Tuhkuri, Joonas

Working Paper

Forecasting Unemployment with Google Searches

ETLA Working Papers, No. 35

Provided in Cooperation with:
The Research Institute of the Finnish Economy (ETLA), Helsinki

Suggested Citation: Tuhkuri, Joonas (2016) : Forecasting Unemployment with Google Searches, ETLA Working Papers, No. 35, The Research Institute of the Finnish Economy (ETLA), Helsinki

This Version is available at:
https://hdl.handle.net/10419/201250

ETLA Working Papers

No. 35

2 March 2016

Joonas Tuhkuri

Forecasting Unemployment with Google Searches

Suggested citation: Tuhkuri, Joonas (2.3.2016). "Forecasting Unemployment with Google Searches". ETLA Working Papers No 35. http://pub.etla.fi/ETLA-Working-Papers-35.pdf

ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY

ISSN-L 2323–2420
ISSN 2323–2420 (print)
ISSN 2323–2439 (online)

Forecasting Unemployment with Google Searches

Joonas Tuhkuri*

March 2, 2016

Abstract

Data on Google searches help predict the unemployment rate in the U.S. But the predictive power of Google searches is limited to short-term predictions, the value of Google data for forecasting purposes is episodic, and the improvements in forecasting accuracy are only modest. The results, obtained by (pseudo) out-of-sample forecast comparison, are robust to a state-level fixed effects model and to different search terms. Joint analysis by cross-correlation function and Granger non-causality tests verifies that Google searches anticipate the unemployment rate. The results illustrate both the potentials and limitations of using big data to predict economic indicators.

Keywords: Big Data, Google, Internet, Nowcasting, Forecasting, Unemployment
JEL Codes: C22, C53, C55, C82, E27

*<joonas.tuhkuri@helsinki.fi>, University of Helsinki, Department of Political and Economic Studies, P.O. Box 17 (Arkadiankatu 7), Helsinki 00014, Finland, and ETLA, the Research Institute of the Finnish Economy. I would like to thank Christian Gourieroux, Antti Kauhanen, Markku Lanne, Philip Oreopoulos, as well as Ari Ojansivu and Johanna Wahlroos at Google, and seminar participants at 2015 New Techniques and Technologies for Statistics Conference, 4th Baltic-Nordic Conference on Survey Statistics 2015, ETLA, University of Helsinki, and Aalto University for helpful comments and conversations. Part of the paper was written while I was visiting the Department of Economics at the University of Toronto. Financial support from the U.S. State Department (ARC Grant) and the University of Helsinki (International Movement Scholarship and Research Station Grant) is gratefully acknowledged. This research was finalized as a part of the ongoing collaboration of BRIE, the Berkeley Roundtable on the International Economy at the University of California at Berkeley, and ETLA. All errors and omissions are mine.

1 Introduction

There are over 100 billion searches on Google every month.[1] Could data from Google searches help predict the unemployment rate in the United States?

Traditional labor statistics are available with at least a one-month lag. A more timely estimate of the unemployment rate would be valuable. From a policy perspective, more accurate and timely knowledge could inform better labor market and monetary policy that would help workers—especially during an economic crisis.

But data on Google searches are publicly available in real time. Each search is someone expressing an interest in or demand for something (Brynjolfsson 2012). This information could help *nowcast* the present unemployment rate, which is unknown. Furthermore, Google search queries could be associated with future expectations and thus help *forecast* the future unemployment rate. Sudden changes in Google search activity could help identify sudden changes—the turning points—in the unemployment rate as well.

New large-scale and high-frequency data sets have been presented in the academic literature with the promise of being able to improve macroeconomic measurement (see, for example, Aruoba and Diebold 2010). Previously, early studies have shown that Internet search query data might help predict influenza epidemics (Ginsberg et al. 2009), video game sales (Goel et al. 2010), and housing market transactions (Wu and Brynjolfsson 2015). However, the data has only been used in a handful of studies.

More recently, several studies have suggested that Google search volumes could help predict the unemployment rate (Askitas and Zimmermann 2009; Choi and Varian 2012). The first studies on the topic, while certainly important, only point out the dataset's potential for forecasting. The literature has not developed much from its first explorations in Germany and in the U.S. The previous papers mostly report strong country-level correlations between relevant Google search volumes and the unemployment rate, while our knowledge on the topic is still limited—the main difference in previous studies is that they were performed in different countries.[2] In particular, little is known about the actual predictive importance of Google searches.

Against this background, in this paper I address three novel questions.

[1] Source: Google Internal Data, 2014.

[2] The studies have been performed in Germany (Askitas and Zimmermann 2009), the U.S. (Choi and Varian 2012; D'Amuri and Marcucci 2012), the UK (McLaren and Shanbhogue 2011), Israel (Suhoy 2009), Finland (Tuhkuri 2014), Italy (D'Amuri 2009), Norway (Anvik and Gjelstad 2010), Turkey (Chadwick and Sengul 2012), France (Fondeur and Karamé 2013), Spain (Vicente et al. 2015), Czech Republic, Hungary, Poland, and Slovakia (Pavlicek and Kristoufek 2014).

First, how far into the future could Google searches predict the unemployment rate? Most previous papers only study monitoring the current conditions with real-time search data, but not predicting the future. In particular, we do not know whether the predictive power of Google searches is limited to only very short-term predictions.

Second, when could Google data be useful? Earlier studies report only historical averages. What we do not know is whether an improvement in prediction accuracy is episodic or stable over time. Google searches could be particularly useful during a recession, when the economic indicators are hard to predict.

Third, how much could Google searches improve unemployment forecasts? Previous studies report a wide range of results—from small to very large improvements in prediction accuracy. Importantly, most previous studies do not ascertain whether the improvements they found are significant.

Since each previous study was based on country-level data around 2008 global economic crisis, their results are based on almost a single event: a sharp increase and subsequent gradual decrease in the unemployment rate. In contrast to previous studies, this paper constructs a state-level panel data set to study the robustness of the results at a more granular level, which allows me to exploit the sharp geographic and temporal variation in the unemployment rate induced by the 2008 economic crisis. Furthermore, this paper takes a novel approach by using statistics for the actual search volumes on Google; previous studies only used normalized variation within a single keyword over time. I use that data for variable selection—that is, to identify popular search terms. Without that knowledge, we were earlier only guessing which terms people actually used. This approach also allows us to weight search terms by their prevalence.

The questions of this paper are relevant: Google data is one of the largest data sets ever collected. Forecasters and researchers alike need to know how useful it actually is. The answers, in turn, will illustrate the potentials of Google search data for economic research and for real economic agents. The questions are also more generally relevant since none of them have been discussed in-depth in other contexts in which Internet search data could be useful.

This paper uses a (pseudo) out-of-sample forecast comparison methodology. The main model contains a variable, *Google Index*, constructed from Google data using approximately 35 million[3] search queries related to unemployment benefits. The underlying idea is that Google searches on these topics could be associated with actual filings for unemployment benefits. That is, more

[3]Source: Google AdWords, 2014.

searches could signal higher unemployment. In contrast to previous studies, this paper also provides descriptive joint analysis to describe the intertemporal relationship between relevant Google searches and the unemployment rate.

As I said earlier, one of the motivations to use timely data, such Google data, is that the traditional statistics are released with a lag. In that sense, this paper is closely related to the more general and rapidly expanding literature on macroeconomic monitoring and real-time data analysis (see, Croushore 2006; Aruoba and Diebold 2010; Bańbura et al. 2013, and the references therein). Real-time assessment of current macroeconomic activity is also called *nowcasting* (Giannone et al. 2008).

But real-time data sources could also have practical relevance for several economic agents. For example, central banks are interested in acquiring real-time information on the economy, and recently, several central banks have shown interest in using Internet search data for economic forecasting (see, for example, Suhoy 2009 and McLaren and Shanbhogue 2011). Several other government institutions and NGOs worldwide, such as national unemployment offices, would also be better equipped if they had more timely information on the unemployment rate.

This paper is also related to several other strands of literature. Current studies document that the Internet plays an important role in the U.S. labor market (see, for example, Kuhn and Skuterud 2004; Stevenson 2008; Kroft and Pope 2014; and Kuhn and Mansour 2014). The Internet is used to search for jobs in a variety of ways, including contacting public employment agencies and submitting job applications (Kuhn and Mansour 2014). In particular, Google searches could offer information on the unemployment rate and labor market activity (Baker and Fradkin 2014).

More generally, Varian (2010) reminds us that previously unrecorded activity is now recorded by computers. For example, we get information about private actions in the labor market through Internet search logs. This nanodata (Wu and Brynjolfsson 2015) arising as a by-product might help improve unemployment forecasts. These new data sources are sometimes called *big data*. It is a broad term that refers to new massive data sets—the amount of information created until 2003 is now created every two days (Einav and Levin 2013, and the references therein). The broad theme underlying this paper is whether big data could improve macroeconomic forecasts.

