Can Google data help predict French youth unemployment?☆

Y. Fondeur a, F. Karamé a,b,⁎

a Centre d'Etudes de l'Emploi, France
b GAINS-TEPP (Le Mans University and FR n°3435, CNRS), DYNARE Team (CEPREMAP), France

Abstract

According to the growing “Google econometrics” literature, Google queries may help predict economic activity. The aim of our paper is to test whether these data can enhance predictions of youth unemployment in France. Because we have weekly series on web search queries and monthly series on unemployment for 15- to 24-year olds, we use the unobserved components approach in order to exploit all available information. Our model is estimated with a modified version of the Kalman filter, taking into account the twofold issue of non-stationarity and multiple frequencies in our data. We find that including Google data improves unemployment predictions relative to a competing model that does not employ search data queries.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Economic time-series are usually published with a significant delay and may still be revised afterwards. Unemployment data are obviously subject to such delays. In France, these data are published on a monthly basis from an administrative source, the claimant count (“Demandeurs d’Emploi en Fin de Mois”, DEFM hereafter) provided by “Pôle Emploi”, the national employment agency responsible for unemployment compensation and jobseeker assistance. These data are available on the website of the French Ministry of Labor after the 24th day of the following month: for example, the claimant count at the end of November is available online on December 24th.

In light of such publication delays, there remains a strong and abiding demand for the real-time estimation of unemployment dynamics. Thus, there is a growing literature on nowcasting, i.e., predicting the present (see, for instance, Castle et al., 2009; Giannone et al., 2008; Schumacher and Breitung, 2008, among many others). In the search for new types of data, real-time data from the Internet may be an important tool for nowcasting. A growing portion of all economic activity is conducted via the Internet at one point or another, leaving traces on the Internet or in the information systems of the various web actors. Google Inc. publishes real time aggregated data for the search volumes of user-entered keywords. Choi and Varian (2009a) show that models including relevant Google data tend to outperform (in terms of predictions) models ignoring such data. The gain can even be quite substantial, in some cases. Many very recent studies follow this approach on various topics.

The aim of our paper is to apply a similar approach to an investigation of French unemployment. Google query data for well-chosen keywords may be connected with the online job search behaviors of employed or unemployed people and can then provide relevant real-time information regarding the current labor market situation. Taking these data into account may produce better forecasts and/or nowcasts.

Additionally, we use an unobserved variables approach. This methodology allows us to disentangle the components of the variables in order to identify potential relations between some of these components (the evolution of their respective trends, for instance). In this paper, we...
model the unemployment slope as a function of the Google data slope for the same month. This method could prove crucial for anyone wanting to exploit Google real-time data as a leading or coincident indicator able to anticipate turning points in unemployment. Our model is estimated using a modified version of the Kalman filter, taking into account the twofold issue of non-stationarity and multiple frequencies in our data.

The paper proceeds in four parts. First, we describe our dataset and discuss the keywords. Second, we present the estimation methodology and the model. Third, we present the out-of-sample forecasting exercise and the tools we need to evaluate it. The last section presents and discusses the results.

2. The data

2.1. Google data and the literature

Google Inc. publishes real-time aggregated data on the search volume for keywords “that receive a significant amount of traffic” (Google does not give any information regarding this threshold). Weekly time series starting in 2004 are available at the end of the week on Google Insights for Search (www.google.com/insights/search). Based on a portion of Google web queries, data reflect the number of searches that have been conducted for a particular term relative to the total number of searches conducted over time. Query results are scaled to the maximum over the selected period.

The construction of a Google index raises the question of the representativeness of the series. Using Google data is particularly relevant for France because its search engine centralizes almost all queries made in the country, with a stable market share of 90% for several years (as opposed to a smaller but growing market share in the US that increased from 60% in 2006 to 70% in 2010).

