CENTER FOR RESEARCH IN ECONOMICS AND MANAGEMENT

CREAM Publication No. 11 - 2010

"Just Google it".

Forecasting Norwegian unemployment figures with web queries

Christian Anvik and Kristoffer Gjelstad

BI NORWEGIAN SCHOOL OF MANAGEMENT

BI Norwegian School of Management – Thesis

M.Sc. in Business and Economics

"Just Google It!"

Forecasting Norwegian unemployment figures with web queries

Date of submission:
01.09.2010

Campus:
BI Oslo

Supervisor:
Professor Espen R. Moen

Students:
Christian Anvik (0797628)
Kristoffer Gjelstad (0737058)

Acknowledgements

This project has been a rewarding, fun, educational and exciting journey for both of us. We were both a bit nervous last spring when we chose to follow the tracks of Hal Varian and Hyunyoung Choi and challenged ourselves to write our thesis on forecasting economic variables based on Internet web queries. Their work was the first on the field, and we had apparently only a vague idea about how to tackle the challenge. However, we have never regretted our choice. It has been a true joy to write about this topic since little has been done so far and results are apparently promising. It has made us to think out of the box and to learn new methods we had never touched upon before. Though, we would have not been where we are today without certain persons who we would like to thank for their valuable support and advice.

We would especially thank Professor Espen R. Moen who has, through his brilliant logic and knowledge, guided us throughout this journey. His engagement in our project, with challenging comments and advice, has been very important for our work.

Furthermore we would also like to thank Professor Hilde C. Bjørnland for her support on the methodology and general comments on the thesis. We also greatly appreciate the advice and feedback from Bjørn Roger Wilhelmsen at First Securities.

Lastly, I, Christian, would like to thank Linda for her support by taking more than her fair share of common responsibilities at times when work has been intense. I also truly appreciate your patience when “economics talk” has been overwhelming.

That being said, all the work in this thesis is to be considered our own and we alone answer for any conclusions drawn. This goes also for any errors that could be found in our work.

OSLO. 12. August 2010

Christian Anvik & Kristoffer Gjelstad

Abstract

This thesis explores whether online search queries, represented by Google search queries, contain information useful in forecasting short term unemployment figures in Norway or not. Based on earlier work utilizing online web queries this should be possible, even in small countries. Looking at job search theory supplied with intuition, words from the Norwegian Welfare Administration (NAV) and counseling from the Language Council of Norway we create four Google Indicators that we add to baseline models to check if this reduces the forecasting error (RMSE) of the models. Our findings supports our hypothesis, that Google search contain information useful when predicting short term changes in unemployment. Our top performing model improves the forecasting accuracy compared to its baseline model by 18.3% on average over twelve months. Our best models also outperform the leading indicator "published job advertisements". These are remarkable results given the noise in our data.

Content

1. Introduction

"Prediction is very difficult, especially if it's about the future" -- Niels Bohr

During the last decades the Internet has rose to become one of the top information sources for people around the globe. People go online to read their newspaper, do product investigations, shop for clothes and airline tickets, get their education or search for a new job among thousands of other things. When surfing the Internet search engines play an essential role. Search engines effectively scan cyberspace for websites containing information related to what you are looking for. It is fast and convenient. You reveal your true intentions.

Imagine the possibilities if you could get a grip on what people intend to do in the future. It is an appealing idea which has been of interest to society, and especially trend experts, for ages. Due to the prevalent adoption of search engines it is increasingly possible to capture highly disaggregated data from millions of people and trillions of intentions. Based on this opportunity Hal Varian and Hyunyoung Choi released an article in the spring of 2009 where they argue that fluctuations in the frequency with which people search for certain words or phrases online can improve the accuracy of econometric models used to predict economic indicators. Their work was the basic inspiration for why we chose to pursue this master thesis.

1.1 Objective and Research Question

Online searching is conducted every day, every hour, every second by millions of users. Because search is generally not strategic it provides honest signals of decision-makers' intentions as they happen. Search is not like a survey or any other questionnaire where the provided answers could be affected by surrounding noise or a personal agenda, but it is what the individual truly want to explore or know about a topic, service, product or any other issue. Search reveals consumers and businesses true intentions.

Now it is possible to observe this micro-behavior online instead of relying on surveys or census data that usually come with a substantial lag. The information is obtainable at literarily zero cost through the fairly new tool Google Insights for

Search. Based on the literature review in section two, we see no other study of the relationship between online search behavior and underlying economic variables in Norway, or in any small country with limited amount of search. By studying the Norwegian labor market we hope to shed light on the possible link between the two. It is our ambition to contribute to the understanding of how micro-behavior can be linked to movements in macroeconomic variables. In particular, our paper is an exploratory study where we aim to investigate the possible relationship between search data and movements in unemployment in Norway

Search data in Google Insights for Search are gathered on the basis of search conducted on google.com and related search engines. Given Google's market share in Norway, 81% (GoogleOperatingSystems 2010), the data should be representative for online search behavior in Norway. There is also a positive trend in the Norwegian population to use the Internet to search for vacancies, see exhibit 1.1, and as such it is reasonable to believe that there exists a relationship between actual behavior and search queries on Google. This gives us the following research question:

Do online search queries, represented by Google search queries, contain information useful in forecasting short term unemployment figures in Norway?

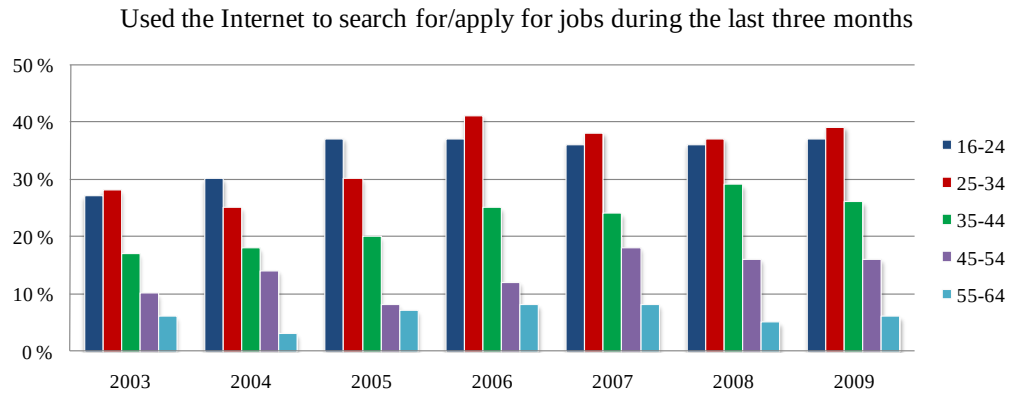


Exhibit 1.1[1]

The methodology we intend to use is based on the ARIMA framework developed by Box and Jenkins. This framework will form the basis for our estimations and

[1] *Source: Statistics Norway. Statistics are based on a yearly questionnaire regarding Norwegians' online habits. Data is divided into age segments.*

models. Google Indicators will be constructed by grouping together keywords analyzed and derived from Google Insights for Search. The indicators will be added to the baseline ARIMA models to form our final models. Next we will indentify the overall top ten best performing models in terms of forecasting ability. We will further investigate if the Google Indicators improve the forecasting performance of the models and finally carry out a robustness test against "published job vacancies", a well known indicator of short term unemployment.

1.2 Thesis Structure

The thesis is divided into three parts and eight chapters. The first part contains the literature review and background theory. In this part we review previous work that has utilized search queries to predict various economic indicators. We also go into job search theory to investigate how individuals and firms conduct search activities. The second part comprises the research design, the data description and the analysis. Here we go through the applied methodology, our data along with how Google Insights for Search work and our estimations and forecasts. The final part covers the discussion and the conclusion. References and appendixes may be found at the end.

The structure of the thesis is as following:

Part I:	Literature review and background theory
Part II:	Applied analysis
Part III:	Discussion and conclusion

Part I – Literature review and Background Theory

2. Literature Review

In the past, prediction of social and economical phenomena was mostly done by using complex mathematical models. The importance of high quality and detailed data to be used in these intricate models was and is immense, and the outcomes are of great interest for both governments and businesses. The complexity of the forecasting models and the tedious gathering of data may now be reduced by the introduction of Google Insights for Search and the utilization of people's search habits. Google Insights for Search is a fairly new and innovative tool in terms of monitoring and predicting economic activity and accordingly there is limited research conducted employing this tool to this date. However, since Varian and Choi's breakthrough article *"Predicting the Present with Google Trends"* (2009), which was reviewed in *The Economist* in the spring 2009, some authors have made significant contributions to the field, as presented in exhibit 2.1. Tough, it shall be pointed out that most articles are either discussion papers or drafts and they are not published in any well known journals, except Gingsberg et. al's work which was published in *Nature*. In this section we present the major contributions (to the extent of our knowledge) to the use of Google Insights for Search.

Contributions to the use of Google Insights in predicting economic indicators				
	Consumption/sales	Unemployment	Housing Market	Other
Google Insights	Varian and Choi (2009) Schmidt and Vosen (2009) Schmidt and Vosen (2009)	Askitas and Zimmermann (2009) D'Amuri (2009) D'Amuri and Marcucci (2009) Varian and Choi (2009) Suhoy (2009)	Wu and Brynjolfson (2009)	Gingsberg et. al. (2009) Constant and Zimmermann (2008)

Exhibit 2.1: Summing up the literature review

2.1 Consumption/Sales

The break-through article by Hal Varian, professor in economics at U.C. Berkeley and Chief Economist at Google, and Hyunyoung Choi was published in April 2009. They argue that fluctuations in the frequency with which people search for certain words or phrases online can improve the accuracy of econometric models used to predict for instance retail sales, automotive sales, home sales or travel. To understand if web queries improve the forecasting accuracy of econometric models they use a seasonal autoregressive model (seasonal AR model of order 1) as a baseline model and add a query index to the baseline model as an extra

explanatory variable, see equation 2.1 for a mathematical presentation. Throughout the thesis we refer to the baseline model as the model without the Google Indicator and the extended model as the model including the Google Indicator.

$$\log(y_t) \sim \log(y_{t-1}) + \log(y_{t-12}) + e_t \quad \text{Baseline model} \tag{2.1}$$
$$\log(y_t) \sim \log(y_{t-1}) + \log(y_{t-12}) + x_t{}^2 + e_t \quad \text{Extended model}$$

Furthermore, they have monthly sales data available and, as we will discuss in chapter 5 under data description, they solve the issue regarding weekly Google data by taking the query index of the first week each month to represent search data for that month. This approach gives emphasis to the simplicity of Google Insights.

By extending the baseline model with the Google index Varian and Choi obtain an improvement in the average absolute values of the prediction errors (MAE) varying from a few percentage points to 18% for motor vehicles and parts and 12% for home sales on a one month forecast (Varian 2009). This is a striking result for any analysts interested in estimating economic activity, sales or production planning among other variables and it was the basic inspiration why we chose to pursue our own master thesis.

In two other studies Schmidt and Vosen (2009 & 2009) compare how well Google Trends forecasts private consumption compared to survey-based indicators. The studies are of particular interest since the robustness of the Google indicator when comparing it to other indicators is tested as well. The first paper was a draft where they looked into consumption in Germany. The monthly survey based indicators are the consumer confidence indicator and the retail trade confidence indicator both conducted on national level on behalf of the European Commission. The Google indicator is constructed with the aid of the category filter in Google Insights and is intended to measure product search activity. It is useful to notice that they solve the issue about weekly Google data by computing monthly averages. Exactly how is not written. Furthermore, they use a seasonal

[2] x_t represents the Google index/indicator of the respective time period and topic of analysis.

autoregressive model as their baseline model in an ordinary OLS regression and look at the percentage change in consumption, i.e. the growth rate, from one quarter to the consecutive quarter. In their estimations the authors investigates whether the extra indicator increases the forecasting power of the baseline model and then if the Google indicator performs better than the survey based indicators in terms of increased forecasting performance. It turns out that Google Insights beats the survey based indicators on all performance measures. In addition to contain valuable information the authors point out that Google Insights is especially helpful since it can be used to predict current levels of consumption as the data is available up to date. However, as they also highlight, due to the limited number of observations in the Google data they are not able to test if the Google indicator is a better indicator than other macroeconomic indicators, only if the Google indicator alone is able to forecast consumption.

Just a couple of months later Schmidt and Vosen did a similar study on consumption in the United States. In this paper they follow the same methodology as above, in addition to extending the baseline model with somewhat arbitrary macroeconomic variables (real income, three-months interest rate and stock prices). They use monthly year on year growth rates instead of seasonally adjusted data or monthly growth rates of consumption due to the recent economic turbulence. The US survey based indicators are the University of Michigan Consumer Sentiment Index and the Conference Board Consumer Confidence Index. The Google Indicator was constructed in the same manner as above. Once again the Google Indicator improves the forecasting performance of the baseline model and outperforms the survey based indicators. When they use the extended macroeconomic model the Google Indicator's information content diminishes, but it remains significant. The problem with lack of observations is not mentioned, as they argued for in the paper on German consumption.

2.2 Unemployment

To our knowledge, there are three papers so far that directly predict the rate of unemployment using web queries. The first paper was conducted by Askitas and Zimmermann (2009) as an exploratory study on the German unemployment rate. Their aim is to demonstrate how web queries can be used to predict economic behavior measured by traditional statistical methods. They construct four groups

of keywords that are used as independent variables in different combinations to find the best model to predict unemployment rates. Weekly Google data is averaged into groups of two weeks, creating Google indicators for week 1+2 and 3+4 for each month. Then week 1+2 of the current month is used to predict unemployment for the current month while the second half of the current month is used to predict the unemployment rate for the next month for each search category. Then they evaluate whether searches in week 1+2 of the current month or searches in week 3+4 of the former month is the best predictor for current month's unemployment. The reason why they divide search data in this way is due to the computation and release of the German unemployment data.

The best model, evaluated in the context of parsimony, prediction success, usefulness and sound economic logic, includes Google data where the keywords "unemployment office OR agency" (K1) and "Stepstone OR Jobworld OR Jobscout OR Meinestadt OR menie Stadt OR Monster Jobs OR Monster de OR Jobboerse" (K4) (German job search engines) are used as indicators employing data from week 3+4 of the former month. Askitas and Zimmermann expect the first indicator (K1) to be connected with people having contacted or being in the process of contacting the unemployment office and as such, they say, it should have something to do with the "flow into unemployment". The second indicator (K4) is expected to be related to job search activities, and they claim that it should be associated with the "flow out of unemployment". They also emphasis the choice of keywords as websites may come in and out of existence, languages change, social and economic levels and other factors which may cause keywords to be invalid. It is therefore important, they say, to choose keywords which remain constant over the time period investigated. However, they do not report a strong theoretical basis for the final choice of keywords, a choice which seems to be based on intuition and trial and error.

Moving on, in a paper written by Francesco D'Amuri (2009) he investigates if a Google indicator has empirical relevance in Italy where unemployment data is released on a quarterly basis. He constructs the Google indicator by using queries for "job offers" ("offerte di lavoro") which is transformed from weekly to quarterly data by taking intra quarter averages. Following a normal ARIMA selection procedure, including minimization of AIC (Akaike's Information

Criterion) and BIC (Bayesian information criterion), an ARIMA (1,1,0) is the preferred benchmark model. The models including the Google indicator performs better than those models without the indicator measured in terms of Mean-Squared-Error (MSE).

However, D'Amuri points to the fact that Google data can be driven by on-the-job search activities as well as the fact that not all workers use the Internet to search for a job and they might not be randomly selected. D'Amuri does not link this point further to relevant job search theory which could be an interesting connection. Despite these issues the indicator constructed performs well in predicting the evolution of unemployment in Italy, and it is superior to other widely accepted leading indicators such as employment expectation surveys and the industrial production index according to D'Amuri.

Together with Juri Marcucci, D'Amuri has done a similar forecast experiment in the US (D'Amuri and Marcucci 2009) where they suggest that the Google index is the best leading indicator to predict the US employment rate. They use the keyword "jobs" as their indicator because "jobs" has high search volumes and is widely used across the range of job seekers, according to the authors. The reason why they do not include other job-related keywords in the indicator is because they are afraid the information conveyed by other keywords could bias the values of the indicator and reduce its predictive ability.

Furthermore, computation of the monthly indicator is aligned with unemployment data released by the government. In their modeling they use different ARIMA models which include and do not include the Google indicator as an exogenous variable, similar to the work of Varian and Choi. Then they run a horserace between the models to check which one is the best in terms of lowest mean-squared-error (MSE). The best model, as they hypothesized, includes the Google indicator and it also outperforms forecasts by the Survey of Professional Forecasters conducted by the Federal Reserve Bank of Philadelphia.

Subsequent to their first work Varian and Choi published a second paper in July 2009 where they predict initial claims for unemployment benefits which is considered to be a well known leading indicator of the US labor market. Initial

claims track the number of people filed for unemployment benefits and as such it is an indication of unemployment. Initial claims data is released with a one week lag meaning that Google data is available 7 days prior to the government's release schedule. Varian and Choi follow the same methodology as in their first paper and apply standard ARIMA selection procedures to select AR(1) as their baseline model. Then they add the Google Insights series, which is constructed by using the category filter for "Jobs" and "Welfare & Unemployment", to see how much this improves predictions. The results show that there is a positive correlation between initial claims and search related to "Jobs" and "Welfare & Unemployment". The forecasts are improved both in the short run (12.9% decrease) and the long run (15.74% decrease) measured by out-of-sample mean-absolute-error (MAE).

At the same time as Varian and Choi published their article on initial claims, Suhoy (2009) came out with her work on predictions of economic downturns in Israel. The aim of the paper is to discover whether Israeli query indices can be helpful for economic monitoring purposes. Her logic is that if the rate of economic activity, measured by Google categories, declines from its long-run trend, the probability of recession increases. In the analysis she investigates the short term predictive ability of query indices with regard to monthly rates of real growth of industrial production, retail trade and service revenue, consumer imports, service exports and employment rates. This resulted in six query categories: human resources (recruiting and staffing), home appliances, real estate, food and drink, and beauty and personal care. She then proposes that it is possible to predict the monthly unemployment rate using the human resources category (which should increase in popularity with increasing unemployment which in turn is an indication of recession) and that the five other categories can be used to measure consumer confidence (which is weakened in bad times and strengthen in good times). Finally the probability of a recession is estimated by using the categories.

The results indicate that the recent economic downturn is captured by all categories. The human resources (recruiting and staffing) category turned out to be the most predictive in determining the probability for a downturn in the economy. For our purpose this suggests that queries about employment may be

well suited to predict the level of unemployment. She also performs a monthly projection of the unemployment rate by applying an ARMA (2,2) model. The fit is greatly improved and the root-mean-squared-error (RMSE) is reduced by adding the human resources category to the baseline model.

