Big Data: Do Google Searches Predict Unemployment?

Joonas Tuhkuri

University of Helsinki
Faculty of Social Sciences, Economics

Master's Thesis

May 2015

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Faculty	Department
Faculty of Social Sciences	Department of Political and Economic Studies

Author
Joonas Tuhkuri

Title
Big Data: Do Google Searches Predict Unemployment?

Subject
Economics

Level	Month and year	Number of pages
Master's Thesis	May 2015	84

Abstract

There are over 100 billion searches on Google every month. This thesis examines whether Google search queries can be used to predict the present and the near future unemployment rate in the US. Predicting the present and near future is of interest, as the official records of the state of the economy are published with a delay. To assess the information contained in Google search queries, the thesis compares a simple predictive model of unemployment to a model that contains a variable, Google Index, constructed from Google data. In addition, descriptive cross-correlation analysis and Granger non-causality tests are performed. To study the robustness of the results, the thesis considers state-level variation in the unemployment rate and Google Index using a fixed effects model. Furthermore, the sensitivity of the results is studied with regard to different search terms. The results suggest that Google searches contain useful information on the present and the near future unemployment rate. The value of Google data for forecasting purposes, however, tends to be time specific, and the predictive power of Google searches appear to be limited to short-term predictions. The results demonstrate that big data can be utilized to forecast economic indicators.

Keywords
big data, Google, Internet, nowcasting, forecasting, unemployment, time-series analysis

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto	Laitos
Valtiotieteellinen tiedekunta	Politiikan ja talouden tutkimuksen laitos

Tekijä
Joonas Tuhkuri

Työn nimi
Big data: Ennustavatko Google-haut työttömyyttä?

Oppiaine
Taloustiede

Työn laji	Aika	Sivumäärä
Pro gradu -tutkielma	Toukokuu 2015	84

Tiivistelmä

Google-hakuja tehdään kuukausittain yli 100 miljardia. Tämä tutkielma selvittää, voiko Google-hauilla ennustaa nykyhetken ja lähitulevaisuuden työttömyyttä Yhdysvalloissa. Nykyhetken ja lähitulevaisuuden ennustaminen on kiinnostavaa, sillä viralliset tiedot talouden tilasta julkaistaan viiveellä. Google-hakujen sisältämän informaation arvioimiseksi tutkielmassa vertaillaan yksinkertaista työttömyyttä kuvaavaa mallia sellaiseen malliin, johon on lisätty Google-aineistosta muodostettu muuttuja, Google Index. Tämän lisäksi tarkastellaan muuttujien välisiä ristikorrelaatioita ja suoritetaan Granger-kausaalisuustesti. Tulosten herkkyyden tarkastelemiseksi tutkielmassa tarkastellaan osavaltiotason vaihtelua työttömyydessä ja Google Indexin arvoissa hyödyntäen kiinteiden vaikutusten mallia. Tämän lisäksi tarkastellaan tulosten herkkyyttä valittujen hakusanojen suhteen. Tuloksen viittaavat siihen, että Google-haut sisältävät hyödyllistä informaatiota nykyhetken ja lähitulevaisuuden työttömyydestä. Google-hakujen sisältämän informaation arvo vaikuttaa kuitenkin vaihtelevan ajanhetkestä riippuen ja sen ennustekyky näyttää rajoittuvan lyhyen aikavälin ennusteisiin. Tulokset kuitenkin osoittavat että big dataa voidaan hyödyntää taloudellisten indikaattorien ennustamiseksi.

Avainsanat
big data, Google, Internet, nykyhetken ennustaminen, ennustaminen, työttömyys, aikasarja-analyysi

Contents

1 Introduction **1**

2 Literature **4**

3 Data **10**

4 Methods **21**

5 Results **41**

6 Robustness **55**

7 Discussion **62**

8 Conclusion **71**

Bibliography **73**

A Appendix **83**

Chapter 1

Introduction

There are over 100 billion searches on Google every month.[1] Could data from Google searches help to 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. More timely estimate of the unemployment rate would be valuable, especially during an economic crisis. From a policy perspective, more accurate knowledge could inform better decisions that might help to reduce unemployment. Moreover, the unemployment rate is a good indicator for the state of the economy.

However, data on Google searches are publicly available in real time. Real-time information might help to nowcast the present unemployment rate, which is uncertain. Furthermore, Google search queries might be associated with the future expectations and thus help to forecast the future unemployment.

In this thesis, I utilize simple autoregressive models to find out whether it is possible to predict the US unemployment rate using Google searches. I also perform descriptive joint analysis of the series to describe the intertemporal relationship between relevant Google searches and unemployment. Furthermore, I construct a state-level panel data set to study the robustness of the results. I explore the sensitivity of the findings to the selected search terms and compare the results with results from different countries. The main model contains a variable, Google Index, constructed from Google data using approximately 35 million[2] search queries, which are related to searches for unemployment benefits. The underlying idea is that Google searches in these topics might be related to actual filings for unemployment benefits. I focus on very short-term predictions.

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). That is why Google searches might be able to offer information especially on the unemployment rate. For example, Kuhn and Mansour (2014) document that the proportion of the unemployed in the United States who looked for work online was high during 2008–2009. The Internet is used in various ways

[1] Source: Google Internal Data, 2014.
[2] Source: Google AdWords, 2014.

in job search, including contacting public employment agencies and submitting job applications (Kuhn and Mansour 2014). Kroft and Pope (2014) point out that online portals, such as Craigslist, have become important platforms for job search. They find, for example, that the introduction of Craigslist significantly lowered the amount of classified job advertisements in newspapers. That is, job search has shifted to the Internet. Stevenson (2008) also documents this shift in the United States.

More generally, Varian (2010) points out that previously unrecorded activity is now recorded by computers. For example, we get information on private actions on labor market through Internet search logs. From an economic perspective, each search is an interest in or demand for something (Brynjolfsson 2012). This novel nanodata[3] (Wu and Brynjolfsson 2015) might help to improve unemployment forecasts.

One of the motivations to use timely data from Google is that the first official releases of the unemployment rate are released with a lag. In this sense, this thesis is closely related to the more general literature on macroeconomic monitoring and real-time data analysis (see, e.g., Croushore 2006 and Bańbura et al. 2013, and the references therein). Real-time assessment of current macroeconomic activity is also called *nowcasting* (Giannone et al. 2008). The underlying idea is that real-time data could help to nowcast the current level of an economic indicator.

However, "nowcasting is not just contemporaneous forecasting" (Castle et al. 2009, pp. 87), but it also poses interesting econometric challenges. These include, but are not limited to, mixed frequency estimation and variable selection (Choi and Varian 2012). The solutions for these puzzles might be of interest for researchers in many fields of economics. Furthermore, Oh and Waldman (1990), Rodríguez Mora and Schulstad (2007), and Castle et al. (2009) illustrate the importance of timely information by suggesting that the initial estimates of economic data could even have a significant impact on the actual economic activity. Lanne (2007) provides an overview on the relevance of economic statistics for empirical macroeconomic research and the economy.

New real-time data sources could also have practical relevance for several economic agents. Central banks are interested in acquiring real-time information on the economy (Bańbura et al. 2013; Aruoba and Diebold 2010). Recently, several central banks have shown interest in using Internet search data for economic forecasting. McLaren and Shanbhogue (2011) illustrate the Bank of England's efforts to use Google data for predicting the unemployment rate and housing prices, while Artola and Galan (2012) of the Bank of Spain provide an example of predicting inflows of British tourists to Spain. The central banks of Israel (Suhoy 2009), Italy (D'Amuri and Marcucci 2012), and Turkey (Chadwick and Sengul 2012) have also explored using Google search volumes in predicting economic indicators.

Indeed, the real-time data from Google might turn out to be useful for central

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

banks. The central banks adjust their monetary policy partly based on the most recent unemployment figures as well as on their unemployment forecasts. A more accurate assessment of the current state of economy, including the level of unemployment, could help to execute more effective monetary policy.

Several other government institutions, such as unemployment offices, would be better of if they had more timely information on unemployment. Up-to-date economic statistics are also important for many businesses (McAfee and Brynjolfsson 2012). This thesis presents methods that businesses might apply in order to use Google search data for predicting various metrics.

This thesis relates also to a long line of previous research on unemployment forecasting. Previous approaches include, but are not limited to, time series models (see, for instance, Montgomery et al. 1998; Koop and Potter 1999; Papell et al. 2000, and the references therein), factor models using a large number of macroeconomic variables (e.g., Stock and Watson 2002), macroeconomic models maintained by central banks and other institutions, and more recently the flow model proposed by Barnichon and Nekarda (2012). Closest to this thesis are methods that extend time series models with exogenous variables such as initial jobless claims (see Montgomery et al. 1998, for an example).

Big data is a broad term that refers to massive data sets. A way to illustrate the magnitude of these new data sets is to note that the estimated amount of original stored information until 2003 was approximately 5 exabytes, and that the same amount of information is now created every two days (Einav and Levin 2013, and the references therein). For this thesis, big data also means new measurement. We are able to measure previously unmeasurable activity. The broad question underlying this thesis is whether big data can be used to improve economic forecasts. I approach the broad question by answering a more specific one. That is, do Google searches predict unemployment in the Unites States?

To be clear, I do not claim any clear causality in this thesis. However, Google searches might offer a signal of the future unemployment rate. A new data set could also offer new insights on unemployment in the United States.

The remainder of the thesis is organized as follows. Chapter 2 offers a brief review of the relevant literature. Chapter 3 describes the data, and Chapter 4 explains the methods for answering the research question: whether Google searches predict unemployment. Subsequently, Chapter 5 presents the results. Chapter 6 explores the robustness of the findings, Chapter 7 discusses the results, and Chapter 8 concludes this thesis.

Chapter 2

Literature

Recent work suggests that search query data might be useful in economic forecasting. However, the topic is new and relatively little studied. To my knowledge, Ettredge et al. (2005) were the first to suggest the use of Internet search data in forecasting. Ettredge et al. (2005) point out a relationship between volumes of certain Internet search terms and the US unemployment rate. Since 2005, many 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) and sales for video games (Goel et al. 2010). More generally, a large part of the previous literature on forecasting with search data focus on micro-level predictions, while applying the data for macroeconomic forecasting has not yet gained as much attention. New big data sources, such as Google data, might prove beneficial for that purpose. As an early example, 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.

This literature review serves two purposes. First, I provide an outlook on the previous research on forecasting unemployment with Internet search data and explain the contribution of this thesis to the previous studies. Second, I discuss other applications where search data has been used to forecast economic indicators. The literature could tell us whether Internet search activity associates with traditional economic activity more generally and whether it is predictive of economic behavior.

I restrict this review to studies that use data from search engine logs although other big data sources might be useful as well. Closely related sources include Twitter tweets (Bollen et al. 2011, Antenucci et al. 2014) and website usage patterns (Moat et al. 2013). We are able to measure previously unmeasurable activity through these data sources.

The topic is still relatively new, and most studies are still exploratory. For this reason, I am not able to describe a development of ideas on the topic

preceding this study, but rather a cross section of what has been done on the first stage.

Previous literature on forecasting unemployment with Internet search data suggests that Google search volumes could be useful in predicting the unemployment rate in Germany (Askitas and Zimmermann 2009), the United Kingdom (McLaren and Shanbhogue 2011), Israel (Suhoy 2009), and Finland (Tuhkuri 2014), as well as in predicting the initial claims for unemployment benefits in the United States (Choi and Varian 2012).

In addition, D'Amuri (2009), Anvik and Gjelstad (2010), Chadwick and Sengul (2012), Fondeur and Karamé (2013), and Vicente et al. (2015) confirm that Google searches could be useful in forecasting unemployment in Italy, Norway, Turkey, France, and Spain. Pavlicek and Kristoufek (2014) find that volumes of relevant Google search queries improve unemployment forecasts in the Czech republic, Hungary, Poland, and Slovakia. Most relevantly for this thesis, D'Amuri and Marcucci (2012) offer evidence that Google searches might help to predict the US unemployment rate.

The methodology in most previous papers on the topic, most importantly in Choi and Varian (2012) and D'Amuri and Marcucci (2012) in the US, consist essentially of two steps: search term selection and forecast comparison. First, the authors select search terms that might describe individual labor market actions. These include, but are not limited to, searches for jobs (e.g., D'Amuri and Marcucci 2012), unemployment benefits (McLaren and Shanbhogue 2011; Choi and Varian 2012), employment offices (Askitas and Zimmermann 2009), and social and welfare issues (Choi and Varian 2012). Second, the studies compare out-of-sample forecasts from time-series models that include relevant Google variables to univariate models that exclude Google variables.

From a methodological perspective, Diebold (2015) reminds that this type of (pseudo) out-of-sample forecast comparison requires maintaining a variety of assumptions. Indeed, previous authors on the topic choose different assumptions. Most commonly, the models are linear, nested, and estimated by ordinary least squares (OLS) using a rolling window. In most studies, both unemployment and Google series are modified, for example, by removing the seasonal variation using seasonal adjustment techniques. I provide further discussion on the methodological differences in Chapter 7.

On the other hand, studies on forecasting unemployment with Google data are surprisingly similar to each other in terms of methodology. The fundamental hypothesis that is tested is that Internet search behavior corresponds to the actual labor market activity of individuals. The main difference in previous studies on the topic is that they have been performed in different countries.

The main result from the earlier literature is unambiguous: Google data are found to predict unemployment. There are no major qualitative differences in the results at the country level.

There are several limitations in the literature, however. Although the recent literature already covers many countries, it has not developed much from the seminal studies in the US (Choi and Varian 2012) and Germany (Askitas and Zimmermann 2009). Two quite general questions remain unanswered.

First, the previous literature does not tell how far into the future Google searches could predict unemployment. 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). A few studies extend the forecast horizon, however. Tuhkuri (2014) provides evidence that Google searches contain useful information on unemployment on a six-months-ahead forecast horizon in Finland. The informational value of search volumes appears to be strongest for three-months-ahead forecasts. In the US, D'Amuri and Marcucci (2012) report that Google data could increase the prediction accuracy for two months ahead. These studies, although reporting short-term predictive accuracy from Internet search data, do not answer how long horizons Google searches could help to predict unemployment.

Second, earlier studies report that it is possible on 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 might be time specific, and occasionally the signal from search activity might be misleading, even though Google searches were useful on average. In contrast to other studies, Choi and Varian (2012), explore the idea briefly in their work on predicting the number of initial unemployment claims.

This study extends the previous literature on the topic by addressing the two limitations. First, I study how far into the future it is possible to improve unemployment forecasts by using Google search data. In other words, I study if there is a limit after which relevant Google search volumes do not offer useful information on the unemployment rate. D'Amuri and Marcucci (2012) report relative large gains from Google data at two-months-ahead forecast horizon in the US. It is relevant to ask whether the advantage is limited to two-months-ahead predictions. Second, I explore if the improvement in forecasting accuracy is constant or if it varies over time. I also ask when the Google data could be most helpful.

Answers to the subquestions have both practical and academic relevance. Forecasters would need to know on which occasions Google search volumes could offer advantage and on which forecast horizons the data could be useful. Academically, the answers could help to describe Internet search behavior in the labor market and illustrate properties of the data for other potential uses for the data. These two questions have also more general relevance since they are not widely discussed in the context of predicting other economic indicators with Internet search data either.

Empirical studies always include several restricting assumptions. This study could also be useful in addressing the robustness of the previous findings on forecasting unemployment with Google searches. Hamermesh (2007) emphasizes the importance of re-evaluating work on important topics for improving the credibility of empirical economic research. From this perspective, a new study using a different sample and different models could be valuable and relevant.

Specific to the US context, D'Amuri and Marcucci (2012) report larger im-

provements in forecasting accuracy for the United States than studies for other countries. I explore whether the results for the United States are as high as previously reported.

D'Amuri and Marcucci (2012, pp. 30) also claim that a variable, which they construct based on search volumes on keyword "jobs" would be "the best leading indicator to predict the US monthly unemployment rate". My approach is somewhat less ambitious. With billions of potential predictors[1] and no clear guidance from economic theory, overfitting is a serious concern. In addition, Google data does not have to be the best in order to be useful. I do not try to find the best forecast method for the US unemployment rate, but answer if real-time Google search volumes could help in the task.

Besides unemployment, there are also many other variables of interest, which Internet search data could help to predict.

In economics, Wu and Brynjolfsson (2015) suggest that the data from search engines such as Google could provide an accurate but simple way to predict the housing market. They observe that Google searches are predictive not only of the present but also of future housing market sales and prices. McLaren and Shanbhogue (2011) confirm that search volumes contain useful information on the housing market in the United Kingdom.

More generally, the study by Wu and Brynjolfsson (2015) demonstrates that it depends on the particular variable of interest whether search data helps to forecast future values of the variable or only predict the present. The planning phase preceding a buying decision in the housing market can be long. That is why housing-related Internet searches might be able to give an early signal on future housing transactions. Another example of this type of behavior is travel. Future holidays are often planned well ahead of time. Choi and Varian (2012) provide a brief discussion on the topic using travel to Hong Kong as an example.

Google data provides new measurement on private actions. One application is forecasting private consumption. Vosen and Schmidt (2011) find that forecasts that include relevant Google variables tend to outperform forecasts based on survey indicators in the United States (Vosen and Schmidt 2011) and in Germany (Vosen and Schmidt 2012). Kholodilin et al. (2010) confirm the findings for the US. Tuomisto (2015) provides mixed evidence on whether Google search data could help to predict private consumption in Sweden.

