

Google Searches Predict Unemployment: How Far, When, and How Much?

Joonas Tuhkuri*

August 26, 2015

Abstract

Data on Google searches help predict the unemployment rate in the US. However, 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 modest. The results, obtained by (pseudo) out-of-sample forecast comparison, are robust to a state-level fixed effects model and not sensitive to different search terms. Joint analysis by cross-correlation function and Granger non-causality tests reveal that Google searches precede the unemployment rate. The results still demonstrate that big data can be used to forecast economic indicators.

*<joonas.tuhkuri@helsinki.fi>, University of Helsinki, Faculty of Social Sciences, Economics and ETLA, the Research Institute of the Finnish Economy. I would like to thank Daniel Ershov, Christian Gourieroux, Antti Kauhanen, Markku Lanne, Matti Mitrunen, Philip Oreopoulos, Petri Rouvinen, Tobias Ylömäki as well as Ari Ojansivu and Johanna Wahlroos at Google and participants at the NTTS 2015 conference session in Brussels and at seminars in ETLA, University of Helsinki, Ministry of Trade and Employment and at Aalto University Statistical Natural Language Processing course for conversations and helpful comments. Part of the article was written while I was visiting the Department of Economics at the University of Toronto. Financial support from the US State Department (ARC Grant) and University of Helsinki (International Exchange Scholarship and Research Station Grant) is gratefully acknowledged. ETLA, the Research Institute of the Finnish Economy kindly provided facilities for finishing the work. I thank participants of the ETLA now project. 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?

Official labor statistics are released on a monthly basis. However, the data are available with almost a one-month lag, and more timely estimate of the unemployment rate would be valuable. From a policy perspective, more accurate knowledge could inform better labor market and monetary policy decisions that might help workers—especially during an economic crisis.

In contrast to traditional statistics, data on Google searches are publicly available in real time. Real-time information might help *nowcast* the present unemployment rate, which is uncertain. Furthermore, Google search queries might be associated with the future expectations and help *forecast* the future unemployment rate. Sudden changes in Google search activity might also help identify sudden changes in the unemployment rate.

The main result from the earlier literature is unambiguous: Google data are found to predict unemployment. However, two general questions remain unanswered.

First, the previous literature does not tell how far into the future Google searches could predict the unemployment rate. Most previous studies on the topic have only conducted studies on assessing the current conditions with real-time search data (nowcasting), but not predicting the future (forecasting).

Second, earlier studies report it is possible to improve unemployment forecasts by using information on Google search volumes, but only on an average sense. What the majority of previous studies do not address, however, is whether the improvement in prediction accuracy is episodic or stable over time. The advantage from search data may be time specific, and the signal from search activity occasionally misleading—even if on average Google searches were useful.

This study extends the literature by addressing the two limitations. Answers to the questions have both practical and academic relevance. Furthermore, the limitations also have more general relevance since they are not widely discussed in the context of predicting other economic indicators with Internet search data either.

The paper uses (pseudo) out-of-sample forecast comparison to answer the two questions. It constructs a state-level panel data set to study the robustness of the results at a finer level, and provides descriptive joint analysis of the series to describe the intertemporal relationship between relevant Google searches and the unemployment rate. Neither of the latter approaches are provided in the previous studies. The main model contains a variable, Google Index, constructed from Google data using approximately 35 million² search queries related to searches for unemployment benefits. The underlying idea is that Google searches in these topics could be related to actual filings for unemployment benefits. This is the first study that uses statistics from the actual

¹Source: Google Internal Data, 2014.

²Source: Google AdWords, 2014.

search volumes on Google; Previous studies use only normalized variation within a keyword over time.

2 Literature

Recent work suggests that search query data might be useful in economic forecasting. The topic is new. However, since Ettredge et al. (2005), several studies have discussed the use of Internet search data in various contexts.

For example, the previous research suggests that Google searches could be useful in predicting influenza epidemics (Ginsberg et al. 2009), sales for video games (Goel et al. 2010), and the housing market (Wu and Brynjolfsson 2015). Choi and Varian (2009a, 2009b, 2012) use Google search data to predict economic indicators, such as initial claims for unemployment benefits and consumer confidence index. Their seminal work provides an overview on using Google data for short-term economic forecasting.

Studies pointing out that Google search volumes could be useful in predicting the unemployment rate span over several countries.³ However, the literature has not developed much from the seminal studies in Germany (Askatas and Zimmermann 2009) and in the US (Choi and Varian 2012). The main difference in the previous studies is that they have been performed in different countries.

The papers most closely related to this study are Choi and Varian (2012) and D'Amuri and Marcucci (2012) in the US. Their methodology consists of two steps: search term selection and forecast comparison. First, the authors select search terms that might describe individual labor market actions. The terms include searches for jobs (D'Amuri and Marcucci 2012) as well as for unemployment benefits and social and welfare topics (Choi and Varian 2012). Second, the studies compare (pseudo) out-of-sample forecasts from models that include relevant Google variables to univariate time-series models that exclude Google variables.

This paper is related to a several other strands of literature. Current studies document that the Internet plays an important role in the US labor market (see, for example, Stevenson 2008; Kroft and Pope 2014; and Kuhn and Mansour 2014). The Internet is used in various ways in job search, including contacting public employment agencies and submitting job applications (Kuhn and Mansour 2014), and Google searches might offer information particularly on the unemployment rate.

More generally, Varian (2010) points out that previously unrecorded activity is now recorded by computers. As a result, we get information on private actions on labor market through Internet search logs. Each search is an interest in or

³The countries include Germany (Askatas and Zimmermann 2009), US (Choi and Varian 2012; D'Amuri and Marcucci 2012), 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).

