DOCUMENT DE RECHERCHE

EPEE

CENTRE D'ETUDES DES POLITIQUES ECONOMIQUES DE L'UNIVERSITE D'EVRY

Can Google Data Help Predict French Youth Unemployment?

Y. Fondeur and F. Karamé

12-03

www.univ-evry.fr/EPEE

Université d'Evry Val d'Essonne, 4 bd. F. Mitterrand, 91025 Evry CEDEX

Can Google Data Help Predict French Youth Unemployment?

Y. Fondeur† & F. Karamé‡

First version: December 2010

Second version: November 2011

Abstract

According to the rising "Google econometrics" literature, Google queries may help predict economic activity. The aim of our paper is to test if these data can enhance predictions for youth unemployment in France.
As we have on the one hand weekly series on web search queries and on the other hand monthly series on unemployment for the 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 issues of non-stationarity and multiple frequencies in our data.
We find that including Google data improves unemployment predictions relatively to a competing model without search data queries.

Keywords: Google econometrics, forecasting, nowcasting, unemployment, unobserved components, diffuse initialization, Kalman filter, univariate treatment of time series, smoothing, multivariate models.

JEL classification: C22, C51, E32, E37.

† Centre d'Etudes de l'Emploi.

‡ EPEE (university Evry Val d'Essonne), TEPP (FR CNRS n°3126), DYNARE Team (Cepremap) and Centre d'Etudes de l'Emploi.

Corresponding author: F. Karamé, Bur. 346, Bat IDF, 4 bd. F. Mitterrand, 91025 Evry CEDEX, FRANCE. Mailing address: frederic.karame@univ-evry.fr. The authors thank the participants of *EPEE* seminar (university of Evry Val d'Essonne), the *ETEPP* winter school (Aussois), the *ERUDITE* seminar (university Paris-Est Val de Marne) and the *Computational and Financial Econometrics* (CFE'2011, London) for their remarks and comments. We remain responsible for errors and omissions.

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, they are published on a monthly basis using an administrative source, the claimant count ("Demandeurs d'Emploi en Fin de Mois", DEFM hereafter) provided by "Pôle Emploi", the national employment agency in charge of unemployment compensation and jobseekers assistance. These data are available on the website of the French Ministry of Labor after the 24th of the following month: for example, the claimant count at the end of November is online on December, 24th.

Due to publication delay, providing real-time estimation of unemployment dynamics is a real stake. There is a growing literature about nowcasting, *i.e.* predicting the present (see for instance Giannone *et al.* 2008, Schumacher & Breitung, 2008, Castle *et al.* 2009 among many others). In the query for new kind of data, real-time data from Internet may be a precious tool for nowcasting. A growing part of the economic activity is going through Internet at one time or another, leaving traces on it or in the information systems of the Internet players. Google Inc. publishes in real time aggregated data for the volume of search for keywords. Choi & Varian (2009a) show that models including relevant Google data tend to outperform models ignoring them in terms of predictions. 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 this kind of approach to French unemployment. Google query data for well-chosen keywords may be connected with the online job search behaviors of employed or unemployed people and then bring relevant real-time information on labor market situation. Taking these data into account may produce better forecasts and/or nowcasts.

For that matter, we use unobserved variables approach. This methodology will allow us disentangling the components of the variables in order to identify potential relations between some of these components (the evolutions of their respective trends for instance). In this paper, we model the unemployment slope as a function of the Google data slope of the same month. This point could reveal crucial for anyone willing 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 issues of non-stationarity and multiple frequencies in our data.

The paper falls into 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. Last section presents and discusses the results.

1 The data

1.1 Google data and the literature

Google Inc. publishes in real time aggregated data on search volume for keywords "that receive a significant amount of traffic" (Google does not give any information about 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 are reflecting how many searches have been done for a particular term relatively to the total number of searches done over time. Query results are scaled by 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 90 % market share for several years (*vs* a lower but growing market share from 60 % in 2006 to 70 % in 2010 in the US).

Since 2009, a handful of papers have used these data in various fields. The seminal work of Choi & Varian (2009a) give 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 (Ginsberg *et al.* 2009, Doornik 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 & Vosen, 2009 for a comparison between Google indicators and the usual survey-based indicators for the US and Suhoy, 2010 for the consumption in Israel). Kulkarni *et al.* (2009) aim at developing a leading indicator of US housing prices. Dealing with the labor market issue, Askitas & Zimmermann (2009) found strong correlations between keywords searches and the unemployment rate in Germany (see also Choi & Varian, 2009b for the US, Suhoy, 2009 for Israel, D'Amuri, 2009 for Italy and D'Amuri & Marcucci, 2009 for the US).

We intend to use Google search data to improve French unemployment forecasts and/or nowcasts. But, beside the fact we focus on the French case, our approach will differ in two ways: the keyword choice strategy and the econometric approach.

1.2 The choice of keywords

The choice of keywords is of course crucial for the study. It thus requires some discussion.