Guzman (2011) examines the use of Google data in predicting inflation, and Rose and Spiegel (2012) propose a novel way to measure the change in sovereign debt default risk using Google search data. These examples suggest that a search may represent different matters in terms of economic behavior. While Guzman (2011) argues that search data are useful in modeling inflation expectations, Rose and Spiegel (2012) suggest search activity as a measure for interest in sovereign risk. In the housing market, Wu and Brynjolfsson (2015) in turn demonstrate that search activity could be seen as demand for housing. These viewpoints are not mutually exclusive.

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

Quite recently, Google data have been utilized in various academic fields other than economics. In epidemiology, Ginsberg et al. (2009) suggest that Internet searches could predict influenza epidemics more accurately than the traditional methods. This highly influential study has been followed by several studies about the use of Internet search data in epidemiology including Brownstein et al. (2009) and Hulth et al. (2009). However, the advantage provided by Internet data in predicting influenza epidemics has recently been questioned by Lazer et al. (2014). The most prominent practical application of using search data in epidemiology is *Google Flu Trends*, which is a real-time Influenza forecast service provided by Google.

More generally, Goel et al. (2010) offer a review on the use of Internet search data in forecasting and describe the restrictions of using search data. Goel et al. (2010) note that although search data are timely and frequently improve the predictions, the improvements are often limited. In specific, univariate backward-looking time-series models generally outperform models using only search data. However, univariate models extended with search data often give more precise forecasts. Moreover, Choi and Varian (2012) emphasize that even small improvements could be valuable in some cases. For example, small improvements in prediction accuracy in stock market can result in large monetary payoffs.

As a matter of fact, Preis et al. (2010, 2013) together with Curme et. al (2014), in a series of consecutive papers, claim that they are able to predict stock prices using Google data. However, finding profitable strategies from the historical data does not imply that the strategy could be used successfully in the future. Furthermore, a closer look at the Preis' et al. (2013) study reveals that toward the end of the observation period in 2011, the Google data has not seemed to offer any profitable advantage in the stock market. This could be due to the more general phenomenon in economic forecasting, emphasized by Fama (1965) in a seminal paper, that sometimes once certain information becomes publicly available, it is not helpful anymore in trying to make a profit in the market.

Already Merton (1948) and many subsequent authors have pointed out that various kinds of feedback loops exists between the economic forecasts and the economy, for example, through expectations or economic policy. However, this kind of mechanism is unlikely to play an important role when forecasting the unemployment rate with Google data. Part of the reason for this is that, to my knowledge, despite a few exceptions, Google data have not been widely used in unemployment forecasting so far. Hence, even if there were a feedback loop between the existing unemployment forecasts and the subsequent unemployment rate, which is rather unlikely, it should not affect the additional predictive power of the Google data.

A wide range of literature exists on selecting adequate or even the best predictors from high-dimensional data (see, for example, Varian 2014 for a brief and recent survey). The research in this area is not limited to conventional statistics or econometrics, as insights from computer science have also been influential. In the context of this thesis, the challenge to find the appropriate

time series from Google data has been emphasized in recent papers by Scott and Varian (2014, 2015) and Vosen and Schmidt (2011). Scott and Varian (2014, 2015) suggest Bayesian variable selection methods for nowcasting economic time series with big data. They illustrate the method in Scott and Varian (2014) with an application to initial claims for unemployment benefits. Furthermore, both relying on machine learning techniques, Ginsberg et al. (2009) and Curme et al. (2014) present two different types of automatic methods for variable selection in the context of Google data. However, the efforts to tackle this issue have been limited. In the earlier work featured in this review, the selected search terms are usually based on economic intuition or trial and error. There is a risk for data mining when using simple trial-and-error techniques. In the studies discussed here, the search terms are, however, reasonable from a practical point of view.

From a different perspective, although using Bayesian methods, Koop and Onorante (2013) argue that Google searches might be useful in selecting which nowcasting model should be used at each point in time instead of using variables based on Google searches as regressors. Their intuition is that even though it is hard to point out a clear causal relationship between the search volumes and certain macroeconomic variables, the Google data might be able to suggest the most relevant explanatory variables. Koop and Onorante (2013) report that this approach is successful in nowcasting nine monthly US macroeconomic variables. What Koop and Onorante (2013) do not do, which would be relevant, is study whether their method would be able to also forecast macroeconomic variables into the future as well. Neither do they perform analysis on the instances where search data is useful and where it is not.

The literature on forecasting with web search data is still in its early stages. Several explorative studies on predicting different economic indicators exist, but with few exceptions, the literature lacks more in-depth analysis on the specific contexts. This thesis aims to contribute in one of these contexts: forecasting unemployment with Google search data. In summary, the previous literature hints that the variation in volumes of relevant Internet search terms could reveal the intentions or sentiment of the population that uses the Internet. There have been efforts to study the predictive power of Internet search, but much remains to be done to better understand the phenomenon.

Chapter 3

Data

The primary data sources for this thesis 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. To my knowledge, this is the first study that utilizes statistics from the actual search volumes on Google, as will be described below in detail. These data come from *Google Internal Data* and *Google AdWords*.

3.1 Unemployment

The US Bureau of Labor Statistics publishes the official unemployment rate[1] monthly. However, the statistical releases are available with almost a one-month reporting lag. For example, the unemployment rate for May is released in June. To put it clearly, we do not know what the current unemployment rate is.

A real-time comparison of the unemployment rate and Google data would require using vintage data, which incorporates the revision history of the series. The federal-level unemployment rate revisions, however, have been small: 0.3 percent on average over the prior 10 years.[2] Therefore, for brevity, I employ only the most recent vintage. The literature on data vintages (see, for example, Croushore 2006 and references therein) does not unambiguously suggest the use of vintage data for the unemployment rate. One caveat arises, however. During the observation period, the revisions have not been constant but larger around turning points. Nonetheless, ignoring the data vintage may bias my results downwards but should not threaten the validity of the findings. Earlier studies on the topic, say Choi and Varian (2012) or D'Amuri and Marcucci (2012), do not employ vintage data.

The unemployment rate is available seasonally and not seasonally adjusted.

[1] The official US unemployment rate is referred as U3. According to the Bureau of Labor Statistics the unemployed are defined as people that are without jobs and they have actively looked for work within the past four weeks. The exact definition is available from ILO (1982).

[2] Source: The Bureau of Labor Statistics, Current Population Survey Vintage Data, 2015.

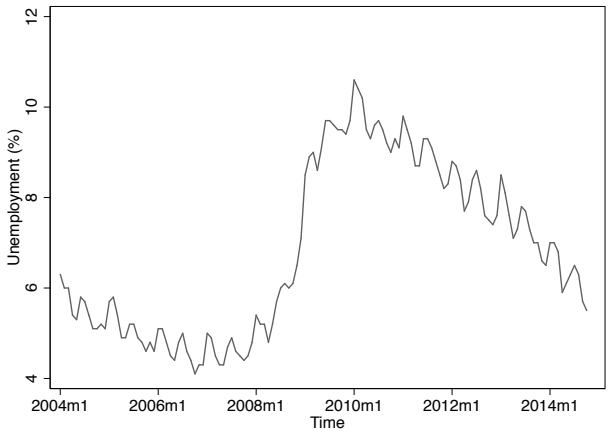

Figure 3.1.1: Unemployment rate in the United States 2004–2014. Not seasonally adjusted. Source: The Bureau of Labor Statistics.

I use the non-seasonally adjusted unemployment rate. This is because we are interested in short term predictions in this thesis. Seasonally adjusted data are based partially on forecasts and are subject to larger revisions than the non-seasonally adjusted data. As this thesis focuses on relatively short-term predictions – even the present – it is reasonable to use the most current and accurate values. Askitas and Zimmermann (2009) use non-seasonally adjusted numbers when studying whether Google searches predict unemployment in Germany.

Judging by the sheer economic significance, the US unemployment rate is arguably the most important unemployment measure. Therefore, I study whether Google searches predict unemployment in the US context. In addition, the large US data allows for state-level panel analysis.

There are also other measures for unemployment than the official unemployment rate. For example, Choi and Varian (2012) study whether Google searches predict the initial claims for unemployment benefits. The initial unemployment claims are widely considered as a leading indicator for the unemployment rate (e.g., Choi and Varian 2012). Web search volumes, say for unemployment benefits, might more accurately predict the weekly number of people who have filed for unemployment benefits for the first time than the actual unemployment rate, which also depends on other factors. However, for practitioners, the primary variable of interest is the unemployment rate.

Figure 3.1.1 describes the evolution of the unemployment rate in the United States from January 2004 until October 2014.[3] Typically, the unemployment rate exhibits seasonal variation, although that variation is not obvious from the Figure 3.1.1. The evolution of the unemployment rate is characterized by a

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

relatively sudden change between 2008 and 2010, which was associated with the economic crisis. The abrupt increase in unemployment was hard to predict, or at least, many of the predictions failed.[4] New big data sources, such as Internet search data, might help to produce more accurate forecasts.

3.2 Google

The *Google Trends* database measures volumes of Google searches. In specific, 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. This is done by analyzing a part of the web searches performed on Google. In practice, Google collects the data using the IP addresses. However, *Google Trends* does not report the exact number of search queries made with a specific keyword, but an index, which describes the intensity of the search at a weekly level. The *Google Trends* data are available globally from 2004 onwards. In the US, the data is published at the state level. In summary, the data set consists of what the Internet users search on Google.

Earlier studies have less data at their disposal because the Google data are available only from 2004 onwards. Compared to the previous study on forecasting unemployment in the US (D'Amuri and Marcucci 2012), I have included over three more years of data in my study. This adds up to a 44 percent increase from 90 to 130 in the number of monthly observations.

This section, which covers Google data, consists of three parts. I first select relevant search terms, then construct a variable, named *Google Index*, which describes search volumes for these terms and finally discuss the properties of *Google Trends* data.

The focus of this thesis is to find out whether it is possible to improve unemployment forecasts using Google searches. However, the number of different Google searches is large.[5] In order to use the Google data, I have to select, which specific search terms to use. Therefore, the first task is to select a set of relevant search terms that could be compared to the official unemployment statistics.

I come up with 125 search terms that might be 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 (Frakes 1992,

[4] This has been well documented by Krugman (2009) and Silver (2012), and many others.

[5] In 2014, there were over 15 billion new search terms every month. Source: Google Internal Data, 2014.

[6] Source: Google AdWords, 2014.

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 Chapter 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* while the time-series for the analysis come from *Google Trends*, which reports an index of search intensity. The advantage of this approach is to be able to select the most salient search terms. In the previous literature, for example, Askitas and Zimmermann (2009) utilize search terms with relatively low search volumes. This increases the uncertainty of results, which is however hard to measure, at least without knowledge about the search volumes.

In this thesis, the variable selection is based on prior economic knowledge of labor market and judgment. The underlying idea is that unemployment and the perceived subjective risk of getting unemployed affect the Google searches made by the individual. Varian and Stephens-Davidowitz (2014, pp. 5) suggest that "using Google means that you are looking for information." Many recent papers argue (Baker and Fradkin 2013, Stephens-Davidowitz 2014, Kearney and Levine 2014) that Google searches yield information about who is most interested in the topic.

Typically, a displaced worker learns about unemployment benefits, and looks for work online (Kuhn and Mansour 2014). For example, the proportion of young unemployed in the US who looked for work online was 74.4 percent in 2009 (Kuhn and Mansour 2014). It is reasonable to think that the unemployed or individuals that face an increased risk of getting unemployed make unemployment-related searches more often than average. A part of these actions leave a trace. This is because Internet browsing is usually done with using an Internet search engine (Broder 2002). However, there has been a strong growth in Internet use in the labor market (Kuhn and Mansour 2014). In 1999 only 24.2 percent of young unemployed looked for work online (Kuhn and Skuterud 2004).

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

[7] Source: Google AdWords, 2014.

[8] This 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.

and Zimmermann 2009) suggests that searches for unemployment benefits have the potential to predict the unemployment rate. I explore this 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.

However, there are still many other terms that could be used. For example, the other terms considered were searches related to job loss, welfare, employment websites, jobs, career, recruiting, and foreclosures. Nonetheless, the research question is not about finding the best set of search terms to predict unemployment, but whether Google searches predict unemployment in the first place. I leave the question of optimal variable selection for further work. I expect this will depend on the particular time series, geographical area, and point of time in question. However, I will explore the sensitivity of the results with regard to the selected search terms.

There are many advanced variable selection methods as well as methods for dimension reduction, which could be used for selecting search terms. These include Gets,[9] LASSO, Spike-and-Slab Regression, and Bayesian model averaging, to name a few. Varian (2014) offers a useful survey on methods for using large high-dimensional data sets, that is, big data. Yet, in order to answer the research question of this thesis, many of these methods are unnecessary and even harmful. Strictly speaking, I do not want the results to depend on specific well-chosen terms. On the contrary, the results should not depend on specific search terms. Rather, they should hold for many search terms that essentially describe the same phenomena. This is not to say that the advanced variable selection methods might not be useful in making actual forecasts or answering other kinds of questions using Google data.

In summary, the economic intuition is that when a person becomes unemployed, she is more likely than not to file a claim for unemployment benefits. A part of these individual actions leaves a trace to the Google data. And that is essentially what drives this study.

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. The index is constructed in the following way, within limits of the *Google Trends*.

First, the search terms are combined by a boolean search operator OR. That is, the index includes searches containing the terms unemployment benefits OR unemployment office OR unemployment claim and so on (Frakes 1992, Silverstein et al. 1999). In other words, it is a sum. The advantage of this method is that it gives each search term a weight based on its search volume, even though the actual search volumes are not directly available from *Google Trends*. The boolean algorithm is not previously widely used in the literature using Google data.

Second, the number of search queries made with the selected keywords is

[9] General To Specific

divided by the number of all search queries, which has been made in the same period of time and in the same geographical area. The resulting figure is a proportion of all Google searches that were made with the selected keywords.

Third, the data are normalized to the scale of 0–100. The normalization is made so that every value of the data is divided by the largest value of the series and multiplied by 100. As a result, the point where the intensity for Google searches is highest attains the value of 100. There are not necessarily any zeros.

Finally, the *Google Trends* data are released on a weekly basis. To match the monthly unemployment rate, the Google Index is aggregated to a monthly level by using the month in which the week begins so that the search intensity of representative week is counted to the month in which the week begins. Some monthly averages include information from four weeks and some from five weeks. Several previous studies on the topic, including D'Amuri and Marcucci (2012) and McLaren and Shanbhogue (2011), also perform the week-to-month aggregation. Choi and Varian (2012) use weekly data for Google search volumes because initial unemployment claims, which they predict, are released on a weekly basis. Google Index does not necessarily obtain a value of 100 on 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 $t = 1, 2, \ldots, f$. 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

$$I_{t,i} = \left\{ \frac{\frac{K_{t,i}}{G_{t,i}}}{\max_t \left(\frac{K_{t,i}}{G_{t,i}} \right)} \right\} \times 100, \tag{3.2.1}$$

where

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

From the Equation 3.2.1 it is easy to see that the search intensity I depends on keyword k, geography i, and length of the period f, which are fixed by the researcher.[11] The search intensity for selected terms varies across time also because of variation in the total amount of search queries. Including new observations will change the previous values since I depends on f through normalization. The starting point of the observation period also affects the values of the search intensity I.

As I said earlier, in this thesis I construct a single Google Index that simultaneously describes the evolution of several search terms that are related to unemployment. The basic idea is that variation in search volumes on one

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

[11] See, for example, Askitas and Zimmermann (2009) and Varian and Stephens-Davidowitz (2014) for other definitions.

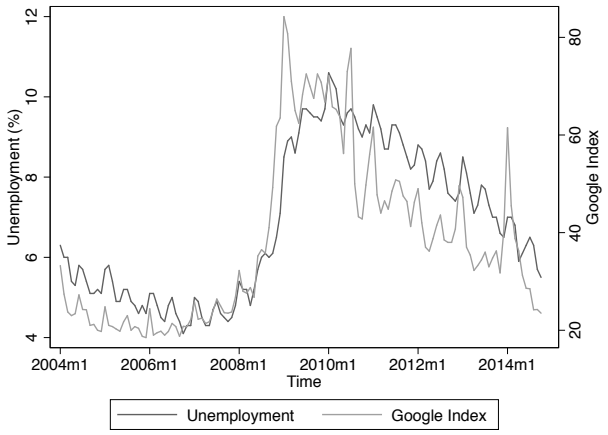

Figure 3.2.1: Unemployment rate and Google Index, which describes search activity for unemployment benefits in the United States 2004–2014. Source: The Bureau of Labor Statistics and *Google Trends*.

particular unemployment-related search term could be related to many random factors. The common variation of many search terms, however, might be able to capture the part of the variation that is connected with the actual interest to unemployment benefits. Using several search terms also reduces the risk that the results would be driven by a specific well-selected keyword rather than a more general association between search trends and unemployment. The sample size is also larger when using a larger number of search terms. Most previous studies explore many keywords but actually use search volumes for only one search term (say, D'Amuri and Marcucci 2012 and McLaren and Shanbhogue 2011). In simple terms, using many search terms could reduce noise in the Google Index.

Figure 3.2.1 describes the evolution of the Google Index and the unemployment rate from January 2004 until October 2014.[12] The search intensity $I_{t,\mathrm{i}}$ 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 3.2.1 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

[12]Google 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 3.2.1: Descriptive statistics for the unemployment rate and Google Index 2004–2014. Source: The Bureau of Labor Statistics and *Google Trends.*

abruptly in January 2014. I repeat the analysis of this thesis while also controlling for the two spikes in Chapter 7. Part of the variation in search intensity may be driven by variation in total amount of search queries $G_{t,i}$.

