Nowcasting with Big Data: is Google useful in Presence of other Information?*

Xinyuan Li
London Business School
xli@london.edu

First Draft: March 20, 2015
This Draft: January 28, 2016

Abstract

I study the usefulness of Google Trends data in nowcasting the US jobless initial claims and employment in a factor model. In contrast to what has been established in the literature that relevant Google search terms or search term categories improve the forecast performance, I show the improvement is minimal, if any, when also considering other conventional macroeconomic data. This paper provides a useful framework incorporating new data sources, which could be at weekly or daily frequency, for nowcasting purpose. Some suggestion on the use of Google Trends series is given.

Key words: forecast, nowcast, factor model, big data

JEL classification: C32, C53, E17, E24, E27

1 Introduction

Monetary policy decisions in real time are made based on economic statistics. Because economic data are often released with significant delay and subsequently revised, how to form early and accurate predictions of economic conditions in the current quarter (nowcasting) and in a longer run (forecasting) becomes crucial to central banks and markets.

The recent surge in big data may largely help in forming these predictions. Particularly, web-based search data have drawn lots of attention of academic researchers and central bankers, and they have been used as extra explanatory variables in forecasting models and shown to improve the forecast performance. See, for instance, the seminal work by Choi and Varian (2012), and working papers of Bank of England (McLaren and Schabhogue, 2011), Bank of Italy (D'Amuri and Marcucci, 2015), Bank of Spain (Artola and Galan, 2012), Central Bank of Chile (Carrière-Swallow and Labbé, 2010), Central Bank of the Republic of Turkey (Chadwick and Sengül, 2012), Bank of Isreal (Suhoy, 2009).

*I thank Nikoleta Anesti, Fabrizio Dell'Acqua, Raffaella Giacomini, Matthew Harding, Sebastian Hohmann, Silvia Miranda Agrippino, Michael McCracken, Michele Modugno, Athanasios Orphanides, Lucrezia Reichlin, the seminar participants at Now-casting, DIW Macroeconometric workshop 2015, LBS Brownbag, for helpful comments and discussion. Part of this paper was written during my visit to the Bank of England. I thank for their hospitality and the view of this paper does not necessarily reflect those of the Bank of England. Email: xli@london.edu.

However, two aspects of search data are of great interest but they have not been addressed in the literature. Intuitively, these data can be helpful in roughly two ways: first, they might be able to measure some economic activities that traditional measures or methods fail to capture, and second, these data are released without much delay. The first point refers to their actual efficacy. Bean (2015) discusses the falling response rates to traditional survey questionnaires and internet-based services that are not picked up by conventional approaches. This is where query data can be helpful. But given they may already have a large set of economic data at hand that may contain very similar information, what can policy makers or economic decision makers make of this new type of data? Most literature has said nothing about this, as the forecasting models so far used in the literature are not able to incorporate data series of different frequencies and of large dimensions.

The second point is timeliness. National statistical institutes usually publish economic data with material delay. For example, the first estimate of the US gross domestic product (GDP) is published about 30 days after the reference quarter. In fact, traditional data sources face a trade-off between timeliness and accuracy: early estimates based on incomplete information will be less reliable than later ones based on more complete information. Bean (2015, Chart 2.A) shows the quantity of information available for measures of UK GDP increases following the end of the reference quarter. The first (preliminary) estimate, which comes out 25 days after the reference quarter, includes roughly 47% of output data for that quarter. By the time the third estimate is published, 89 days after the end of the reference quarter, well over 90% of the data is available. But the longer a decision maker has to wait for the statistics, the less useful are they likely to be. Survey data are more timely: it is possible to get survey results around five to ten days before the reference period. And previous studies have shown the indispensable role of survey data for GDP forecast when their timely publication is taken into account properly (see, for instance, Bańbura and Rünstler, 2011). New data sources, such as web-based query data, have a big advantage in terms of timeliness. Thank to technological advances, these data are ready for publication nearly instantaneously upon occurrence of real economic activities and they do not require subsequent revisions. This could be exploited to get more timely predictions. Again, in presence of other variables, does their timely publication improve forecast? This question requires a model that can take into account constant information flows.

In this paper, I address these two questions using a state-of-art factor model as in Bańbura et al (2013). I need to emphasize the difference between this approach and what has been established in the literature. In most of the papers that study the use of query data in nowcasting for forecasting (see, e.g. Choi and Varian, 2012; D'Amuri and Marcucci, 2015, Doornik, 2009), only very simple time series models are used, such as AR or ARMA, and search data enter the models as extra predictors. These models have a few drawbacks in our nowcasting context: they are not able to take into account of a large number of series and facilitate comparison of their information content; they can use only a monthly

series of search data (typically weekly) by selecting one or two specific weeks or by averaging weeks over a month or a quarter to retain the same frequency as the forecast target leads to impoverishment of the data (Fondeur and Karamé, 2013).

The factor model used in this paper can overcome these drawbacks. The idea is that in this model the unconventional search data are put together with other key economic variables and the data are potentially (i) of large dimension and (ii) of mixed frequencies and are released (iii) at different publication lags ('ragged edge problem). The forecast is updated upon every release and forecast accuracy is measured. This mimics the information flows in real time and facilitates assessment of the usefulness of the new data. Furthermore, this model allows for counterfactual analysis in the sense that we can easily push forward or backward the publication date so that the timeliness of data can be promoted or deduced. Then we will be able to discern whether improvement in forecast accuracy, if any, is attributed to efficacy or timeliness.

Factor models, featured as a parsimonious scheme, have been shown as successful in summarizing information, as well as nowcasting, and they have been widely used in academic literature and in central banks (see also Boivin and Ng, 2005; Forni et al., 2005b; D'Agostino and Giannone, 2006; Giannone et al., 2008; Marcellino et al., 2003; Stock and Watson 2002a, b).

Within this framework, I show an application of Google Trends data to the US labour market. To be exact, I take the two indices used in an influential paper in this literature, Choi and Varian (2012), put them together with other 29 economic variables, which are often used in a factor model for nowcasting purpose, and forecast the US jobless initial claims and employment. Then I compare the forecast performance of the model with Google search data and without. This exercise has various purposes. First, Google Trends query data have received an enormous amount of attention and have led to rather fruitful research. Second, the application to labour variables is steadily backed by the theory: in a classic search-and-match model, the number of newly matched pairs of vacancies and workers is a function of the workers that have been searching for a job and the vacancies that have been posted by companies (see, e.g. Pissarides, 2000). Therefore, Google Trends series that are a direct measure of search intensity should shed some light on employment condition. This application seems somehow more natural than GDP forecasting.

The framework is so flexible that applications can be easily done to other forecast target, regardless of its frequency and publication lags, or to data of other countries, though it might require some careful selection of the search keywords. It is likely that this type of new data should be more helpful when data quality is poor or publication lags are significant.

Another contribution of this paper is a rather comprehensive literature review. I discuss the platform of Google Trends and of Google Correlate, and their applications in different fields. Several issues that may influence robustness of forecast results, such as variable selection and sampling error, are discussed

in detail and suggestion to tackle these issues is provided.

The remaining part of the paper is organized as follows: Section 2 gives an introduction to Google Trends and Google Correlate, Section 3 discusses various issues of the data, and reviews the applications and the methodology, Section 4 sets up a dynamic factor model, Section 5 reviews the data, Section 6 shows the empirical results, and Section 7 concludes.

2 Google Trends and Google Correlate

2.1 Google Trends

Understanding the data series we are working with is necessary, especially when the data source is unconventional. So in this section I give an introduction to how Google Trends and Google Correlate are constructed and what can be achieved from the their websites. A part of the description is borrowed from D'Amuri and Marcucci (2015).

The Google Trends website gives how often a particular term is searched relative to the total search volume over a certain period of time in a certain geographical region. It gives an index, instead of the absolute number of searches, due to privacy reason. Particularly, the index is constructed in the following way: the search share $S_{d,r}$ for a particular keyword in day d, in region r, is given by the number of web searches containing that keyword, $V_{d,r}$, divided by the total number of web searches performed through Google in the same day in that area $T_{d,r}$, i.e. $S_{d,r} = V_{d,r}/T_{d,r}$. A week is defined to start on Sunday and end on Saturday. Then the search share of week w is given by the weekly average

$$S_{w,r} = \frac{1}{7} \sum_{d \in w, d=Sun}^{Sat} S_{d,r}.$$

Upon request, the Google Trends website produces an index, with the largest $S_{w,r}$ in the requested period scaled to 100. To be precise, if a user requests the Google Trends index over weeks that are in a set $[\underline{w}, \overline{w}]$, then the Google Trends website will give the index $GI_{w,r}$ as

$$GI_{w,r} = \frac{100}{\underset{i \in [\underline{w}, \overline{w}]}{max} S_{i,r}} S_{w,r}, \qquad w \in [\underline{w}, \overline{w}].$$

Data are gathered using IP address information and are made available to the public if the number of search terms exceeds a certain undeclared threshold. Repeated queries from a single IP address within a short period of time are eliminated to avoid a single person or a robot sending identical queries over a short period to bias the sample.

The data is available since Jan 04, 2004. The index for j-th week is available at the beginning (with Sunday as the first day) of the $(j+1)$-th week. This timely release enables us to conduct early forecast,

as argued in Choi and Varian (2012). By default, the data exported from the website is weekly if the requested period is longer than three months. For instance, if one requests the search index for the word 'Jobs' from Jan 04, 2004 to present, the dataset will be weekly. However, the website will generate daily index if the requested time span is less or equal to three months. With the transformation specified in Appendix B, it is possible to achieve a daily version of the index. Depending on the purpose, one can decide which frequency to use. Obviously high-frequency data contains more noise, for instance, the search category `'Jobs'` exhibits strong seasonality during a week, with clear troughs on weekends.

Apart from time span, Google Trends also provides options on countries, regions, cities, categories, and languages, if one has particular interest in these features. There are altogether 25 categories and about 200 subcategories. When a category is requested, the index will be given in the percentage change of the index (over the previous period) with the observation of the week starting on Jan 04, 2004 initialized as 0%.

2.2 Google Correlate

Google Correlate is a tool that allows users to upload their data series of interest, and it computes the Pearson Correlation between the data of interest and the search intensity of every query in Google database, then shows the queries whose search intensity over time is most correlated with the data of interest (Vanderkam et al, 2013).

Google Correlate provides data of two kinds: spatial and temporal. The spatial one allows us to upload a dataset by US state. Then it gives the queries whose search intensity across states is most correlated with the data of interest. The temporal one allows us to upload a time series starting from Jan 04, 2004. It allows for weekly and monthly frequency. And therefore it gives the queries who search intensity over time is most correlated and these queries are of the same frequency as the series of interest. The search intensity again is an index instead of an absolute number. Figure 1 gives an example of a Google Correlate request. The target variable, the US unemployment rate (seasonally unadjusted) has been uploaded and Google Correlate produces a list of 100 search terms whose search intensity is most positively correlated with the target. And beneath the list, the plot shows the search intensity of one of the terms, together with the target.

3 Literature: issues, applications and methods

In this section, I discuss the various issues with Google Trends and Google Correlate data and review how the literature tackles these issues. In the meantime, I review the applications and methods that have been established in the literature.

3.1 Variable Selection

Variable selection has to do with two questions. One question is how Google applies its algorithm to millions of queries and form the indices in Google Trends and Google Correlate. It is unknown to us how the algorithm is designed and whether the algorithm is stable over time. Lazer et al (2014) discuss this algorithm dynamics problem.

Another question is how we economists select variables from the output of Google Trends and Google Correlate. Judgment is involved in most of the existing literature when it comes to select predictors. See, for instance, Choi and Varian (2012). Da et al (2015) take the well-known dictionaries in the finance and textual analytics literature (Tetlock 2007) as a starting point, where the latter provides a better ground for keyword selection. However, there are a few studies trying to automatize the variable selection procedure. In their seminal work, Ginsberg et al (2009) automatically pick 45 queries from 50 million candidate queries based on the fit against out-of-sample influenza-like illness data. So their method is automatic but exhaustive. Scott and Varian (2014) also make the attempt to construct a more robust and automatic system selecting predictors for nowcasting weekly initial claims and monthly retail sales. They build a state space model with a time series part that captures the trend and seasonality in the data and a regression part that includes predictors from Google Trends and Google Correlate. Bayesian shrinkage is used in the regression part for variable selection. They show that adding the regression part can reduce the forecast error. However, there is still another problem with Google Correlate, which is that the 100 queries might not have economic meaning. As Scott and Varian (2014) argue, in case of nowcasting weekly initial claims (seasonally unadjusted), 'of the 100 top predictors from Google Correlate, 14 were queries for unemployment for a specific state'. But this is not true in case of the weekly jobless claims (seasonally adjusted) or unemployment rate (seasonally adjusted or unadjusted). Take unemployment rate (seasonally unadjusted) as an example. Figure 1 shows the Google Correlate output when the US unemployment rate (seasonally unadjusted) is the target: out of the top ten most relevant search terms, only 'alabama unemployment' is economically meaningful. If we included all these terms in our regression model, the regression would be spurious and it is quite unlikely that the predictors can add any predictive power. Though they do not discuss this problem explicitly, Scott and Varian (2014) add Google Trends data with economic meaning as predictors when nowcasting monthly retail sales, which does slightly better than only using Google Correlate data.

3.2 Sampling Error

To increase the response speed, Google currently calculates the index based on a random sample from the historical data and this will result in sampling error. This is hard to detect because if a user sends the same request from the same gmail account, from the same IP address on the same day, Google will give the same index value. This is why most papers in the literature do not mention this. Choi and Varian

(2012) mention that 'Google Trends data is computed using a sampling method and therefore vary a few percent from day to day'. Da et al (2015) also realize the sampling issue but they believe that the impact of such sampling error is small for their study and it should bias against finding significant results.

On the other hand, Carrière-Swallow and Labbé (2010) notice that the sampling appears to take place daily, such that requesting an identical query on different days returns slightly different series. They downloaded the series on 17 occasions during May and June 2010 and compute the cross-sectional mean at each time t. They also plot the signal-to-noise ratio of the series and they conclude the signal is strong according to the Rose criterion. D'Amuri and Marcucci (2015) point out that the indices can vary depending on the download date and the IP address. Therefore, they take the average of 24 downloads carried out on 12 different days from two IP addresses, and they argue the correlation between these downloads are never lower than 0.99. McLaren and Schabhogue (2011) also notice the sample is drawn daily and they take the average of the data generated on seven consecutive days. However, in their application to housing market, they notice Google Trends data of certain search terms vary significantly when downloaded on different days, perhaps because of low search volumes and this volatility affects the robustness of the result. Instead, they choose a search term more stable when downloaded on different days to circumvent this problem.

