stepwise feature selection. Let us start with a set of 174 queries. In the first step, we estimate Model (3a–3b) for the train sample with just one potential explanatory variable without lags in (3a), keeping the structure in the noise Eq. (3b). We repeat this step for each GT variable in our initial set, which implies estimating a model for each indicator. Once the estimation loop is finished, we sort the models by the lowest information criterion used. The authors recommend AIC for this purpose.[4] This permits us to get the best in-sample model out of all the estimates, according to AIC. Second, we compute the one-step-ahead out-of-sample forecasts in the evaluation sample (here 2016/01–2018/09) based on the

[2] Shin–Fuller's unit root test rejects the null hypothesis of non-stationarity for the transformed series in differences, and so the $\nabla\nabla_{12}$ transformation is confirmed. Additionally, the null hypothesis for normality and homoskedasticity are not rejected on the residuals.

[3] The same models were identified if we use $\log(u_{it}^*)$ instead of u_{it}^* as the endogenous variable. The results of the paper do not change significantly when the log transformation is applied to all the series.

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

estimates of the chosen model. The root-mean-square error (RMSE) is then calculated from the previous forecasts.[5] We repeat this process, by adding a new indicator to the previous model, as long as the RMSE is lower than the one obtained with the benchmark. For this, we rerun the model selection loop and choose the next predictor whose model minimizes AIC. The process stops when the inclusion of an additional indicator, whose estimated model yields the lowest information criterion, does not improve the RMSE benchmark model. The RMSE is then only used to make the algorithm stop, i.e., to determine J in Eq. (3a).

We run the procedure detailed above to find alternative models for all the combinations of genders and groups of age.[6] We choose this method among others existing in the literature because of two main reasons: (1) It is computationally simple and fast enough to be applied to a large amount of models and indicators: We work with 18 endogenous variables and more than 170 potential predictors, and (2) it has proved to be able to find good predictors and remarkable forecasting gains. A more detailed discussion of this method against close alternatives can be found in Mulero and Garcia-Hiernaux (2021).

Hence, the alternative models, whose forecasts will be compared against the benchmarks, are represented by Eqs. (3a–3b), where x_{ijt} with $j = 1, 2, \ldots, J$ denotes the predictors chosen by the feature selection method, for each endogenous variable u^*_{it} (with $i = 1, 2, \ldots, 18$) presented in Fig. 2.

5 Main findings

This section analyzes the results of applying the previous models to forecast the Spanish unemployment by gender and age groups in an out-of-sample validation of 33 periods. However, the purpose of this section is not to merely predict the unemployment using Interned mined data. Instead, the main intention is to study whether the inclusion of GT predictors reveals information about a potential age and/or gender *digital divide* when forecasting the unemployment. Therefore, the forecasting performance is compared for Eqs. (2) and (3a–3b), which include GT data, across all the combinations of age groups and genders. For this comparison, the RMSE and the relative RMSE against the corresponding benchmark model are computed. All the forecasting models converge adequately and show no evidence of poor specification.

As the out-of-sample size is relatively small, there could be non-negligible uncertainty in the RMSEs. To incorporate this uncertainty in our evaluation of the predictive capacity, the forecast comparisons include the Diebold and Mariano (1995) test and its p value.[7] The null hypothesis of this test is that the two predictions (coming from the benchmark and the alternative model) are equally accurate. Hence, a small p value evi-

[5] 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 $\left(\frac{1}{L}\sum_{l=1}^{L}\hat{a}^2_{l+1|l}\right)^{1/2}$.

[6] The Python code for the forward stepwise feature selection algorithm as well as the forecasting analysis presented in Sect. 5 is available from the authors upon request.

[7] Although Diebold and Mariano test does not account for parameter estimation error, it is appropriate in this application as in all our cases the out-of-sample size is small relative to the in-sample size. In these

Table 1 RMSEs for best alternative models and its corresponding benchmark by age (all)

Age group	J	RMSE			Diebold–Mariano test	
		Model (J)	Model (0)	Relative (%)	Model (J) versus model (0) Statistic	*p* value
<25	3	0.5328	0.5908	90.19	1.130	0.113
25–45	1	1.1787	1.4418	81.75	1.714	0.048
>45	1	0.5670	0.6791	83.50	2.002	0.027
>25	1	1.6052	1.9824	80.97	1.987	0.028
<45	3	1.4284	1.8571	76.91	2.150	0.020
All	4	1.8276	2.4316	75.16	2.333	0.013

dences that the alternative model predicts better than the benchmark with a particular significance level.

