Aalto University
School of Business

DO BIG DATA PREDICT CONSUMER CONFIDENCE?

An empirical study with the Finnish consumer confidence indicator and Google searches

Master's Thesis
Sami Torniainen
Aalto University School of Business
Economics
Spring 2020

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076
AALTO
www.aalto.fi

Abstract of master's thesis

Author Sami Torniainen	
Title of thesis Do Big Data Predict Consumer Confidence	
Degree Master of Science (Economics and Business Administration)	
Degree programme Economics	
Thesis advisor(s)	
Year of approval 2020 **Number of pages** 49	**Language** English

Abstract

There are 3.5 billion searches globally on Google every day. This thesis analyses whether Google search queries can be used to predict the present and the near future value of the consumer confidence indicator in Finland. This is interesting since the official statistics of consumer confidence are published with a reporting lag.

In order to assess the information contained in Google search queries, this study compares a simple predictive model of consumer confidence to a model that contains variables formed from Google data. When compared to a simple benchmark, Google search queries improve the prediction of the present by 5 % measured by mean absolute error. Moreover, the results show that current search activity provides useful information for the consumer confidence predictions up to six months ahead.

However, the predictive ability Google data for forecasting purposes appear to vary over time. When the consumer confidence fluctuates more suddenly, Google data improves the accuracy of nowcasts over the benchmark more than on the periods when the fluctuations are modest. More generally, the results of this thesis suggest that Google searches contain useful information of the present and the near future consumer confidence indicator in Finland.

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

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Tekijä Sami Torniainen		
Työn nimi Ennustaako Big Data Kuluttajien Luottamusta		
Tutkinto Kauppatieteiden maisteri		
Koulutusohjelma Taloustiede		
Työn ohjaaja(t)		
Hyväksymisvuosi 2020	**Sivumäärä** 49	**Kieli** Englanti

Tiivistelmä

Päivittäin tehdään 3.5 miljardia Google-hakua. Tässä opinnäytetyössä selvitetään voiko Google-hakuja käyttää nykyhetken ja lähitulevaisuuden kuluttajien luottamuksen ennustamiseen Suomessa. Tämä on kiinnostavaa sillä viralliset tiedot kuluttajan luottamuksesta julkaistaan viiveellä.

Google-hakujen ennustekykyä tutkitaan vertailemalla yksinkertaista kuluttajien luottamusta kuvaavaa mallia sellaiseen malliin mihin on lisätty Google-aineistosta muodostetut muuttujat. Yksinkertaiseen vertailukohtaiseen malliin nähden Google-haut tarkentavat nykyhetken ennustetta 5 % absoluuttisella keskivirheellä mitattuna. Lisäksi havaitaan että Google-hakujen lisääminen malliin parantaa keskimäärin kuusi kuukautta eteenpäin tehtyjä ennusteita.

Toisaalta Google-hakujen sisältämän informaation arvo vaikuttaa kuitenkin vaihtelevan ajanhetkestä riippuen. Kun kuluttajien luottamus vaihtelee äkillisemmin, Google-haut parantavat nykyhetken ennusteiden tarkkuutta vertailukohtaan nähden keskimäärin enemmän kuin ajanjaksoina, jolloin vaihtelut ovat vaatimattomia. Tämän tutkielman tulokset viittaavat yleisemmin siihen, että Google-haut sisältävät hyödyllistä tietoa nykyisestä ja lähitulevaisuuden kuluttajien luottamuksesta Suomessa.

Avainsanat big data, Google, Internet, nykyhetken ennustaminen, ennustaminen, kuluttajien luottamus, aikasarja-analyysi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

There are 3.5 billion Google searches are made every day[1]. Could the data from Google help to predict consumer confidence in Finland?

Consumer confidence is considered as one of the key indicators since it provides accurate information about consumer views with regard to the general economic situation and the financial situation of the consumers' own household. For example, if consumers have a low confidence in the economy, they will put a hold on their purchases of e.g. durable goods such as electronics and cars. If consumers start postponing their purchases for a year or longer, this can lead to disastrous economic consequences. Therefore, to avoid economic crunches, it is important for governments to manage consumer confidence.

Many institutions monitor consumer confidence and use it in decision making. A decreasing trend in the consumer confidence suggests that consumers have a negative view on the general economic situation as well as on their own financial situation. For example, manufacturers than may expect consumers to avoid big ticket purchases and they might bring down their inventories and postpone investments in new projects. Likewise, decreases in the consumer confidence may give a signal for governments to take fiscal or monetary actions to stimulate the economy. On the other hand, an increasing trend in consumer confidence indicates an improvement in the buying patterns of consumers. In this case, manufacturers can increase production and hiring. Furthermore, banks may expect an increasing demand for credit and construction companies might anticipate increasing investments in real estate.

Official consumer confidence statistics are published on a monthly basis, but with almost a one-month reporting lag. In order to make better policy decisions, especially during times of economic crisis, more timely estimates of consume confidence would be valuable. The search data from Google is available almost in real time basis. The real time information contained in the search data may be helpful in nowcasting and forecasting consumer confidence.

[1] Google inc.

In this thesis, we use simple autoregressive models to assess whether it is possible to predict the consumer confidence indicator published by Statistics Finland using Google searches. Our main model contains variables constructed from the data from Google Trends that are extended to the simple benchmark model. In order to determine the performance, we compare the results of our model over the benchmark model by properties such as information criteria, statistical significance, squared root and the forecasting accuracy in terms of mean absolute mean error.

The rest of the thesis is structured as follows. Section 1.1 reviews the relevant previous research on the topic of using Internet search data in forecasting. Chapter 2 contains an overview of the theory of consumer choice and consume confidence. Chapter 3 describes the data and research methodologies to answer the research question: do Google searches predict consumer confidence. Chapter 4 presents the results of the empirical part of this study. Chapter 5 discusses the results and limitations, and Chapter 6 concludes this thesis.

1.1 Previous Research

As far as we know, the first published paper that suggested that web search data was useful in forecasting economic indicators was Ettredge et al. (2005), which studied the U.S. unemployment rate. At the same time Cooper et al. (2005) described using Internet search volume for cancer-related topics. Since then there have been several papers that have examined web search data in various fields. For example, improvements in the prediction accuracy have been found for U.S. inflation (Guzman, 2011), the U.K. housing market (McLaren & Shanbhogue, 2011), Swedish private consumption (Lindberg, 2011), German unemployment (Askitas & Zimmermann, 2009) and U.S. private consumption (Vosen & Schmidt, 2011). In the field of epidemiology, Polgreen et al. (2008) and Ginsberg et al. (2009) showed that search data could help predict the incidence of influenza-like diseases.

In economics, Choi & Varian (2009a & 2009b) described how the search data from Google Trends can be used to predict econometric metrics including, for example, automotive sales, unemployment claims and vacation destinations. Choi & Varian (2012), which is an updated version of those working papers, shows how Google Trends predicts the survey-based Australian consumer confidence index. Their approach contains two main phases: variable selection and statistical testing.