In order to gauge the magnitude of the sampling error and to assess its impact on forecast performance, I did two exercises. The first exercise is to download multiple samples on the same day. The purpose of this is, first, to mimic the real-time nowcasting practice, that is, once a data point comes out, it should be immediately incorporated into the model, as opposed to downloading the data over multiple days as in Carrière-Swallow and Labbé (2010) and McLaren and Schabhogue (2011). Of course, one can argue that the latter approach might help mitigate the sampling error, but there is always a trade off between timeliness and signal precision. The second purpose is that if the sampling error turned out to be large, we could consider using sample mean of these multiple downloads as the data to incorporate into the forecasting model instead of a single download. Precisely, I focus on the same two categories, `'Jobs & Education\Jobs'` and `'Law & Government\Social Services\Welfare & Unemployment'` as in Choi and Varian (2012). And I downloaded 52 samples, using four different gmail accounts, from 13 different IP address, on Dec 25, 2015. In this way, most of the samples are identical [1]. The original data used in Choi and Varian (2012) is available from Hal Varian's website. In Figure 2 and 3, the upper panels show the raw data series from Choi and Varian (2012) and the sample mean of 52 samples and the lower panels show the standard deviation of the samples. Three points emerge: first, the within-day sampling error is not as large, compared to the magnitude of the index. The standard deviation of `Jobs` stabilizes around 0.6, and the standard deviation of `Welfare & Unemployment` ranges from 1 to

[1] I also tried downloading the data from different gmail accounts, from different IP, but on different days in November 2015 and it turns out that samples between days vary a bit more than within days. The standard deviation of `Jobs` is around 1.5 while the standard deviation of `Welfare & Unemployment` still ranges from 1 to 5. If plotted, though, these samples look nearly identical because the variation is yet small compared to the absolute magnitude of the indices.

5. Second, the between-day variation seems more important, and particularly, the variation caused by dynamic algorithm as mentioned in Section 3.1 seems to be the main source of such variation, instead of between-day sampling error. This is because the data downloaded over a shorter period of time, say, a month, vary much less than the data downloaded over a longer elapse (e.g. compare my samples with Choi and Varian (2012)). Third, the magnitude of sampling error may vary a lot across indices. As mentioned in McLaren and Schabhogue (2011), some search terms might be too volatile to facilitate any forecast. Therefore a careful study into the sampling error is essential, prior to any application of such data.

The second exercise is to assess the impact of sampling error on nowcasting. To this end, I do a baseline forecast evaluation, following Choi and Varian (2012), using the data available at these two vintages. Table 1 shows the mean absolute error (MAE) and root mean square error (RMSE) of 1-step-ahead forecast of initial claims using a random walk without drift (RW), a random walk with drift (RW with drift), an AR(1), an AR(1) augmented by Google Trends series `'Jobs'` and `'Welfare & Unemployment'`. Further, the augmented AR model takes the Google Trends series with three variations: (a) the same data used in Choi and Varian (2012), deseasonalized using `stl` command with the smoothing parameter s.window='periodic', (b) the same data used in Choi and Varian (2012), deseasonalized using `stl` command with the smoothing parameter s.window=7, (c) one data sample [2] that was downloaded on Dec 25, 2015, deseasonalized using `stl` with the smoothing parameter s.window='periodic'. The sample period is Jan 10, 2004 to Jul 01, 2011 and the evaluation periods are listed in the table. These are the same in Choi and Varian (2012). As we can see, their original result is not very robust to different vintages, and several 1-ratios that are less than one become larger with the new data.

From the previous exercises it is clear that depending on the search term, sampling error can vary and the forecast performance may be affected by sampling error. Given the sampling approach of Google, downloading the series from multiple IP addresses over a short period of time and getting the average seems a good solution. However, when the target variable of forecast comes at a high frequency, e.g. weekly, then downloading over days or weeks is not timely any more; it has to be done with a day or so. This, in turn, requires a large number of gmail accounts and IP addresses and hence web crawling might be worth trying. But again, as we do not know the 'true' dataset, there is no way of know what is the sufficient sampling size.

3.3 Individual and Social Searching

Ormerod et al (2014) point out that people may search for a phrase because they genuinely want the information (independent searching) or they may search simply because others are searching for it (social searching), and this may affect forecasting result. Using the Bass diffusion model, they show that the

[2] The result using the sample mean is very similar to the result using one single sample.

independent search for information motive was much stronger in the cases of accurate prediction than in the inaccurate ones and social search was stronger in the cases of inaccurate prediction. Then an important task is therefore to give some indication about the relative importance of the individual and social motives for the searches, in the early phase of a rise in search activity. Bentley and Ormerod (2010) show that using the Bass model a rapid rise followed by a slow decline indicates more independent motives, whereas a symmetrical outcome indicates more social motives. But they do not provide a way for diagnosis ex ante.

3.4 Seasonality

Seasonality is a prominent feature of Google Trends data. Figure 2 shows the strong seasonal component in `'Jobs'`, while it is not the case with `'Welfare & Unemployment'` in Figure 3.

In presence of such seasonality, it is natural to deseasonalize the data before using it. Choi and Varian (2012) use a command `stl` in `R` for seasonal adjustment, which is essentially a local regression method (Cleveland et al, 2009). Shimshoni et al (2009) also use the `stl` command, but they choose the smoothing parameter values to minimize the MAE. D'Amuri and Marcucci (2015) claim that they use X-13 ARIMA-SEATS method to deseasonalize monthly and weekly variables [3].

However, there are some issues with this type of methods. First, the methods mentioned above are essentially applying a two-sided filter on the raw data, which might cause material impact on the value at the end of the sample, while the value at the end of the sample turns out to be crucial to our forecast. Therefore applying a two-sided filter might not be appropriate in the context of forecast. Second, the choice of smoothing parameter value might also be a bit arbitrary and perhaps influence the forecast performance.

To examine the impact of seasonal adjustment on forecast performance, in Table 1 I show the 1-step-ahead forecast performance using (b) the same data used in Choi and Varian (2012), deseasonalized using `stl` command with the smoothing parameter s.window=7 and their original result (a) is not very robust. Figure 4, 5, 6, and 7 plot the raw data against the trend, seasonal and remainder part of the data, using different values for the smoothing parameter s.window. It is clear that when s.window='periodic', there is still some seasonal component in the remainder part of `'Jobs'`, while it is much reduced when s.window=7. This is why I choose this value of 7 as an experiment. Overall, the choice of this parameter requires careful inspection into the characteristics of the data and can be decided by a diagnostic method described in Cleveland et al (1990).

Given the pitfalls of filter-type seasonal adjustment, perhaps a better approach is to model the seasonal component explicitly. Fondeur and Karamé (2013) incorporate trend and the seasonal component

[3] But X-13 ARIMA-SEATS is not designed to deseasonalize weekly variables. It can only deseasonlize monthly or quarterly variables. Therefore the Bureau of Labor Statistics uses their own program MoveReg to deseasonalize weekly variables, e.g. initial claims.

in a state-space model. Scott and Varian (2014) also model the trend, seasonal and the regression part in a state space as in Harvey (1990). This leaves us with some important questions: how does seasonal adjustment affect forecast? How to do seasonal adjustment properly in order to get the optimal forecast? Given the surging interest in low frequency econometrics and its application in forecasting and uncertainty measurement (e.g. Müller and Watson, 2013, 2015), this could yield very important theoretical contribution, which I leave for future research.

3.5 Nowcasting Applications

Starting from Ginsberg et al (2009) predicting influenza-like illness, one main strand of application of Google Trends and Google Correlate data is nowcasting. The main idea is that Google search data reflects people's attention or interest, which is usually hard to measure. The information content of the search data could thus be unique and useful. And also because Google produces the search data in a timely fashion as elaborated in Section 2, the timeliness can be exploited for nowcasting.

There have been nowcasting applications on various topics: labor, consumption and consumer sentiment, housing, tourism, epidemics. Now I review the literature on these applications by topics.

One leading application is the nowcast of the unemployment rate or initial claims. Ettredge et al (2005) examine whether rates of employment-related searches by Internet users are associated with unemployment levels disclosed by the U.S. government in subsequent monthly reports. A positive, significant association is found between the job-search variables and the official unemployment data. They also observe longer lead times are associated with lower explanatory power. Following this branch of interest, Askitas and Zimmermann (2009) find strong correlations between keyword searches and monthly German unemployment rate. D'Amuri and Marcucci (2015) find models augmented with Google data outperform the traditional ones in predicting the US unemployment rate. Fondeur and Karamé (2013) get similar result from a state space model nowcasting youth unemployment in France. With Israeli data, Suhoy (2013) concludes Google query indices, from human resources (amongst others) are helpful in drawing inferences about the state of current economic growth, given the fact that official data are released with a delay. Choi and Varian (2012) use the query category `'Jobs'` and `'Welfare & Unemployment'` from Google Trends to help predicting the initial claims in the US. Scott and Varian (2014) use a structural time series model with a regression component capturing the contribution of Google Correlate data and forecast initial claims. Bughin (2010) does the nowcasting exercise for Belgian, Chadwick and Sengül (2012) for Turkey, and McLaren and Schabhogue (2011) for the UK.

On consumer behavior, Della Penna and Huang (2009) construct a consumer sentiment index for U.S. using Google Searches. Goel et al (2010) show that what consumers are searching for online can also predict their collective future behavior days or even weeks in advance. Kholodilin, Podstawski and Siliverstoves (2011) find Google search activity can help in nowcasting the year-on-year growth rates

of monthly US private consumption using a real-time data set. Vosen and Schmidt (2009) introduce a new indicator for private consumption based on search queries provided by Google Trends and this indicator outperforms other survey-based indicators both in-sample and out-of-sample. Carrière-Swallow and Labbé (2010) conduct similar out-of-sample evaluation for forecasting automobile sales in Chile. Scott and Varian (2014) use a structural time series model to nowcast retail sales. Choi and Varian (2012) show various examples for nowcasting motor sales, travel, and consumer confidence using Google Trends data.

On housing market, Kulkarni et al (2009) use Google search index at the city level to predict to predict change in the seasonally adjusted Case-Shiller index for 20 cities. Wu and Brynjolfsson (2014) find that a housing search index is strongly predictive of future housing market sales and prices. McLaren and Schabhogue (2011) use housing related search terms to predict housing price in the UK.

Other nowcasting applications are: inflation expectation (Guzman, 2011), epidemics (Ginsberg et al, 2009, Dukic et al, 2012), Birth (Billari, D'Amuri and Marcucci, 2013), tourism (Song, Pan and Ng, 2009, Choi and Liu, 2011), and special events (Schmidt and Vosen, 2012).

So far the literature seems to have suggested the usefulness of Google Trends and Google Correlate data in nowcasting, however, we can easily see why it might not be the case by reviewing the models primarily used in the literature. Typically, most of the papers in the literature use the following model:

$$\phi(L)y_t = \alpha + \beta x_t + \theta(L)\epsilon_t. \tag{1}$$

where y_t is the variable to forecast, $\phi(L)$ and $\theta(L)$ are lag polynomials. Some papers take $\phi(L) = 1 - aL$ and $\theta(L) = 1$ (e.g. Choi and Varian, 2012) or use information criteria to select the model (e.g. Schmidt and Vosen (2009)). D'Amuri and Marcucci (2015) run a horserace among various AR and ARMA models and select the one with the smallest MSE. x_t stands for extra predictors consisting of Google queries. It is important to keep in mind that the reason we can do this regression with contemporaneous x_t is that the release of x_t is more timely than y_t (Choi and Varian, 2012). x_t can either be the query index of a single word, such as 'jobs', or it can be the index for relevant categories or subcategories, such as `'Luxury goods'`, `'Home furnishing'` as in Della Penna and Huang (2009), or principal components extracted from the high dimensional Google search series as in Schmidt and Vosen (2009) and Kholodilin et al (2010). Then the out-of-sample performance of this augmented model is compared with a benchmark AR(1) model. [4]

[4]There are some exceptions though. Fondeur and Karamé (2013) use weekly series of web queries and monthly series of unemployment, extract the unobserved components with a modified Kalman filter so that both nonstationarity and multiple frequencies are taken into account. But just as Fondeur and Karamé (2013) admit in their paper that 'the choice of keywords is of course crucial for the study', in these works researchers select predictors using their own judgment of relevance to the particular prediction problem. Apparently, a big problem with this approach is that it does not easily scale to models where the number of possible predictors exceeds the number of observations. For this reason, Scott and Varian (2014) adopt three Bayesian techniques: Kalman filtering, spike-and-slab regression, and model averaging. Koop and Onorante (2013) conduct nowcasting with dynamic model selection (DMS) method, which allow for model switching between time-varying parameter regression models and therefore it might be helpful in an environment of coefficient instability and over-parameterization.

Clearly, there are several issues with this approach in the context of nowcasting. The most prominent problem is that this approach has not taken into account other available information that could be potentially useful for nowcasting, whereas using large panels is very standard in the nowcasting literature and proves to be successful. See, e.g. Giannone et al (2008), Bańbura et al (2013). It is unlikely that in presence of other information, Google query can still be useful. One easy way to think of this is that, instead of Equation (1), we consider

$$\phi(L)y_t = \alpha + \beta x_t + \gamma F_t + \theta(L)\epsilon_t \tag{2}$$

where F_t is the principal component of a large panel (See e.g. Stock and Watson, 2008) and it is very likely that β will not be significantly different from zero and therefore x_t will add minimal marginal predictive power.

Second, the mixed frequency has not been addressed. It is crucial to keep in mind, in line with the title of Castle et al (2009), that 'nowcasting is not just contemporaneous forecasting', as nowcasting makes use of the release of contemporaneous data and thus the treatment to these timely but not necessarily accurate time series is crucial in forecasting accuracy. When the forecast target variable and the regressors are of different frequencies, most papers simply take the monthly or quarterly average of the weekly Google data or select one or two specific weeks (See e.g. Choi and Varian, 2012, D'Amuri and Marcucci, 2015, Doornik, 2009). Fondeur and Karamé (2013) realize the dataset is generally 'impoverished' by doing so. Therefore their state space model is very suitable to solve the mixed-frequency problem. The state space model in Scott and Varian (2014) should also be able to incorporate mixed frequencies, but they do not really tackle this issue.

