Creating an Unemployment Rate Indicator for Brazil using Google Trends Data

Talitha Speranza[*,a], Raíra Vieira[a], Pedro Costa Ferreira[a]

[a] *Instituto Brasileiro de Economia, Fundação Getúlio Vargas (FGV/IBRE). Rua Barão de Itambí 60, Rio de Janeiro, RJ, Brazil.*

Abstract

This is the abstract.

It consists of two paragraphs.

Introduction

The use of search engine data has a high potential for guiding and improving public policy. This statement is based on four facts about the nature of these data. First, the frequency with which we can obtain them is limited only by the computational power that is available: in theory, any granularity is possible - monthly, weekly, daily, hourly, minute to minute or even fractions of minutes. This is already an obvious advantage in relation to what exists today, since most of the socioeconomic variables are only available in frequencies higher than the weekly ones.

Secondly, the cost of extracting information is very low compared to laborious research conducted by public statistical institutes around the world. The censuses produced every ten years by IBGE are examples of this type of high-cost work. The population coverage of census data is, of course, much higher than that of social networking data. However, the increasing number of users of these virtual spaces already makes them potentially more fertile than usual samples. In the National Household Sample Survey (PNAD), for example, the IBGE sample is about 360 thousand people, while half of the Brazilian population has Facebook accounts.

In this project, we intend to develop indicators for the Brazilian labor market. Our intention is, above all, to overcome the low frequency with which the IBGE releases the data related to the theme. Official employment measures in Brazil are released about 1 month after the reference date. If government entities anticipated these variables more quickly, they could respond with increased effectiveness.

An extra knowledge can also be provided by these data when separating them by different regions or states of the federation. With this information, the

[*] Corresponding Author

Email addresses: talitha.speranza@fgv.br (Talitha Speranza), raira.vieira@fgv.br (Raíra Vieira), pedro.guilherme@fgv.br (Pedro Costa Ferreira)

government could design solutions tailored to specific locations before the problem worsens and at low cost. Currently, the PNAD can only be disaggregated in this way in its quarterly (hence late) disclosure and there are very few observations available: the research, as it is today, only began in 2012.

Therefore, our goal is to create models that generate reliable signals of Brazilian labor market conditions in real time before official statistics are available. The Brazilians would only have to win.

Previous Literature

- Suhoy (2009) tests the predictive ability of Google Trends' thematic search indexes for Israel's business cycles. In particular, Suhoy calculates the probabilities of deceleration in the said series of Google with Bayesian estimation techniques and finds that these probabilities do indeed increase in the proximity of the last Israeli economic crises.

- Choi and Varian (2009) seek to determine whether Google Trends' thematic indexes can increase the predictive power of the number of people entering the Initial Jobless Claims in the US. This series has weekly frequency and is considered a coincident indicator of unemployment. Using an AR (1) model as the baseline, the authors show that adding the search index (which precedes Initial Jobless Claims within 7 days) considerably improves prediction accuracy by 12.9% in the worst case and 15.74% in the best case .

- Askitas and Zimmermann (2009) demonstrate significant correlations between Google-specific keyword search indices and the monthly unemployment rate in Germany. The adjusted values and forecasts of the developed model have a lot of adherence to the official data, as well as anticipating changes in trends in certain important events, such as during the 2008 crisis. An additional advantage of the model is the possibility of generating quality forecasts two weeks before official publication of the unemployment rate by the German government.

- D'Amuri and Marcucci (2010) argue that the Google Trends index for job-related searches is the best predictor of predicting the US unemployment rate. Several models of forecasting out of the sample are compared and evaluated in detail, adopting or not the indicator among the explanatory variables. The bottom line is that models that include the Google index perform better, even if the predictions are compared to those of the Survey of Professional Forecasters.

- Choi and Varian (2012) seek to create baseline models to demonstrate the potential of the Google Trends thematic indexes for nowcasting ("present forecast"). In the five provided examples, all with economic variables, the model with the index selected from Google among its regressors performed better than its merely autoregressive pair.

- Antenucci et al. (2013) present an algorithm for choosing signals derived from text fragments containing terms of interest. The motivation of the authors is the difficulty that potential users of social networking data encounter in selecting relevant time series to solve their problems. In our case, the drawback would be to identify which, among all possible phrases containing "unemployed" or "turned off", for example, would be highly correlated with the unemployment rate. The article shows that phrases determined by humans have low explanatory power, but the proposed algorithm is able to overcome these choices only marginally.