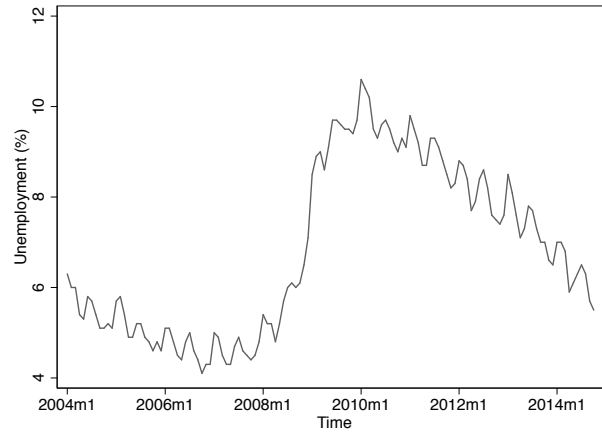


Figure 3.1: Unemployment rate in the United States 2004–2014. Not seasonally adjusted. Source: The Bureau of Labor Statistics.

demand for something (Brynjolfsson 2012), and the novel nanodata⁴ (Wu and Brynjolfsson 2015), arising as a by-product, might help improve unemployment forecasts.

Big data is a broad term that refers to massive data sets. The estimated amount of information created until 2003 was approximately 5 exabytes, and the same amount of information is now created every two days (Einav and Levin 2013, and the references therein). The broad question underlying this paper is whether big data can be used to improve economic forecasts. I approach the broad question by answering a more specific one: Do Google searches predict unemployment in the United States?

3 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 published by the US Bureau of Labor Statistics.

3.1 Unemployment

Figure 3.1 describes the evolution of the unemployment rate in the United States from January 2004 until October 2014.⁵ Typically, the unemployment rate exhibits seasonal variation, although that variation is not obvious from the

⁴The origin of the term dates back to Arrow (1987), who referred analysis of individual transactions as nanoeconomics.

⁵The unemployment data for this study was retrieved from the the US Bureau of Labor Statistics website on Dec 15th 2014.

Figure 3.1. 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 relatively sudden change between 2008 and 2010 associated with the economic crisis. The abrupt increase in unemployment was hard to predict, or at least, many of the predictions failed. New big data sources, such as Internet search data, might help produce more accurate forecasts.

3.2 Google

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

This section covers Google data and consists of two parts. I first select relevant search terms, then construct a variable, named *Google Index*, that describes search volumes for these terms.

I come up with 125 search terms that are related to unemployment benefits, and select 13 search terms with the highest search volumes.⁶ These search terms are: unemployment benefits, unemployment office, unemployment claim, unemployment compensation, unemployment insurance, apply for unemployment, applying for unemployment, filing for unemployment, unemployment online, unemployment office locations, unemployment eligibility, ui benefits, and unemployment benefit. This is the highest amount of search terms that *Google Trends* database allows to export on one session. Dividing the export to multiple sessions would not allow to use boolean search operators (Silverstein et al. 1999) later in constructing a variable from the search volumes. I explore the sensitivity of the results to the selected search terms in Section 6.

From 2004 to 2014, there were approximately 270, 000 monthly search queries with the selected search terms.⁷ In other words, the analysis is based on approximately 35 million Google searches.⁸ 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.

To my knowledge, this is the first study that uses the actual Google search volumes in variable selection. This is done by combining two data sets: data for the actual volumes come from *Google AdWords*, and the time-series for the analysis come from *Google Trends*, which only reports an index of search intensity. The advantage of the approach is to be able to select the most salient search terms.

The selected search terms are specifically related to unemployment benefits, because these are likely to be the first searches that a displaced worker types.

⁶Source: Google AdWords, 2014.

⁷Source: Google AdWords, 2014.

⁸This is an approximate number. The volume of Google searches have been increasing over time, but the search volumes from the early years are not available.

Furthermore, particularly the unemployed are likely to search for unemployment benefits. Besides that, for instance searches for jobs, in contrast, might increase for many reasons that are not related to unemployment. Previous research from the United Kingdom (McLaren and Shanbhogue 2011) and Germany (Askitas and Zimmermann 2009) suggests that searches for unemployment benefits have the potential to predict the unemployment rate. I explore the potential further. The previous study for the US (D’Amuri and Marcucci 2012) does not utilize search terms related to unemployment benefits but only one term: jobs.

One point is worth emphasizing. With billions of potential predictors⁹ and no clear guidance from economic theory, overfitting is a serious concern. I do not try to find the best forecast method or set of keywords for the US unemployment rate, but answer if real-time Google search volumes could help in the task. Google data does not have to be the best in order to be useful.

The following part describes the construction of a variable, which I give a name *Google Index*, from the selected search terms. Google Index represents aggregate search activity for the selected unemployment-related search queries, and its ability to predict the unemployment rate is possible to test in an econometric model.

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 the method is it gives 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 has been 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 to a monthly level.

In summary, let $K_{t,i}$ denote the amount of searches with a set of keywords k for a given geography i and time period t , where $t = 1, 2, \dots, f$. Let also $G_{t,i}$ denote the total amount of search queries in geography i at time t .¹⁰ Then the unit of measurement for search intensity $I_{t,i}$ of the Google Index is

$$I_{t,i} = \left\{ \frac{\frac{K_{t,i}}{G_{t,i}}}{\max_t \left(\frac{K_{t,i}}{G_{t,i}} \right)} \right\} \times 100, \quad (3.1)$$

where

$$\begin{aligned} K_{t,i} &\in (K_{1,i}, K_{2,i}, \dots, K_{t,i}, \dots, K_{f,i}) \\ G_{t,i} &\in (G_{1,i}, G_{2,i}, \dots, G_{t,i}, \dots, G_{f,i}). \end{aligned}$$

⁹For example, there are over 15 billion new search terms searched on Google every month. Source: Google Internal Data, 2014.