To select the relevant Google searches, Choi & Varian (2012) use methods ranging from personal judgement to more advanced technical analysis. In case of the consumer confidence, they utilize a Bayesian method known as “spike and slab” regression in the variable selection that delivers three Google Trends categories: Crime & Justice, Hybrid & Alternative Vehicles and Trucks & SUVs. For the last two ones they find correlation with oil prices, but for the first one they have no explanation.

To statistically test the usefulness of Google Trends data in forecasting, Choi & Varian (2012) construct a simple autoregressive baseline model and augment that with the search data. These models are used to generate one-step-ahead predictions (nowcasts). After that, they assess the predictive ability of Google Trends data by comparing the mean absolute error between the baseline model and the model extended with the search data.

Across all examples, Choi & Varian (2012) report that the models extended with the Google Trends data tend to outperform the baseline models by 5 percent to 20 percent. For the Australian consumer confidence index, the one-step ahead mean absolute error goes from 3.63 percent to 3.29 percent, yielding an improvement of 9.3 percent. For the U.S. unemployment claims they report that the baseline model performs better than the model augmented with the Google Trend data, which is the opposite result than in the other examples. However, when looking only at the recession periods in the unemployment claim forecasts, they found that using the Google Trends data reduces the mean absolute error from 3.98 percent to 3.44 percent, an improvement of 13.6 percent. Based on this, they suggest that Google data might be useful in identifying the turning points of economic time series, which are hard to predict.

Niesert et al. (2018) examine the predictive ability of Google search data on unemployment, consumer price index and consume confidence for the U.S., U.K., Canada, Germany and Japan. Their paper utilizes the same out-of-sample nowcasting than used by Choi & Varian (2018). Niesert et al. report that Google searches provides useful information for unemployment predictions, but not for consumer price index or consumer confidence. They argue that online search behaviour is a relatively reliable gauge of an individual’s personal situation (employment status), but less reliable when it comes to variables that are unknown to the individual (consumer price index) or too general to be linked to specific search terms (consumer confidence).

Some papers take a somewhat different approach and construct the consumer confidence index entirely from Google search data. For example, Della Penna (2009) creates a consumer sentiment index for the U.S., which consists of highly correlated components with the Index of Consumer Sentiment from the University of Michigan and the Consumer Confidence Index from the Conference Board. According to Della Penna, the results are promising as among the three sentiment indices, the Google search-based index leads in time and predicts other indices.

Tuhkuri (2015) studies whether Google search queries can be used to predict the present and near future unemployment rate in the United States. Tuhkuri utilizes mainly the same methods than Choi & Varian (2012), but tests also for Granger causality and examines forecasts up to six months ahead in addition to nowcasts. Tuhkuri chooses 13 unemployment-related searches and constructs a single variable from those, called a Google Index. A seasonal autoregressive model is created for the baseline, which is augmented with the Google Index. These models are used to generate nowcasts and forecasts while the predictive accuracy of those are evaluated using a mean absolute percentage error.

Tuhkuri (2015) reports that the model augmented with Google data delivers 4.32 percent improvement in the nowcasts and 7.48 percent improvement in the two months ahead forecasts. The Google Index is also found to Granger-cause the U.S. unemployment rate. During the recession period in 2007-2009, the Trends data reportedly improves the predictive accuracy four times more than on average. This aligns with the results of Choi & Varian (2012) that suggest Google data may help to predict the turning points of an economic time series.

In Finland, ETLA[2] has been experimenting with big data in forecasting house prices (Widgrén, 2016) and the unemployment rate (Tuhkuri, 2014). According to Widgrén, Google searches improve the prediction of the present house price index by 7.5 percent measured by mean absolute error when compared to a benchmark model. For the unemployment rate, Tuhkuri (2014) reports that Google searches improve the prediction of the present by 10 percent measured by mean absolute error when compared to a benchmark model. In addition, both of these papers report that search queries improve the near future forecasts from two to six months ahead.

[2] ETLA - Research Institute of the Finnish Economy

This thesis utilizes the same methods than Choi & Varian (2012), Tuhkuri (2014 & 2015) and Widgrén (2016) but applies those to a different use case. We select the Google searches that potentially represent the Finnish consumer confidence indicator and statistically test the predictive ability of those.

2 THEORETICAL BACKGROUND

The theory of consumer choice examines the trade-offs that people face in their role as consumers (Mankiw, 2004). When a consumer buys more of one good, she can afford less of other goods. When she spends more time enjoying leisure and less time working, she has lower income and can afford less consumption. When she spends more of her income in the present and saves less of it, she must accept a lower level of consumption in the future. The theory of consumer choice examines how consumers facing these trade-offs make decisions and how they respond to changes in their environment.

The next section describes what consumer can afford, what she wants and what is her optimal choice.

2.1 Budget Constraint, Preferences and Optimal Choice

Suppose there are two goods, x_1 and x_2, that consumer can buy. With the two goods, the consumer's consumption bundle can be described as (x_1, x_2). The bundle tells how much the consumer is choosing to consume good 1, and how much the consumer is choosing to consume good 2 (Mankiw, 2004). Furthermore, the prices of the two goods are observed as p_1 and p_2 while the amount of money the consumer can afford to spend is m. The budget constraint of the consumer can be written as:

$$p_1x_1 + p_2x_2 \leq m \tag{1}$$

where $p_1x_1$is the amount of money the consumer is spending on good 1 and $p_2x_2$is the amount of money consumer is spending on good 2. The budget constraint of the consumer requires that the amount of money spenton the goods cannot be more than the total amount the consumer can afford to spend. The consumer's affordable consumption bundles are those that don't cost any more than m.

The consumer's choices, however, depend not only on her budget constraint but also on her preferences regarding the two goods (Mankiw, 2004). The consumer's preferences allow her to choose among different bundles of the goods. For example, if the consumer is offered two different bundles both of which she can afford, she chooses the bundle that best suits her tastes. If the two bundles suit her tastes equally well, it is said that the customer is indifferent between the bundles.

The indifference curve shows the various bundles of consumption that make the consumer equally happy (Mankiw, 2004). The slope at any point on an indifference curve equals the rate at which the consumer is willing to substitute one good for the other. This rate is called the marginal rate of substitution (MRS).

For example, in this case, the MRS measures how much good 1 the consumer requires to be compensated for a one-unit reduction in good 2 consumption. Because of the indifference curves are not straight lines, the MRS is not the same at all points on a given indifference curve. To summarize the above, marginal rate of substitution can be written as:

$$MRS = \frac{\Delta x_2}{\Delta x_1} \tag{2}$$

where the ratio of Δx_2 and Δx_1measures the rate in which the consumer is willing to substitute good 2 for good 1. (Varian, 2003)

Since we know how much the consumer can afford to spend (budget constraint) and what she wants to spend it on (preferences), we can consider consumer's decision about what to buy. The highest indifference curve that the consumer can reach is the one that touches her budget constraint (Mankiw, 2004). The point at which this indifference curve and the budget constraint touch is called the optimum. This is illustrated in Figure 1 as the optimal choice. The optimum represents the best bundle of good 1 and good 2 available to the consumer.