- Daas and Puts (2014) find that there is a strong correlation between mood and Dutch consumer confidence. A mood proxy is obtained from Dutch posts on Facebook and Twitter, which are automatically classified as "positive" or "negative" by a local business. The series are also cointegrated and Granger's causality tests have shown that changes in consumer confidence probably precede those in the mood in 7 days. This allows the mood series to be used to create a matched indicator for trust, but the authors do not come up with this idea.

- Antenucci et al. (2014) use Twitter post data to create job loss, search, and job offer rates for the US. This data is filtered through employment-related phrases such as "I lost my job," and the main components of the generated signals are identified to construct the indexes. The new job loss measure is validated against official data and shows good adherence and predictability in real time. The other two indexes also behave as expected, but their assessments are made only in qualitative terms, comparing their movements with those of proxies and the newly created index of loss.

- Llorente et al. (2015) measure how and what behavioral characteristics of the population can be extracted from data of social networks and, therefore, related to the economic activity of geographic regions. More specifically, they develop a method to measure the incidence of unemployment in Spanish territories through geolocated posts on Twitter. The model is based on differences in daytime rhythms, mobility patterns and communication styles of Spanish users. The conclusion is that the unemployment rate is lower where there are more diversified mobility flows, starting everyday earlier and more accurate grammar. The model reproduces with satisfactory accuracy the incidence of employment in the regions of Spain, taking as input only the data of Twitter.

- Biorci et al. (2016) use a database of about 4 million tweets in Italian, filtered by keywords related to the job market. The objective is to compare the signs obtained with the traditional indicators of employment. Although the results have been hampered by noise in the series constructed from tweets, the authors bring some news: the selection of search terms is derived from advanced linguistic tools and the location of the tweets, together with hierarchical clustering techniques, allowed the territorial analysis of data.

Google Trends as an Information Source

The data provided by the Google Trends API is monthly and standardized within the requested time window, on a scale of 0 to 100. The maximum value of the window is reported as 100 and the remainder is calculated in simple proportion. In the help manual of said API, the following description of the adjustment of the data is found:

> Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, places with the most search volume would always be ranked highest. The resulting numbers are then scaled to a range of 0 to 100 based on a topic's proportion to all searches on all topics. Different regions that show the same number of searches for a term do not always have the same total search volumes.

If, for example, the user requests the observations from January 2017 to March 2017 and receives the series $s = \{50, 100, 25\}$, then the February search volume was the highest, double the volume of January and the quadruple of the volume of March. It is not possible to extract from this series the true proportions, that is, the number of searches for a given term in the desired locality and period divided by the total number of searches in the same locality and period.

At the outset, it was thought that this way of spreading the data would undermine any kind of continuous analysis, since all the values of the series can change when a new data is added, if it is the maximum value of the period. However, this difficulty can be circumvented if observations are transformed into monthly variations. A simple demonstration of this fact is developed below.

Suponha que a série de que o Google dispõe, mas não divulga, seja $G = \{g_{t-3}, g_{t-2}, g_{t-1}\}$ no mês t. Os valores de G são proporções entre o número de buscas por um termo de interesse e o total de buscas, por quaisquer termos, em uma dada região durante os meses $t-3$ a $t-1$. Seja a série transformada e divulgada pelo Google $T(G) = \{T(g_{t-3}), T(g_{t-2}), T(g_{t-1})\}$. Se $max(G) = g_{t-2}$, então $T(g_{t-2}) = 100$ e, por regra de três, $T(g_{t-1}) = 100 \cdot g_{t-1}/g_{t-2}$ e $T(g_{t-3}) = 100 \cdot g_{t-3}/g_{t-2}$.