Since 2009, a handful of papers have used these data in various fields. The seminal work of Choi and Varian (2009a) gives examples of nowcasting for car and home sales in the US and for travel to Hong Kong. Several papers use Google search data for influenza virus surveillance (Doornik, 2009; Ginsberg et al., 2009). Suhoy (2009) examines the ability of Google queries to predict the 2008 downturn in real time in Israel. Kholodilin et al. (2010) apply a factor model on a large set of Google queries to extract principal components that improve nowcasts of US private consumption (see also Schmidt and Vosen, 2009) for a comparison between Google indicators and the usual survey-based indicators for the US, and see Suhoy (2010) on consumption in Israel). Kulkarni et al. (2009) aim to develop a leading indicator able to anticipate turning points in unemployment. Our choice strategy and the econometric approach.

We tried several keywords related to the French labor market, and we consider “EMPLOI” (which means “job” but also “employment” in French) and strong correlations between keyword searches and the number of jobs.

Fig. 1. Monthly data.
to be the best choice for our purpose. Google activity associated with this term is expected to be directly connected with job searches because it is the simplest way to find websites where jobs are posted. It may also reflect a more general concern by firms regarding the labor market situation. Fig. 1(a) shows the monthly series for the Google index by week.

2.3. Unemployment data

We use raw data (not seasonally adjusted) from the claimant count (“Demandeurs d’Emploi en Fin de Mois”, categories A, B and C for continental France). These data are provided by Pôle Emploi, the national employment agency in charge of unemployment compensation and jobseeker assistance. This variable describes the inventory of unemployed people at the end of each month.

Because Internet use is most likely affected by a generation bias, we disaggregate the DEFM series by age, presuming that a potential relation with Google data may be stronger for young claimants. This point echoes the selection bias issue raised by D’Amuri (2009). According to him, the Google index is not fully representative because not everyone uses the Internet, particularly not as a job-search tool. This criticism applies to the other studies on this issue. We suppose it could most likely be attenuated in our case because we focus on the young claimant count. Fig. 1(b) shows the DEFM for the age group 15–24 years on a monthly basis from January 2004 to July 2011.

2.4. Characterization of the data

Our data are characterized by at least four important features:

• There is a clear seasonal pattern in the data. We choose to work with raw instead of seasonally adjusted data because the Google index displays an obvious break from 2009 (see Fig. 2(b)). Taking into account this unstable seasonality compels us to choose a flexible representation of seasonality rather than using an automatic seasonal-adjustment procedure that may influence our results.
• The series are non-stationary.
• As highlighted by the common turning point at December 2008, the trends of the series seem to be strongly related.
• Data do not display the same frequency. In our case, the DEFM series is monthly, while the Google series is weekly. The current literature using Google data generally displays a limitation we want to surpass: the use of standard time series models does not allow researchers to address this multi-frequency issue. This limitation is the reason why the dataset is generally ‘poorly’ retained by the monthly (or even quarterly) frequency and using only a Google monthly series (by selecting one or two specific weeks or by averaging weeks over a month or a quarter) (see Choi and Varian, 2009a; D’Amuri, 2009 or Doornik, 2009, for instance). In our approach, we want to circumvent this limitation and use all of the available information. Ferrara et al. (2010) present an interesting review of the literature on these issues. They develop a robust non-parametric approach to nowcast GDP using mixed-frequency (i.e. monthly data to nowcast quarterly GDP) and ragged-edge data (at the end of the sample, some variables can be missing due to their publication delays).

As we can see, the first three points make reference to some unobserved components of our time series. Additionally, the Google index may contain an important noise element because it may also include search queries unrelated to the labor market. For this reason, we choose an unobserved-variables decomposition-based approach.

3. Our approach

3.1. The econometric methodology

Unobserved components models are generally treated with the Kalman filter and estimated with the maximum likelihood, which allows us to restore unobservable components and estimate unknown parameters, even in the presence of missing data. This choice is interesting for several reasons.