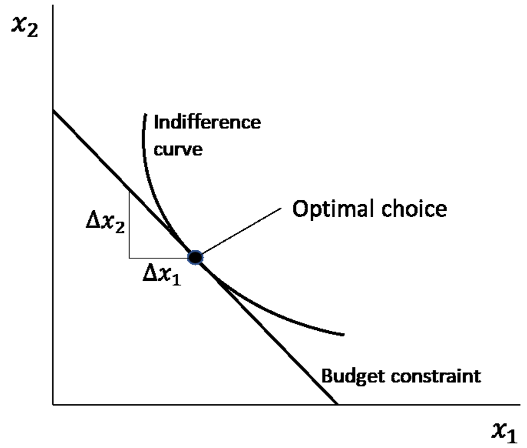

Figure 2.1. Budget constraint, indifference curve and the optimal choice.

At the optimum, the slope of the indifference curve equals the slope of the budget constraint (Varian, 2003). It can be said that the indifference curve is tangent to the budget constraint. The slope of the indifference curve is the marginal rate of substitution between good 1 and good 2, and the slope of the budget constraint is the relative price of good 1 and good 2.

Thus, the consumer chooses consumption of the two goods so that the marginal rate of substitution equals the relative price:

$$MRS = -\frac{p_1}{p_2} \qquad (3)$$

In the next section we will present how consumer's behaviour will change in response to changes in economic environment.

2.2 Income and Price Effect on Consumption

Let's suppose that income increases from m to m' while the prices stay constant. With higher income, the consumer can afford more of both goods. The increase in income, therefore, shifts the budget constraint outward. Because the relative price of the two goods has not changed, the slope of the new budget constraint is the same as the slope of the initial budget constraint (Mankiw, 2004). Thus, an increase in income leads to a parallel shift in the budget constraint as illustrated in Figure 2.2 (a).

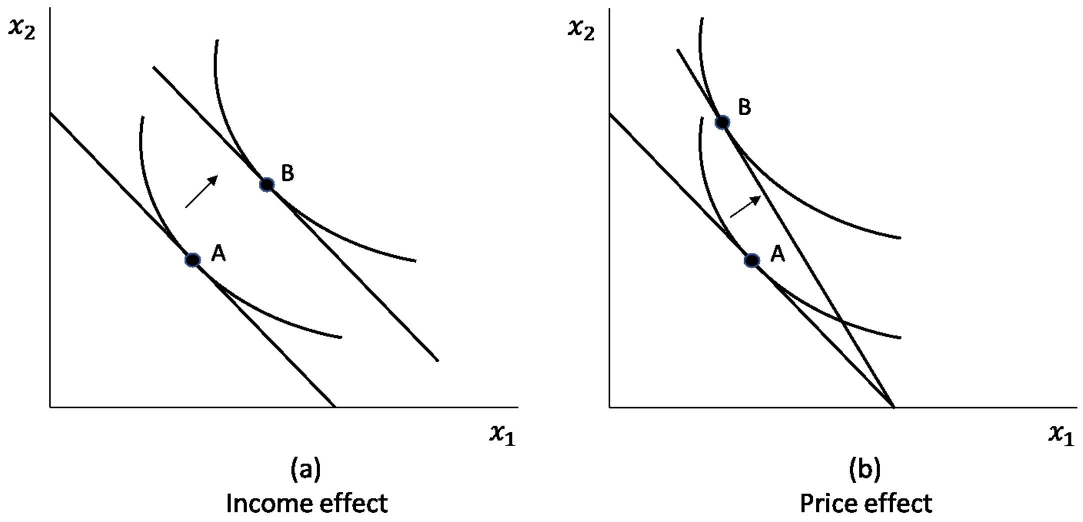

Figure 2.2. (a) Income effect and (b) price effect.

The expanded budget constraint allows the consumer to choose a better combination of good 1 and good 2, one that is on a higher indifference curve. Given the shift in the budget constraint and the consumer's preferences as represented by her indifference curves, the consumer's optimum moves from initial optimum (point A) to new optimum (point B).

The above described situation is most common. If a consumer wants more of a good when her income rises, the good in question is a normal good. However, this is not always the case. There can also be situations where consumption of a good decreases when income increases. Such goods are called inferior goods (Varian, 2003). For example, bus rides are inferior goods.

Let's now consider how a change in the price of one of the good alters the consumer's choices. For example, when the price of good 2 fall to half of its original price, it will expand the consumer's buying opportunities (Mankiw, 2004). The consumer moves from the initial optimum to the new optimum, which changes her purchases of both good 1 and good 2. In this case, as in Figure 2.2 (b), the quantity of good 2 consumed rise and the quantity of good 1 consumed fall.

In other words, fall in the price of any good shift budget constraint forward. However, now the outward shift in budget constraint changes its slope. This differs from the case of income effect in which the prices stayed the same.

In addition to the income and price changes, consumer's expectations of overall state of the economy and her personal financial situation can also have effect on the consumption decisions. These expectations can be described as consume confidence, which is presented in the next section.

2.3 Consumer Confidence

Consumer confidence is an economic indicator that measures the degree of optimism that consumers feel about the overall state of the economy and their personal financial situation. If the consumer has confidence in the immediate and near future economy and her personal finance, then the consumer tends to spend more than save.

In the economics literature, consumer confidence is considered originating from John Maynard Keynes (1936) who defined the term "animal spirits". Animal spirits is the description that Keynes gave to irrationality, uncertainty and confidence. He wrote the following example where he stressed the fundamental role of animal spirits in businessmen's calculations:

" Our basis of knowledge for estimating the yield ten years hence for a railway, a copper mine, a textile factory, the goodwill of a patent medicine, an Atlantic liner, a building in the City of London amounts to little and sometimes to nothing. If people are so uncertain, how are decisions made? They "can be taken as a result of animal spirits." They are the result of

"spontaneous urge to action." They are not, as rational economic theory would dictate, "the outcome of a weighted average quantitative benefits multiplied by quantitative probabilities."(Keynes, 1936).

The spontaneous and inconsistent element in the businessmen's behaviour, that Keynes refers as animal spirits, can be considered to reflect consumer confidence in a broader context.

Empirical measures of consumer confidence are primarily captured with surveys in which households give their expectations for income, unemployment and economic situation. The two best known surveys of consumer attitudes, the University of Michigan Index of Consumer Sentiment and the Conference Board Consumer Confidence Index, are widely tracked by policymakers, financial analysts, and journalists (Bram, 1998). For instance, stock markets react very quickly to the release of the confidence index, and in the same way, central bankers' decisions to modify the monetary policy is partly based on the evolution of this index (Beltran, 2008). Also, studies (Adams, 1965; Juster & Wachtel, 1972a and b) have found that consumer sentiment significantly affects expenditures on consumer durables such as motor vehicles, electronics, etc.