2.3 Housing Market

One can imagine that Google Insights could be helpful in improving predictions about present and short term outcomes of the housing market by employing queries related to real estate. At least this was the idea Brynjolfson (a prominent professor at the MIT Sloan School of Management) and Wu (2009) had when they wrote about how Google searches foreshadow housing prices and quantities in the United States. The aim of the paper is to show the power of search queries and that they will play an important role in the future of forecasting. As such they employ a basic econometric framework that can easily be applied across markets. Nevertheless, Wu and Brynjolfsson demonstrate that even a simple framework can be effective and the results they obtain should be given attention.

They use a seasonal autoregressive methodology (as Varian and Choi) to predict both current and future levels of financial indicators (sales volume and price index) of the housing market. The Google indicator is constructed by utilizing the category filter for "Real Estate" on state level which is then added to the baseline model. They run a correlation analysis to examine the relationship between the housing market indicators and the corresponding Google searches which turns out positive. Furthermore they apply fixed-effect specifications to eliminate any influence from time invariant characteristics as well as adding dummy variables to control for seasonality. The final results tell us that the current search index does not have a statistical significant relationship to housing sales while the past search index do. This demonstrates that past search activity has the ability to predict current housing sales. When it comes to the price index both the current and the past search indexes are positively correlated.

Wu and Brynjolfson demonstrate how search queries can be used to make predictions about prices and quantities months before they actually change in the market. An important and interesting comment is made that "search not only precedes purchase decisions, but also is a more honest signal of actual interest and

preferences since there is no bargaining, gaming or strategic signaling involved, in contrast to many market-based transactions. As a result, these digital traces left by consumers can be compiled to reveal comprehensive pictures of the true underlying economic activity" (Wu and Brynjolfsson, 2009). The implication is that businesses and governments can make more effective and efficient decisions.

2.4 Other

Here we report some other interesting studies that have been done by using Google Insights, but which are not directly relevant for our thesis.

A study by Gingsberg et. al. (2009), which received much attention in the media, analyzes the breakout of influenza epidemics. They obtain strong historical correlation between the reported visits to physicians with an influence-like illness and their estimated visits based on a probability model employing Google search queries. Through such monitoring of health-seeking behavior the ability to detect early breakouts of diseases is significantly improved since the reporting lag is only one day (the time it takes for searches to be updated in Google Insights) compared to the 1-2 weeks reporting lag on government data.

Google Insights may also be used to predict other topics commonly interesting to the society like the winner of the presidential election (Constant and Zimmermann 2008) and the winner of American Idol (Nosek 2009).

To wrap up the literature review, we agree with the authors that Google Insights for Search is a powerful new tool which gives insights into intentional and unfiltered individual behavior which is a breakthrough in terms of speed, availability and breadth of coverage. Work done so far is mainly exploratory and the thoroughly empirical papers are yet to be published. This thesis investigates the tool's ability to predict the unemployment level in Norway.

3. Job search theory

Along with the internet-boom at the end of last decade numerous job search engines such as Monster and HotJobs emerged online as to improve the matching process between those who seek work and employers looking for certain skills. With the explosive growth of at-home Internet many economists started to show interest in the effect on the labor market. The sudden opportunity to browse through available jobs based on specific characteristics of the firm and the position itself has drastically reduced workers' cost of job searching. Moreover, people can easily post applications 24 hours a day and they may upload résumés to CV-databases readily available to future employers targeting people with particular skills; online technologies which have also reduced the cost of hiring compared to traditional hiring methods.

It is important to acquire thoroughly knowledge of search theory to understand how search activities are performed and why they exist from the perspective of both workers and firms. Along with the growth of the Internet new cost reducing tools have emerged that have shifted the way individuals and firms conduct search due to a change in the cost of search. This has made the Internet one of the primary sources for job search activities. Hence it is likely that online search queries are an appropriate way of analyzing the movements in the job-search-market, and implicitly the movements in unemployment.

The first section of this chapter is dedicated to derive the theoretical framework for the relationship between search intensity and change in search costs for individuals. We start with McCall (1970) who was the first to present the job search process mathematically and provide a model easily interpretable visually. Being criticized for taking too many exogenous assumptions (ex. Rothschild 1973, MacMinn 1980) we turn to one of the leading theories of the matching process between firms and workers developed by Christopher Pissarides (2000). Having the same intuition both frameworks are then used to summarize the effect of the introduction of online search tools on individual search behavior. We then move on to argue that search activity has shifted towards relatively cheaper online methods in addition to increase the effort put into the process due to a reduction in the cost of searching from such a shift. Secondly we argue that already employed persons perform on-the-job online searching which appears to follow a pro-

cyclical behavioral pattern, that is, employees search more in good times. In the third section we turn back to Pissarides (2000) to review the other half of the job-matching model, namely the firms and their presence in the market and response to reduced hiring costs and unemployed individuals' search activity. We sum up by operationalize the theory into three main subjects of search behavior to be used as basis when defining Google Indicators.

3.1 Individual Job search

3.1.1 McCall's search model

McCall (1970) was the first to mathematically derive the reservation wage in a search model. His simplified model, intentionally meant to describe the stopping strategies where an individual decides to accept a job offer rather than to continue the search process, provides us with a simple framework well suited to visually interpret the effect of reduced search costs on the amount of effort put into the search process for a certain individual. The jobs are independent random selections from a distribution of wages. These offers occur periodically and are either accepted or rejected. The result of the model is intuitive: stop the search process whenever there is a job offer exceeding the lowest wage an individual would accept defined as the *reservation wage*. The optimal search strategy is given by the key equation (3.1):

$$c = \int_{\varepsilon}^{\infty} (x - \varepsilon)\phi(x)dx = H(\varepsilon) \tag{3.1}$$

where:

x = *a random variable denoting the job offer in terms of utility (one per period)*

c = *cost per period of search*

ε = *reservation wage*

$\phi(x)$ = *probability density function of x*

The interpretation is straight forward; the marginal cost of generating another job offer is equal to the increase in the return of generating an additional job offer. $H(\varepsilon)$ is strictly decreasing with ε and convex hence the reservation wage has a unique solution as shown in figure 3.1.

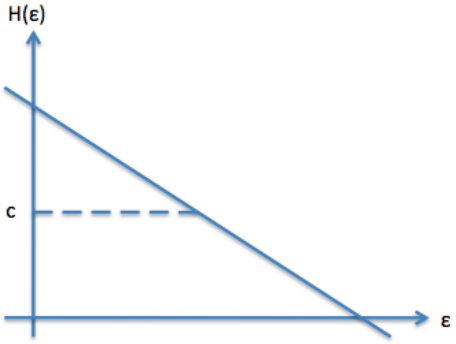


Figure 3.1

There are two ways of increasing the search activity (periods of search), either reduce the cost of search or increase the offer distribution. The probability density function ($\phi(x)$) is known to the individual, hence a larger variance while holding the mean constant would increase the search time for an unemployed. This is due to a larger upside of wage offers increasing the average job offer above the reservation wage. The effect of cost reduction will be discussed in section 3.3 where we introduce the Internet as a search method which lowers search costs.

Roethschild (1973) is questioning the rationale behind profit maximizing firms posting wages above the general market equilibrium. In addition, Diamond (1971) observed that wage dispersion was impossible in a market where employers know the search strategies of the individuals which face a positive cost of search, are equally productive, face the same value of leisure and search randomly without recall among the offers given. With these assumptions, firms setting the wage above the stopping rate can lower the wage without affecting the acceptance decision made by the searcher. Generally, presenting wages as an exogenous variable in this model is in violation with classical economic theory which says that prices should be equal in competitive equilibriums.

3.1.2 Pissarides – The matching process

As McCall's search model lacks the convincing explanations for changes in the flows of unemployment we turn to Christopher Pissarides that with help from Dale Mortensen has in the book *"Equilibrium Unemployment Theory"* (2000) written one of the most influential books on unemployment used in macroeconomic labor theory. His main contribution was the introduction of a matching function describing the formation of relationships between unemployed

workers and firms with vacancies in a setting which allows for market frictions. In this section we will go through the theory required to acquire a thoroughly understanding of how changes in search costs and real shocks in the labor market affects the search intensity of unemployed individuals.

The central idea in Pissarides book is that trade in the labor market is an economic activity. He assumes that there is a well behaved *matching function* that gives the number of jobs formed at any moment in time as a function of the number of workers looking for jobs, the number of firms looking for workers and a small number of other variables. Separation between workers and firms results from firm-specific shocks, such as changes in relative demand or in technology, providing a flow into unemployment. Equilibrium in the system is defined by a state in which firms and workers maximize their respective objective functions, subjected to the matching and separation technologies, and in which the flow of workers into unemployment is equal to the flow of workers out of unemployment. It is assumed to be a unique unemployment rate at which these two flows are equal. The job matching function per unit time in a model where firms and individuals set their level of search intensity is given by:

$$m = m(su, av) \tag{3.2}$$

where:
s = *level of search intensity*
a = *level of job advertising*
u = *unemployment rate*
v = *vacancy rate*

The level of a and s are market averages, and in equilibrium no agent will find it advantageous to change his or her intensity, given that all other agents are in equilibrium. The matching function $m(.,.)$ is assumed to be increasing in both its arguments, concave and homogenous of degree one. Homogeneity, or constant returns to scale, is an important property as it is the only assumption that can ensure a constant unemployment rate along the balanced-growth path. This matching is not a perfect process as some vacancies will receive several applications and others none, creating frictions in the market. The transition probabilities of workers and firms are derived in the following sections.

3.1.3 Pissarides – optimal search intensity

Define the "efficiency units" of searching workers as su. For each efficiency unit supplied, there is a Poisson process transferring workers from unemployment to employment at the rate $m(su, av)/su$. That is the total amount of matches in a given time period per efficiency unit of search provided in the market. From this we can derive the following probability rate that an unemployed individual i will move into a vacant position:

$$q^w = \frac{s_i}{su} m(su, av) \tag{3.3}$$

where:

$m(.)$ = *matching function between unemployed individuals and vacancies*

u = *unemployment rate*

v = *vacancy rate*

s_i = *search intensity for individual i*

The transition rate depends on efficiency units in the market and the individual efficiency search units of worker i. The more units provided by the individual the larger the probability of obtaining a vacant position. In equilibrium all individuals search with the same intensity, s, and firms have the same level of advertising, a. This result gives us the following transition rate for the representative worker:

$$q^w = \frac{s}{su} m(su, av) = m(s, a\theta) \tag{3.4}$$

where:

$m(.)$ = *matching function between unemployed individuals and vacancies*

u = *unemployment rate*

v = *vacancy rate*

θ = *labor tightness (v/u)*

With m new jobs each period and u unemployed individuals the probability rate of obtaining a vacant position for the representative worker is given by m/u. The probability rate is increasing in the three arguments s, a, and θ.

The unemployment rate in equilibrium is affected through the labor tightness, the rate of flow into unemployment, λ, and the transition rate,q^w:

$$u = \frac{\lambda}{\lambda + q^w(s, a, \theta)} \tag{3.5}$$

Equation (3.5) provides us with the rate of unemployment that equates flows into unemployment with flows out of it, when there is no growth in the labor force. The job specific shock (λ) may be caused by structural shifts in demand that changes the relative price of the good produced by a job, or by productivity shocks that change the unit costs of production. In either case they are real shocks associated with a change in technology and will affect the search intensity of an individual.

All work pairs; firms and employees are equally productive. If separated each must undergo an expensive process of search to identify a new match. A realized match yields some economic rent that is equal to the sum of the expected search cost of the firm and the worker including forgone profits for both parties. Wages need to share this economic rent and the cost from forming the job. The rent is shared according to the Nash bargaining game. The individual wage rate derived from the Nash bargaining solution is the w_i that maximizes the weighted product of the worker's and the firm's net return from the job. The first order condition helps us to derive the aggregate wage equation that holds in equilibrium. It can be shown that the maximization process yields:

$$w = (1 - \beta)z + \beta p(1 + c\theta) \tag{3.6}$$

where:

w = wage rate

z = income while unemployed (leisure value, unemployment benefits)

p = value of a job` s output

c = cost variable

β = constant in Nash bargaining solution

According to Pissarides pc is assumed to be the fixed hiring cost experienced by a firm with a vacant position, hence the cost of a vacancy. The hiring costs are proportional to productivity making it more costly to hire more productive workers. Hence, the relationship $pc\theta$ is the average hiring cost of each unemployed worker (since $pc\theta = pcv/u$ and pcv is total hiring cost in the economy). The result implies that workers are rewarded for any reduced hiring costs enjoyed by the firm when a job is formed. β could be interpreted as bargaining power in the Nash bargaining game. Increased β implies a larger bargaining power for the individual workers and increased wages.

For any individual to have an incentive to work the value of the wage rate has to be equal or larger than the value of leisure, $w \geq z$. We assume that the cost of search for an individual i increases in the margin and on average, i.e. raising search intensity is costly. The cost of s_i units of search along with forgone leisure value is given by $\sigma_i = (s_i\,, z)$, hence a person's income during unemployment is given by the difference between leisure income and search cost, $z - \sigma_i$.

We define the present discounted market value of unemployment and employment respectively for U and W. The present value of employment is common to all workers. An unemployed worker chooses the intensity of search, s_i, to maximize the present-discounted value of their expected income during search.

$$rU_i = z - \sigma_i(s_i\,, z) + q^w(s_i; .)(W - U_i) \tag{3.7}$$

The equation states that the present-discounted value is depending on the income during unemployment and the expected gain from a change of searching, given by the transition rate of obtaining a job multiplied with the increased value of obtaining a job.

It can be shown that the optimal s_i satisfies the FOC:

$$\sigma_s(s, z) = \frac{w - z + \sigma(s, z)}{r + \lambda + q^w(s, a, \theta)} \frac{q^w(s, a, \theta)}{s} \tag{3.8}$$

Optimal search intensity is found where the marginal cost of an efficiency unit is

equal to the contribution of one efficiency unit of search to expected net worth. This gain is given by the wage when employed, w, minus the income during unemployment, $z - \sigma_i$, discounted with an effective discount rate consisting of the time rate, r, the rate at which job destruction shocks arrive, λ, and the probability transition rate, q^w, times the marginal change in the transition rate given a change in search intensity of the representative worker given by the last term.

From equation (3.8) we can read out the effects of changes in the labor market on search behavior. Reduced search costs are affecting the individual through two channels. First through lower individual search costs, σ, which obviously increases the optimal equilibrium search, s. And second, indirectly through increased wages.

A comparative-static analysis, holding all other endogenous variables constant, shows that a wage increase has a positive effect on the search intensity, s, because the relative income from work is now higher. We will argue in section 3.5, when analyzing the optimal search activity of firms, that reduced search costs for firms make them raise the vacancy rate, v, resulting in an increased labor tightness, θ. Allowing for changes in the vacancy rate improves the bargaining position of the representative individual as the outside option improves, which raises the wage rate as argued in (3.6). While we know increased wages will reduce the number of profitable vacancies, the net effect of reduced search costs for firms is assumed to be an increase in individual search activity.

A real technology shock that increases the flow into unemployment, an increase in λ, is also affecting the intensity in two ways. First, directly as an increased discount factor in equation (3.7) which is due to the fact that the value of obtaining a job is less given that the probability of losing it has increased. And secondly, it would decrease the individual search activity as it reduces the labor tightness, θ. Workers search less intensively when the ratio of jobs to workers declines, since the chances that they will locate a job declines. This effect is similar to an increase in the discount rate. The effect of a change in labor tightness on the wage will be opposite compared to the last paragraph, and the net effect is assumed to be negative on the search activity.

3.3 Reduction in search costs – The introduction of the Internet

As cost of search is assumed to be mostly related to the transportation costs and the value of forgoing job offers it is reasonable to hypothesize that the introduction of the Internet has reduced this cost. Stevenson (2008) argues that search activity and the growth of search methods developed through the introduction of the Internet have made the search process more extensive. Her data over the relevant period shows that job-search activity among unemployed in the US has increased in the period 1994 through 2003, along with the rate of unemployed actively searching for jobs which has almost doubled from 17% in 1994 to 30% in 2003.

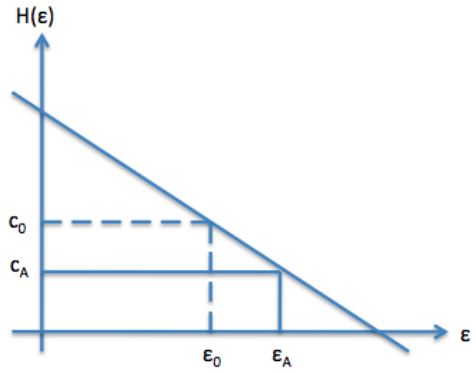


Figure 3.2

This is in line with both McCall's and Pissarides' theory, arguing that a reduction in the search cost from c_0 to c_A (Figure 3.2) increases the amount of effort used on search. Stevenson (2008) also argues that the actual search activity is changed toward "cheaper" search methods often found on the Internet. Though whether these new online tools replace "traditional" job search methods or not depends on whether the methods are complementary or substitutes (Kuhn 2004). Either way, the rapid growth and share of job search activity ongoing on the Internet, alongside the market share of Google, support the hypothesis that actual search behavior would be traceable through Google Insights.

3.4 On-the-job search

Following Burdett and Mortensen (1998) wage dispersion is a robust outcome if information about an individual's job offer is incomplete. Given the assumption of wage dispersion one could extend the interpretation of McCall's model implying

that employed individuals are likely to conduct on-the-job search when the expected level of wages is sufficiently high relative to the cost of search. This is in line with McCall`s own article stating that unemployment could be viewed as just another occupation. Given the argument of decreased search cost through the introduction of the Internet and the Internet usage among Norwegians (Exhibit 1.1) we would observe already employed individuals performing job search when the expectations about increased salary exceeds the search cost. Stevenson (2008) confirms this and emphasizes that already employed persons constitute the vast majority of the job search activity on the Internet. This movement between jobs increases in good times and decreases in bad times as upturns in the business cycle are marked by an intensification of the reallocation of workers among jobs (Cahuc and Zylberberg 2004).

3.5 The Firm

Individual job-searchers represent only one part of a two-sided search market. Employers with vacancies are gathering information as well in order to reduce the risk associated with hiring workers with limited knowledge about their productivity (Spence 1973). The risk is present due to the asymmetric information existing when the seller (the worker) knows more about their own skills than the buyer (the firm). Without gathering of information about the potential candidates the job market turns into a classic lemon market (Akerlof 1970). Training, compatibility with current employees and contract clauses often make wrong hiring decisions an expensive affair.

