Economics and Business Letters

8(4), 209-214, 2019

An approach to predict Spanish mortgage market activity using Google data

Marcos González-Fernández[*] • Carmen González-Velasco

University of Leon, Leon, Spain

Received: 14 April 2019
Revised: 12 September 2019
Accepted: 18 Septemeber 2019

Abstract

The aim of this paper is to use Google data to predict Spanish mortgage market activity during the period from January 2004 to January 2019. Thus, we collect monthly Google data for the keyword hipoteca, the Spanish expression for mortgage, and then, we perform a regression and an out-of-sample analysis. We find evidence that the use of Google data significantly improves prediction accuracy.

Keywords: mortgage market; Google data; Spain; out-of-sample
JEL Classification Codes: G2, E27, G40

1. Introduction and theoretical background

The mortgage market has been one of the most important drivers of the Spanish financial sector and economy during recent decades. Although the financial crisis generated considerable distress in the Spanish mortgage market and supposed the end of the Spanish real estate boom, in 2010 (just after the financial turmoil), mortgage loans accounted for 620 billion Euros (Pérez Montes, 2014). Currently, the mortgage market remains an important driver of the Spanish financial sector and represents an important indicator of the condition of the economy. Therefore, in this paper, we attempt to predict mortgage market evolution in Spain using Google data as a predictor.

The use of Google data has rapidly spread in the literature to predict other economic indicators such as unemployment (Choi and Varian, 2012; D'Amuri and Marcucci, 2017; Fondeur and Karamé, 2013; González-Fernández and González-Velasco, 2018), to analyze their impact on stock markets (Ben-Rephael, Da, and Israelsen, 2017; Da, Engelberg, and Gao, 2011, 2015), and to study bond markets (Dergiades, Milas, and Panagiotidis, 2015; Milas, Panagiotidis, and Dergiades, 2018) or their impact on commodities (Han, Li, and Yin, 2017; Peri, Vandone, and Baldi, 2014) with special attention to oil (Bampinas, Panagiotidis, and Rouska, 2019; Han, Lv, and Yin, 2017). However, related to the aim of this paper, some studies have shown conflicting results. On the one hand, Limnios and You (2016) analyze the U.S. housing market using

[*] Corresponding author. E-mail: mgonf@unileon.es.

Citation: González-Fernández, M., and González-Velasco, C. (2019) An approach to predict Spanish mortgage market activity using Google da-ta, *Economics and Business Letters*, 8(4), 209-214.

DOI: 10.17811/ebl.8.4.2019.209-214

Google data. They find that the out-of-sample estimations do not show a higher predictive power with the inclusion of Google data. On the other hand, Beracha and Wintoki (2013) find that Google data are useful to predict changes in housing price in different U.S. cities. Moreover, Wu and Brynjolfsson (2015) demonstrate that Google data are related to future sales and prices in the U.S. housing market. Similarly, Mclaren and Shanbhogue (2011) show that Google data improves the forecast of house prices in the UK.

Following this line of research, our paper analyzes this relationship in Spain. The contribution of the paper is threefold. First, to the best of our knowledge, our paper is the earliest one to address this topic in the context of the Spanish mortgage market, which is one the most active mortgage markets in Europe and has considerable importance for the Spanish financial sector and economy. Second, most of the papers about this topic analyze the U.S. mortgage market, which presents different characteristics compared to the Spanish one. In contrast, we also provide an updated study that covers a long period before and after the financial crisis, which severely affected the Spanish mortgage market. Specifically, our analysis covers the period from January 2004 to January 2019. Similar papers, such as Mclaren and Shanbhogue (2011), use data up to January 2011. Therefore, they do not considered the evolution and prediction of the mortgage market years after the financial crisis. Third, our results demonstrate the ability of Google data to improve mortgage forecasts. This finding might have important implications for the financial sector in order to anticipate the demand of mortgage loans. Moreover, the inclusion of Google data enhances prediction, especially when the data are filtered by category, considering searches within the finance category.