Although there remains some confusion about the forces driving the consumer confidence index, its role in business cycle analysis has been increasing over time. This fact is supported by the large consensus today over the explanatory power of this index in modelling consumption and income. In particular, the Beltran (2008) Jennings (1994) and Ludvigson (2004) have shown that the inclusion of the consumer confidence index improves the forecasts of real consumption, future labour income and non-stock market wealth growth. Furthermore, there is evidence that higher confidence levels could be related to future consumption growth if households are liquidity constrained so that greater income is closely tracked by greater consumption, or if some households follow a "rule-of-thumb" of consuming some fraction of their current income in every period (Campbell, 1989).

Recently, there has been a growing interest in the literature of understanding the parameters that may influence consume confidence. In the next section, we'll review the related empirical studies to get more insights of what causes the movements in consumer confidence.

2.4 Determinants of Consumer Confidence

The papers uncovering the determinants of consumer confidence have studied the effect of multiple factors including, for example, macroeconomics variables, stock market, commodity prices, political events, negative news coverage and household financial distress.

For example, Praet and Vuchelen (1989) show that increases in the price of oil depress the consumer confidence. The same occurs when the currency depreciates and interest rate rises. These movements could reflect fears of rising inflationary expectations. However, the effect is not always uniform. According to Praet and Vuchelen (1989), the increases in the value of the U.S. dollar affects consumer confidence in the U.S., Germany and France, but not in the U.K. and Italy. Furthermore, they find changes in the U.S. stock market index to affect positively German confidence, but negatively the sentiment in the U.K. and to have no effect in France and Italy. The stock market's influence on the consumer confidence in the U.S. is also confirmed by Beltran (2003).

However, according to Beltran (2003) the parameters affecting the confidence are changing over time. The variables of business cycle such as growth indicator and wages seem to be the main forces that have driven the confidence index during the eighties. Moreover, during this period the information from the stock price fluctuations overlaps the one contained in the growth indicator. The structural breakeven of this reportedly took place in the beginning of nineties.

Fuhrer (1993), Lovel (1975) and Throop (1992) show that economic variables such as the unemployment rate explain the consumer sentiment fairly well in times of usual political and economic activity. However, Throop (1992) points out that this tends to be reversed at times of an unusual economic or political events like the Persian Gulf War. During this type of events, the sentiment is affected by unusual factors and becomes detached from usual economic variables.

The paper by Mandal & McCollum (2013) study the short-term and long-run relationship between the unemployment rate and the consumer confidence index in five metropolitan statistical areas of New York State. They use panel cointegration and panel error-correction models, developed by Engle and Granger (1987), to explore the causal relationship between these two indicators. The results indicate a negative causality from unemployment to consumer sentiment and vice versa, indicating that unemployment and consumer sentiment reinforce each other in the short run. In the long-run, there is significant negative causality from consumer confidence to unemployment. However, the direction of causality from unemployment to consumer confidence is not significant.

Hollanders (2010) study the relationship between consumer confidence and economic news coverage in national newspapers in the Netherlands. The results show that Media-attention for economic developments is associated with consumer confidence with more negative news decreasing consumer confidence. However, the relationship appears to differ depending on the state of the business cycle. For example, the effect was particularly strong for the months following the beginning of the credit-crisis in 2007.

Ekici (2016) analyses the psychological and socio-economic determinants of consumer confidence. The results indicate that subjective financial distress, which measures e.g. how much stress consumers have about their current debt obligation and their capability to pay it off, might be useful to explaining consumer confidence. Higher distress is found to increase the probability of reporting negative sentiment. According to Ekici (2016), this suggests that there are also psychological factors in the formation of consumer confidence in addition to economic variables.

This section reviewed the previous research of the determinants of consumer confidence. As a summary, there are several factors that have effect on consumer confidence. These factors can affect consumer confidence simultaneously, but with different directions and various magnitudes. Furthermore, some factors have more effect on certain time periods, while the others have more effect on specific markets.

In next chapter we'll take a closer look on the data and research methodologies.

3 RESEARCH METHODOLOGY

This chapter describes the methods that we use in assessing whether Google data is helpful in predicting consumer confidence. The methods, which are the same than used by Choi & Varian (2012) and Tuhkuri (2014, 2015) and Widgrén (2016), can be split into two phases: variable selection and statistical testing. However, before introducing the methods, we will present the data being studied.

3.1 Data

3.1.1 Consumer Confidence

Statistics Finland publishes the official Consumer Confidence Indicator (CCI) on a monthly basis. Like in case of many other economic indicators, the statistical releases of CCI are available with reporting lag of nearly a one-month and therefore its current value is unknown.

CCI expresses consumers' views and expectations about the development of their own and Finland's general economic situation. The data is collected with the consumer survey, which contains 17 questions that ask about consumer's own economy now and in one year's time, Finland's economy in one year's time and consumer's spending money on major (durable) purchases within one year. The questions are harmonized among the consumer surveys fielded in the other EU countries. In most of the questions, respondents are presented with a choice of five alternative answer options. Once data collection is completed, the balance figure, which describes the respondents' average opinion, is calculated for each question. The arithmetic mean of the balance figures is called the consumer confidence indicator (CCI). The confidence indicator values can range between -100 and +100. The higher the figure, the brighter is the view on the economy.[3]

The consumer survey is carried out with a mixed-mode data collection method, i.e. with a web questionnaire and by telephone interviews. The rotating panel design is used in the survey: everybody answers twice within six months. Each month, the target is a random sample of 2,200 persons, of whom one half are included for the first time and one half for the second time. The target area is the whole country and the respondents of the survey

[3] Official Statistics of Finland (OSF): Consumer Confidence [e-publication].
ISSN=2669-8889. Helsinki: Statistics Finland [referred: 10.12.2019].
Access method: http://www.stat.fi/til/kbar/kas_en.html

represent the population aged 18 to 74 in Finland according to age, gender, area of residence and native language. On average more than 1,200 persons respond to the survey each month. Figure 3.1 shows the CCI from Jan 2004 to April 2019.

The data collection and content of the consumer survey changed in May 2019, which will also affect to the CCI time series from that date onwards. Therefore, in this paper we will examine the CCI time series prior the changes, ranging from January 2004 to April 2019 to provide more consistent results.[4]

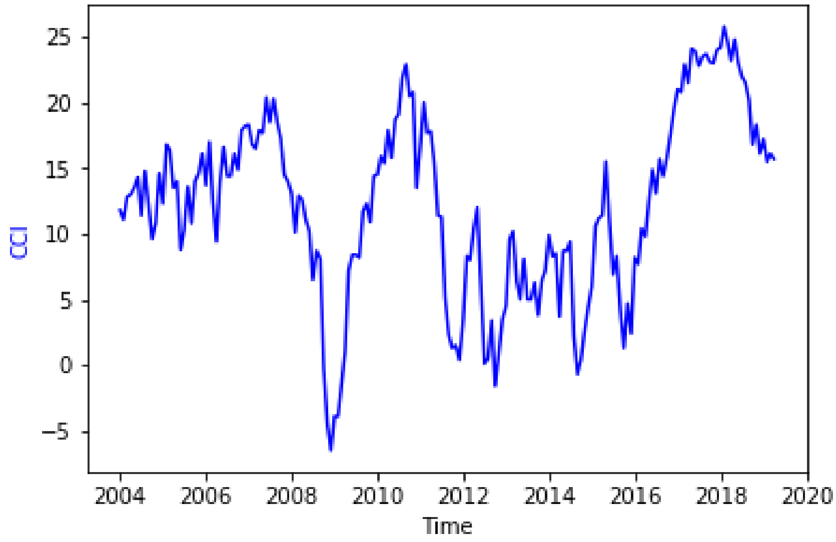

Figure 3.1. Consumer Confidence Indicator in Finland 2004-2019. Source: Statistics Finland.