The interrelationship between firms' cost reduction and search intensity is among others analyzed by Pissarides (2000). In his book he assumes firms are small and each job is either vacant or occupied by a worker. The matching between individuals and firms is according to the matching function derived earlier. The number of jobs is endogenous and determined by profit maximization. When a position is vacant firms search for employees with a cost $c(a_i)$ per unit where a_i is the level of advertising for the vacancy. The hiring cost has the same properties as the individual search cost. We argued earlier that workers are rewarded for the saved hiring costs which the firm enjoys when a job is formed (3.10), but, despite that argument, we mostly ignore the effect of decreased hiring cost on wages throughout this section. Pissarides proved this to be an innocuous simplification.

The transition rate for the firms (the flow of workers into employment) is similar to the transition rate of individuals. Given the Nash equilibrium all firms will choose the same level of advertising, a, resulting in the following transition rate for the representative firm:

$$q^f = \frac{a}{av} m(su, av) = m(\frac{s}{\theta}, a) \tag{3.9}$$

q is increasing in s and a but decreasing in θ as an increase in the number of vacancies relative to the number of unemployed makes it less likely for firms to fill their positions.

The firm's expected profit from one more job vacancy is given by:

$$rV_j = -pc(a_j) + q^f(a_j; .)(J - V_j), \tag{3.10}$$

where:

V_j = the present-discounted value of expected profit from a vacant job

J = represents the present-discounted value of expected profit from an occupied job

r = rental rate for capital k

p = value of a job` s output

c = cost variable

The hiring cost of the firm is now dependent both on the level of productivity but also on the level of advertising. The expected profit of a vacant position depends on the hiring cost of a vacant position, $-pc(a_j)$, and the expected increase in profit of filling the vacancy given the transition rate of the firm, $q(a_j; .)(J - V_j)$. The firm chooses its individual level of advertising, a_j, to maximize V_j. It can be shown that the FOC with respect to a_j at the equilibrium level of advertising is:

$$pc'(a) = \frac{p - w + pc(a)}{r + \lambda + q^f(s, a, \theta)} \frac{q^f(s, a, \theta)}{a} \tag{3.11}$$

where:

w = wage rate

z = income while unemployed (leisure value, unemployment benefits)

p = value of a job`s output

c = cost variable

β = constant in Nash bargaining solution

This equation is similar to the condition for optimal search intensity for individuals (3.8). Holding all other variables fixed it is quite straight forward to see that an increase in the marginal product of labor, a decrease in wage, a decrease in the interest rate and a decrease in the rate of job separation increase job advertising because they increase the expected profit from the job. The amount of time spent on job advertising is positively related to the search intensity of individuals due to positive trading externalities. If unemployed search more often in the market, firms respond by increasing the level of advertising. Labor market tightness has a negative effect on advertising, also due to the trading externalities. More jobs per unemployed worker reduce the chances of finding a worker, making firms advertise less. Hence decreased individual search costs, increased individual search intensity and increased unemployment that decrease labor tightness is related to increased search activity among firms.

With freedom to entry and exit firms will continue to exploit all profit opportunities in the market until the expected profit of a vacant job is equal to zero. Hence in equilibrium the supply of jobs necessitates $V = 0$ simplifying the profit function (3.9) to:

$$J = \frac{pc(a_j)}{q(a_j;.)} \tag{3.12}$$

for all optimal a_j. As $q^f(a_j;.)$ equals the transition rate the expression $1/q^f(a_j;.)$ would give us the time it takes for a vacant position to be filled. Multiplying the time with the hiring costs per unit time should equal the present-discounted gain from an occupied job in optimum.

Since we are interested in the equilibrium where all firms choose the same advertising intensity we substitute $a_j = a$ into the maximization equation. By using (3.10) we find that

$$\frac{c'(a)a}{c(a)} = 1 \tag{3.13}$$

When firms optimize the number of vacancies the level of advertising is chosen such that the elasticity of the cost of advertising is equal to one. Hence, the optimal level of advertising is not affected by any proportionality between the cost of advertising and wages. This implies that firms would never find it optimal to use advertising as an instrument to attract workers when it can adjust the vacancy rate. Though, it is worth to mention that a change in the properties of the cost function might alter the optimal level of advertisement for all firms in equilibrium.

As with the workers we see that the introduction of the Internet has reduced the cost of search and hence decreased $c(a)$. Decreased costs make the firm increase the amount of vacancies in the market as they find more profitable vacant positions, and hereby increase the search intensity. An increase in the vacancy rate would improve individual wage as argued in equation (3.6), and again reduce the amount of vacancies posted by the firms. Although the effect on individual search activity in (3.8) is ambiguous due to increased wages the net effect is assumed to be positive for the representative individual.

Though while it probably does, it is not obvious that the new online tools actually decrease hiring firms' costs. Fountain (2005) hypothesizes that increased search activity among workers due to lowered search costs would increase the number of applicants to each position. Increased number of applicants would increase the amount of time spent to sort out information for the firm. The screening cost would then work in an inverse relationship with the individual search cost.

3.5 Hypotheses and operationalization

Utility maximizing individuals are likely to gather information in the most effective manner and therefore use the Internet as a tool in the search process. A reduction in the cost of search increases individual search intensity. Negative shocks, λ, that increase the flow into unemployment increase overall search intensity. Even though it also lowers the wage for the representative worker

making it less attractive to search, the net effect is likely to be positively related to unemployment; when more people face unemployment the total search volume is likely to increase for unemployment-related search terms.

As the majority of job searches on the Internet are conducted by people already employed, and this variable is procyclical, it is likely that search for new jobs would be negatively correlated with the unemployment rate, while search for unemployment offices and unemployment benefits would be positively correlated with the unemployment rate.

Profit maximizing firms would prefer to use the vacancy rate as a response on search intensity to a change in the cost function. In periods with increased search activity among individuals we would, based on theory, see an increase in search activity among firms due to positive spillover effects. When the level of advertising is high, unemployed workers are more likely to come across a vacant job and respond by increasing their search intensity, and vice versa. Negative real shocks to the economy, which reduce the labor tightness, also increase the search intensity among firms. Which of these effects that dominates is uncertain, though it is likely that the vacancy rate is procyclical indicating a negative correlation with unemployment.

One might believe that the timing of the search activity of individuals is relevant, that is, if the individual search more in the start of the unemployment period or if it continues throughout the whole period at a constant level. One way to assess this is if we assume, as Pissarides does, that changes in unemployment are driven by real shocks, λ, which are shocks such as changes in technology or structural shifts in demand. In such cases the flow into unemployment would be more or less proportional to the number of unemployed. Hence, it would be irrelevant whether the individual searches throughout its period of unemployment or mostly at the time he or she loses the job as the changes would be traceable and could be linked to the changes in unemployment.

Based on the discussion in this chapter we believe that search observed on the Internet reflect more or less the overall search activity in the labor market. The rationale behind this argument is that the introduction of the Internet has

introduced new search tools which have lowered the cost of information making it likely that profit maximizing firms and utility maximizing individuals shift their search activity towards less expensive and more effective tools, hence their search activity on the Internet should be representative for their overall search activity. Given that internet search is representable for total search activity we operationalize the theory in this chapter into different areas of search behavior that can be used as a basis for identifying relevant Google categories in the applied part of the thesis. Relating matching theory to the real world we would believe that job search can be divided into three main subjects of interest:

1. *The job market* – A category where both firms and individuals search for matches, which is related to the flow out of unemployment. This category is complex and has several effects pulling in each direction. First, recalling on-the-job search, we would believe that on-the-job search increases in periods with higher vacancy rate as the outside option motivates search for currently employed individuals making it procyclical with unemployment. Secondly, firms are likely to post more vacancies in good times as more vacancies become profitable, making their search activity procyclical as well. Thirdly, individuals are likely to search more when the unemployment rate increases as more people become unemployed. In total this subject is likely to be negatively related with the unemployment rate though positive when accounting for unemployed individuals only.
2. *Unemployment institutions and offices* – This subject is related to people having contacted or who are in the process of contacting the unemployment institution; hence a flow into unemployment and it is positively correlated with the unemployment rate. However, as The Norwegian Labor and Welfare Administration also provide information about vacant position the correlation might be weakened.
3. *Unemployment benefits* – This is the most intuitive search area when heading into unemployment. It should be positively correlated with the unemployment rate.

Linking this with relevant search theory we propose the following hypothesis about general search behavior among individuals and firms that would later be used to identify relevant search terms:

H: By collecting data about peoples' job search on the Internet from Google we would be able to extract information useful when predicting short-term changes in unemployment in Norway.

In the next part of the paper we will operationalize the specific areas of the job market to test this hypothesis.

Part II – Research Design

In this part we give a brief introduction to time series and especially ARIMA models needed in order to follow the Box-Jenkins framework that will be applied to answer the research question and hypothesis put forward. The Box-Jenkins approach is chosen as we not only intend to identify the usefulness of introducing the Google Indicator in forecasting but also want to identify the best model to predict unemployment. The Box-Jenkins framework is viewed as suitable for this purpose.

The first chapter of the section is dedicated to a short introduction of the properties of time series and the ARIMA model and basic forecasting. The second gives an introduction to statistical data which will be utilized in order to perform the analysis required to answer the research question. The third chapter outlines the basic framework for ARIMA forecasting introduced by Box and Jenkins and includes general penalty function methods to identify the best models, before finally assessing the out-of-sample forecasting ability and comparing them. The chapter relies mainly on the book *"Econometric Models and Economic Forecasts"* (1998) written by Pindyck and Rubinfeld.

4. Introduction to ARIMA forecasting

In some way or another, the purpose of forecasting is to improve decision making. The method though could vary from simple guessing to advanced structural models. Univariate time series is one example, where the idea is to predict future values using only prior values of the time series and the error term. One contribution to univariate modeling was made by Box and Jenkins (1970). They effectively put together, in a comprehensive manner, the relevant information required to understand and use univariate time series for the purpose of forecasting. This is known as ARIMA models, and is described by Box and Jenkins (1970) and later by Box, Jenkins and Reinsel (1994). In this section we provide a short introduction to the basic properties of the ARIMA model along with a section on how to utilize ARIMA models in order to make forecasts to be able to follow the analysis conducted in section 6.

4.1 Properties of ARIMA models

An *Autoregressive Integrated Moving Average* (ARIMA) model is a ***homogenous nonstationary time series*** that describes the value of a time series in time *t* explained as a function of prior values of the same time series (AR - autoregressive) and a combination of random disturbances (MA – moving average). The integrated component (I) refers to the number of times a nonstationary time series must be differentiated to become stationary. The model is specified as ARIMA(p,d,q) where *p* and *q* are the order of the AR(p) and MA(q) components and *d* is the number of differentiations.

A general ARMA model without any differentiation and with p AR lags and q MA lags can be written as:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \delta + \varepsilon_t - \psi_1 \varepsilon_{t-1} - \cdots - \psi_q \varepsilon_{t-q} \quad (4.1)$$

or by introducing the backward shift operator B, where B imposes a one-period time lag each time it is applied to a variable, we can rewrite (4.1) to

$$\phi(B) y_t = \delta + \psi(B) \varepsilon_t \quad (4.2)$$

Differentiating the ARMA model *d* times to achieve stationarity provide us with the general ARIMA model which we can write as:

$$\phi(B) \Delta^d y_t = \delta + \psi(B) \varepsilon_t \quad (4.3)$$

Where $\Delta^{\mathrm{d}} y_t$ is a stationary series and $\Delta^{\mathrm{d}} = (1 - B)^d$ is the number of regular differences required to induce stationarity in y_t.

$\phi(B)$ represents the AR(p) process defined as

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \quad (4.4)$$

and $\psi(B)$ the MA(q) process

$$\psi(B) = 1 - \psi_1 B - \psi_2 B^2 - \cdots - \psi_q B^q \quad (4.5)$$

ARMA processes that also include current and/or lagged exogenously determined variables are called ARMA processes with exogenous variables and are denoted by ARMAX processes, or ARIMAX if integrated (Rachev et al. 2007). Denoting the exogenous variable by x_t, an ARMAX process has the form:

$$\phi(B)y_t = \delta + \psi(B)\varepsilon_t + \omega(B)x_t \tag{4.6}$$

or with *r* different exogenous variables $x_{1,t}, x_{2,t}, \dots, x_{r,t}$ affecting y_t, then equation (4.6) can be generalized to

$$\phi(B)y_t = \delta + \psi(B)\varepsilon_t + \sum_{i=1}^{r} \omega_i(B)x_{i,t} \tag{4.7}$$

where $\omega_i(B)$ is the lag operator of degree n_i that is associated with variable $x_{i,t}$. The challenge is to determine the order of *p,q* and *d* that best describes the time series. This is addressed in chapter 6 through the Box-Jenkins approach. We refer to the book of Pindyck and Rubinfeld (1998) for further information about the characteristics of AR, MA and ARIMA models.

Explaining a univariate time series as a function of prior values of the same time series and a combination of random disturbance has both benefits and disadvantages. The method is easy to apply making it cheap and practical if several models are to be forecasted. ARIMA models have proven to be relative robust when conducting short term forecasts. Montgomery (1998) shows the power of ARIMA models by comparing a simple model to more sophisticated linear and non-linear models when predicting unemployment in the short run. As mentioned, D`Amuri et al. (2009) has similar findings and show how ARIMA models are useful when comparing different models with and without additional exogenous variables, so called ARIMAX models, and how they outperform other widely accepted leading indicators of unemployment dynamics, such as employment expectations surveys and the industrial production index (D`Amuri 2009). Cecchetti (2000) finds that a simple AR model is performing better than leading indicators when predicting inflation, and Bell (1993) has similar results compared to basic structural models. As a methodology for building forecast

models the ARIMA model has proved as good as and even superior to much more elaborate specifications (Greene 2008).

While performing well on a short term forecasts the model says nothing about the causality of the changes in the time series and provides little value except for the forecast in itself. It is not embedded with any theory or underlying structural relationships. Hence, it falls under the general Lucas critique (1976) as the ARIMA model lacks autonomy related to changes in policy and is generally poor at forecasting turning points unless it lies in the long-run equilibrium of the time series. In addition, the traditional identification of ARIMA models is based on subjective analysis of the autocorrelation and partial autocorrelation function and dependent on experience and skills from the forecaster.

4.2 General Forecasting

Our objective is to predict future values of the time series with as little error as possible. As the forecast error is a stochastic variable, we minimize the expected value. Thus, we wish to forecast $\hat{y}_T(l)$ so that $E[e_T^2(l)] = E\{[y_{T+1} - \hat{y}_T(l)]^2\}$ is minimized. This forecast is given by the conditional expectation of y_{T+1},

$$\hat{y}_T(l) = E(y_{T+1}|y_T y_{T-1}, \dots, y_1) \tag{4.8}$$

The computation of the forecast $\hat{y}_T(l)$ can be done recursively by using the estimated ARIMA model. This involves first computing a forecast one period ahead, then using this forecast to compute a forecast two periods ahead and then continuing until the *l*-period forecast has been reached. Let us write the ARIMA(*p*,*d*,*q*) model with the transformed time series, w_t, as

$$w_t = \phi_1 w_{t-1} + \cdots + \phi_p w_{t-p} + \delta + \varepsilon_t - \psi_1 \varepsilon_{t-1} - \cdots - \psi_q \varepsilon_{t-q} \tag{4.9}$$

where

$$y_t = \sum^d w_t \tag{4.10}$$

To compute the forecast $\hat{y}_T(l)$, we begin by computing the *one-period* forecast of w_T, $\widehat{w}_T(1)$. To do so, we write eq. (4.9) with the time period modified

$$w_{t+1} = \phi_1 w_T + \cdots + \phi_p w_{T-p+1} + \delta + \varepsilon_{T+1} \\ -\psi_1 \varepsilon_T - \cdots - \psi_q \varepsilon_{T-q+1} \tag{4.11}$$

Next we calculate our forecast $\hat{y}_T(1)$ by taking the conditional expected value of w_{T+1} in equation (4.11):

$$\widehat{w}_T(1) = E(w_{T+1}|w_T, \ldots) \\ = \phi_1 w_T + \cdots + \phi_p w_{T-p+1} + \delta - \psi_1 \hat{\varepsilon}_T - \cdots - \psi_q \hat{\varepsilon}_{T-q+1} \tag{4.12}$$

where the $\hat{\varepsilon}_T$, $\hat{\varepsilon}_{T-1}$, etc., are observed residuals. Note that the expected value of ε_{T+1} is 0. Now, using the one-period forecast $\widehat{w}_T(1)$, we can obtain the *two-period* forecast $\widehat{w}_T(2)$:

$$\widehat{w}_T(2) = E(w_{T+2}|w_T, \ldots) \\ = \phi_1 \widehat{w}_T(1) + \phi_2 w_T + \cdots + \phi_p w_{T-p+2} + \delta - \\ \psi_2 \hat{\varepsilon}_T - \cdots - \psi_q \hat{\varepsilon}_{T-q+2} \tag{4.13}$$

The two periods forecast is then used to produce the three-period forecast, and so on until the *l*-period forecast $\widehat{w}_T(l)$ is reached:

$$\widehat{w}_T(l) = \phi_1 \widehat{w}_T(l-1) + \phi_1 w_T + \cdots + \phi_p w_{T-p+l} + \delta \\ -\psi_l \hat{\varepsilon}_T - \cdots - \psi_q \hat{\varepsilon}_{T-q+l} \tag{4.14}$$

Once the differenced series w_t has been forecasted, a forecast can be obtained for the original series y_t simply by applying the summarization operation to w_t, that is, by summing w_t *d* times. Suppose, for example, that *d* = 1. Then our *l*-period forecast of y_t is given by

$$\hat{y}_T(l) = y_T + \widehat{w}_T(1) + \widehat{w}_T(2) + \cdots + \widehat{w}_T(l) \tag{4.15}$$

An example of the estimation is provided in appendix 9.3.

5. Data description

5.1 Google Insights for Search

In this section we present the idea and the background behind Google Insights for Search, how it works, a description and discussion of the Google data and a presentation of the search queries we decide to apply in our analysis.

5.1.1 Background

In October 2006 Google launched Google Trends, a new product added to the information overload available online today. Marissa Mayer, Vice President of Search Products and User Experience, stated in her Google blog that the purpose of the tool is to sort several years of search queries from around the world to get a general idea of everything from user preferences on ice-cream flavors to the relative popularity of politicians in their respective cities or countries (Mayer 2006).