Given the limitations of the literature, I aim to answer three questions in this paper:

1. Are Google query data useful if we do not include other variables? (Sometimes yes, but not always robust. See Section 3.2 and 3.4)

2. Are Google query data useful if we do include other variables, given Google is more timely? (No.)

3. Are Google query data useful if we do include other variables, given Google's timeliness is removed? (No. This is not surprising given the answer to the second question is no. But if the answer to the second question was yes, then this question would clarify whether Google's advantage is due to timeliness or informativeness.)

3.6 Other Applications

There are also other applications of Google Trends and Google Correlate data, which do not fall into the realm of nowcasting: investor sentiment (Da et al, 2011, Da et al 2014, Siganos, 2013), market

volume (Bordino et al, 2012), volatility (Risteski and Davcev, 2014, Vlastakis and Markellos, 2012), risk diversification (Kristoufek, 2013), home bias in international investment (Mondria et al, 2010), forward looking and GDP (Preis et al, 2012), unemployment insurance and job search (Baker and Fradkin, 2015, Stevenson, 2008), Birth (Billari et al, 2013). Given the rich structure of Google Trends and Google Correlate, that is, they allow for spatial and temporal comparison and comparison across search terms, I think exploiting these variation could lead to very promising and fruitful research.

4 The Model

In this section I show how to assess the usefulness of Google Trends data in presence of other information source by using a dynamic factor model. Before I start with the specification of the model, it is worthwhile to emphasize the features of a nowcasting problem with information flows arriving constantly and in a non-synchronized fashion. A dynamic factor model is suitable to tackle such a problem and using a dynamic model to assess the usefulness of Google search data seems appropriate.

First, as mentioned in Section 3.5, it is standard in the nowcasting literature to use large panels (large n) to extract information and a suitable model should be able to handle a large panel and yet remain parsimonious. For instance, in the practice of nowcasting real economic activities, the number of variables is normally of two digits (e.g. Bańbura et al, 2013, $n = 24$, Giannone et al, 2008, n is about 200, Bańbura and Modugno, 2013, n ranges from 14 to 101, Bańbura and Rünstler, 2011, $n = 76$), and hence the factor model emerges as a parsimonious way of summarizing information from these variables, compared to the VAR model, which produces unstable estimates as the number of lags varies, which by essence is a typical syndrome of over-parameterization.

Second, the variables used for nowcasting are released in a nonsynchronized manner. In the context of standard forecasting, we do not care too much of the release dates of different variables, as timeliness is not the focus of medium-term or long-term forecasting, while for nowcasting, timeliness becomes important and therefore newly released data must be taken into account to revise the nowcast figure. With 'jagged-edged' data, Kalman filter is a natural tool to take into account of the missing value. This is implemented by putting a variance of infinity in the measurement equation when the datum is missing, or by not aggregating this datum in the Kalman filter iteration[5].In the same way, the missing values in the beginning of the time series due to different availability of the data can be properly handled. Typically the data used for economic forecasting are of different lengthy and it is not justifiable to disregard any useful information. Apart from the unavailable data at the beginning and at the end of the times series, the Kalman filter can also deal with missing data in any arbitrary pattern. This point is made clear by

[5]This is to say, in practice, one can put a very large number for the variance. Or equivalently, when the Kalman filter gets updated, the time series which is not yet released at the time of estimation is omitted from the iteration. The equivalence is due to the way the observed data enter the equation: they enter the measurement equation as precision-weighted sum, and therefore a datum with infinite variance will enter the equation as zero. This will become clear in the remaining part of this section.

Bańbura and Modugno (2013).

Third, variables are of mixed frequencies. This problem was first addressed by Mariano and Murasawa (2003), in which they combine monthly time series with quarterly GDP growth and construct a new coincident index of business cycles. Their motivation is to exploit the information content in real GDP, which turns out to be one of the most important coincident business cycle indicators but somewhat ignored by previous literature. With similar rationale, we want to exploit the information from variables at different frequencies. However, The difficulty of having multiple frequencies, e.g. weekly and daily, is that Mariano and Murasawa (2003) approach will result in a high dimensional state space, with the dimension of the state variables increasing at a rate of N. Bańbura et al (2013) solve this problem with a recursive representation that gives rise to a smaller state vector, and therefore faster computation [6]. This idea can date back to Harvey (1990).

As emphasized earlier, the real-time data flow is inherently high dimensional and therefore the factor model, featured as parsimonious, is particularly suitable in this context. The model used in this paper is borrowed from Bańbura et al (2013), but I will lay out the key equations here.

In a dynamic factor model, the observed time series are decomposed into two orthogonal parts: one is the common components, which is a product of the factor loadings and the factors, and the other is the idiosyncratic terms. The idea is that the factors should capture most of the variation if the observed series do co-move to a large extent. To be precise, we denote Y_t^d as the most frequently (demeaned, standardized) observed variables (in this paper, the most frequently observed data are daily, denoted with a superscript of d), Λ^d the factor loading, and F_t^d the factors, E_t^d the idiosyncratic term, and we allow for an autoregressive structure on the factors with $A(L)$ as the lag polynomial and U_t^d as the error term:

$$Y_t^d = \Lambda^d F_t^d + E_t^d \tag{3}$$

$$F_t^d = A(L)F_{t-1}^d + U_t^d \tag{4}$$

with

$$\begin{pmatrix} E_t^d \\ U_t^d \end{pmatrix} \sim i.i.d.\ \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} R_t & 0 \\ 0 & Q_t \end{pmatrix}\right) \tag{5}$$

Now we specify how to integrate data of different frequencies into one state-space representation. First we distinguish between stock and flow variables. Modugno (2011) develops a nowcasting model with daily, weekly and monthly data. In the same fashion, Bańbura et al (2013) seek for a representation of the less frequently observed variables as an exact aggregation of their corresponding latent, unobserved

[6] In Mariano and Murasawa (2003), with monthly and quarterly data, $N = 5$, and one factor with AR(1) dynamics, the dimension of the state variable is $5+5N = 30$. If with daily frequency, the dimension would be roughly $5+5(2\times30-1) = 300$.

variables. This treatment is aimed to reduce the dimension of the state variable. Particularly, assume that for a variable $Y_t^{k,\cdot}$ that is observable every k period, for $k = k_w, k_m, k_q$. The subscripts w, m, q stand for weekly, monthly and quarterly, respectively, and the dot in the superscript can either be f for a flow variable or s for a stock variable. Y_t^d is the corresponding underlying unobservable series. For instance, initial jobless claims is a weekly flow variable, $k_w = 7$ days, $Y_t^{k_w,f}$ is the flow variable, with Y_t^d as the underlying unobservable daily jobless claim. As shown in Appendix A.1, the less frequently observed variables can be represented with a weighted sum of the underlying:

$$Y_t^{k,s} = \sum_{i=0}^{k-1} \underbrace{1}_{=w_i^s} Y_{t-i} = \sum_{i=0}^{k-1} w_i^s (\Lambda^d F_{t-i}^{k_d} + E_{t-i}^{k_d})$$
$$Y_t^{k,f} = \sum_{i=-k+1}^{k-1} \underbrace{(k-|i|)}_{=w_i^f} Y_{t-k+1+i} = \sum_{j=0}^{2k-2} w_{j-k+1}^s (\Lambda^d F_{t-j}^{k_d} + E_{t-j}^{k_d})$$

Then the measurement equation can be written as follows (see Appendix A.2 for details):

$$\begin{bmatrix} Y_t^{k_q,f} \\ Y_t^{k_q,s} \\ Y_t^{k_m,f} \\ Y_t^{k_m,s} \\ Y_t^{k_w,f} \\ Y_t^{k_w,s} \\ Y_t^{k_d} \end{bmatrix} = \begin{bmatrix} \tilde{\Lambda}^{q,f} & & & & & & \\ & \Lambda^{q,s} & & & & 0 & \\ & & \tilde{\Lambda}^{m,f} & & & & \\ & & & \Lambda^{m,s} & & & \\ & & & & \tilde{\Lambda}^{w,f} & & \\ & 0 & & & & \Lambda^{w,s} & \\ & & & & & & \Lambda^d \end{bmatrix} \begin{bmatrix} \tilde{F}_t^{k_q,f} \\ F_t^{k_q,s} \\ \tilde{F}_t^{k_m,f} \\ F_t^{k_m,s} \\ \tilde{F}_t^{k_w,f} \\ F_t^{k_w,s} \\ F_t^{k_d} \end{bmatrix} + \begin{bmatrix} E_t^{k_q,f} \\ E_t^{k_q,s} \\ E_t^{k_m,f} \\ E_t^{k_m,s} \\ E_t^{k_w,f} \\ E_t^{k_w,s} \\ E_t^{k_d} \end{bmatrix} \tag{6}$$

I need to emphasize that with daily variables, i.e. the variables that are observable at the highest frequency, we do not distinguish between flow and stock variables, as whether they are stock or flow variables will not matter variables always observed, i.e. not regularly missing. The transition equation is

$$\begin{bmatrix} I_{2r} & 0 & 0 & 0 & 0 & 0 & W_t^{k_q,f} \\ 0 & I_r & 0 & 0 & 0 & 0 & W_t^{k_q,s} \\ 0 & 0 & I_{2r} & 0 & 0 & 0 & W_t^{k_m,f} \\ 0 & 0 & 0 & I_r & 0 & 0 & W_t^{k_m,s} \\ 0 & 0 & 0 & 0 & I_{2r} & 0 & W_t^{k_w,f} \\ 0 & 0 & 0 & 0 & 0 & I_r & W_t^{k_w,s} \\ 0 & 0 & 0 & 0 & 0 & 0 & I_r \end{bmatrix} \begin{bmatrix} \tilde{F}_t^{k_q,f} \\ F_t^{k_q,s} \\ \tilde{F}_t^{k_m,f} \\ F_t^{k_m,s} \\ \tilde{F}_t^{k_w,f} \\ F_t^{k_w,s} \\ F_t^{k_d} \end{bmatrix} = \begin{bmatrix} \mathcal{I}_t^{k_q,f} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathcal{I}_t^{k_q,s} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathcal{I}_t^{k_m,f} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathcal{I}_t^{k_m,s} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathcal{I}_t^{k_w,f} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathcal{I}_t^{k_w,s} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A \end{bmatrix} \begin{bmatrix} \tilde{F}_{t-1}^{k_q,f} \\ F_{t-1}^{k_q,s} \\ \tilde{F}_{t-1}^{k_m,f} \\ F_{t-1}^{k_m,s} \\ \tilde{F}_{t-1}^{k_w,f} \\ F_{t-1}^{k_w,s} \\ F_{t-1}^{k_d} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ U_t \end{bmatrix} \tag{7}$$

Equation (3)-(7) complete the model and the estimation is done by the EM algorithm. The consistency properties of the estimator are studied in Doz et al (2012).

5 Data

We consider 29 macroeconomic series, which include components of output, labor market indicators, price indices, surveys on the economic outlook, equity and commodity price indices. Table 2 gives the name of the variables, their frequencies, the publication delay in number of days to the reference period, transformation of the data and whether they are stock or flow variables. After the transformation indicated, variables then get standardized to have zero mean and unit standard deviation.

There are a few remarks on the choice of the data series. First, depending on the objective of forecast (in this case employment and initial claims), I choose only the real variables, as nominal variables do not add too much predictive power if the target of forecast is real. Second, I choose only the 'headline' variables. By this I mean instead of including all the subcategories, I tend to only include the aggregates. For instance, I use industrial production total index but not the sectoral disaggregates. Giannone, Reichlin and Small (2008) use about 200 variables and show that disaggregates do not add too much marginal predictive power. Bańbura and Modugno (2010) and Bańbura, Giannone and Reichlin (2011) analyse the marginal influence of disaggregate data on the nowcast precision and show that it is minimal, which supports the argument that the markets only focus on the headline of each report. The same author also show that inclusion of these disaggregates does not deteriorate the forecast result. This, in turn, supports the robustness of the factor model to data selection. Third, only monthly, weekly and daily variables are used in the model while one important quarterly variable, GDP, is left out. This is due to the consideration of quarterly variables might contain too many missing values at a daily frequency and therefore we might risk inaccurate estimation of the parameters, while someone might have in mind, instead, Okun's Law that tells us when an economy's unemployment rate falls by 1 %, its GNP rises by 3 %, which might be a reason to include GDP growth.

In terms of the query data, regardless the various issues that have been discussed in Section 3, in this paper I use the same series as in Choi and Varian (2012) and process the data in the same way, aiming to separate these issue from nowcasting. To be precise, I downloaded the two Google Trends series `'Jobs'` and `'Welfare & Unemployment'` on Dec 25, 2015 from one single IP address and seasonally adjusted them using the `stl` function with the same smoothing parameters as Choi and Varian (2012). These are the series that finally enter the model.

6 Empirical Results

6.1 Forecasting Initial Claims with a Dynamic Factor Model

Now I show the forecast performance of initial claims (FRED ticker: ICSA) from the daily dynamic factor model, with or without the Google Trends data. The forecast exercise is designed in the following

way: the forecast starts two weeks before the reference week and the backcast continues till one week after the reference week as the initial claims data will be released the following Thursday. In the model with Google Trends data, I evaluate the forecast on Mondays when Google Trends data are released and on Thursdays when initial claims are released, while in the model without, we evaluate the forecast only on Thursday when initial claims are released. Therefore we will have a two-step-ahead forecast (W(-2)MON, W(-2)THU), a one-step-ahead forecast (W(-1)MON, W(-1)THU), a nowcast (W(0)MON, W(0)THU) and a backcast (W(+1)MON, W(+1)THU). Table 3 shows the root mean forecast squared error (RMSE) of these two specifications and Figure 8 is the bar chart of this result with time on the x coordinate. Table 3 also shows the RMSE of an AR(1) model and a random walk (RW) with drift on Thursdays when initial claims data is released.