Many studies evoked previously use a large bunch of Google queries. In order to retain pertinent and tractable information, their dimensionality is reduced by extracting their principal components which enter as exogenous variables in some (standard) time-series models such as ARMAX for instance (see Suhoy 2009, 2010, Kholodilin *et al.,* 2010, Schmidt & Vosen, 2009). Choi & Varian (2009b) use two indicators based on Google Trends categories "*jobs*" and "*welfare & unemployment*". Askitas & Zimmermann (2009) use four groups from one to eight keywords with boolean operator "OR". The first group is composed of two keywords related to the German federal employment agency (*"Arbeitsamt"* OR *"Arbeitagentur"*). This group is expected to be connected with people having contacted or being in the process of contacting the employment agency. The second group is simply composed of a single keyword meaning "unemployment rate" (*"Arbeitlosenquote"*). The third group is composed of two keywords relative to HR consulting (*"Personalberater"* OR *"Personalberatung"*). It is expected to proxy high-skilled workers reacting for fears of layoffs and companies preparing layoffs or reorganizations. The last group is composed of eight keywords corresponding to the most popular job boards in Germany (*"Stepstone"* OR *"Jobworld"* OR *"Jobscout"* OR *"Meinestadt"* OR *"meine Stadt"* OR *"Monster Jobs"* OR *"Monster.de"* OR *"Jobboerse"*). It is expected to capture job searching activities. Variables resulting from these four groups are then used as regressors in a monthly model aiming at forecasting unemployment rate. For their study on US unemployment, D'Amuri & Marcucci (2009) simply use the keyword "*jobs*".

We tried several keywords relative to French labor market. We thus consider *"EMPLOI"* (which means *"job"* but also *"employment"* in French) to be the best choice for our purpose. Google activity along this term is expected to be directly connected with job searches, as it is the simplest way to find websites where jobs are posted. It may also reflect a more general concern of firms about labor market situation. Figure (1-a) shows the monthly series for the Google index by week.

1.3 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, DEFM hereafter). These data are provided by *Pôle Emploi*, the national employment agency in charge of unemployment compensation and jobseekers assistance. This variable describes the unemployed people inventory at the end of each month.

As Internet use is probably affected by a generation bias, we disaggregate DEFM series by age, presuming that a potential relation with Google data may be stronger for young claimants. This point makes an echo to the issue of the selection bias evoked by D'Amuri (2009). According to him, the Google index suffers from a lack of representativeness since not everybody uses Internet, particularly as a job-search tool. This criticism applies to the other studies on this issue. We guess it could probably be attenuated in our case since we focus on young claimant count. Figure (1-b) shows the DEFM for the 15-24 years on a monthly basis from January 2004 to July 2011[1].

1.4 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 figure 2-b). Taking into account this instable seasonality induces us to choose a flexible representation of seasonality instead of using an automatic seasonal-adjustment procedure that may influence our results.
- The series are non-stationary.
- As highlighted by their common turning point in December 2008, it can be interesting to find a relation between the trends of the series that 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 mainly displays a limitation we want to surpass: the use of standard time series models does not allow dealing with this multi-frequency issue. This is the reason why the dataset is generally 'impoverished' by retaining the monthly (or even quarterly) frequency and using only a Google monthly series (by selecting one or two specific

[1] Data are available at http://www.travail-emploi-sante.gouv.fr/etudes-recherche-statistiques-de,76/statistiques,78/chomage,79/.

weeks, or by averaging weeks over a month or a quarter) (see Choi & Varian 2009a, Doornik 2009 or D'Amuri 2009 for instance). In our approach, we want to circumvent this limitation and use all the available information. This choice implies the recourse to a specific econometric methodology.

As we can see, the first three points make reference to some unobserved components of our time series. Besides, the Google index may contain an important noise since it may also include search queries unrelated to the labor market. This is the reason why we choose an unobserved-variables decomposition-based model.

2 Our approach

2.1 The econometric methodology

Unobserved components models are generally treated with the Kalman filter and estimated with the maximum likelihood, which allows restoring unobservable components and 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 non parametric filters (like HP, Bandpass, …). This approach is expected to be more congruent with the data. Furthermore, unlike nonparametric filters, it allows forecasting as it is based on a model. Second, as the decomposition is based on a maximum likelihood estimation, it allows providing standard errors for unknown parameters, confidence bands for unobserved variables and parameter testing. Third, as the data are non-stationary, the use of the standard Kalman filter is not the best choice since it is based on an incompatible initial condition of stationary distributions for unobserved components. Previous articles used it without taking into account this point (see for instance Clark 1987, 1989, Kuttner 1994, Kim & Nelson 1999, Koopman & Franses 2002, among many others) or circumvented the problem by using stationary variables or rewriting the model in terms of stationary variables (see for instance Stock & Watson 1991, Doz & Lenglart 1999 among many others). In order to obtain efficient estimations for parameters and evaluation for state variables, we use the diffuse Kalman filter proposed by Durbin & Koopman (2001, 2003). It provides a specific treatment due to the diffuse initial conditions of the filter. Once the effect of initial conditions vanished, the filter becomes a standard Kalman filter.