¹⁰For some purposes one could define analogously g as a set of all possible search queries where $k \in g$.

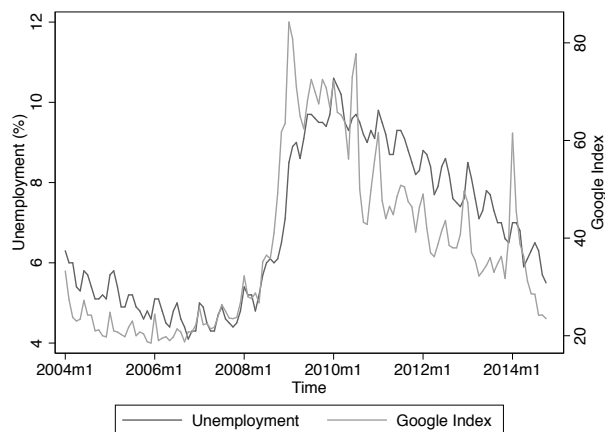


Figure 3.2: Unemployment rate and Google Index, which describes search activity for unemployment benefits in the United States 2004–2014. Source: The Bureau of Labor Statistics and *Google Trends*.

Figure 3.2 describes the evolution of the Google Index and the unemployment rate from January 2004 until October 2014.¹¹ The search intensity $I_{t,i}$ for selected unemployment-related searches compared to other searches exhibits no clear trend between 2004 and 2008. After 2008, there is a sudden increase, possibly related to the economic crisis. Following the initial increase, there are several spikes that may reflect suddenly increased interest in unemployment benefits, for example, because of changes in the unemployment benefit system. Possible events might include discussions and news about the extension of unemployment benefits. For example, President Barack Obama signed the Unemployment Compensation Extension Act of 2010 into law in July 2010. Figure 3.2 shows that there is a rapid increase in search activity around July 2010. The second spike coincides with the Congress ending the same act rather abruptly in January 2014. I repeat the analysis of this study while also controlling for the two spikes.

The series seem to behave in a similar manner; the correlation between monthly unemployment and the Google Index is 0.87. The Google Index does not seem to exhibit any clear seasonality. Table 3.1 gives descriptive statistics for the Google Index and the unemployment rate.

¹¹Google data for this study was retrieved on December 12th 2014. Formally for the US federal-level Google Index k is the 13 search terms, $i = USA$, starting point is 1/2004 and length of period $f = 130$.

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

4 Methods

This section presents the main methods used in this paper to answer whether Google searches predict unemployment. To be clear, there are two distinct questions closely related to the topic. The first considers the joint analysis of the two series. The second asks, how to improve short-term forecasts of the unemployment rate. I focus on the latter question by comparing the performance of models for forecasting US unemployment with and without Google data but also provide descriptive joint analysis of the series.

4.1 Joint Analysis

I analyze the series jointly by performing Granger (1969) non-causality tests and by studying the cross-correlation function. The analysis of 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 allows also to study whether the level of unemployment rate predicts the volumes of Google searches over and above its own history. If not, then Google data might offer genuinely new information on the unemployment rate.

4.2 Model

In this section I present the main models, which are used for answering whether Google searches predict unemployment. In specific, I am interested in finding out about the incremental predictive ability of Google Index over and above that of the own history of the unemployment rate. The models in this section also aim to answer two subquestions. First, how far into the future is it possible to improve unemployment forecasts by using Google data? Second, are the potential improvements in forecasting accuracy constant or do they vary over time?

The first step is to construct a relevant benchmark model for the unemployment rate. The benchmark model is extended with the Google Index, and the models and their forecast performance are compared. This type of model and (pseudo) out-of-sample forecast comparison methodology is associated with

West (1996) and Clark and McCracken (2001). In context of Google data, the approach traces to the work of Choi and Varian (2009a, 2009b, 2012) and Goel et al. (2010). However, Diebold (2015) notes that (pseudo) out-of-sample forecasts do not, for example, provide protection against overfitting. The selected method is not the last word on the topic.

I limit the set of potential benchmark models to autoregressive (AR) models. The idea is to impose as little structure as possible to minimize assumptions. Furthermore, starting from 2004, there are only 130 monthly observations of Google data available, which means that complicated or high-order models are not necessarily estimated accurately. With only a limited amount of data at hand, overfitting is a serious concern (see, for instance, Varian 2014). A simple univariate autoregressive model also is a common benchmark in the forecasting literature. Empirical research has shown that simple models often yield better out-of-sample predictions than complex models (Makridakis et al. 1979; Mahmoud 1984).

I use both variables, the unemployment rate and the Google Index, 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, during the last one hundred years, the US unemployment rate had no visible trend, and economic theory does not suggest it should have had one (Cochrane 1991, Montgomery et al. 1998).

I include a seasonal autoregressive term y_{t-12} to 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 for example Choi and Varian (2012) and Wu and Brynjolfsson (2015) apply the same approach. Additionally, I perform a logarithmic transformation for the unemployment series since changes in unemployment rate are most naturally discussed in percentage terms, and because logarithmic transformation helps stabilize the variance of the series (Lütkepohl and Xu 2012).

The task is to find the autoregressive AR(p) model that sufficiently captures the dynamics of the given time series. My approach relies on the Box-Jenkins methodology (Box et al. 2008, Chapter 1) in the sense that it consists of three steps: identification, estimation, and diagnostics. However, in application to a short sample, part of the traditional Box-Jenkins tools for model specification need to be applied with caution.

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 seems to be a cutoff at the 13th lag, after which the partial autocorrelations remain statistically insignificant at the 5% level.