2 Data

The primary data sources for this paper are the *Google Trends* database by *Google Inc.* and the Labor Force Statistics from the Current Population Survey by the U.S. Bureau of Labor Statistics.

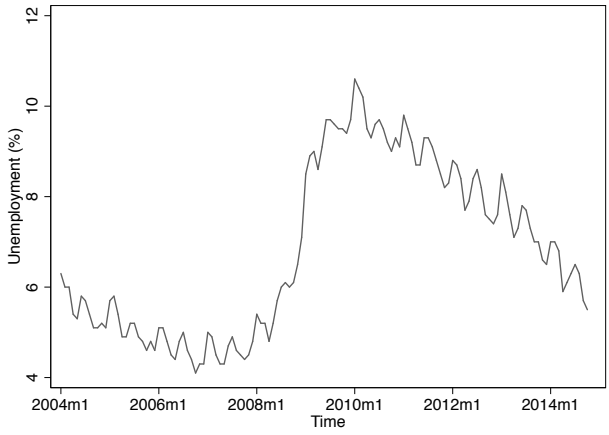

Figure 2.1: Unemployment rate in the United States 2004–2014. Not seasonally adjusted. Source: The Bureau of Labor Statistics.

2.1 Unemployment

Figure 2.1 describes the evolution of the unemployment rate in the United States from January 2004 until October 2014.[4] Typically, the unemployment rate exhibits seasonal variation, although this is not obvious in Figure 2.1. I use the non-seasonally adjusted unemployment rate as we are interested in short-term predictions. The evolution of the unemployment rate is characterized by a sudden increase between 2008 and 2010 that was associated with the economic crisis. The abrupt increase in unemployment was hard to predict—or at least, many predictions failed. New big data sources, such as Internet search data, could help produce more accurate forecasts.

2.2 Google

The *Google Trends* database measures volumes of Google searches. It tells us how many searches on certain search terms have been made, compared to the total number of Google search queries in the same period. The data are publicly available from 2004 onwards. In the U.S., the data is published at the state level.

In this section, I first select relevant search terms and then construct a variable that describes search volumes for the terms. I give the variable a name, *Google Index*.

I come up with 125 search terms that are related to unemployment benefits, and selected 13 terms with the highest search volumes. The search terms are: unemployment benefits, unemployment office, unemployment claim, unemployment compensation, unemployment insurance, apply

[4]The unemployment data for this study was retrieved from the the U.S. Bureau of Labor Statistics website on 15 December 2014.

for unemployment, applying for unemployment, filing for unemployment, unemployment online, unemployment office locations, unemployment eligibility, ui benefits, and unemployment benefit. This is the highest number of search terms that the *Google Trends* database allows to be export in one session. Exporting the data in one session allows me to use Boolean search operators (see, for example, Silverstein et al. 1999) later to construct a variable from the search volumes. The data for identifying the search terms comes from the *Google AdWords* database. It has not been used in the previous literature.

From 2004 to 2014, there were approximately 270,000 monthly search queries with the selected search terms.[5] In other words, the analysis is based on approximately 35 million Google searches. Distribution of the search volume with respect to the search terms is steep: 50 percent of the searches in the set were made with the most popular search term: unemployment benefits. Only 0.6 percent were made with the 13th most popular (and misspelled) term: unemployment benefit.

The selected search terms are related to unemployment benefits because they are likely to be the first searches that a laid-off worker types into Google. In contrast, searches for jobs might increase for many reasons not related to unemployment. Previous research from Germany (Askitas and Zimmermann 2009) and the U.K. (McLaren and Shanbhogue 2011) suggests that searches for unemployment benefits have the potential to predict the unemployment rate. The previous study in the U.S. (D'Amuri and Marcucci 2012) did not use search terms related to unemployment benefits, but only one term: jobs. I use many search terms instead of one in order to extract a more robust signal, and I explore the sensitivity of the results to the selected search terms later, in Section 5.

One point is worth emphasizing. There are over 15 billion new search terms typed into Google every month.[6] With billions of potential predictors and no clear guidance from economic theory, overfitting is a serious concern. I do not try to find the best set of keywords for predicting the U.S. unemployment rate, but rather examine if real-time Google search volumes could help in the task. Google data do not have to be the best to be useful.

The following section describes the construction of Google Index from the selected search terms. Google Index represents aggregate search activity for the selected unemployment-related search queries.

First, the search terms are combined by a Boolean search operator OR. The index includes searches containing the terms unemployment benefits OR unemployment office OR unemployment claim and so on (Silverstein et al. 1999). It is a sum. The advantage of this method is that it gives

[5]Source: Google AdWords, 2014.
[6]Source: Google Internal Data, 2014.

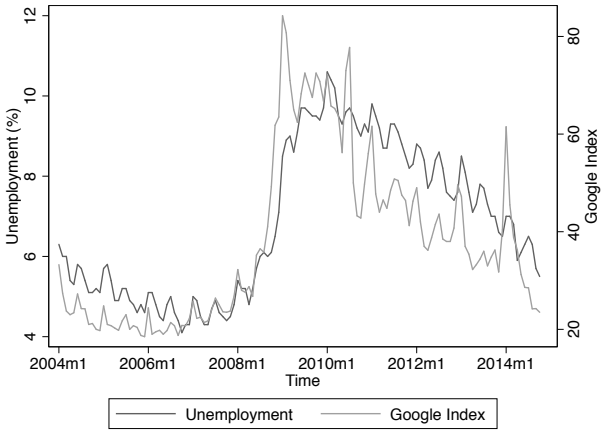

Figure 2.2: Unemployment rate and the Google Index that describes search activity for unemployment benefits in the United States 2004–2014. Source: The Bureau of Labor Statistics and *Google Trends*.

each search term a weight based on its search volume, even when the actual search volumes are not directly available from *Google Trends*. Second, the number of search queries made with the selected keywords is divided by the number of all search queries, which were made in the same period of time and in the same geographical area. Third, the data are normalized to the scale of 0–100. Finally, the Google Index is aggregated from weekly to monthly level.

In summary, let $K_{t,i}$ denote the number of searches with a set of keywords k for a given geography i and time period t, where $t = 1, 2, \ldots, f$. Let also $G_{t,i}$ denote the total number of search queries in geography i at time t. Then the unit of measurement for search intensity $I_{t,i}$ of the Google Index is

$$I_{t,i} = \left\{ \frac{\frac{K_{t,i}}{G_{t,i}}}{\max_t \left(\frac{K_{t,i}}{G_{t,i}} \right)} \right\} \times 100, \tag{2.1}$$

where

$$K_{t,i} \in (K_{1,i}, K_{2,i}, \ldots, K_{t,i}, \ldots, K_{f,i})$$
$$G_{t,i} \in (G_{1,i}, G_{2,i}, \ldots, G_{t,i}, \ldots, G_{f,i}).$$

Figure 2.2 describes the evolution of the Google Index and the unemployment rate from January 2004 until October 2014.[7] The series seem to behave in a similar manner; the correlation between monthly unemployment and the Google Index is 0.87. The search intensity for the selected

[7]Data was retrieved on 12 December 2014.

Variable	n	μ	σ	σ^2	sk	k	min	max
Unemployment (%)	130	6.84	1.89	3.55	0.25	1.62	4.1	10.6
Google Index	130	38.1	17.2	295.4	0.83	2.64	18.5	84.2

Sample period Jan 2004–Oct 2014, n = sample size, μ = mean, σ = standard deviation, σ^2= variance, sk = skewness, k = kurtosis, min = smallest value, and max = largest value.

Table 2.1: Descriptive statistics for the unemployment rate and Google Index 2004–2014. Source: The Bureau of Labor Statistics and *Google Trends.*

unemployment-related searches exhibits no clear trend between 2004 and 2008. After 2008, there is a sudden increase coincident with the economic crisis. Following the initial increase, there are several spikes. President Barack Obama signed the Unemployment Compensation Extension Act of 2010 into law in July 2010, and the figure shows a rapid increase in search activity around that time. The second spike coincides with Congress ending the same act in January 2014. I repeated the analysis, this time controlling for the two spikes. Table 2.1 gives descriptive statistics for the Google Index and the unemployment rate.

3 Methods

This section presents the methods for determining whether Google searches predict unemployment, and how far into the future, when in time, and by how much Google data could improve unemployment forecasts. First, I outline my methods for descriptive joint analysis of the series, and then specify the models selected to answer the questions more directly.