The remainder of this paper is organized as follows. Section 2 summarizes the data. In section 3, we show the main results. Finally, the last section concludes the paper and offers some implications for the results obtained.

2. Data

To conduct our analysis, we need data for two variables of interest: Google data related to the mortgage market and data of real mortgages. Regarding the latter, the data were obtained from the Spanish National Institute of Statistics. Specifically, we gather monthly data for the number of mortgages granted from January 2004 to January 2019. This serves as a measure of the evolution of the mortgage market in Spain. Regarding Google data, we obtain the Google Search Volume Index (GSVI) from Google trends (http://www.google.com/trends/). Google provides data for those searches that include our keywords of interest.

$$GSVI_t = \frac{V_{j,t}^s}{V_{a,t}^s} \tag{1}$$

Notably, the GSVI does not represent the total number of searches; rather, it is an index. For this purpose, Google divides the number of searches (s) that contain our keyword (j) by a random sample[1] of all searches (a) for the same period (t), as shown in Eq. (1). Then, this ratio is normalized into a scale that ranges from 0 to 100, where a higher value indicates a higher level of attention.

One important step is the choice of keywords (j). We have selected the keyword *hipoteca*, which is the Spanish word for mortgage, as we expect that it reveals the attention of users related to the mortgage market[2]. We download the data from Google trends for each month between

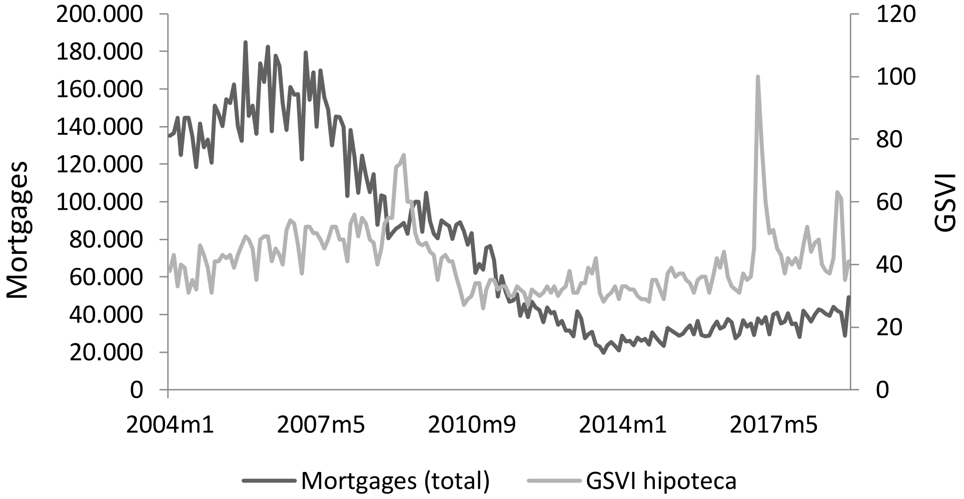
[1] To increase response speed, Google uses a random sample instead of the total number of searches (Da et al., 2015). Thus, Google data downloaded on different dates can differ slightly. To address this issue we download the data on three different dates (March 4th, 8th and 9th, 2019) and use the average value.

[2] We choose the word *hipoteca* since it is specific enough to contain noise that is either relatively small or constant over time and it is not used to describe any other concept (Dzielinski, 2012). Moreover, using the keyword alone, we also consider the results for other longer searches that include the word *hipoteca*. For instance, a query in which

January 2004 and January 2019. Moreover, Google data can be filtered by territory or category. Thus, we have limited our searches to the geographic area of Spain. Regarding the category, we download the data from the *general* category and from the *finance, loans and credits* categories. This filtering process ensures that Google data are actually related to the financial sector and are used as a measure of robustness for the Google data.

