Improving prediction of unemployment statistics with Google trends: part 2

Pedro Ferreira

Abstract

A preliminary experiment using query terms searched on Google as predictors of unemployment was carried out at Eurostat. This is the follow-up of that preliminary work. In this report, a new approach to treat google trends data is taken. Dynamic factor model is used to extract a latent variable which is a good proxy for the unemployment dynamics. Prediction models for unemployment that make use of the estimated latent variable have performed better than the proposed approaches in previous works, in particular during a period where there was an abrupt change in the trend. Based on the results presented in this report, some future lines of research are proposed to further increase forecasts precision.

1 Introduction

This report is a follow up of the initial results obtained from applying Google trends time series to improve the prediction of unemployment statistics[1]. It was shown that using three query terms searched on Google regarding unemployment in France improved predictive performance as compared to an auto-regressive model. Those conclusions were in line with similar analysis for unemployment in the United States[2].

The analysis presented on this paper will analyze unemployment in Portugal for two reasons. First, being the author a portuguese native speaker, it was easier to identify potential good candidates for query terms and popular job search websites. Second, there was a significant change on the official unemployment figure from 2013 onwards. Unemployment have been dropping since 2013 mainly due to an increasing number of unemployed persons who emigrate and also a significant number of discourage persons who stopped looking for a job (and therefore considered as inactive). It would be interesting to test the prediction power of the model during an abrupt change in the trend.

For the moment only non-seasonal data will be taken into account. Seasonality will be dealt in the follow-up of this work.

This paper goes one step forward as compared to the previous analysis[1, 2] since it will make use of dynamic factor models to derive one latent variable from the google trends time series which is believe to represent the driving force behind the query terms searched.

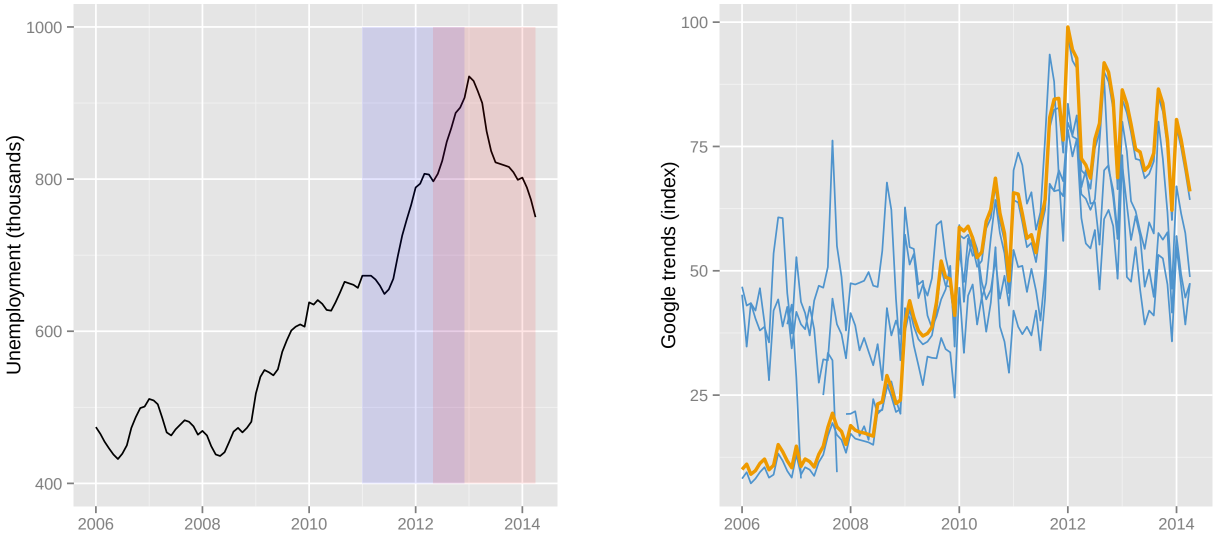
2 Available data

Official data

Monthly number of unemployed persons, non-seasonal adjusted, from January 2006 to April 2014 (source: Eurostat).