Using sequential testing, the first AR(p) model that has a statistically significant p :th coefficient at the 5% level is AR(13). The lag order p is also estimated using the Akaike (AIC) and Schwarz (BIC) information criteria. Both criteria give the smallest value for AR(13) when using maximum lag of 20 and including

a seasonal lag for every model. One explanation for this, is that the seasonal lag might not be able to accommodate the seasonality in the series. Furthermore, both information criteria decrease almost monotonously until the 13th lag. However, in order to use the AR(13) model as a benchmark, I have to estimate 14 coefficients, while there are only 130 observations in the unemployment series. In part for this reason, the AR(13) is not reasonable as the only benchmark for a out-of-sample forecast comparison.

My solution is following. I use a naïve seasonal AR(1) benchmark for the main specification. That model uses only the previous period and seasonal effects to predict the unemployment rate. There are five reasons for this.

First, a simple model serves as a first test to ascertain whether Google data offer any advantage on 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 ACF 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} .¹²

Third, judging by the determinant coefficient of determination, the R^2 is already 0.96 for the seasonal AR(1) model. This implies that additional lags are not able to improve the fit of the model considerably.

Fourth, during the observation period from 2004 to 2014, the evolution of the US unemployment rate was dominated by an abrupt increase followed by 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.

Fifth, Montgomery et al. (1998) suggest a first order autoregressive model for short-term unemployment forecasting. They point out that this specification is also commonly used to model the unemployment rate.

However, 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 is adequate to describe our data-generating process (DGP) by providing several diagnostic checks.

I estimate the seasonal AR(1) model by 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 might 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 evaluate formally whether most of the temporal dependence has been removed from the residuals I compute the Ljung–Box (1978) portmanteau statistic for the residuals. The portmanteau test statistic Q_K computed with $K = 12$

¹²In the previous literature on forecasting with Google data, Choi and Varian (2012) use this argument to motivate the use the AR(1) benchmark.

and $K = 24$ lags does reject the null hypothesis of no serial correlation at 1% level. Furthermore, the Ljung–Box Q_K test statistic calculated with $K = 12$ and $K = 24$ lags rejects the null hypothesis of no autocorrelation in the residuals at 1% significance level for every $AR(p)$ model until the 13th order AR model. The reason for this is possibly that the 12th lag, which was included in every model, is not capable to accommodate 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 rather 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.

I also estimate alternative models up till fourth order seasonal AR model, and find that when limiting to lower than fourth-order seasonal AR models, the higher order models do not give clear advantage against the seasonal AR(1) model, judging by the estimated autocorrelation functions of the residuals. For the reasons above, among lower than fourth-order autoregressive models, the seasonal AR(1) model seems to be adequate benchmark, however not perfect.

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 sensitivity of the results for the selected benchmark, I also estimate the results using seasonal AR(2) and AR(3) benchmark models in Section 6.

To make a long story short, I use the seasonal AR(1) model as a benchmark because the more complicated benchmark models do not necessarily offer a marked advantage against the simple one.

4.2.1 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. The models are nested and linear.

I account for the remaining autocorrelation in the residuals by using heteroskedasticity- and autocorrelation-consistent (HAC) standard errors developed

by Newey and West (1987). The number of lags for the robust standard errors is selected by method proposed by Newey and West (1994) for every estimation window. The models are estimated by the quasi-maximum likelihood (QML) method under the normality assumption.

There is a reason for caution 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—I have to demonstrate that the model with the Google Index performs at least better than a benchmark model using lagged data and seasonal effects (Varian and Stephens–Davidowiz 2014; Goel et al. 2010). In a traditional non-time series regression framework, one could think of the Google Index x_t as a regressor of interest and lagged unemployment data y_{t-1} as well as seasonal effects of unemployment y_{t-12} as control variables.

I start by reporting estimation results from the entire observation period. These results could provide some evidence on the fit of the benchmark and extended models and give information on the statistical properties of the US unemployment rate. I compare the fit of the models measured by coefficient of determination R^2 , as well as other properties, such as Akaike (1973) and Bayesian (Schwarz 1978) information criteria, statistical significance, and the magnitude of the parameters.

To answer whether Google searches could help to forecast the unemployment, 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 both models (0.0) and (1.0). For each month from 2008, I train the model using past 48 observations, and then evaluate the out-of-sample predictions by comparing the forecasted values to the realized values of the unemployment rate. The window of 48 months is chosen to make sure that there are enough observations to estimate the models, and that the evaluation period is long enough.

I use mean absolute percentage error (MAPE) as a measure of forecasting accuracy. It is defined as

$$\text{MAPE} = \frac{1}{T} \sum_{t=1}^T |E_t|, \quad (4.1)$$

where

$$E_t = \frac{\hat{y}_t - y_t}{y_t} \times 100,$$

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 selected error measure with mean squared error in Section 6.

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).

4.2.2 Forecasting the Future

So far we have considered only predicting the present, that is, nowcasting. In this section, I present methods to study whether Google searches also predict the unemployment rate in the near future. This would be the case if searches for unemployment benefits now would help to predict the future unemployment rate.

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.

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

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

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

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

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

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

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

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

The unemployment rate is denoted by y_{t-k} and the Google Index by x_{t-l} , where the subscripts refer to the date of observation. Coefficients and constant terms are denoted by β :s with different subscripts, while error terms are denoted by e_t . I utilize heteroskedasticity- and autocorrelation-consistent (Newey and West 1987; 1994) standard errors as earlier.