The solution to the multi-frequency issue consists in considering monthly data as partially-observed weekly data[2]. In our framework, as the measurement variables are partially observed, we have to use the univariate treatment version of the diffuse Kalman filter. It allows evaluating the state vector by incorporating information from observables when available. Furthermore, it considerably speeds up the estimation of large models by manipulating scalars instead of matrices (Durbin & Koopman 2000, 2001)[3].

2.2 The models

2.2.1 The common framework

We consider a bivariate approach with weekly and monthly data. We define y_t as a 2×1 vector of partially-observed variables:

$$y_t = ln\begin{pmatrix} Google_t \\ DEFM_t \end{pmatrix}$$

$Google_t$ stands for the weekly Google index. $DEFM_t$ is the claimant count for the 15-24 year-old. We use January 2004-August 2011 weekly series (see figure 2-a). The DEFM series then stops in July.

We decompose each element of y into respectively a trend, a seasonal and an irregular component:

$$y_{i,t} = T_{i,t} + S_{i,t} + \varepsilon_t^{y_i} \qquad i = 1,2, \qquad t = 1,...T$$

with:

$$\varepsilon_t^{y_i} \approx N(0, \sigma_{y_i}^2) \qquad i = 1,2$$

Irregular components are assumed to be independent.

For Google data, seasonality is modelled with a Fourier decomposition series:

$$S_{1,t} = \sum_{j=1}^{S/2} \left\{ a_{j,t}\, cos\left(\tau_t \frac{2j\pi}{S}\right) + b_{j,t}\, sin\left(\tau_t \frac{2j\pi}{S}\right) \right\}$$

---

[2] The monthly DEFM observation corresponds to the last week of the month.

[3] For the US, Mariano & Murazawa (2003) deal with this issue to build a coincident index of business cycle based on both monthly and quarterly stationary data. See also Cornec (2006) and Cornec & Deperraz (2006) for a similar approach for France.

with S the seasonality period (here 52.25 weeks) and τ_t the number of weeks elapsed from the beginning of the current year. This representation, inspired from Dordonnat *et al.* (2008), is flexible and allows capturing breaks in the seasonality since weights $(a_{j,t}, b_{j,t})$ are random walks:

$$\begin{cases} a_{j,t} = a_{j,t-1} + \varepsilon_t^{a_j} & \varepsilon_t^{a_j} \approx N(0, \sigma_{a_j}^2) \\ b_{j,t} = b_{j,t-1} + \varepsilon_t^{b_j} & \varepsilon_t^{b_j} \approx N(0, \sigma_{b_j}^2) \end{cases} \qquad j = 1, \ldots \frac{S}{2}$$

This pattern will be particularly important for the last two years of the sample, as observed in figure (1-b). In order 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:

$$S_{2,t} = \sum_{j=1}^{S/2} \left\{ a_j \cos\left(\tau_t \frac{2j\pi}{S} \right) + b_j \sin\left(\tau_t \frac{2j\pi}{S} \right) \right\} \qquad \text{with } S = 12, \text{ if the series is observable}$$

$$S_{2,t} = 0 \qquad \text{otherwise.}$$

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

2.2.2 The benchmark model

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

$$\begin{cases} T_{i,t} = T_{i,t-1} + d_{i,t-1} + \varepsilon_t^{T_i} & \varepsilon_t^{T_i} \approx N(0, \sigma_{T_i}^2) \\ d_{i,t} = d_{i,t-1} + \varepsilon_t^{d_i} & \varepsilon_t^{d_i} \approx N(0, \sigma_{d_i}^2) \end{cases} \qquad i = 1,2$$

which is usual in the literature. This representation amounts to estimate the Google and DEFM series simultaneously but independently. It will constitute a benchmark as regards the following bivariate representation including the Google effect.

2.2.3 The bivariate model

In this section, we modify the benchmark model in order in take into account the Google effect. We tested several specifications and finally modify the previous equations for DEFM trend and slope as follows; the trend is now:

$$T_{2,t} = T_{2,t-1} + d_{2,t} + \varepsilon_t^{T_2} \qquad \varepsilon_t^{T_2} \approx N(0, \sigma_{T_2}^2)$$

with:

$$d_{2,t} = \alpha_0 + \alpha_1 d_{1,t} + \varepsilon_t^{d_2} \qquad \varepsilon_t^{d_2} \approx N(0, \sigma_{d_2}^2)$$

The DEFM slope instantaneously depends on the Google slope, with parameter α_1 measuring the Google potential impact if any. The DEFM slope (and consequently its trend) now benefits from the real-time information of Google data, while in the benchmark model, it only depends on its own past. As a side-effect, estimation will provide an evaluation for the claimant count on a weekly basis.

2.3 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 be both treated with the diffuse Kalman filter and maximum likelihood. Indeed, conditionally on 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 conditionally to estimated parameters to obtain filtered variables. A third step can be added based on the Kalman smoother. It provides a more stable and accurate evaluation of unobserved variables because contrarily to filtered components, it is no more based on past and present sample information but on the whole sample information. Durbin & Koopman (2001, 2003) modify the Kalman smoother in order 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 information on observed variables at date *t* is partially available or not). The smoother also provides evaluations for state and measure innovations. 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).

3 Forecasting and nowcasting

As 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 gain of using Google data for the DEFM predictions.