Google Insights for Search (from now on referred to as Google Insights) was introduced in the summer of 2008 when Google added new and scientific features to Google Trends intended for more professional use (Helft 2008). The new services included the possibility to download search data directly to a spreadsheet making it easier to analyze the data, identifying regional interest and rising/top search and the ability to filter search queries into location, time and seasons, categories and subcategories. The category filter made it possible to distinguish the brand Apple from the fruit apple. Unfortunately, the category filter is currently not available in Norway.

5.1.2 How It Works

Here we give a description of how you may extract and start to use search data from Google. To use this new tool you go online to Google Insights' website: http://www.google.com/insights/search/#. The front page of Google Insights looks like this (Exhibit 5.1):

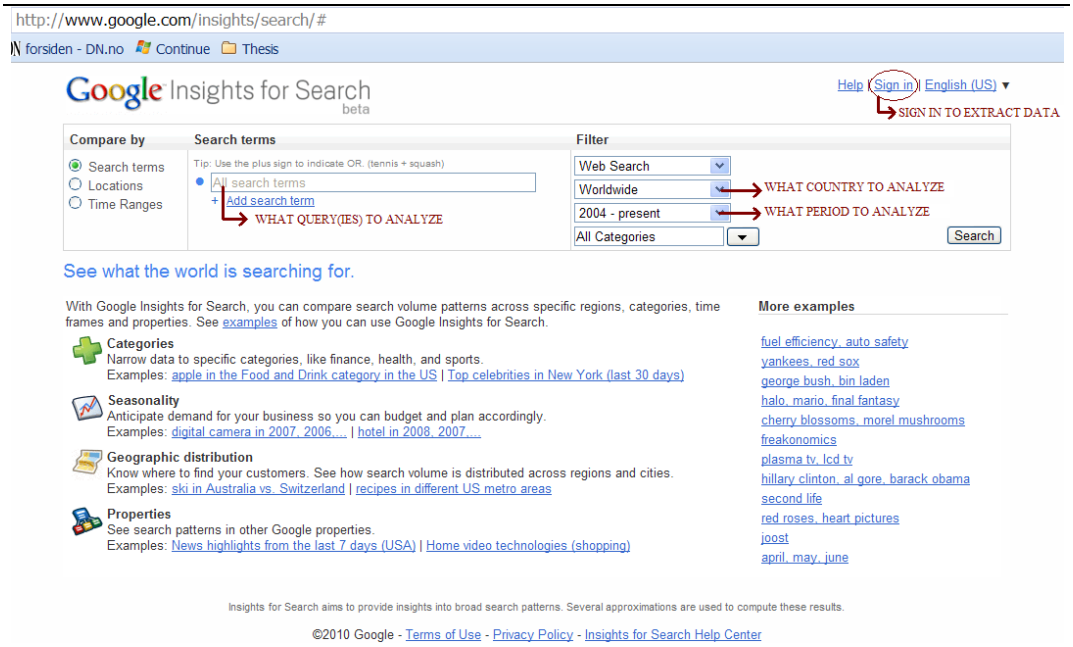


Exhibit 5.1

On the front page you can choose the country/region you want to analyze, the time period under investigation and type in the search queries you want to explore. In order to extract and analyze data you have to create your own Google account and then log into the account before using the tool. After signing in and typing in your preferences and search queries you get this picture (Exhibit 5.2):

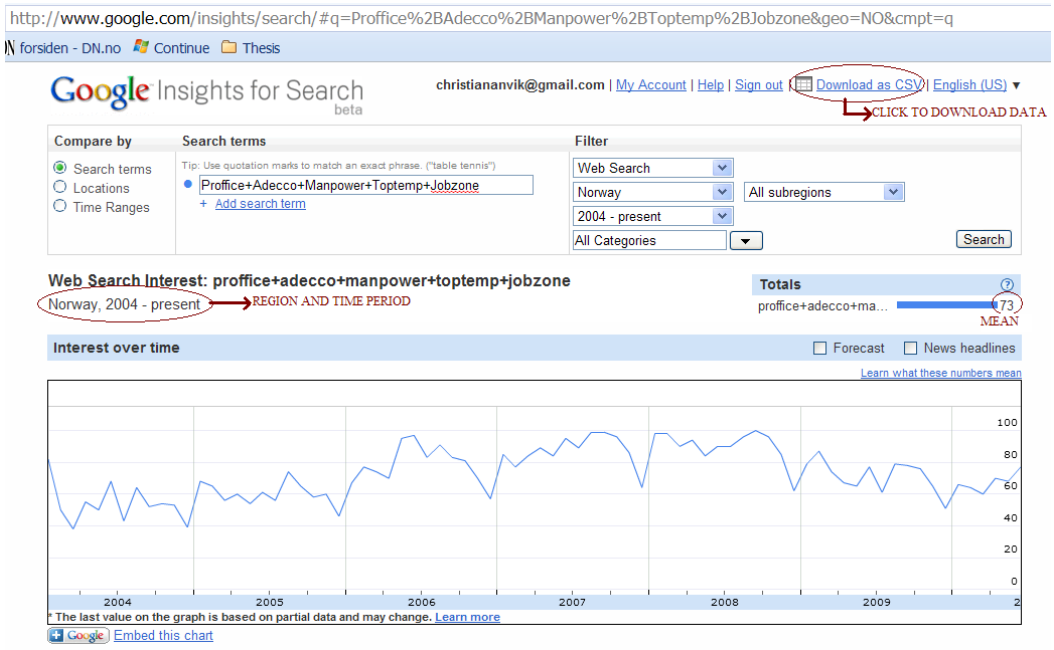


Exhibit 5.2

When analyzing a search term Google Insights uses a portion (based on a random sample) of worldwide Google web searches from all Google domains to compute

how many searches have been done for the terms you have entered, relative to the total number of searches done on Google over time in the specified geographical area (Google 2010). This is called the query index. It is presented in the diagram in exhibit 5.2. The query index, underlying the graph, may be downloaded as a CSV file. This is done by clicking on "Download as CSV file". Results are only shown for search terms that receive a significant amount of traffic. You may also analyze several terms (maximum 5) at the same time and across locations for comparison. This is done by clicking "Add search term". If you want to group multiple terms together (maximum 25), like "Proffice", "Adecco", "Manpower", "Toptemp" and "Jobzone" this is done by using the + sign: "Proffice+Adecco+Manpower+Toptemp+Jobzone". This feature enables you to cover people's overall interest in a topic, as in this case, people's interest in services offered by professional employment bureaus. It also enables you to treat misspellings as for instance in the analysis of "center" you may include "center+centre+centere".

5.1.3 Data

In this section we describe the data you get in the CSV file introduced in the last section. Google's database of "keywords" is updated daily. The database stretches back to January 1st 2004 which provides us with over 6 years of data. Data is normally reported weekly, but for low search volumes monthly data is reported to avoid large variation. In cases of too low volumes nothing is reported. To determine regional correspondence IP address information is used to make an educated guess about query origin.

Furthermore, the data is normalized by dividing the sets of data by a common variable in the certain area, see equation 5.1. The common variable is the total number of search conducted in that area in the specific time period. Normalization is done in order to identify the underlying characteristics of the data which would not be as easily done with absolute values. For example, a search for iPod in Norway seems to be on average higher than in the United States, though this does not mean that Norway has a higher absolute search volume for iPods, but rather that Norwegians are Googling iPods on a more regular basis. Presenting data in absolute values will therefore be less productive as geographical locations with high density will dominate less dense areas which says little about the underlying

trend in the two areas. The core of Google Insights is to identify peoples' propensity to search for a specific term or topic on Google on a regular basis (Google 2009).

$$\text{Normalized value} = \left(\frac{\text{actual search term volum}}{\text{total search volum}}\right) \tag{5.1}$$

In addition to being normalized the data is also scaled for easier interpretation. Equation 5.2 is a mathematically demonstration the scaling procedure. The scale is presented in numbers from 0-100 where 100 represents the search peak. You may have noticed this from the diagram in exhibit 5.2. Insufficient search volume displays a 0 on the scale. Every point on the scale is created by dividing the value by the highest point or 100. The average value over the time period chosen is shown on the right side of Insight Value Index under "Totals" (see the notation "mean" in exhibit 5.2). When comparing and analyzing more than one term, subsequent terms are scaled relative to the term with highest volume.

$$\text{Scale} = \text{Google index} = \left(\frac{\text{normalized value}}{\text{highest normalized value}}\right) * 100 \tag{5.2}$$

There is a possible validity issue to our dataset. When Google Insights derives a portion of Google web searches for a specific term, Google Insights analyzes the likelihood of a random user to search for a particular search term from a certain location at a certain time. For example, if you want to analyze the query "jobb" in Norway during March 2010, Google Insights examines a percentage of all searches for "jobb" within the same time and location parameters. This is the root of a potential validity issue. The user of Google Insights will notice that the data will vary dependent upon the date of extraction. For example, data for the queries "proffice+adecco+manpower+toptemp+"top temp"+jobzone" for Norway from 01.01.2004 – present extracted at 14.06.2010 and 21.06.2010 shows two different time series with variation up to |20| for specific weeks, see exhibit 5.3 for a visual example.

Norway; 2004 - present			
Interest over time			
Week	proffice+adecco+manpower+toptem		
	14.06.2010	21.06.2010	Difference
2004-01-04 - 2004-01-10	47	51	-4
2004-01-11 - 2004-01-17	71	72	-1
2004-01-18 - 2004-01-24	56	56	0
2004-01-25 - 2004-01-31	69	77	-8
2004-02-01 - 2004-02-07	51	71	-20
2004-02-08 - 2004-02-14	44	46	-2
2004-02-15 - 2004-02-21	29	26	3
2004-02-22 - 2004-02-28	29	27	2
2004-02-29 - 2004-03-06	29	28	1
2004-03-07 - 2004-03-13	29	29	0
2004-03-14 - 2004-03-20	29	30	-1
2004-03-21 - 2004-03-27	32	33	-1
2004-03-28 - 2004-04-03	35	35	0

Exhibit 5.3

This problem is intensified for greater variance in the search volumes. For more stable series this problem is dampened. When looking at the query “jobs” in the United States for the period 01.01.2004–present, extracted at the same dates as above, variation is at maximum |3| for the whole period. The reason why we observe this variation dependent upon the date of extraction lies in the nature of Google Insights. Based on the information published on Google Insights’ website we can think of mainly two reasons. The first reason is connected to the actual search volumes Google Insights uses to do the normalization and scaling. As stated above, Google Insights estimates a user’s propensity to search for a specific term in a given location at a given time by analyzing a percentage randomly drawn. If there is larger variation in search volumes we observe larger variation in the series for each daily recalculation of the series. One way of thinking of this could be as follow:

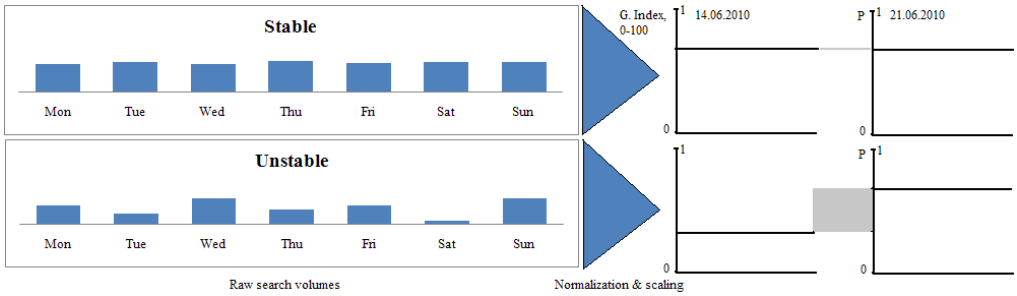


Exhibit 5.4

In the stable scenario there have been almost constant search volumes throughout the week. When Google Insights randomly takes a portion of this specific search term for this week and performs the normalization and the scaling, the reported

search activity would be more stable independent of the date of extraction. In the unstable scenario one would experience the opposite. Google Insights is designed this way in order to identify trends.

The second reason we can think of is found in the normalization and scaling procedure. The data is presented as a query index which starts with the total query volume for a given search term in a given geographical area divided by the total number of queries in that area at a point in time. Then the data is scaled by dividing the numbers by the largest number. This creates a relative relationship between the numbers. If tomorrows search volume is the historical peak, all numbers would be rescaled relative to tomorrow's search volume, and hence, historical data "change" accordingly, as can be seen from exhibit 5.5:

Term "X" in country "Y"				Today	
Week	1	2	3	4	
Term volume	40	100	30	50	
Total volume	1000	1100	1300	900	
Normalization	0,04	0,091	0,023	0,056	
Scale = index	44	100	25	61	
New data available				Today	Tomorrow
Week	1	2	3	4	5
Term volume	40	100	30	50	200
Total volume	1000	1100	1300	900	800
Normalization	0,04	0,091	0,023	0,056	0,250
Scale = index	16	36	9	22	100

Exhibit 5.5

Example – value of index week 1:

Before week 5 search is known $\frac{0,04}{0,091} * 100 = 44$

After week 5 search is known $\frac{0,04}{0,250} * 100 = 16$

In this example the Google Insights query index is shown on the last line in exhibit 5.5. When data is available and shown for only 4 weeks week 2 represents the peak and the numbers are scaled accordingly to week 2. When data for week 5 becomes available ("tomorrow") all data is rescaled because week 5 represents the new peak. This reason, combined with changing numerators in the normalization

process due to the randomization, makes the data change dependent upon the day of extraction. That is, it is not only possible new search peaks in the data as more data become available that make the data slightly unstable, the search peak could also occur in previous periods when Google randomly draws the number to represent search for a specific time point, i.e. the numerator in the normalization process.

The extraction dependent data variation could be a threat to the validity of our data, and is a factor we have to consider when choosing which keywords to include in our analysis. If we are to trust the results from our models they have to be relatively stable in the short run, and not change dramatically on a day to day basis. Therefore we should seek keywords which have relatively high search volumes over time, i.e. they are popular, and show short term stability.

Data used by Google Insights is aggregated from millions of Google users without personally identifiable information. The system also eliminates repeated queries from a single user over a short period of time, so that the level of interest is not artificially impacted by these types of queries.

5.2.1 Google Data Transformation

In this section we explain how we transform weekly Google data to monthly data. The unemployment data we use in our analysis is on a monthly basis while Google data in the CSV file is on a weekly basis, from Sunday to Sunday. Therefore we must, in some way, adjust our Google data. There are several ways of doing this. Varian and Choi (2009) use the first week of the month to represent that month's search data, while others use a plain average. Askitas and Zimmerman (2009) use a slightly more sophisticated method as explained in the literature review (see chapter 2.2). This procedure enables them to capture data and forecast the unemployment rate before current month's unemployment data is announced. The procedure also fully utilizes the power of search data since data is available before announcement, that is, people search before they become registered as unemployed. Both Varian & Choi and Askitas & Zimmerman can use these rather straightforward procedures because search data is quite stable in the US and in Germany.

Inspired by Askitas and Zimmerman's method we have taken our own approach. We wanted to capture the advantage of having data before the release of the unemployment data and at the same time we did not want to lose any data in our transformation, which could happen in the approach by Askitas and Zimmerman since some months contain data involving five weeks. Additionally we had to take into consideration that search data is more unstable in Norway than in the US and in Germany, especially in the period 2004-2006.

We have transformed the weekly data to monthly data by taking a weighted average dependent upon how many days there are in a specific month. Our months run from the 15th of the previous month to the 14th of the current month. Hence, we use approximately two weeks of last month and two weeks of the current month to align our search data with current month's unemployment figures. Figure 5.1 illustrates the concept.

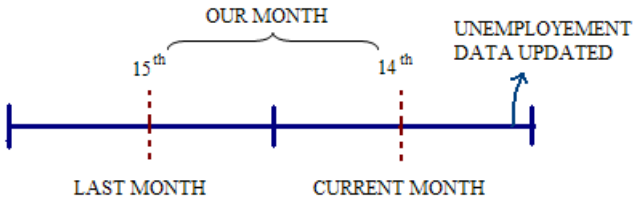


Figure 5.1

To further illustrate the transformation we include an example. Assume we want to construct data for January 2010 for our category "C1: Unemployment and benefits" (see section 5.2.3 for applied search queries). The month will include data from 15.12.2009 until 14.01.2010. That is the last 17 days of December and the first 14 days of January; a month consisting of 31 days. For this specific period we have Google data for the week 13.12.2009-19.12.2009 plus the next three weeks plus 10.01.2010-16.01.2010. Now we need to use a fraction of the first week, specifically 5/7, and a fraction of the last week, also 5/7, together with the three weeks in between to construct a weighted average.

Exhibit 5.6 contains data for the specific weeks. The calculation becomes:

$$\text{January} = \left(\frac{5}{7} * 43 + 30 + 32 + 63 + \frac{5}{7} * 58\right) * \frac{7}{31} = 44{,}5 \tag{5.3}$$

Week	GI weekly	January
2009-12-13 - 2009-12-19	43	44,5
2009-12-20 - 2009-12-26	30	
2009-12-27 - 2010-01-02	32	
2010-01-03 - 2010-01-09	63	
2010-01-10 - 2010-01-16	58	

Exhibit 5.6

The transformation ensures that we do not exclude any data and we are able to forecast unemployment data before they become available.

5.2.2 Smoothing

Generally, there is a slight lack of observations in Google data from 2004 to approximately 2006, which applies to several query series. This is due to lower search volumes at that time and it is a general weakness of using search queries at this early stage. This implies that our Google Indicators will have large variation in the beginning of the series. In the future one could eliminate the first years of the series without having a lack of observations in the analysis, but we cannot do this since we would have too few observations in our analysis. However, there is indication of underlying trends in our data. Additionally, since we use seasonally adjusted unemployment data we should have seasonally adjusted Google series as well. Hence, we want a smoothing algorithm that adjusts for noise and other inaccuracies in our series. There are several methods to do this, though for simplicity we choose double exponential smoothing (Brooks 2008). LaViola (2003) has empirically showed that this method performs equivalently to the Kalman filter and the extended Kalman filter in addition to be faster and easier to implement. Another feature in favor of double exponential smoothing is its sole dependence on current and past values and not any future values of the series, as is the case for instance in the Hodrick-Prescott filter.

Double smoothing of a series is formally defined by the recursions:

$$S_t = \alpha y_t + (1-\alpha) S_{t-1} \qquad (5.4)$$
$$D_t = \alpha S_t + (1-\alpha) D_{t-1}$$

S is the single smoothed series and D is the double smoothed series. α is a smoothing parameter that measures the weight put on former values in the series.