3.1 Joint Analysis

I analyze the series jointly by performing Granger (1969) non-causality tests and by studying the cross-correlation function. The analysis of the cross-correlation function helps resolve the lead-lag relationship between search volumes and unemployment. The advantage of the Granger non-causality test is that it also allows us to study the relationship the other way around; that is, whether the unemployment rate also predicts Google searches. If not, Google data might offer genuinely new information about the unemployment rate.

3.2 Model

This paper uses (pseudo) out-of-sample forecast comparison methodology associated with West (1996) and Clark and McCracken (2001). For Google data, the approach traces to the work of Choi and Varian (2012) and Goel et al. (2010). The first step is to select a relevant benchmark model for the unemployment rate. That model is extended with the Google Index, and the models and their forecast performance are compared. However, recent work by Diebold (2015) reminds us that (pseudo) out-of-sample forecasts do not automatically provide protection against overfitting.

This paper uses a seasonal AR(1) model as the main benchmark. It uses only the previous period and seasonal effects to predict the unemployment rate, as presented below.

$$\text{Model (0.0): } log(y_t) = \beta_0 + \beta_1 log(y_{t-1}) + \beta_2 log(y_{t-12}) + e_t$$

There are three main reasons for selecting the seasonal AR(1) model.

First, a simple model serves as a first test to ascertain whether Google data offer any advantage for predicting the unemployment rate. If Google data fail to offer any improvement against the naïve benchmark, then it is not likely to improve the more sophisticated models either.

Second, visual analysis of the autocorrelation function in Figure A.1 suggests that the unemployment rate follows almost a random walk process. For pure random walk processes, the best univariate forecast for y_t would be only y_{t-1}.[8]

Third, during 2004–2014 observation period, the evolution of the U.S. unemployment rate was dominated by an abrupt increase following the financial crisis of 2007–2008. There is uncertainty on how the dynamics of the unemployment series should be modeled within such a short and historically idiosyncratic sample.

The estimated autocorrelation function (ACF) and partial autocorrelation function (PACF) of the series are provided in Figures A.1 and A.2. The ACF has a slow decay but eventually tails off, reaching zero. The first lag of the PACF has a relatively high partial autocorrelation compared to the other lags, and there appears to be a cutoff at the 13th lag, after which the partial autocorrelations remain statistically insignificant at the 5% level.

Using sequential testing, AR(13) is the first model that has a statistically significant p:th coefficient at the 5% level. Also, both Akaike (1973) and Bayesian (Schwarz 1978) information criteria

[8] In the previous literature on forecasting with Google data, Choi and Varian (2012) use the same argument to motivate the use the AR(1) benchmark.

give the smallest value for AR(13) when using a maximum lag of 20 and including a seasonal lag for every model. A reason for this may be that the seasonal lag is not able to accommodate the seasonality in the series. Both information criteria decrease almost monotonously until the 13th lag. However, AR(13) is not a reasonable main benchmark for the (pseudo) out-of-sample forecast comparison. We would have to estimate 14 coefficients, while there are only 130 observations in the unemployment series.

The selected model is autoregressive in order to impose as little structure as possible to minimize assumptions. With only a limited amount of data at hand, overfitting is a serious concern (see, for instance, Varian 2014) and complicated models are not necessarily estimated accurately. Furthermore, empirical research has shown that simple models often yield better out-of-sample predictions than complex models (Mahmoud 1984). That is why a simple univariate autoregressive model is a relevant benchmark in a (pseudo) out-of-sample forecasting environment. More to the point, Montgomery et al. (1998) document that an autoregressive model is appropriate for short-term unemployment forecasting.

Both variables, the unemployment rate and the Google Index, are measured in levels rather than in differenced values, because both are bounded between 0 and 100. For this reason, they cannot exhibit global unit root behavior (Koop and Potter 1999). Furthermore, Cochrane (1991) argues that during the last one hundred years, the U.S. unemployment rate has had no visible trend, and economic theory does not suggest it should have had one.

A seasonal autoregressive term, y_{t-12}, is included in the benchmark AR model to make sure that a possibly observed relationship between the unemployment rate and the Google Index would not be entirely driven by common seasonality. In the previous literature on assessing the relevance of Internet data sources, Choi and Varian (2012) and Wu and Brynjolfsson (2015) apply the same approach.

Additionally, I perform a logarithmic transformation on the unemployment series since changes in unemployment rate are most naturally discussed in percentage terms, and also because logarithmic transformation helps stabilize the variance of the series (Lütkepohl and Xu 2012).

But the seasonal AR(1) model is almost certainly not identical to the true model. As a minimum protection against such problems, I check that the fitted model was adequate to describe our data-generating process (DGP) by providing several diagnostic checks.

I estimate the seasonal AR(1) model by using a quasi-maximum likelihood (QML) method under normality assumption. Figures A.3 and A.4 outline the autocorrelation functions of the residuals

and squared residuals for the baseline seasonal AR(1) model. There is still a small amount of autocorrelation in the residuals, but not necessarily conditional heteroskedasticity. The residual autocorrelation may be due to remaining seasonality in the residual series. Nonetheless, the autocorrelation in the residual series abates as the lag increases. There does not appear to be unit root problems.

To formally evaluate whether most of the temporal dependence have been removed from the residuals, I compute the Ljung–Box (1978) portmanteau statistic for the residuals. The portmanteau test statistic Q_K computed with $K = 12$ and $K = 24$ lags does reject the null hypothesis of no serial correlation at the 1% level. Furthermore, the same test statistic rejects the null hypothesis of no autocorrelation in the residuals at the 1% level for every AR(p) model until the 13th order AR model. Again, the reason for this is possibly that the 12th lag, which was included in every model, is not capable of accommodating the seasonality in the series. I also formally test for the conditional heteroskedasticity in the residual series. Although the squared residuals in Figure A.4 seem serially uncorrelated, the McLeod-Li (1983) test statistic, computed as Q_K for the squared residuals, rejects the null hypothesis of no conditional heteroskedasticity with $K = 12$ and $K = 24$ lags. However, when estimating alternative benchmark models up until the fourth-order seasonal AR model, I find no clear advantage against the seasonal AR(1) model, judging by the estimated autocorrelation functions of the residuals.

For the reasons listed above, among lower than fourth-order autoregressive models, the seasonal AR(1) model appears to be an adequate benchmark. I use the seasonal AR(1) model as a benchmark because the more complicated benchmark models do not offer a marked advantage against the simple one. I account for the remaining autocorrelation in the residuals by using heteroskedasticity- and autocorrelation-consistent (HAC) standard errors developed by Newey and West (1987, 1994). To explore the sensitivity of the results for the selected benchmark, I also estimate the results using seasonal AR(2) and AR(3) benchmark models in Section 5.

3.2.1 Predicting the Present

Google data are available a month earlier than the official unemployment statistics. It gives the Google data a meaningful forecasting lead (Choi and Varian 2012). Searches for unemployment benefits now could help predict the current unemployment rate, which is not known at the date of prediction.

The main specifications for evaluating nowcasting performance are the benchmark Model (0.0)

and the extended Model (1.0), which are presented below.

$$\text{Model (0.0): } log(y_t) = \beta_0 + \beta_1 log(y_{t-1}) + \beta_2 log(y_{t-12}) + e_t$$
$$\text{Model (1.0): } log(y_t) = \beta_{00} + \beta_{10} log(y_{t-1}) + \beta_{20} log(y_{t-12}) + \beta_{30} x_t + e_t$$

The unemployment rate in the present month t is denoted by y_t, in the previous month by y_{t-1}, and a year ago by y_{t-12}. The contemporaneous value of the Google Index is denoted by x_t. Moreover, e_t stands for the error term. Coefficients and constant terms are denoted by β:s using different subscripts.

Caution should be exercised when studying whether a new indicator predicts economic activity. In many cases, a model using only the previous period and seasonal effects will explain more than 90 percent of the variance in a dependent variable (Goel et al. 2010). It is not enough to illustrate that Google searches are correlated with current or future unemployment—it must be demonstrated that the model with the Google Index performs at least better than a benchmark model using lagged data and seasonal effects (Goel et al. 2010).

I begin by estimating the models for the entire observation period. These results could provide some evidence about the fit of the benchmark and extended models and give information on the statistical properties of the U.S. unemployment rate. I compare the fit of the models measured by coefficient of determination R^2, as well as other properties, such as Akaike (AIC) and Bayesian (BIC) information criteria, statistical significance, and the magnitude of the parameters.

To answer whether Google searches could help to forecast the unemployment rate, I conduct a (pseudo) out-of-sample forecast comparison. In specific, I am interested in finding out about the incremental predictive ability of the Google Index over and above lagged and seasonal effects of the unemployment rate itself. I generate a series of one-step-ahead out-of-sample predictions using a rolling window of 48 months for models (0.0) and (1.0). For each month beginning in 2008, I train the model using 48 past observations, and then evaluate the out-of-sample predictions by comparing the forecasted values to the realized values of the unemployment rate. The 48-month window is chosen to make sure that there are enough observations to estimate the models, and that the evaluation period is long enough.