We start by focusing on the comparison of different age groups with no gender disaggregation, denoted by *All* in tables and figures. First, Fig. 5 (top) shows the relative RMSEs against the benchmark's for model (3a–3b), by age groups. Notice that the best predictions in terms of lowest RMSEs are usually found with a few GT variables. This is also revealed by the (low) values of J in Table 1 and seems to be consistent across genders and age groups. Second, both Fig. 5 (top) and Table 1 show no evidence of the youth bias found by Naccarato et al. (2018) and Fondeur and Karamé (2013) for Italy and France, respectively, as all the groups of age present a statistically significant benefit (around 10% level or less) of using GT searches as predictors. Third, contrary to the literature, the lowest forecasting gain when including Internet searches is found for the youngest age group, clearly rejecting an age *digital divide* effect in the total (male plus female) unemployed population.

Nonetheless, when those groups of age are disaggregated by gender, the results exhibit a completely different picture. First, the *digital divide* effect in favor of the youngest group is now clearly perceptible in men, as groups with age under 25 and under 45 are those with a higher forecasting improvement; see Fig. 5 (middle). This is confirmed in Table 2, where the gain in terms of RMSE is statistically significant at 10% only for these two groups. In fact, as the improvement for the group 25–45 is not significant, one could conclude that the benefit of using GT predictors for males' unemployment only occurs for those under 25. Conversely, Fig. 5 (bottom) shows that GT searches systematically improve the forecasts obtained from the benchmark model for all the women age groups. When looking at Table 3, we observe that these improvements are indeed statistically significant at 5%. In fact, the benefit of including GT queries as predictors for the unemployment is much greater in females than males. Therefore, one can conclude that most of the gain found for the total unemployment series comes from the women side. This result has no precedence in the literature.

Footnote 7 continued
situations, according to West (2006), assuming there is no estimation error may be considered as a good approximation in the forecasting evaluation exercise.

