Economics Letters 170 (2018) 42–45

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Can Google econometrics predict unemployment? Evidence from Spain

Marcos González-Fernández *, Carmen González-Velasco

Department of Business Economics and Management, Faculty of Economics and Business, University of León, Campus de Vegazana, 24071 León, Spain

HIGHLIGHTS

- We analyze whether Google data can predict Spanish unemployment.
- We perform an in and out-of-sample forecast evaluation.
- The results indicate a high correlation between Google searches and unemployment.
- The results indicate that Google data search volumes improve unemployment prediction.

ARTICLE INFO

Article history:
Received 27 November 2017
Received in revised form 2 May 2018
Accepted 27 May 2018
Available online 1 June 2018

JEL classification:
E24
E27
J64
G17
C53

Keywords:
Google econometrics
Unemployment
Spain
Forecasting

ABSTRACT

The aim of the paper is to analyze the ability of internet activity, what has been called Google econometrics, to predict unemployment in Spain. We include a new predictor for Spanish unemployment based on internet information provided by Google Trends. Using monthly data from January 2004 to November 2017 we found evidence of a high correlation between internet queries and unemployment. Besides that, the inclusion of internet activity enhances model's prediction performance.

1. Introduction and theoretical background

The internet has become the most important source of information. Most economic or financial decisions are preceded by a search of information on the internet. Researchers have noticed this fact, and the number of papers regarding this topic has increased sharply during the last decade in all fields of academic literature. In this paper, we address this issue by analyzing the ability of internet activity to predict unemployment in the Spanish labor market.

For this purpose, we use what has been called Google econometrics (Fondeur and Karamé, 2013) to measure internet activity regarding the labor market in Spain. Specifically, Google econometrics refers to the data obtained from the Google Trends tool. This instrument provides indexed data from the number of queries that result from specific keywords over time.

The use of Google econometrics to predict economic variables is not a new concept in the literature, but it is relatively recent. Many papers regarding Google econometrics exist in other disciplines, such as medicine (Ginsberg et al., 2009). However, the economic literature has only in the last few years started to address this topic (Dergiades et al., 2015; Siganos, 2013 among others). Arguments concerning the relationship between Google searches and economic decisions can be found in the theory of buyer behavior (Howard and Sheth, 1969) or in arguments from Barber and Odean (2008) regarding stock markets. In this context, more examples are apparent, such as Vlastakis and Markellos (2012) or Joseph et al. (2011) in the US market, Moussa et al. (2017) in France or Bank et al. (2011) in Germany. Another field within the economic area that has addressed this issue in a prolific way has been the commodities market (Li et al., 2015; Peri et al., 2014).

* Corresponding author.
E-mail address: mgonf@unileon.es (M. González-Fernández).

https://doi.org/10.1016/j.econlet.2018.05.031
0165-1765/

Related to the aim of this paper, Fondeur and Karamé (2013) analyzed whether Google econometrics could predict unemployment in France between 2004 and 2011. They found that Google data enhanced predictions for unemployment. Similarly, Choi and Varian (2012) study unemployment in the US and state that models, including Google econometrics, outperform baseline prediction models. Similarly, McLaren and Shanbhogue (2011) perform a study on the unemployment and housing markets in the UK. They found that Google data is useful for predicting both economic indicators. To the best of our knowledge, our paper is the first that addresses this issue in the context of the Spanish labor market.

2. Data

To approximate internet activity, we use Google data search volumes. We use the Google Search Volume Index (SVI) since it is used in almost all the previous literature, and it is the web browser with the greatest diffusion in Spain (more than 95% of users in Spain use Google rather than other options). Google provides data for the queries that include the selected keywords. Nevertheless, the data SVI_t do not represent the total number of searches for a certain keyword in t. SVI_t is calculated by dividing the volume of queries that contain the keywords V_t^q by a sample,[1] of the entire volume of queries submitted during the same period $V_{e,t}^q$.

$$SVI_t = \frac{V_t^q}{V_{e,t}^q} \tag{1}$$

Subsequently, this value is normalized. Thus, the final index ranges from zero to 100. The data can be filtered for a territory or a category. Since we analyze the Spanish labor market, we have collected the data for Spain between January 2004 and November 2017, which represents the most current sample available. The frequency of the data is monthly. As it is expected that people who fear losing their jobs are more prone to make queries about this eventual situation, we have selected the word *desempleo,* which is the Spanish word for unemployment.[2]

The data for the labor market were obtained from Eurostat every month. Specifically, we collected the unemployment series as a percentage of the active population for Spain from January 2004 to November 2017. Fig. 1 represents the evolution of unemployment data and the Google searches for the word *desempleo*.

3. Results and discussion

First, we analyzed the correlation between the Google data series and unemployment. In Fig. 1, a similar trend in unemployment data and Google SVI is observed for *desempleo*. Table 1 shows the correlations between unemployment in the period *t+1* and the previous three months of Google searches for our keyword, verifying that the series are highly correlated.