To further explain why search data do not add much predictive power, we need to further look through the evolvement of our forecast. Bańbura and Modugno (2014) denote Ω_v as the information set available at vintage v, and $I_{v+1} = \Omega_{v+1} \backslash \Omega_v$ as the updated information between v and $v+1$. Then from v to $v+1$ the revision to the forecast of a variable's realization at time t, y_t, is given by

$$\underbrace{E(y_t|\Omega_{v+1})}_{\text{new forecast}} = \underbrace{E(y_t|\Omega_v)}_{\text{old forecast}} + \underbrace{E(y_t|I_{v+1})}_{\text{revision}}$$

and the revision part can be further decomposed as

$$E(y_t|I_{v+1}) = E(y_t I_{v+1}) E(I_{v+1} I'_{v+1})^{-1} I_{v+1}$$

This is saying that the revision depends on two components, news and weight. News is defined as the difference between the realization and the projection on the previous vintage, that is to say, the news is the part that has not been captured by the model. And weight is the correlation between news and the forecast target. Basically these two measures can tell us why or why not some certain variable makes a difference to the forecast: in case that a variable leads to a large revision to the forecast, it is either because the variable contains news that is large in its magnitude, i.e. this variable provides very important new information, even on top of other variables, or because the news is very correlated with the forecast target and therefore even small surprise can lead to a large revision to the forecast. Figure 10 and Figure 11 show the news and weights of some selected variables. In these two graphs, both news and weights are adjusted by variables' standard deviations so that the bars are comparable. From Figure 10 we can see that the magnitude of news are similar across all the selected variables while Figure 11 gives us quite a different picture: variables, such as Industrial Production Index, Unemployment rate, employment, have rather big weights and therefore the impact of these variables on revision to the forecast is large; variables, such as the search series `Jobs` and `Welfare & Unemployment`, oil price, have smaller weights, which are less than 0.01. It is interesting to see that in terms of weight, search data is similar to financial

data. This tells us the reason why search data fail to improve the forecast because though they contain significant amount of surprise the surprise is not very correlated to the forecast target itself. This result is robust to other specifications of the model: removing all the financial variables, or putting real and nominal (in this case only financial variables) in separate blocks.

6.2 Forecasting Employment with a Dynamic Factor Model

It is also interesting to discern whether the Google Trends data will improve the forecast of another key variable, employment (FRED ticker: PAYEMS). The forecast starts from one month before the reference month and the backcast continue till one month after the reference month as the employment data will be released seven days after the reference month. And we do the evaluation on the 7th, 14th, 21st and 28th of a month to look closely at the impact of information release. Figure 9 shows the root mean forecast squared error (RMSE) of these two specifications. Similarly to the result of initial claims, the specifications yields similar forecast performance.

7 Conclusion

This paper studies the usefulness of Google Trends in forecasting the US weekly jobless claims and monthly employment payrolls. To assess, particularly, the usefulness of these search data in presence of other conventional macroeconomic data, the econometric framework used is a dynamic factor model that can take into account (i) a large panel, (ii) mixed frequencies and (iii) non-synchronized publication lags. I show that at least in the US Google Trends data do not improve the forecast accuracy of initial claims and employment significantly.

This result contradicts the general conclusion in the literature, however, the key difference between my approach and the ones prevailing in the literature is that instead of solely considering search data in over-simplified econometric models, I embed them in a large dataset, which is very conventional in todays forecast practice of central banks and markets, and assess their additional predictive power.

In turn, this calls for our reflection on the potential usefulness of search data. Typically, for countries whose data quality is good and publication delay is moderately, the US being a leading example, search data might not help too much. It is hard to expect that these new data sources will be better than national statistical agencies. But for countries, whose data quality is poor or where there are very few forward-looking variables (surveys) and data publication is subject to severe delay, search data might provide extra information and its timely publication can be beneficial in forecasting or even cross-validating the real economic data. This is particularly true in emerging market economics. For instance, Greek unemployment is published about three months after the end of the reference month. Therefore applying the same methodology detailed in this paper to some emerging market economies might give encouraging

results. Another potential use is that, in the similar spirit as Henderson et al (2012), search data may reckoned as a measure of some economic activity, with measurement error, and together with other measures also subject to measurement error, they may help improve data quality, given the two errors are orthogonal.

This paper also raises a few questions for future research. First, variable selection could be more automatic. Promising directions are the following: first, Google Correlate selects the words whose search intensity is most correlated to the interested series, but it will still require some judgment to remove the meaningless words and then further compress the data as done in Varian and Scott (2014). Second, some linguistic literature could be borrowed to pre-select the choices of words.

Second, to model seasonality in a factor model, or more generally, in any model used for forecast purpose, becomes necessary. New data can be seasonal, while traditional seasonality adjustment process might not be applicable to be applied to these data. Some countries statistical agencies produce only seasonally unadjusted variables. This often requires us pre-adjust the seasonal series or model them with other series. Despite an old topic, seasonality has regained some attention recently (Wright, 2013; Manski, 2014) as the recent financial crisis causes distortion to the estimation of seasonal factors using X-ARIMA type of procedure. Hence explicitly modelling the seasonality component seems a better approach.

References

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. German Council for Social and Economic Data (RatSWD) Research Notes, (41).

Bai, J. (2003). Inferential theory for factor models of large dimensions.Econometrica, 135-171.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. Econometrica, 70(1), 191-221.

Baker, S. R., & Fradkin, A. (2011). What drives job search? Evidence from Google search data. Evidence from Google Search Data (April 15, 2011).

Baker, S. R., & Fradkin, A. (2014). The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data. Available at SSRN 2251548.

Bańbura, M. M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow. Handbook of Economic Forecasting, 2, 195-237.

Bańbura, M., & Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. Journal of Applied Econometrics, 29(1), 133-160.

Bańbura, M., & Rünstler, G. (2011). A look into the factor model black box: publication lags and the role of hard and soft data in forecasting GDP.International Journal of Forecasting, 27(2), 333-346.

Billari, F., D'Amuri, F., & Marcucci, J. (2013). Forecasting births using google. In Annual Meeting of the Population Association of America.

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. Journal of Forecasting, 32(4), 289-298.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. Google Inc, 1-5.

Choi, H., & Varian, H. (2012). Predicting the present with google trends.Economic Record, 88(s1), 2-9.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1), 3-73.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. The Journal of Finance, 66(5), 1461-1499. DAmuri & Marcucci (2009), Google it! Forecasting the US unemployment rate with a Google job search index. ISER WP 2009-32.

D'Amuri, F., & Marcucci, J. (2012). The predictive power of Google searches in forecasting unemployment. Bank of Italy Temi di Discussione (Working Paper) No, 891.

Della Penna, N., & Huang, H. (2010). Constructing consumer sentiment index for US using Google searches (No. 2009-26).

Doz, C., Giannone, D., & Reichlin, L. (2012). A quasimaximum likelihood approach for large, approximate dynamic factor models. Review of economics and statistics, 94(4), 1014-1024.

Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. Journal of Econometrics, 164(1), 188-205.

Fondeur, Y., & Karamé, F. (2013). Can Google data help predict French youth unemployment?. Economic Modelling, 30, 117-125.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The Generalized Factor Model: Identification and EstimationV. The Review of Economics and Statistics.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2004). The generalized dynamic factor model consistency and rates. Journal of Econometrics, 119(2), 231-255.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. Journal of Monetary Economics,55(4), 665-676.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012-1014.

Goel, S., Hofman, J., Lahaie, S. Pennock, D., and Watts, D., Predicting consumer behavior with web search, PNAS Early Edition (2010).

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. The Journal of Economic and Social Measurement, 36(3).

Harvey, A. C. (1990). Forecasting, structural time series models and the Kalman filter. Cambridge university press.

Kholodilin, K. A., Podstawski, M., & Siliverstovs, B. (2010). Do Google searches help in nowcasting private consumption? A real-time evidence for the US. KOF Swiss Economic Institute Working Paper, (256).

Koop, G., & Onorante, L. (2013). Macroeconomic nowcasting using Google probabilities. mimeo.

Koopman, S. J., & Harvey, A. (2003). Computing observation weights for signal extraction and filtering. Journal of Economic Dynamics and Control, 27(7), 1317-1333.

Kristoufek, L. (2013). Can Google Trends search queries contribute to risk diversification?. Scientific reports, 3.

Kulkarni, R., Haynes, K., Stough, R., and Paelinck, J., Forecasting housing prices with Google econometrics, George Mason University School of Public Policy Research Paper 10 (2009).

Manski, C. (2014), "Communicating Uncertainty in Official Economic Statistics," National Bureau of Economic Research Working Paper 20098.

Mariano, R. S., & Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. Journal of Applied Econometrics, 18(4), 427-443.

Müller, U., & Watson, M. W. (2013). Measuring Uncertainty about Long-Run Prediction (No. w18870). National Bureau of Economic Research.

Müller, U. K., & Watson, M. W. (2015). Low-Frequency Econometrics (No. w21564). National

Bureau of Economic Research.

Pissarides, C. A. (2000). Equilibrium Unemployment Theory. MIT Press Books, 1.

Schmidt, T., & Vosen, S. (2012). Using Internet data to account for special events in economic forecasting. Ruhr economic paper, (382).

Scott, S. L., & Varian, H. (2014). Bayesian variable selection for nowcasting economic time series. In Economic Analysis of the Digital Economy. University of Chicago Press.

Song, H., Pan, B., & Ng, D. (2009). Forecasting demand for hotel rooms with search engine query volume data. College of Charleston Working Paper.

Stevenson, B. (2008). The Internet and job search (No. w13886). National Bureau of Economic Research.

Suhoy, Tanya (2009), Query Indices and a 2008 Downturn. Bank of Israel Discussion Paper (06).

Vanderkam, D., Schonberger, R., Rowley, H., & Kumar, S. Nearest Neighbor Search in Google Correlate.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. Journal of Forecasting, 30(6), 565-578.

Wu, L., & Brynjolfsson, E. (2014). The future of prediction: How Google searches foreshadow housing prices and sales. In Economic Analysis of the Digital Economy. University of Chicago Press.

Evaluation period		RW	RW with drift	AR(1)	AR(1) Aug.(a)	AR(1) Aug.(b)	AR(1) Aug.(c)
from	to	MAE	1-ratio	1-ratio	1-ratio	1-ratio	1-ratio
03 Nov 2007	02 Jul 2011	0.0313	1.0038	1.1117	1.1778	1.1750	1.1562
01 Dec 2007	30 Jun 2009	0.0312	0.9982	1.2754	1.1024	1.0630	1.0116
01 Mar 2009	01 May 2009	0.0301	1.0013	1.0182	0.7956	0.8541	0.9806
01 Dec 2009	01 Feb 2010	0.0354	0.9980	1.0072	0.8827	0.9390	1.1939
15 Jul 2010	01 Oct 2010	0.0255	1.0037	0.9911	0.9591	0.9676	0.9805
01 Jan 2011	01 May 2011	0.0516	1.0007	0.9959	0.9894	0.9691	0.9816
from	to	RMFSE	1-ratio	1-ratio	1-ratio	1-ratio	1-ratio
03 Nov 2007	02 Jul 2011	0.0402	1.0017	1.0996	1.1214	1.1187	1.1119
01 Dec 2007	30 Jun 2009	0.0413	0.9986	1.2139	1.0421	1.0241	0.9865
01 Mar 2009	01 May 2009	0.0376	1.0085	1.0132	0.7637	0.8063	0.9880
01 Dec 2009	01 Feb 2010	0.0423	0.9987	1.0085	0.9750	0.9742	1.1203
15 Jul 2010	01 Oct 2010	0.0296	1.0002	1.0032	1.0351	1.0120	1.0497
01 Jan 2011	01 May 2011	0.0586	0.9998	0.9976	0.9664	0.9519	0.9784

Table 1: Baseline forecast evaluation: 1-step-ahead, sample= 10 JAN 2004 - 02 JUL 2011, rolling window = NO. (a) data used in Choi and Varian (2012), with s.window=periodic. (b) data used in Choi and Varian (2012), with s.window=7. (c) data available on Dec 25, 2015, first cut to 10 JAN 2004 - 02 JUL 2011 then deseasonalized, with s.window='periodic'.

No.	Variable	Frequency	Publication delay	Transformation	Stock or Flow
1	Industrial Production Index	M	14	1	F
2	Capacity Utilization: Total Industry	M	15	2	F
3	ISM Manufacturing: PMI Composite Index	M	3	2	F
4	Real Disposable Personal Income	M	29	1	F
5	Civilian Unemployment Rate	M	7	2	F
6	All Employees: Total Nonfarm Payrolls	M	7	1	F
7	Real Personal Consumption Expenditures	M	29	1	F
8	Housing Starts: Total: New Privately Owned Housing	M	19	1	F
9	New One Family Houses Sold: United States	M	26	1	F
10	Manufacturers' New Orders: Durable Goods	M	27	1	F
11	Producer Price Index by Commodity for Finished Goods	M	13	1	S
12	Consumer Price Index for All Urban Consumers: All Items	M	14	1	S
13	Exports of Goods and Services, Balance of Payments Basis	M	43	1	F
14	Imports of Goods and Services: Balance of Payments Basis	M	43	1	F
15	PA Fed Manufacturing Business Outlook Survey: general activity	M	-10	2	F
16	Conference Board consumer confidence index	M	-5	2	F
17	Real Retail and Food Services Sales	M	14	1	F
18	Value of Manufacturers' Total Inventories for Durable Goods Industries	M	27	1	F
19	Value of Manufacturers' Unfilled Orders for Durable Goods Industries	M	27	1	F
20	Bloomberg consumer comfort index	W	4	2	F
21	Chicago Fed National Financial Conditions Index	W	3	2	F
22	Initial Claims	W	4	1	F
23	Continued Claims (Insured Unemployment)	W	9	1	S
24	Covered Employment	W	14	2	F
25	Google search category: jobs	W	1	2	F
26	Google search category: welfare & unemployment	W	1	2	F
27	Crude Oil: West Texas Intermediate (WTI) - Cushing, Oklahoma	D	1	1	S
28	10-Year Treasury Constant Maturity Rate	D	1	2	S
29	3-Month Treasury Bill: Secondary Market Rate	D	1	2	S
30	Trade Weighted Exchange Index: Major Currencies	D	1	2	S
31	S&P 500	D	1	1	S

Table 2: List of variables. M=monthly, W=weekly, D=daily. Publication delay is given in the number of days compared to the reference period. For example, Industrial Production Index is published after 14 days after the referred month, i.e. May's datum is published on 14th June. Transformation: 1=log difference, 2=difference. F=flow variable, S=stock variable.

	W(-2)MON	W(-2)THU	W(-1)MON	W(-1)THU	W(0)MON	W(0)THU	W(+1)MON	W(+1)THU
with Google	3.904524	3.929212	3.921109	3.914779	3.922659	3.894901	3.895136	3.844584
without Google		3.929678		3.916533		3.895516		3.844744
RW		5.363574		4.902846		4.019929		
AR(1)		7.397058		6.030544		4.420277		

Table 3: RMSE: initial claims (ICSA). The root mean forecast squared error from a dynamic factor model with and without Google Trends data, together with a random walk without drift and an AR(1) model. Sample=01FEB2004-02JUL2011, evaluation=03NOV2007-02JUL2011. Figure 8 result.