Trends data

Indexed data for 4 terms regarding unemployment (source: Google trends):



(a) Unemployment in thousands. Two time periods, from January 2011 to December 2012 (shaded in blue) and from May 2012 to April 2014 (shaded in red), will be analyzed.

(b) Google trends and their estimated common factor. blue lines represent the four query terms considered ("desemprego", "net emprego", "ofertas emprego" and "subsidio desemprego"), while the orange line represents the common factor of those series estimated by a dynamic factor model.

Fig. 2.1: Available data.

1. desemprego;
2. net empregos (popular portuguese website for searching job offers);
3. ofertas emprego;
4. subsidio desemprego.

As shown in Figure 2.1(a), the number of unemployed persons in Portugal was somehow stable from 2006 to mid 2008, became twice as big on the period mid 2008 until 2012 and since beginning 2013 there has been a steady drop. Two periods will be considered for the analysis of the performance of the prediction models:

1. Period 1: from January 2011 to December 2012, which is a period where the positive trend is stable (shaded in blue);
2. Period 2: from May 2012 to April 2014, which covers the change in the trend recorded in the beginning of 2013 (shaded in red).

Figure 2.1(b) shows the 4 google trends series (in blue) and the common factor that these series share (in orange). Details on how to derive the common trend will be explained in the next sections.

3 Model

Four models will be analyzed. Two of them were presented in previous works[1, 2]. The third model will make use of a common factor to the four google trends series that was estimated by a dynamic factor model. This common factor will then be used as a regressor. The fourth model is a simple extension of the third, where lagged values on the common factor will be taken into account as well.

3.1 Base line-type models

Model 1: Base line model

$$y_t = \beta_0 + \beta_1 y_{t-1} + a_t$$

where y_t is the unemployment at period t and a_t is white noise.

Model 2: Base line adjusted

$$y_t = \beta_0 + \beta_1 y_{t-1} + \sum_i \phi_i x_t^i + a_t$$

where y_t is the unemployment at period t, x_t^i is the i-th index data from Google trends and a_t is white noise.

3.2 Dynamic factor analysis

Dynamic factor model is able to describe a multivariate time series with a smaller number of uncorrelated factors. Dynamic factors can be identified with some latent driving force of the whole multivariate process.

A dynamic factor model was applied to the 4 time series from google trends and the first common factor was considered to be an estimate of a latent variable that, in principle, should reflect unemployment. In fact, “being unemployed” is the driving force to search on google words related to finding a job and / or information about the unemployment.

The model applied was the following:

$$\begin{aligned} u_t &= u_{t-1} + w_t \\ \mathbf{o}_t &= \mathbf{Z} u_t + \mathbf{v}_t \end{aligned}$$

where $\mathbf{o}_t$ is a vector of observed time series (in this case, the four Google trends indices) which are considered to be a function of one unobserved trend (u_t) and a factor loadings matrix ($\mathbf{Z}$), $w_t \sim N(0, \sigma_w)$ and $\mathbf{v}_t \sim MVN(0, \mathbf{R})$.

The latent variable u_t was rescaled to have the same mean and range has y_t for presentation purposes, e.g. case if one needs to present in the same plot both the common factor and unemployment:

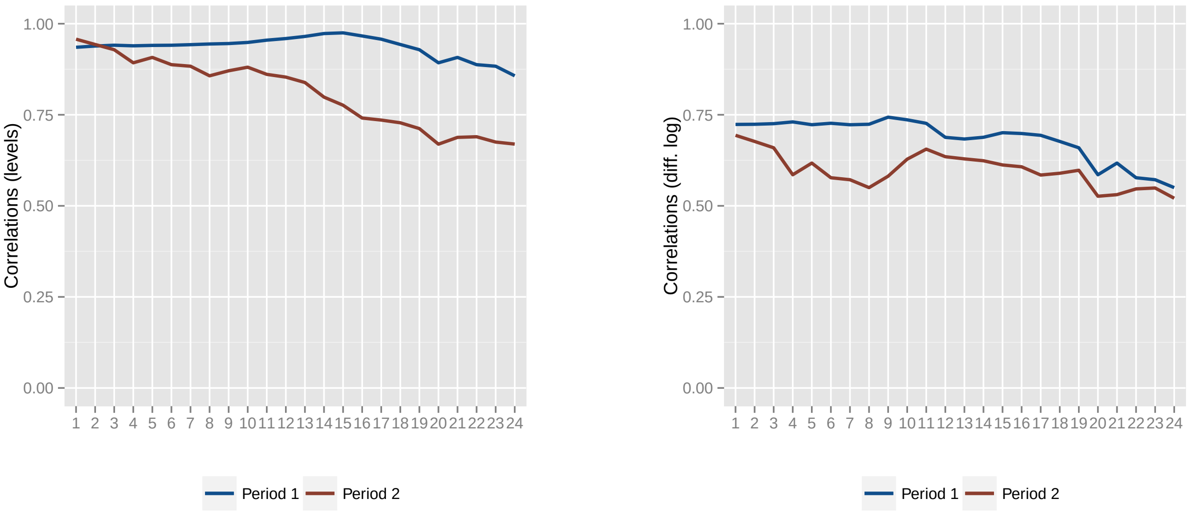
$$\tilde{u}_t = \bar{y}_t + \frac{\max(y_t) - \min(y_t)}{\max(u_t) - \min(u_t)} u_t$$

There is a strong correlation between $\tilde{u}_t$ and y_t, i.e. between the estimated latent variable and unemployment, as shown in Figure 3.1. This correlation is strong and significant both on the level of unemployment and on their differences of the log.

Model 3: DFM model

$$y_t = \beta_0 + \beta_1 y_{t-1} + \phi \tilde{u}_t + a_t$$

where y_t is the unemployment at period t, $\tilde{u}_t$ is the re-scaled common factor and a_t is white noise.



(a) With unemployment $(cor(y_t, \tilde{u}_t))$

(b) Differences of the log: $cor\,(\Delta\log(y_t), \Delta\log(\tilde{u}_t))$

Fig. 3.1: Correlations between unemployment and the common factor for the 60 observations used in the pseudo-out-of-sample forecast.

Model 4: DFM model + ADL model

There seems to be some correlation between lagged values of the latent variable $\tilde{u}_t$ and y_t. As such, a variation of Model 3 was also considered where lagged values for the common factor was included:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \phi_0 \tilde{u}_t + \phi_1 \tilde{u}_{t-1} + a_t$$

4 Results

A pseudo-out-of-sample forecast was conducted for the two different time periods described before, using a rolling window of 60 observations.

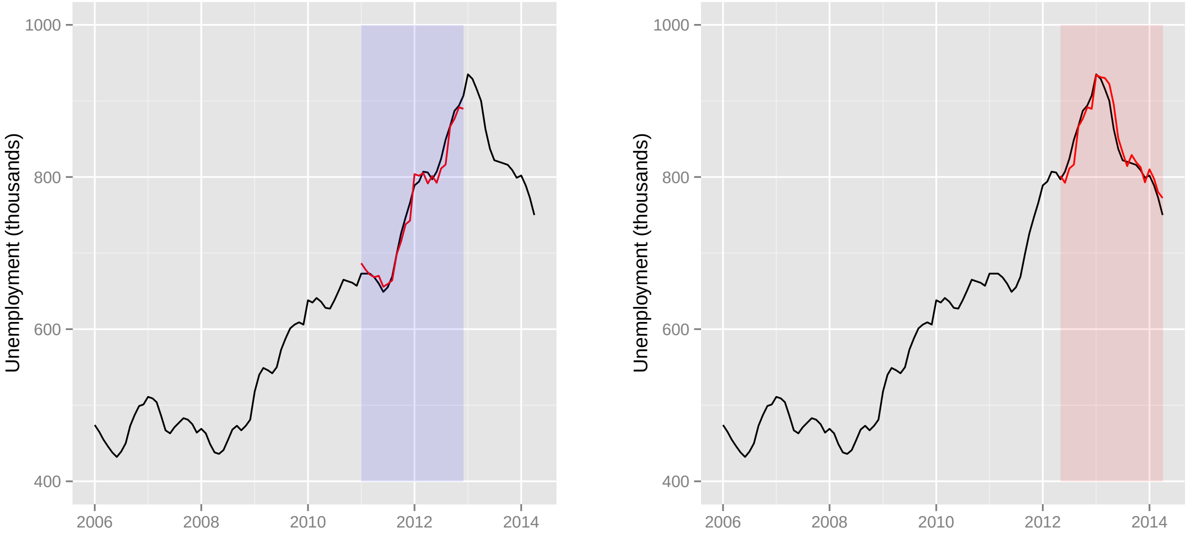
The Mean Absolute Forecast Percentage Error (MAFPE) was computed for each one of the models described in the previous section. Results are presented in Table 1.