Following Choi and Varian (2012), we ran regression models to test whether the Google SVI can be useful for short-term economic

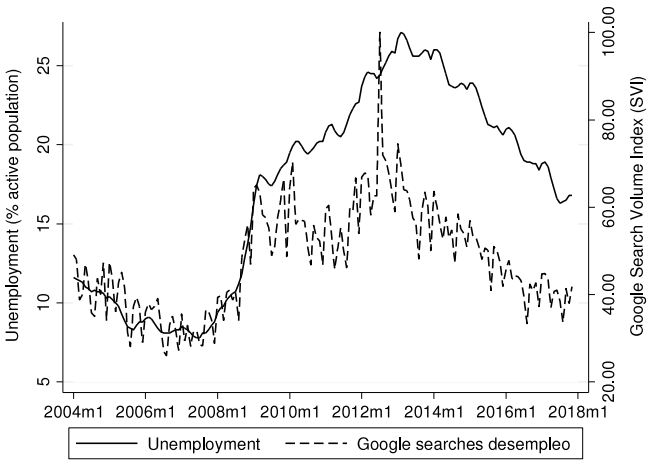

Fig. 1. Unemployment and Google SVI for *desempleo*.

predictions. For this purpose, we used SVI itself (model I) and lagged one period (model II). The direct inclusion of SVI allows the simplification of the estimation. Nevertheless, to be thorough in the analysis, we also include other robustness checks. Namely, we include the abnormal SVI (ASVI) following Da et al. (2011). We calculate the ASVI as the log of the SVI during the current month minus the log median SVI during the previous six months[3]:

$$ASVI(median)_t = log\,(SVI_t) - \log(median(SVI_{t-1}, SVI_{t-2}, SVI_{t-3}, SVI_{t-4}, SVI_{t-5}, SVI_{t-6})) \tag{2}$$

We also use the average (ASVI(average)$_t$) instead of the median and the log of the six months moving average (ASVI(MA)$_t$) as robust measures for ASVI. These measures allow us to identify extreme changes in SVI that could predict unemployment other than the SVI data itself. Table 2 displays the main estimation results for testing the ability of Google Trends data to forecast unemployment in Spain. The first model (baseline) represents the AR (1) baseline model. It is worth noting that the coefficient is close to one, which suggests that it is a random walk process (Choi and Varian, 2012). This means that the best univariate forecast is itself lagged. Nevertheless, we incorporate SVI data and the robustness measures to test whether the baseline model will improve.

Models I and II represent the contemporaneous and lagged SVI for *desempleo* which represent the main models of the analysis. Subsequently, we incorporated the ASVI robustness measures (models III through V). All the coefficients for SVI data are highly significant with the expected positive sign, and the inclusion of the SVI slightly enhances the predictive accuracy of the models since the R^2 coefficient and the *RMSE* indicators improve slightly. The results for the robustness checks, i.e., ASVI, were also as expected. To test the forecast accuracy we used Diebold–Mariano test (Diebold and Mariano, 1995). We obtained the prediction for the unemployment series for the whole sample for each of the models. Then, we ran the test, which compared the actual unemployment data and two competing predictions. The results indicate that models which include Google, in general, better forecast unemployment rather than the baseline, with model I showing the highest accuracy in the prediction.

To complete the analysis, an out-of-sample forecast evaluation was also performed. We used it to compare how the models predict unemployment. First, each model was estimated up until

[1] Google uses a random sample to calculate the index. Therefore, the data provided by Google can slightly differ if we collect the data on different days (Carrière-Swallow and Labbé, 2013; Choi and Varian, 2012; McLaren and Shanbhogue, 2011). Taking this into account, we collected the data on ten different days and calculated the average value.

[2] We assume that *desempleo* summarizes the queries about the situation in which people lose their jobs. Other Spanish expressions intimately related to unemployment are *prestación por desempleo* and *subsidio por desempleo,* which are the Spanish expressions that refer to the payment that unemployed people receive from the government, i.e., unemployment benefits. Therefore, those expressions also include the word *desempleo,* and their queries will be taken into account in Google search data even if we only consider the word *desempleo* Another keyword to consider is *paro*, which is a Spanish colloquial word for *desempleo*. Nevertheless, *paro* can lead to misunderstandings since it has other meanings that do not involve unemployment.

[3] We selected six months heuristically. We also performed the analysis using other time spans, such as three months, and the results remained robust.

Table 1
Correlation analysis.

	$Unemployment_{t+1}$	$SVI\ desempleo_{t}$	$SVI\ desempleo_{t-1}$	$SVI\ desempleo_{t-2}$
$Unemployment_{t+1}$	1.000			
$SVI\ desempleo_{t}$	0.783***	1.000		
$SVI\ desempleo_{t-1}$	0.791***	0.819***	1.000	
$SVI\ desempleo_{t-2}$	0.795***	0.772***	0.819***	1.000

Notes: SVI refers to the Google Search Volume Index for the associated keyword.
*** Significant at 1%.

Table 2
Regression analysis.