	M(-1)D7	M(-1)D14	M(-1)D21	M(-1)D28	M0D7	M0D14	M0D21	M0D28	M(+1)D7	M(+1)D14	M(+1)D21	M(+1)D28
with Google	0.1654	0.1505	0.1537	0.1637	0.1675	0.1430	0.1454	0.1444	0.1162	0.0967	0.0982	0.1002
without Google	0.1657	0.1507	0.1539	0.1643	0.1675	0.1431	0.1456	0.1448	0.1162	0.0964	0.0979	0.1001

Table 4: RMSE: employment. The root mean forecast squared error from a dynamic factor model with and without Google Trends data, together with a random walk without drift and an AR(1) model. Sample=01FEB2004-27FEB2015, evaluation=28NOV2008-27FEB2015. Figure 9 result.

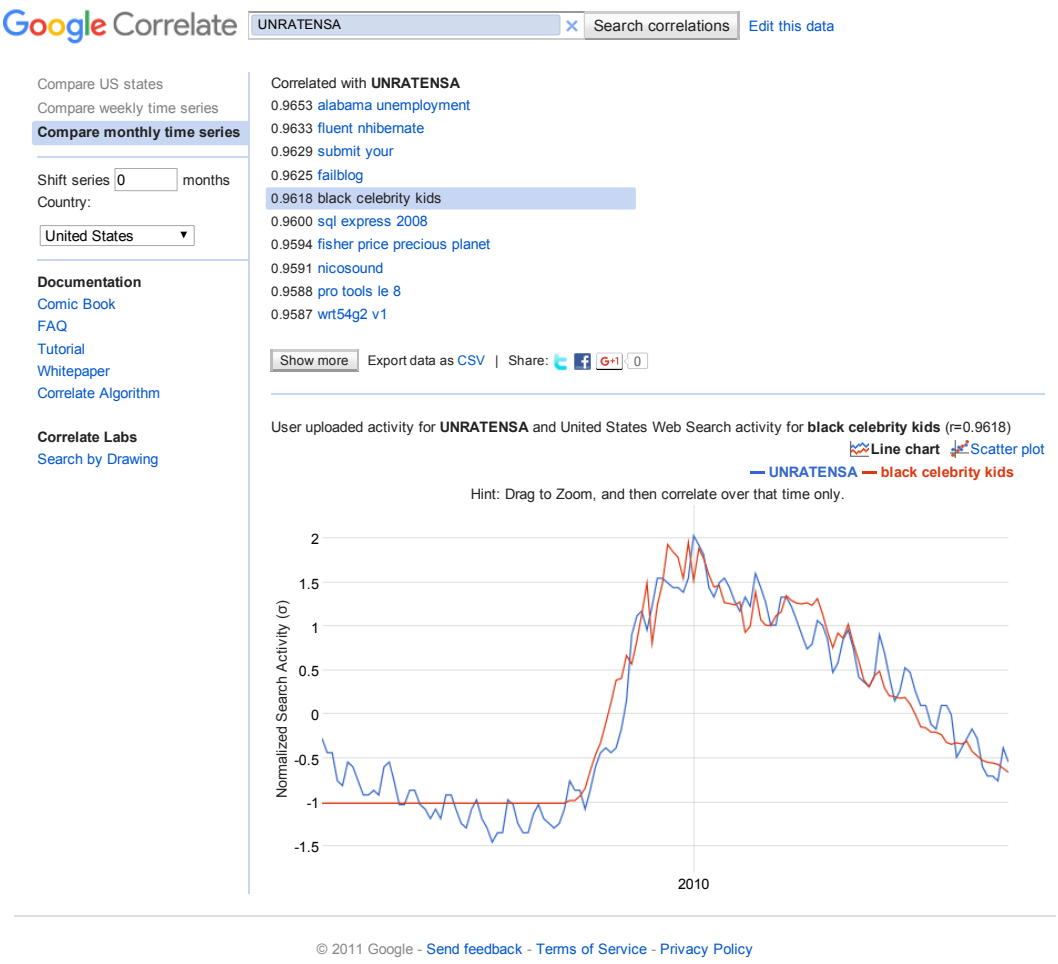

Figure 1: Google Correlate: Unemployment rate, seasonally unadjusted

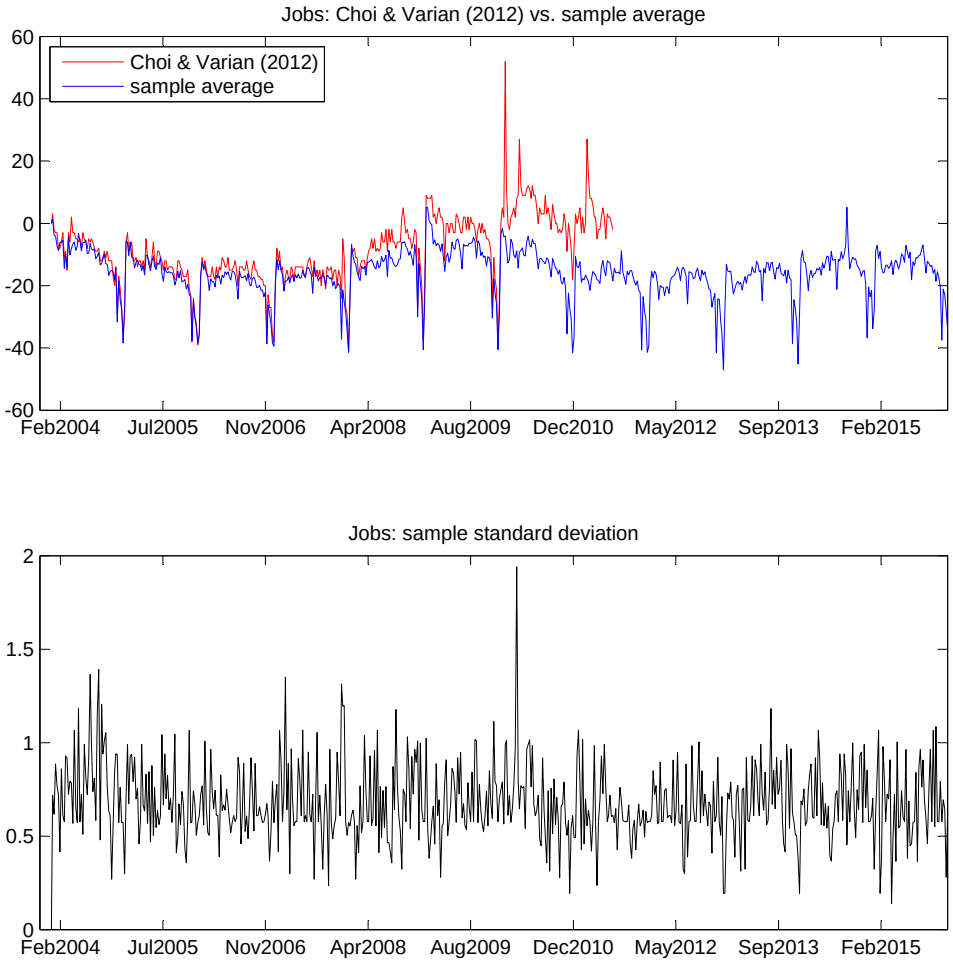

Figure 2: Google Trends: Jobs. Upper panel: data used in Choi and Varian (2012) and the average of 50 samples available on 25DEC2015. Lower panel: standard deviation of the 50 samples.

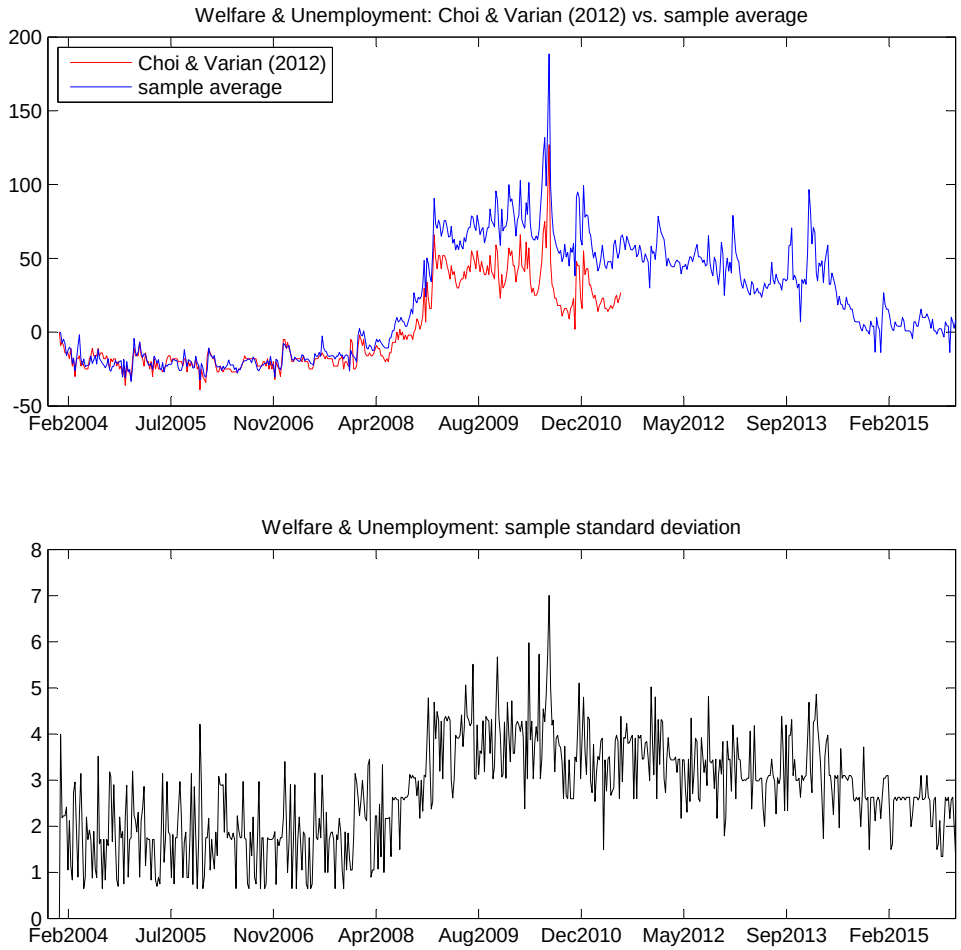

Figure 3: Google Trends: Welfare & Unemployment. Upper panel: data used in Choi and Varian (2012) and the average of 50 samples available on 25DEC2015.

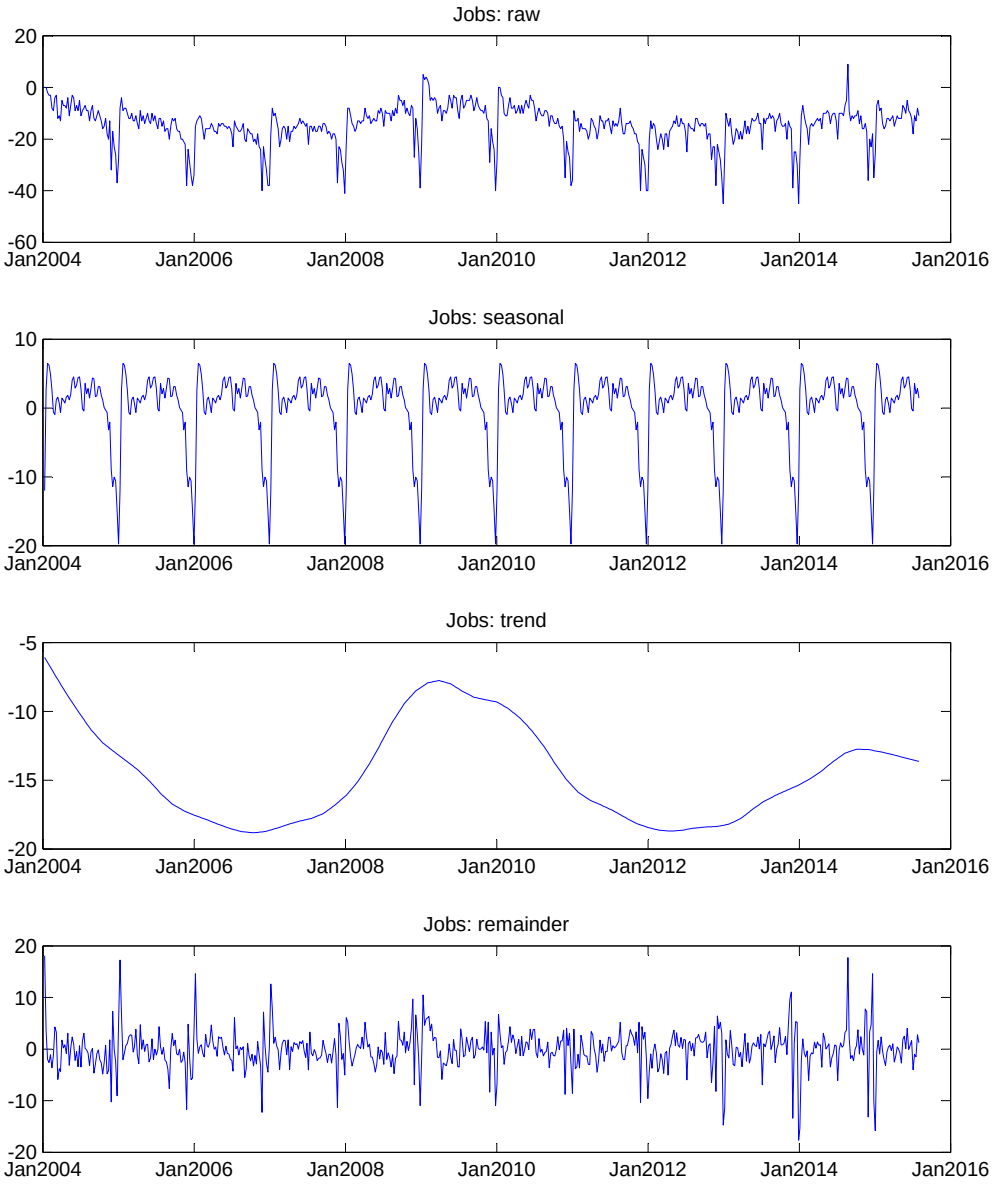

Figure 4: Jobs: seasonally adjusted with stl, s.window='periodic'