The unemployment rate and the Google Index seem to behave in a similar manner. However, after 2010, there is a larger spread between the series. The Pearson correlation coefficient between monthly unemployment and the Google Index is 0.87. The Google Index does not seem to exhibit any clear seasonality. Table 3.2.1 gives descriptive statistics for the Google Index and the unemployment rate. Compared to the Gaussian benchmark, both series exhibit thinner tails than the normal distribution, as kurtosis is smaller than 3. The Google Index has a lower signal-to-noise ratio μ/σ than the unemployment rate. The numbers are 2.22 and 3.62, respectively. Also evident from Figure 3.2.1 that the Google Index is more volatile than the unemployment rate.

Although there have been some efforts to use the data on Google searches in economic research, the Google data is not well documented in the previous literature.

First of all, the data set is large. For example, there were over 100 billion searches globally on Google every month in 2014.[13] According to comScore, 12 billion of these searches were made in the United States in October 2014.[14] The total number of Internet searches in the US was 18 billion, which means that 67 percent of search queries conducted in October 2014 were made with Google.[15] The second most popular search engine was *Microsoft Bing* with 19.5 percent of the Internet searches in the US.[16] According to the US Census Bureau, in 2013, 74.4 percent of all households reported Internet use.[17] Google is the only major search engine that makes historical search data publicly available. For this reason, it is a natural choice for studying Internet searches.

There are many potential advantages in using Google data. One of the main advantages is related to forecasting: Google search data are available in almost real time. Most of the data are available on a weekly basis, but the value of the

[13]Source: Google Internal Data, 2014.
[14]Source: comScore qSearch, October 2014.
[15]Source: comScore qSearch, October 2014.
[16]Source: comScore qSearch, October 2014.
[17]Source: The US Census Bureau, Computer and Internet Use in the United States, 2013.

latest week is updated on a daily basis. *Google Trends* data are usually available well ahead of official data.

Unlike survey data, statistics from Google searches arise as a by-product from Internet use (Varian 2010). In a sense, it is aggregated register data. This may reduce issues usually associated with surveys: attrition, non-response, and measurement error. It is also accessible and free, which makes it possible to replicate previous results. In addition, the research questions need not be decided in advance, but the data are collected constantly, covering a wide range of questions. As a result, Internet search data can help to analyze suddenly emerging phenomena.

Furthermore, Stephens-Davidowitz (2014) finds little evidence of social desirability bias in Google searches. This makes sense: individuals have an incentive to report truthfully about the information that interests them (Stephens-Davidowitz 2014, Varian and Stephens-Davidowitz 2014).[18]

Despite the advantages of Internet searches data, there are many issues associated with *Google Trends*. The Internet is still a relatively new phenomenon, so the statistics from Google searches are available only from 2004 onwards. This is a short period of time compared to many other economic indicators, and it is a short time period in terms of many econometric applications. For example, when using Google data on a monthly level, there are only 130 observations during 2004–2014.

Another technical challenge is that the interpretation search intensity is not straightforward. The search intensity is based on the fraction

$$\frac{K_{t,i}}{G_{t,i}}, \tag{3.2.2}$$

where $K_{t,i}$ denotes the amount of searches that were made with the selected keyword and $G_{t,i}$ denotes the total amount of search queries at the same time. This measure does not thus tell the actual number of searches. It varies with both $K_{t,i}$ and $G_{t,i}$. However, there is a reason for the way of measurement. Almost all search terms are associated with a rising trend in absolute values, due to the fact that the search volumes have increased many-fold since 2004. By dividing $K_{t,i}$ by $G_{t,i}$, *Google Trends* removes the rising trend. On the other hand, the method may produce a new downward trend. Many search terms exhibit a constantly decreasing pattern, as the Internet is used for many new purposes that generate new search queries.

More to the point, 15 percent of daily searches are new.[19] That is, for these searches, the exact combination of terms has never before been typed into Google. This poses also another challenge for the usage. The used search queries change over time, and this has to be taken into account when constructing statistical variables from the text corpus.

Google Trends data are based on a daily sample of all web searches performed on Google. For this reason, they vary little from day to day. The less

[18] Moreover, it can be argued that truthful reporting is a dominant strategy in search behavior.

[19] Source: Google Internal Data, 2014.

common the keyword is, the greater the variation is likely to be. It is a common practice to sample massive big data sets to reduce computation time. However, Google claims that the sample is large enough to give relatively precise results (Varian and Stephens-Davidowitz 2014). Thus, many samples are not likely to be necessary.

In order to make the data more representative of Internet users, Google removes repeated queries that were made from a single IP address within a short period of time. The adjustment is also made to ensure that most search queries were made by actual people, not machines.

All the data are anonymous and aggregated. That is a privacy protection issue. *Google Trends* database allows researchers to access and explore sometimes sensitive data while respecting research ethical and confidentiality issues.

For confidentiality reasons, *Google Trends* has a privacy threshold, θ, which Google does not disclose publicly.[20] If the number of searches is smaller than the threshold, that is $K_{t,i} < \theta$, a 0 will be reported. The privacy threshold is based on absolute volumes. Varian and Stephens-Davidowitz (2014) point out that an important implication of this is that smaller places and earlier periods will more frequently show zeros. While using US federal-level data, the search volumes are above the threshold. However, previous studies using only single search terms often face limitations from this threshold. Stephens-Davidowitz (2014) presents an algorithm to overcome the privacy threshold, but the algorithm does not remove the issue of a low actual search volume for the observational unit. For robust inference, I would not recommend building the analysis on keywords with low query volumes. Heffetz and Ligett (2014) provide a discussion on privacy and data-based research covering also issues associated with the use of Internet search data.

There are also several more general issues with Google data. While the data are extensive, they are not necessarily representative or a non-random sample. Internet use is still strongly associated with the education level and place of residence of the individuals.[21] In 2013, Internet use was most common in young and educated, high-income Asian or White households in metropolitan areas.[22] *Google Trends* does not offer any background characteristics for the people that make the search queries, even on an aggregate level. It is also worth mentioning that for brevity, I only use search terms in English, although in 2009, English was the main language of only 80 percent of Americans.[23] The result from this is that search queries by an English-speaking population may be overrepresented in the sample of this study. The representativeness issues may lead to less accurate forecasts for the whole population (D'Amuri 2009, Fondeur and Karamé 2013) Thus, my findings might be biased downwards compared to an approach that

[20] There is, however, anecdotal evidence on this threshold. The Chief Economist of Google, Hal Varian, told during a public lecture in 2012 that the threshold would be "fifty or so" (Varian 2012, "Using Google Data for Short-term Economic Forecasting", Strata 2012).

[21] Source: The US Census Bureau, Computer and Internet Use in the United States, 2013.

[22] Source: The US Census Bureau, Computer and Internet Use in the United States, 2013.

[23] Source: The US Census Bureau, American Community Survey, Language Spoken at Home, 2009.

would take this into account.

There are also a lot of economic activity in which the role of Internet searches is limited. For example, it is unlikely that corporate investments in the future would appear as clearly detectable search queries in Google search statistics. For this reason, Google searches might be comparatively useful in understanding and forecasting consumer-based topics, such as consumption decisions (Goel et al., 2010, Vosen and Schmidt 2012), the housing market (Wu and Brynjolfsson 2015, McLaren and Shanbhogue 2011, Choi and Varian 2009b), job search (Baker and Fradkin 2014), or the behavior of the unemployed (Choi and Varian 2012, Askitas and Zimmermann, 2009, Baker and Fradkin 2014).

One of the main issues with Google data is that the data-generating process is not fully understood. There are many reasons for this. Different users may use different search terms, even if their intention is the same. On the contrary, users with completely different intentions may search the Internet using very similar keywords. For example, it is plausible that individuals are more likely to search for "unemployment benefits" when they are actually willing to file for unemployment benefits. However, there might be a number of other reasons for searches using this keyword. As discussed earlier, uncertainty, for example, might make people search for more information. Similarly, Preis et al. (2010) together with Bordino et al. (2012) find that search activity related to a particular company may be associated with either positive or negative financial performance, or even both depending on the context. As a result, the search statistics should be interpreted with caution.

In summary, the greatest advantage of Google data is that it is available in real time. The main weakness is that the data-generating process is not entirely clear. In other words, there is potential real-time correlation with economic phenomena. For this reason, one of the most natural uses for Google data is short-term forecasting. This serves as a motivation for the research setting in this thesis.

Chapter 4

Methods

This chapter presents the main methods used in this thesis to answer whether Google searches predict unemployment.

To be clear, there are two distinct questions closely related to this 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. This approach is outlined in Section 4.2. However, to first describe the relationship, I analyze the series jointly by performing Granger non-causality tests and studying the cross-correlation function in Section 4.1. The results for both approaches are presented in Chapter 5. I explore the robustness of the results with different methods in the Chapter 6.

From a methodological perspective, this thesis offers somewhat novel analysis in several ways. Most previous studies employ a (pseudo) out-of-sample forecast comparison methodology, where a univariate benchmark is extended with Google data. This thesis analyzes the series also jointly. Furthermore, in Chapter 6 this thesis applies US state-level data on unemployment and Google search volumes in order to address the robustness of federal-level findings, using panel data methods. Through state-level data I am able to exploit the geographic and temporal variation in the level of the unemployment rate induced by the 2008 economic crisis. Panel data methods provide also an opportunity to control for unobserved factors in the relationship between Google searches and unemployment.

4.1 Joint Analysis

In this section, I present the joint analysis methods, which aim to provide descriptive analysis on the relationship between Google Index and the unemployment rate. Google Index describes search activity for unemployment benefits.

4.1.1 Cross-correlation

Descriptive information on the dynamics of the unemployment rate and Google searches is provided by their cross-correlation function (CCF). The cross-correlation function is the joint autocorrelation function of two series. Intuitively, the cross-correlation function tells whether, for example, current Google search volumes are more strongly correlated with the future unemployment than with the present. Gourieroux and Jasiak (2001, Chapter 3) define the empirical cross-correlation function as follows:

$$\hat{\rho}_{i,j,T}(h) = \frac{\hat{\gamma}_{i,j}(h)}{\hat{\gamma}_{i,i}^{1/2}(0)\hat{\gamma}_{j,j}^{1/2}(0)}, \tag{4.1.1}$$

where h is the lag between the variables and $\hat{\gamma}_{i,j}(h)$ denotes the (i,j)th element of the autocovariance function $\hat{\Gamma}_T(h)$, which is given as

$$\hat{\Gamma}_T(h) = \frac{1}{T}\sum_{t=h+1}^{T}(Y_t - \hat{m}_T)(Y_{t-h} - \hat{m}_T)'. \tag{4.1.2}$$

In Equation 4.1.2, $\hat{m}_T$ denotes the sample mean

$$\hat{m}_T = \frac{1}{T}\sum_{t=1}^{T} Y_t. \tag{4.1.3}$$

The cross-correlation function is estimated from its empirical counterpart. The analysis of the cross-correlation function could help to resolve the lead–lag relationship between search volumes and unemployment.

4.1.2 Granger Causality

This study employs the Granger (1969) non-causality test as a descriptive measure to determine whether the lags of the Google Index have useful linear predictive content above and beyond the unemployment rate itself. The advantage of the Granger non-causality test is that it allows to study the direction of temporal dependence also the other way around. In other words, I also 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. To my knowledge, previous studies on the topic, except for Tuhkuri (2014), do not perform Granger non-causality tests.

The Granger non-causality test examines whether the lagged values of one variable in a vector autoregressive model (VAR) help to predict another variable. A time series x_t Granger causes y_t if the past values of x_t help forecast y_t over and above the own history of y_t.[1] In a bivariate VAR model, Granger causality

[1] Granger causality does not imply causality in traditional sense, for example, in terms of potential outcomes described by Rubin (2005).

from x_t to y_t can be tested by testing the significance of the coefficients of the lags of x_t in the equation of y_t.

I estimate the following VAR(1) model to study the Granger causality:

$$\boldsymbol{y}_t = \boldsymbol{\Theta}_1 \boldsymbol{y}_{t-1} + \boldsymbol{e}_t, \tag{4.1.4}$$

where the weak white noise admits a variance-covariance matrix,

$$\boldsymbol{\Omega} = \begin{pmatrix} w_{11} & w_{21} \\ w_{21} & w_{22} \end{pmatrix}. \tag{4.1.5}$$

In the matrix notation, $\boldsymbol{y}_t$ denotes a vector of dependent variables, the unemployment rate and Google Index, $\boldsymbol{\Theta}_1$ is the 2×2 autoregressive coefficient matrix and, $\boldsymbol{e}_t$ denotes the error term vector. The model is estimated by the ordinary least squares (OLS) method equation by equation (see, for example, Gourieroux and Jasiak 2001, Chapter 3). I select the first-order VAR model, which, based on the Schwarz criterion (BIC), is a statistically adequate simplification of second- and third-order VAR models.

The Granger causality statistic is an F-statistic testing a null hypothesis that the coefficients on all lags of that variable are zero. In the first-order VAR model utilized in this thesis, the Granger causality test from x_t to y_t simplifies to testing the significance of the coefficient of x_{t-1} in the equation of y_t. Consequently, the corresponding null hypothesis H_0 is that $\theta_{i,12} = 0$ in

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \theta_{1,11} & \theta_{1,12} \\ \theta_{1,21} & \theta_{1,22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix}, \tag{4.1.6}$$

where y_t is the unemployment rate and x_t is the Google Index.

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. The idea is that the current value of the Google Index might Granger cause unemployment even when the lagged value would not.

The asymmetric vector autoregressive model (VAR) for the second specification is following:

$$\begin{pmatrix} y_t \\ x_{t+1} \end{pmatrix} = \begin{pmatrix} \theta_{1,11} & \theta_{1,12} \\ \theta_{1,21} & \theta_{1,22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_t \end{pmatrix} + \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix}. \tag{4.1.7}$$

To explore the sensitivity of the results, I also estimate the models by using fourth-order VARs. The described Granger causality analysis is implemented in a linear framework and does not capture possible nonlinearities in the relationship.

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 section is organized as follows. The first step is to construct a relevant benchmark model for the unemployment rate. The benchmark model is then extended with the Google Index. Subsequently, the models and their forecast performance are compared.

The specific question is whether including data on Google searches improves the benchmark unemployment model. In particular, I aim to answer whether the Google Index improves (pseudo) out-of-sample forecasts in the short run. The (pseudo) out-of-sample forecast comparison methodology and the other methods for comparing the models are explained for nowcasting in Section 4.2.1 and extended to cover also forecasting in Section 4.2.2. The type of model and (pseudo) out-of-sample forecast comparison methodology used is this thesis is associated with West (1996) and Clark and McCracken (2001). West (2006) provides a lengthy survey on the selected method. Out-of-sample forecast comparison methodology is selected because it provides an illustrative and convenient way to assess the potential value of Google data for unemployment forecasting. 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.

As I have mentioned earlier, in the previous literature on forecasting with Internet search data a majority of the studies compare models that include relevant Google variables to univariate models that exclude Google variables. In specific, this approach for studying the topic traces to the work of Choi and Varian (2009a, 2009b, 2012) and Goel et al. (2010). The framework of Choi and Varian (2012) consists essentially of a comparison of two nested models: one that contains Google data and one that does not. However, this is a standard procedure in the forecasting literature.

The in this section models 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 methods used for answering these two subquestions are presented in Sections 4.2.2 and 4.2.3.

The rest of this part concentrates the benchmark model selection. The selected benchmark will be used throughout the analysis in this section. Within the chosen framework it is necessary to choose a relevant benchmark model for the US unemployment dynamics in order to answer the whether the Google data could improve performance of the model. However, the focus of this thesis is in the relevance of Google search data for unemployment forecasting not in the unemployment dynamics itself.

I limit the set of potential benchmark models to autoregressive (AR) models. In an autoregressive (AR) time-series model, the dependent variable y_t depends

linearly on its past values. By using a lag operator, the AR(p) can be written as

$$\theta(L)y_t = \varepsilon_t, \tag{4.2.1}$$

where $\theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_p L^p$, and ε_t denotes a white noise process. In the AR(p) model y_t depends on the p most recent values.

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, Elliot and Timmermann 2008, Hawkins 2004). Furthermore, the baseline model still needs to be extended with a Google variable, so I do not want to use too many parameters for the benchmark.

A simple univariate autoregressive model is also 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). A simple autoregressive structure also allows to easily add Google data as an additional predictor. For example, it is more difficult to extend an autoregressive moving average (ARMA) model to additional regressors than it is for AR models.

There is a long line of literature on the dynamics of the US unemployment rate (see, for example, Nelson and Plosser 1982; Montgomery et al. 1998; Papell et al. 2000), but the dynamics of Google data are not well known. In part for this reason, I restrict the analysis to models with less structure. Koop and Potter (1999) and van Dijk et al. (2002) argue that the most prominent feature in the unemployment series is persistence. An autoregressive model is able to approximately capture this attribute.

One option would be to apply an established and historically well performing unemployment forecasting model, such as Barnichon-Nekarda (2012) flow-model, and ask whether including Google search data could improve the model and its forecasting performance. The research question is, however, whether Google searches predict unemployment in general, not whether a particular forecasting model can be improved by using search data. In a more general sense, the advantage from big data does not necessarily come from doing the same old things with new data, but doing things differently (McAfee and Brynjolfsson 2012). That is why extending an existing forecast model with a "big data variable" is not necessarily particularly useful exercise. In this thesis, I focus on assessing the relevance of this new big data source. The actual use of this data source is a question of further research. To illustrate further the approach of this thesis it is useful to think that there is an implicit null hypothesis that Google searches do not predict unemployment. In a sense, one purpose of the thesis is to answer whether we have enough evidence to reject the null hypothesis of no predictive content on unemployment.