Mean absolute percentage error (MAPE) is used as a measure of forecasting accuracy. It is defined as:

$$\begin{aligned} \text{MAPE} &= \frac{1}{T}\sum_{t=1}^{T} \mid E_t \mid, \qquad (3.1) \\ \text{where} & \\ E_t &= \frac{\hat{y}_t - y_t}{y_t} \times 100, \end{aligned}$$

where y_t denotes the official unemployment rate and $\hat{y}_t$ denotes the forecasted value. If the error measure for forecasts computed from the extended model lies below that of the benchmark model, I conclude that Google searches predict unemployment. I explore the sensitivity of results to the selected error measure with mean squared error in Section 5.

Finally, I test whether the difference in forecast accuracy between the two models is statistically significant using the test for equal predictive accuracy of Diebold and Mariano (1995) and West (1996).

3.2.2 Forecasting the Future

Extending the nowcasting framework of the previous section, I construct separate models for each horizon into the future, so that every model uses the most recent information when producing dynamic forecasts for the future. Dynamic forecast means that only values that are known at the date of prediction t are used. The Models (0.0)–(1.6) are presented below.

$$\text{Model (0.0): } log(y_t) = \beta_0 + \beta_1 log(y_{t-1}) + \beta_2 log(y_{t-12}) + e_t$$

$$\text{Model (1.0): } log(y_t) = \beta_{00} + \beta_{10} log(y_{t-1}) + \beta_{20} log(y_{t-12}) + \beta_{30} x_t + e_t$$

$$\text{Model (1.1): } log(y_t) = \beta_{01} + \beta_{11} log(y_{t-1}) + \beta_{21} log(y_{t-12}) + \beta_{31} x_{t-1} + e_t$$

$$\text{Model (1.2): } log(y_t) = \beta_{02} + \beta_{12} log(y_{t-1}) + \beta_{22} log(y_{t-12}) + \beta_{32} x_{t-2} + e_t$$

$$\text{Model (1.3): } log(y_t) = \beta_{03} + \beta_{13} log(y_{t-1}) + \beta_{23} log(y_{t-12}) + \beta_{33} x_{t-3} + e_t$$

$$\text{Model (1.4): } log(y_t) = \beta_{04} + \beta_{14} log(y_{t-1}) + \beta_{24} log(y_{t-12}) + \beta_{34} x_{t-4} + e_t$$

$$\text{Model (1.5): } log(y_t) = \beta_{05} + \beta_{15} log(y_{t-1}) + \beta_{25} log(y_{t-12}) + \beta_{35} x_{t-5} + e_t$$

$$\text{Model (1.6): } log(y_t) = \beta_{06} + \beta_{16} log(y_{t-1}) + \beta_{26} log(y_{t-12}) + \beta_{36} x_{t-6} + e_t$$

Optimal forecasts are produced recursively. For example, Model (1.1) produces the dynamic forecast for horizon $h = 1$. This is done recursively (starting with the one-period forecast) by using

the unemployment rate in the period $t-1$ and $t-12$ and the value of the Google Index at time t for the last forecast horizon. This study uses dynamic forecasts instead of static forecasts because this method is closer to what actual forecasters would do.

I evaluate the models' out-of-sample performance by comparing the dynamic h-step-ahead forecasts using the same methodology described earlier in Section 3.2.1. In specific, if a model that includes the Google Index provides more accurate forecasts than a benchmark model in the (pseudo) out-of-sample environment for horizon h but not for $h+1$, I can infer that the marginal predictive ability of Google searches is limited to horizon h predictions.

3.2.3 Time-specific Forecasts

The value of Google data for forecasting purposes may depend on the date of the forecasts. Real-time data might be especially useful during a recession when the economic indicators are hard to predict. From a practical forecasting perspective, this is an important criterion for the relevance of a new data source.

I study whether the marginal predictive ability of Google data varies over time by analyzing the performance of the models during the 2007–2009 recession in comparison with their historical performance during the whole observation period. I also take a closer look at the topic by constructing a series describing the difference in forecast errors between the two models. That is, I not only consider average improvements in forecasting accuracy, but also when the improvement happens.

4 Results

I begin by reporting the descriptive joint analysis of Google searches and the unemployment rate, then turn to estimating models in order to answer the questions more directly.

4.1 Joint Analysis

4.1.1 Cross-correlation

Do Google search volumes anticipate unemployment? As a simple summary of the temporal relationship between the unemployment rate and the Google Index, Table 4.1 displays the values of the estimated cross-correlation function (CCF).

The main observation is that the values of the cross-correlation function between present unemployment volumes and past Google searches appear to be larger than the that of the opposite

h	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
CCF	0.92	0.91	0.89	0.88	0.89	0.89	0.87	0.82	0.77	0.74	0.72	0.70	0.67

n = 130, h = lag of Google Index, CCF = value of cross-correlation function. The values of CCF on the left-hand side tell the correlation coefficients between past Google search volumes and the present unemployment.

Table 4.1: Cross-correlation function between the unemployment rate and the Google Index.

Null hypothesis							
VAR(1)				VAR(1) using lead of x			
$y \nrightarrow x$		$x \nrightarrow y$		$y \nrightarrow x$		$x \nrightarrow y$	
χ^2	p-value	χ^2	p-value	χ^2	p-value	χ^2	p-value
0.040	0.84	22.83	<0.001***	0.0032	0.96	71.6	<0.001***

y = unemployment rate, x = Google Index.
The sample period is Jan 2004–Oct 2014 (n = 130). Both models estimated are first-order VARs, which, based on the Schwarz criterion, are statistically adequate simplifications of second-order VARs. Asterisks *, **, and *** denote significance at the 5%, 1%, and 0.1% levels, i.e., Granger non-causality $' \nrightarrow '$ is rejected.

Table 4.2: Statistics for testing Granger non-causality.

case; Google search volumes tend to anticipate the U.S. unemployment rate. A closer look reveals that the correlation is strongest between the current search activity and the unemployment rate six months ahead. The temporal dependence revealed by the historical cross-correlation function of the unemployment rate and the Google Index suggests a bivariate structure of the two series, and likely the possibility to outperform the predictions based on an autoregressive model by introducing Google search volumes among the regressors.

4.1.2 Granger Causality

Do Google searches Granger cause unemployment? Table 4.2 gives statistics for testing Granger non-causality. The null hypothesis that Google searches do not Granger cause unemployment can be rejected at the 1% level. A second specification is based on a different VAR model. I use the lead of x instead of x, because the Google Index is available a month before the unemployment rate. That is, in the corresponding VAR model, the explanatory variables represent the most recent observations at the date of prediction. This is a non-standard procedure, but respects the actual information

set available for forecasters. A similar conclusion is drawn when Google data are observed a month earlier than the unemployment rate. In summary, both specifications suggest that Google searches offer useful information in predicting the unemployment rate.

In contrast, according to the Granger non-causality test, lagged values of the unemployment rate do not offer useful information in predicting Google searches over and above the Google series themselves. This suggests that Google searches could offer genuinely new information on unemployment that is not already included in the unemployment series itself. When using fourth-order VAR models, I find no qualitative changes in the results.

4.2 Model

4.2.1 Predicting the Present

Do Google searches help predict the present unemployment rate? The estimation results for Models (0.0) and (1.0) are presented in Table 4.3. The coefficient for the Google Index is statistically significant at the 1% level. The positive sign of the coefficient means that the searches related to unemployment benefits are positively connected to the unemployment rate. More specifically, the coefficient 0.00440 means that the 1 percent increase in current search intensity is associated with a 0.44 percent increase in the current unemployment rate.

The R^2 for model (0.0) is 0.962, which means that the benchmark model alone can explain a large part of the variation in the unemployment rate, as suggested before by Goel et al. (2010). Nonetheless, extending the benchmark model (0.0) with the Google Index decreases the values of both Akaike and Bayesian information criteria. This result suggests that the Google searches offer useful information in explaining variation in the unemployment rate within the estimation sample.

Results from one-step-ahead out-of-sample predictions using a rolling window of 48 months are illustrated in Figure 4.1. The mean absolute percentage errors for nowcasts are given on the first row of Table 4.4. The mean absolute percentage error for forecasts computed from Model (0.0) without Google data is 4.58 percent. The same measure for Model (1.0) with Google data is 4.38 percent. This is an improvement of 4.32 percent for predicting the present unemployment rate. I infer that Google searches help to predict unemployment compared to a univariate benchmark. However, because the benchmark and the loss function are more or less arbitrary, the reported improvement is indicative.

