The power of Google search data; an alternative approach to the measurement of unemployment in Brazil

Fernando Lasso
Erasmus University Rotterdam
FernandoMLassoP@student.eur.nl

(Sebastian Snijders)
Erasmus University Rotterdam
Sebastiansnijders@student.eur.nl

ABSTRACT

By the use of econometric techniques, this paper extends the application of predictive methods for unemployment rates by the use of Google Trends from developed western countries to the case of Brazil. Indeed, Google search volumes of keywords related to job-search turn out to contain significant predictive power and biweekly search data can predict the direction of the unemployment rate with around 80 percent accuracy, beating baseline methods using seasonal patterns by more than 15 percent.

Keywords

Google Trends, Unemployment, Brazil, Econometrics

INTRODUCTION

Since the advent of the internet, there has been an explosion in collected and obtainable data. However, in the domain of economic indicators, a significant amount of the available data is still not being used to its full potential. Several studies have shown how Google search data can be a powerful predictive tool for economic indicators such as GDP, retail sales and many more (Koop et al., 2013; Choi & Varian, 2012). Especially the estimation of unemployment figures has been a topic of interest across a multitude of developed countries, which resulted in methods that are able to put real-time search data into reliable projections. This procedure, dubbed as 'nowcasting' the economy, led to successful econometric models in Germany, Belgium, the United Kingdom and Finland (Askitas & Zimmermann, 2009; Bughin, 2011; McLaren & Shanbhogue, 2011; Tuhkuri, 2014). Approximations of the development of the unemployment rate rely heavily on surveys and official reports of other government entities, which may not always be a reliable source. Especially in developing nations, institutions may not possess the accuracy that may be desired for a fundamental analysis of their economies. Unfortunately, previous research on the nowcasting of unemployment lacks the extension to non-western countries with weaker institutions. This paper will tackle the case of Brazil and attempt to broaden previous models to provide real-time estimates of unemployment in this South American country.

SRC 2016, November 30, 2016, The Netherlands.

Currently, Brazil faces a desperate recession and a political crisis. This unfortunate situation will inevitably result in the loss of thousands if not millions of jobs. It would be of social interest to produce real-time estimates of the unemployment rate or to forecast these figures using alternative methods. The social and economic benefits of such a model range from being a useful tool in policy making to producing early warning signals of economic crises in the future. One simple and accessible source of data to facilitate the estimation of these indicators in a timely manner is the volume of Google keyword searches, which is available in real-time and sorted into fine-grained categories. To contribute to the subject, this paper will explore the following research question:

To what extent do Google keyword search volumes contain explanatory and predictive power for the monthly unemployment rate of Brazil?

After the introduction, the theoretical framework will describe the economic landscape of Brazil, extend on the use of Google search data in past research and lay out the hypotheses of the paper. Then, the data and methodology sections will go into detail on the selection of appropriate sources, the cleansing of data and the use of models. In the results section, two specifications of our model will be presented, each containing a significant degree of predictive power. The final models can predict the direction in which the unemployment rate will move with around 80 percent accuracy halfway through the month. These techniques significantly improve baseline methods based on seasonal patterns, and thus can be deemed useful tools in decision making or fundamental analysis using forecasts. Finally, the conclusion will wrap up the paper with a summary of relevant outcomes and interpretations.

THEORETIAL FRAMEWORK

This paper proposes an alternative method to effectively trace the volatile unemployment figures in Brazil, by making use of publically available Google search trends. The internet plays a growing role as the primary intermediary between employees and businesses. This gives plenty of reason to suspect that variation in search behavior associated with job searches may have a strong correlation with the actual unemployment. Hence, the following hypothesis is investigated.

Hypothesis 1: The volume of Google keyword searches related to unemployment is correlated with the monthly *unemployment rate in Brazil.*

As different countries may have distinct forces that influence their economic environment, it is not possible to extrapolate European models to the Brazilian economy. Furthermore, non-western societies are more likely to face lower literacy rates and limited access to the internet, which may hinder previously proposed models. Therefore, this paper will attempt to fill the void in the existing literature by investigating the feasibility of a predictive model for the case of Brazil.