From another perspective, many previous authors including Montgomery et

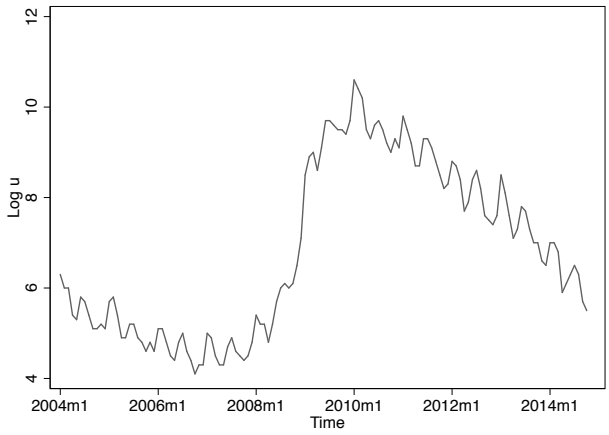

Figure 4.2.1: The evolution of the logarithm of the unemployment rate 2004–2014. Not seasonally adjusted. Source: The Bureau of Labor Statistics.

al. (1998) and van Dijk et al. (2002) suggest an autoregressive structure for an unemployment forecasting model. Furthermore, for example Montgomery et al. (1998) extend an autoregressive model with an exogenous variable, initial claims for unemployment, which is available earlier than the unemployment rate. In that sense, I obtain roughly the approach of previous literature on forecasting unemployment.

The time series of the logarithm of the unemployment are depicted in the Figure 4.2.1. The economic theory does not suggest any particular specification for an AR model of the US unemployment rate (see, for example, Montgomery et al. 1998; Papell et al. 2000). In the following part, I choose and estimate a univariate linear time series model for the monthly US unemployment rate y_t which serves as a benchmark for forecasting comparison against a model, which includes a Google variable. The task is to find the autoregressive AR(p) model that sufficiently captures the dynamics of the given time series.

Before identifying a relevant order of AR(p) benchmark model preparatory data transformations are discussed.

As a starting point, I use both variables, the unemployment rate and the Google Index, in levels rather than in differenced values. For the unemployment rate, the explanation is that it is bounded between 0 and 100. For this reason, the unemployment rate 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 (Montgomery et al. 1998). Nevertheless, there is no clear-cut conclusion in the literature of unemployment dynamics (see, for example, Nelson and Plosser 1982; Montgomery et al. 1998; Papell et al. 2000 for a discussion) whether the unemployment rate should, however, be modeled

as non-stationary and, stationarity-inducing transformation should be applied. For example Rothman (1998) emphasizes that the relative forecasting performance for the unemployment rate may be sensitive to such a transformation. In any case, there is a consensus that the unemployment rate is still highly persistent. Along this line, the unemployment rate might be viewed as a process that is possibly stationary but highly persistent (Montgomery et al. 1998). These processes are often approximated by unit root processes in empirical applications (see, for example, Montgomery et al. 1998). However, according to Koop and Potter (1999) and Papell et al. (2000) it is quite safe to model the unemployment rate in levels.

The Google Index is also bounded between 0 and 100 by construction, and therefore the previous argument holds for the presence of a global unit root. Moreover, because there is a reporting threshold in *Google Trends*, using differences in the Google Index would be problematic. Although there are issues with interpreting the level of the Google Index, it is still more informative than the difference. However, as discussed in Chapter 3, there might be a downward sloping deterministic trend in the Google Index. This would happen if the volume of all Google searches consistently grew at a faster pace than the volume of unemployment-related searches. Nevertheless, a visual analysis of Figure 3.2.1 suggests that the series do not exhibit a clear deterministic trend. In short, as a starting point I use both variables in levels, because I do not find compelling reasons not to do so.

Montgomery et al. (1998) document that the US unemployment rate exhibits seasonality in the long term. Inspection of the sample autocorrelation function depicted in Figure 4.2.2 suggests, however, no clear form of seasonality. The time series plot in Figure 4.2.1 exhibits no evidence of a specific type of seasonality either. However, I will 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.

I perform a logarithmic transformation for the unemployment series for two reasons. The first and most important reason is that changes in unemployment rate are most naturally discussed in percentage terms. The transformation also gives a meaningful interpretation for the coefficient of the non-transformed Google Index. The second reason is that variation in unemployment rate seems to increase with the level of the series. Logarithmic transformation can help to stabilize the variance of a time series. Furthermore, Lütkepohl and Xu (2012) note that logarithmic transformation often helps improve out-of-sample forecasts for levels of the data. However, if the variance is not stabilized by the logarithmic transformation, the transformation may be unnecessary or even harmful (Lütkepohl and Xu 2012). While most of the variation in the unemployment series comes from the increase in 2008, the logarithmic transformation fails to stabilize the variance in the series. Still, since I study the relative performance of a model including a Google variable compared to a benchmark, I

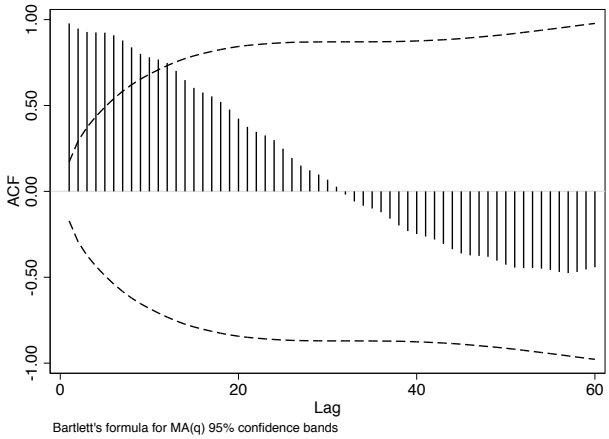

Figure 4.2.2: The estimated autocorrelation function of the logarithm of the unemployment rate 2004–2014.

do not expect the adverse effects from logarithmic transformation to be severe. Logarithmic transformation for the dependent variable has also been proposed by Choi and Varian (2012) for studying whether Google searches predict economic indicators.

After preparatory actions the next step is to identify an adequate autoregressive order of the AR(p) model. In the model specification I follow a principle of parsimony suggested, for example, by Box et al. (2008, Chapter 1). Parsimony refers intuitively to an approach to modeling time series where the aim is to find the simplest model that still describes the series at hand adequately. My approach relies on the Box-Jenkins methodology 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. A one reason for this is that the dynamics of a small sample may not accurately represent the properties of the underlying data-generating process.

The estimated autocorrelation function (ACF) of the series is provided in Figure 4.2.2. The autocorrelation function in Figure 4.2.2 has a slow decay but eventually tails off, reaching zero. The first 12 estimated autocorrelation coefficients are statistically significantly different from zero (at 5% level). Based on the values of the autocorrelation function, the series feature a strong serial dependence, as suggested by previous empirical work (Koop and Potter 1999 and references therein). However, as discussed earlier, from a theoretical perspective, the unemployment rate is not likely to exhibit a global unit root over a longer period of time (see, for example, Koop and Potter 1999).

I study this conjecture by applying the standard Dickey-Fuller (DF) test proposed by Dickey and Fuller (1979) and the nonparametric Phillips-Perron

test introduced by Phillips and Perron (1988) with several different lags. I use a maximum of 13 lags for the Phillips-Perron test because the analysis of the sample autocorrelation coefficients function depicted in Figure 4.2.2 suggests that the first 12 autocovariances are relevant. Hamilton (1994, Chapter 15) suggests this selection procedure. During the sample period from 2004 to 2014, these tests fail to reject the null hypothesis of a unit root at the 5% level against the alternative of stationarity.

Likewise, the augmented Dickey-Fuller test (ADF) fails to reject the null hypothesis of a unit root against the alternative hypothesis of being stationary at the 5% level. I follow the procedure proposed by Hall (1994) and use Akaike (AIC) and Schwarz (BIC) information criteria (Akaike 1973, and Schwarz 1978, respectively) to estimate the lag length for the augmented Dickey-Fuller statistic. As a result, I choose lag length 13, since the both criteria prefer the same model. Moreover, the KPSS test developed by Kwiatkowski et al. (1992) rejects the null hypothesis of stationarity at the 5% level in the same sample with Bartlett kernel weights (Newey and West 1987) but fails to reject the null hypothesis with the alternative quadratic spectral kernel (Andrews 1991) using the bandwidth of 13 lags suggested by AIC and BIC. The lesson from these tests is that locally, during this observation period, the unemployment rate is serially dependent to a high degree and the unit root tests fail to reject the null hypothesis of stationarity.

A visual analysis of the series and the estimated autocorrelation function depicted in Figures 4.2.2, and 4.2.2 confirms the observation. I also test for the unit root in the first differenced series using the procedure described earlier. I observe that the differenced series are not likely to exhibit unit root behavior. However, a time series may be non-stationary for other reasons than having at least a one unit root (Gourieroux and Jasiak 2001, Chapter 5; Gourieroux and Robert 2006).

Again, in the long run, the unemployment rate in levels has not exhibited either a deterministic or stochastic trend (Montgomery et al. 1998). Furthermore, Cochrane (1991) argues that the unemployment rate does not have a unit root even though the unit root tests might suggest locally that there were one. More generally, Cochrane (1991) points out that in finite samples the unit root tests should not be applied blindly without considering the underlying process and whether an assumption of a unit root is relevant. In the previous studies on forecasting unemployment with Google searches, for example Choi and Varian (2012) and D'Amuri and Marcucci (2012) model the unemployment rate in levels. For these reasons, despite the results of the unit root tests, I continue to use levels of unemployment without imposing a unit root in my main benchmark.

Sample partial autocorrelation function (PACF) could help to identify a possible order of an AR model (Box et al. 2008, Chapter 6; Gourieroux and Jasiak 2001, Chapter 2). The sample partial autocorrelations of logarithmic unemployment are provided in Figure 4.2.3. It is evident from Figure 4.2.3 that the first lag has a relatively high partial autocorrelation compared to the other lags. Furthermore, there seems to be a cutoff at the 13th lag, after which the partial autocorrelations remain statistically insignificant at the 5% level. A

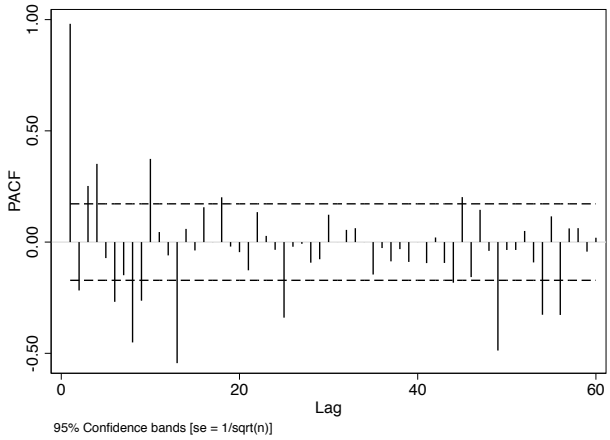

Figure 4.2.3: The estimated partial autocorrelation function of the logarithm of the unemployment rate 2004–2014.

visual inspection of the sample partial autocorrelation function in Figure 4.2.3 suggests that an adequate AR model could be of order 13 or lower (see, for example, Gourieroux and Jasiak 2001, Chapter 2). The main observation from both autocorrelations and partial autocorrelations is that the ACF is infinite in extent and tails off as the lag increases while the PACF is close to zero for lags larger than 13. The conclusion from this observation is that an AR(p) representation might be acceptable for the purposes of this study.

An examination of estimated autocorrelation and partial autocorrelation functions offers a starting point for choosing a particular autoregressive order for a benchmark. Using sequential testing, I start by estimating a relatively high order AR(15) model. 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. This means that when using either of the information criteria with lower values than 13 for maximum lag, the selected lag order would be the maximum itself.

However, in order to use the AR(13) model as a benchmark, I have to estimate 14 coefficients. On the other hand, 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 for this is following. I decide to use a naïve seasonal AR(1) benchmark for the main specification. This model uses only the previous period and seasonal effects to predict the unemployment rate. There are five reasons for this.

First, the main reason is that 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, we cannot reject the implicit null hypothesis that Google searches do not predict the unemployment rate. Furthermore, if the Google data do not improve prediction accuracy against the seasonal AR(1) benchmark, then it is not likely to improve the more sophisticated models either.

Second, the unit root tests and the visual analysis of the ACF suggest that the unemployment rate follows almost a random walk process. This is a typical feature for macroeconomic data (Nelson and Plosser 1982). For pure random walk processes, the best univariate forecast for y_t would be only y_{t-1}.[2] Therefore, a seasonal AR(1) model could offer a reasonable comparison. In the previous literature on forecasting with Google data, Choi and Varian (2012) use this argument to motivate the use the AR(1) benchmark.

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, although the Akaike and Schwarz information criteria might suggest higher order models.

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. Within the short sample, there is uncertainty on how the dynamics of the unemployment series should be modeled as a univariate time series process. Furthermore, a search for the true data-generating process may be misleading within such a short and historically idiosyncratic sample. From an economic point of view, a complex and high-order model is hard to justify. However, a simple seasonal AR(1) model is able to capture the high persistence in unemployment rate, which is the most prominent feature of the series.

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

It is worth mentioning, however, that 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. If the selected model adequately captures the dynamics of the time series, the residuals should be approximately white noise. At least, reasonable benchmark model would not have too much autocorrelation in the residuals. I check this by using the following three methods: visual inspection of the residual series (this may also reveal potential outliers), estimating the ACF r_k $(k = 1, 2, ..., K)$ of the residuals, and estimating the ACF of the squared residuals for remaining conditional heteroskedasticity. I also formally test both autocorrelation and conditional

[2] Series with this property are also called martingales (Gourieroux and Jasiak 2001, Chapter 2).

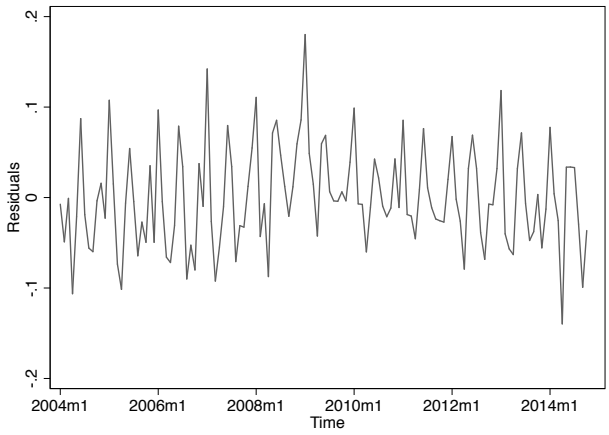

Figure 4.2.4: The evolution of the residuals of seasonal AR(1) model 2004–2014.

heteroskedasticity in the residuals.

I estimate the seasonal AR(1) model by quasi-maximum likelihood (QML) method under normality assumption. Figure 4.2.4 describes the evolution of the residual series to detect potential outliers. There is an increase in the residual series at the time of high unemployment depicted in Figure 3.1.1. The volatility of the residual series seems to decrease after 2009. Figures 4.2.5 and 4.2.6 outline the autocorrelation functions of the residuals and squared residuals for the baseline seasonal AR(1) model. In seasonal AR(1) benchmark, 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 seems to abate as the lag increases. The estimated autocorrelation function for the residuals suggests that there does not appear to be unit root problems. The Q-Q plot presented in Figure A.0.1 in the Appendix suggests that the residuals are quite normally distributed.

I estimate alternative models up till fourth order seasonal AR model. I find that when limiting to lower than fourth-order seasonal AR models, the higher order models do not give clear advantages against the seasonal AR(1) model, judging by the estimated autocorrelation functions of the residuals.

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 $K = 12$ and $K = 24$ lags does reject the null hypothesis of no serial correlation at 1% level. The number lags are selected to take into account possible remaining seasonality. 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.

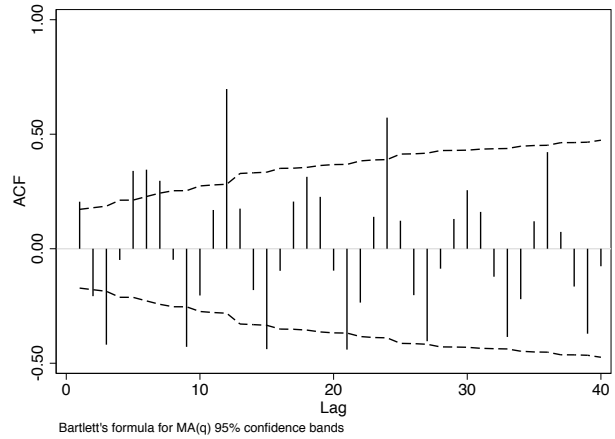

Figure 4.2.5: The estimated autocorrelation function of the residuals for seasonal AR(1) model.

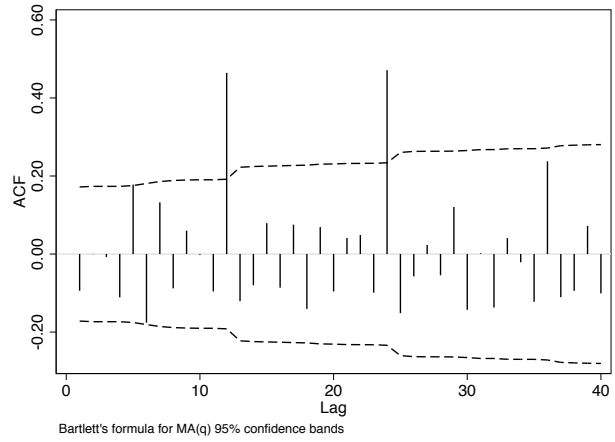

Figure 4.2.6: The estimated autocorrelation function of the squared residuals for seasonal AR(1) model.