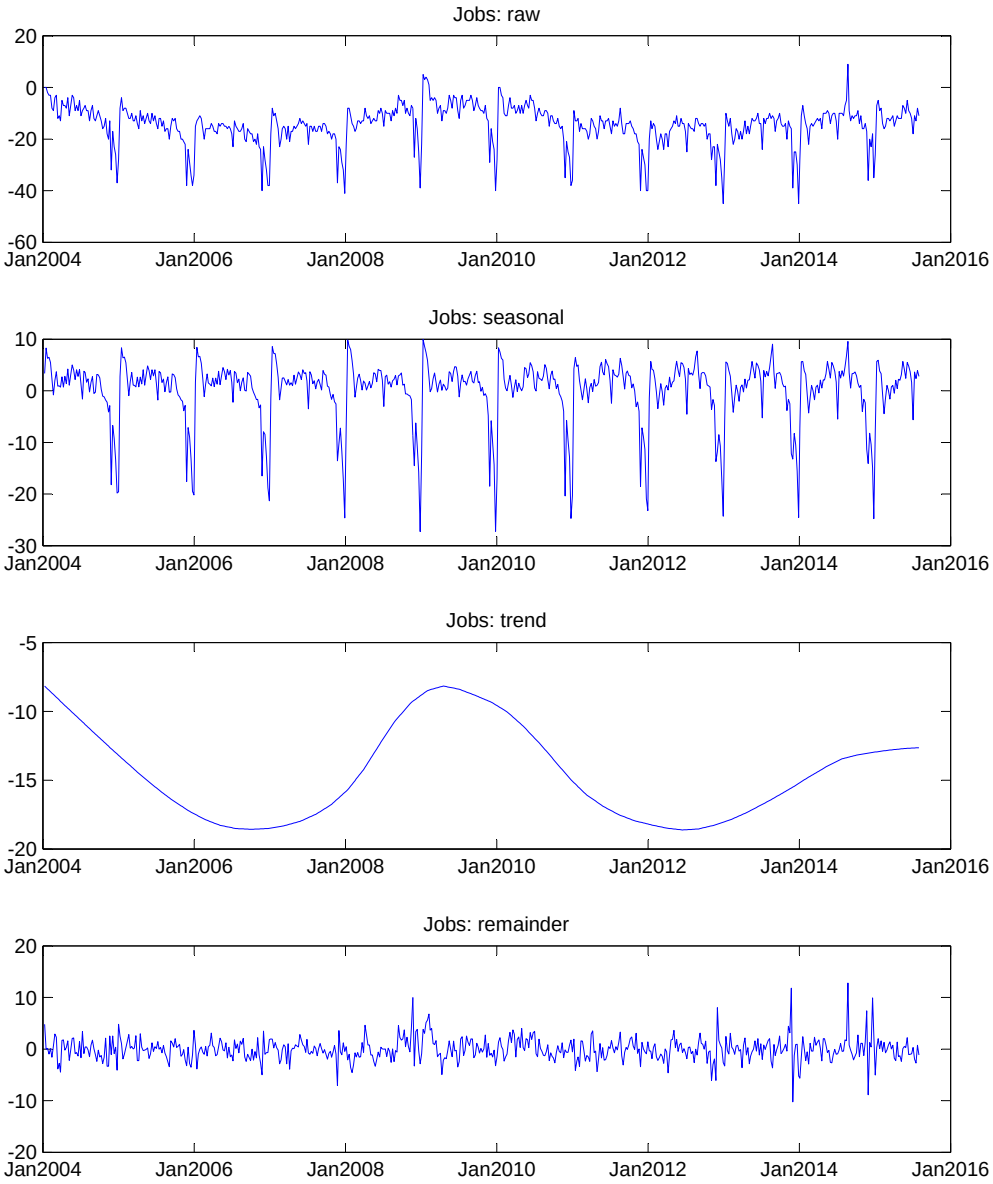

Figure 5: Jobs: seasonally adjusted with stl, s.window=7

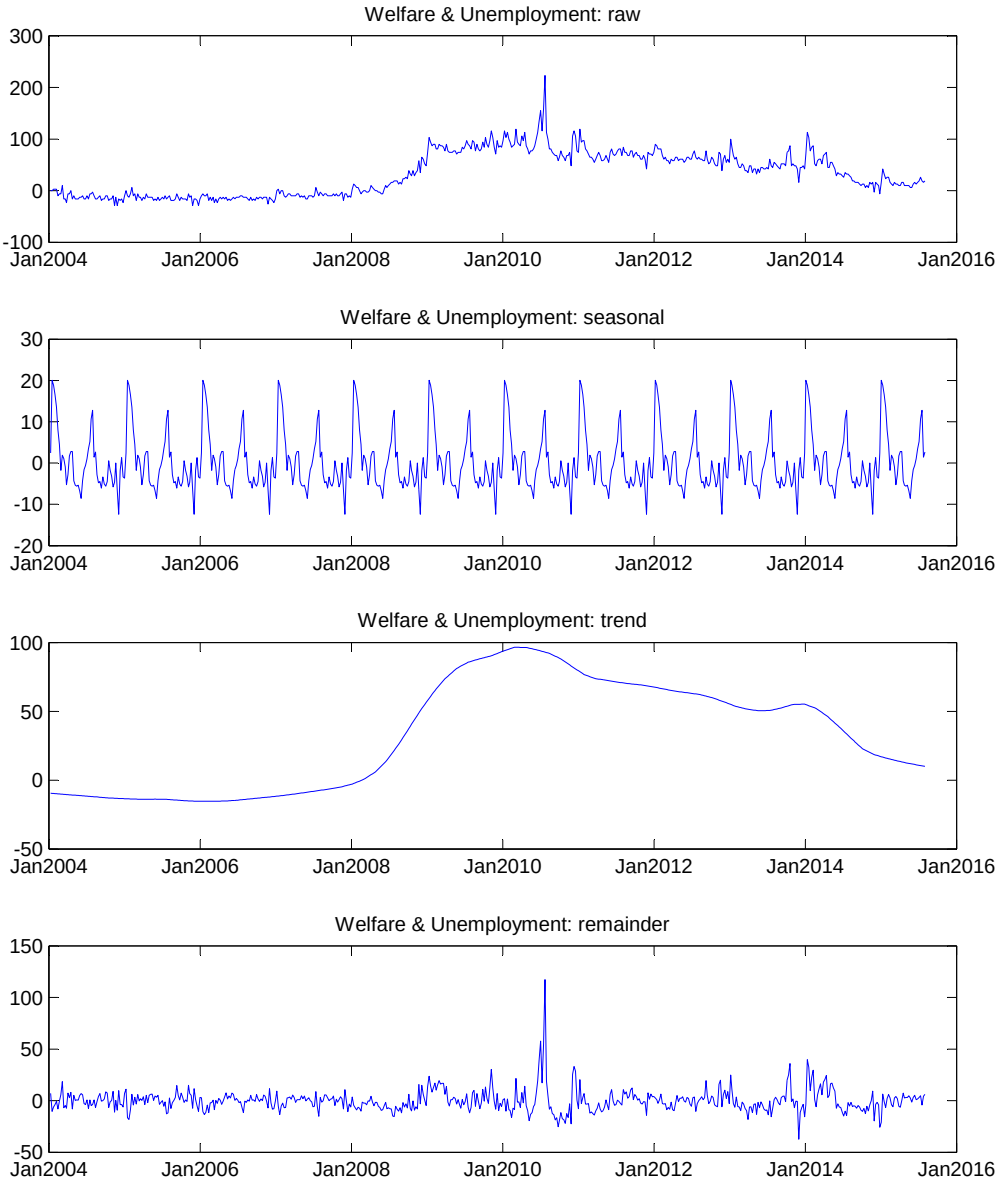

Figure 6: Welfare & Unemployment: seasonally adjusted with stl, s.window='periodic'

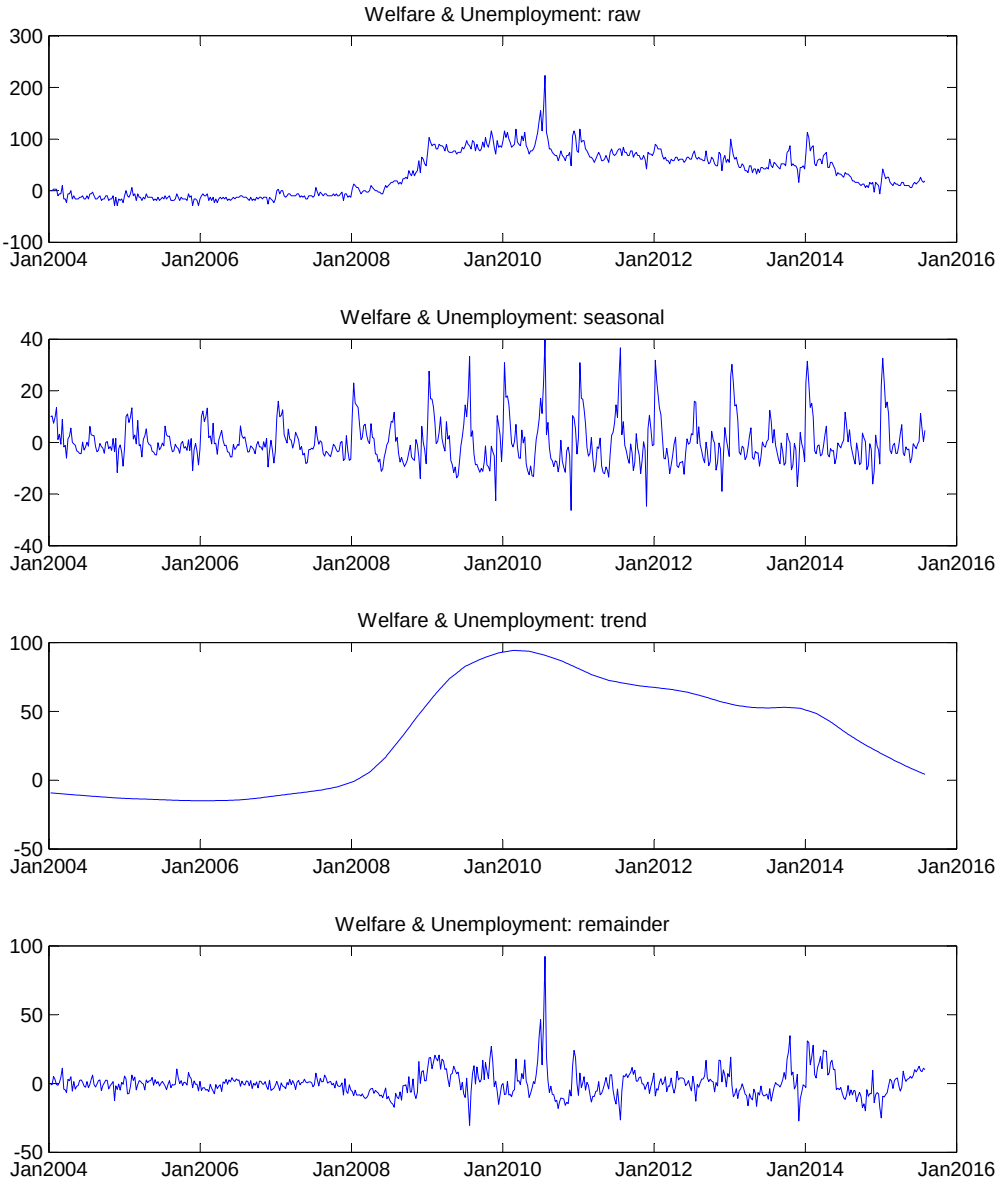

Figure 7: Welfare & Unemployment: seasonally adjusted with stl, s.window=7

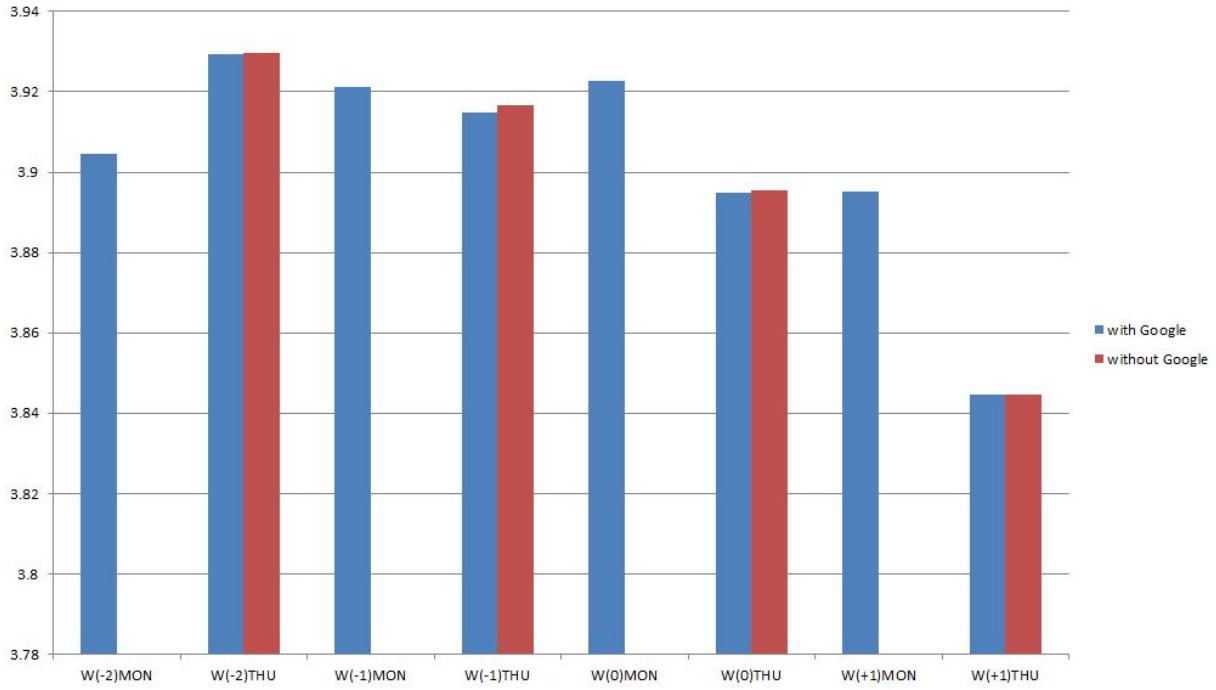

Figure 8: RMSE: ICSA. Sample=01FEB2004-02JUL2011, evaluation=03NOV2007-02JUL2011. The RMSE of a random walk model is 4.019929 and an AR(1) 4.420277, over the same period for one-step-ahead forecast.