3.1.2 Google

Google Trends provides a time series index of the volume of queries that users enter into Google in a given geographic area. The query index is based on a query share: the total query volume for the search term in question within a particular geographic region divided by the total number of queries in that region during the time period being examined (Choi & Varian, 2012). The maximum query share in the time period specified is normalised to be 100, and the query share at the initial date being examined is normalised to be zero. This is illustrated in equation (4):

[4] Official Statistics of Finland (OSF): Consumer Confidence [e-publication]. ISSN=2669-8889. Helsinki: Statistics Finland [referred: 10.12.2019]. Access method: http://www.stat.fi/til/kbar/kbar_2019-03-25_uut_001_en.html

$$I(K_t) = \left\{ \frac{\frac{K_t}{G_t}}{max\left(\frac{K}{G}\right)} \right\} * 100 \tag{4}$$

where K_t denotes the total query volume for the search term k and G_t denotes total number of queries in that region during the time period t. The data go back to January 1, 2004 and is available on weekly basis. In Finland, the data is published at the state level.

Google classifies search queries into approximately 30 categories at the top level and approximately 250 categories at the second level using a natural language classification engine (Choi & Varian, 2012). For example, the search term [car tire] would be assigned to category "Vehicle Tires" which is a subcategory of "Auto Parts" which is a subcategory of "Automotive". However, the assignment is probabilistic in the sense that a search term such as [apple] could be partially assigned to "Computers & Electronics", "Food & Drink" and "Entertainment".

There are many potential benefits related to Google data. From a forecasting point of view, one of the main benefits is the fact that Google searches are available almost in real time. In Google Trends, the search volumes can be obtained as weekly and monthly time series while the value for latest week is updated on daily basis. Because of this, the Google data is ahead of official statistics releases.

However, despite the benefits, there are also shortcomings associated with Google data. One of the main issues is that the Google searches are available only from 2004 onwards. This can be considered a relatively short period of time if comparing against other economic indicators. For example, when using Google data on a monthly level, there are only 184 observations during Jan 2004 – Apr 2019. Because of this, there might be limitations to its usefulness in economic forecasting.

It is not obvious which Google search categories or search terms would be most helpful in predicting consumer confidence. In the next section we will discuss on the variable selection in more detail.

3.2 Variable Selection

There are several methods described in the literature that can be used in variable selection. These include e.g. Bayesian model averaging, LASSO and Spike-and-Slab Regression, the latter of which is used by Choi & Varian (2012). However, in this paper we will use the consumer confidence determinants presented in Section 2.4 to identify the initial set of Google search volumes. After that, we examine the correlations between CCI and the search volumes to select the final set of variables into the model. We use both the Google search categories and the search terms as the model variables since not all consumer confidence drivers can be identified reliably at category level in Finland.

Using our personal judgement, we come up with four Google search categories and 13 Google search terms that were identified to relate with the consumer confidence determinants. Next, we will assess how the search categories and search terms correlate with the CCI. Figure 3.2 presents the calculated correlations between the CCI and the search categories and search terms. In Figure 3.2, negative correlation is represented by dark shades and positive correlation by lighter shades while the label “[C]” in the variable name indicates Google search category (the search terms have no labels assigned).

From Figure 3.2, we can observe that the search categories “Hybrid and Alternative Vehicles” and “Trucks and Suv” and search term trade-in cars (“vaihtoautot”) are top three with the highest positive correlation with the CCI. For the first two ones, this is not surprising since according to Choi & Varian (2012) they are correlated with the price of gasoline in the U.S., which is known to impact consumer confidence. Also Della Penna (2009) reports the link of the “Hybrid and Alternative Vehicles” category to public concern of energy costs.

Similarly, there are several Google search categories and search terms identified with unemployment and financial distress that are negatively correlated with the CCI like e.g. earnings-related daily allowance (“ansiosidonnainen päiväraha”), unemployment benefit (“työttömyyskorvaus”) and general unemployment fund (“ytk”). The first two are used by Tuhkuri (2014) in predicting the unemployment rate in Finland while, to the best of my knowledge, the latter one has not been used earlier in the literature.

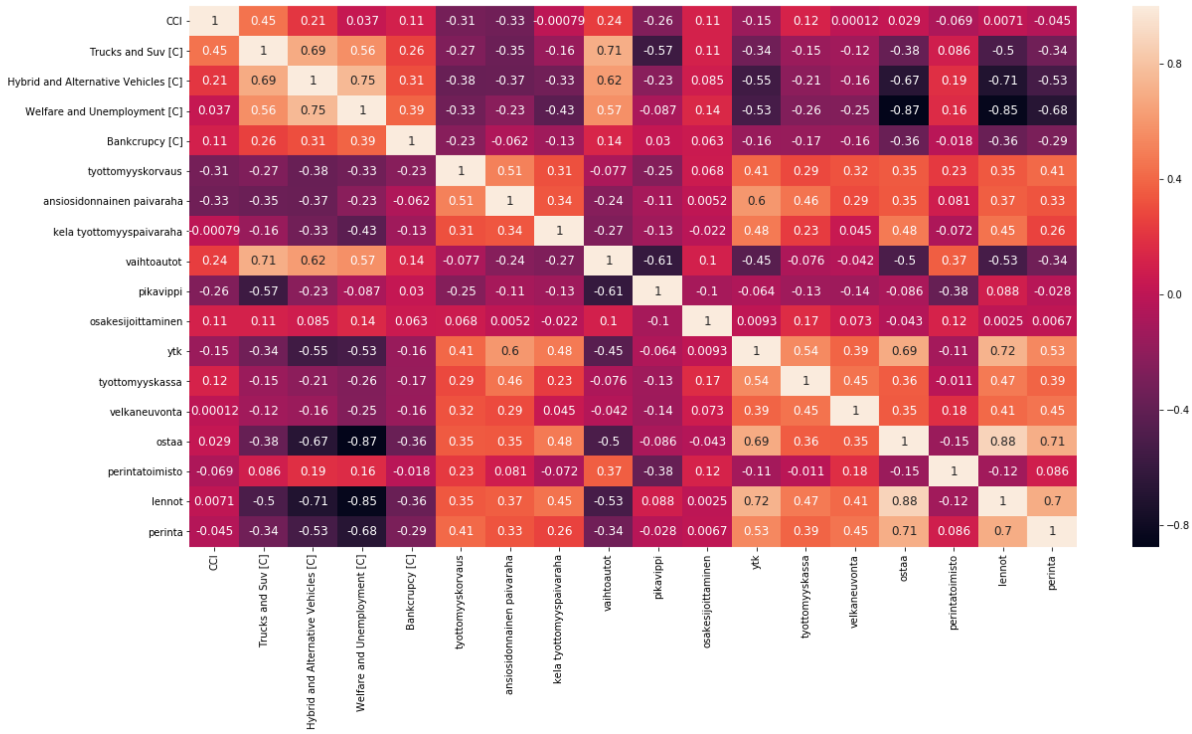

Figure 3.2. The correlation between CCI and Google searches.