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 4.2.6 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. Referring to the discussion earlier, the variance of the unemployment rate has not been constant from 2004 to 2014. Thus, there could still be weak conditional heteroskedasticity in the residual series when a simple autoregressive model is estimated for the series.

The estimated PACF of the unemployment series in Figure 4.2.3 also shows that the first lag is distinctively higher than the following lags, although the PACF cuts off only after 13 lags. This, in turn, suggests the benchmark could be either seasonal AR(1) or considerably higher AR(13).

For these reasons, among lower than fourth-order autoregressive models, the seasonal AR(1) model seems to be adequate, however not perfect, compared to the higher order models. Finally, I accept the seasonal AR(1) model as a main benchmark for predicting the unemployment rate.

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).

It is still worth emphasizing that seasonal AR(1) model might be too restricting benchmark. To explore sensitivity of the results for the selected benchmark I also estimate the results using seasonal AR(2) and AR(3) benchmark models and an AR(13) model in Chapter 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

The Google Index is available in real time, while the unemployment rate is only available after the end of each month. In simple terms, Google data are available a month earlier than the official unemployment statistics. This gives the Google data a meaningful forecasting lead (Choi and Varian 2012). More to the point, in the monthly level forecasting model, Google data are available at the date of prediction t, but the unemployment rate is available with a one-month lag, $t - 1$. The difference in publication lag is one of the main motivations in terms of why Google data might improve the predictions of the present unemployment rate. In other words, I study whether searches for unemployment benefits "now" could help to predict the current unemployment rate, which is not known at the date of prediction.

Previous literature suggests that Google searches are typically associated with present economic activity (see, for example, Baker and Fradkin 2014; Da et al. 2011, 2015; Hall 2011), which means that data on the popularity of search queries might be helpful for predicting the present (Choi and Varian

2012). However, the Google search queries might be correlated with future expectations and thus help to forecast future unemployment. I present methods to explore whether Google data can help predict the future unemployment rate in the next section.

In this section, I present methods to study whether Google searches predict the present unemployment. In summary, I include the most recent value of Google Index to the benchmark model as an additional predictor. Then I compare the properties of the two models. In specific, I study how the out-of-sample forecasts improve by using a rolling window forecast.

The resulting 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.[3]

However, 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). Strictly speaking, for forecasting purposes, it is not enough to illustrate that Google searches are correlated with current or future unemployment. To show that Google searches predict the unemployment rate, 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 (see, for example, Angrist and Pischke 2009, Chapter 3), 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.

Even though my main objective is to examine the predictive power of the Google searches for unemployment, 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

[3]See, for example, Gourieroux and Jasiak (2001, Chapter 2).

properties of the US unemployment rate. In specific, I compare the fit of the models measured by coefficient of determination R^2, as well as other properties, such as information criteria, statistical significance, and the magnitude of the parameters.

This study employs Akaike (1973) and Bayesian (Schwarz 1978) information criteria to compare the models. The information criteria address the tradeoff between goodness of fit and number of parameters (Verbeek 2012, Chapter 8). The Akaike information criterion for this context is given by

$$\text{AIC} \quad = \quad log\hat{\sigma}^2 + 2\frac{p+1}{T}, \tag{4.2.2}$$

where $\hat{\sigma}^2$ is the estimated variance of the error term e_t, $p$is the number of coefficients, and T is the number of observations. The alternative Bayesian information criterion is given by

$$\text{BIC} \quad = \quad log\hat{\sigma}^2 + \frac{p+1}{T}logT, \tag{4.2.3}$$

with the same notation. It is said that the information criteria prefer a model with the smallest value. Values of information criteria are computed for both estimated models. The candidate models are estimated over the same sample. Hannan (1980) demonstrates that BIC is a consistent estimate of the true model, while AIC prefers larger models and is thus not consistent. Nonetheless, if the true model is not among the models considered, the value of information criteria is limited (Verbeek 2012, Chapter 8).

I also compare residual diagnostics for the models. In specific, diagnostic tests by Ljung and Box (1978) and McLeod and Li (1983) for the null hypotheses of no remaining autocorrelation and conditional heteroskedasticity in the residuals are performed for both residual series. Not too much weight should be given to the above in-sample measures, as the focus of this thesis is on forecasting.

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. In practice, 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). In other words, 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 also long enough.

Pesaran and Timmerman (2004) offer a review on the costs of ignoring breaks in forecasting time series. However, for brevity, I do not model potential breaks explicitly. One reason for this is that the breaks are not necessarily known to

forecasters a priori and that the break might arise from a gradual change in the parameters coming from the measurement of the Google Index. I use a relatively short and fixed width for the rolling window instead of an expanding window to account for the possible structural changes in the series.

I use the same models for each rolling estimation window. Therefore, the forecasts are not out of sample in the sense that the model selection was made using the whole sample. Relevant previous papers on this topic (Askitas and Zimmermann 2009; Choi and Varian 2012; D'Amuri and Marcucci 2012) utilize the same approach as this thesis. Alternatively, it is also a common practice in the forecasting literature to perform the model selection recursively, that is, to use a certain selection procedure, say the Schwarz information criterion, for each information set. For our data generating-process, the Akaike and Schwarz information criteria and sequential testing procedure tend to select relatively high-order models (AR(13) with few exceptions), which are not reasonable benchmarks for the purposes of this study.

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

$$\begin{aligned} \text{MAPE} &= \frac{1}{T}\sum_{t=1}^{T} \mid E_t \mid, \qquad (4.2.4) \\ \text{where} & \\ E_t &= \frac{\hat{y}_t - y_t}{y_t} \times 100, \end{aligned}$$

where y_t denotes the official unemployment rate and $\hat{y}_t$ denotes the forecasted value. A lower error measure means more accurate predictions. The error is defined as percentage deviation.

The selected error measure has the advantage of being scale-independent. The disadvantage of being not defined when $y_t = 0$ does not apply for the unemployment rate. Mean absolute percentage error (MAPE) is selected instead of, for example, the mean squared error (MSE) also because it does not give as much weight to the occasional large errors. In the literature, Choi and Varian (2009a) use mean absolute percentage error for evaluating whether Google searches help to predict initial unemployment claims. However, Lanne (2009) points out that professional forecaster's loss functions do not necessarily match commonly used error measures. Again, it is worth noticing that I am not looking for optimal or the "best" forecast for unemployment based on specific loss function, but answering whether Google searches could help to forecast it overall. I explore the sensitivity of results to selected error measure with mean squared error in Chapter 6.

By comparing the accuracy of (pseudo) out-of-sample forecasts by the benchmark Model (0.0) and the extended Model (1.0) using selected error measure, I can evaluate whether Google data could help to predict the present. In specific, if the value of the error measure for forecasts computed from the extended model with the Google Index lies below error measure values of the benchmark

univariate AR model, I conclude that the Google searches might be useful in nowcasting unemployment.

Finally, I test whether the difference in forecast accuracy between the two models is statistically significant. For this purpose, I use the test for equal predictive accuracy of Diebold and Mariano (1995) and West (1996). The Diebold-Mariano (DM) test is a way to compare the predictive accuracy of two or more competing forecasts. This is done by comparing differences in the error measures of the forecasts and the actual series. The null hypothesis is that there is no difference in the accuracy (Diebold and Mariano 1995).

Generally speaking, in a two-forecast case the test is based on the following loss differential:

$$d_t = g(e_{1,t}) - g(e_{2,t}),$$

where $e_{1,t}$ and $e_{2,t}$ denote series of forecast errors $\hat{y}_t - y_t$, and g is the selected loss function. The forecasts have equal predictive accuracy if the loss differential d_t has an expectation of zero. The null hypothesis is

$$H_0 : \quad E(d_t) = 0$$

against a two-sided alternative that the expectation is non-zero.[4] The test uses the asymptotic critical values presented in Diebold and Mariano (1995) and West (1996).

For the purposes of this study, the results are obtained for the full out-of-sample period. I allow for serial correlation by employing heteroskedasticity and autocorrelation robust (HAC) standard errors, as in Newey and West (1987) for performing the test. The number of lags $h - 1$ for computing variance are selected according to the h-step-ahead forecast horizon, as proposed by Diebold and Mariano (1995).

There are also other tests for comparing predictive accuracy. These include, but are not limited to Hansen (2005) and Hansen et al. (2011). West (2006) provides a comprehensive review on tests for comparing predictive accuracy. Diebold (2015) describes the potential pitfalls in using the Diebold-Mariano test in a (pseudo) out-of-sample environment. The test comes with a potential power loss compared to full-sample alternatives (Diebold 2015). Hansen and Timmermann (2012) reach similar a conclusion. In addition, Chadwick and Sengul (2012) in the context of Google data, and West (2006) more generally argue that the asymptotic properties of the Diebold-Mariano test are not necessarily satisfied for comparing nested models. For these reasons the test should be interpreted with caution. However the Diebold-Mariano test is convenient, widely used, and to some extent describes whether the incremental predictive ability from the Google Index is statistically significant. Moreover, Diebold (2015) concludes that despite the issues, the Diebold-Mariano test still provides direct information on the comparative historical predictive performance of the

[4] Verbeek (2012, Chapter 8), for example, uses slightly different notation.

forecasts. Relevant extensions, such as Harvey et al. (1997), which attempt to accommodate the rather short time series could be proposed, however.

D'Amuri and Marcucci (2012) also use the Diebold-Mariano test to study whether Google searches could improve unemployment forecasts. In contrast, Chadwick and Sengul (2012) conduct the Harvey et al. (1997) small sample modification of Diebold-Mariano test for equal predictive accuracy. With few exceptions, most of the previous literature on using Internet search data for economic forecasting, for example, Choi and Varian (2012) and Goel et al. (2010), do not employ formal tests for comparing predictive accuracy.

In addition to these methods, constructing forecast intervals for the forecasts would be relatively straightforward. However, forecast errors are more informative. Whether Google data could help to improve density forecasts would be another, perhaps interesting, question.

During the modeling process, I have made several assumptions. I address these in Chapter 6 by estimating different models and seeing if the qualitative conclusions change.

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. Corresponding to the previous section, the models are nested and linear. I utilize heteroskedasticity- and autocorrelation-consistent (Newey and West

(1987) standard errors as earlier. The number of lags for the robust standard errors is selected by method proposed by Newey and West (1994) for each model and every estimation window.

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. A dynamic procedure also captures the real-time aspect of Google data. Previous studies on the topic by D'Amuri (2009) and D'Amuri and Marcucci (2012), which study forecasting the future and not only nowcasting, use dynamic forecasts as well.

I evaluate the models' out-of-sample performance by comparing the dynamic h-step-ahead forecasts by using the same methodology, which described earlier in Section 4.2.1. Each extended model is compared to the recursive h-step-ahead forecasts computed from the same benchmark Model (0.0).

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 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, with economic data, forecasters are typically interested in the turning points. However, Kling (1987) points out that predicting the turning points in macroeconomic data using only past observations is difficult. Yet, real-time information, such as Google data, might give a signal on potential turning points (Bańbura et al. 2013). Furthermore, real-time data might be especially useful during a recession when the economic indicators are sometimes 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.

Chapter 5

Results

5.1 Joint Analysis

In this section I provide descriptive evidence on the relationship between the Google Index and the unemployment rate by studying the cross-correlation function of the series and performing Granger (1969) non-causality tests.

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 5.1.1 displays the values of the estimated cross-correlation function (CCF). The pattern of cross correlations are plotted against the lag order h in Figure 5.1.1.

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 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

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

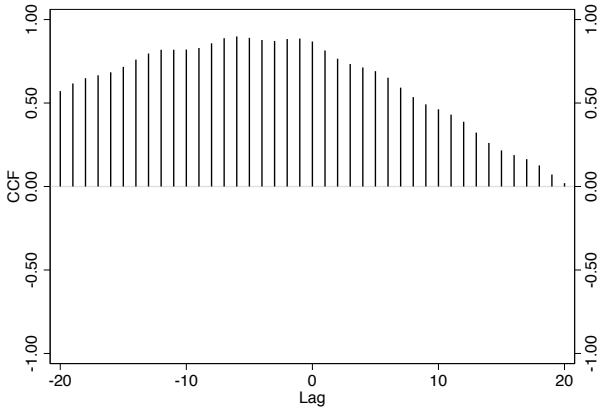

Figure 5.1.1: The estimated cross-correlation function between the unemployment rate and the Google Index against the lag order h of the Google Index 2004–2014.

of a classical leading indicator, that is, the Google search volumes are better correlated with the future than current unemployment. Bordino et al. (2012) and Wu and Brynjolfsson (2015) observe a similar pattern in the stock market trading volumes and the housing market prices respectively.

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 5.1.2 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 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 unemployment 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. In other words, Internet search

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 5.1.2: Statistics for testing Granger non-causality.

activity seems to precede unemployment. This result is in line with the previous analysis of cross correlations.

When using fourth-order VAR models, I find no qualitative changes in the results. Toda and Phillips (1993) argue that there are problems with performing for Granger non-causality when the time series are non-stationary and possibly cointegrated, which means that the results should be interpreted with caution. Although searches predict unemployment, the unemployment potentially causes the searches. The causal relationship is however far from clear. Despite the evident endogeneity, using Google searches to predict the unemployment rate is of interest because the data are available before the unemployment statistics, and seem to be associated with the future unemployment rate.

5.2 Model

This section presents the estimation results for the models described in Section 4.2. Furthermore the section summarizes the performance of the models with and without the Google Index for forecasting the US unemployment rate, using a (pseudo) out-of-sample forecast comparison methodology.

5.2.1 Predicting the Present

Do Google searches help to predict the present unemployment rate? The estimation results for Models (0.0) and (1.0) are presented in Table 5.2.1. The coefficient for the Google Index is statistically significant at the 1% level.[1] The positive sign of the coefficient means that the searches related to unemployment

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

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). The sample period is Jan 2004 – Oct 2014.

Table 5.2.1: Estimation results of the benchmark seasonal AR(1) model (0.0) and the extended model (1.0), which includes Google Index.

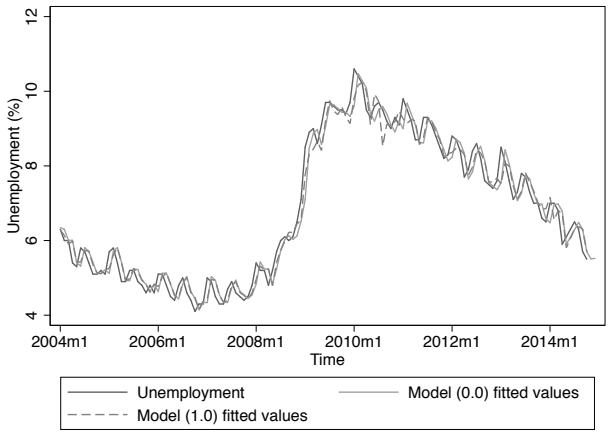

Figure 5.2.1: Unemployment rate and the fitted values for the benchmark model (0.0) and extended model (1.0), which includes Google Index 2004–2014.

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 (Verbeek 2012, Chapter 3). 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.

If we look at the autocorrelation functions of the extended model's residual and squared residual series in Figures 5.2.2 and 5.2.3, we see that including the Google variable decreases the temporal dependence in the residuals compared to the benchmark model described in Figures 4.2.5 and 4.2.6.

However, Ljung-Box and McLeod-Li tests with 12 and 24 lags reject the null hypotheses of no remaining autocorrelation and conditional heteroskedasticity in the residual series at the 5% significance level. I infer that the Google Index does not remove the residual autocorrelation and conditional heteroskedasticity of the benchmark model. As noted earlier, I account for the remaining autocorrelation in the residuals by using heteroskedasticity- and autocorrelation-consistent (HAC) standard errors (Newey and West 1987).

Figure 5.2.1 presents the fitted values for both models (0.0) and (1.0). Precise fitted values will not necessarily mean, however, accurate forecasts (Verbeek 2012, Chapter 8). Do Google searches help to forecast the present unemploy-

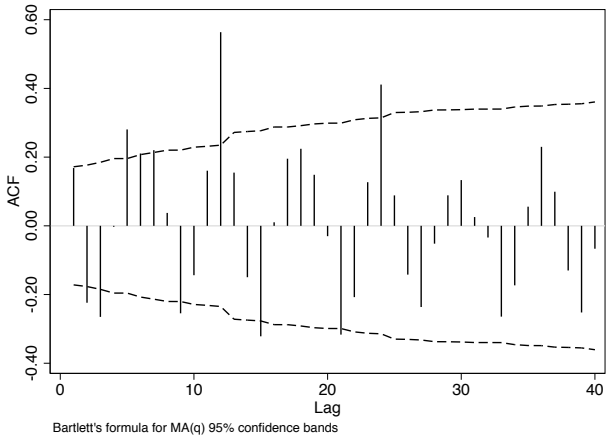

Figure 5.2.2: The estimated autocorrelation function of the residuals for Model (1.0), which includes Google Index.