Figure 1 displays the evolution of the total number of mortgages granted, which includes all kinds of constructions[3] and Google searches for the word *hipoteca* from January 2004 to January 2019. We can observe a similar pattern in both series and a similar trend, especially after 2010. The GSVI shows a different pattern in the final part of the time span due to changes in the mortgage market law, which are reflected in the peak in the GSVI at the beginning of 2017. In the next section, we analyze the relationship between both series through an in-sample and out-of-sample analysis.

Figure 1. Evolution of total mortgages and GSVI for *hipoteca*.



3. Results and discussion

Table 1 shows the correlation between the total number of mortgages, the number of mortgages granted for urban households and the GSVI for our keyword (*hipoteca*). Ex-ante expectations are for Google searches anticipating mortgage loans. This idea is based in the economic psychology literature, according to which, people search for information before an economic decision, especially when uncertainty about price increases (Lemieux and Peterson, 2011). The table shows that mortgage data are significantly correlated with the GSVI, regardless of the category we consider. However, the GSVI filtered by category shows a higher correlation, approximately 45%, than those performed in the general category.

Table 2 displays the results[4] for an in-sample evaluation through a regression analysis considering an AR (1) model as the baseline model (Choi and Varian, 2012; McLaren and Shanbhogue, 2011). This is expected to be the best univariate forecast since most macroeconomic series are characterized by random walks (Nelson and Plosser, 1982). We observe that the baseline model shows a coefficient close to one. This finding confirms that the variable representing mortgages is a random walk, and thus, the baseline is an adequate forecast model. However,

somebody searches for *precio hipoteca*, which is the translation of mortgage price, was considered. Besides that the selected keyword should be broad enough and not excessively specific (Beracha and Wintoki, 2013).

[3] The total number includes rustic and urban households, grounds, depots, garages and other constructions in which a mortgage is involved.

[4] For the sake of brevity, we only display the results for the total number of mortgages. The results for mortgages for urban households are similar to those displayed in Table 2.

we might improve this baseline forecast by including other predictors, namely, the GSVI (Choi and Varian, 2012). In models 1 to 3, we include the GSVI for the keyword *hipoteca.* The coefficients for the GSVI show the expected positive and significant sign in all models. The values for the R^2 and RMSE slightly improve with the presence of the GSVI. This result indicates that the addition of Google data slightly enhances the predictive power of the baseline model.

Table 1. Pairwise correlations between the number of mortgages granted and Google data for the keyword *hipoteca*.

	1	***2***	***3***	***4***	***5***
Mortgages (total) (1)	1.000				
Mortgages (urban households) (2)	0.996***	1.000			
GSVI *hipoteca* (3)	0.355***	0.351***	1.000		
GSVI *hipoteca* (Finance) (4)	0.435***	0.432***	0.989***	1.000	
GSVI *hipoteca* (Credits and Loans) (5)	0.463***	0.464***	0.980***	0.995***	1.000

Notes. GSVI *hipoteca* refers to the searches in the general category, GSVI *hipoteca* (Finance) are the searches within the finance category and GSVI *hipoteca* (Credits and Loans) are the searches within the credits and loans category. *** Significant at 1%.

Table 2. Regression analysis.

Dependent variable: total mortgages (logs)	***Baseline***	***Model 1***	***Model 2***	***Model 3***
Total mortgages$_{t-1}$ (logs)	0.969***	0.946***	0.938***	0.935***
	(0.017)	(0.018)	(0.019)	(0.019)
GSVI *hipoteca$_t$* (logs)		0.163***		
		(0.048)		
GSVI *hipoteca$_t$* (Finance) (logs)			0.167***	
			(0.046)	
GSVI *hipoteca$_t$* (Credits and Loans) (logs)				0.167***
				(0.046)
Constant	0.328*	-0.021	0.057	0.088
	(0.195)	(0.212)	(0.198)	(0.196)
R^2	0.947	0.949	0.950	0.950
RMSE	0.155	0.978	0.973	0.971

Notes. The table shows the regressions for the baseline and other models including the GSVI for the keyword *hipoteca.* All variables are expressed in logs. As in D'Amuri and Marcucci (2017), we display the *RMSE* value for the baseline model, whereas for the rest of the models we report the ratio of each model and the baseline. Thus, a ratio below one suggests that the model improves the baseline prediction. *Significant at 10% *** Significant at 1%.