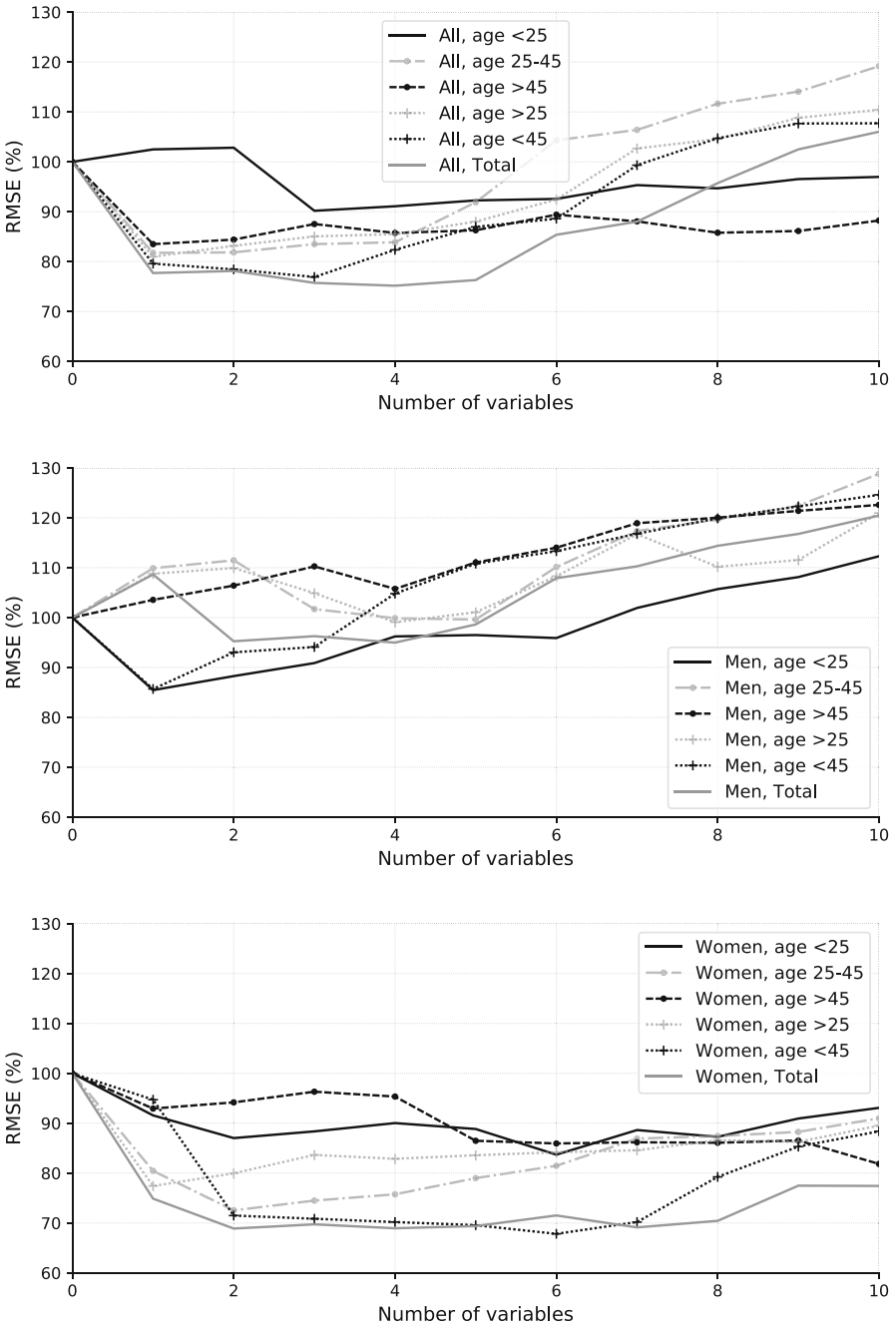

Fig. 5 Forecasting accuracy of the alternative models. Relative RMSEs comparison by age and gender

Table 2 RMSEs for best alternative models and its corresponding benchmark by age (men)

Age group	J	RMSE			Diebold–Mariano test	
		Model (J)	Model (0)	Relative (%)	Model (J) versus model (0) Statistic	*p*value
<25	1	0.2986	0.3494	85.45	1.750	0.045
25–45	5	0.8387	0.8420	99.60	0.059	0.477
>45	0	0.4475	0.4475	100.00	–	–
>25	4	1.2970	1.3091	99.08	0.324	0.374
<45	1	0.9958	1.1622	85.68	1.513	0.070
All	0	1.4955	1.4955	100.00	–	–

Table 3 RMSEs for best alternative models and its corresponding benchmark by age (women)

Age group	J	RMSE			Diebold–Mariano test	
		Model (J)	Model (0)	Relative (%)	Model (J) versus model (0) Statistic	*p* value
<25	6	0.2217	0.2648	83.70	1.918	0.032
25–45	2	0.4968	0.6850	72.53	2.315	0.014
>45	10	0.2717	0.3320	81.84	1.723	0.047
>25	1	0.7532	0.9729	77.43	1.898	0.033
<45	6	0.5927	0.8739	67.82	2.501	0.009
All	2	0.7905	1.1476	68.89	2.651	0.006

Table 4 Relative RMSE with respect to its corresponding benchmark for women, men and differentials by age groups

Age group	Women (%)	Men (%)	Women–men (%)
<25	83.70	85.45	−1.76
25–45	72.53	99.60	−27.07
>45	81.84	100.0	−18.16
>25	77.43	99.08	−21.65
<45	67.82	85.68	−17.86
All	68.89	100.0	−31.11

Digging into this finding, Table 4 and Fig. 6 emphasize the differences in the forecasting improvement, in terms of RMSE with respect to each benchmark, between women and men by age group. The gain is clearly greater for females in all groups of age. This differential seems not to be very relevant for the youngest group (1.8 percentage points), but it is remarkable for the rest of the groups (ranging from 18 to 31 p.p. of higher gain in favor of women, according to the age group; see Table 4). Besides, this result is independent of the number of predictors used to build the forecasts. Figure 6 shows that the women relative RMSE is consistently under the men's one, no matter the age group or the number of predictors.