First, the identification of the signal components is based on relative general specification choices for components and not on a priori values for traditional nonparametric filters (such as HP, Bandpass, etc.). This approach is expected to be more congruent with the data. Furthermore, unlike nonparametric filters, it allows forecasting because it is based on a model. Second, as the decomposition is based on a maximum likelihood estimation, it allows us to provide standard errors for unknown parameters, confidence bands for unobserved variables and parameter testing. Third, because the data are non-stationary, the use of the standard Kalman filter is not the best choice because it is based on an incompatible initial condition of stationary distributions for unobserved

---

1 Data are available at http://www.travail-emploi-sante.gouv.fr/etudes-recherche-statistiques-de,76/statistiques,78/chomage,79/.
components. Previous articles used the Kalman filter without taking this point into account (see, for instance, Clark, 1987, 1989; Kim and Nelson, 1999; Koopman and Franses, 2002; Kuttner, 1994, among many others) or circumvented the problem by using stationary variables or rewriting the model in terms of stationary variables (see, for instance, Doz and Lenglart, 1999; Stock and Watson, 1991, among many others). To obtain efficient estimations for parameters and to evaluate for state variables, we use the diffuse Kalman filter proposed by Durbin and Koopman (2001) and Koopman and Durbin (2003). This filter provides a specific treatment due to its diffuse initial conditions. Once the effect of initial conditions vanishes, the filter becomes a standard Kalman filter.

The solution to the multi-frequency issue consists of considering monthly data as partially observed weekly data. Because the measure-

yi \approx \text{walks:}

\text{of matrices (Koopman and Durbin, 2000; Durbin and Koopman, 2001).}

This pattern will be particularly important for the last two years of the sample, as observed in Fig. 2(a) and (b). To obtain a parsimonious representation, we will restrict the variances that are naturally found close to zero and test these restrictions.

The DEFM series seem to display a stable seasonal pattern. We then retain the same representation:

\[ S_{2,i} = \sum_{j=1}^{S/2} \left\{ a_j \cos \left( \tau_j \frac{2\pi j m}{52} \right) + b_j \sin \left( \tau_j \frac{2\pi j m}{52} \right) \right\} \]

with constant weights \((a_j, b_j).\) This representation is less parsimonious but easier to handle (because the DEFM are partially observed in this model) than the standard stochastic seasonality representation.

3.2.2. The benchmark model

All stochastic trends are represented with a random walk with a time-varying drift

\[ T_{iz} = T_{iz-1} + d_{iz-1} + \epsilon_{iz}^{T} \]

\[ d_{iz} = \alpha_0 + \alpha_t d_{iz-1} + \epsilon_{iz}^{d} \]

with

\[ \epsilon_{iz}^{T} \approx N(0, \sigma_{iz}^2) \]

\[ \epsilon_{iz}^{d} \approx N(0, \alpha_{iz}^2) \]

To estimate these models, we built a complete Gauss library that contains the Kalman filter and smoother, both in their diffuse and traditional versions and including the univariate treatment of time-series (when the DEFM series seem to display a stable seasonal pattern. We then retain the same representation:)}

3.2.3. The bivariate model

In this section, we modify the benchmark model to take into account the Google effect. We test several specifications and finally modify the previous equations for DEFM trend and slope as follows. The trend is now

\[ T_{iz} = T_{iz-1} + d_{iz-1} + \epsilon_{iz}^{T} \]

\[ d_{iz} = \alpha_0 + \alpha_t d_{iz-1} + \epsilon_{iz}^{d} \]

\[ \epsilon_{iz}^{T} \approx N(0, \sigma_{iz}^2) \]

\[ \epsilon_{iz}^{d} \approx N(0, \alpha_{iz}^2) \]

The DEFM slope instantaneously depends on the Google slope, with parameter \(\alpha_t\) measuring the potential impact of Google, if any. The DEFM slope (and consequently its trend) now benefits from the real-time information of Google data, while in the benchmark model, that slope depends only on its own past. As a side effect, estimation will provide an evaluation for the claimant count on a weekly basis.