3.1 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 gain of adding it in a dynamic model. We select the last m observations of the sample. For each of these observations ($\tau = T–m, \ldots T$), we re-estimate the models and calculate the forecasts for several horizons h: $\hat{y}_{\tau+h|\tau}$.

The univariate model resulting from the exclusion of Google data is used as a benchmark. The timing is as follows. We first re-estimate the model and calculate the forecast for the end of the month before and after the 24th of the current month, *i.e.* without and with the DEFM observation of the previous month.

For the representation including Google data, namely the bivariate model, Google information is available instantaneously each week of the month. We then revise the prediction for the end of the month by adding this new Google observation each week. Three forecasts are produced along the following timing. Before the 24th of the current month, we perform a prediction called '*2 weeks ahead*', *i.e.* without the DEFM observation of the previous month but with the Google current observation. After the 24th of the current month, our information set is extended to the DEFM observation from the previous month (that is now available) and to two Google observations: it conducts to the revision of the previous prediction on DEFM that we call '*1 week ahead'* and '*nowcast'*. The procedure is repeated until the last week of our sample.

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

3.2 Assessing the quality of the predictions

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

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

In order to assess the predictive performance of the models, we first question a systematic predictive bias by testing the significance of the predictive error at each horizon:

$$H_0 : E(e_{\tau+h}) = 0$$

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

$$UB(h) = \frac{\bar{\hat{e}}_h}{\sqrt{V_{as}(\bar{\hat{e}}_h)}}$$

with

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

and its long-term variance estimated with the Newey & West (1989) estimator:

$$\hat{V}_{as}(\bar{\hat{e}}_h) = \frac{1}{m-h+1} V(\hat{e}_{\tau+h}) + \frac{2}{m-h+1} \sum_{k=1}^{l} \left(1 - \frac{k}{l+1}\right) \cdot COV(\hat{e}_{\tau+h}, \hat{e}_{\tau+h-k})$$

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

$$\hat{e}^{SQ}_{\tau+h} = (y_{\tau+h} - \hat{y}_{\tau+h|\tau})^2$$
$$\hat{e}^{APE}_{\tau+h} = \left| \frac{y_{\tau+h} - \hat{y}_{\tau+h|\tau}}{y_{\tau+h}} \right|$$

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

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

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

At last, we test if the predictions produced by two competing models M_1 and M_2 are significantly the same:

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

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

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

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

$$DM^*(h) = \frac{\bar{\hat{d}}_h}{\sqrt{\gamma \cdot V_{as}(\bar{\hat{d}}_h)}}$$

where $V_{as}(\bar{\hat{d}}_h)$ is the long-term variance of $\bar{\hat{d}}_h$ and $\gamma = \frac{T+1-2T^{1/4}+T^{1/4}(T^{1/4}-1)T^{-1}}{T}$.

4 The results

4.1 Estimation

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 shows the interest to use a sophisticated representation to take into account the break in seasonality of the Google index.

Table 2 displays the estimation results for both the univariate and the bivariate models. The two models satisfy all specification tests at 5% (table 2-c). Estimation results are stable for the Google equation (table 2-a). The fluctuations of the estimated slopes are very close (figure 3-f), which seems to indicate that the assumed relation between Google index and DEFM in the bivariate representation is quite natural and pertinent, and does not affect the evaluation of Google components. We had seven dummy variables ($\delta_i, i=1,...7$) in order 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, results are also very stable (table 2-b and figure 3-b). We observe a decrease in the trend and slope variances between the univariate and the bivariate models. The first column of figure (3-e) simultaneously displays the original DEFM series (not seasonally adjusted and partially observable on a weekly basis) and the trend component. Figure (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%.

4.2 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 Google information outperforms the univariate model which ignores it, whether we dispose information on the previous DEFM observation or not. As expected, we observe an important decrease of both RMSE and MAPE performance indicators once the previous DEFM observation is known. Second, while 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 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 relatively to the "best" univariate model), we observe that using Google data improves the quality of the prediction up to nearly 27% for the studied period. Furthermore, the prediction accuracy is also greatly improved since the associated standard-error (that is provided by the Kalman filter) decreases by 40% in average for the bivariate model (and by 49% at maximum).

Table 5 displays the modified Diebold & Mariano statistics for the significance of predictions between 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. Besides, 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 & Mariano test for the five predictions is coherent 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 that without it, whether for the univariate or bivariate models. 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 the absence of the previous DEFM observation as regards the univariate representation used after

the 24th. Third, there is no statistical difference between the 'one-week ahead' predictions and the nowcast for the bivariate model.

Conclusion

The aim of this paper was to test if Google real-time information could enhance the predictions for the claimant count of 15-24 years in France. In order 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 issues of non-stationarity and multiple frequencies in our data. In order 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 15-24 years unemployed people, both in level and accuracy.

The same forecasting exercise has been carried out for the 25-49 year-old and the 50 and over (the results are not reproduced here). The RMSE (calculated from an equivalent bivariate model relatively to their univariate counterpart) are respectively enhanced by 17.5% and 9.7%. This observation probably illustrates the selection bias evoked by D'Amuri (2009) in favour of the young as regards the Internet job search. □

References

Askitas N. & K.F. Zimmermann, 2009, Google Econometrics and Unemployment Forecasting, Applied Economics Quarterly, 55(2), 107-120.