Next, we will analyse if Google search volumes will anticipate the movements in the CCI. Table A.1 in Appendix 1 displays the correlations between different lags and leads of the CCI and the above-mentioned Google search categories and search terms. In Table A.1, the columns indicate lag orders ranging from -6 to 6 month while absolute values tell the correlations between the Google search volumes at the associated lag order and the present CCI. From the table, we can observe that the absolute values of the correlations with lag of -6 are more often higher than those with lag of 0. This implies that Google searches appear to anticipate the movements in the CCI. Bordino et al. (2012), Wu & Brynjolfsson (2009) and Tuhkuri (2014, 2015) observe a similar pattern in the stock market trading volumes, the housing market prices and unemployment rate, respectively.

Based on the correlations in Table A.1, we select the "Trucks & Suv" and "Hybrid & Alternative Vehicles" search categories for the predictors to the final model since those have the highest positive lagged correlations with the CCI: 0.505 with lag of -6 and 0.280 with lag of -6, respectively. Also, when estimating the model, we found that the coefficients of these two searches are statistically significant at 5% level whereas the coefficient of trade-in cars was not. In case of unemployment, we select only one search term or category as the predictor the for sake of simplicity. Our choice is unemployment benefit ("työttömyyskorvaus") since it has the highest negative lagged correlation with the CCI among the other unemployment related searches: -0.391 with lags -5 and -6.

Table 3.1 contains the descriptive statistics of CCI and the selected Google search volumes. The trends lines of CCI and the selected Google search volumes are presented in the figures 3.3, 3.4 and 3.5. For example, we can observe from Figure 3.3 that the CCI and unemployment series are clearly inversely correlated over the period 2004-2019. Decreasing CCI values seems to precede an increasing number of unemployment related searches, and vice versa.

Next, we will analyse the covariance-stationarity with the augmented Dickey-Fuller (1979) tests (ADF). Based on the results of the ADF, we cannot reject the null hypothesis of unit root for CCI nor the Google search volumes. Furthermore, we conduct the ADF test also against the first differences of the same series. The test rejects the null hypothesis of unit root for CCI and for two Google searches on 5% significance level but cannot reject for "Trucks & Suv". This indicates that the first differences appear to be stationary for all variables, except the last. Finally, we will analyse possible trend-stationary using KPSS-test (Kwiatkowski, 1992), the null hypothesis of which states that the time series is stationary around a deterministic trend. We cannot reject the null hypothesis for any of the series. Based on these results it would be preferable to use first differences of the series instead of levels.

However, for example a study by Stock (2001) argues that it is not necessary to use first differences in the forecasting model if the prediction horizon is relatively short compared to the whole period of the series. Indeed, this is the case in our study since we aim to predict 6 periods in advance while the total number of observations of the series is 184. Based on this, we decide to use levels of the variables instead of first differences. Also, the papers by Tuhkuri (2014, 2015) and Widgrén (2016) follow the same approach.

As the variable selection is now completed, next we will look at the model in more detail.

Table 3.1.The descriptive statistics of CCI and Google search volumes in 2004-2019.

Variable name	n	μ	σ	min	25%	50%	75%	max
CCI	184	12.2	7.0	-6.5	8.1	12.7	16.9	25.8
Trucks & Suv	184	64.7	15.4	42.0	50.8	62.0	78.0	100.0
Hybrid & Alternative Vehicles	184	30.8	11.3	17.0	22.0	28.5	36.0	100.0
Unemployment benefit	184	45.8	17.4	11.0	34.0	44.5	57.3	100.0

n = sample size, μ = mean, σ= standard deviation, min = smallest value, 25%=25th percentile, 50% = 50th percentile, 75%=75th percentile and max = largest value. Sample period Jan 2004 – Apr 2019.

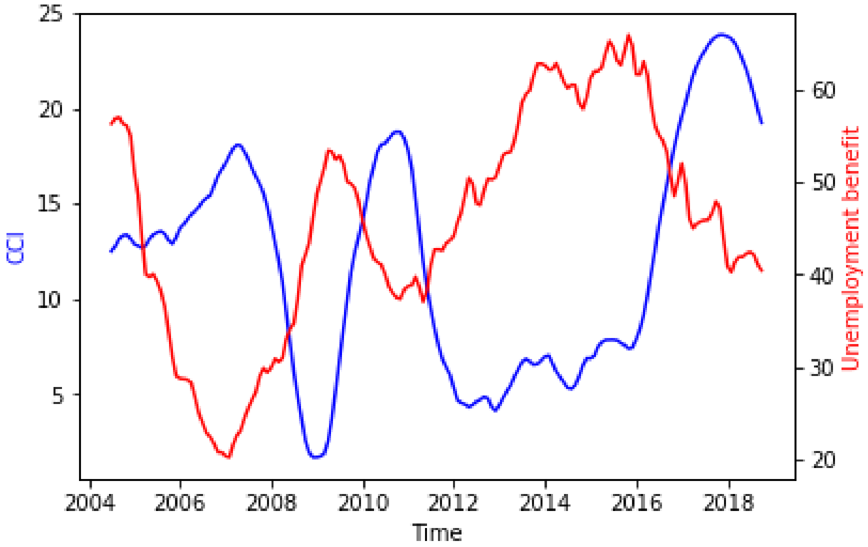

Figure 3.3. Trend lines of CCI and search term "unemployment benefit".

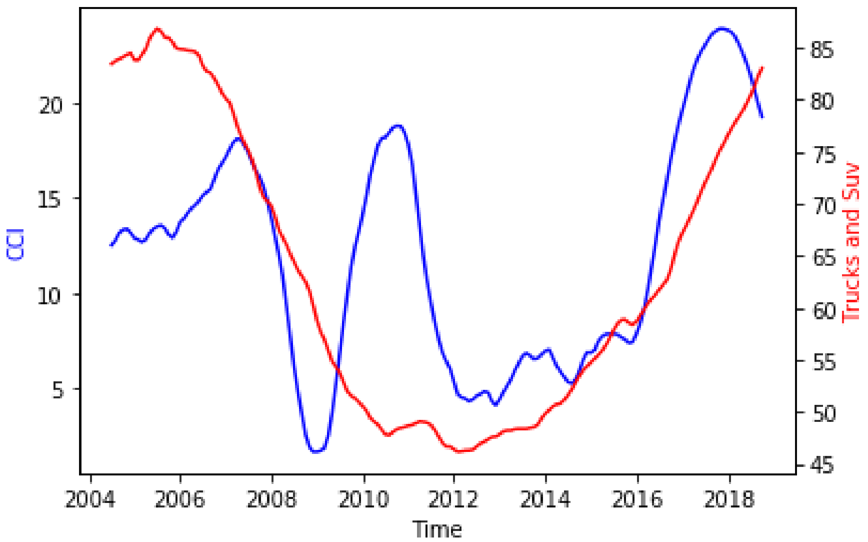

Figure 3.4. Trend lines of CCI and search category "Trucks and Suv".