Model 1, which didn't use google trends data, was the one that performed the worst. This result is in line with previous analysis[1, 2]: google trends improves the prediction of unemployment. Model 3 and 4 which have used as a regressor the common factor of the google trends estimated by a dynamic factor model performed better than Model 2. This is particularly true during period two where a change in the trend was recorded. Based on this results **it seems that estimating latent variables from google trends should be considered instead of using trends directly**.

Figure 4.1 shows the 1-step ahead forecasts produced by Model 4 for the two different time periods. Each point represented by the red line is an out-of-sample forecast based on the previous 60 observations.

MAFPE	Model 1	Model 2	Model 3	Model 4
Period 1	1.372	1.275	**1.234**	1.315
Period 2	1.670	1.526	1.358	**1.329**

Tab. 1: Mean Absolute Forecast Percentage Error (MAFPE)



(a) Period 1: January 2011 to December 2012

(b) Period 2: May 2012 to April 2014

Fig. 4.1: Unemployment (y_{t+1}) and 1-step ahead forecasts $(\hat{y}_{t+1|t})$ using Model 4

5 Conclusions and future research

The conclusions that can be taken from this analysis are scarce as compared to the lines of future research that they've created. Basically, this reports re-enforces the conclusion that Google trends can help the prediction of unemployment and draw a new conclusion compared with the previous works: **searching for latent variables tend to increase the predictive power of as compared to using the trends without any treatment**.

Some examples of future research areas that can be carried out in the follow up of this analysis:

- Scale this analysis to an european aggregate, e.g., euro area;
- The use of Partial Least Squares should be taken into account as well. PLS is used to find the fundamental relations between two matrices, i.e. a latent variable approach to modeling the covariance structures if these two spaces. As such, it this analysis is applied to a big country like Germany, one could modeled a matrix of dependent variables, e.g. number of unemployed persons per region, with a matrix of independent variables, e.g. Google trends data;
- Take in to account different temporal desegregation. Google trends are available weekly but for this report monthly averages were taken instead. It is possible that by averaging weekly data worsened the extraction of latent variables, so deriving latent variables at week level should also be looked at;
- Due to lack of time, seasonality was not even considered. This analysis used non-seasonal data only. Seasonality is an important aspect of unemployment and several google trends series showed significant seasonal patterns, which should be taken into consideration in the future;
- Using benchmark techniques to derive a weekly coincident indicator of unemployment. Depending on the quality of seasonal adjustments and on the extraction of latent variables on weekly data, it could be possible to derive an early and with an higher frequency estimate of unemployment.

References

[1] Perduca, V. (2014). Improving prediction of unemployment statistics with Google trends: preliminary experiments. Eurostat.

[2] Choi, H. & Varian, H. (2011). Predicting the Present with Google Trends. Web Document: http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf

[3] Holmes, E.E., Ward, E.J. & Scheuerell, M.D. (2014). Analysis of multivariate time- series using the MARSS package. User guide: http://cran.r-project.org/web/packages/MARSS/vignettes/UserGuide.pdf