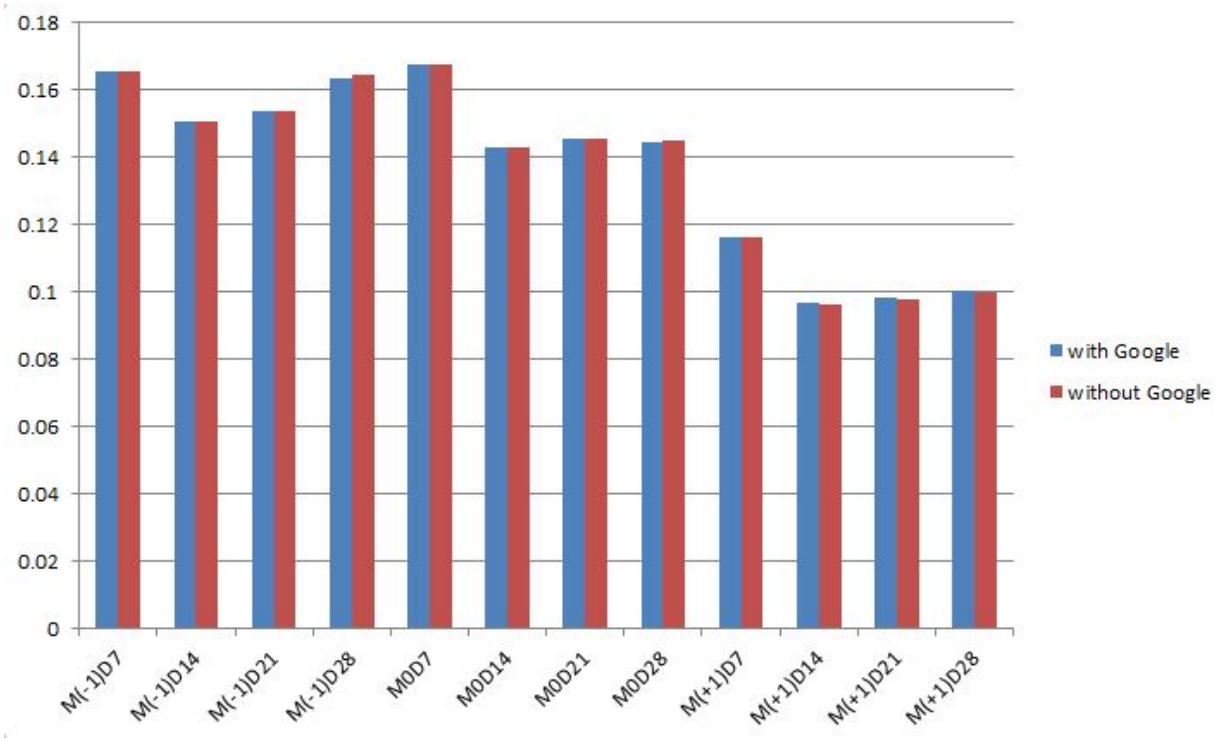

Figure 9: RMSE: employment. Sample=01FEB2004-27FEB2015, evaluation=28NOV2008-27FEB2015.

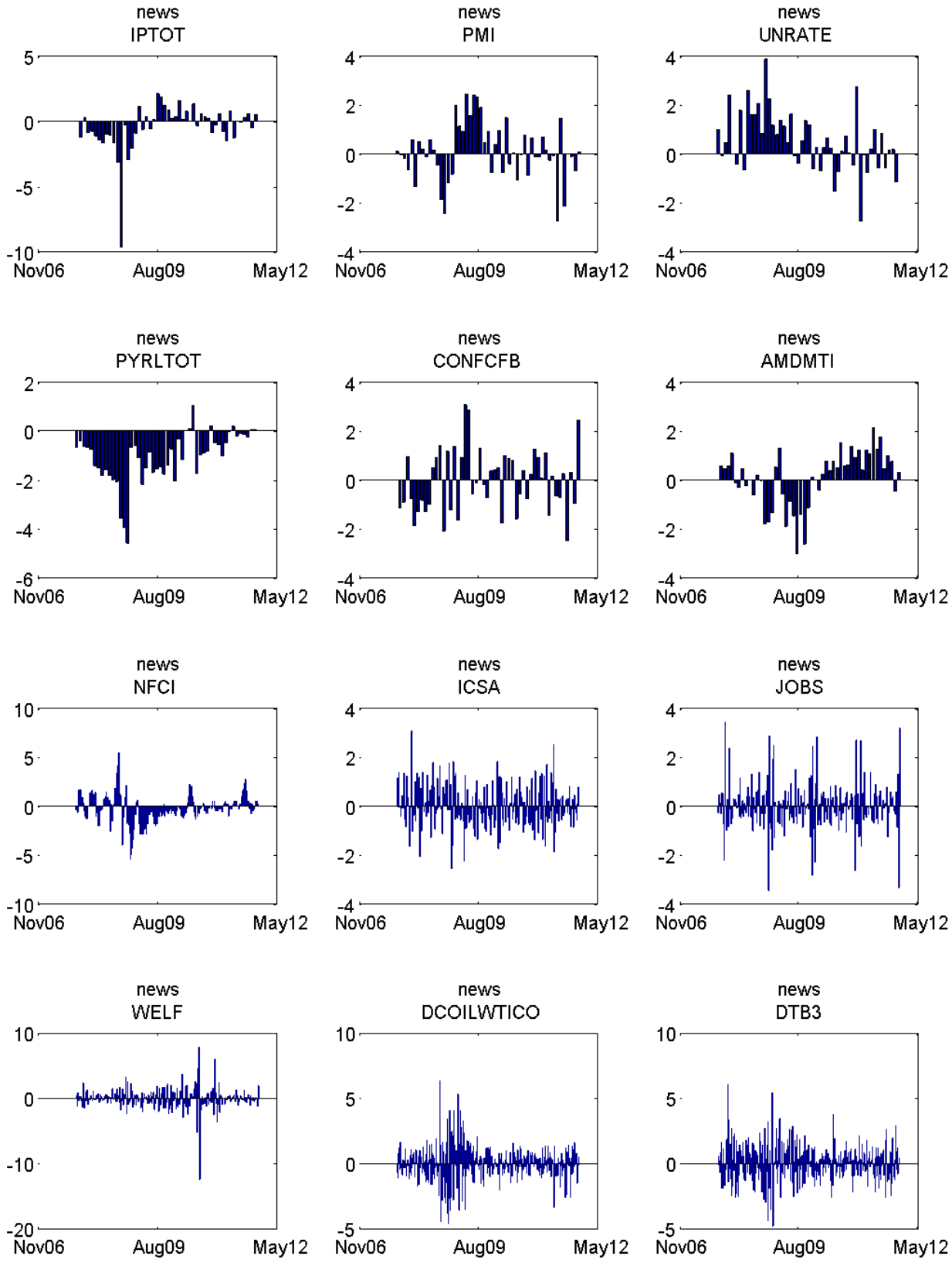

Figure 10: News of selected variables. Standardized. Sample=01FEB2004-02JUL2011, evaluation=03NOV2007-02JUL2011.

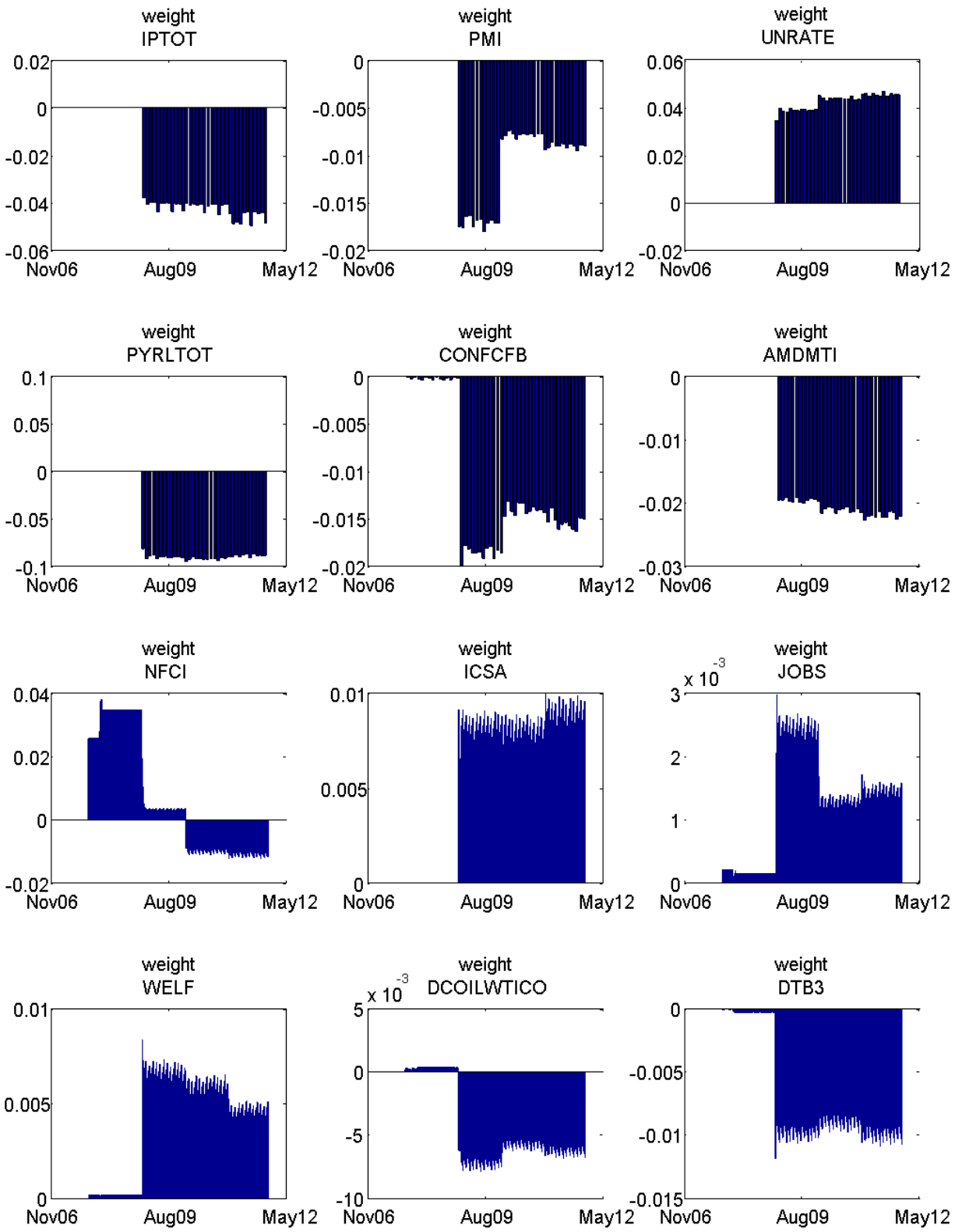

Figure 11: Weights of selected variables. Standardized. Sample=01FEB2004-02JUL2011, evaluation=03NOV2007-02JUL2011.

APPENDICE

A Details on the Factor Model with Mixed Frequencies and Missing Data

A.1 Temporal Aggregation of Stock and Flow variables

Denote raw series as z_t^k, and the latent series as z_t, where k is the number of intervals between the observed z_t^k.

Stock variables For a stock variable, the observed variable equals the latent variable

$$z_t^k = z_t$$

and if y_t^k is the difference of z_t^k[7], then

$$\begin{aligned}
y_t^k &= z_t^k - z_{t-k}^k \\
&= z_t - z_{t-k} \\
&= (z_t - z_{t-1}) + (z_{t-1} - z_{t-2}) + \cdots + (z_{t-k+1} - z_{t-k}) \\
&= \Delta z_t + \Delta z_{t-1} + \cdots + \Delta z_{t-k+1} \\
&=: y_t + y_{t-1} + \cdots + y_{t-k+1} \\
&= \sum_{i=0}^{k-1} \underbrace{1}_{=w_i^s} \cdot y_{t-i}
\end{aligned}$$

Flow variables For a flow variable, the observed variable equals the sum of latent variable over the past interval

$$z_t^k = z_t + z_{t-1} + \cdots + z_{t-k+1}$$

[7]The reason we work with y_t^k instead of z_t^k is because z_t^k is oftentimes nonstationary, and it must be transformed to a stationary series. However, whether z_t^k is a stock or flow variable will influence the relation between y_t^k and y_t. We thank Alberto Caruso for bringing up this point.

and if y_t^k is the difference of z_t^k, then

$$
\begin{aligned}
y_t^k &= z_t^k - z_{t-k}^k \\
&= (z_t + z_{t-1} + \cdots + z_{t-k+1}) - (z_{t-k} + z_{t-k-1} + \cdots + z_{t-k-k+1}) \\
&= (z_t - z_{t-k}) + (z_{t-1} - z_{t-k-1}) + \cdots + (z_{t-k+1} - z_{t-k-k+1}) \\
&= [(z_t - z_{t-1}) + (z_{t-1} - z_{t-2}) + \cdots + (z_{t-k+1} - z_{t-k})] \\
&+ [(z_{t-1} - z_{t-2}) + (z_{t-2} - z_{t-3}) + \cdots + (z_{t-k} - z_{t-k-1})] \\
&+ \ldots \\
&+ [(z_{t-k+1} - z_{t-k}) + (z_{t-k} - z_{t-k-1}) + \cdots + (z_{t-k-k+2} - z_{t-k-k+1})] \\
&= (\Delta z_t + \Delta z_{t-1} + \cdots + \Delta z_{t-k+1}) \\
&+ (\Delta z_{t-1} + \Delta z_{t-2} + \cdots + \Delta z_{t-k}) \\
&+ \ldots \\
&+ (\Delta z_{t-k+1} + \Delta z_{t-k} + \cdots + \Delta z_{t-k-k+2}) \\
&= (y_t + y_{t-1} + \cdots + y_{t-k+1}) \\
&+ (y_{t-1} + y_{t-2} + \cdots + y_{t-k}) \\
&+ \ldots \\
&+ (y_{t-k+1} + y_{t-k} + \cdots + y_{t-k-k+2}) \\
&= \sum_{i=-k+1}^{k-1} \underbrace{(k - |i|)}_{=w_i^f} y_{t-k+1+i}
\end{aligned}
$$

Notice in the second line, there are k items in each parentheses; in the third line, there are k parentheses and it uses a standard trick to pair z_t with its k-order lag; the last line resembles the weights in spectral density [8]: y_{t-k+1} appears k times, while y_t and $y_{t-k-k+2}$ appear only once.

Notice the similarity and difference of these two aggregations. They are both sums of the past latent variables. We denote the weights in front as $w_i^{\cdot}$. However, for stock variables, the sum only goes from y_t back to y_{t-k+1} while for flow varibles, the sum goes from y_t back to $y_{t-k-k+2}$. Remember the period over which we observe the variables is k, so this will make a difference in the state-space representation for flow variables, as we will show next in Appendix A.2.