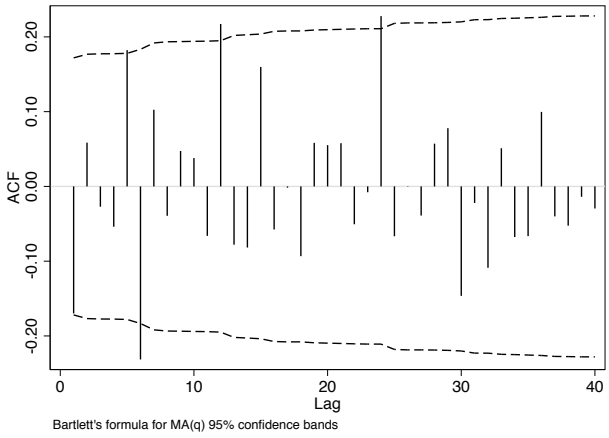

Figure 5.2.3: The estimated autocorrelation function of the squared residuals for Model (1.0), which includes Google Index.

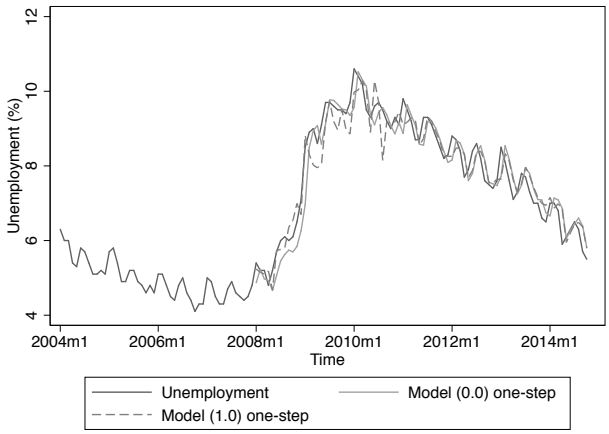

Figure 5.2.4: Unemployment rate 2004–2014 and the one-step-ahead nowcasts with a rolling window of 48 months for the univariate benchmark model (0.0) and the extended model (1.0), which includes Google Index 2008–2014.

ment rate?

Results from one-step-ahead out-of-sample predictions using a rolling window of 48 months are illustrated in Figure 5.2.4. The mean absolute percentage errors for nowcasts are given in Table 5.2.2.

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.

Model	MAPE	Δ
(0.0)	4.58%	4.32%
(1.0)	4.38%	

MAPE = mean absolute percentage error

Δ = improvement in forecasting accuracy

The evaluation period is Jan 2008 – Oct 2014.

Rolling window of 48 months is employed.

Table 5.2.2: Nowcasting (one-step-ahead) accuracy of the seasonal AR(1) benchmark model (0.0) and the extended model (1.0), which includes Google Index 2008–2014.

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. In other words, the test may fail to reject the null hypothesis even if the alternative were true. The low power in finite samples is noted by 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 benchmark represents a more general difference "in population" or merely an observation "in sample."

Presence of a potential unit root may impose restrictions on the interpretation of the results. There is a risk that the observed relationship might be partly driven by unit root problems (Granger 1974). I discuss this issue further in Chapter 7.

5.2.2 Forecasting the Future

So far, we have considered only nowcasting. Could Google searches be useful in forecasting as well? Table 5.2.3 summarizes the mean absolute percentage errors of out-of-sample dynamic forecasts up to the horizon $h = 6$. We can see from Table 5.2.3 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. Therefore, a model with the Google Index might provide more accurate forecasts than a benchmark model using lagged data and seasonal effects for horizon h but not for $h + 1$ where h is relatively small. Results in the Table 5.2.3 support this idea. The results indicate that Google data might help to predict unemployment for horizon $h = 1$, but not necessarily much further. That is, there is a decline in predictive accuracy in longer than one-month-ahead forecast horizons when using the extended model instead of a univariate benchmark. There is an exception on a six-month horizon, where we find 9.9 percent improvement in predictive accuracy. This observation, however, may not have much structural significance.

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. The evolution of two-steps-ahead out-of-sample predictions using a rolling window of 48 months is illustrated in Figure 5.2.5. The Google Index seems to give a timely signal on the increase in the unemployment rate in 2008. Majority of the previous literature on forecasting with Internet search data, including Choi

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 5.2.3: 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

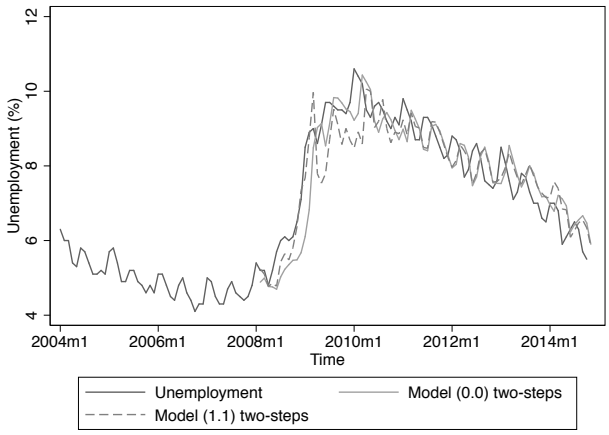

Figure 5.2.5: Unemployment rate 2004–2014 and the dynamic two-steps-ahead forecasts with a rolling window of 48 months for the univariate benchmark model (0.0) and extended model (1.0), which includes Google Index 2008–2014.

and Varian (2012), Goel et al. (2010), and Askitas and Zimmermann (2009), consider only assessing current conditions.

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. As discussed earlier, a likely explanation to this is a rather short time series available from Google data.

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 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 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 5.2.6 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

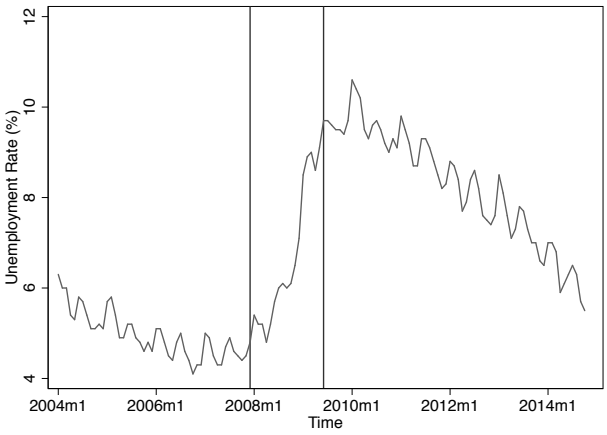

Figure 5.2.6: The recession.

recession at the point where it was later determined to have begun.[2]

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 5.2.4 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. 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.

The Diebold-Mariano tests for comparing predictive accuracy support the finding that the improvements in prediction accuracy are larger in the recession. Table 5.2.4 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? Earlier Figures 5.2.4 and 5.2.5 describe the evolutions of one-step-ahead

[2]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 5.2.4: 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 from 12/2007 until 6/2009.

Figure 5.2.7: 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.

and two-steps-ahead forecasts. Looking more closely at the series, Figure 5.2.7 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. In a similar manner, Figure 5.2.8 visualizes the difference in two-steps-ahead foresting errors for the baseline model and the extended model with the Google Index for each month. 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. 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.

To sum up, the informational value of search data appears to be time specific. Google search queries improve the prediction accuracy especially during the 2007–2009 recession in the United States.

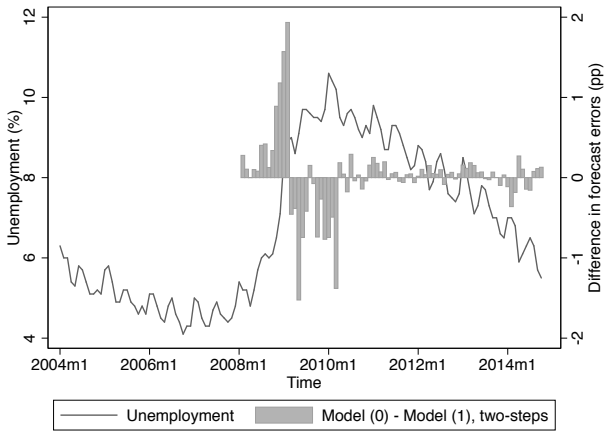

Figure 5.2.8: The difference in absolute forecast errors for two-steps-ahead forecasts of univariate benchmark model (0.0) and extended model (1.0), which includes Google Index 2008–2014 and the unemployment rate 2004–2014. The vertical bars are positive when the extended model performs better.

Chapter 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. The panel is strongly balanced. Compared to the previous data set, we now have 5,900 observations instead of 130. That is, the cross-sectional data help compensate for the relatively short time series that are available from *Google Trends* from 2004 onwards, as outlined by Choi and Varian (2012). 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:

$$log(y_{i,t}) = \beta_1 log(y_{i,t-1}) + \beta_2 log(y_{i,t-12}) + \beta_3 x_{i,t} + \alpha_i + e_{t,i}, \qquad (6.1.1)$$

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 $x_{i,t}$. The model is also know as a dynamic panel data model.

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). Angrist and Pischke (2009, Chapter 8) argue that in empirical applications the robust standard errors can fall below the conventional standard errors, and suggest selecting the maximum of the two standard errors. I follow this procedure. The standard errors are not clustered because the number of groups equals with the number of panels. Arellano (2003) points out that the robust standard errors coincide with the "clustered" standard errors, which are clustered over the panel variable. Angrist and Pischke (2009, Chapter 8) note that how best to approach serial correlation in panel data models is an area of ongoing research and current practices might be limited.

In the model the state-level lagged dependent variables $y_{i,t-1}$ and $y_{i,t-12}$ are correlated with the unobserved panel-level effects by construction. As first noted and shown by Nickell (1981), this makes the standard within estimators, such as the ordinary least squares (OLS) estimator, inconsistent. For this reason I use an asymptotically consistent generalized method of moments (GMM) type estimator derived by Arellano and Bond (1991) to estimate the parameters. However, the estimator of Arellano and Bond (1991) is aimed for datasets with a large number of panels i and few periods t. In Arellano and Bond (1991), the asymptotic properties are derived for a fixed number of periods and number panels $N \to \infty$. Alvarez and Arellano (2003) discuss a case where both number of panels and periods is large and point out that Arellano-Bond estimator is asymptotically consistent when both the number of panels $N \to \infty$ and periods $T \to \infty$. Judson and Owen (1999) point out that in a case where the number of periods T is also large but the number of panels N is smaller than T, a within panel fixed effects estimator, however, might perform better than than the Arellano-Bond estimator. I check the results also by employing a within estimator using the ordinary least squares (OLS) method with the robust standard errors of Arellano (1987).

The fixed effects model is selected instead of alternative random effects model because the panel contains all US states and is not a random sample of larger amount of states. A similar argument is made by Judson and Owen (1999). The Hausman (1978) test also rejects the null hypothesis that the random effects model is consistent at 1% significance level and thus suggests a fixed effects model. There are also many alternative specifications for the panel data model. For example, I could include time-fixed effects or allow for heterogeneity over state panels in the coefficients. However, the purpose of this model is to explore the robustness of federal-level results, and the additional complexity is not necessary.

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.0.2 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 the this further, I have constructed a map displayed in Figure 6.1.1, which visualizes the US state-level differences in the popularity of unemployment-related Google searches between November 2009 and February 2010. The darker colors refer to higher values of the Google Index.

The results from state fixed effects model are given in the second column of Table 6.1.1 together 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 controlling for the state-level lagged and seasonal effects. The pattern – Google searches predict unemployment – seems to be repeated at state level. This lends additional credibility for the main results in Chapter 5. The

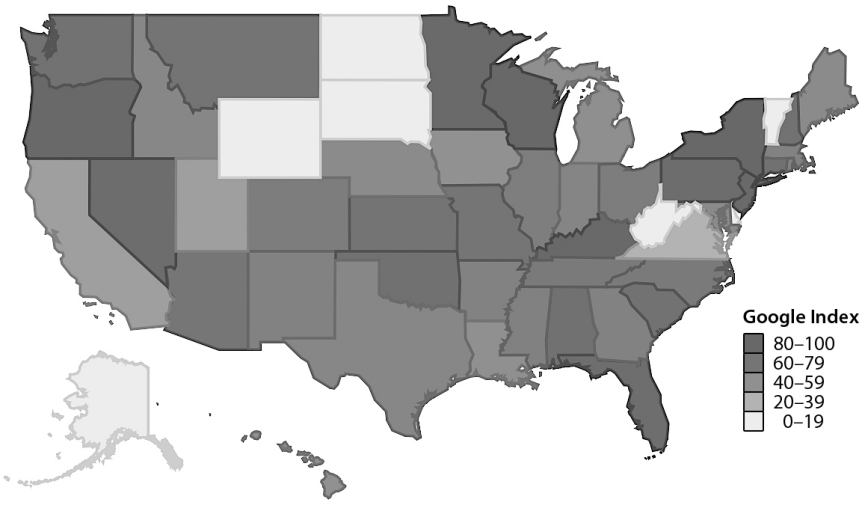

Figure 6.1.1: Relative popularity of unemployment-related Google searches. Average between Nov 2009 and Feb 2010. Source: *Google Trends*.

results from fixed effects regression also confirm that the Google searches predict unemployment, not only on a temporal but also on a geographical level.

The estimation results from within estimator using the ordinary least squares (OLS) method are reported in the third column of Table 6.1.1 and 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.

There are many other issues with this specification. For example, the state-level Google Index occasionally obtains the value of zero for privacy reasons. Therefore, the observed month-to-month variation in the Google Index between zero and non-zero values may be partly noise. This is an issue because fixed effects estimates are sensitive to measurement error (Angrist and Pischke 2009, Chapter 5). However, the fixed effect model serves as a reasonable robustness check. To address the issues associated with the threshold, I construct an indicator variable for the threshold and include the indicator to the model as a control. The associated coefficient for the Google Index is somewhat larger (0.0023) and statistically significant at the 1% level.

Measurement of Google data as an index also poses a reason for caution. The index captures the relative popularity for certain search queries compared

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 5.2.1. The sampling period is Jan 2004 – Oct 2014.

Table 6.1.1: Estimation results of the extended autoregressive model (1.0) and the fixed effects model (FE).

to other points in time. Thus, search intensity measures are not directly comparable on the state level, but the patterns are.

The Internet penetration varies geographically at the US state level.[1] States with high GDP per capita, such as California and Massachusetts, are also states with high levels of Internet use. Low GDP-per-capita and rural states, such as Louisiana and Alabama, exhibit low levels of Internet use.[2] Furthermore, the unemployment rate and Internet adoption are seemingly correlated on the state level.[3] This thesis measures, however, the relative popularity of unemployment-related search queries not the pure volumes. Therefore, the state-level results are not likely to be driven by pure differences in the level of Internet use. Yet, it is relevant to see whether the results hold for states with different levels of Internet use.

Part of the state-level Google data are only available on a monthly frequency. This is for privacy protection reasons because of the low search volumes in the beginning of the observation period. For this reason, the method of temporal aggregation may be different for these states, since it is done by Google.

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 very sensitive to the choice of the set of search terms. I explore this 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.[4] 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*.

[1] Source: The US Census Bureau, Computer and Internet Use in the United States, 2013.

[2] Source: The US Bureau of Economic Analysis, State Personal Income 2014 and the US Census Bureau, Computer and Internet Use in the United States, 2013.

[3] Source: The US Census Bureau, Computer and Internet Use in the United States, 2013 and the Bureau of Labor Statistics, Current Population Survey, 2014.

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

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. Some of these choices may lead to different results (e.g., Diebold 2015). In the following, I explore the sensitivity of the results to some of the most restricting assumptions, but many more remain.

As the main benchmark against which Google data are compared, this thesis employs a seasonal AR(1) model. If I have underestimated the autoregressive order, the dynamics are misspecified, and the benchmark model does not capture enough temporal dependence in the series (e.g., Gourieroux and Jasiak 2001, Chapter 2). In that case, the positive results could arise from misspecified benchmark model.

To account for this possibility, I estimate the results also using higher order benchmarks, seasonal AR(2) and AR(3) models, as well as using the AR(13) model. Although seasonal AR(2) or AR(3) models do not offer markedly better representation of the data, I include additional lags to the benchmark model to find out whether the results are driven by a possibly misspecified benchmark. The AR(13) model is used because it is preferred by Akaike and Bayesian information criteria among AR(p) models. However, to keep things simple, I do not go through every possible benchmark.

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.

In comparison with the AR(13) model, the Google Index included in that model as an exogenous regressor is not statistically significant at the 10% level, but the model including Google Index is still preferred by Akaike information criterion. However, it is not preferred by the Bayesian criterion, which imposes a tighter penalty for additional parameters. Google variable does not improve out-of-sample predictions against the AR(13) benchmark, but for our data-generating process the AR(13) model yields less accurate predictions than the seasonal AR(1) model. Furthermore, a forecast comparison using a rolling window of 48 observations is not necessarily feasible with a large model.

Are the results robust to the selected loss function? One concern is that the mean absolute percentage error is not symmetric but gives more heavier penalty on negative errors ($y_t < \hat{y}_t$) than on positive errors (see, for example, Makridakis 1993). Here is a chance for confusion: if a forecast is too large the error is negative. However, an asymmetric loss function is not necessarily a bad thing (Lanne 2009). For a robustness check, I compute the forecast errors using other commonly used error measure, mean squared error (MSE). The results are essentially the same as those from using mean absolute percentage error (MAPE). For the whole out-of-sample period, Google searches have marginal predictive ability over and above that of the own history of unemployment on the present and one-month-ahead forecast horizons. Compared to results obtained

with mean absolute percentage error, improvement for predicting the present is slightly lower (3.16%) but somewhat higher (8.12%) for forecasting one month ahead. These numbers are computed against the previous univariate benchmark presented in Chapter 4. During the recession, the improvements in prediction accuracy are higher when measured by mean squared error as opposed to mean absolute percentage error. Improvement for horizon $h = 0$ is 27.4%, for $h = 1$ 45.14%, and for $h = 2$ 18.6%, Furthermore, when using mean squared error, the test of Diebold and Mariano (1995) and West (1996) for comparing predictive accuracy rejects the null hypothesis for equal predictive accuracy during the recession for the one-month-ahead horizon $h = 1$ at the 1% level as reported previously. I conclude that changing the loss function to mean squared error does not affect the qualitative conclusions.