The results from the Diebold-Mariano test, however, display no statistical significance (at the 10% level) on the difference between the forecasts. There are two apparent reasons for this. First,

Model	(0.0)	(1.0)
Variables		
$log(y_{t-1})$	0.983**	0.955**
	(0.0295)	(0.0356)
$log(y_{t-12})$	-0.0103	0.0156
	(0.0300)	(0.0368)
x_t		.00440**
		(0.000656)
Constant	1.848**	1.692**
	(0.191)	(0.150)
Summary		
R^2	0.962	0.969
AIC	-371.2	-396.6
BIC	-359.7	-382.3
n	130	130

y = unemployment rate, x = Google Index. Asterisks * and ** denote statistical significance at 5% and 1% levels using a two-sided test. The standard errors of Newey and West (1987) of the estimated coefficients are given in parentheses. The number of lags for the standard errors is selected as in Newey and West (1994). Estimation by QML. The sample period is Jan 2004–Oct 2014.

Table 4.3: Estimation results of the benchmark seasonal AR(1) model (0.0) and the extended model (1.0), which includes Google Index.

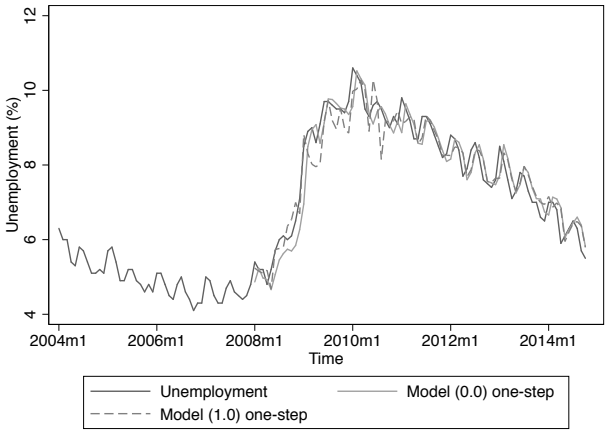

Figure 4.1: Unemployment rate 2004–2014 and the one-step-ahead nowcasts with a rolling window of 48 months for the univariate benchmark model (0.0) and the extended model (1.0), which includes Google Index 2008–2014.

the observation period is short—only 130 monthly observations. Thus, the power of the test is low (Diebold 2015). Second, the observed improvement is small. A small improvement combined with a low-power test makes it hard to distinguish whether the incremental predictive accuracy against the benchmark represents a more general difference "in population" or merely an observation "in sample".

4.2.2 Forecasting the Future

Do Google searches help forecast the future unemployment rate? Table 4.4 summarizes the mean absolute percentage errors of out-of-sample dynamic forecasts up to the horizon $h = 6$. We can see from Table 4.4 that the two-step-ahead forecasts improve 7.48 percent on average when we add in the Google data, compared to a 4.32 percent improvement for the one-step-ahead forecasts.

But if we predict the unemployment rate two months ahead we get a decline of 3.92 percent in forecast accuracy. The results indicate that Google data might help to predict unemployment for horizon $h = 1$, but not necessarily much further. Still, the series of (pseudo) out-of-sample predictions demonstrate that the current Internet searches for unemployment benefits are likely to offer information on the next month's unemployment rate, not only on the present one. Note that forecasts are on average less accurate than nowcasts, as they should be. Increasing the forecast horizon decreases the forecasting accuracy for both models.

Are the differences between the forecasts statistically significant? In line with the results in the previous section for nowcasting accuracy, the Diebold-Mariano test for equal predictive accuracy

Horizon	Model	MAPE	Δ
$h = 0$	(0.0)	4.58%	4.32%
	(1.0)	4.38%	
$h = 1$	(0.0)	7.57%	7.48%
	(1.1)	7.01%	
$h = 2$	(0.0)	9.48%	-3.92%
	(1.2)	9.85%	
$h = 3$	(0.0)	10.4%	-6.28%
	(1.3)	11.06%	
$h = 4$	(0.0)	11.1%	-17.22%
	(1.4)	13.02%	
$h = 5$	(0.0)	11.96%	-13.22%
	(1.5)	13.54%	
$h = 6$	(0.0)	13.40%	9.93%
	(1.6)	12.07%	

MAPE = mean absolute percentage error

Δ = improvement in forecasting accuracy

Estimated values are computed recursively using dynamic n-step-ahead forecasts with a rolling window of 48 months for each model. The evaluation period is Jan 2008–Oct 2014.

Table 4.4: Nowcasting and forecasting accuracy of the seasonal AR(1) benchmark model (0.0) and the extended models (1.0)–(1.6) that include Google Index 2008–2014.

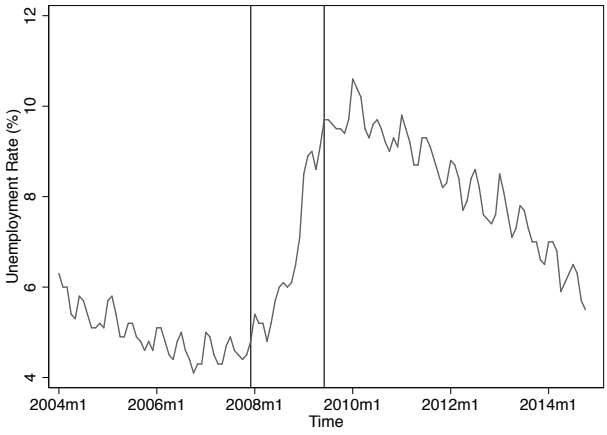

Figure 4.2: The recession.

reports at the 10% level no statistically significant differences between the forecasts. This suggests that the improvements in forecasting accuracy from using Google data may not be large.

Earlier, the descriptive cross-correlation analysis suggested that the correlation is strongest between the current search activity and the unemployment rate six months ahead. On the other hand, the (pseudo) out-of-sample forecast comparison does not find an advantage from Google data beyond one month ahead. If there is a longer-term link, it tends to be overshadowed by other factors. One explanation for the discrepancy is that a substantial share of the correlation is driven by a large increase and a subsequent decrease in the series. The Google search activity peaks six months before the initial increase in unemployment in 2008. However, the lead-lag relationship may not be consistent.

4.2.3 Time-specific Forecasts

Does the marginal predictive ability of Google data vary over time? From 2004 to 2014, there was only one contraction phase, according to the National Bureau of Economic Research (NBER) Business Cycle Dating Committee. The recession happened from December 2007 until June 2009 and lasted for 18 months. The vertical lines in Figure 4.2 highlight the economic crisis. During that time, official statistics were revised frequently, and there was a genuine need for more accurate information. A majority of professional forecasts failed to identify the recession at the point where it was later determined to have begun.[9]

However, previous studies by Choi and Varian (2012) and Goel et al. (2010) conjecture that

[9]Source: Federal Reserve Bank of Philadelphia, Survey of Professional Forecasters, 2015 and National Bureau of Economic Research, Business Cycle Dating Committee, 2015.

Horizon	Model	MAPE	Δ
$h = 0$	(0.0)	7.17%	17.95%
	(1.0)	5.88%	
$h = 1$	(0.0)	11.69%	34.50%***
	(1.1)	7.66%	
$h = 2$	(0.0)	15.60%	4.53%
	(1.2)	14.89%	
$h = 3$	(0.0)	20.57%	-25.57%*
	(1.3)	25.57%	
$h = 4$	(0.0)	26.07%	-35.06%
	(1.4)	35.06%	

MAPE = mean absolute percentage error
Δ = improvement in forecasting accuracy
Estimated values are computed using dynamic n-step-ahead forecasts with a rolling window of 48 months for each model. The statistical significance of the differences in the mean absolute percentage errors is tested using the test of Diebold and Mariano (1995) and West (1996). In the table, *, **, and *** denote the rejection of the null hypothesis of equal predictive performance at 10%, 5% and 1% significance levels, respectively. The evaluation period is Dec 2007–June 2009.

Table 4.5: The recession. Nowcasting and forecasting accuracy of the seasonal AR(1) benchmark model (0.0) and the extended models (1.0)–(1.6) that include Google Index 12/2007–6/2009.

sudden changes in search intensity could help identify sudden changes in economic time series. Table 4.5 gives the mean absolute percentage errors of dynamic forecasts up to $h = 4$ from December 2007 until June 2009. When we look at one-step-ahead forecasts during the recession, we find that the mean absolute percentage error goes from 7.17 percent using the baseline forecast to 5.88 percent using the Google data, which is a 17.95 percent improvement in prediction accuracy. Additionally, in the two-steps-ahead out-of-sample forecasts, there is 34.50 percent improvement. Even at the three-steps-ahead horizon, there is a gain of 4.53 percent; while on average, Google data do not improve three-steps-ahead forecasts.