[8] The spectral density function of a zero-mean stationary time series with autocovariance function $\gamma(.)$ satisfying $\sum_{h=-\infty}^{+\infty} |\gamma(h)| < \infty$, is $f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} e^{-ih\lambda}\gamma(h)$, for $\lambda \in \mathbb{R}$. One can show $f_N(\lambda) := \frac{1}{2\pi N} E\left(\left|\sum_{r=1}^{N} X_r e^{-ir\lambda}\right|^2\right) = \frac{1}{2\pi N} \sum_{h=-\infty}^{+\infty} (N - |h|) e^{-ih\lambda}\gamma(h)$

A.2 State-space Representation

After sorting out the different between stock and flow variables, next we show how to write these less frequently observed variables into a state-space representation. And we aim at a lower dimensional state variable.

The prime idea turning the presumably large state space, as in Mariano and Murasawa (2003), into the current small state space, is to aggregate the daily factor and its lags into "quarterly", "monthly" or "weekly" factors and in turn the observables load on these aggregated factors. On the one hand, this reduces the dimension of the state variables and hence the parameterization. On the other hand, this avoids linear restrictions on the factor loadings, also as in Mariano and Murasawa (2003), which can be cumbersome.

Take a generic lower-frequency factor, $F_t^{k,\cdot}$ as an example. As is shown before, as observable variable $Y_t^{k,\cdot}$ is an aggregate of the latent variable Y_t and its lags, the corresponding factor $F_t^{k,\cdot}$ should also be an aggregate of the latent higher-frequency factor $F_t^{k_d}$, with the same weights. Namely,

$$\begin{aligned} F_t^{k,s} &= \sum_{i=0}^{k-1} \underbrace{1}_{=w_i^s} F_{t-i}^{k_d} \\ F_t^{k,f} &= \sum_{i=-k+1}^{k-1} \underbrace{(k-|i|)}_{=w_i^f} F_{t-k+1-i}^{k_d} \end{aligned}$$

Let us emphasize again that for the stock variable, the sum goes from t to $t-k+1$, while for the flow variable, the sum goes from t to $t-2k+2$. And these will cause a different manipulation of the state space. What is also important is, these factors are defined as such only when the corresponding variables are observed. How these factors are defined only the corresponding variables are not observed is at our discretion. Now we want to write the factor at time t, $F_t^{k,\cdot}$, into an AR(1) representation. For the stock variable

$$F_t^{k,s} = \sum_{i=0}^{k-1} w_i^s F_{t-i}^{k_d} = \sum_{i=1}^{k-1} w_i^s F_{t-i}^{k_d} + w_0^s F_t^{k_d} =: F_{t-1}^{k,s} + w_0^s F_t^{k_d}, \quad \text{if } Y_t^{k,s} \text{ observable}$$

for the flow variable,

$$F_t^{k,f} = \sum_{i=-k+1}^{k-1} w_i^f F_{t-k+1-i}^{k_d} = \sum_{i=-k+2}^{k-1} w_i^f F_{t-k+1-i}^{k_d} + w_{-k+1}^f F_t^{k_d} =: F_{t-1}^{k,f} + w_{-k+1}^f F_t^{k_d}, \quad \text{if } Y_t^{k,f} \text{ observable}$$

Notice the way we define F_{t-1} will ensure the finiteness of the processes. Otherwise, they would be nothing but a random walk and thus diverge. So the sequences should be "reset" to some initial values periodically. Since the period is just k, the intervals at which $y_t^{k,\cdot}$ is observed, the factor processes should be naturally reset at this frequency. Suppose at $t = k, 2k, 3k, \ldots$, $Y_t^{k,\cdot}$ is observed. For stock varibles, it

is natuaral to define

$$F_t^{k,s} = \begin{cases} w_0^s F_t^{k_d} & \text{if } t = 1, k+1, 2k+1, \ldots \\ F_{t-1}^{k,s} + w_0^s F_t^{k_d} & \text{otherwise.} \end{cases} \tag{8}$$

then

$$I_{2r} F_t^{k,s} - w_{k-1}^{k,f} F_t^{k_d} = \begin{cases} 0 F_{t-1}^{k,s}, \text{if } t = 1, k+1, 2k+1, \ldots \\ I_r F_{t-1}^{k,s}, \text{otherwise} \end{cases}$$

for stock variables and the weight matrix is defined as

$$W_t^{k,s} = -w_{k-1}^{k,s} I_r$$

and indicator matrix

$$\mathcal{I}_t^{k,s} = \begin{cases} 0, & \text{if } t = 1, k+1, 2k+1, \ldots \\ I_r, & \text{otherwise} \end{cases}$$

For flow variables, things are slightly more complex: we want to set the factors back to initial values every $2k$ periods but also want the observables to load on these factors every k periods. This requires an auxiliary state variable $\bar{F}_t^{k,f}$,

$$\bar{F}_t^{k,f} = \begin{cases} 0 & \text{if } t = 1, k+1, 2k+1, \ldots \\ \bar{F}_{t-1}^{k,s} + w_{R(k-t,k)+k}^s F_t^{k_d} & \text{otherwise.} \end{cases} \tag{9}$$

$\bar{F}_t^{k,f}$ is actually a partial sum series, reset to 0 at $t = nk + 1, n \in \mathbb{Z}$. It is illuminating to write out its values over one period:, k,

$$
\begin{aligned}
\bar{F}_1^{k,f} &= 0 \\
\bar{F}_2^{k,f} &= \bar{F}_1^{k,f} + 1F_2^{k_d} \\
\bar{F}_3^{k,f} &= \bar{F}_2^{k,f} + 2F_3^{k_d} \\
&\dots \\
\bar{F}_{k-1}^{k,f} &= \bar{F}_{k-2}^{k,f} + (k-2)F_{k-1}^{k_d} \\
\bar{F}_k^{k,f} &= \bar{F}_{k-1}^{k,f} + (k-1)F_k^{k_d} \\
\bar{F}_k^{k,f} &= 0 \\
&\dots
\end{aligned}
$$

and the actual factor that the observables will load on will be

$$
F_t^{k,f} = \begin{cases} \text{initial value, } F_1 & \text{if } t = 1 \\ \bar{F}_{t-1}^{k,f} + w_{k-1}^s F_t^{k_d} & \text{if } t = k+1, 2k+1, \dots \\ F_{t-1}^{k,f} + w_{R(k-t,k)}^s F_t^{k_d} & \text{otherwise.} \end{cases} \tag{10}
$$

Similary, we write out its values over on period, k,

$$
\begin{aligned}
F_{k+1}^{k,f} &= \bar{F}_k^{k,f} + kF_{k+1}^{k_d} \\
F_{k+2}^{k,f} &= F_{k+1}^{k,f} + (k-1)F_{k+2}^{k_d} \\
F_{k+3}^{k,f} &= F_{k+2}^{k,f} + (k-2)F_{k+3}^{k_d} \\
&\dots \\
F_{2k}^{k,f} &= F_{2k-1}^{k,f} + 1F_{2k}^{k_d} \\
F_{2k+1}^{k,f} &= \bar{F}_{2k}^{k,f} + kF_{2k+1}^{k_d} \\
&\dots
\end{aligned}
$$

From this, we can see $F_t^{k,f}$ loads exactly what we want for $t = k, 2k, 3k, \dots$. Denote

$$
\tilde{F}_t^{k,f} = \begin{pmatrix} F_t^{k,f} \\ \bar{F}_t^{k,f} \end{pmatrix}
$$

then

$$I_{2r}\begin{pmatrix} F_t^{k,f} \\ \bar{F}_t^{k,f} \end{pmatrix} + \begin{pmatrix} -w_{k-1}^{k,f} \\ 0 \end{pmatrix} F_t^{k_d} = \begin{pmatrix} 0 & I_r \\ 0 & 0 \end{pmatrix} \begin{pmatrix} F_{t-1}^{k,f} \\ \bar{F}_{t-1}^{k,f} \end{pmatrix} + 0, \text{if } t = 1, k+1, 2k+1, \ldots$$

$$I_{2r}\begin{pmatrix} F_t^{k,f} \\ \bar{F}_t^{k,f} \end{pmatrix} + \begin{pmatrix} -w_{R(k-t,k)}^{k,f} \\ -w_{R(k-t,k)+k}^{k,f} \end{pmatrix} F_t^{k_d} = I_{2r}\begin{pmatrix} F_{t-1}^{k,f} \\ \bar{F}_{t-1}^{k,f} \end{pmatrix} + 0, \text{otherwise}$$

and the weight matrix is defined as

$$W_t^{k,f} = \begin{cases} \begin{pmatrix} -w_{k-1}^{k,f} \\ 0 \end{pmatrix} & \text{if } t = 1, k+1, 2k+1, \ldots \\ \begin{pmatrix} -w_{R(k-t,k)}^{k,f} \\ -w_{R(k-t,k)+k}^{k,f} \end{pmatrix} & \text{otherwise.} \end{cases}$$

and indicator matrix

$$\mathcal{I}_t^{k,f} = \begin{cases} \begin{pmatrix} 0 & I_r \\ 0 & 0 \end{pmatrix} & \text{if} t = 1, k+1, 2k+1, \ldots \\ I_{2r} & \text{otherwise.} \end{cases}$$

Then the transition equation is

$$\begin{bmatrix} I_{2r} & 0 & 0 & 0 & 0 & 0 & W_t^{k_q,f} \\ 0 & I_r & 0 & 0 & 0 & 0 & W_t^{k_q,s} \\ 0 & 0 & I_{2r} & 0 & 0 & 0 & W_t^{k_m,f} \\ 0 & 0 & 0 & I_r & 0 & 0 & W_t^{k_m,s} \\ 0 & 0 & 0 & 0 & I_{2r} & 0 & W_t^{k_w,f} \\ 0 & 0 & 0 & 0 & 0 & I_r & W_t^{k_w,s} \\ 0 & 0 & 0 & 0 & 0 & 0 & I_r \end{bmatrix} \begin{bmatrix} \tilde{F}_t^{k_q,f} \\ F_t^{k_q,s} \\ \tilde{F}_t^{k_m,f} \\ F_t^{k_m,s} \\ \tilde{F}_t^{k_w,f} \\ F_t^{k_w,s} \\ F_t^{k_d} \end{bmatrix} = \begin{bmatrix} \mathcal{I}_t^{k_q,f} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathcal{I}_t^{k_q,s} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathcal{I}_t^{k_m,f} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathcal{I}_t^{k_m,s} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathcal{I}_t^{k_w,f} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathcal{I}_t^{k_w,s} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A \end{bmatrix} \begin{bmatrix} \tilde{F}_{t-1}^{k_q,f} \\ F_{t-1}^{k_q,s} \\ \tilde{F}_{t-1}^{k_m,f} \\ F_{t-1}^{k_m,s} \\ \tilde{F}_{t-1}^{k_w,f} \\ F_{t-1}^{k_w,s} \\ F_{t-1}^{k_d} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ U_t \end{bmatrix} \tag{11}$$

What is worth noticing is that only the dynamics of $F_t^{k_d}$ involves error term while the relation between the lower-frequency factor F_t^k , for $k = k_q, k_m, k_w$ and $F_t^{k_d}$ are exact, due to the recursive representation.

B Create Daily GI Data from Daily and Weekly

This part is based on the blog of http://erikjohansson.blogspot.fi/.

Oftentimes we want to utilize all the data available from the Google Trends and moreover we want

to compare the forecast performance using series of different frequencies. By default, if a requested time span is greater than three months, the Google Trends website will generate a weekly index, while if a requested time span is less or equal to three months, the Google Trends website will generate instead a daily index. This gives us the possibility of compiling a daily version of the index so that we can explore the index of both weekly and daily frequency. We illustrate this transformation with the word 'jobs'. First we download a set of daily index with no time overlapping. As mentioned before, Google only gives daily index for time span requested shorter or equal to three months, and therefore the datasets shall be produced every three months, e.g. from 01Jan2004 to 31Mar2004, from 01Apr2004 to 30Jun2004, from 01Jul2004 to 30Sep2004, from 01Oct2004 to 31Dec2004, etc.. The daily data look as follows:

Day	Daily index
18/01/2004	64
19/01/2004	86
20/01/2004	87
21/01/2004	88
22/01/2004	85
23/01/2004	76
24/01/2004	66
25/01/2004	70
26/01/2004	81
27/01/2004	82
28/01/2004	81
29/01/2004	78
30/01/2004	72
31/01/2004	61

Then we download the weekly index from 04Jan2004 to present. The weekly data look like this:

Week	Weekly index
2004-01-04 - 2004-01-10	36
2004-01-11 - 2004-01-17	36
2004-01-18 - 2004-01-24	35
2004-01-25 - 2004-01-31	32
2004-02-01 - 2004-02-07	33
2004-02-08 - 2004-02-14	31
2004-02-15 - 2004-02-21	32
2004-02-22 - 2004-02-28	32

We merge the weekly index with the daily, with the first day of the week (Sunday) as the merging variable. Take the weeks 18Jan2004-24Jan2004 and 25Jan2004-31Jan2004 as an example:

Day	Daily index	Weekly index
18/01/2004	64	35
19/01/2004	86	
20/01/2004	87	
21/01/2004	88	
22/01/2004	85	
23/01/2004	76	
24/01/2004	66	
25/01/2004	70	32
26/01/2004	81	
27/01/2004	82	
28/01/2004	81	
29/01/2004	78	
30/01/2004	72	
31/01/2004	61	

Then we create a column of 'adjustment factor', which equals the weekly index divided by the daily index if both are available for the day, and equals the last value if the weekly index is missing. Then 'adjusted daily index' is the product of the raw daily index and adjustment factor. This is what the daily index we need for the analysis.

Day	Daily index	Weekly index	Adj. factor	Adj. daily
18/01/2004	64	35	0.546875	35
19/01/2004	86		0.546875	47.03125
20/01/2004	87		0.546875	47.57813
21/01/2004	88		0.546875	48.125
22/01/2004	85		0.546875	46.48438
23/01/2004	76		0.546875	41.5625
24/01/2004	66		0.546875	36.09375
25/01/2004	70	32	0.457143	32
26/01/2004	81		0.457143	37.02857
27/01/2004	82		0.457143	37.48571
28/01/2004	81		0.457143	37.02857
29/01/2004	78		0.457143	35.65714
30/01/2004	72		0.457143	32.91429
31/01/2004	61		0.457143	27.88571