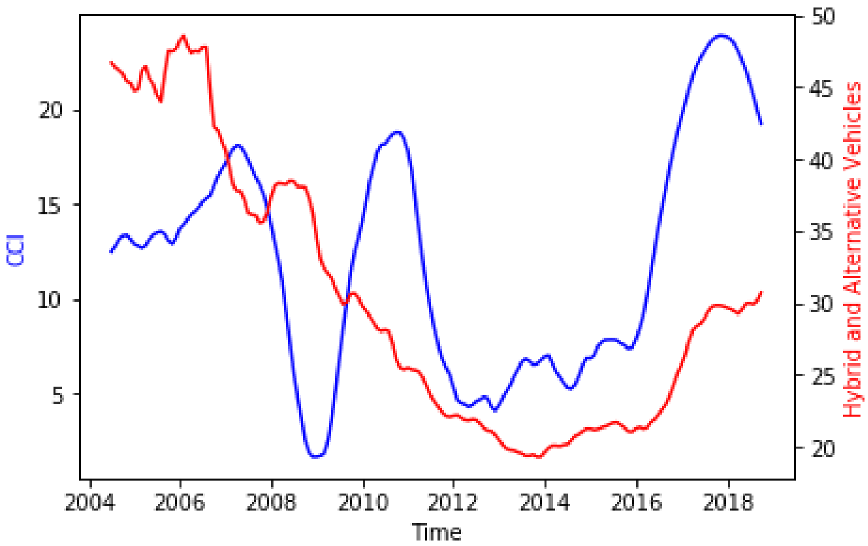

Figure 3.5. Trend lines of CCI and search category "Hybrid and Alternative Vehicles".

3.3 The Econometric Model

This section presents the econometric models that we use to analyse whether Google searches predict consumer confidence. More specifically, we are interested in finding out the incremental predictive ability of Google searches over the consumer confidence indicator.

The section is structured as follows. First, we construct a relevant benchmark model for the consumer confidence indicator. After that, the benchmark model is extended with the Google search variables, which were discussed in the previous section. Furthermore, the models and their forecast performance are compared. We follow the same approach as Choi & Varian (2012), Tuhkuri (2014, 2015) and Widgrén (2016) and use a simple model as the baseline: current month CCI is regressed on previous month(s) CCI value(s). This is known as an autoregressive (AR) model in literature. Also Clar et al. (2007) state that autoregressive models were performing best in most of the cases when forecasting consumer confidence.

We begin the specific benchmark model selection by analysing the autocorrelation function and the partial autocorrelation function of the CCI. These are presented in Figure 3.3 and Figure 3.4 respectively. The autocorrelation function appears to decrease slowly while the partial autocorrelation function is statistically significant only in the first lag. The visual analysis of the autocorrelation functions support the selection of a first order autoregressive model.

Next, we will evaluate the model against Akaike (AIC) and Schwarz (BIC) information criteria (Akaike 1973 & Schwarz 1978). The both criteria have minimum value when lag has value one, which also suggest the selection of AR(1) model. After that we analyse visually the autocorrelations of the model residuals, which are presented in Figure 3.5 and Figure 3.6. We don’t observe clear autocorrelation or squared autocorrelation at any lag which indicates that there isn’t any seasonality in the error term. In order to assess the residual seasonality more formally, we calculate the Ljung–Box (1978) portmanteau statistic for the residuals with lags up to K = 40. As an outcome of this, the null hypothesis of no serial correlation cannot be rejected at the 5% level, which also supports the above-mentioned argumentation of AR(1) model selection.

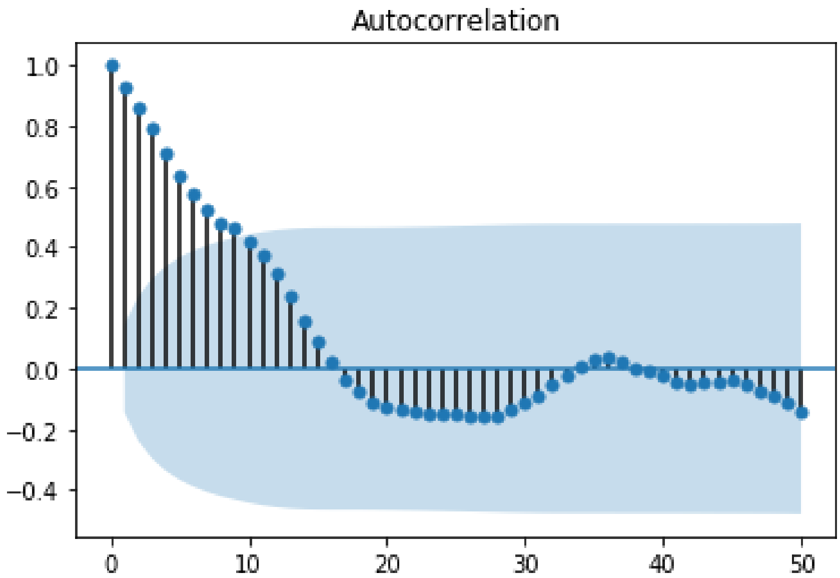

Figure 3.6. Autocorrelation function of CCI.

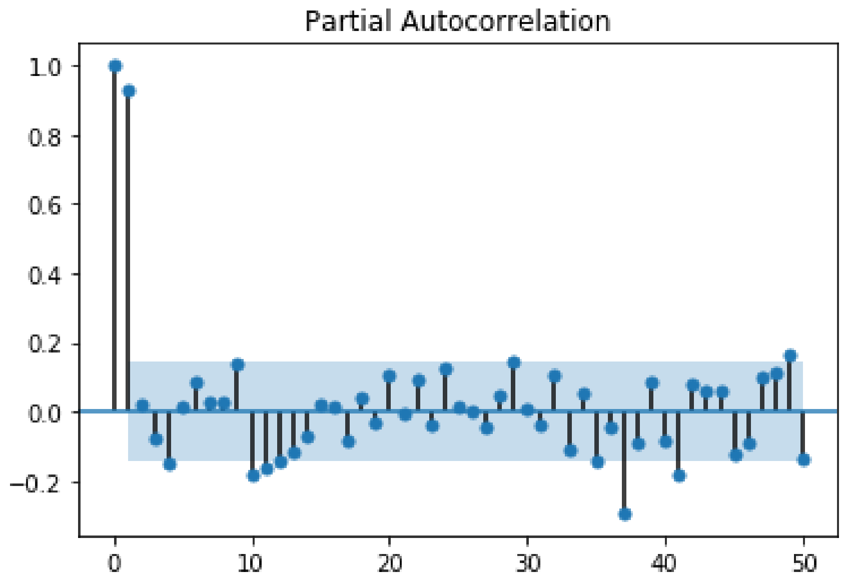

Figure 3.7. Partial autocorrelation function of CCI.