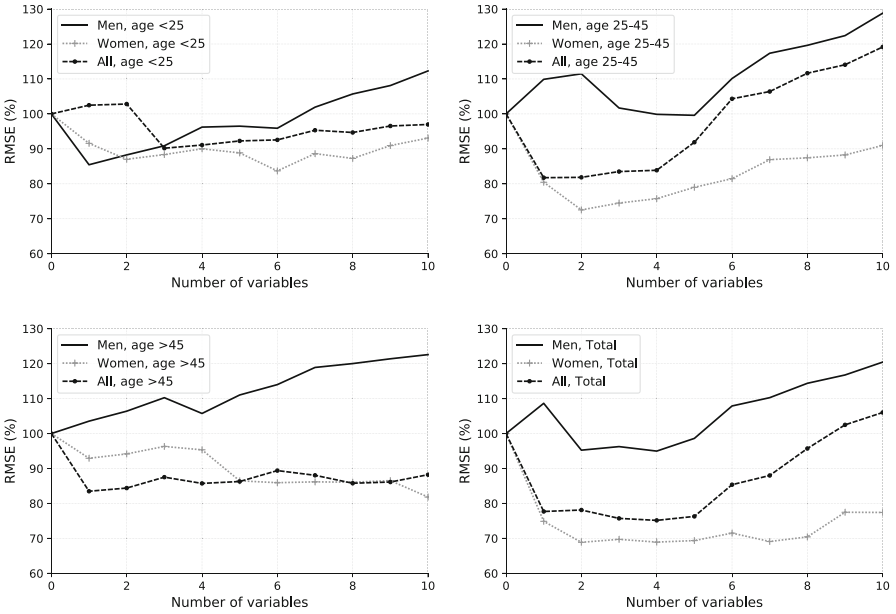

Fig. 6 Forecasting accuracy of the alternative models. Relative RMSEs comparison of men, women and total unemployment by age groups

Beyond these main findings, it is interesting to analyze the queries that provide these results. Thus, we now discuss which are the best GT queries in terms of predictive power by gender and group of age. We will focus on two particular GT queries for each model: (1) the first predictor chosen by the feature selection algorithm, described in Sect. 4.2, and (2) the predictor that yields the highest forecasting gain in terms of RMSE reduction. Table 5 offers this information for men, women and total unemployment by age groups. When looking at male unemployment, there is not much variability in the best GT predictors found: either *LinkedIn* or *Orange vacancies* is the first feature selected by the algorithm or the best one, in terms of RMSE reduction for all age groups. When the prediction improvement is statistically significant (men younger than 25 or younger than 45), the only relevant GT query is *LinkedIn*, evidencing the importance of this social media networking site when it comes to searching a job, particularly for men. This is also noticeable in Fig. 5 (middle), where only one step-down is perceptible for men in most of the relative RMSEs.

Interestingly, this picture varies considerably when looking at female unemployment, where the variability of the best predictors is higher. Table 5 shows a diversity of terms (e.g., *curriculum vitae*, *job vacancy*, *virtual employment office*, *job offers*, etc.), suggesting that women are more likely to expand their job search to different queries, related to websites, firms and public institutions. Accordingly, Fig. 5 (bottom) shows several step-downs in women's relative RMSEs. Contrary, none of the previous query terms seem to be informative enough to predict male unemployment. *LinkedIn* remains a very good predictor also for women unemployment in some age

Table 5 Best GT predictors by gender and group of age

Unemployment series (age)	Men			Women			Total		
	First Predictor	Best predictor[a]	RMSE gain%	First predictor	Best predictor[a]	RMSE gain %	First predictor	Best predictor[a]	RMSE gain%
<25	LinkedIn	LinkedIn (1)	14.55	Curriculum vitae	Media markt jobs (6)	16.33	Job vacancies	LinkedIn (3)	9.82
25–45	Orange jobs	Orange vacancies (5)	0.41	LinkedIn	Job vacancies (2)	27.47	LinkedIn	LinkedIn (1)	18.25
>45	–	–	–	Virtual job office	CV (10)	18.15	LinkedIn	LinkedIn (1)	16.50
>25	Orange jobs	LinkedIn (4)	0.92	LinkedIn	LinkedIn (1)	22.58	LinkedIn	LinkedIn (1)	19.03
<45	LinkedIn	LinkedIn (1)	14.32	Job vacancies	Ikea jobs (6)	31.96	LinkedIn	Ikea vacancies (3)	23.09
Total	Orange jobs	LinkedIn (2)	3.46	LinkedIn	Job vacancies (2)	31.11	LinkedIn	HTFJ[b] (4)	24.84

Best predictor[a] is the predictor with the highest reduction in RMSE. The position in which it was found is in parentheses. HTFJ[b] stands for *How to Find a Job*

groups, which explains it remains the best predictor for most of the age groups when considering the total unemployment sample.