	Baseline	I	II	III	IV	V
$Unemployment_{t-1}$	0.996***	0.957***	0.975***	0.994***	0.996***	0.971***
	(0.004)	(0.008)	(0.008)	(0.003)	(0.004)	(0.010)
SVI_{t}		0.026***				
		(0.005)				
SVI_{t-1}			0.013***			
			(0.003)			
$ASVI\ (median)_{t}$				1.528***		
				(0.230)		
$ASVI\ (average)_{t}$					1.502***	
					(0.235)	
$ASVI\ (MA)_{t}$						0.697**
						(0.269)
Constant	0.105	−0.501***	−0.204**	0.153**	0.111	−2.158**
	(0.071)	(0.130)	(0.089)	(0.069)	(0.068)	(0.864)
R^2	0.9963	0.9974	0.9966	0.9973	0.9972	0.9964
RMSE	0.3845	0.3232	0.3708	0.3316	0.3359	0.3825

The dependent variable is the percentage of unemployment over the active population in months *t*. SVI indicates Google searches for the keyword *desempleo*. ASVI indicates the Abnormal Search Volume index for the keyword *desempleo*.
*** Significant at 1%.
** Significant at 5%.

Table 3
Out-of-sample forecast evaluation for unemployment.

	Baseline	I	II	III	IV	V
RMSE	0.4129	0.3343	0.3888	0.3595	0.3663	0.4172

Each model is estimated for the period up to December 2010, which represents half of the sample. Then, we generate a one-step-ahead forecast up to the end of the sample for each model. The results are robust for other starting periods.

December 2010, and then a prediction was produced for the next month (January 2011). Then, the difference between the prediction and actual unemployment data was recorded. We repeat this procedure estimating the model up until January 2011 and with a prediction for unemployment for February 2011 being compared with the real data again. This is replicated until the end of the sample (McLaren and Shanbhogue, 2011). The results are shown in Table 3. Upon observation, the model including contemporaneous SVI itself (model I) produces the smallest errors in the prediction of unemployment. This result is in line with the previous Diebold and Mariano test indicating that solely the inclusion of Google data for *desempleo*, without any other transformation, enhances the unemployment prediction in Spain. Therefore, although the inclusion of more sophisticated proxies, such as the ASVI (model III) also slightly enhances unemployment prediction, the simpler measure provides the best results.

In short, the presented results are consistent with the previous literature (Choi and Varian, 2012; Fondeur and Karamé, 2013; McLaren and Shanbhogue, 2011), indicating that the inclusion of Google econometrics enhances accuracy and prediction of models. This finding can help predict the evolution of the labor market in Spain and other countries or regions with the mere inclusion of contemporaneous SVI data for unemployment.

4. Conclusions

In this paper, we address whether Google search data, what we have called Google econometrics, is useful to predict Spanish unemployment. The results indicate that Google data not only are helpful but also enhance the predictive power of models, indicating that the inclusion of the SVI for queries containing the word *desempleo* improves unemployment predictions. These results note the importance of internet search activity data as a useful tool in economic forecasts that should become widespread in economic research fields in the future.

References

Bank, M., Larch, M., Peter, G., 2011. Google search volume and its influence on liquidity and returns of german stocks. Financ. Mark. Portfolio Manage. 25 (3), 239–264.
Barber, B.M., Odean, T., 2008. All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors. Rev. Financ. Stud. 21 (2), 785–818.
Carrière-Swallow, Y., Labbé, F., 2013. Nowcasting with google trends in an emerging market. J. Forecast. 32 (4), 289–298.
Choi, H., Varian, H., 2012. Predicting the present with google trends. Econ. Rec. 88 (s1), 2–9.
Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. J. Finance 66 (5), 1461–1499.
Dergiades, T., Milas, C., Panagiotidis, T., 2015. Tweets, Google trends, and sovereign spreads in the GIIPS. Oxford Econ. Pap. 67 (2), 406–432.
Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13 (3), 253–263.
Fondeur, Y., Karamé, F., 2013. Can google data help predict french youth unemployment?. Econ. Modell. 30, 117–125.
Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457 (7232), 1012–1014.
Howard, J.A., Sheth, J.N., 1969. The Theory of Buyer Behavior. Wiley, New York.
Joseph, K., Babajide Wintoki, M., Zhang, Z., 2011. Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. Int. J. Forecast. 27 (4), 1116–1127.

Li, X., Ma, J., Wang, S., Zhang, X., 2015. How does google search affect trader positions and crude oil prices?. Econ. Modell. 49, 162–171.
McLaren, N., Shanbhogue, R., 2011. Using internet search data as economic indicators. SSRN Electron. J..
Moussa, F., Delhoumi, E., Ouda, O. Ben, ., 2017. Stock return and volatility reactions to information demand and supply. Res. Int. Bus. Finance 39, 54–67.
Peri, M., Vandone, D., Baldi, L., 2014. Internet, noise trading and commodity futures prices. Int. Rev. Econ. Finance 33, 82–89.
Siganos, A., 2013. Google attention and target price run ups. Int. Rev. Financ. Anal. 29, 219–226.
Vlastakis, N., Markellos, R.N., 2012. Information demand and stock market volatility. J. Bank. Finance 36 (6), 1808–1821.