3.3. The estimation method

Both representations can be written in the linear state-space form with time-varying parameters. We have a twofold problem of unobserved variables and unknown parameters that can both be treated with the diffuse Kalman filter and maximum likelihood. Indeed, given a particular value of parameters, the filter is able to recursively provide (i) an evaluation of unobserved variables and (ii) the log-likelihood. The problem can then be solved in two steps. First, we maximize the likelihood provided by the filter with respect to unknown parameters. Second, we run the filter one more time, conditional to estimated parameters, to obtain filtered variables. A third step can be added based on the Kalman smoother. This provides a more stable and accurate evaluation of unobserved variables because – as opposed to the filtered components – it is no longer based on past and present sample information but on the information of the whole sample. Durbin and Koopman (2001) and Koopman and Durbin (2003) modify the Kalman smoother to include the treatment of diffuse initial conditions and of partially observed measurement variables. To estimate these models, we built a complete Gauss library that contains the Kalman filter and smoother, both in their diffuse and traditional versions and including the univariate treatment of time-series (when the DEFM series seem to display a stable seasonal pattern. We then retain the same representation:)}
information on observed variables at date $t$ is partially available). The smoother also provides evaluations for state and measure innovations. The estimation is realized by (quasi-) maximum likelihood (using a BFGS algorithm) and provides the standard specification tests on normalized one-step-ahead prediction residuals (autocorrelation, heteroskedasticity, normality).

4. Forecasting and nowcasting

Because the DEFM are made available with a one-month delay, we can use our bivariate model to forecast and even nowcast youth unemployment using real-time Google weekly data. In this section, we briefly describe the out-of-sample exercise and the tools for evaluating the benefit of using Google data for the DEFM predictions.

4.1. Description of the out-of-sample exercise

Instead of adding extra macroeconomic explanatory variables, this exercise will focus on Google information and then quantify the benefit of adding it in a dynamic model. We select the last $m$ observations of the sample. For each of these observations ($r = T - m, \ldots, T$), we re-estimate the models and calculate the forecasts for several horizons $h$: $y_{T+h|T}$.

The univariate model resulting from the exclusion of Google data is used as a benchmark. We then have to perform 1 prediction for Google.

Before the 24th of the current month, we perform a prediction called '2 weeks ahead': we first estimate our models at week 2 (with the Google information included). We then also have to perform 2 predictions for Google. As we are two weeks before, the DEFM observation for the previous month is not available.

The estimation is only available up to this current week and is used both for estimation and prediction. The nowcast simply consists in evaluating the DEFM for the end of the current month.

With such an approach, we will assess the predictive gain relative to the univariate model, i.e., the same model of monthly DEFM excluding Google data. It will also allow us to know whether the inclusion of new Google information each week produces a significant revision/improvement of the forecast of the current month or not.

4.2. Assessing the quality of the predictions

We calculate $\hat{e}_{T+h}$, the prediction error at horizon $h$:

$$\hat{e}_{T+h} = y_{T+h} - \hat{y}_{T+h|T}.$$ 

To assess the predictive performance of the models, we first investigate a systematic predictive bias by testing the significance of the prediction error at each horizon:

$$H_0 : E(\hat{e}_{T+h}) = 0.$$

This assessment can be performed by simply testing the nullity of the prediction error mean:

$$UB(h) = \frac{\hat{e}_h}{\sqrt{\hat{V}_w(\hat{e}_h)}}$$

with

$$\hat{e}_h = \frac{1}{m-h+1} \sum_{r=T-m}^{T-h+1} \hat{e}_{r+h}$$

and its long-term variance, estimated with the Newey and West (1987) estimator:

$$\hat{V}_w(\hat{e}_h) = \frac{1}{m-h+1} \sum_{r=T-m}^{T-h+1} \left(1 - \frac{k}{h+1}\right) \text{COV}(\hat{e}_{r+h}, \hat{e}_{r+h-k}).$$