Hypothesis 2: The volume of Google keyword searches related to unemployment contains predictive power for the monthly unemployment rate in Brazil.

Since it is the aim to approximate real unemployment rates, reliable estimates of the actual figures will be needed to test the model against. Unfortunately, official government estimates have been prone to adjustment in the past (Reuters, 2014). This is largely due to the issue of hidden unemployment. The adoption of new methods and further revision caused official unemployment estimates to be volatile, even many years after their initial publishing. Therefore, a special procedure will be applied to address the issue of unreliable government data. For the purpose of this paper, the range of the analysis will be split into period one, the time before Rousseff was elected in 2011 and period two, the years following her election. As the unemployment rates in period one have had sufficient time to be revised, the data will be fitted onto the official figures between 2004 and 2011. By making the assumption that the official government report error terms are not negatively or positively biased, the model will be used to reliably forecast the unemployment rate in period two.

To assert the effectiveness of the model, it must be compared to alternative techniques of estimation. The methodology section will lay out two alternative baseline methods to test for the last hypothesis.

Hypothesis 3: The model for estimating Brazils unemployment rates by Google search volumes significantly improves baseline forecasting techniques.

DATA

To begin with, data regarding the Brazilian unemployment rate were obtained from the Brazilian Institute of Geography and Statistics, Instituto Brasileiro de Geografia e Estatstica (IBGE). To answer the research question of this paper, the published monthly “unemployment rate” will be utilized. Although unemployment rates are often seasonally adjusted to account for hiring and layoff patterns during the calendar year, the data will not be adjusted in order to reflect the real unemployment. Besides, it is especially interesting to analyze whether Google search volumes follow these seasonal patterns. To match the Google search data, the Brazilian unemployment rates from January 2004 to February 2016 will be considered.

The independent variables for the models were obtained using the free database “Google Trends”, provided by Google Inc. The data contains the relative interest for particular Google search keywords over a specified range of time in customized geographic areas, normalized to an interest factor between a value of 0 and 100.

For this research, a multitude of keyword interest series related to unemployment has been downloaded on both a weekly and monthly level. Next to using their absolute levels, first differences and seasonally adjusted duplicates of the search data are taken. The seasonal adjustment eliminates patterns throughout the data and transforms undulations into steady trends. For a list of the investigated keywords, refer to Table I. More recent data is available, but has been truncated because the unemployment data of Brazil is available only until February 2016.

Table I. Description of used keywords

Keyword	Description
Décimo terceiro salário (DTS)	Social security program
Empregos	Jobs
FGTS	Severance Indemnity Fund
INSS	National Institute for Social Security
Seguro desemprego	Employment Insurance
Unemployment & Social benefits (index)	General interest in unemployment
Job vacancies (index)	General interest in job searches

METHODOLOGY

The model that will be investigated is a standard linear model with a combination of search levels and first differences:

$$\Delta U_t = \beta_0 + \beta_1 S_{1,t} + \ldots + \beta_k S_{k,t} + \gamma_1 \Delta S_{1,t} + \ldots + \gamma_p \Delta S_{p,t} + \varepsilon_t$$

The unemployment rate at the end of a certain month is defined as U_t and can be inferred from the change in unemployment of that month, defined as ΔU_t. In addition, the search level of keyword k in month t is defined as $S_{k,t}$. $\Delta S_{k,t}$ and ΔU_t are defined as the change in the search level of the keyword k and in the unemployment rate between month t and month t − 1 respectively. Finally, the coefficients are denoted by the parameter β, where β_0 represents the constant.

Since the Google search data is provided in both weekly and monthly intervals, the model will be specified on a monthly and on a semi-monthly level. For the semi-monthly data, only the first 15 days of the month will be examined, in order to investigate whether this data can be used to forecast the unemployment figure of the entire month.