In summary, during the recession, the improvements are about four times larger than on average. This observation suggests that Google search queries tend to improve the prediction accuracy especially during the recent recession, that is, the most recent turning point. On the other hand, the

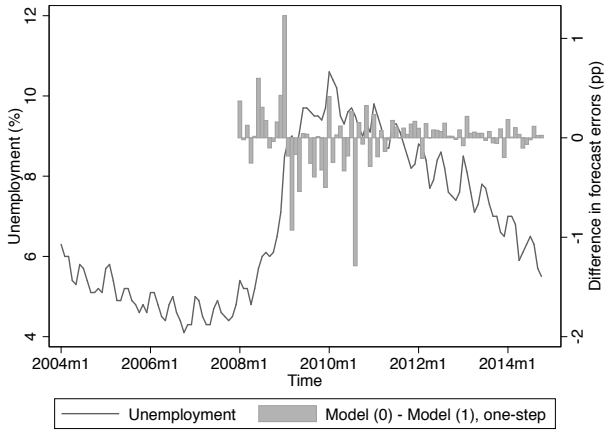

Figure 4.3: The difference in absolute forecast errors for one-step-ahead nowcasts of the univariate benchmark model (0.0) and the extended model model (1.0), which includes Google Index 2008–2014 and the unemployment rate 2004–2014. The vertical bars are positive when the extended model performs better.

models using Google data improve predictions markedly only until $h = 1$, even during the recession. Furthermore, both models give less accurate predictions during the recession than on average.

Diebold-Mariano tests for comparing predictive accuracy support the finding that the improvements in prediction accuracy are larger in the recession. Table 4.5 reports that there is a statistically significant difference between the forecasts (at the 1% level) at the one-month horizon, when the improvement is at its largest. However, this is the only significant improvement at the 10% level.

More generally, when does the Google Index help forecast the unemployment rate? Looking more closely at the series, Figure 4.3 describes the difference in one-step-ahead forecast errors for the baseline model and the extended model with the Google Index for each month. The difference is positive when the model with the Google Index produces more accurate predictions and negative when the benchmark is more accurate. The main observation is that while the Google search data identifies the initial recession spike, the extended model underpredicts the unemployment immediately after. The forecast performance of the extended model with Google data tends to be episodic.

The observation period is short, and there is essentially only one major source of variation in the unemployment series. Therefore, this approach is limited in its ability to answer when the Google data are especially useful. But despite the benefits of Google data, including the Google Index as an additional predictor occasionally makes the benchmark model's out-of-sample predictions not better but worse.

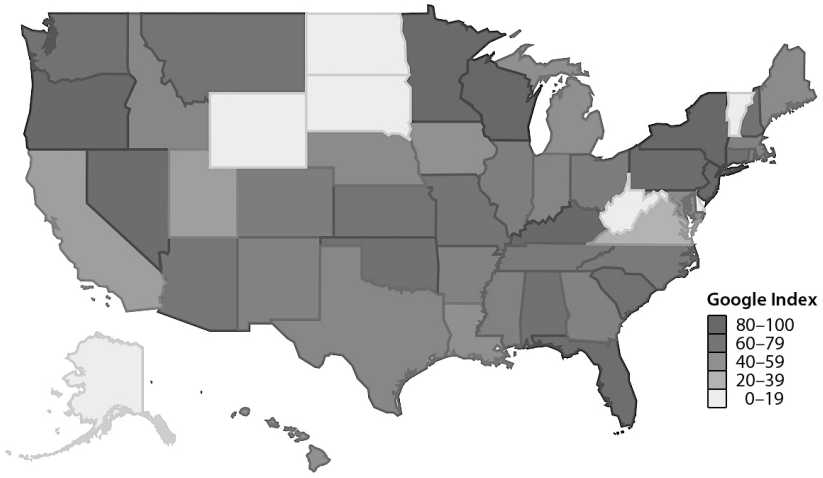

Figure 5.1: Relative popularity of unemployment-related Google searches. Average between Nov 2009 and Feb 2010. Source: *Google Trends*.

5 Robustness

5.1 Panel Data

The U.S. federal-level results in the earlier literature—and in this paper—are based on almost a single event: the economic crisis. Based on that evidence, it is not clear whether the pattern will hold in the future. However, the unemployment rate and Google searches had somewhat different patterns in each U.S. state. Figure A.5 in the Appendix illustrates the evolution of these differences. For example, during 2004–2014 in Illinois, both the unemployment rate and the Google Index increased earlier than in North Dakota. To illustrate this further, a map displayed in Figure 5.1 visualizes the U.S. state-level differences in the popularity of unemployment-related Google searches between November 2009 and February 2010.

To exploit the geographic and temporal variation in the unemployment rate induced by the 2008 economic crisis, I construct a state-level panel data set to study the robustness of the results. In this panel data set, we have 50 cross-section units for 130 time periods. Compared to the previous data set, we now have 5,900 observations instead of 130. To my knowledge, this is the first attempt to construct and study a panel data set using Google searches in the forecasting literature.

This paper uses the following fixed effects model with lagged dependent variables:

$$log(y_{i,t}) = \beta_1 log(y_{i,t-1}) + \beta_2 log(y_{i,t-12}) + \beta_3 x_{i,t} + \alpha_i + e_{t,i}, \tag{5.1}$$

where $i = 1, \dots 50$ and $t = 1, \dots, 118$. Each state is denoted by i. The fixed effects model has 50

Model	(1.0)	FE (AB)	FE (OLS)
Variables			
$log(y_{t-1})$	0.955**	0.825**	0.832**
	(0.0356)	(0.00555)	(0.0062)
$log(y_{t-12})$	0.0156	0.0678**	0.0673**
	(0.0368)	(0.00442)	(0.00499)
x_t	.00440**	.00176**	.00167**
	(0.000656)	(0.000058)	(0.000066)
Constant	1.692**		
	(0.150)		
Summary statistics for FE (OLS)			
R^2	within	0.935	
	between	0.998	
	overall	0.956	
F test that state fixed effects = 0		5.51	
		(<0.0001)	

y = unemployment rate, x = Google Index.
Asterisks * and ** denote statistical significance at 5% and 1% levels using a two-sided test with standard errors of Arellano (1987). In the second column, the model is estimated by method of Arellano and Bond (1991). In the third column, the model is estimated by the ordinary least squares (OLS) method. The results for Model (1.0) in the first column come from Table 4.3. The sampling period is Jan 2004–Oct 2014.

Table 5.1: Estimation results of the extended autoregressive model (1.0) and the fixed effects model (FE).

different intercepts denoted by α_i, one for each state. The model is otherwise similar to the Model (1.0) and follows the same logic. Again, the unemployment rate is denoted by $y_{i,t}$ and the Google Index by $x_{i,t}$. I account for the the remaining within-panel serial correlation in the state-level error term $e_{i,t}$ by employing heteroskedasticity- and autocorrelation-robust standard errors developed by Arellano (1987). Furthermore, I use an asymptotically-consistent generalized method of moments (GMM) type estimator derived by Arellano and Bond (1991) to estimate the parameters, but also check the results by employing a within estimator using the ordinary least squares (OLS) method.

The results from the state fixed effects model are given in the second column of Table 5.1 , with earlier results of the extended autoregressive model in the first column. In summary, the coefficient of the Google Index is significant at the 1% level, although smaller than in the Model (1.0). The

state level analysis suggests that the Google searches are associated with the unemployment rate even when controlling for the state-level fixed, lagged, and seasonal effects. The pattern—Google searches predict unemployment—seems to be repeated at the state level.

The estimation results from within estimator, reported in the third column of Table 5.1, are similar to that of the Arellano-Bond method. The coefficient of the Google Index is statistically significant at the 1% level with both methods.

Panel data methods provide an opportunity to control for unobserved factors in the relationship between Google searches and unemployment. This may explain the smaller coefficient in the state-level fixed effects model than in the federal-level autoregressive model. However, this is also a limitation against the model specification, because it is not entirely clear what the unobserved variables are.

In practice, using a cross-sectional dimension in the Google data might prove beneficial for forecasting. A forecaster might be able to produce more accurate predictions by predicting unemployment at the state level and then aggregating to the federal level.

5.2 Variables

One concern would be that the results were sensitive to the choice of the set of search terms. I explore the sensitivity by estimating the aggregate-level models with different search terms. I construct an alternative Google Index by using only one of the most salient terms, "unemployment benefits", alone. I also study the validity of the results by using search intensity for the search term "*facebook*" as a fake Google Index. The keyword "*facebook*" was the most popular search term on Google in 2014.[10] The idea is that the fake index, based on an irrelevant search term, should not help in predicting the unemployment rate.

I find that the models using the search term "unemployment benefits" alone yield very similar results. A variable describing query volumes for the keyword "unemployment benefits" is statistically significant at the 1% level. In addition, I find no statistical significance at the 10% level for the fake Google Index or improvement in prediction accuracy by using search intensity for *Facebook*.