To assess the relative predictive performances of the models, we build two simple performance indicators, the square prediction error and the absolute prediction error at horizon $h$:

$$\hat{e}^{SQ}_{T+h} = (y_{T+h} - \hat{y}_{T+h|T})^2$$

$$\hat{e}^{APE}_{T+h} = \left|\frac{y_{T+h} - \hat{y}_{T+h|T}}{y_{T+h}}\right|.$$ 

We then calculate the Root Mean Square Error (RMSE) and the Mean Absolute Predictive Error (MAPE):

$$\text{RMSE}(h) = \left(\frac{1}{m-h+1} \sum_{r=T-m}^{T-h+1} \hat{e}^{SQ}_{T+h}\right)^{1/2}$$

$$\text{MAPE}(h) = \frac{1}{m-h+1} \sum_{r=T-m}^{T-h+1} \hat{e}^{APE}_{T+h}.$$ 

These two indicators are homogeneous in the predicted variable. They could differ because they do not give the same weight to large errors.

Finally, we test whether the predictions produced by the two competing models $M_1$ and $M_2$ are significantly similar:

$$H_0 : E(d_{T+h}) = 0$$

with $d_{T+h}$ as a loss (whether squared or absolute) function based on the predictive error:

$$d_{T+h} = \begin{cases} \hat{e}^2_{T+h|M_1} - \hat{e}^2_{T+h|M_2} & \text{Squared error loss} \\ \left|\hat{e}_{T+h|M_1} - \hat{e}_{T+h|M_2}\right| & \text{Absolute error loss} \end{cases}$$

We use the Diebold and Mariano (1995) modified statistics $DM^*(h)$ proposed by Harvey et al. (1997) for small samples:

$$DM^*(h) = \frac{\bar{d}_h}{\gamma \hat{V}_w(\bar{d}_h)},$$

where $\hat{V}_w(\bar{d}_h)$ is the long-term variance of $\bar{d}_h$ and

$$\gamma = \frac{T+1}{T^2} \sum_{r=1}^{T-1} \left(1 + \frac{r}{T+1}\right)^{-1}. $$
Bold means significant at 5%.

Table 1 displays the likelihood contribution of estimated representations in three cases: (i) for the general representation of the seasonality (the 52 weights are random walks); (ii) for the most constrained representation of seasonality (all weights are constant); (iii) for the specification we finally retained (24 weights are random walks, and 28 are constant). The most constrained model is clearly rejected by the data. The retained model is largely accepted, which suggests the virtue of using a sophisticated representation to take into account the break in seasonality of the Google index.

Tables 2a, 2b and 2c display the estimation results for both the univariate and the bivariate models. The two models satisfy all specification tests at 5% (Table 2c). Estimation results are stable for the Google equation (Table 2a). The fluctuations of the estimated slopes are very close (Fig. 3(f)), which seem to indicate that the assumed relation between the Google index and the DEFM in the bivariate representation is quite natural and pertinent and does not affect the evaluation of the Google components. We used seven dummy variables ($\hat{\delta}_i, i = 1, \ldots, 7$) to avoid outliers, control for the potential volume effects of holidays or bridge days and reach normality in the Google equation.

In the DEFM equation, the results are also very stable (Table 2b and Fig. 3(b)). We observe a decrease in the trend and slope variances between the univariate and the bivariate models. The first column of Fig. 3(e) simultaneously displays the original DEFM series (not seasonally adjusted and partially observable on a weekly basis) and the trend component. Fig. 3(g) displays the seasonally adjusted DEFM series on a weekly basis and the trend component. The representation seems to be very efficient. The so-called Google impact (from the Google slope to the DEFM slope) appears significant at 5%.

Table 2c specifies the tests for prediction errors. The table displays p-values for specification tests of the model implemented on prediction errors. Bold means rejection of the null of good specification at 5%.
Fig. 3. (Smoothed) components for the Google data (bivariate model).
5.2. The predictions

We select the last 31 DEFM observations of the sample to implement the out-of-sample forecasting exercise (from January 2009 to July 2011).