Table 3. Out-of-sample forecast evaluation.

Dependent variable: total mortgages (logs)	***Baseline***	***Model 1***	***Model 2***	***Model 3***
RMSE	0.168	0.977***	0.973***	0.977**

Notes. The table shows the RMSE values for the out-of-sample estimates. We first run a regression until December 2010, and then, we generate a one-month-ahead forecast up to the end of the sample for each model. Model 1 includes the keyword *hipoteca* within the general category; Model 2 includes the keyword *hipoteca* within the finance category; Model 3 includes the keyword *hipoteca* within the credits and loans category. As in D'Amuri and Marcucci (2017) we display the RMSE value for the baseline model, whereas for the rest of the models we report the ratio of each model and the baseline. Thus, a ratio below one suggests that the model improves the baseline prediction. *** and ** indicate rejection at the 1% and 5% respectively of the Diebold and Mariano (1995) test under the null hypothesis that the forecast adequacy is equal between the baseline and each of the models.

Table 2 indicates that Google data improve the prediction of the mortgage market in Spain. However, in-sample forecasting implies perils (Choi and Varian, 2012). Therefore, to confirm this result, an out-of-sample procedure is needed. We perform a one-month-ahead forecast to test the ability of Google data to predict the mortgage market. Specifically, we estimate each model until December 2010, and then, we predict the next month. This procedure repeats one month at a time until the end of the sample. Eventually, we collect the predictions for all the models and perform the Diebold and Mariano (1995) test to check whether the forecasts are significantly different. The results are reported in Table 3.

The out-of-sample analysis reveals that the models which include the GSVI outperform the baseline model. Moreover, the Diebold and Mariano (1995) test shows that the forecast accuracy is different between the models indicating that GSVI inclusion significantly improves the prediction. Notably, the GSVI obtained from the finance category seems to have greater ability to enhance the prediction. This is an interesting finding since previous studies do not perform this filtering process (Beracha and Wintoki, 2013; Limnios and You, 2016; McLaren and Shanbhogue, 2011; Wu and Brynjolfsson, 2015) and it indicates that filtering Google data by category and restricting the searches for the finance category provides better forecasting results than the general category.

4. Conclusions

This paper analyses the ability of Google data to predict the mortgage market in Spain. The results confirm that the inclusion of Google data enhances the predictive power of the estimations. These findings are in line with previous studies and support the importance of Google data as an alternative proxy to measure investor attention in the economic field.

These empirical findings may have important implications for the Spanish housing market. First, we have demonstrated that Google data can predict mortgages' evolution. Therefore, considering that Google data are easily available and are more transparent than other information sources (Da et al., 2015), the data can provide important insights to banks and financial institutions involved in the housing market to anticipate changes in mortgages' demand. In this sense, a deeper analysis of Google data by regions can provide a clearer picture of the impact of Google searches on the Spanish mortgage market. This is one of the future research projects for which this paper has served as starting point. Second, Google data have proven to be a useful indicator of economic behavior (McLaren and Shanbhogue, 2011). Thus, those data ought to be monitored by the Bank of Spain as it is done in other countries, such as the UK, to help with the prediction of the Spanish economy.

Acknowledgements

We would like to thank the Editor and the referees for providing us with constructive comments to improve the initial version of this paper. This work was supported by the Ministerio de Economía, Industria y competitividad, Gobierno de España [research project number ECO2017-89715-P, entitled "El Análisis del Riesgo en los Mercados Financieros"].