Last, to establish the generality of these findings, future research should analyze whether the results reported in this paper are robust to the phase of the business cycle in which the forecasts are computed. This does not seem obvious, as the unemployment series for women and men show more dissimilarities during the recoveries (and less in the recessions), except for the youths (see Fig. 3). This could partially explain the lower gender impact of the GT queries found for the unemployed younger than 25. Unfortunately, due to sample limitations, our forecast exercise only covers an unemployment recovery phase (shaded area in Fig. 3).

6 Discussion and concluding remarks

This paper studies whether data mined from Internet, collected in the form of time series from GT queries, reveal some information about age and/or gender effects of the *digital divide*, when forecasting the Spanish unemployment. To analyze this fact, we disaggregate the unemployment series by age groups and gender and use more than 170 GT series as potential predictors.

Some papers emphasize the idea that the Internet access is not yet universal even in advanced economies, and so this *digital divide*, either by age, gender or race, yields a selection bias in the use of data mined from GT, compared to other indicators. D'Amuri (2009), Fondeur and Karamé (2013) and Naccarato et al. (2018) find a *digital divide* effect in favor of the youngest unemployed (over the rest of ages), while Dilmaghani (2018) finds it in favor of Whites (over Hispanics and African-Americans) and (white) males over (white) females. The latter study only focuses on unemployed younger than 25 years of age.

Surprisingly, our research only finds an age *digital divide* effect in males unemployment. Its gain of 14.5%, in terms of RMSE, for the youngest unemployed is similar to the range 9-16% found by Fondeur and Karamé (2013), 15% by D'Amuri (2009), 15% by Vicente et al. (2015) and 10–19% by France and Shi (2018).

On the contrary, results on female and total unemployment suggest no evidence of any age *digital divide* effect. Further, GT queries turn out to perform much better as predictors for women unemployment than for men's, for every group of age. To our knowledge, these results are unprecedented in the literature. When we examine these findings in contrast with the data supplied by the Spanish National Institute of Statistics (INE 2019), we observe that the digital gap by age was only significant in 2015 for people older than 45 years of age (a gap of 29% points of usual Internet use, with respect to the youngest people). In 2018, this gap had fallen to 13 p.p. When looking at the digital gap by gender—in the same survey—this has been closed from 2015 to 2018 for all working ages and even become slightly negative (in favor of women) for most of them. This could partially explain our results. However, we believe that the relation between GT variables and unemployment series is too much complex to be explained by just a *digital divide* effect, either by age or gender. Many other factors come into play here, specially when the measure of *digital divide*, as the difference of terms of usual Internet use, is small. For instance, the fact that women look up on search

engines more than men (Gargallo-Castel et al. 2010), that the higher is the education level, the higher is the Internet use (Gómez 2019), and the country-specific structure of the labor market are likely additional variables of importance in order to explain the results obtained in this paper. In fact, van Dijk (2020) proposes that three main factors contribute to the *digital divide*: personal categories (e.g., age, gender, ethnicity, etc.), positional categories (e.g., labor, education, household, etc.) and resources. In this sense, only conclusions about the first factor can be drawn from this exercise.

In summary, the contribution of this research to the literature is twofold: (1) The paper is the first to provide evidence that Internet search predictors improve the predictability of unemployment much more in females than males. This result is robust to different groups of age. Moreover, this comes together with the fact that females unemployed seem to use a more diverse good predictors than men's; (2) contrary to the literature, the gain in predictive power obtained by GT searches for the total unemployment series does not evidence an age *digital divide* effect. However, when disaggregating by gender, this effect is observed for males, but clearly rejected for women.

Finally, future research should analyze how the results reported in this paper depend on the phase of the business cycle undergone during the validation sample.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Alfredo Garcia-Hiernaux gratefully acknowledges financial support from UCM-Santander Grant Ref. PR75/18-21570.

Data availability Publicly available data.

Code Availability Replication code available from authors upon request.

Declaration

Conflict of interest There are no financial and non-financial competing interests to declare.

Ethical approval Not applicable.