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 appears to be time specific. This result, although quite expected, is not emphasized in most previous literature on the topic. With an expansive window starting with 48 months, the results are virtually the same as with a rolling window reported in Chapter 5.

Chapter 7

Discussion

There are still some concerns. First of all, the improvements in prediction accuracy are in most cases only modest. This has also been noted by Goel et al. (2010) in previous work on forecasting with Google data. Most likely, the main reason for this is that in many cases, simple models already give quite good forecasts in the short run (see, e.g., Makridakis et al. 1979; Elliot and Timmerman 2008). For example, Leitch and Tanner (1991) point out that the univariate time series approach sometimes outperforms even professional forecasts. In specific, previous studies on unemployment forecasting (see, for example, Montgomery et al. 1998) report reasonably low forecast errors in the short term using time series models.

For practical use, it is not necessarily discouraging that the improvements are not large within the whole sample. My results suggest that Google search queries improve the prediction accuracy, especially during the 2007–2009 recession in the United States, when the need for more timely information was evidently considerable. Sudden changes in Internet search volumes may also help forecast especially the turning points in economic time series, which are hard to identify with the autoregressive baseline models. Further research might tell whether Google data could help in this purpose.

However, the modest improvements contrast with the existing study (D'Amuri and Marcucci 2012) on predicting the unemployment rate with Google data in the US. D'Amuri and Marcucci (2012) report markedly higher improvements than I find. In specific, D'Amuri and Marcucci (2012) claim that compared to their benchmark, the forecast accuracy increases at two months ahead by 40 percent. On the contrary, I do not find any consistent improvement in prediction accuracy beyond one-month-ahead predictions. My results suggest that D'Amuri and Marcucci (2012) are perhaps too optimistic in their view. The authors walk through over 500 models and report the results for the best performing model within the estimation sample. The best performing model is selected ex post, and there is no guarantee that this specific model would make the most accurate predictions in the future. In other words, there is a chance for a specific type of overfitting using their method as recently described by

Diebold (2015). At least, the quantitative results seem to depend on specific modeling choices.

In terms of magnitude, my findings are more in line with Choi and Varian (2012). They find that including the *Google Trends* data can help improve the predictions for the initial claims for unemployment benefits in the US. However, when they look at their whole sample from 2004 to 2011 they do not find any improvement in prediction accuracy compared to the univariate AR(1) benchmark, even at the one-step-ahead horizon. Nonetheless, they report improvement during the 2007–2009 recession. I also find stronger gains from Google data during the recession. More generally, my results and the findings by Choi and Varian (2012) suggest that the utility of search data seems to depend on the time-specific context. The relatively low improvements are also in line with previous results from the UK (McLaren and Shanbhogue 2011), Norway (Anvik and Gjelstad 2010), and Spain (Vicente et al. 2015).

Beyond small improvements, Goel et al. (2010) together with Choi and Varian (2012) are concerned that the improvements in prediction accuracy could be driven by common seasonality, not necessarily by actual predictive power. However, my analysis suggests that at least controlling for seasonal effects does not have a qualitatively meaningful impact on the results. As a matter of fact, the coefficient for the seasonal term is small and not statistically significant. The only previous study on whether Google predict the US unemployment rate (D'Amuri and Marcucci 2012) uses seasonally adjusted data. Furthermore, several other studies on forecasting unemployment with Google data in Italy (D'Amuri 2009) and the UK (McLaren and Shanbhogue 2011) employ official seasonally adjusted data. In contrast, Askitas and Zimmermann (2009) use seasonally unadjusted monthly unemployment rate for Germany. Compared to these earlier studies, I do not find major qualitative differences in the results depending on the use of seasonal adjustment.

Another concern is that the simple autoregressive models utilized in this thesis sometimes provide reasonable predictions but occasionally produce very bad forecasts. This issue is visualized in Figures 5.2.7 and 5.2.8. Koop and Onorante (2013) together with Lazer et al. (2014) have emphasized the importance of this issue, although from different perspectives.

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. Part of this change is initiated by Google itself. Lazer et al. (2014) point out that, for example, Google's recommended search algorithm may increase the relative volumes of certain search queries. The search behavior is thus not exogenously determined, but also endogenously related to the search engine. Consequently, it is relevant to understand the search algorithm in order to produce robust economic forecasts.

Koop and Onorante (2013), in turn, point out that the issue is not only with the interpretation of the Google data, but also with the properties of simple autoregressive models. The authors claim that the causal autoregressive models occasionally fail to accommodate a situation where the economy is not constant, such as the recent economic crisis. However, according to Koop and Onorante

(2013), the Google data could be used to select a suitable model at each point in time using, for example, dynamic model selection (DMS) techniques. Comparing my results to Koop and Onorante (2013), I observe that the advantage of Google data in forecasting seems to depend on the specific use of the information. The Google data are "suggestive, not definitive" (Stephens-Davidowitz 2015, pp. SR1), and a good forecasting method would take this into account. That is, my results suggest that Google searches help to predict unemployment, especially during the recession. However, another type of model might be able to improve unemployment forecasts during times when the model of this study fails.

Despite the differences, the underlying principle is common among Koop and Onorante (2013) and Lazer (2014). The bad forecasts seem to arise from the inability to identify the changes in the data-generating process.

My own take on this is more directly targeted on the specific 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 thesis 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.

As a matter of fact, several previous studies report this phenomenon. Choi and Varian (2012) note that an AR model extended with Google data predicts the numbers of initial claims for unemployment better during the recession, but the benchmark model without Google data fits better immediately after. The Google model underpredicts the initial claims for unemployment during the recovery. Furthermore, Askitas and Zimmermann (2009) observe that unemployment forecasts from Google searches in Germany become increasingly less accurate after the year 2008. The authors attribute this to changes in labor market policy that followed the economic crisis. D'Amuri and Marcucci (2012) report a similar pattern of underpredicting the unemployment after the crisis when predicting the US unemployment rate with Google data. The results from predicting the Finnish unemployment rate with Google search volumes (Tuhkuri 2014) are an exception. In my previous study on the topic, I find a contrary pattern, where the model with Google data underpredicts the unemployment rate before the crisis and overpredicts the unemployment rate immediately after. Nonetheless, the models utilized in this thesis as well as in the previous literature are relatively simple, and we should avoid far-reaching conclusions from these results.

My broader argument is that specifics may matter. This argument is not limited to unemployment forecasting. Knowledge on the particular phenomenon, which the series describe, is potentially valuable for forecasting purposes. A successful forecasting strategy may well depend on the particular variable in case and vary over time. In short, forecasting is context specific.[1] A recent example of this approach is the Barnichon and Nekarda (2012) flow model that tries to describe the unemployment dynamics more carefully. These unemploy-

[1] This has been powerfully argued by Silver (2012).

ment flows, however, might be better approximated using auxiliary data sources such as Google data. However, their model does underpredict the outflows to unemployment as well. One reason for this could be that after 2007 the unemployment outflows have been low by historical comparison. Furthermore, it is not clear that a forecasting methodology that utilizes two or more variables and then calculates their sum would yield better forecasts than a pure time-series approach (Verbeek 2012, Chapter 8).

However, the methods presented in this thesis do not necessarily represent the ways actual forecasters would or should use the information from the Google searches. The contribution of this thesis is to point out that relevant Google searches predict unemployment. The actual use of this information is an interesting topic for further research. Lanne and Nyberg (2015) note, that further macroeconometric methodological work is still needed so that the new data sources could be utilized effectively.

One of the issues that we are always going to run into is the changes in search behavior. For example, "unemployment benefits" might be a good predictor for unemployment as long as the underlying reason for search activity is actually getting unemployed. Two spikes in search activity are apparent in Figure 3.2.1 and they are presumably associated with changes in the federal-level unemployment benefits extension and not with the level of unemployment rate. The sudden spikes coincide with the bad forecasts depicted in Figures 5.2.7 and 5.2.8. After controlling for the two events by using separate indicator variables for these months, the improvements from Google data compared to the benchmark are 10 percent higher on average than the improvements reported earlier. The controlling was done ex post, and it is relevant to ask whether this information would have been available for the forecasters at the date of prediction.

How about the external validity of the results: Are these findings US specific or more general? To my knowledge, the literature on unemployment forecasting with Google data contains no negative results. However, this could arise from some sort of publication bias. In the end, finding positive correlations from billions of search queries is not hard.

In Germany, Askitas and Zimmermann (2009) report strong and significant correlations between the unemployment rate and Google search activity. They find a good fit between the forecasts and realized unemployment rate. However, they do not quantify the value of Google data in out-of-sample forecasting in Germany. Askitas and Zimmermann (2009) control for lagged values of the series, but do not have a baseline model without Google data for comparison.

McLaren and Shanbhogue (2011) discuss the use of Internet search data as economic indicators in the United Kingdom. According to their results, Google searches contain relevant information for forecasting UK unemployment. McLaren and Shanbhogue (2011) find 12.6 percent improvements on average in root mean squared error (RMSE) compared to the univariate benchmark when using data on Google searches.

In Israel, Suhoy (2009) finds a statistically significant coefficient on a variable extracted from Google data when predicting the unemployment rate. Suhoy

(2009) controls for the lagged effects by using the autoregressive–moving-average ARMA(2,2) specification with an additional Google variable. The author does not assess the incremental improvements in prediction accuracy from Google data.

Previous authors have also studied whether Google searches predict unemployment in Norway, Turkey, France, and Spain. In Norway, Anvik and Gjelstad (2010) report -0.2–15.3 percent improvements in root mean squared error for one-step-ahead predictions compared to the AR(1) benchmark, depending on selected search terms. Chadwick and Sengul (2012) note that a model that contains Google search query data gives 38.3 percent more accurate one-step-ahead predictions than a univariate benchmark in terms of root mean squared error in Turkey. Fondeur and Karamé (2013) use Google search data to predict youth unemployment in France. They point out that using Google data improves the accuracy of the one-step-ahead forecasts by almost 27 percent when compared to a simple baseline. In Spain, Vicente et al. (2015) claim a15 percent decrease on average in several forecast error measures compared to a moving average benchmark. Vicente et al. (2015) emphasize that the results also hold for Spain, where the level of unemployment and the number of discouraged workers is high compared to the countries of previous studies. However, each of these papers focus on predicting the present only.

Several studies extend the forecasting horizon up to three steps ahead in countries other than the US. In Italy, D'Amuri (2009) finds improvements in forecasting accuracy up to two months ahead when data on Google searches are included in the model. D'Amuri (2009) reports a wide range of results but suggests that the improvement against a simple benchmark is around 30 percent. Pavlicek and Kristoufek (2014) forecast the unemployment rate with Google data in the Czech Republic, Hungary, Poland, and Slovakia. The authors find a 30 percent gain in accuracy on average for predicting the present and 20 percent for predicting the following month.

However, these studies are not directly comparable with this study since they have been made using different methods. In particular, the reported improvements against a benchmark depend on specific modeling choices, such as error criterion and the benchmark model. My previous work on using Google data to predict the unemployment rate in Finland (Tuhkuri 2014) provides an opportunity for a direct comparison to evaluate the external validity of this study. The institutional framework in Finland is different from the US. For example, the application process for unemployment benefits is quite different, partly because of the historical Ghent system documented by Böckerman and Uusitalo (2006). I use a comparable set of search terms in Finnish to construct a Google Index for Finland and estimate the models with Finnish data.[2]

While using the same forecast comparison methods, my results from Finland suggest that, compared to a simple baseline, Google search queries improve the prediction of the present by 10 percent, measured by mean absolute percentage error. Moreover, predictions using search terms perform 39 percent better over

[2]Source: Statistics Finland, Labor Force Survey, 2014.

the benchmark for near future unemployment three months ahead. Google search queries also tend to improve the prediction accuracy during the global recession. The results from Finland are similar to those of the US. The Google searches improve the prediction accuracy, however, more so in Finland than in the US. One reason for this might be that in Finland, 72 percent of all claims for unemployment benefits were made online in 2014.[3] Internet use is also more common: in 2014, 96 percent of all residents aged 16–64 reported Internet use.[4] Still, the results for both US and Finland might depend on the composition of the users of Google.

To conclude this country comparison, the qualitative results seem to hold for several developed countries. Furthermore, the country-level differences do not seem to be substantial.

However, these studies do not discern whether Internet search data would help to predict the unemployment rate in countries where Internet use and coverage is limited. Nevertheless, in an analysis the previous literature in Europe, I do not find a relationship between the popularity of Internet use at a country level and the reported improvements in forecasting accuracy.[5] Fondeur and Karamé (2013) assess the selectivity issue by comparing the predictive accuracy for different age groups. Their idea is that younger generations are more likely to use the Internet as a tool in the labor market. Consistent with the idea, the authors find a smaller advantage from Google data for predicting unemployment for older generations. Fondeur and Karamé (2013) argue that this illustrates the impact of selection bias, emphasized previously by D'Amuri (2009). However, there could be many other reasons for this. For example, from 2004 to 2012, the volatility of youth unemployment was higher than that of the older age groups.[6] Popularity of Internet use is not the only factor that would affect the forecasting accuracy. The results may differ, for example, for China, where Internet use is relatively common, but Internet search behavior might be different due to political restrictions.

As I said earlier, the reported figures are not directly comparable across countries because previous authors have made different modeling choices. Meanwhile, the flip side is that we see if the results hold regardless of different methods. The comparison is not perfect, as the other factors are not fixed.

A large majority of the literature, as well as this study, uses a (pseudo) out-of-sample forecast environment to compare the predictive accuracy of models with and without a variable constructed from Google data. However, the number of models varies from one baseline and the alternative (Tuhkuri 2014) to over 500 models (D'Amuri and Marcucci 2012). Most of these studies use autoregressive integrated moving average (ARIMA) models as a baseline. The level of sophistication of models varies among previous papers as well. Many studies, including Tuhkuri (2014), Varian (2009a, 2012), and this thesis, adopt seasonal

[3]Source: Finnish Social Insurance Institution Internal Data, 2014

[4]Source: Statistics Finland, Use of information and communications technology by individuals, 2014.

[5]Source: Eurostat, Internet use and frequency of use 2014.

[6]Source: Eurostat, 2015.

or nonseasonal AR(1) for the main benchmark. In comparison, McLaren and Shanbhogue (2011) employ the AR(2) benchmark, while Anvik and Gjelstad (2010) utilize the AR(3) model, among other benchmarks. Vicente et. al (2015) select a moving average (MA) model of order 2 for a benchmark describing the unemployment dynamics. Most of the previous studies identify their models by utilizing the whole sample since 2004 under the implicit assumption of no structural change. However, for example, Vicente et al. (2015) account for structural change in the unemployment series by imposing a level shift. None of these modeling choices appear to have a meaningful impact on the results.

Various other approaches have also been proposed to answer whether Google searches predict the unemployment rate. For example, Askitas and Zimmermann (2009) argue that a stationary long-term relationship should exist between the unemployment rate and Google searches. For this reason, they use an error-correction model specification developed by Engle and Granger (1987). In this case, however, cointegration is not necessarily plausible. This arises from the mechanics of Google data, which are normalized between 0 and 100, and the value of 100 is applied on every observation period. In contrast, the unemployment rate is not subject to such transformation. Although there exists a relationship between the two variables, it is not necessarily stationary. Visual analysis of the series in Askitas and Zimmermann (2009) as well as in Figure 3.2.1 supports this conjecture. Furthermore, in the long run, neither of the variables is likely to exhibit a global unit root, since they are bounded within a fixed interval (Koop and Potter 1999).

Still, McLaren and Shanbhogue (2011), Suhoy (2009), and D'Amuri (2009) favor the use of the unemployment rate in differences rather than in levels. These studies find that volumes of relevant Google searches predict changes in the unemployment rate. Previous literature suggest that the results are not probably driven by a "spurious regression" described by Granger and Newbold (1974) and Phillips (1986). The concept refers to a false relationship between two integrated or nearly integrated time series, when the series are actually independent of each other (Granger and Newbold 1974).

From another perspective, Chadwick and Sengul (2012) account for model uncertainty utilizing the framework of Bayesian model averaging (BMA). They emphasize that a large number of potential predictors increases the chance of model misspecification.

Google data are released on a weekly basis, while the unemployment rate is usually released on a monthly basis. In this study, I aggregate the Google data for simplicity to a monthly level. The temporal aggregation is also performed in many previous studies, including D'Amuri and Marcucci (2012), Goel et al. (2010), and Choi and Varian (2012). As opposed to my approach, Fondeur and Karamé (2013) estimate a model with mixed frequency data to use all the available information. For this purpose, they use state-space representation. Their approach is highly relevant because one of the main advantages from Internet search data, and big data in general, is that it is usually available on a higher frequency than traditional data. I do not employ their approach in this study because it is not necessary in order to answer the research question.

More generally, Google data consist of a large number of variables observed at a relatively high frequency and fine geographical level. A successful forecasting method would take advantage of those dimensions.