Hence, the new series that is created from double smoothening the original series consist partly of current values and partly of former values. This transformation gives us values on time point whose before were missing in addition to dampen outliers. To better understand how the smoothing works we have added an example. Exhibit 5.7 contains the unfiltered category 1 "C1: Unemployment and benefits" and the seasonally adjusted unemployment figures:

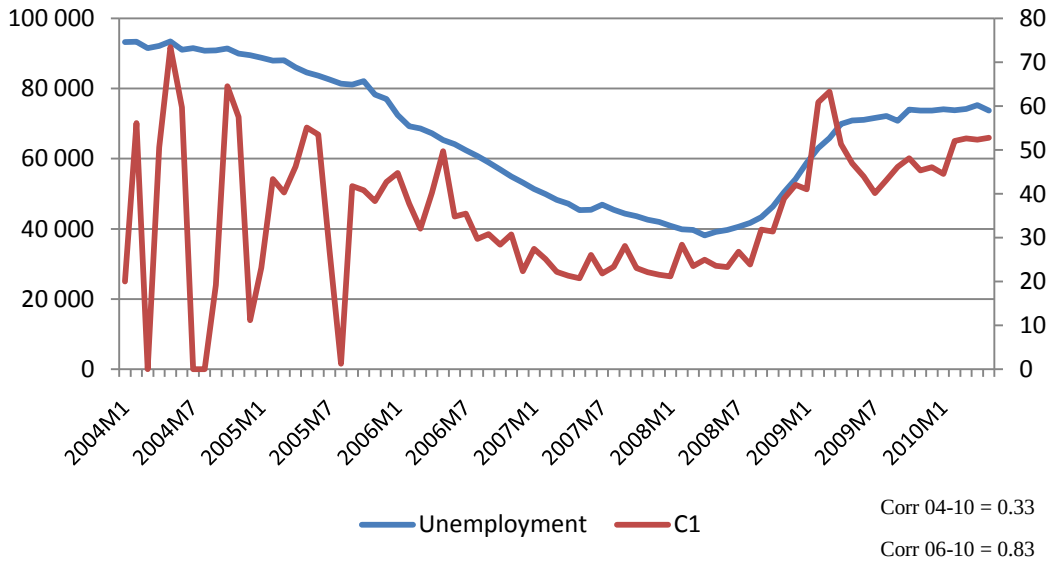


Exhibit 5.7

The series lack some observations in the beginning. However, there is indication of an underlying trend in the data. Exhibit 5.8 presents the after double exponential smoothing graph:

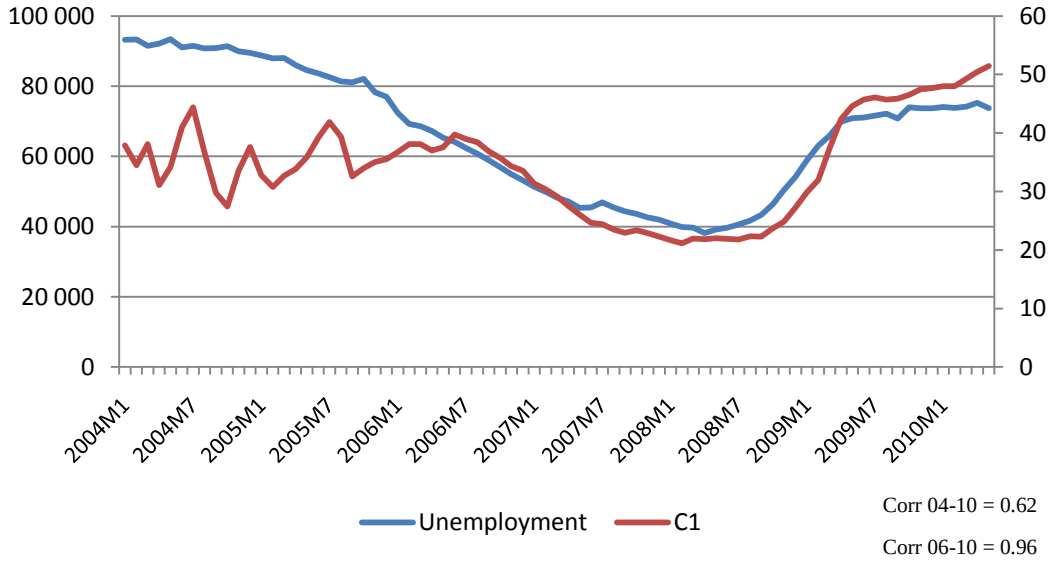


Exhibit 5.8

After smoothing the data we clearly capture the underlying trend better than before. This is also reflected by the correlation coefficient with the seasonally adjusted unemployment data which changes from 0.33 to 0.62. The correlation coefficient had presumably been even better if we had observations at the points whose lacking.

5.2.3 Queries Applied

Now we turn to the queries we actually applied in our analysis and, most importantly, how we chose them. First of all we based our selection on the three categories defined in chapter 3.5. This approach made sure that our selection was grounded in sound economic theory. Next we followed parts of Askitas and Zimmerman's (2009) way of selecting queries by grouping together keywords related to the different categories. That is, we connected several keywords to one category. Having theory in mind we supplied our procedure with intuition, words from the website of the Norwegian Labor and Welfare Administration (NAV) and counseling from the Language Council of Norway. Additionally we used Google's table of "Top Searches" which is presented to you when using Google Insights at the bottom left corner. This is a table of the most popular terms related to the query(ies) you analyze. Finally we checked the popularity and stability of each single keyword as emphasized in section 5.1.3. In the end we got the following categories containing several keywords:

Final categories and keywords				
Categories	C1: Unemployment and benefits	C2: Unemployment institutions and offices	C3: The job market – private employment agencies	C4: Active search
Keywords	Stønad Dagpenger Meldekort Arbeidsledig Arbeidsledighetstrygd	Nav Aetat Nav.no Aetat.no Trygdekontoret Trygdekontor	Manpower Adecco Proffice Toptemp "Top temp" Jobzone	"Ledig stilling" Stillingsannonser

Exhibit 5.9: Final categories and keywords

Each category represents a "Google Indicator". That is, we have used the "+" sign to group together keywords to create each category. Each category (Google Indicator) will be added to a baseline model, in line with earlier work on this field, to investigate whether this improves the model's forecast ability.

Notice that category 2 includes both "NAV" and "Aetat". The Norwegian Labor and Welfare Administration changed its name in the summer of 2006 from Aetat to NAV and made several services available online at the same time. The change is known as "the NAV reform". The name change is likely to create some noise in our data because of the massive media coverage and the public interest. The reform could also explain why we experience the positive trend in the beginning of the series in category 2, see exhibit 5.11.

Below we present the four categories together with the seasonally adjusted unemployment figures. The left hand side axis represents the number of unemployed while the right hand side axis represents the Google index. Correlation coefficients for the periods 2004-2010 and 2006-2010 may be found beneath the horizontal axis in the exhibits.

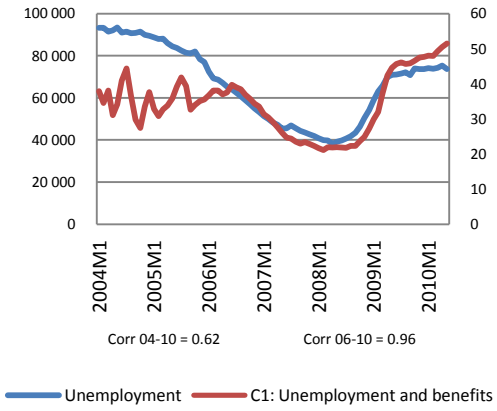


Exhibit 5.10

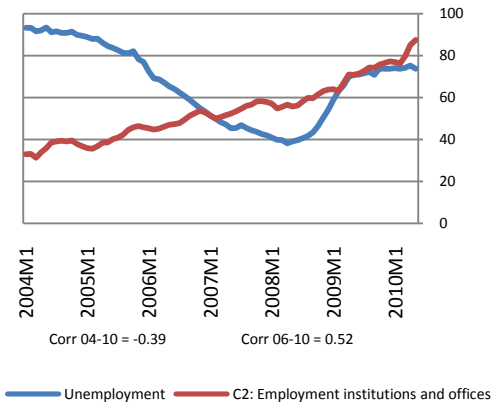


Exhibit 5.11

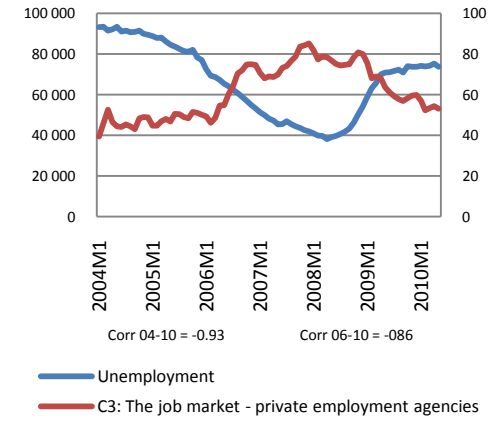


Exhibit 5.12

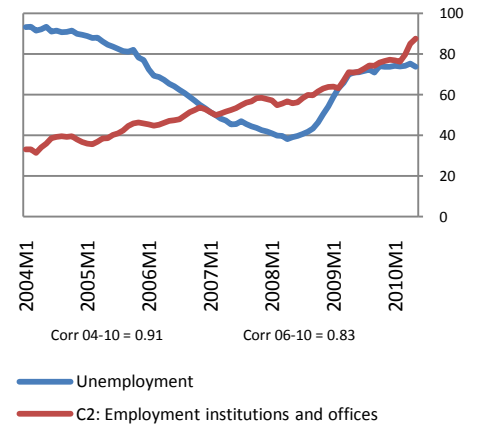


Exhibit 5.13

Most of the categories are intuitive, in line with theory and behave as expected. Category 1, "Unemployment and benefits", moves together with unemployment which is in line with theory form chapter 3. We expected this category to follow unemployment since more people search for unemployment benefits as they become unemployed or are in the state of becoming unemployed. As noticed above, category 2 has a positive trend in the beginning of the series which was expected since more services became available online after the NAV reform. The last part of the series is also in line with theory for the same reason given for category 1.

However, there are some issues that deserve discussion and clarification. Category 3, "The job market – private employment agencies" has a correlation coefficient with unemployment data equal to -0.93. Category 4 "Active search" has a correlation coefficient equal to 0.91. These results might appear counterintuitive since both categories include queries linked to job search, but on the other hand they could have a rather easy explanation. Let us start with category 3. This category captures search from both firms and workers interested in employment agencies or a specific job. We expected this category, in accordance with theory, to correlate negatively with unemployment as it is characterized by hiring firms and on-the-job search, though with noise from unemployed individuals. As there are more vacancies reported to unemployment agencies in good times than in bad times we experience an increase in search activity in good times. Hence, the category behaves countercyclical to unemployment. Category 4 includes search for "ledig stillinger" (available jobs) and "stillingsannonser" (job ads). We expected this category to capture search conducted by workers only, not firms. In bad times the number of unemployed increases which will increase the overall search intensity as theorized in chapter 3. Hence, workers' propensity to search for random available jobs with terms such as "ledig stilling" (available position) and "stillingsannonse" (job ad) will increase as they become unemployed. Thus, we experience category 4 as procyclical to unemployment. In summary, category 3 and 4 move the opposite of each other which is due to what kind of search activity the two categories measure.

5.3 Unemployment data

We now turn to describe the unemployment data that we use in the thesis. Two sources of unemployment data are available in Norway. The first is the Norwegian Labour and Welfare Administration which is called NAV. They report the number of registered unemployed on a monthly basis. The second source is Statistics Norway which in addition to registered unemployed bases its numbers on a survey called "Arbeidskraft Undersøkelsen" (AKU). This survey includes unemployed persons who are not registered in NAV's database and people who are on labor market measures (StatisticsNorway 2010). Statistics Norway's data is released on a quarterly basis. Unemployment data from both Statistics Norway and NAV is based on the three international principles for defining unemployed persons which can be found on nav.no:

1. The person is totally unemployed.
2. The person must have recently tried to acquire a job.
3. The person must be available for immediate employment

Contrary to the numbers from Statistics Norway, NAV's data has two additional requirements. First, the person must not be on a labor market measure such as job training (not including ordinary unemployment benefits) and secondly he must have applied for a job through the NAV system.

Furthermore, there are some additional differences between the two data sets due to the application of dissimilar measurement methods. In the AKU survey it is the person itself that evaluates whether he has applied for a job and is available for work, while in the NAV system this evaluation is done by professionals employed by NAV. AKU numbers will in addition have short term variation due to uncertainty in the sample selection and it also includes persons who have not applied for a job through the NAV system; typically students.

In this thesis we will use data from NAV. The data is available on their website. These numbers do not have any biases related to the population as the data is based on persons registered as unemployed and not a survey. Monthly data is also more useful in predicting short term developments. However, by using NAV's numbers we do not include those who are not registered with NAV. On the other

hand it is more reasonable to use data on registered unemployed in this thesis as these persons are more likely to Google terms related to being unemployed compared to people not searching for a new job because they have chosen to stay out of the labor market and enjoy the benefits of not working. NAV's data could also be said to be more reliable as the evaluation of whether a person is unemployed or not is more objective compared to Statistics Norway's data.

For our purpose it is also reasonable to use seasonally adjusted unemployment data. We do not try to forecast seasonal variations in the series, and since we use ARIMA models accompanied with leading indicators we would have to remove seasonality. The models would otherwise produce unstable forecasts. NAV release seasonally adjusted unemployment figures at the same time as they release the absolute numbers; downloadable in an Excel file from their website. They use the common Cenus-X12-ARIMA method to adjust the data.

6. Forecasting Framework – The Box-Jenkins Methodology

This chapter is dedicated to the analysis. Based on theory and the Google Indicators chosen we apply a simple autoregressive model of order one and add the Google Indicators to investigate if Google search contains information which improves the predictive power of this simple baseline model. The analysis is then extended to identify the best model for predicting unemployment based on the well known Box-Jenkins methodology before we conduct a test of robustness by comparing Google information with today's leading indicator in short term prediction of unemployment, namely the publication of new job adds.

6.1 Test of content

In line with the work of Varian and Choi we utilize the AR(1) model as the baseline model and add the Google Indicator to check if the forecast error (RMSE – defined in section 6.2.3) is reduced, hence if our Google data contains any information of interest. Each Google Indicator is added separately and we try to estimate the level of unemployment in May 2010. The equations are as following:

$$u_t = c + \phi_1 u_{t-1} + e_t \quad \text{Baseline} \qquad (6.1)$$

$$u_t = c + \phi_1 u_{t-1} + \omega_1 GI_t + e_t \quad \text{Extended with Google Indicator (GI)}$$

where

u_t = *unemployment level at time t*
c = *constant*
ϕ_1 = *AR coefficient order 1 on previous values of u*
e_t = *error term at time t*
ω_1 = *Google Indicator coefficient*
GI_t = *Google Indicator value at time t*

The results may be found in table 6.1. Only category 1 failed to reduce the root-mean-square-error compared to the baseline model. This result indicates that there is valuable information contained in the Google Indicators. Knowing this we will use the next section to apply the Box-Jenkins methodology to identify which models that best predict unemployment.

Baseline	Category 1		Category2		Category 3		Category 4	
RMSE	RMSE	% change	RMSE	% change	RMSE	% change	RMSE	%Change
2218,053	2223,019	-0,2 %	1878,473	15,3 %	2171,107	2,1 %	2126,419	4,1 %

Table 6.1

6.2 The Box-Jenkins method

The Box-Jenkins method applies to the use of ARMA and ARIMA models to make forecasts of time series based on past values of the same series. The approach to modeling the time series consists of three phases and is summarized in figure 6.1.

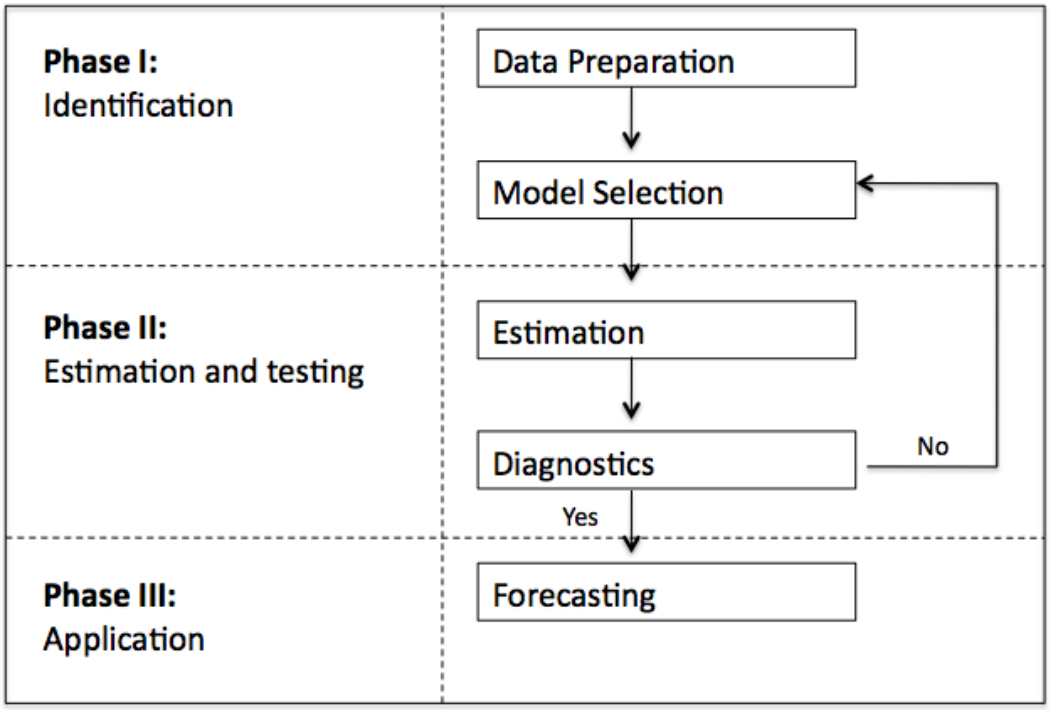


Figure 6.1

The first phase starts with collecting data and examining them graphically and statistically. This includes testing for stationarity. Once the data is stationary we try to identify the correct ARIMA model. The identification of the correct models will be based on both the principles of the Box-Jenkins method along with general penalty functions. Diagnosing the model is usually conducted through sensitivity analysis and residual tests. If a model passes the second phase it can be used to construct forecasts in the third and final phase. This is not necessarily a sequential process as step four and five usually is conducted to evaluate the models chosen in stage three; hence in more advanced ARIMA models this could be a circular process. To evaluate the predictive power of the different models chosen we will apply a pseudo-out-of-sample forecast and compare it against a holdout set, that is, an omitted part in the end of the time series.

6.2.1 Phase I: Identification

The identification process aims to identify the degree of *p,d* and *q* in the ARIMA model. The main tools are testing for stationarity and analysis of the ACF and PACF.