Globally, there is no systematic bias in the prediction of our models (Table 3).

At least four points should be highlighted from the analysis of the performance indicators (Table 4). First, the model using the Google information outperforms the univariate model, which ignores that information whether we exclude information on the previous DEFM observation or not. As expected, we observe an important decrease in both RMSE and MAPE performance indicators once the previous DEFM observation is known. Second, while both positive, these two indicators do not always deliver a clear message about the relative performances of the two models. Third, the comparison of predictive horizons leads us to determine the “optimal” date to produce the forecast, which is one week ahead in our case; adding an extra Google observation is useless. Fourth, when calculating the maximum gain (as the percentage of the minimum indicator relative to the “best” univariate model), we observe that using Google data improves the quality of the prediction by up to nearly 27% for the studied period. Furthermore, the prediction accuracy is also greatly improved because the associated standard error (that is provided by the Kalman filter) decreases by 40% on average for the bivariate model (and by 49% at maximum).

Table 5 displays the modified Diebold and Mariano statistics for the significance of the compared predictions by the two competing models. Confirming the decrease in RMSE or MAPE resulting from the use of Google data, we notice that there is a statistically significant difference between the predictions from the univariate and the bivariate models at all horizons. Moreover, this difference can also be found significant for predictions made at different horizons for the same model and for the same horizon for different models. The ranking obtained from the Diebold and Mariano test for the five predictions is consistent with previous results. Our findings can be summed up as follows. First, as expected, it is always better to predict DEFM with the previous DEFM observation than without it, whether for the univariate or bivariate model. Second, predictions from the bivariate model after the 24th significantly outperform those from the univariate model before and after the 24th. However, the use of current Google information in the bivariate model (before the 24th) does not compensate for the absence of the previous DEFM observation in regard to the univariate representation used after the 24th. Third, there is no significant difference between the ‘one-week ahead’ predictions and the nowcast for the bivariate model: so the knowledge of current Google information does not help to nowcast unemployment.4

6. Conclusion

The aim of this paper was to test whether Google’s real-time information could enhance the predictions for the claimant count of the 15- to 24-year old demographic in France. To exploit all available information, we use an unobserved components approach. We propose a statistical model estimated with a modified version of the Kalman filter that takes into account the twofold issue of non-stationarity and multiple frequencies in our data. To use real-time information, we model the DEFM slope as a function of the current Google slope.

We conclude that Google data contribute to enhance predictions and nowcasts for the 15- to 24-year old unemployed population, in terms of both level and accuracy.

The same forecasting exercise has been carried out for the 25- to 49-year old group and the 50 and over population (the results are not reproduced here). The RMSE (calculated from an equivalent bivariate model relative to its univariate counterpart) are enhanced by 17.5% and 9.7%, respectively. This observation most likely illustrates the selection bias noted by D’Amuri (2009) in favor of the young regarding the Internet as a job search tool.

---

4 We specify a simple monthly model as another benchmark. The DEFM growth rate is modeled as an AR(5). The Fourier representation employed in the paper is used to capture the seasonality (this representation is much more congruent with the data than the specifications used by Cho and Varian for instance). We introduce the seasonally-adjusted Google index growth rate as an explanatory variable. We choose the third week of the current month because it provides the best fit and (slightly) the best forecasting indicators. We find that the Google impact is only significant at 10%. This approach almost displays the same behavior that the weekly bivariate model for the nowcasting exercise: no systematic bias, lesser RMSE and MAPE indicators but not statistically significantly different. Results (not reproduced here) are available upon request.
References


Kulkarni, R., Haynes, K., Stough, R., Paelinck, J.H.P., 2009. Forecasting housing prices with Google econometrics, mimeo, School of Public Policy, George Mason University, Fairfax, VA 22030.