Most previous work, as well as this thesis, maintain a simplifying assumption that Google data are available a month earlier than the unemployment rate. The assumption is reasonable, depending on the institutional framework. Implicitly, several previous studies justify the assumption by the fact that the publication lag for the unemployment rate is actually longer than one month. If so, then their results on the usefulness of Google data might have been underestimated. Askitas and Zimmermann (2009) account for the statistical release schedule in Germany by constructing two monthly level variables: one describing search activity in the first half and the other describing the activity in the second. The unemployment rate is released in the middle of the month, and the authors want to utilize the most recent information at the date of prediction.

The lesson from each of these different methods discussed here is that the selected approach does not seem to affect the qualitative conclusions. This indicates that the finding that Google searches predict unemployment is quite robust. This observation also favors the use of relatively simple methods in this thesis, as the more complicated methods do not appear to offer new insights on this research question. None of the methods utilized in this thesis 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.

To facilitate the analysis in this thesis and demonstrate the practical implications of my results I have constructed a forecast model ETLAnow and an associated website.[7] ETLAnow predicts automatically the unemployment rate for each EU-28 country for three months ahead using data from *Google Trends* database and *Eurostat*, publishing the results every morning. The model utilizes real-time data on the volumes of unemployment-related Google searches as well as the latest official figures on the unemployment rate. ETLAnow forecasts are updated on a daily basis. To my knowledge ETLAnow is the first publicly available forecast model that utilizes data created by search engine logs.

ETLAnow provides also a portal through which the users can modify the set of search terms utilized for forecasting. This idea of crowd-sourcing for selecting variables from big data may turn out to be useful, but much more work remains. Brynjolfsson et al. (2014) present a crowd-sourcing based variable selection method for improving predictions when using Google search data. In specific, human interaction with the model might help to identify when the language is changing. In essence, *Google Trends* data are text data, as are many other big data as well. Methodological and applied work on text analysis might help to improve economic forecasts, which use big data.

[7]The website is maintained by ETLA, The Research Institute of the Finnish Economy, and can be accessed at https://www.etla.fi/en/etlanow-eu28/ with username and password "etlanow2015".

Through this crowd-sourcing exercise I have managed to construct a data set covering search terms for all EU-28 member countries in 25 languages, and the search volumes for these terms. I find strong correlations between the unemployment rate and search volumes for unemployment benefits at the country level.[8] The exercise is still preliminary, as the search terms might be misspecified for some languages.[9] The preliminary exercise, however, demonstrates that Internet search data might be utilized to produce international unemployment forecasts.

There are still limitations for the results of this thesis. The methods utilized in this thesis are relatively simple and do not necessarily represent the ways actual forecasters would use this data. Our understanding on Internet search is still 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. This thesis is not the last word on the topic.

An important caveat also arises for practical implications. The econometric models with Google data may not be the best forecasting methods for unemployment. Recent work surveyed by Snowberg et al. (2013) suggests that prediction markets, for example, could possibly produce more accurate forecasts. Snowberg et al. (2013) 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 thesis 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). Google data give information on these private actions. More generally, Einav and Levin (2014) argue that the bigger the data, the more important it becomes to have a solid framework to organize and reduce the dimensionality of the data in order to draw meaningful inference.

This thesis 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 might prove beneficial.

[8] Source: Eurostat, harmonized unemployment rate, 2015 and *Google Trends* 2015.

[9] Most search terms are, however, confirmed by native speakers.

Chapter 8

Conclusion

This thesis analyzes whether data on Google search volumes could help to 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 to predict the present and near-future unemployment rate.

Two somewhat 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 might be smaller than previously thought.

In summary, using data from the Internet activity is of interest because the Internet plays a major role in the economy (Edelman 2012; Einav and Levin 2013). We can use Google search data to measure activity that is otherwise hard to measure.

Considering the importance of accurate economic forecasts and the scale of Google data, this study answers an important question. The thesis 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.

Bibliography

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.

Alvarez, J. and Arellano, M. (2003). The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators. *Econometrica*, 71(4):1121–1159.

Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3):817–858.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

Antenucci, D., Cafarella, M., Levenstein, M. C., and Shapiro, M. D. (2014). Using Social Media to Measure Labor Market Flows. *NBER Working Paper 20010*.

Anvik, C. and Gjelstad, K. (2010). "Just Google it". Forecasting Norwegian unemployment figures with web queries. *CREAM Publication 11*.

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.

Artola, C. and Galan, E. (2012). Las huellas del futuro están en la web: construcción de indicadores adelantados a partir de las búsquedas en Internet. *Banco de Espana Documentos Ocasionales N. 1203*.

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.

Böckerman, P. and Uusitalo, R. (2006). Erosion of the Ghent system and union membership decline: Lessons from Finland. *British Journal of Industrial Relations*, 44(2):283–303.

Bollen, J., Mao, H., and Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., and Weber, I. (2012). Web search queries can predict stock market volumes. *PloS ONE*, 7(7):e40014.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2008). *Time series analysis: forecasting and control*. John Wiley & Sons., 4th edition.

Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum*, 36(2):3–10.

Brownstein, J. S., Freifeld, C. C., and Madoff, L. C. (2009). Digital disease detection – harnessing the Web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157.

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

Brynjolfsson, E., Geva, T., and Reichman, S. (2014). Crowd-Squared: Amplifying the Predictive Power of Large-Scale Crowd-Based Data. *Working Paper, MIT*.

Castle, J. L., Fawcett, N. W. P., and Hendry, D. F. (2009). Nowcasting Is Not Just Contemporaneous Forecasting. *National Institute Economic Review*, 210(1):71–89.

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.

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.

Curme, C., Preis, T., Stanley, H. E., and Moat, H. S. (2014). Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences of the United States of America*, 111(32):11600–11605.

Da, Z., Engelberg, J., and Gao, P. (2011). In Search of Attention. *Journal of Finance*, 66(5):1461–1499.

Da, Z., Engelberg, J., and Gao, P. (2015). The Sum of All FEARS: Investor Sentiment and Asset Prices. *Review of Financial Studies*, 28(1):1–32.

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*.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, 74(366):427–431.

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.

Edelman, B. (2012). Using Internet Data for Economic Research. *Journal of Economic Perspectives*, 26(2):189–206.

Einav, L. and Levin, J. (2014). Economics in the age of big data. *Science*, 346(6210):1243089.

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

Elliott, G. and Timmermann, A. (2008). Economic Forecasting. *Journal of Economic Literature*, 46(1):3–56.

Engle, R. F. and Granger, C. W. J. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55(2):251–276.

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.

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *Journal of Business*, 38(1):34–105.

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

Frakes, W. B. (1992). Introduction to information storage and retrieval systems. *Space*, 14(10).

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.

Gourieroux, C. and Jasiak, J. (2001). *Financial Econometrics*. Princeton University Press.

Gourieroux, C. and Robert, C. Y. (2006). Stochastic unit root models. *Econometric Theory*, 22(06):1052–1090.

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

Granger, C. W. J. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2(2):111–120.

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement*, 36(3):119–167.

Hall, A. (1994). Testing for a Unit Root in Time Series with Pretest Data-Based Model Selection. *Journal of Business & Economic Statistics*, 12(4):461–470.

Hall, R. (2011). The Long Slump. *American Economic Review*, 101(2):431–469.

Hamermesh, D. S. (2007). Viewpoint: Replication in economics. *Canadian Journal of Economics*, 40(3):715–733.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hannan, E. J. (1980). The Estimation of the Order of an ARMA Process.

Hansen, P. R. (2005). A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics*, 23(4):365–380.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The Model Confidence Set. *Econometrica*, 79(2):453–497.

Hansen, P. R. and Timmermann, A. (2012). Equivalence Between Out-of-Sample Forecast Comparisons and Wald Statistics. *EUI Working Paper ECO 2012/24*.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291.

Hausman, J. (1978). Specification Tests in Econometrics. *Econometrica*, 46(6):1251–1271.

Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12.

Heffetz, O. and Ligett, K. (2014). Privacy and Data-Based Research. *Journal of Economic Perspectives*, 28(2):75–98.

Hulth, A., Rydevik, G., and Linde, A. (2009). Web Queries as a Source for Syndromic Surveillance. *PloS ONE*, 4(2):e4378.

ILO (1982). Resolution concerning statistics of the economically active population, employment, unemployment and underemployment, adopted by the Thirteenth International Conference of Labour Statisticians. (October).

Judson, R. A. and Owen, A. L. (1999). Estimating dynamic panel data models: a guide for macroeconomists. *Economics Letters*, 65(1):9–15.

Kearney, M. S. and Levine, P. B. (2014). Media Influences and Social Outcomes: The Effect of MTV's "16 and Pregnant" on Teen Childbearing. *NBER Working Paper 19795*.

Kholodilin, K. A., Podstawski, M., and Siliverstovs, B. (2010). Do Google searches help in nowcasting private consumption? A real-time evidence for the US. *DIW Berlin Discussion Papers 997*.

Kling, J. L. (1987). Predicting the turning points of business and economic time series. *Journal of Business*, 60(2):201–238.

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.

Krugman, P. (2009). How did economists get it so wrong? *New York Times*, Sept. 6:MM36.

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.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series are nonstationary? *Journal of Econometrics*, 54(1):159–178.

Lanne, M. (2007). Taloustilastojen merkitys empiiriselle makrotaloudelliselle tutkimukselle ja makrotaloudelle. *Kansantaloudellinen aikakauskirja*, 103(4):424–430.

Lanne, M. (2009). Ennustajien tappiofunktiot ja BKT-ennusteiden rationaalisuus. *Kansantaloudellinen aikakauskirja*, 105(4):416–421.

Lanne, M. and Nyberg, H. (2015). Suomen kansantalouden suhdanneindeksi 2009-2014. *Kansantaloudellinen aikakauskirja*, 111(1):6–15.

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.

Leitch, G. and Tanner, J. E. (1991). Economic Forecast Evaluation: Profits Versus the Conventional Error Measures. *American Economic Review*, 81(3):580–590.

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. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529.

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.

McAfee, A. and Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*, 90(10):3–9.

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.

Merton, R. K. (1948). The self-fulfilling prophecy. *Antioch Review*, 8(2):193–210.

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., and Preis, T. (2013). Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports*, 3(1801):1–5.

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.

Nelson, C. R. and Plosser, C. I. (1982). Trends and Random walks in Macroeconomic Time Series: Some Evidence and Implications. *Journal of Monetary Economics*, 10(2):139–162.

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.

Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica*, 46(6):1417–1426.

Oh, S. and Waldman, M. (1990). The Macroeconomic Effects of False Announcements. *Quarterly Journal of Economics*, 105(4):1017–1034.

Papell, D. H., Murray, C. J., and Ghiblawi, H. (2000). The Structure of Unemployment. *Review of Economics and Statistics*, 82(2):309–315.

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

Pesaran, M. H. and Timmermann, A. (2004). How costly is it to ignore breaks when forecasting the direction of a time series? *International Journal of Forecasting*, 20(3):411–425.

Phillips, P. C. B. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics*, 33(3):311–340.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.

Preis, T., Moat, H. S., and Stanley, H. E. (2013). Quantifying Trading Behavior in Financial Markets using Google Trends. *Scientific Reports*, 3(1684):1–6.

Preis, T., Reith, D., and Stanley, H. E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A*, 368(1933):5707–19.

Rodríguez Mora, J. V. and Schulstad, P. (2007). The effect of GNP announcements on fluctuations of GNP growth. *European Economic Review*, 51(8):1922–1940.

Rose, A. K. and Spiegel, M. M. (2012). Dollar illiquidity and central bank swap arrangements during the global financial crisis. *Journal of International Economics*, 88(2):326–340.

Rothman, P. (1998). Forecasting Asymmetric Unemployment Rates. *Review of Economics and Statistics*, 80(1):164–168.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

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

Scott, S. L. and Varian, H. R. (2014). Predicting the Present with Bayesian Structural Time Series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1):4–23.

Scott, S. L. and Varian, H. R. (2015). Bayesian Variable Selection for Nowcasting Economic Time Series. In Goldfarb, A., Greenstein, S., and Tucker, C., editors, *Economic Analysis of the Digital Economy*, pages 119–136. University of Chicago Press.

Silver, N. (2012). *The Signal and the Noise: The Art and Science of Prediction*. Penguin Group.

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.

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics*, 118:26–40.

Stephens-Davidowitz, S. (2015). Searching for Sex. *New York Times*, Jan. 25:SR1.

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

Stock, J. H. and Watson, M. W. (2002). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

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

Toda, H. and Phillips, P. C. B. (1993). Vector autoregressions and causality. *Econometrica*, 61(6):1367–1393.

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

Tuomisto, L. (2015). Nowcasting Swedish Private Consumption with Google Search Data. *Master's Thesis, University of Helsinki*.

van Dijk, D., Franses, P., and Paap, R. (2002). A nonlinear long memory model, with an application to US unemployment. *Journal of Econometrics*, 110(2):135–165.

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. *Unpublished Manuscript, Google*.

Verbeek, M. (2012). *A Guide to Modern Econometrics*. John Wiley & Sons, 4th edition.

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.

Vosen, S. and Schmidt, T. (2011). Forecasting Private Consumption: Survey Based Indicators vs. Google Trends. *Journal of Forecasting*, 30(6):565–578.

Vosen, S. and Schmidt, T. (2012). A monthly consumption indicator for Germany based on Internet search query data. *Applied Economics Letters*, 19(7):683–687.

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

West, K. D. (2006). Forecast Evaluation. In Elliott, G., Granger, C. W. J., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 99–134. Elsevier.

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.

Appendix A

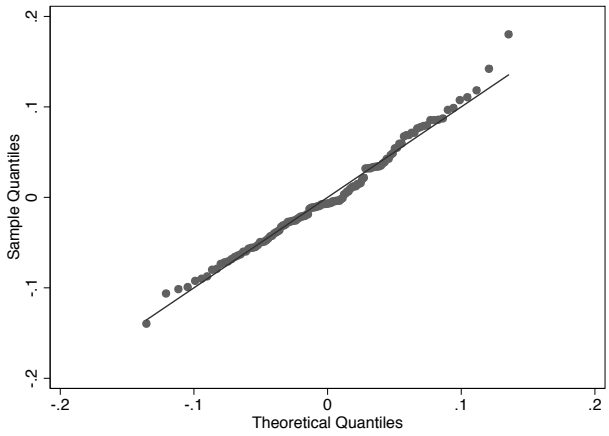

Figure A.0.1: The Q-Q plot of the residuals of seasonal AR(1) model.

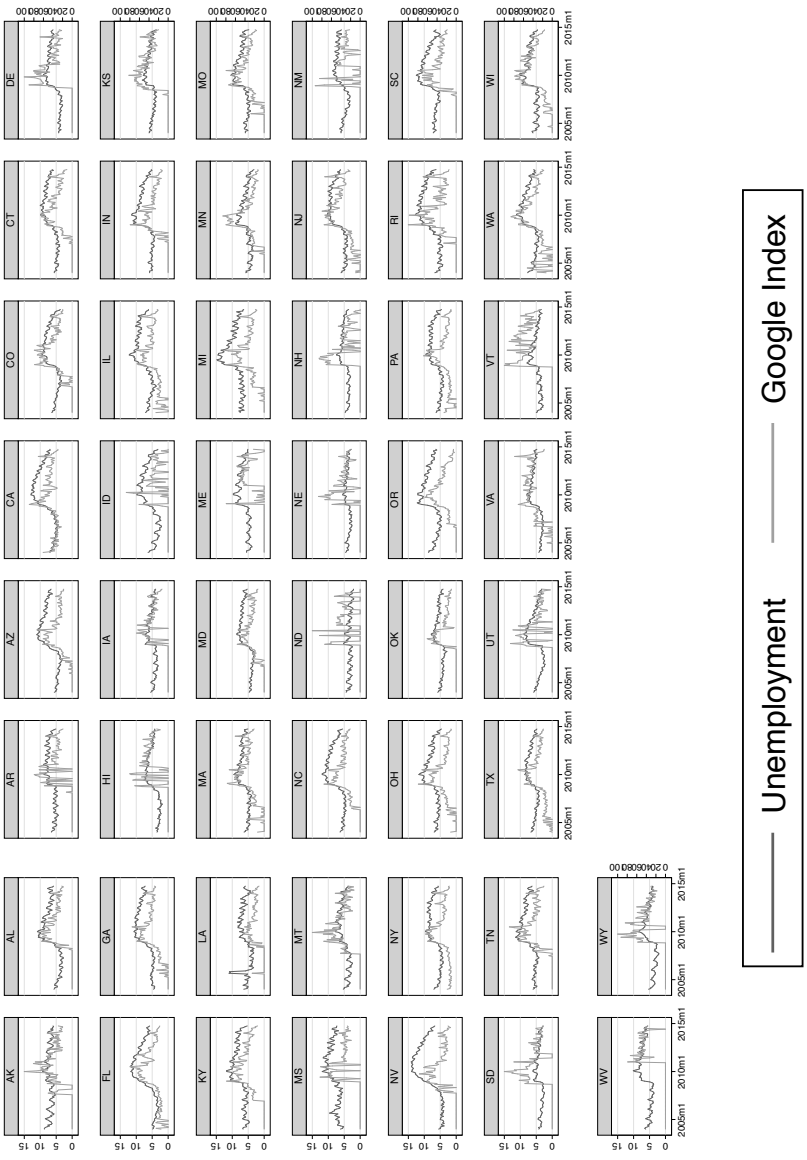

Figure A.0.2: Unemployment rate and Google Index in the United States 2004–2014. Source: The Bureau of Labor Statistics and *Google Trends*.