The models are estimated by the quasi-maximum likelihood (QML) method under normality assumption, and 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 Google Index at time t for the last forecast horizon. This study uses dynamic forecasts instead of static ones 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 by using the same methodology described earlier in Section 4.2.1. In specific, if a model that includes the Google Index provides more

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

accurate forecasts than a benchmark model in the (pseudo) out-of-sample environment for horizon h but not for $h + 1$, I infer that the marginal predictive ability of Google searches is limited to horizon h predictions. On the other hand, Google searches might then help to forecast the future unemployment rate h steps ahead.

4.2.3 Time-specific Forecasts

The value of Google data for forecasting purposes may depend on the date of the forecasts. For example, real-time data might be especially useful during a recession when the economic indicators are usually harder to predict. From a practical forecasting perspective, this is an important criterion for the relevance of 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 this improvement happens.

5 Results

5.1 Joint Analysis

5.1.1 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 5.1 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

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 5.2: Statistics for testing Granger non-causality.

be larger than the that of the opposite case. The implication is that Google search volumes tend to anticipate the US unemployment rate. In other words, the variables are interconnected, but the Google Index presents the pattern of a classical leading indicator, that is, the Google search volumes are better correlated with the future than current unemployment.

A closer look at the cross correlations reveals that the historical cross-correlation function attains it maximum at the sixth lag of the Google Index. Therefore, 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 autoregressive model by introducing Google search volumes among the regressors.

5.1.2 Granger Causality

Do Google searches Granger cause unemployment? Table 5.2 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 un-

employment rate do not offer useful information in predicting Google searches over and above the Google series themselves. This suggests that Google searches might 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.

5.2 Model

5.2.1 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 5.3. The coefficient for the Google Index is statistically significant at the 1% level.¹³ 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 current unemployment rate.

The R^2 for model (0.0) is 0.962, which means that the benchmark model can alone explain a large part of the variation in the unemployment rate, as suggested before (Goel et al. 2010). Including the Google Index increases the R^2 , although not markedly. Issues associated with interpreting the R^2 are well known. 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 of 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 5.1. The mean absolute percentage errors for nowcasts are given on the first row of Table 5.4.

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 most apparent reasons for this. First, the observation period is short, and there are only 130 monthly observations. Thus, the power of the test is low (Diebold 2015). In other words, the test may fail to reject the null hypothesis even if the alternative were true. 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 benchmark represents a more general difference “in population” or merely an observation “in sample.”

¹³The potential loss of power from using HAC standard errors does not overshadow the statistical significance of the coefficient.

Model	(0.0)	(1.0)
<hr/>		
Variables		
$\log(y_{t-1})$	0.983** (0.0295)	0.955** (0.0356)
$\log(y_{t-12})$	-0.0103 (0.0300)	0.0156 (0.0368)
x_t		.00440** (0.000656)
Constant	1.848** (0.191)	1.692** (0.150)
<hr/>		
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). The sample period is Jan 2004–Oct 2014.

Table 5.3: Estimation results of the benchmark seasonal AR(1) model (0.0) and the extended model (1.0), which includes Google Index.

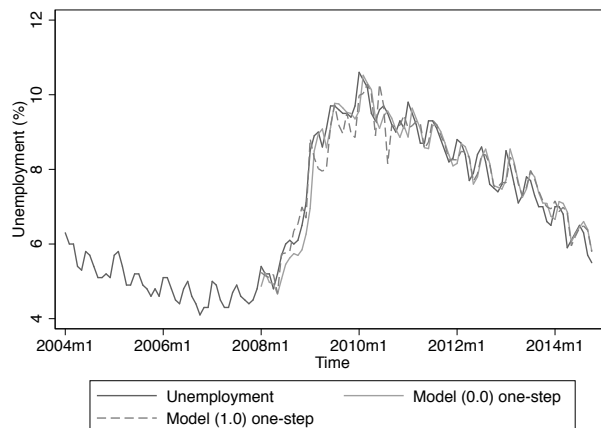


Figure 5.1: Unemployment rate 2004–2014 and the one-step-ahead nowcasts with a rolling window of 48 months for the univariate benchmark model (0.0) and the extended model (1.0), which includes Google Index 2008–2014.

5.2.2 Forecasting the Future

Do Google searches help forecast the future unemployment rate? Table 5.4 summarizes the mean absolute percentage errors of out-of-sample dynamic forecasts up to the horizon $h = 6$. We can see from Table 5.4 that the two-step-ahead forecasts improve 7.48 percent on average when we add in the Google data, compared to 4.32 percent improvement for the one-step-ahead forecasts.

However, if we predict the unemployment rate two months ahead we get a decline of 3.92 percent in forecast accuracy. Notice that nowcasts are on average more accurate than forecasts, as they should be. Increasing the forecast horizon decreases the forecasting accuracy for both models.

The alleged mechanical relationship between volumes of relevant Internet searches and the level of unemployment rate would suggest that the forecasts are limited to short-term predictions. Results in the Table 5.4 support this idea. 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.

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 reports at the 10% level no statistically significant differences between the forecasts.

The descriptive cross-correlation analysis suggests 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 compar-

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 5.4: 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

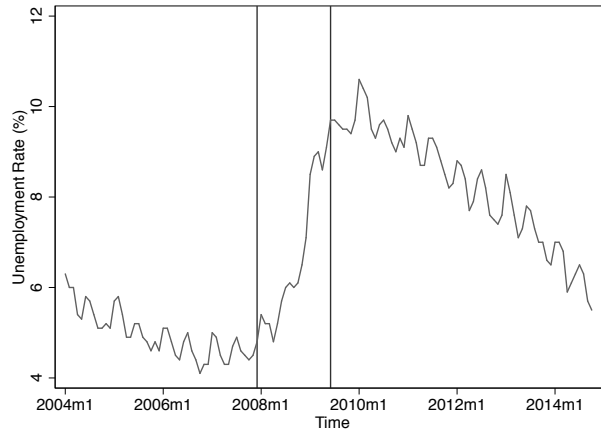


Figure 5.2: The recession.

ison 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 following decrease in the series. The Google search activity peaks six months before the initial increase in unemployment in 2008. However, the lead-lag relationship might not be consistent.