Castle J.L., N.W.P. Fawcett & D.F. Hendry, 2009, Nowcasting is not just Contemporaneous Forecasting, forthcoming in the National Institute Economic Review.

Choi H. & H. Varian, 2009a, Predicting the Present with Google Trends, *mimeo*, Google Inc.

Choi H. & H. Varian, 2009b, Predicting Initial Claims for Unemployment Benefits, *mimeo*, Google Inc.

Clark P.K., 1987, The Cyclical Component of US Economic Activity, Quarterly Journal of Economics, 797-814.

Clark P.K., 1989, Trend Reversion in Real Output and Unemployment, Journal of Econometrics, 40, 15-32.

Cornec M., 2006, Analyse factorielle dynamique multifréquence appliqué à la datation de la conjoncture française, Economie & Prévision, 172, 29-43.

Cornec M. & T. Deperraz, 2006, Un nouvel indicateur synthétique mensuel résumant le climat des affaires dans les services en France, Economie & Statistique, 395-396, 13-38.

D'Amuri F., 2009, Predicting unemployment in short samples with internet job search query data, MPRA Paper No. 18403.

D'Amuri F. & J. Marcucci, 2009, "Google it!" Forecasting the US unemployment rate with a Google job search index, ISER Working paper series, 2009-32.

Diebold F.X. & R.S. Mariano, 1995, Comparing Predictive Accuracy, Journal of Business and Economic Statistics, 13(3), 134-144.

Doornik J.A. 2009, Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data, mimeo, University of Oxford.

Dordonnat V., S.J. Koopman, M. Ooms, A. Dessertaine & J. Collet, 2008, An Hourly Periodic State Space Model for Modelling French National Electricity Load, International Journal of Forecasting 24, 566–587.

Doz C. & F. Lenglart, 1999, Analyse factorielle dynamique : test du nombre de facteurs, estimation et application à l'enquête de conjoncture dans l'industrie, Annales d'Economie et de Statistique, 54, 91-127.

Durbin J. & S.J. Koopman, 2001, Time Series Analysis by State-Space Methods, Oxford.

Ferrara L. & S.J. Koopman, 2010, Common Business and Housing Market Cycles in the Euro Area from a Multivariate Decomposition, working Paper Banque de France 275.

Giannone D., L. Reichlin & D. Small, 2008, Nowcasting: the Real-time Informational Content of Macroeconomic Data, Journal of Monetary Economics, 55, 665-676.

Ginsberg J., M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski & L. Brilliant, 2009, Detecting Influenza Epidemics Using Search Engine Query Data, Nature, 457(19), 1012-1015.

Harvey D., S. Leybourne & P. Newbold, 1997, Tests of Equal forecast Accuracy and Encompassing for Nested Models, International Journal of Forecasting, 13, 281-291.

Kholodilin K.A., M. Podstawski & B. Siliverstovs 2010, Do Google Searches Help in Nowcasting Private Consumption? A Real-Time Evidence for the US, KOF Working Papers n°256, April, Zurich.

Kim C.J. & C.R. Nelson, 1999, State-Space Models with Regime-Switching: Classical and Gibbs-Sampling Approaches with Applications, MIT Press.

Koopman S. J. & J. Durbin, 2000, Fast Filtering and Smoothing for Multivariate State Space Models, Journal of Time Series Analysis 21(3), 281-296.

Koopman S. J. & J. Durbin, 2003, Filtering and Smoothing of State Vector for Diffuse State-Space Models, Journal of Time Series Analysis, 24(1), 86-98.

Koopman S. J. & H.P. Franses, 2002, Constructing Seasonally Adjusted Data with Time-varying Confidence Intervals, Oxford Bulletin of Economics and Statistics, 64(5), 509-526.

Kulkarni R., K. Haynes, R. Stough & J.H.P. Paelinck, 2009, Forecasting Housing Prices with Google Econometrics, mimeo, School of Public Policy, George Mason University, Fairfax, VA 22030.

Kuttner K.N., 1994, Estimating Potential Output as a Latent Variable, Journal of Business and Economic Statistics, 12, 361-368.

Mariano R.S. & Y. Murasawa, 2003, A New Coincident Index of Business Cycles Based on Monthly and Quarterly Series, Journal of Applied Econometrics, 18, 427-443.

Schmidt T. & S. Vosen, 2009, Forecasting Private Consumption: Survey-based Indicators vs. Google Trends, RUHR Economic Paper 155.

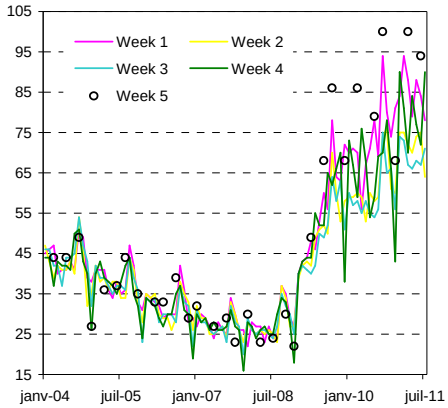
Schumacher C. & J. Breitung, 2008, Real-time Forecasting of German GDP based on a Large Factor Model with Monthly and Quarterly Data, International Journal of Forecasting, 24, 386-398.