One of the issues that we are always going to run into is changes in search behavior. Two spikes in search activity, depicted in Figure 2.2, were presumably associated with news about changes in the labor market policy, not with the level of unemployment. After controlling for the two events, the improvements from Google data compared to the benchmark are on average 10 percent higher

[10]Source: *Google Trends*, 2014.

than the improvements reported earlier.

5.3 Model Specifications

In a (pseudo) out-of-sample forecast comparison environment, it is necessary to make a variety of assumptions and choices in modeling. I explore the sensitivity of the results to some of the most restricting assumptions.

Against seasonal AR(2) and AR(3) benchmarks the Google Index is statistically significant at the 1% level, improves in-sample fit, is preferred by both Akaike and Bayesian information criteria, and does offer improvement in out-of-sample forecast comparison. The improvements, however, are slightly smaller than against the AR(1) benchmark.

The results of another commonly used error measure, mean squared error (MSE), are essentially the same as those using a mean absolute percentage error (MAPE). I also explore the sensitivity of the results to the selected rolling window size with several widths, including 24 and 60 months, and find that the magnitude of the results is somewhat sensitive to the selected width. However, this underlines the observation that the advantage from Google data is time specific.

6 Discussion

There are still some concerns. First of all, the improvements in prediction accuracy are only modest. This finding contrasts with some of the earlier literature on the topic in the U.S. D'Amuri and Marcucci (2012) find a 40 percent improvement in forecasting accuracy compared to their benchmark—on a two-months-ahead horizon. However, I do not find any consistent improvement in prediction accuracy beyond one-month-ahead predictions. A possible reason for the discrepancy is that the authors walk through over 500 models and report the results for the best performing model within the estimation sample. This paper avoids overfitting by using the simplest models possible. Additionally, compared to D'Amuri and Marcucci (2012), I have included over three more years of data. This adds up to a 44 percent increase in the number of observations. In terms of magnitude, my findings are more in line with the modest improvements reported by Choi and Varian (2012).

Another concern is that the simple autoregressive models used in this paper sometimes provide reasonable predictions but occasionally produce very bad forecasts. Lazer et al. (2014) argue that the Google search algorithm is constantly changing, and it is hard to train the forecasting model using past data. My take more directly targets unemployment forecasting. The unemployment rate

is a function of new cases, exits, and duration (see, for example, Barnichon and Nekarda 2012). The method used in this paper may make it harder to predict duration or changes in duration, which may explain why I underpredict unemployment after the initial recession spike—I miss longer-term unemployment.

But the methods used in this paper are relatively simple and do not necessarily represent the ways actual forecasters would use this data. For example, Koop and Onorante (2013) point out that Google variables could be useful in model selection rather than as additional regressors. This paper, however, sheds new light on the usefulness of the Google data.

A common criticism about forecasting with big data is that with vast amounts of data, it is easy to mistake a noise for a signal. Are the findings of this paper something meaningful, or only a random and interesting pattern that happens to be true in the past but might not have that much structural significance? At least, there is a solid background for the findings—we can predict the unemployment rate because individuals actually use the Internet as a tool in the labor market (Stevenson 2008; Kuhn and Mansour 2014). None of the methods used in this study alone would give an unambiguous answer as to whether Google searches predict unemployment. However, several methods combined together with the earlier literature on the topic indicate that Google data do contain useful information on current and near future unemployment, and that this information can be used to predict the U.S. unemployment rate.[11]

This paper disentangles the almost mechanical relationship between Google searches for unemployment benefits and the unemployment rate. Google data might also provide new insights, for example, on the behavior of the unemployed on the Internet. An early example of this is a work by Baker and Fradkin (2014). Fine-grained Internet data allow us to measure individual actions that previously have been hard to measure. At the same time, the Internet and the digitalization of the labor market also create new activities. To understand these activities, Internet data sources such as Google search logs will prove beneficial.

7 Conclusion

This paper has analyzed whether data on Google search volumes could help predict the unemployment rate. I was interested in the specifics. Using (pseudo) out-of-sample forecast comparison, I have found that models with relevant Google variables produce on average more accurate forecasts

[11] See our real-time implementation *ETLAnow* and additional materials maintained by ETLA, The Research Institute of the Finnish Economy, at www.etla.fi/en/etlanow-eu28, with username and password *etlanow2015*.

than the same models without Google data. That is, Google searches predict unemployment. Extending the previous literature, I have also shown that the pattern holds on a more granular U.S. state level, even after controlling for the state-level fixed effects. Joint analysis of the series verified that Google searches anticipate the unemployment rate, and the results were found to be robust to different model specifications and search terms. In contrast to previous literature on Internet searches, I included data on the actual search volumes on Google.

Three novel findings arise. First, improvements in predictive accuracy from using Google data are limited to short-term predictions. Second, the informational value of search data is time specific—Google search queries improve prediction accuracy especially during the recent recession. Third, compared to the previous results from the U.S. (D'Amuri and Marcucci 2012), I found that the improvements in forecasting accuracy from Google data may be smaller than previously thought. Yet, the qualitative results on nowcasting potential are in line with the previous findings on Google searches and unemployment by Askitas and Zimmermann (2009), Choi and Varian (2012), and D'Amuri and Marcucci (2012)—Google searches do predict unemployment in short term.

More generally, the results illustrate both the potentials and limitations of using big data to predict macroeconomic indicators. Big data does not necessarily mean that one single data source, such as Google data, would be able to improve economic forecasts in a large measure. But big data comes from billions of such data sources. Big data grows from little things, and better forecasts grow from little improvements. Just being able to measure previously unmeasurable activity is an extraordinary thing. We are in a position to make discoveries that no one has yet imagined.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In Petrov, B. N. and Cszaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Arellano, M. (1987). Computing Robust Standard Errors for Within-groups Estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434.

Arellano, M. and Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58(2):277–297.

Aruoba, S. B. and Diebold, F. X. (2010). Real-time macroeconomic monitoring: Real activity, inflation, and interactions. *American Economic Review*, 100(2):20–24.

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

Baker, S. R. and Fradkin, A. (2014). The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data. *Working Paper, Stanford University.*

Banbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Now-casting and the real-time data flow. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, pages 195–237. Elsevier.

Barnichon, R. and Nekarda, C. J. (2012). The Ins and Outs of Forecasting Unemployment: Using Labor Force Flows to Forecast the Labor Market. *Brookings Papers on Economic Activity*, Fall:83–132.

Brynjolfsson, E. (2012). Big Data: A revolution in decision-making improves productivity. *MIT Sloan Experts.*

Chadwick, M. G. and Sengul, G. (2012). Nowcasting unemployment rate in Turkey: Let's ask Google. *Central Bank of the Republic of Turkey Working Paper 12/18*, (June).

Choi, H. and Varian, H. R. (2012). Predicting the Present with Google Trends. *Economic Record*, 88(s1):2–9.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110.

Cochrane, J. H. (1991). A critique of the application of unit root tests. *Journal of Economic Dynamics and Control*, 15(2):275–284.

Croushore, D. (2006). Forecasting with Real-Time Macroeconomic Data. In Elliott, G., Granger, C. W. J., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 961–982. Elsevier.

D'Amuri, F. (2009). Predicting unemployment in short samples with internet job search query data. *MPRA Working Paper 18403*.

D'Amuri, F. and Marcucci, J. (2012). The Predictive Power of Google Searches in Forecasting Unemployment. *Bank of Italy Working Paper 891*.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests. *Journal of Business & Economic Statistics*, 33(1):1–24.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 20(1):134–144.

Einav, L. and Levin, J. D. (2013). The Data Revolution and Economic Analysis. *NBER Working Paper 19035*.

Fondeur, Y. and Karamé, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, 30:117–125.

Giannone, D., Reichlin, L., and 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., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–14.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America*, 107(41):17486–90.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Koop, G. and Onorante, L. (2013). Macroeconomic Nowcasting Using Google Probabilities. *Working Paper, University of Strathclyde and ECB*.

Koop, G. and Potter, S. M. (1999). Dynamic Asymmetries in U.S. Unemployment. *Journal of Business & Economics Statistics*, 17(3):298–312.

Kroft, K. and Pope, D. G. (2014). Does Online Search Crowd Out Traditional Search and Improve Matching Efficiency? Evidence from Craigslist. *Journal of Labor Economics*, 32(2):259–303.

Kuhn, P. and Mansour, H. (2014). Is Internet Job Search Still Ineffective? *Economic Journal*, 124(581):1213–1233.

Kuhn, P. and Skuterud, M. (2004). Internet Job Search and Unemployment Durations. *American Economic Review*, 94(1):218–232.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203–1205.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Lütkepohl, H. and Xu, F. (2012). The role of the log transformation in forecasting economic variables. *Empirical Economics*, 42(3):619–638.