5.2.3 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. This recession happened from December 2007 until June 2009 and lasted for 18 months. The vertical lines in Figure 5.2 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.¹⁴

However, previous studies by Choi and Varian (2012) and Goel et al. (2010) conjecture that sudden changes in search intensity could help identify sudden changes in economic time series. Table 5.5 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

¹⁴Source: 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 5.5: 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.



Figure 5.3: 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.

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. On the other hand, the 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 5.5 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 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 5.3 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. Google search queries improve the prediction accuracy especially during the 2007–2009 recession in the United States.

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. Despite the benefits of Google data, including the Google Index as an additional predictor to the benchmark model occasionally makes the out-of-sample predictions not better but worse.

6 Robustness

6.1 Panel Data

I manually 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.

I estimate the following fixed effects model with lagged dependent variables:

$$\log(y_{i,t}) = \beta_1 \log(y_{i,t-1}) + \beta_2 \log(y_{i,t-12}) + \beta_3 x_{i,t} + \alpha_i + e_{t,i}, \quad (6.1)$$

where $i = 1, \dots, 50$ and $t = 1, \dots, 118$. Each state is denoted by i . The fixed effects model has 50 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, unemployment rate is denoted by $y_{i,t}$ and Google Index by $x_{i,t}$.

I account for 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, and check the results also by employing a within estimator using the ordinary least squares (OLS) method.

I am able to exploit the geographic and temporal variation in level of the unemployment rate induced by the 2008 economic crisis. The unemployment rate and Google searches have somewhat different patterns in each 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 increase earlier than in North Dakota. To illustrate this further, a map displayed in Figure 6.1 visualizes the US state-level differences in the popularity of unemployment-related Google searches between November 2009 and February 2010.

The results from state fixed effects model are given in the second column of Table 6.1 with earlier results of the extended autoregressive model in the first column. In summary, the coefficient of the Google Index is significant at 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

Model	(1.0)	FE (AB)	FE (OLS)
<hr/>			
Variables			
$\log(y_{t-1})$	0.955** (0.0356)	0.825** (0.00555)	0.832** (0.0062)
$\log(y_{t-12})$	0.0156 (0.0368)	0.0678** (0.00442)	0.0673** (0.00499)
x_t	.00440** (0.000656)	.00176** (0.000058)	.00167** (0.000066)
Constant	1.692** (0.150)		
<hr/>			
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 5.3. The sampling period is Jan 2004–Oct 2014.

Table 6.1: Estimation results of the extended autoregressive model (1.0) and the fixed effects model (FE).

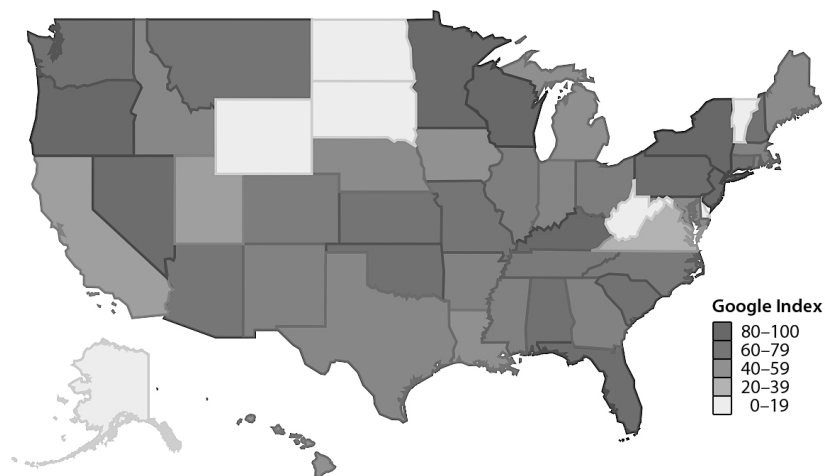


Figure 6.1: Relative popularity of unemployment-related Google searches. Average between Nov 2009 and Feb 2010. Source: *Google Trends*.

controlling for the state-level lagged and seasonal effects. The pattern—Google searches predict unemployment—seems to be repeated at state level.

The estimation results from within estimator, reported in the third column of Table 6.1, are similar to that of the Arellano-Bond method. The coefficient of the Google Index is statistically significant at 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 unobserved variables are. It is also relevant to note that this type of model is usually used for analyzing causal effects. Google searches hardly cause the unemployment.

In practice, utilizing 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.

6.2 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 2014.¹⁵ 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 the changes in search behavior. Two spikes in search activity, depicted in Figure 3.2, were presumably associated with changes labor market policy, not unemployment. After controlling for the two events, the improvements from Google data compared to the benchmark are 10 percent higher on average than the improvements reported earlier.

6.3 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 are slightly smaller than against the AR(1) benchmark, however.

The results using other commonly used error measure, mean squared error (MSE), are essentially the same as those from using mean absolute percentage error (MAPE). I 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.

7 Discussion

There are still some concerns. First of all, the improvements in prediction accuracy are only modest. This finding contrasts with earlier literature on the topic in the US. D’Amuri and Marcucci (2012) find 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. In terms of magnitude, my findings are more in line with modest improvements reported by Choi and Varian (2012).