Stock J.H. & M.W. Watson, 1991, New Indexes of Coincident and Leading Economic Indicators, NBER Macroeconomics Annual, 4, 351-409.

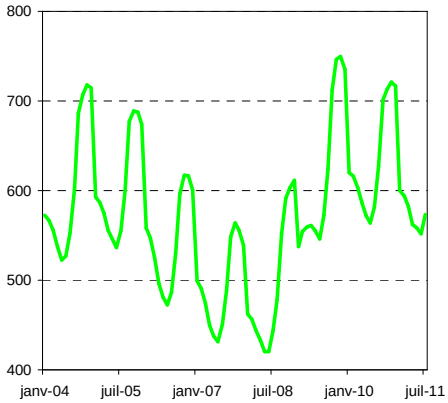
Suhoy T., 2009, Query Indices and a 2008 Downturn: Israeli Data, Discussion Paper No. 2009.06, Bank of Israel.

Suhoy T., 2010, Monthly Assessments of Private Consumption, Discussion Paper No. 2010.09, Bank of Israel.

Figure 1: Monthly data

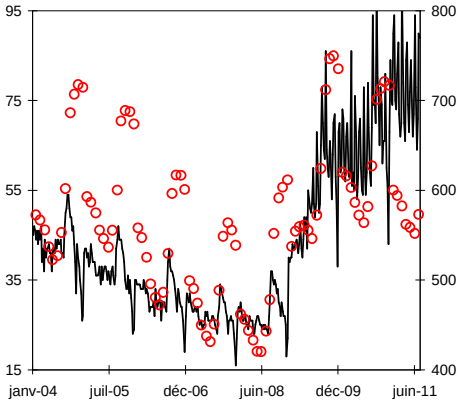


(a): Google indexes (NSA, by week)

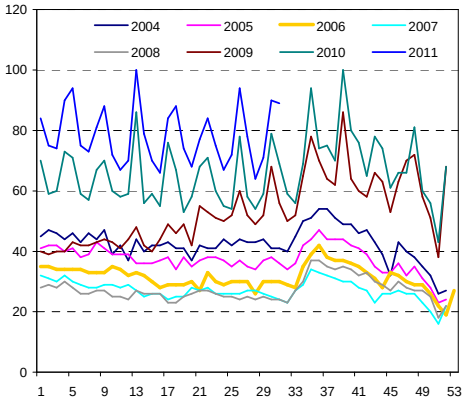


(b): DEFM (15-24 years, NSA, in thousand)

Figure 2: Weekly data

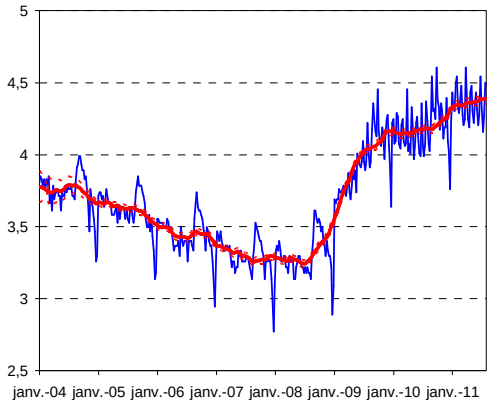


(a): the Google index (NSA) and the DEFM (15-24 years, NSA, in thousand)

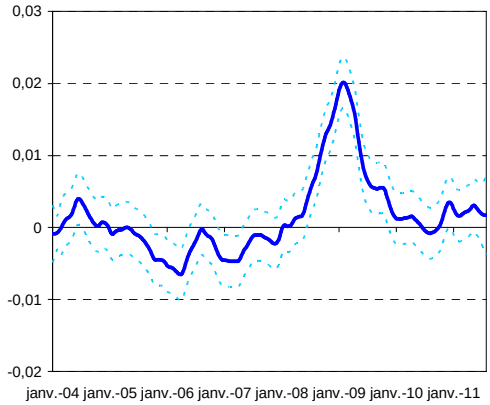


(b): The Google index (NSA, split by year)

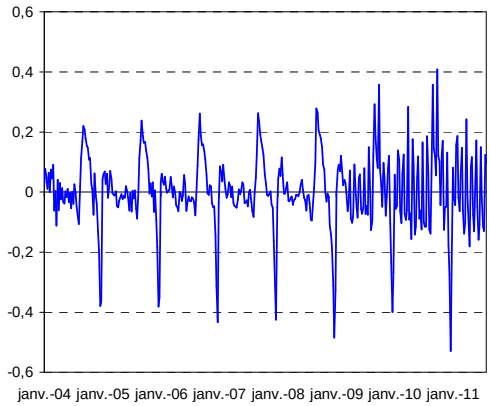
Figure 3: (smoothed) components for the Google data (bivariate model)