6.2.1.1 Data Preparation

The data should be examined both graphically and statistically. A sample above 50 observations is often necessary when conducting univariate time series forecasting. To be able to apply the B-J methodology, we must have a time series that is stationary or a series that is stationary after one or more differencing. This is because the object of B-J is to identify and estimate a statistical model which can be interpreted as having generated the data. If this estimated model is then to be used for forecasting we must assume that the features of this model are constant through time, and particularly over future time periods. Thus the simple reason for requiring stationary data is that any model which is inferred from these data can itself be interpreted as stationary or stable, therefore providing valid basis for forecasting (Gujarati 2003).

A stationary series is defined by a constant mean, constant variance and a constant autocovariance. In a nonstationary time series we can study the series behavior only for the time period under consideration. Each set of time series data will therefore be for a particular episode. As a consequence, it is not possible to generalize it to other time periods. Therefore, for the purpose of forecasting, such (nonstationary) time series may be of little practical value unless it is differentiated (Gujarati 2003).

In the B-J framework this is done by analyzing the plotted series which should also reveal possible data error and structural breaks that might need intervention. If a time series is stationary it should decay rapidly from the initial value at lag zero. Sub-sections of the dataset should also be analyzed for outliers, data errors, structural shifts or temporary effects where intervention analysis using dummy variables might be necessary.

From appendix 9.1 we observe that the dataset shows no sign of seasonal peaks and troughs. The autocorrelation do not converge to zero as the number of lags

increase and are outside the standard deviation bounds. This is an indication that the time series is nonstationary and we should try to difference the model in order to obtain stationarity.

The differentiated series (Appendix 9.2) indicates that the autocorrelation function converges to zero more rapidly as the number of lags increases, though the autocorrelation are at some large lags outside the standard deviation bounds. This process is rather subjective and it is usual to apply certain formal tests to determine whether the time series is stationary or nonstationary.

The Dickey-Fuller test could be applied to investigate whether a unit root is present in the time series and hence identifying *d* and *D* in the general ARIMA(*p,d,q*)x (*P,D,Q*). Suppose we have a variable Y_t which has been growing over time and is described as follow:

$$Y_t = \alpha + \beta t + \rho Y_{t-1} + \varepsilon_t \quad (6.2)$$

One possibility would be that the time series is growing because it has a positive trend ($\beta > 0$) but would be stationary after detrending (i.e., $\rho < 1$). In this case Y_t could be used in a regression. Another possibility would be that Y_t has been growing because it follows a random walk with positive drift, in this case a detrending would not make the time series stationary, and inclusion of Y_t in a regression could lead to spurious results. We test the null hypothesis $\rho = 1$ against the alternative hypothesis $|\rho| < 1$. Under the null hypothesis the time series follows a random walk. The test is conducted by subtracting Y_{t-1} on both sides of (6.2). Using OLS, one first runs the unrestricted regression:

$$Y_t - Y_{t-1} = \alpha + \beta t + (\rho - 1) Y_{t-1} + \varepsilon_t \quad (6.3)$$

and then the restricted regression

$$Y_t - Y_{t-1} = \alpha \quad (6.4)$$

Dickey and Fuller derived the distribution for the estimator $\hat{\rho}$ that holds when $\rho = 1$ and generated statistics for a simple *F* test of the random walk hypothesis.

Applying a normal *t*-statistics on the estimator $\hat{\rho}$ could lead one to incorrectly reject the random walk hypothesis. A time series though usually have autocorrelation between the residuals, ε_t. This violates on of the assumptions of the OLS-regression and could lead to unbiased but inefficient estimates. The augmented Dickey-Fuller test was introduced to test for a unit root by including lagged changes in Y_t on the right hand side of (6.3):

$$Y_t = \alpha + \beta t + \rho Y_{t-1} + \sum_{j=1}^{p} \lambda_j \Delta Y_{t-j} + \varepsilon_t \tag{6.5}$$

The unit root is tested in the same way as above;

$$Y_t - Y_{t-1} = \alpha + \beta t + (\rho - 1) Y_{t-1} + \sum_{j=1}^{p} \lambda_j \Delta Y_{t-j} \tag{6.6}$$

and then the restricted regression

$$Y_t - Y_{t-1} = \alpha + \sum_{j=1}^{p} \lambda_j \Delta Y_{t-j} \tag{6.7}$$

The unit root test is carried out by testing the joint null hypothesis of $\rho = 1, \beta = 0$. How many lags the time series should have is usually done by experimentation. A rule of thumb is to chose p as low as possible not to lose degrees of freedom while at the same time large enough to remove any possible autocorrelation in the residuals. It is important to note that the power of the test is limited. It only allows to reject (fail to reject) the hypothesis that a variable is *not* a random walk. A failure to reject (especially at a high significance level) provides only weak evidence in favor of the random walk hypothesis.

	ADF	ADF with intercept	ADF with constant and intercept
Seasonal adjusted unemployment	-2,3635**	-2,3662	-2,5709
Critical Value	-1,9453	-2,9012	-3,4717

Table 6.2

Following the estimation listed in table 6.2 the ADF test fails to reject the null hypothesis when we include an intercept and both an intercept and a constant term, while the joint null hypothesis is rejected when testing if unemployment is a random walk at the 95% confidence interval. The time series containing seasonally adjusted unemployment is difference stationary of order one.

6.2.1.2 Model Selection – Identifying p *and* q

The PACF and ACF provide guidelines on how to select pure AR(p) and MA(q) models as stated previous. If $q = 0$ then the time series w_t is said to be an autoregressive model of order *p*, and the autocorrelation will damp out. If $p = 0$ then the time series is said to be a moving average model of order q. Though if both p and q are different from zero the matter of identification become difficult and often subjective relying on training and experience. There will usually be more than one plausible model identified where we need additional methods to determine the best statistical fit. Box, Jenkins and Reinsel (1994) discuss the use of parameters in an ARIMA model and recommend that one should choose the smallest number of parameters for adequate representation, as they exemplify through the well known airline model (Box et al. 1994), known as the concept of parsimony.

A common pitfall when selecting ARIMA models is to over-specify the model through *data mining*, which improves the explanatory power when using in-sample selection criteria such as the root mean squared error (RMSE) but lead to poor out-of-sample forecasting. In-sample criteria are a biased estimator of the out-of-sample prediction error variance. Hence there is a need to use selection criteria which penalize the in-sample residual variance by taking into account the degrees of freedom in the model. Some of the most common criteria are the Akaike`s Information Criterion (AIC) and Schwartz Criterion (SC/BIC). Both

criteria, as we will see, seek to minimize the residuals sum of squares and add a penalty term which takes into the account the number of estimated parameters. The advantage of applying a penalty model is that it is objective and easy to apply, while on the other hand it can only be used to compare different ARIMA models, it has no guidelines for testing and the differences between the values are often marginal.

These criteria are generally written as

$$\text{AIC} = \log\left(\frac{\sum \hat{\varepsilon}_i^2}{N}\right) + \frac{2\,k}{N} \tag{6.8}$$

$$\text{SC/BIC} = \log\left(\frac{\sum \hat{\varepsilon}_i^2}{N}\right) + \frac{k \log N}{N} \tag{6.9}$$

where

$\sum \hat{\varepsilon}_i^2$ = *residual sum of squares*

k = *number of coefficients estimated*

N = *number of observations*

It is quite straight forward to see that the BIC will penalize models harder than AIC whenever $\log N > 2$, which occurs when the number of observations (N) is above seven. Hence, the SC is likely to be more parsimonious in the model selection than the AIC criteria. Both are useful when identifying potential models. ARIMA modeling can become quite extensive if not the maximum order is limited. We will limit our study to models up to three autoregressive and moving average lags while combining up to two GI in the extended model.

Neither AIC nor SC provides any clear statistical test when comparing different ARIMA models and the differences between the statistic might be marginal. Meyler et al. (1998) points out that BIC is favorable compared to the AIC as AIC will usually result in an overparameterised model; that is a model with too many AR or MA terms. Makridakis et al. (1998) states that differences in the AIC values of 2 or less is not regarded as substantial and one may wish to choose a simpler model either for simplicity, or for the sake of getting a better model fit.

Hence we decide to apply a general approach in the identification process choosing the top 10 models in both the AIC and BIC criteria for further investigation. By choosing the top performing models based on both criteria we will assess the models by how they perform in a pseudo-out-of-sample forecast. Table 6.3 summarizes the results.

AIC			
Rank	Model	Category	AIC
1	(3,1,3)	3&4	16,994
2	(3,1,3)	1&3	17,012
3	(3,1,3)	3	17,027
4	(3,1,3)	2&3	17,059
5	(3,1,2)	3	17,116
6	(3,1,1)	2	17,121
7	(3,1,2)	1&3	17,122
8	(3,1,1)	1&2	17,125
9	(3,1,2)	2	17,127
10	(1,1,2)	1&2	17,134

BIC			
Rank	Model	Category	BIC/SC
1	(3,1,3)	3&4	17,279
2	(3,1,3)	3	17,280
3	(3,1,3)	1&3	17,297
4	(1,1,2)	2	17,297
5	(2,1,0)	2	17,298
6	(1,1,1)	2	17,299
7	(3,1,0)	-	17,305
8	(3,1,0)	2	17,307
9	(3,1,1)	2	17,311
10	(2,1,0)	-	17,311

Table 6.3

We observe that there are only ARIMA models containing a Google Indicator among the top ten AIC rated models, while only two among the best rated models with respect to the BIC are pure ARIMA models. As expected, the BIC prefer more parsimonious models on average.

6.2.2 Phase II: Estimation and testing

After analyzing, identifying and estimating the models the result must be diagnosed to assure that the chosen model(s) fulfill the requirements for a univariate time series, that the residuals are white noise. This is done through a *Ljung-Box Q-test* based on the autocorrelation plot; a formal portmanteau test for linear dependence in time series.

$$Q = n(n+2)\sum_{k=1}^{h}\frac{r_k^2}{n-k} \tag{6.10}$$

where n is the number of observation, h is the maximum lag being considered and r_k is the autocorrelation of the residuals. The null hypothesis is that none of the autocorrelation coefficients up to lag h are different from zero; that the data are random. If the residuals are white noise, the statistic Q has a chi-squared (χ^2) distribution with $(h-m)$ degrees of freedom where m is the number of parameters in the model. It is normal to conclude that the data are not white noise if the value of Q lies in the extreme 5% of the right-hand tail of the χ^2 distribution (Makridakis 1998). Both Makridakis (1998) and Pindyck and Rubenfeld (1998) argue that the test is usually done at lag 15 to 20 for low order models, either lag length would have yield the same conclusion and we chose the former.

AIC				
Rank	Model	Category	r^2	Prob
1	(3,1,3)	3&4**	0.674	0.766
2	(3,1,3)	1&3**	0.668	0.86
3	(3,1,3)	3**	0.654	0.834
4	(3,1,3)	2&3*	0.652	0.277
5	(3,1,2)	3**	0.611	0.9
6	(3,1,1)	2**	0.598	0.977
7	(3,1,2)	1&3**	0.619	0.911
8	(3,1,1)	1&2**	0.607	0.946
9	(3,1,2)	2	0.606	0.996
10	(1,1,2)	1&2*	0.585	0.96

BIC				
Rank	Model	Category	r^2	Prob
1	(3,1,3)	3&4**	0.674	0.766
2	(3,1,3)	3**	0.654	0.834
3	(3,1,3)	1&3**	0.668	0.86
4	(1,1,2)	2**	0.570	0.971
5	(2,1,0)	2**	0.560	0.766
6	(1,1,1)	2*	0.543	0.804
7	(3,1,0)	-	0.550	0.976
8	(3,1,0)	2**	0.575	0.992
9	(3,1,1)	2**	0.598	0.977
10	(2,1,0)	-	0.513	0.648

** a category which is significant at 5% level

* a category which is significant at 10% level

Table 6.4

We see from the Ljung-Box LB statistics that the sum of the 15 squared autocorrelations are not statistically significant, indicating that the residuals estimated from the different ARIMA models are purely random, that they are *white noise*. Hence the models are a reasonable fit to the data. It is worth mentioning that the fit is better, larger Q statistic/lower p-value, when including a constant term.

Furthermore, most of the models have a Google Indicator that is significant at the 5% level. In the models with two GIs, only one of them is significant. Google category two seems to be performing well among the models estimated. The R-square is on average lower for models chosen by the BIC which is due to its penalty term being larger than for the AIC. Anyways, all models have an R-square above 0.5 indicating that over 50% of the variance in unemployment is explained.

6.2.3 Phase III: Application

6.2.3.1 Evaluation of the forecast

We will use a holdout set when assessing the out-of-sample prediction ability and comparing the selected models. That is, the end of the time series is omitted to be used as a benchmark for how well the ARIMA models perform when estimating. Since we compare the models on their genuine prediction ability we simply compare the RMSE of the different models on the holdout set. The RMSE over *T* periods is generally calculated by

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_t^2} \tag{6.11}$$

where the error e_t is given by

$$e_t = \hat{y}_T(l) - y_t \tag{6.12}$$

RMSE tells us how many individuals, in absolute numbers, we fail to forecast in the unemployment level of the respective months.

To improve the understanding of ARIMA forecasting and evaluation we provide an example of the forecast for unemployed individuals for May 2010 using the ARIMAX model (2,1,0) with Google Indicator 2 as an additional explanatory factor, this could be found in appendix 9.3.

To see if Google improves prediction over time we have estimated a series of one-month ahead predictions on a quasi-out-of-sample and computed the average RMSE through the last twelve months for the top ten models based on the AIC

and BIC/SC. Each forecast uses only the Google information available upon the time of estimation, which is the Google Indicator two weeks into the month in question and previous values of the unemployment. The result follows in table 6.5 and 6.6. The numbers reported in table 6.5 and 6.6 are the root-mean-square-errors (RMSE) for the top ten best performing models based on the BIC and the AIC respectively. As explained in 6.1.2, the BIC punishes inclusion of additional right hand side variables more than the AIC which is the difference between table 6.5 and 6.6. The first line indicates the order of the ARIMA model and the second line, GI, says which Google Indicators that are added to that model. When no GI number is reported it is a pure ARIMA model without any Google Indicator.

Top 10 SC/BIC – RMSE										
Model	(3,1,3)	(3,1,3)	(3,1,3)	**(1,1,2)**	**(2,1,0)**	(1,1,1)	(3,1,0)	**(3,1,0)**	(3,1,1)	(2,1,0)
GI	3&4	3	1&3	**2**	**2**	2	-	**2**	2	-
jun.09	2361	2450	2349	**1838**	**1847**	2243	2221	**2044**	1728	2019
jul.09	420	416	269	**228**	**95**	1816	672	**734**	615	145
aug.09	244	206	181	**319**	**175**	305	138	**5**	407	301
sep.09	2041	1891	1870	**1300**	**1534**	1718	1789	**1558**	1704	1855
okt.09	2847	2805	3301	**3268**	**3329**	3307	3447	**3212**	3443	3589
nov.09	888	807	833	**575**	**739**	1298	896	**691**	965	812
des.09	1356	1412	1391	**1093**	**1324**	513	671	**798**	744	1369
jan.10	485	332	357	**257**	**803**	438	310	**9**	149	537
feb.10	758	761	725	**53**	**129**	300	323	**57**	581	401
mar.10	274	350	322	**347**	**507**	490	454	**329**	915	424
apr.10	317	313	316	**31**	**96**	62	977	**40**	324	1069
mai.10	2376	2282	2441	**1858**	**1871**	1985	1963	**1912**	2011	2183
Avrg	1197	1169	1196	**931**	**1037**	1206	1155	**949**	1132	1225

Table 6.5

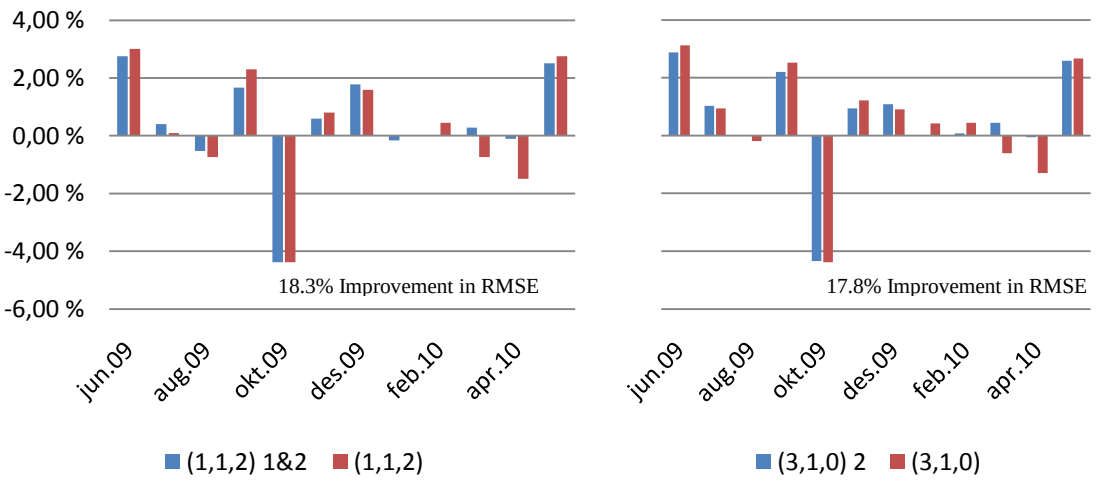
Top 10 AIC – RMSE										
Model	(3,1,3)	(3,1,3)	(3,1,3)	(3,1,3)	(3,1,2)	(3,1,1)	(3,1,2)	(3,1,1)	(3,1,2)	**(1,1,2)**
GI	3&4	1&3	3	2&3	3	2	1&3	1&2	2	**1&2**
jun.09	2361	2349	2450	2115	1380	1728	1686	2064	1539	**1957**
jul.09	420	269	416	251	244	615	106	725	259	**287**
aug.09	244	181	206	228	444	407	706	307	777	**384**
sep.09	2041	1870	1891	1748	1676	1704	1534	1673	1956	**1180**
okt.09	2847	3301	2805	3418	3193	3443	3154	3413	3608	**3290**
nov.09	888	833	807	1647	1128	965	891	817	635	**432**
des.09	1356	1391	1412	245	1095	744	1238	915	956	**1309**
jan.10	485	357	332	374	294	149	281	92	445	**122**
feb.10	758	725	761	1071	823	581	793	538	754	**13**
mar.10	274	322	350	279	376	915	376	895	27	**209**
apr.10	317	316	313	198	328	324	279	283	202	**84**
mai.10	2376	2441	2282	1669	2141	2011	2132	1992	1715	**1847**
Avrg	1197	1196	1169	1103	1094	1132	1098	1143	1073	**926**

Table 6.6

Among the top ten models with respect to AIC and BIC, the four models that return the lowest RMSE number on a twelve month average are highlighted. All of them are ARIMAX models which include one or two Google Indicators, and the best model ((1,1,2) with GI 1&2) performs 19.8% better than the best pure ARIMA model (3,1,0) on average.