References

Bampinas, G., Panagiotidis, T., and Rouska, C. (2019).Volatility persistence and asymmetry under the microscope: the role of information demand for gold and oil, *Scottish Journal of Political Economy*, *66*(1), 180–197.

Ben-Rephael, A., Da, Z., and Israelsen, R. D. (2017) It Depends on Where You Search: Institutional Investor Attention and Underreaction to News, *The Review of Financial Studies*, *30*(9), 3009–3047.

Beracha, E., and Wintoki, M. B. (2013) Forecasting Residential Real Estate Price Changes from Online Search Activity, *The Journal of Real Estate Research*, *35*, 283–312.

Choi, H., and Varian, H. (2012) Predicting the Present with Google Trends, *Economic Record*, *88*(s1), 2–9.

D'Amuri, F., and Marcucci, J. (2017) The predictive power of Google searches in forecasting US unemployment, *International Journal of Forecasting*, *33*(4), 801–816.

Da, Z., Engelberg, J., and Gao, P. (2011) In Search of Attention, *The Journal of Finance*, *66*(5), 1461–1499.

Da, Z., Engelberg, J., and Gao, P. (2015) The Sum of All FEARS Investor Sentiment and Asset Prices, *Review of Financial Studies*, *28*(1), 1–32.

Dergiades, T., Milas, C., and Panagiotidis, T. (2015) Tweets, Google trends, and sovereign spreads in the GIIPS, *Oxford Economic Papers*, *67*(2), 406–432.

Diebold, F. X., and Mariano, R. S. (1995) Comparing predictive accuracy, *Journal of Business & Economic Statistics Jul*, *13*(3), 253–263.

Dzielinski, M. (2012) Measuring economic uncertainty and its impact on the stock market, *Finance Research Letters*, *9*(3), 167–175.

Fondeur, Y., and Karamé, F. (2013) Can Google data help predict French youth unemployment?, *Economic Modelling*, *30*, 117–125.

González-Fernández, M., and González-Velasco, C. (2018) Can Google econometrics predict unemployment? Evidence from Spain, *Economics Letters*, *170*, 42–45.

Han, L., Li, Z., and Yin, L. (2017) The effects of investor attention on commodity futures markets, *Journal of Futures Markets*, *37*(10), 1031–1049.

Han, L., Lv, Q., and Yin, L. (2017) Can investor attention predict oil prices?, *Energy Economics*, *66*, 547–558.

Lemieux, J., and Peterson, R. A. (2011) Purchase deadline as a moderator of the effects of price uncertainty on search duration, *Journal of Economic Psychology*, *32*(1), 33–44.

Limnios, C., and You, H. (2016) Can Google Trends Actually Improve Housing Market Forecasts?, *SSRN Electronic Journal*.

McLaren, N., and Shanbhogue, R. (2011) Using Internet Search Data as Economic Indicators, *SSRN Electronic Journal*.

Milas, C., Panagiotidis, T., and Dergiades, T. (2018) Twitter versus Traditional News Media: Evidence for the Sovereign Bond Markets, *SSRN Electronic Journal*.

Nelson, C. R., and Plosser, C. R. (1982) Trends and random walks in macroeconmic time series, *Journal of Monetary Economics*, *10*(2), 139–162.

Pérez Montes, C. (2014) The effect on competition of banking sector consolidation following the financial crisis of 2008, *Journal of Banking & Finance*, *43*, 124–136.

Peri, M., Vandone, D., and Baldi, L. (2014) Internet, noise trading and commodity futures prices, *International Review of Economics & Finance*, *33*, 82–89.

Wu, L., and Brynjolfsson, E. (2015) Chapter 3 - The Future of Prediction: How Google searches foreshadow housing prices and sales, in *Economic Analysis of the Digital Economy* (pp. 89–118). University of Chicago Press.