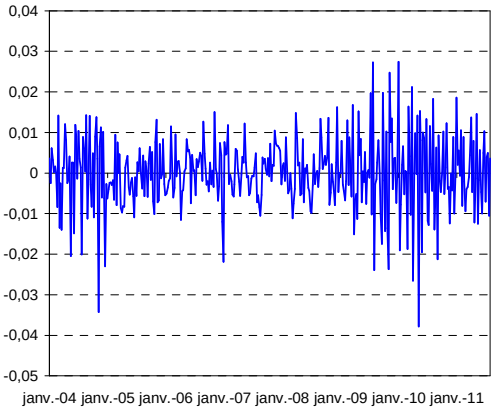
(a) Google index + Trend



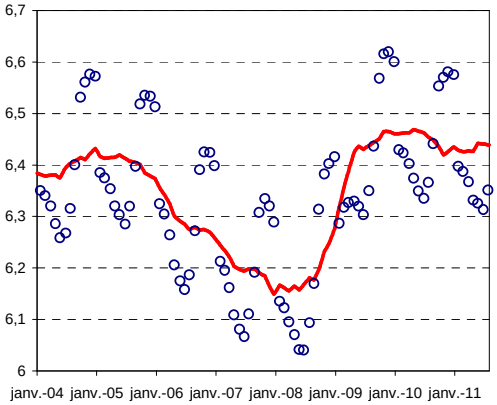
(b): Slope



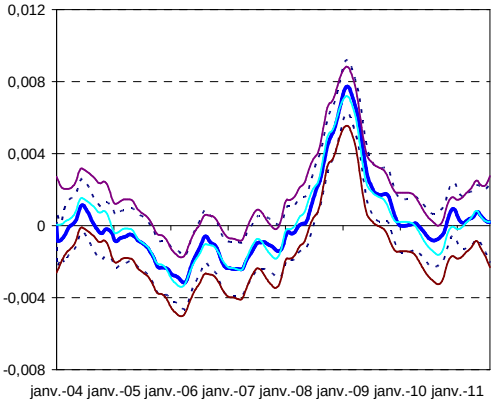
(c) Seasonality



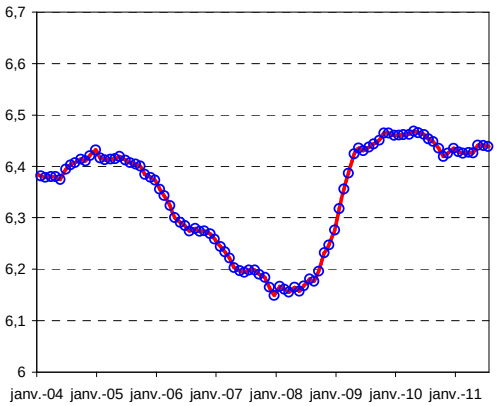
(d) Irregular component



(e) DEFM (NSA) + Trend



(f) Slope from univariate and bivariate models



(g) DEFM (SA) + Trend

Table 1: Specification of the Fourier weights for Google data seasonality

	Likelihood contribution	Number of total estimated parameters	Number of estimated variances	*LR-test (p-value)*
General model (all weights are random walks)	368.062	62	52	-
Most Constrained model (all weights are constant)	212.383	10	0	0
Retained model (24 random walks and 28 constant)	363.273	34	24	0.999

Seasonality is modelled with a Fourier decomposition series:

$$S_{1,t} = \sum_{j=1}^{S/2} \left\{ a_{j,t} \cos\left(\tau_t \frac{2j\pi}{S} \right) + b_{j,t} \sin\left(\tau_t \frac{2j\pi}{S} \right) \right\}$$

with S the seasonality (here 52.25 weeks) and τ_t the number of weeks elapsed from the beginning of the current year. We allow evolutions in the composition of the seasonality since weights ($a_{j,t}$, $b_{j,t}$) are random walks:

$$\begin{cases} a_{j,t} = a_{j,t-1} + \varepsilon_t^{a_j} & \varepsilon_t^{a_j} \approx N(0, \sigma_{a_j}^2) \\ b_{j,t} = b_{j,t-1} + \varepsilon_t^{b_j} & \varepsilon_t^{b_j} \approx N(0, \sigma_{b_j}^2) \end{cases} \qquad j = 1, \ldots \frac{S}{2}$$