6.2.3.2 Robustness of the models

To further investigate the improvement of prediction using GI we compare the four best models with their baseline models in terms of percentage forecast error (Exhibit 6.1):



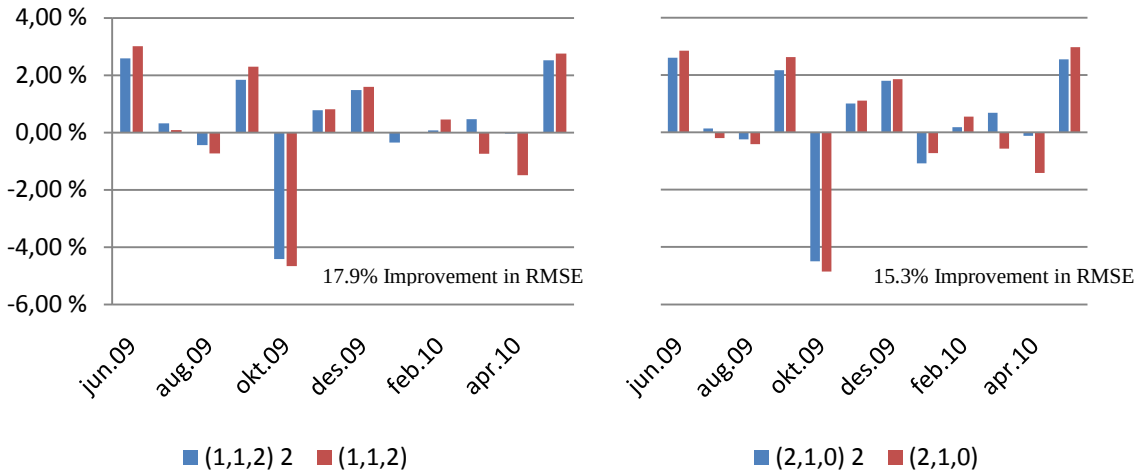


Exhibit 6.1

It is quite straight forward to see that all of these models outperform their baseline model and that this is consistent over the period. It seems to be no clear trend of either over- or underestimating unemployment. It is also peculiar that all the models improve prediction error with at least 15%.

Taking the best performing model, (1,1,2) with GI category 2, and plotting the one period rolling forecasted unemployment against the true unemployment gives us the following graph:

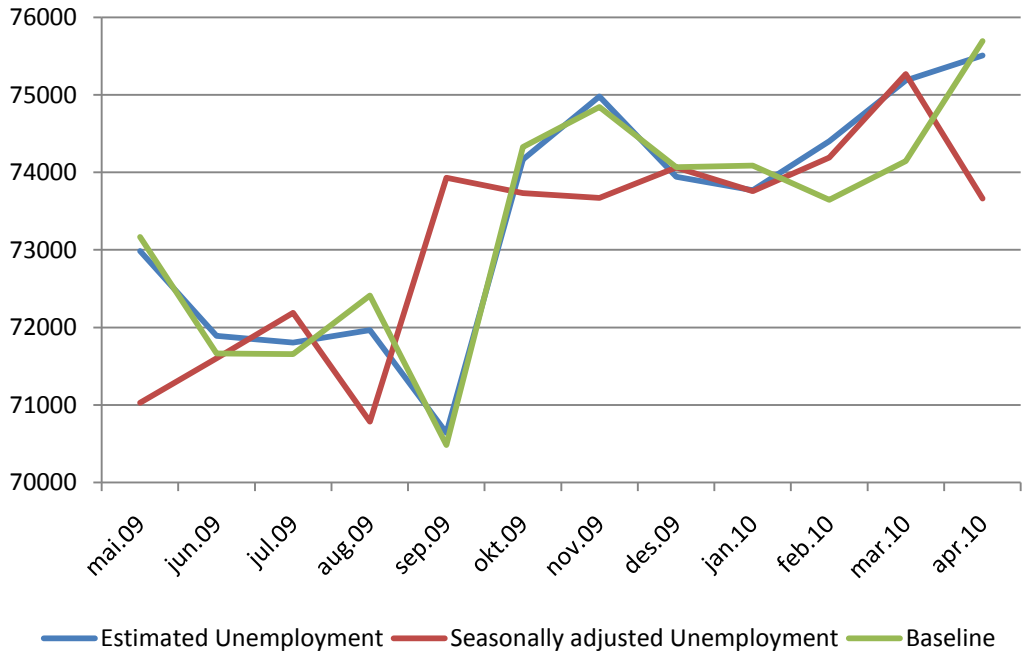


Exhibit 6.2

Both models have difficulties predicting the sudden large increase in unemployment experienced in October 2009. It seems that our model including the Google Indicator is better suited to reduce the error when observing sudden large changes in unemployment, a common feature for our top performers.

Moving on, every month the amount of published job vacancies is registered and published by NAV at the same time as the unemployment figures. The statistic is commonly used by financial institutions, such as First Securities (Bjørn Roger Wilhelmsen 2010), as a leading indicator for unemployment in the short-run and hence it is a suitable comparison when measuring the robustness of the Google queries. The statistic includes all new positions posted at NAV or in any form of media such as newspapers, magazines etc. Due to noise in the time series we use the same smoothening algorithm as for the Google queries to produce values that are closer to the true values of the measurement, and also to be able to compare the GI indicator with posted vacancies on a common ground. We have plotted the series below along with the unemployment below (Exhibit 6.3):

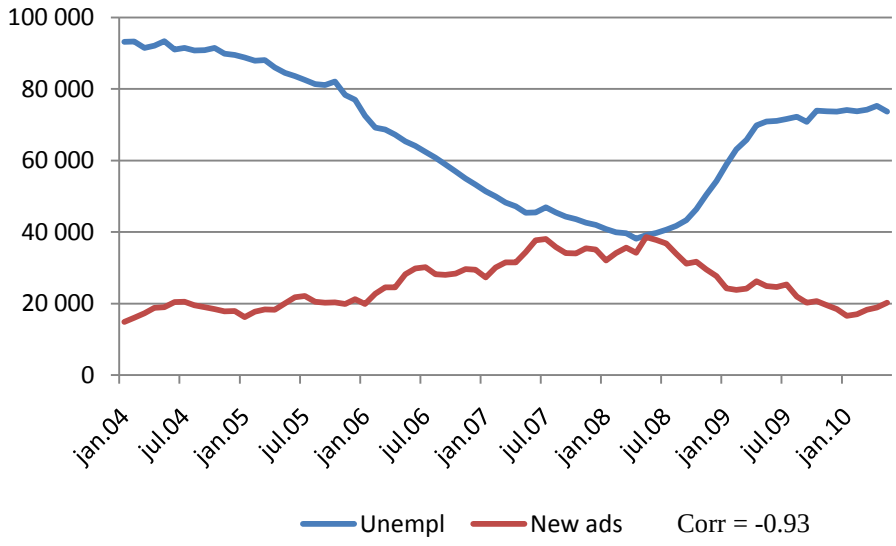


Exhibit 6.3

A correlation coefficient of -0.93 indicates a strong relationship between new published vacancies and the unemployment rate. The estimation and forecast is calculated using the same method as above with a one-month rolling forecast over the period stretching from June 2009 to May 2010. As the statistic is released in the same report as the unemployment figures we have used the number of new ads

in the prior month as explanatory variable for the forecast of this month unemployment. The results are shown in table 6.7 below:

	Google Model		New Ads	
Model	Category	Avrg. RMSE	Avrg. RMSE	% Diff
(1,1,2)	1&2	926	1098	-16 %
(3,1,0)	2	949	1100	-14 %
(1,1,2)	2	931	1098	-15 %
(2,1,0)	2	1037	1171	-11 %

Table 6.7

The ARIMAX models using Google indicators as explanatory variables perform better than the ARIMAX models using published vacancies as explanatory variables with respect to the RMSE over a twelve month average. The best model using Google indicators returns a RMSE 16% lower than that of the published vacancies. It is worth mentioning that all of the ARIMAX models including "new ads" improve the RMSE from the baseline models, but with less than 5% for all four in table 6.7.

Part III – Concluding Remarks and Implications

7. Discussion, Limitations and Concluding Remarks

The hypothesis put forward in the beginning of this thesis was that search behavior on micro level consists of information useful in predicting short-term changes in unemployment. Regression analysis in a simple ARIMA framework lends to support the hypothesis. Furthermore, Google search queries seem robust compared to current leading indicators of short-term fluctuations in unemployment hence strengthening the findings found in the regression analysis. In this section we will discuss some of the major findings along with deviations from the a priori expectations. Furthermore, we will discuss several limitations to using Google search data as an explanatory variable. Finally, we will discuss briefly some implications of the findings and suggestions for future research along.

7.1 Major Findings

In general, the findings in this thesis support the hypothesis put forward; Google search indeed contains information useful when predicting short term changes in unemployment. We have shown that analyzing search data from Google along with a relatively simple model framework yield surprisingly accurate predictions. In addition, the top four models based on information criteria improve the one month ahead prediction accuracy compared to their baseline models with up to 18.3% on average over twelve months.

A leading indicator for short-term changes in unemployment has been the statistics of new published job advertisements. The top four ARIMAX models were used to compare the rolling one-step prediction power to test the robustness of the GIs on a common ground. By replacing the GIs with newly published vacancies we found that our best model containing Google information performed 16% better on average over twelve months than newly published vacancies. This is a remarkable result given the noise in our data series.

7.2 Validity

In addition to discussing the intuition behind the Google Indicators we need to address the validity of the analysis. We should look into the internal and external

validity of the analysis including the causality of the keywords. Internal validity is defined as "the validity of assertions regarding the effects of the independent variable(s) on the dependent variable(s)…In other words, is what has taken place (i.e. phenomenon observed) due to the variables the researcher claims to be operating (e.g., manipulated variables), or can it be attributed to other variables?...The answer depends on the plausibility of alternative explanations." (Pedhazur and Schmelkin 1991). External validity "refers to the generalizability of findings *to* or *across* target populations, settings, times, and the like". (Pedhazur and Schmelkin 1991)

First of all, we need to reassure the internal validity of our search terms otherwise it makes no sense to discuss the external validity of our results. Several questions could be raised regarding the internal validity of the results. Are the keywords we have used correct? Do we measure what we want to measure? Is it even possible to measure unemployment with the use of search queries? Is the causality weak or are there any spurious relationships? As the definition of internal validity states, the answer to these questions depends on the plausibility of alternative explanations. Section 5.2.3 described how we chose keywords to create the Google Indicators. We based our procedure on theory supplied with intuition and help from the Norwegian Language Council, NAV and Google. It was a solid and thoroughly selection procedure so the keywords should be correct and as discussed in chapter 5, they are relatively stable over the period of investigation.

However, it is uncertain whether there is a direct link between the specific keywords and unemployment or if there are any spurious relationships. For example, when we use the keyword "dagpenger" (money as part of unemployment benefits) to represent the demand for unemployment benefits which again measures the flow into unemployment, it is not clear if "dagpenger" measures how many that are unemployed or if "dagpenger" measure how many that receive unemployment benefits. The number of people receiving "dagpenger" is not necessarily the same number of people that are unemployed. "Dagpenger" could merely measure the demand for unemployment benefits and the real variable we measure with this category could be the number of people that actually receives this. The same applies for people searching for "nav" or "nav.no". Since NAV offers several services besides those directly related to

unemployment, these searches capture irrelevant information to our context. One could say that only searches for "arbeidsledig" (unemployed) truly measures unemployment. However, despite this noise, we believe the keywords we have chosen are valid. The reason lies in the fact that Google Insights measures the propensity to search for a certain term which makes it plausible to believe that search for "nav" or similar would be constant unless any particular event disturbs the series. Thus in periods with increased unemployment more people would gear up their propensity to search for queries like "nav". Additionally, even though "dagpenger" might measure demand for unemployment benefits, or those searching for nav.no are not unemployed, there is likely a fairly constant ratio of those searching for the specific keywords that are unemployed or are in the state of becoming unemployed. Given such a ratio the keywords covaries with the number of unemployed and we are able to measure unemployment with search queries. Hence there is a link between the specified search queries and unemployment. This explanation justifies the internal validity of the analysis.

Secondly, now that we have reassured the internal validity, we ask ourselves if the result could be generalized over time. We know that the Google data is sensitive to the date of extraction as explained in chapter 5. All Google search data will vary somewhat since people's search intensity varies over time. This fact could have threatened the external validity of our results. If our results had been in the borderline of improving the forecast ability of the basic econometric models, the threat would have been more severe. However since the Google Indicators greatly improve the predictions the general result that search data contain useful information in forecasting short term unemployment figures in Norway should be valid and could be generalized across time.

7.3 Limitations

In this early phase of Google Insights for Search there are some general limitations to the data, in addition to some specific limitations to our research. First of all, by using web queries we do not capture the behavior of all firms or workers. Some people go directly to specific websites, like nav.no, some do not use the Internet at all for specific services and some firms might prefer traditional hiring methods. This implies that we do not include their behavior through search queries and it could create a bias in our data. However, as our results are

consistently showing improvements in the predictions we see this as a minor problem. This issue is also likely to diminish over time as there is a positive trend in people using the Internet for various tasks according to Statistics Norway (Appendix 9.4)

Moreover, there has been a general problem regarding manipulation of Google Trends data. Users may spam search engines with terms that will appear as rising trends. A famous example is the swastika that appeared on Google Trends as the hottest search during the summer of 2008 (Arrington 2008). However, Google has tried to fix this problem, as written in chapter 5, by eliminating repeated queries from a single user over a short period of time. We feel safe that this problem has not affected our own research, but it could appear as a problem in the future as businesses and researchers come to rely more on search data for decision-making such that there will be incentives for opposing parties to influence the data in their direction or make the data useless through generating false or misleading queries.

In addition to these general limitations to Google Insights for Search we would like to point to some other issues regarding our thesis in particular. Search data for Norway in the period 2004-2006 do not have the desired quality, as indicated throughout the paper. We lack some observations which we had to adjust for through our smoothing algorithm. Smoothing the data is not fortunate as it could make a leading series lagging since the smoothing procedure is based on prior values of the series itself. This could hurt the prediction power of the series. Additionally it makes our framework more complicated. However, we saw the smoothening as necessary due to some of the missing observations from 2004-2006 to capture the underlying trend. It is important to notice that from 2006 and onwards we do not experience this problem in the data which is positive for future research.

As written, Google's market share in Norway was 81% in 2008 according to ComScore, a marketing research company specialized in the Internet business. We do not have any other source on their market share and as such we could question ourselves if the data are representative. We do not know anything about the last 20% either. However, we do know that Google delivers search to other search

engines in Norway, as kvasir.no, and that their global market share is high so it is most likely that the data are representative.

Finally there are two issues regarding our research design. Firstly, we limited ourselves to model maximum 3 autoregressive components, 3 MA components and 2 Google Indicators. There might be better performing models of higher order than those identified, but our main goal was to investigate whether search queries have any prediction power. We were able to achieve our goal within the restrictions we put on ourselves. Secondly, it is important to emphasize that the keywords chosen and the Google Indicators constructed were solely selected by us in the end. People or firms might use other keywords when they search than the ones analyzed. These were not included because of low search volumes or that they simply did not come to our minds.

7.4 Implications and Future Research

Even though the Google Indicators contain noise both in form of capturing irrelevant search, not capturing all relevant search and lacking search volume in the start of the series they still outperform leading indicators used by Norwegian financial institutions when analyzing short-term changes in unemployment. This thesis barely touches upon the potential micro-behavior have in performing both forecasts and nowcasts.

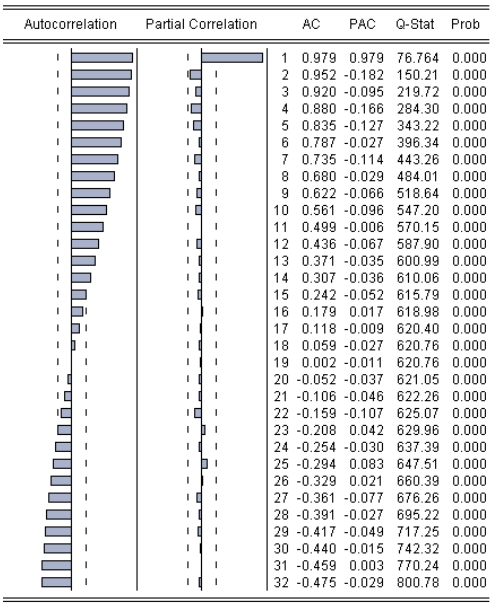
Increased stability in the Google data would make it unnecessary to smooth the time series and likely improve the predictive power in addition to make the framework even more simplistic. Larger data sets would also improve the prediction power, hence we believe to observe a reduction in prediction error as time evolves.

If and when a potential category filter arrives in Norway is not known, but this tool would radically improve the access of precise information without having to construct categories based on discussions with the Norwegian Language council, NAV and gut feeling. A possible route for further research would be to explore the predictive power of Google search in other areas of the economy. With access to the category filter it might even be possible to estimate larger variables such as GDP or interest rates.

Micro-behavior study is in its early stages, and we get the feeling of touching upon an area with enormous potential. Being able to grasp individuals' intentions, rather than being limited to stated preferences through surveys, the field of social science opens up a whole new area of study. The subject of micro-behavior is vastly understudied given its potential and we expect a lot of further research in the coming years as more data become available and the need for information in an ever changing globalized world is increasing. Despite the obvious shortcomings of our design, it offers additional support for using search queries, and hence intentions, to say something about the changes in the economy.