To avoid over-fitting the data by an excessive number of independent variables, it is the aim to find a model with the five most significant explanatory variables (excluding the constant). To determine the optimal combination of keywords, permutation selection will be utilized, by using the explained variance (R-squared) as indicator of model performance. A high R-squared with significant coefficients for the independent variables would be evidence for the existence of a correlation between Google search volumes and the Brazilian unemployment in the case of monthly data and signify predictive power when using biweekly data. The combinatorial regression will be performed in-sample over the period of January 2004 until December 2010 (period one). Once an optimal combination of keywords has been found, model performance will be evaluated using both in-sample (period one) and out-of-sample (period two) statistics.

Finally, to address hypothesis three, the models will be tested against two baseline scenarios. The first one using the change in the previous period as a current estimate for the change in unemployment. It is reasonable to assume that when unemployment increases/decreases in one month, it will follow the same direction in the next one.

Baseline method 1: $\Delta U_t = \Delta U_{t-1}$

The second baseline method makes use of yearly seasonal patterns to establish an educated guess. It uses the exact same rate as 12 months earlier as an estimation for current figures.

Baseline method 2: $\Delta U_t = \Delta U_{t-12}$

By establishing these baselines, root mean square errors (RMSE) and hit rates can be compared with the search-based models in order to examine whether they provide any predictive power. RMSE is a measurement that computes the average distance from the forecasts and reported rates and the hit rate tallies the number of times the forecast took the same direction (increase or decrease) as the reported rate.

RESULTS

The results of model one have been determined in two specifications. The first specification uses the monthly search data while the second specification contains semi-monthly search data. The results of the regressions are displayed in Table II using the RMSE and hit rates as performance measures. With regards to the first two hypotheses, it can be concluded from the R-squared above 0.7 that all search variables are strongly correlated with the unemployment of Brazil. Hence, search behavior seems to exhibit patterns similar to the unemployment rate in the country. The power of the biweekly models is their ability to efficiently approximate the unemployment rate halfway through the month already. The estimated parameters for both models are given in Table III. To illustrate how close the biweekly forecasts of the change in unemployment get to the actual values, a graph is given in Figure I.

Table II. Performance measures
(1% significant values contain '*')

	In-Sample			Out-of-Sample	
	R^2	RMSE	HR	RMSE	HR
Biweekly	0.7	0.27	0.82^*	0.31	0.79^*
Monthly	0.73	0.25	0.87^*	0.32	0.76^*
Baseline 1	0.32	0.67	62	0.49	0.56
Baseline 2	0.45	0.4	72	0.28	0.71

Table III. Estimated model parameters, Jan. 2004 – Dec. 2010
(All coefficients are significant on 1% level, except for 'constant')

	Monthly	Biweekly
Constant	−0.0557	−0.1178
Empregos (U)	0.0520	0.0581
Empregos (A)	−0.0521	−0.0568
DTS (U)(D)	0.0139	0.0078
DTS (A)(D)	−0.0070	
Vacancy Index (U)(D)	0.0124	
Unemp. Index (U)(D)		−0.0050
Empregos (A)(D)		0.0460

Note: 'A' for adjusted, 'U' for unadjusted data and 'D' for first level differences.

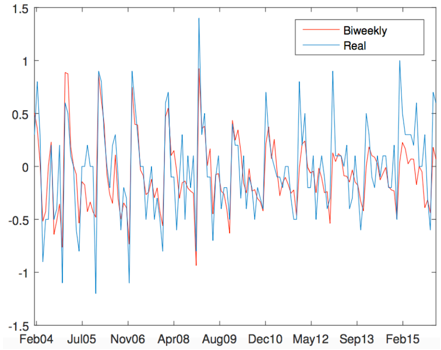

Figure I. Real to biweekly forecasted change in unemployment

Judged by in-sample performance, it is clear that the monthly model performs best on all metrics. The biweekly model also provides a high in-sample variance explanation.