Table 2-a: Results for Google equation

		Univariate model	Bivariate model
Seasonality	σ_{b_3}	5.13E-03	5.29E-03
	σ_{b_4}	9.88E-04	1.00E-03
	σ_{a_5}	1.66E-03	1.61E-03
	σ_{a_6}	9.65E-04	9.65E-04
	σ_{b_6}	9.92E-04	9.90E-04
	σ_{a_7}	1.09E-03	1.09E-03
	σ_{a_8}	2.51E-03	2.53E-03
	σ_{b_8}	6.36E-04	6.44E-04
	σ_{a_9}	1.19E-03	1.19E-03
	$\sigma_{a_{11}}$	8.17E-04	8.28E-04
	$\sigma_{a_{12}}$	2.89E-03	2.90E-03
	$\sigma_{b_{12}}$	6.59E-03	6.59E-03
	$\sigma_{a_{14}}$	1.14E-03	1.15E-03
	$\sigma_{b_{14}}$	1.38E-03	1.38E-03
	$\sigma_{b_{15}}$	9.40E-04	9.46E-04
	$\sigma_{a_{16}}$	9.64E-04	9.63E-04
	$\sigma_{a_{18}}$	8.36E-04	8.35E-04
	$\sigma_{b_{18}}$	1.95E-03	1.95E-03
	$\sigma_{a_{20}}$	6.31E-04	6.29E-04
	$\sigma_{a_{22}}$	9.69E-04	9.74E-04
	$\sigma_{a_{24}}$	4.85E-03	4.85E-03
	$\sigma_{b_{24}}$	1.45E-03	1.44E-03
	$\sigma_{a_{26}}$	1.27E-03	1.27E-03
	$\sigma_{b_{26}}$	3.16E-03	3.15E-03
Trend	σ_{T_1}	1.30E-02	1.28E-02
Slope	σ_{d_1}	1.10E-03	1.21E-03
Innovation	σ_{y_1}	1.95E-02	1.95E-02
Dummies	δ_1	**-5.77E-02**	**-5.76E-02**
	δ_2	**-5.91E-02**	**-5.89E-02**
	δ_3	1.27E-01	1.30E-01
	δ_4	**4.41E-01**	**4.41E-01**
	δ_5	4.29E-03	5.66E-03
	δ_6	5.12E-02	5.18E-02
	δ_7	**2.53E-01**	**2.53E-01**

Dummies variables are used to reach normality of prediction errors and to control for outliers. They concern respectively weeks with holidays, bridge days and 5 outliers.
Bold means significant at 5%.

Table 2-b: Results for DEFM equation

		Univariate model	Bivariate model
Trend	σ_{T_2}	3.68E-03	3.76E-03
Slope	σ_{d_2}	5.27E-04	1.58E-04
Innovation	σ_{y_2}	-	-
Seasonality	a_1	**5.16E-02**	**5.13E-02**
	b_1	**-1.30E-01**	**-1.30E-01**
	a_2	**-3.98E-02**	**-3.99E-02**
	b_2	**-5.73E-03**	**-5.75E-03**
	a_3	**-6.16E-03**	**-6.19E-03**
	b_3	**-8.90E-03**	**-8.89E-03**
	a_4	**-1.57E-02**	**-1.58E-02**
	b_4	**-1.11E-02**	**-1.11E-02**
	a_5	**-1.68E-02**	**-1.68E-02**
	b_5	-9.80E-04	-9.80E-04
	a_6	**-3.81E-03**	**-3.81E-03**
Google impact	α_0	-	-4.97E-04
	α_1	-	**4.09E-01**

Bold means significant at 5%.

Table 2-c: Specification tests on prediction errors

		Univariate model	Bivariate model
	Likelihood	642.005	658.168
Normalized prediction error #1	Box-Pierce(28)	0.843	0.689
	ARCH(28)	0.772	0.777
	Normality	0.950	0.390
Normalized prediction error #2	Box-Pierce(28)	0.314	0.264
	ARCH(28)	0.661	0.477
	Normality	0.568	0.396

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%.

Table 3: Test for no systematic predictive bias

Univariate model	Before 24th	-0.568
	After 24th	-0.672
Bivariate model	2 weeks ahead	-1.160
	1 week ahead	-1.027
	Nowcasting	-1.198

This table displays the *t*-stat for testing the nullity of prediction error mean (the test is bilateral). Bold means rejection of the null at 5%.

Table 4: Predictive performances (RMSE and MAPE)

	Horizon	RMSE	MAPE
Univariate model	Before 24th	3.25E-02	6.08E-02
	After 24th	1.72E-02	4.42E-02
Bivariate model	2 weeks ahead	2.51E-02	5.532E-02
	1 week ahead	**1.22E-02**	**3.84E-02**
	Nowcasting	1.26E-02	3.91E-02
	Max Gain (%)	26.8	11.4

Bold highlights the minimal RMSE or MAPE for each model. The maximum gain is calculated as the percentage of the minimum indicator relatively to the univariate model (which produces predictions for DEFM after the 24th without using Google data).

Table 5: Modified Diebold & Mariano test

		Univariate model	Bivariate model		
Models	Horizon	After 24th	2 weeks ahead	1 week ahead	Nowcast
Squared error loss					
Univariate model	Before 24th	2.44**	2.01**	2.47**	2.47**
	After 24th	-	-2.66**	2.33**	2.27**
Bivariate model	2 weeks ahead	-	-	2.76**	2.76**
	1 week ahead	-	-	-	-1.42
Absolute error loss					
Univariate model	Before 24th	3.60**	1.89*	3.52**	3.54**
	After 24th	-	-4.12**	2.22**	2.10**
Bivariate model	2 weeks ahead	-	-	4.37**	4.48**
	1 week ahead	-	-	-	-1.26

This table displays the *t*-stat for testing the nullity of the loss function mean from two competing predictions (the test is bilateral). Negative (positive) statistics means that the row model out- (under-) performs the column model. ** means rejection of the null at 5%. * means rejection of the null only at 10%.