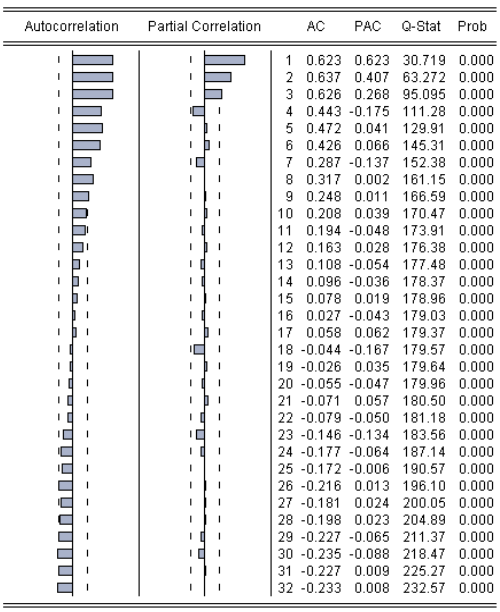
9. Appendix

Appendix 9.1



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.979	0.979	76.764	0.000
		2	0.952	-0.182	150.21	0.000
		3	0.920	-0.095	219.72	0.000
		4	0.880	-0.166	284.30	0.000
		5	0.835	-0.127	343.22	0.000
		6	0.787	-0.027	396.34	0.000
		7	0.735	-0.114	443.26	0.000
		8	0.680	-0.029	484.01	0.000
		9	0.622	-0.066	518.64	0.000
		10	0.561	-0.096	547.20	0.000
		11	0.499	-0.006	570.15	0.000
		12	0.436	-0.067	587.90	0.000
		13	0.371	-0.035	600.99	0.000
		14	0.307	-0.036	610.06	0.000
		15	0.242	-0.052	615.79	0.000
		16	0.179	0.017	618.98	0.000
		17	0.118	-0.009	620.40	0.000
		18	0.059	-0.027	620.76	0.000
		19	0.002	-0.011	620.76	0.000
		20	-0.052	-0.037	621.05	0.000
		21	-0.106	-0.046	622.26	0.000
		22	-0.159	-0.107	625.07	0.000
		23	-0.208	0.042	629.96	0.000
		24	-0.254	-0.030	637.39	0.000
		25	-0.294	0.083	647.51	0.000
		26	-0.329	0.021	660.39	0.000
		27	-0.361	-0.077	676.26	0.000
		28	-0.391	-0.027	695.22	0.000
		29	-0.417	-0.049	717.25	0.000
		30	-0.440	-0.015	742.32	0.000
		31	-0.459	0.003	770.24	0.000
		32	-0.475	-0.029	800.78	0.000

Appendix 9.2



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.623	0.623	30.719	0.000
		2	0.637	0.407	63.272	0.000
		3	0.626	0.268	95.095	0.000
		4	0.443	-0.175	111.28	0.000
		5	0.472	0.041	129.91	0.000
		6	0.426	0.066	145.31	0.000
		7	0.287	-0.137	152.38	0.000
		8	0.317	0.002	161.15	0.000
		9	0.248	0.011	166.59	0.000
		10	0.208	0.039	170.47	0.000
		11	0.194	-0.048	173.91	0.000
		12	0.163	0.028	176.38	0.000
		13	0.108	-0.054	177.48	0.000
		14	0.096	-0.036	178.37	0.000
		15	0.078	0.019	178.96	0.000
		16	0.027	-0.043	179.03	0.000
		17	0.058	0.062	179.37	0.000
		18	-0.044	-0.167	179.57	0.000
		19	-0.026	0.035	179.64	0.000
		20	-0.055	-0.047	179.96	0.000
		21	-0.071	0.057	180.50	0.000
		22	-0.079	-0.050	181.18	0.000
		23	-0.146	-0.134	183.56	0.000
		24	-0.177	-0.064	187.14	0.000
		25	-0.172	-0.006	190.57	0.000
		26	-0.216	0.013	196.10	0.000
		27	-0.181	0.024	200.05	0.000
		28	-0.198	0.023	204.89	0.000
		29	-0.227	-0.065	211.37	0.000
		30	-0.235	-0.088	218.47	0.000
		31	-0.227	0.009	225.27	0.000
		32	-0.233	0.008	232.57	0.000

Appendix 9.3

Let y^* denote the first difference. We then estimate the following regression:

$$y_t^* = \delta + \phi_1 y_{t-1}^* + \phi_2 y_{t-2}^* + \psi_1 GI2_t + u_t \qquad (9.3.1)$$

Computing the regression in Eviews provides us with the following results:

	Coefficient	Std. Error	t-Statistic	Prob.
C	-345.0133	717.1309	-0.481102	0.6320
D(C2SM,1)	232.8319	102.1713	2.278839	0.0258
AR(1)	0.357174	0.107864	3.311320	0.0015
AR(2)	0.433447	0.107734	4.023297	0.0001

R-squared	0.546822	Mean dependent var	-222.2973
Adjusted R-squared	0.527119	S.D. dependent var	1835.451
S.E. of regression	1262.173	Akaike info criterion	17.17229
Sum squared resid	1.10E+08	Schwarz criterion	17.29780
Log likelihood	-622.7887	Hannan-Quinn criter.	17.22231
F-statistic	27.75269	Durbin-Watson stat	2.203309
Prob(F-statistic)	0.000000		

We observe that the estimated coefficient for category 2 is significant at the 5% level and the r-squared value is 0.55. Using equation (4.21) we can set up the following equation for estimating the one period forecast for May 2010:

$$\phi(B)\Delta y_t = c\phi(B) + \beta\Delta GI_t\phi(B) + \varepsilon_t$$

$$(1 - \phi_1 B^1 - \phi_2 B^2)(y_t - y_{t-1}) = c(1 - \phi_1 B^1 - \phi_2 B^2) + \beta(GI_t - GI_{t-1})(1 - \phi_1 B^1 - \phi_2 B^2) + \varepsilon_t$$

$$y_t = c(1 - \phi_1 B^1 - \phi_2 B^2) + y_{t-1} + \phi_1(y_{t-1} - y_{t-2}) + \phi_2(y_{t-2} - y_{t-3}) + \beta(GI_t - GI_{t-1}) + \beta\phi_1(GI_{t-2} - GI_{t-1}) + \beta\phi_2(GI_{t-3} - GI_{t-2}) + \varepsilon_t$$

The expected value of the error term is equal to zero, hence inserting the values from the Eviews output gives us the following estimate:

$$y_t = -72.24 + 75\,269 + 0.36(75\,269 - 74\,191) + 0.43(74\,191 - 73\,757) + 233(87.47 - 85.04) + 233 * 0.36(79.63 - 85.04) + 233 * 0.43(76.14 - 79.63)$$

The estimation indicates that an increase in Google terms from period (*t*) to (*t-1*) embedded in category 2 increases the unemployment in period t, This is in line with our a priori expectation. This means that increased search for unemployment office related terms is positively related to an increase in unemployment. The relationship is also significant.

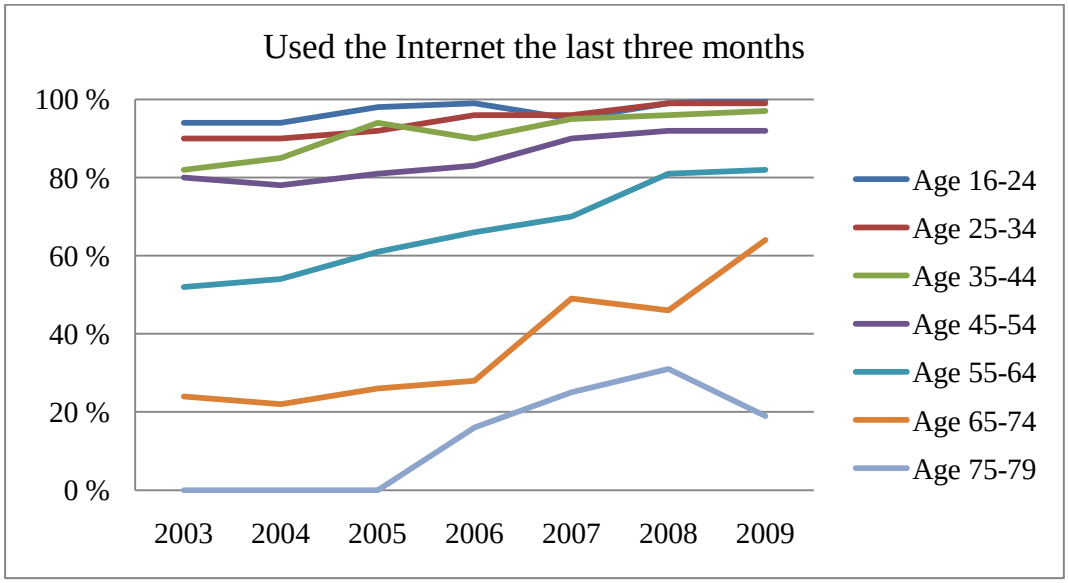
$$y_t = 75\,533{,}82$$

The forecasted level of unemployed individuals in May 2010 is 75 534 individuals while NAV reports the seasonally adjusted number to be 73 662 individuals. This gives us the following *root mean squared error* using equation (6.18):

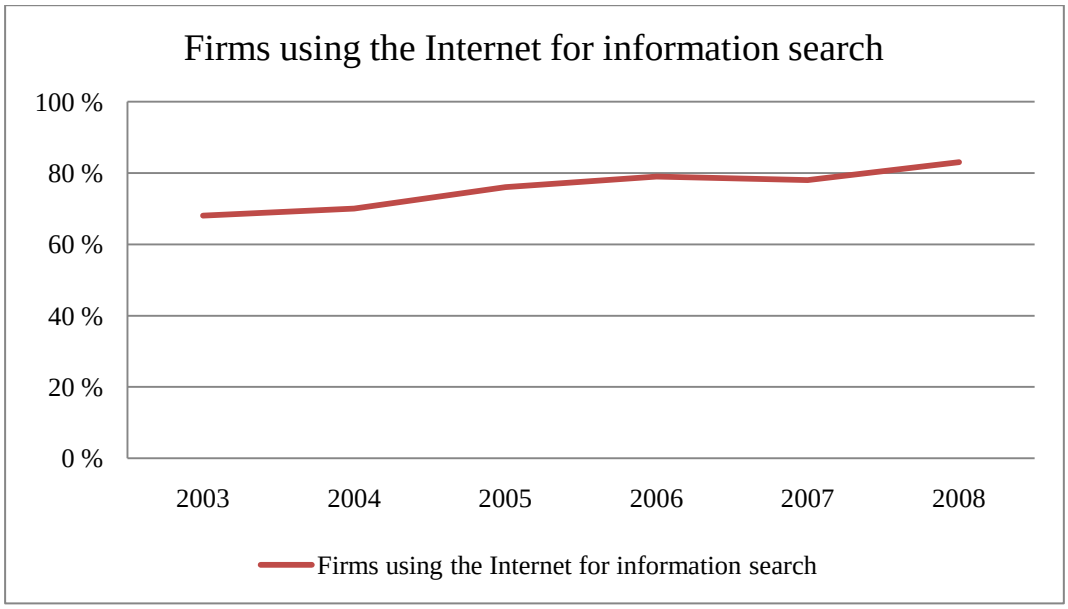
$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_t^2} = \sqrt{(75\,533.82 - 73\,662.44)^2} = 1\,871$$

The root mean square error is 1 871 individuals and our model overestimates the number of unemployed in May 2010 with this amount.

Appendix X.4



The Internet has apparently become an integrated part of Norwegians daily life according to the statistics (StatisticsNorway 2010)



There is a positive trend among Norwegian firms to use the Internet for information search (StatisticsNorway 2010)

10. References

Akerlof, George A. 1970. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics* 84:488-500.

Arrington, Michael. 2008. Swastika Appears On Google Trends. http://techcrunch.com/2008/07/10/swastika-appears-on-google-trends/.

Askitas, Nikolaos, and Klaus F. Zimmermann. 2009. Google Econometrics and Unemployment Forecasting. *IZA DP No. 4201*, http://ftp.iza.org/dp4201.pdf.

Bell, William R. 1993. Empirical Comparisons of Seasonal ARIMA and ARIMA Component (Structural) TIme Series Models. In *Bureau of the Census Statistical Research Division Report Series*. Washington: U.S. Bureau of the Census.

Box, George E. P., Gwilym M. Jenkins, and Gregory C. Reinsel. 2008. *Time Series Analysis: Forecasting and Control*. Hoboken, NJ: John Wiley & Sons, Inc.

Brooks, Chris. 2008. *Introductory Econometrics for Finance*. Second Edition ed. Camebridge: Cambridge University Press.

Burdett, K., and D. T. Mortensen. 1998. Wage differentials, employer size, and unemployment. *International Economic Review* 39 (2):257-273.

Cacchetti, Stephen, Rita S. Chu, and Charles Steindel. 2000. The unreliability of inflation indicators. *Current Issues in Economics and Finance* 6 (4).

Cahuc, Pierre, and André Zylberberg. 2004. *Labor economics*. Edited by A. Zylberberg. Cambridge, Mass.: MIT Press.

Choi, Hyunyoung, and Hal Varian. 2009. Predicting Initial Claims for Unemployment Benefits. http://research.google.com/archive/papers/initialclaimsUS.pdf.

Constant, Amelie F., and Klaus F. Zimmermann. 2008. Face to Face with the Financial Crisis: The U.S. Presidential Election from a Transnational Perspective. *Im Angesicht der Krise: US Präsidentschaftswahlen in transnationaler Sicht*, http://ideas.repec.org/a/diw/diwwrp/wr4-16.html.

D'Amuri, Francesco. 2009. Predicting unemployment in short samples with internet job search query data. http://mpra.ub.uni-muenchen.de/18403/.

D'Amuri, Francesco, and Juri Marcucci. 2009. "Google it!" Forecasting the US unemployment rate with a Google job search index. http://mpra.ub.uni-muenchen.de/18248/.

Diamond, P. A. 1971. Model of Price Adjustment. *Journal of Economic Theory* 3 (2):156-168.

Fountain, C. 2005. Finding a job in the Internet age. *Social Forces* 83 (3):1235-1262.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. http://research.google.com/archive/papers/detecting-influenza-epidemics.pdf.

Google. *Google Insights for Search*. Available from http://www.google.com/insights/search/#.

Google. 2010. Insights for Search Help. *Google Insights for Search*, http://www.google.com/support/insights/bin/topic.py?hl=en&topic=13973.

GoogleOperatingSystems. 2010. *Google's Market Share in Your Country* 2009 [cited 12. January 2010]. Available from http://googlesystem.blogspot.com/2009/03/googles-market-share-in-your-country.html.

Greene, William H. 2008. *Econometric Analysis*. Upper Saddle River, N.J.: Pearson Prentice Hall.

Gujarati, Damodar N. 2003. *Basic Econometrics*. New York: McGraw-Hill.

Helft, Miguel. *Google's New Tool Is Meant for Marketers*. NY Times 2008. Available from http://www.nytimes.com/2008/08/06/business/media/06adco.html?_r=3&ref=business.

Kuhn, Peter J. 2004. The Internet and Matching in Labor Markets. University of California, Santa Barbara.

LaViola, Joseph J. 2003. Double Exponential Smoothing: An Alternative to Kalman Filter-Based Predictive Tracking. In *The Eurographics Association*. Providence, RI: Brown University Technology Center.

Lucas, R. E. 1976. Econometric Policy Evaluation - Critique. *Carnegie-Rochester Conference Series on Public Policy* 1:19-46.

MacMinn, Richard D. 1980. Job Search and the Labor Dropout Problem Reconsidered. *The Quarterly Journal of Economics* 95 (1):69-87.

Makridakis, Spyros, Steven C. Wheelwright, and Rob J. Hyndman. 1998. *Forecasting: Methods and Applications*. 3rd ed. ed: John Wiley & Sons, Inc.

Mayer, Marissa. 2006. Yes, we are still all about search. *The Official Google Blog*, http://googleblog.blogspot.com/2006/05/yes-we-are-still-all-about-search.html.

McCall, J. J. 1970. Economics of Information and Job Search. *Quarterly Journal of Economics.* 84 (No. 1):113-126.

Meyler, Aidan, Geoff Kenny, and Terry Quinne. Forecasting Irish Inflation Using ARIMA Models.

Montgomery, Alan L., Victor Zarnowitz, Ruey S. Tsay, and George C. Tiao. 1998. Forecasting the U.S. Unemployment Rate. *Journal of the American Statistical Association* 93 (442):478-493.

NAV. The Norwegian Labor and Welfare Administration. http://www.nav.no/.

NAV. *Hovedtall om arbeidsmarkedet* 2010. Available from http://www.nav.no/Om+NAV/Tall+og+analyse/Arbeidsmarked/Statistikk/1073745853.cms.

Nosek, Donald. 2009. Google Insights for Search Predicts Kris Allen Next American Idol According to ymarketing's Latest Blog. *ymarketing | Digital.Search*.

Pedhazur, Elazar J., and Liora Pedhazur Schmelkin. 1991. *Measurment, Design, and Analysis. An Integrated Approach*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

Pindyck, Robert S., and Daniel L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts*: McGraw-Hill/Irwin.

Pissarides, Christopher A. 2000. *Equilibrium unemployment theory*. Cambridge,MA: MIT Press.

Rachev, Svetlozar T. 2007. *Financial econometrics: from basics to advanced modeling techniques*. Hoboken, N.J.: Wiley.

Rothschild, Michael. 1973. Models of Market Organization with Imperfect Information: A Survey. *The Journal of Political Economy* 81 (6):1283-1308.

Schmidt, Torsten, and Simeon Vosen. 2009. Forecasting Private Consumption: Survey-based Confidence Indicators vs. Google Trends. https://www.ciret.org/workshops/budapest/papers/Schmidt-Vosen.pdf.

Schmidt, Torsten, and Simeon Vosen. 2009. Forecasting Private Consumption: Survey-based Indicators vs. Google Trends. (#155), http://repec.rwi-essen.de/files/REP_09_155.pdf.

Spence, M. 1973. Job Market Signalling. *The Quarterly Journal of Economics* 87 (3):355-374.

StatisticsNorway. 2010. Arbeidskraftundersøkelsen. http://www.ssb.no/aku/.

StatisticsNorway. 2010. Norwegians' online behavior. http://statbank.ssb.no/statistikkbanken/.

Stevenson, Betsey. 2008. The Internet and Job Search. National Bureau of Economic Research, Inc.

Suhoy, Tanya. 2009. Query Indices and a 2008 Downturn: Israeli Data. *Discussion Paper No. 2009.06*, http://www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf.

Trends, Google. 2010. About Google Trends. http://www.google.com/intl/en/trends/about.html.

Varian, Hal, and Hyunyoung Choi. 2009. Predicting the Present with Google Trends. http://www.google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

Wilhelmsen, Bjørn Roger. 2010. *Personal Interview*.

Wu, Lynn, and Erik Brynjolfsson. 2009. The Future of Predictions: How Google Searches Foreshadow Housing Prices and Quantities. http://aisel.aisnet.org/icis2009/147/.

Center for Research in Economics and Management (CREAM)
Handelshøyskolen BI / Norwegian School of Management
0442 Oslo, Norway

The objective of CREAM is to provide research and analysis in the area of industrial economics and labor economics with applications to management, and provide research-based analysis for decision makers in public and private sector.

ISSN 1891-599X