References

Abbey R, Hyde S (2009) No country for older people? Age and the digital divide. J Inf Commun Ethics Soc 7(4):225–242

Adecco (2016) Infojob-Adecco report on social media and labor market. https://www.adeccogroup.es/wp-content/uploads/2017/11/Informe-2017-Empleo-y-Redes.-Infoempleo-Adecco.pdf (**in Spanish**)

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
Blazquez D, Domenech J (2018) Big data sources and methods for social and economic analyses. Technol Forecast Soc Chang 130:99–113
Borup D, Schütte ECM (2022) In search of a job: Forecasting employment growth using Google Trends. J Bus Econ Stat 40(1):186–200
Box GEP, Jenkins G (1976) Time series analysis: forecasting and control. Holden-Day, San Francisco
Cañón Rodríguez R, Grande de Prado M, Cantón Mayo I (2016) Digital divide: impact on social and personal development. Associated factors. Tendencias pedagógicas 28:115–132
Caperna G, Colagrossi M, Geraci A, Mazzarella G (2020) Googling unemployment during the pandemic: Inference and nowcast using search data. Publications Office of the European Union
Cebrián E, Domenech J (2022) Is Google Trends a quality data source? Appl Econ Lett. (in press)
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
Dilmaghani M (2018) The racial 'digital divide' in the predictive power of Google Trends data for forecasting the unemployment rate. J Econ Soc Meas 43(3–4):119–142
Dilmaghani M (2019) Workopolis or the Pirate Bay: what does Google Trends say about the unemployment rate? J Econ Stud 46(2):422–445
Enoch Y, Soker Z (2006) Age, gender, ethnicity and the digital divide: university students' use of web-based instruction. Open Learn J Open Distance e-Learn 21(2):99–110
Fondeur Y, Karamé F (2013) Can Google data help predict French youth unemployment? Econ Model 30:117–125
France SL, Shi Y (2018) Aggregating Google Trends: Multivariate testing and analysis. arXiv:1712.03152v2
Garcia-Hiernaux A, Casals J, Jerez M (2012) Estimating the system order by subspace methods. Comput Stat 27:411–425
Garcia-Hiernaux A, Casals J, Jerez M (2022) Identification of canonical models for vectors of time series: a subspace approach. https://ssrn.com/abstract=2572931
Gargallo-Castel A, Esteban-Salvador L, Perez-Sanz J (2010) Impact of gender in adopting and using ICTs in Spain. J Technol Manag Innov 5(3):120–128
Gómez DC (2019) An approach to the evolution of the digital divide among the young population in Spain (2006–2015). Revista Española de Sociología 28:27–44
González-Fernández M, González-Velasco C (2018) Can Google econometrics predict unemployment? Evidence from Spain. Econ Lett 170:42–45
Hidalgo A, Gabaly S, Morales-Alonso G, Urueña A (2020) The digital divide in light of sustainable development: an approach through advanced machine learning techniques. Technol Forecast Social Change 150:119754
INE (2019) Free downloadable publications. Survey on equipment and use of information and communication technologies at home. https://www.ine.es/
INE (2020) Free downloadable publications. Women and men in Spain. https://www.ine.es/
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
Khun P, Mansour H (2014) Is internet job search still ineffective. Econ J 124(581):1213–1233
Mulero R, Garcia-Hiernaux A (2021) Forecasting Spanish unemployment with Google Trends and dimension reduction techniques. Ser J Span Econ Assoc 12:329–349
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. Financ Res Lett 30:103–109
Niesert RF, Oorschot JA, Veldhuisen CP, Brons K, Lange R (2020) Can Google search data help predict macroeconomic series? Int J Forecast 36(3):1163–1172

Novak TP, Hoffman DL (1998) Bridging the racial divide on the internet. Science 280(5362):390–392
Pavlicek J, Kristoufek L (2015) Nowcasting unemployment rates with Google searches: evidence from the Visegrad group countries. PLoS ONE 10:5
Simionescu M (2020) Improving unemployment rate forecasts at regional level in Romania using Google Trends. Technol Forecast Soc Change 155:120026
Simionescu M, Cifuentes-Faura J (2022) Can unemployment forecasts based on Google Trends help government design better policies? An investigation based on Spain and Portugal. J Policy Model 44(1):1–21
Simionescu M, Streimikiene D, Strielkowski W (2020) What does Google Trends tell us about the impact of Brexit on the unemployment rate in the UK? Sustainability 12(3):1011
van Dijk JAGM (2020) The digital divide. Polity, Cambridge
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
West KD (2006) Forecast evaluation. In: Elliott G, Granger C, Timmermann A (eds) Handbook of economic forecasting, vol 1. Elsevier, New York, pp 99–134