Mahmoud, E. (1984). Accuracy in Forecasting: A Survey. *Journal of Forecasting*, 3(2):139–159.

McLaren, N. and Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, Q2:134–140.

McLeod, A. I. and Li, W. K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis*, 4(4):269–273.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., and Tiao, G. C. (1998). Forecasting the U.S. Unemployment Rate. *Journal of Americal Statistical Association*, 93(442):478–493.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–708.

Newey, W. K. and West, K. D. (1994). Automatic Lag Selection in Covariance Matrix Estimation. *Review of Economic Studies*, 61(4):631–653.

Pavlicek, J. and Kristoufek, L. (2014). Can Google searches help nowcast and forecast unemployment rates in the Visegrad Group countries? *Working Paper, Charles University*.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464.

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999). Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33(1):6–12.

Stevenson, B. (2008). The Internet and Job Search. *NBER Working Paper 13886*.

Suhoy, T. (2009). Query Indices and a 2008 Downturn: Israeli Data. *Bank of Israel Discussion Paper 2009.06*.

Tuhkuri, J. (2014). Big Data: Google Searches Predict Unemployment in Finland. *ETLA Reports 31*.

Varian, H. R. (2010). Computer Mediated Transactions. *American Economic Review: Papers & Proceedings*, 100(2):1–10.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2):3–36.

Vicente, M. R., López-Menéndez, A. J., and Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting & Social Change*, 92:132–139.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5):1067–1084.

Wu, L. and Brynjolfsson, E. (2015). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. In Goldfarb, A., Greenstein, S., and Tucker, C., editors, *Economic Analysis of the Digital Economy*, pages 89–118. University of Chicago Press.

A Appendix

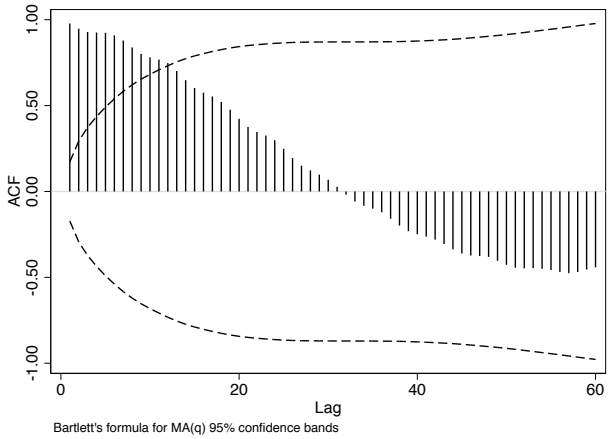

Figure A.1: The estimated autocorrelation function of the logarithm of the unemployment rate 2004–2014.

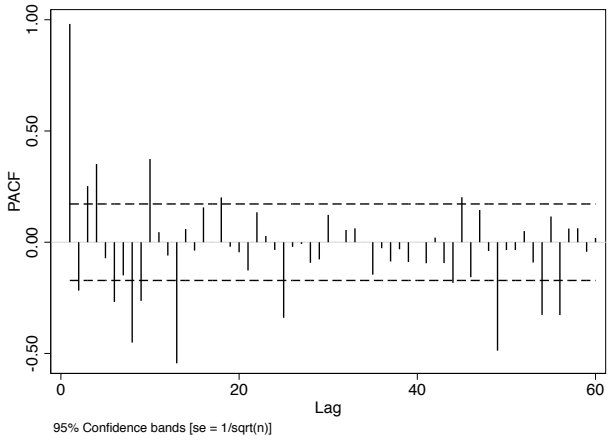

Figure A.2: The estimated partial autocorrelation function of the logarithm of the unemployment rate 2004–2014.

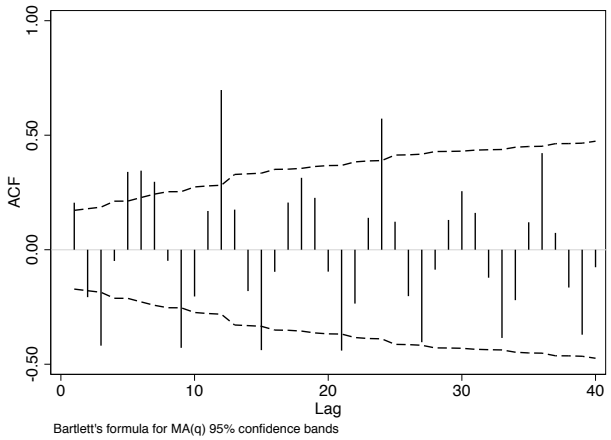

Figure A.3: The estimated autocorrelation function of the residuals for seasonal AR(1) model.

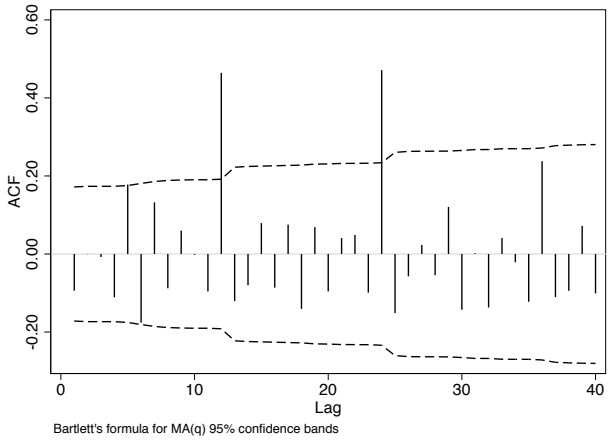

Figure A.4: The estimated autocorrelation function of the squared residuals for seasonal AR(1) model.

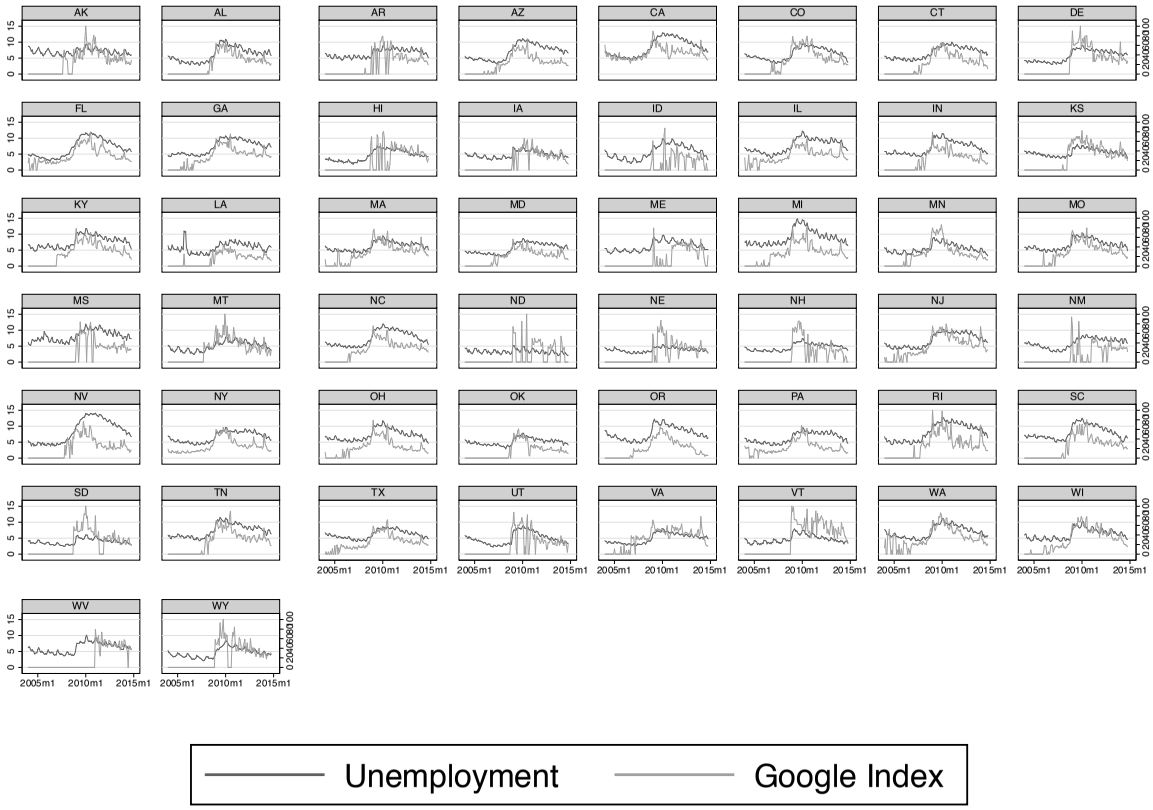

Figure A.5: Unemployment rate and Google Index in the United States 2004–2014. Source: The Bureau of Labor Statistics and *Google Trends*.