A variação percentual mensal pode ser calculada por uma função bivariada f tal que $f(x, y) = (y - x)/x$, onde x é a observação mais antiga e y, a mais atual. Então,

$$\begin{aligned} f(g_{t-3}, g_{t-2}) = f(T(g_{t-3}), T(g_{t-2})) = (g_{t-2} - g_{t-3})/g_{t-3} \\ f(g_{t-2}, g_{t-1}) = f(T(g_{t-2}), T(g_{t-1})) = (g_{t-1} - g_{t-2})/g_{t-2} \end{aligned} \tag{1}$$

No mês $t+1$, um novo dado é adicionado a G, de modo que $G = \{g_{t-3}, g_{t-2}, g_{t-1}, g_t\}$. Se $max(G) = g_t$, então a transformação de G deve ser recalculada:

$$T'(G) = \{100 \cdot g_{t-3}/g_t, 100 \cdot g_{t-2}/g_t, 100 \cdot g_{t-1}/g_t, 100\} \tag{2}$$

Felizmente, T' não modifica os resultados de f:

$$\begin{aligned} f(g_{t-3}, g_{t-2}) &= f(T'(g_{t-3}), T'(g_{t-2})) \\ &= (100 \cdot g_{t-2}/g_t - 100 \cdot g_{t-3}/g_t)/100 \cdot g_{t-3}/g_t \\ &= (g_{t-2} - g_{t-3})/g_{t-3} \end{aligned} \tag{3}$$

Therefore, the application of the monthly percentage variation generates a consistent series. But that requires more care to be taken. It is necessary that the IBGE series (PNAD and PME) be transformed so that their movements have the same supposed basis for the trajectory of Google series. In its original undisclosed form, Google series contain percentages. The staggered form between 0 and 100 is just an integer representation of these percentages. So when the monthly percentage change is calculated over the series reported by Google, the resulting values will be percentage percent variations.

For example, if in a month the search for the term *employment* is done 50 times amid a total of 1000 searches for any other terms in the same month and the next, the percentage will be $(50/1000) \cdot 100 = 0.5\%$. If the same search is done 25 times the following month, the percentage will be $(25/1000) \cdot 100 = 0.25\%$. Google would release data 100 and 50, instead of 0.5 and 0.25, respectively. In any case, the monthly percentage change would be $(0.25 - 0.5)/0.5 = (50 - 100)/100 = -0.5 = -50\%$.

If the premise is that the proportion of searches by the term *employment* is positively correlated to the unemployment rate, the same transformation in monthly percentage variation should be applied in this case, since only then will the comparison with the Google series be valid. In the previous example, the search volume decreased by $0.5 - 0.25 = 0.25$ percentage points. However, such information would not be available. It would only be possible to extract the percentage variation of the percentages, that is, -50%.

Assuming that the unemployment rate measured by the PNAD or PME was 8% and 5% in the same two months, it would not be valid to compare the fall of $8 - 5 = 3$ percentage points with that of 50% calculated for the series of searches on Google. The correct would be to compare with the percentage change, that is, $ (5 - 8) / 8 = -0.375 = -37.5%$. In this case, the fall of 3 percentage points is comparable to that of 0.25 percentage points, but this last data could not be extracted from Google.

Methodology

Results

Concluding Remarks

References

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro. 2014. “Using Social Media to Measure Labor Market

Flows.” NBER Working Papers 20010. National Bureau of Economic Research, Inc.

Antenucci, Dolan, Christopher R?, Michael J. Cafarella, Matthew D. Shapiro, and Margaret C. Levenstein. 2013. “Ringtail: Feature Selection for Easier Nowcasting.”

Askitas, Nikos, and Klaus F. Zimmermann. 2009. “Google Econometrics and Unemployment Forecasting.” IZA Discussion Papers 4201. Institute for the Study of Labor (IZA).

Biorci, Grazia, Antonella Emina, Michelangelo Puliga, Lisa Sella, and Gianna Vivaldo. 2016. “Tweet-tales: moods of socio-economic crisis?” Working Papers 04/2016. IMT Institute for Advanced Studies Lucca.

Choi, Hyunyoung, and Hal Varian. 2009. “Predicting Initial Claims for Unemployment Benefits.”

———. 2012. “Predicting the Present with Google Trends.” *The Economic Record* 88 (s1): 2–9.

Daas, Piet J.H., and Marco J.H. Puts. 2014. “Social media sentiment and consumer confidence.” Statistics Paper Series 5. European Central Bank.

D’Amuri, Francesco, and Juri Marcucci. 2010. “Google it! Forecasting the US Unemployment Rate with a Google Job Search index.” Working Papers 2010.31. Fondazione Eni Enrico Mattei.

Llorente, Alejandro, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. “Social Media Fingerprints of Unemployment.” Edited by Yamir Moreno. *PLOS ONE* 10 (5). Public Library of Science (PLoS): e0128692.

Suhoy, Tanya. 2009. “Query Indices and a 2008 Downturn: Israeli Data.” Bank of Israel Research Department.