Another concern is that the simple autoregressive models utilized in this paper sometimes provide reasonable predictions but occasionally produce very bad forecasts. Lazer et al. (2014) argue that the Google search algorithm is

¹⁵Source: *Google Trends*, 2014.

constantly changing, and it is hard to train the forecasting model using past data. My take is more directly targeted on the context of unemployment forecasting. The unemployment rate is a function of new cases, exits, and duration (see, for example, Barnichon and Nekarda 2012). The method in this paper may be 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 and discouraged workers.

There are further limitations for the results of this paper. The methods utilized in this paper are relatively simple and do not necessarily represent the ways actual forecasters would use this data. Our understanding on Internet search is limited, and interpreting changes in search volumes is difficult. Moreover, the observation period is short, and there is only one major increase and subsequent decrease in the unemployment rate. That is, based almost on one event, it is not clear whether this pattern would hold in the future.

A caveat also arises for practical implications. The econometric models with Google data may not be the best forecasting tools for the unemployment rate. Recent work surveyed by Snowberg et al. (2013) suggests that prediction markets, for example, could produce more accurate forecasts. The authors provide evidence that a prediction market is weakly more accurate than survey forecasts for initial unemployment claims.

A common criticism toward forecasting with big data is that with vast amounts of data, it is easy to mistake a noise for a signal. Is the finding 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 because individuals actually use the Internet as a tool in the labor market (Stevenson 2008; Kuhn and Mansour 2014). None of the methods utilized 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 contain useful information on the current and near future unemployment, and that information can be used to predict the US unemployment rate.

This paper only describes almost a mechanical relationship between Google searches for unemployment benefits and the actual 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 work by Baker and Fradkin (2014). Fine-grained Internet data allow us to measure individual actions that have been previously hard to measure. At the same time, the Internet and digitalization of the economy also create new activity. To understand these activities, Internet data sources such as Google search logs may prove beneficial.

8 Conclusion

This paper analyzes whether data on Google search volumes could help predict the unemployment rate. I have found that autoregressive models with relevant

Google variables tend to produce, on average, more accurate forecasts than the same models without those predictors. Joint analysis of the series suggests that changes in Google searches, which are related to unemployment benefits more often than not, precede changes in the unemployment rate. The results suggest that Google searches could help predict the present and near-future unemployment rate.

Two novel findings arise. First, improvements in predictive accuracy from using Google data appear to be limited to short-term predictions. Second, the informational value of search data tends to be time specific. The results appear quite robust to different model specifications and search terms, and Google search volumes are also associated with the unemployment rate on the US state level. I conclude that Google searches predict unemployment.

The qualitative results are in line with the previous findings on Google searches and unemployment by Askitas and Zimmermann (2009), McLaren and Shanbhogue (2011), Choi and Varian (2012), and D'Amuri and Marcucci (2012). Compared to previous results from the US (D'Amuri and Marcucci 2012), I find that the improvements in forecasting accuracy from Google data may be smaller than previously thought.

Considering the importance of accurate economic forecasts and the scale of Google data, this study answers an important question. The paper highlights the potential of Internet searches in predicting economic indicators, and I propose that Google searches could offer useful information for predicting the US unemployment rate in the short run. The results also demonstrate that big data can be utilized to forecast official statistics. However, it is important to emphasize that the predictive power of Google searches seems to be limited to relatively short-term predictions, and the improvements are modest.

More generally, 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. However, big data consist of 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 imagined yet.

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.
- Arrow, K. (1987). Reflections on the Essays. In Feiwel, G. R., editor, *Arrow and the Foundations of the Theory of Economic Policy*, volume 2, pages 727–734. Palgrave Macmillan.
- 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*.
- 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.
- Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2008). *Time series analysis: forecasting and control*. John Wiley & Sons., 4th edition.
- 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. (2009a). Predicting Initial Claims for Unemployment Benefits. *Technical Report, Google*.
- Choi, H. and Varian, H. R. (2009b). Predicting the Present with Google Trends. *Technical Report, Google*.
- 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.
- 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*.
- Ettredge, M., Gerdes, J., and Karuga, G. (2005). Using Web-based Search Data to Predict Macroeconomic Statistics. *Communications of the ACM*, 48(11):87–92.
- Fondeur, Y. and Karamé, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, 30:117–125.
- 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.
- Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review: Papers & Proceedings*, 105(5):491–495.
- Koop, G. and Potter, S. M. (1999). Dynamic Asymmetries in U.S. Unemployment. *Journal of Business & Economic 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.
- Makridakis, S., Hibon, M., and Moser, C. (1979). Accuracy of Forecasting: An Empirical Investigation. *Journal of the Royal Statistical Society. Series A (General)*, 142(2):pp. 97–145.
- 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 American 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.
- Snowberg, E., Wolfers, J., and Zitzewitz, E. (2013). Prediction Markets for Economic Forecasting. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, pages 657–684. Elsevier.

- 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.
- Varian, H. R. and Stephens-Davidowitz, S. (2014). Google Trends: A Primer for Social Scientists. *Working Paper, Google*.
- 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

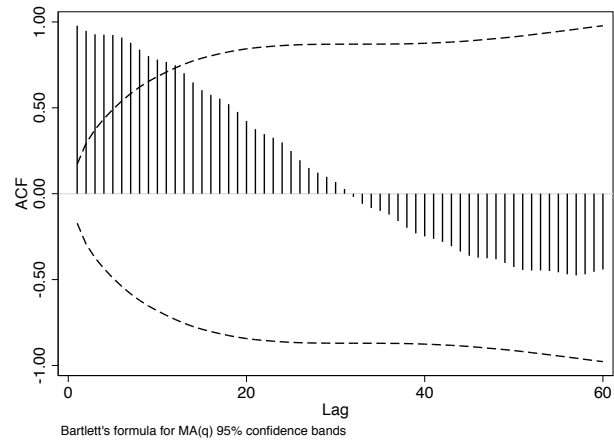


Figure A.1: The estimated autocorrelation function of the logarithm of the unemployment rate 2004–2014.