When examining out-of-sample performance, it must first be noted that the second baseline (lagged by one year) has the lowest root mean squared error (RMSE). Hence, there seems to be a seasonal component in play. Moreover, the null hypothesis of equal RMSE is rejected for all other models (with $p < 0.01$), except for the biweekly model (using a 5% significance level). Subsequently, there is no reason to assume that the biweekly model provides worse forecasts than baseline 2 in terms of RMSE out of sample. In addition, the biweekly model also proves to behave significantly better when judged by its out-of-sample hit rate of 0.76, which measures the fraction of times the model was able to forecast the correct direction of the move (up or down). Using a two-proportion z-test, all models were found to provide better forecasts than baseline model 1 in terms of hit rates. Both models significantly improved on the baselines in terms of hit rates. The ratio of the correctly predicted direction is around 15 percent higher and reaches an accuracy of above 75%. If one only considers the last year of the sample, the hit rate even surpasses 80%.

In Table III, one can observe that the parameters of the keywords “Empregos” and “Décimo terceiro salário” and their seasonally adjusted versions to (partly) cancel each other out. The remainder of this equation is the seasonal patterns that appears to be a strong predictor of unemployment. Therefore, it can be concluded that a large part of the explained variation is the prediction of seasonal patterns. Still, this outcome is very interesting, as it provides strong evidence that Google searches indeed directly correlate with real unemployment.

Summarizing, it can be concluded that, when examining both in-sample and out-of-sample performance, all models exhibit significant predictive power. Their predictive power exceeds the seasonal patterns from the baselines, since they were successful in producing correct estimations in 2015, when most seasonal pattern were disrupted due to worsening economic conditions. For these reasons, the first two hypotheses cannot be rejected and thus the models prove to contain explanatory and predictive power for the Brazilian unemployment. Since the biweekly method seems to improve current baseline methods on all metrics, also the third hypothesis is not rejected.

CONCLUSION

The era of big data has just begun, and once more we can catch a glimpse of its immense power. As proven by this paper, it is possible to predict the direction of the monthly Brazilian unemployment rate halfway through the month with more than 80 percent accuracy, solely based on Google search volumes. Surprisingly, the results found in this paper do not significantly differ from previous research in western nations. It is a worthy addition to this line of research to conclude that the social and economic climate of Brazil does not affect the predictive power of Google-based forecasting models. The implications of models containing such predictive power entail an efficient and simple way of measuring the economic climate of Brazil. Further research could strive to optimize the techniques and models as prescribed in this paper by examining a larger pool of potential keywords or improved models to extract more information from the data. For instance, the models used in this paper were based on data from 2004 to 2011, but as time passes, one could increase the time range and re-calibrate the models.

In addition, the possibility of adaptive calibration could be further investigated. Moreover, the methods could be altered to facilitate forecasting multiple months ahead. Eventually, because these models proved to be especially effective during the economic downturn in 2015, these techniques could be extended into useful tools to create early warning signals for volatile unemployment rates and recessions. In any case, it is clear that Google search data reflects real economic behavior and could be a valuable device in the future of econometrics.

ROLE OF THE STUDENT

When Sebastian proposed to investigate the power of Google Search Data, Fernando initially intended to target the Netherlands, on which Dr. Franses proposed to focus on a developing country. Subsequently, both Authors spent a lot of time on the structure and choice of the research and its methods. Finally, while Sebastian specialized in the cleansing and organization of the Data and its application in eViews, Fernando took on most of the writing.

ACKNOWLEDGEMENT

Special thanks to Professor Franses, dean of the Erasmus School of Economics for the supervision of the paper. His contributions and ideas were vital for the project.

REFERENCES

1. Askitas, N., & Zimmermann, K.F. (2009). Google econometrics and unemployment forecasting. German Council for Social and Economic Data (RatSWD) Research Notes, 41.
2. Bughin, J.R. (2011). ’Nowcasting’ the Belgian Economy. Available at SSRN 1903791.
3. Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. Economic Record, 88(s1), 2-9.
4. Koop, G., & Onorante, L. (2013). Macroeconomic nowcasting using Google probabilities. University of Strathclyde.
5. McLaren, N., & Shanbhogue, R. (2011). Using internet search data as economic indicators. Bank of England Quarterly Bulletin, Q2.
6. Reuters (2014). Brazil’s jobless rate revised up under new methodology. Thomson Reuters.
7. Tuhkuri, J. (2014). Big Data: Google Searches Predict Unemployment in Finland. ETLA Reports, 31.