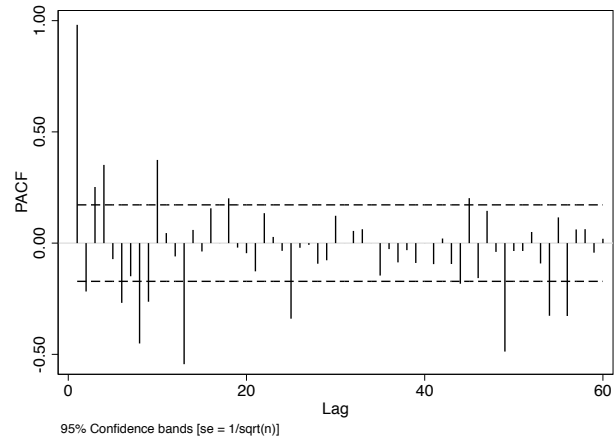


Figure A.2: The estimated partial autocorrelation function of the logarithm of the unemployment rate 2004–2014.

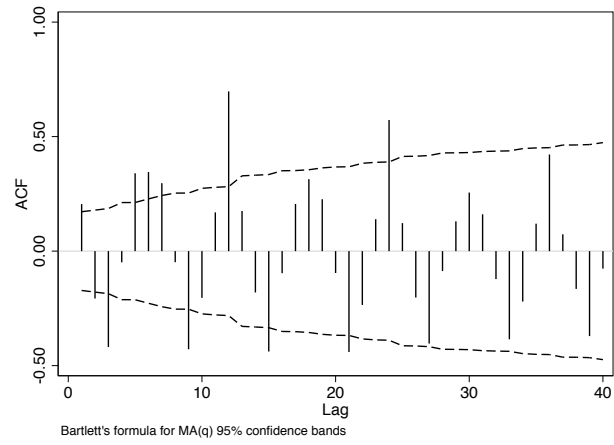


Figure A.3: The estimated autocorrelation function of the residuals for seasonal AR(1) model.

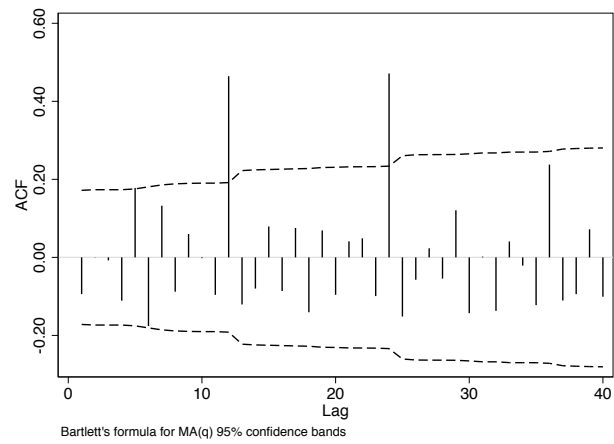
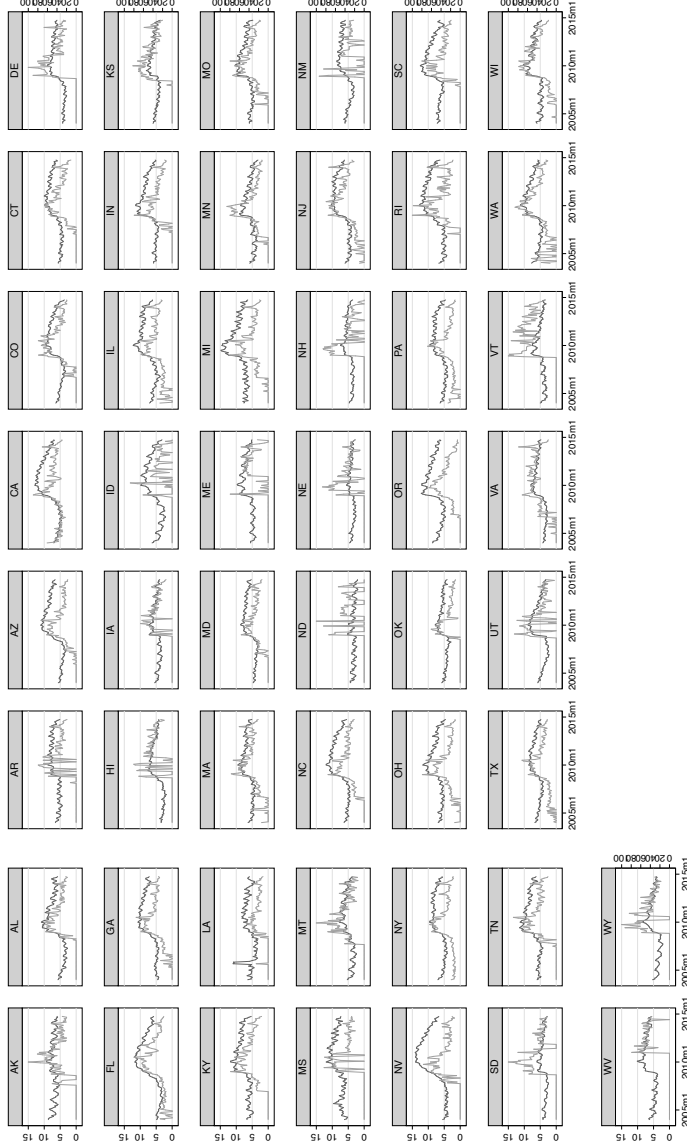


Figure A.4: The estimated autocorrelation function of the squared residuals for seasonal AR(1) model.



— Unemployment - - - Google Index

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