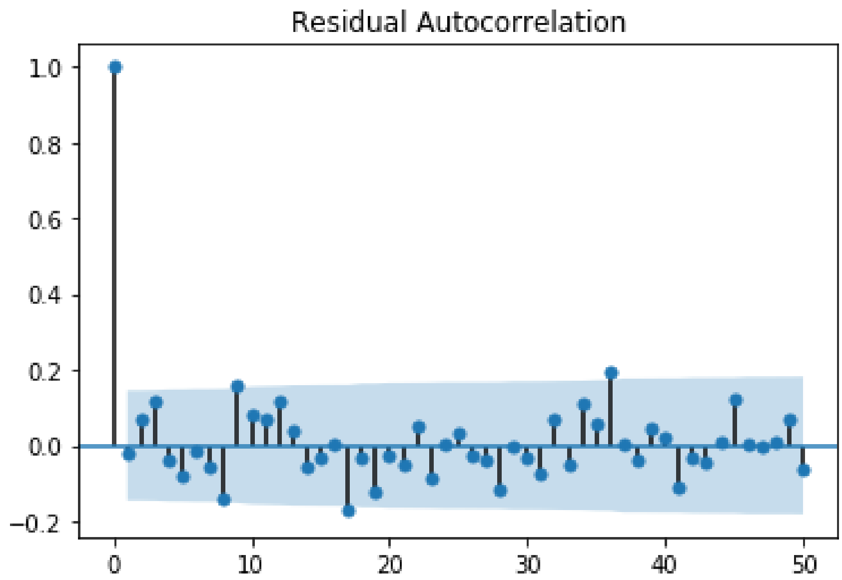

Figure 3.8. Autocorrelation function of residuals.

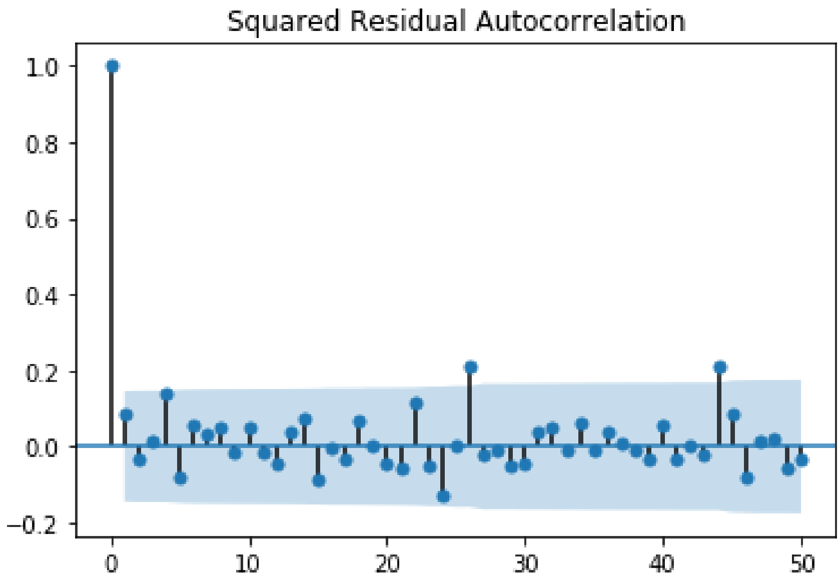

Figure 3.9. Squared autocorrelation function of residuals.

Based on the above findings we decide to use a first order autoregressive model as the benchmark model. Also, in order to minimize the effect of seasonality between Google predictors and CCI, we include the lagged 12-month CCI value as a predictor to the model. This means that current CCI value is predicted using the previous month's CCI value plus the 12-month lagged value. The same AR(1) model is also used by Choi & Varian (2012) and Tuhkuri (2015). The benchmark model is presented in equation (5). The model including Google searches is created by extending the benchmark model with the selected Google predictors. Since Google data is available almost in real time, we can use the present value of the Google predictors. Due to the real time availability, the Google data can provide timely signals of the changes in consumer confidence. The benchmark model extended with the Google predictors is presented in equation (6).

Model (0) $$y_t = y_{t-1} + y_{t-12} + \epsilon_t \tag{5}$$

Model (1) $$y_t = y_{t-1} + y_{t-12} + x_{1,t} + x_{2,t} + x_{3,t} + \epsilon_t \tag{6}$$

where y_t is CCI, x_1=Trucks & Suv, x_2=Hybrid & Alternative Vehicles and x_3=Unemployment benefit are the Google predictors and ϵ_t is error term.

The models are estimated with the maximum likelihood method. After estimation, we compare the fit of the models measured by properties such as information criteria, R^2, magnitude and statistical significance of the variable coefficients. However, our main focus is on the comparison of forecasting accuracy of the models that is presented in the next section.

To forecast the present, we use previous period CCI and lagged 12-month CCI value with current period's Google predictors values. To generate a series of one-step-ahead (pseudo) out-of-sample predictions, we use a rolling window of 48 months for both models. This means for each month from 2008, we train the model using past 48 observations, and then evaluate the out-of-sample predictions by comparing the forecasted values to the realized values of the CCI. The window size is set to 48 in order to have enough observations to train the models while also to include the time period of U.S. financial crisis (2008 - 2009), which had impact on consume confidence in Finland, into the predictions

In addition to forecasting the present, we will also predict the six months ahead to the future. In practice the method for predicting future is similar than in forecasting the present. The only difference is that available data is now one period older.

This section presented the models used for predicting consumer confidence. Next, we will describe how the forecasting accuracy can be measured.

3.4 Measuring Forecast Accuracy

To compare the accuracy of (pseudo) out-of-sample forecasts by the benchmark Model (0) and the extended Model (1) we use the mean absolute error (MAE). In specific, if the value of the error measure for forecasts computed from the extended model lies below error measure values of the benchmark model, we can conclude that Google searches might be useful in predicting the CCI. The mean absolute error is an average of the distance between predicted value and actual value as illustrated in equation (7):

$$MAE = \frac{1}{T}\sum_{i=1}^{T}|\hat{y}_i - y_i| \tag{7}$$

where $\hat{y}_i$ denotes predicted value and y_i denotes actual value.

Initially we planned to use mean absolute percentage error (MAPE) as a measure to compare the forecast accuracy since it is scale independent and appears to be more informative than MAE. However, the time series of the CCI was limiting the use of MAPE since there can be zero values, which can lead to division by zero and thus biased results. On the other hand, another candidate that we considered, the mean squared error (MSE), has the disadvantage of heavily weighting outliers. This is a result of the squaring of each term, which effectively weights large errors more heavily than small ones. This can be undesirable in our case since the CCI appear to fluctuate greatly in e.g. recession. Due to the above-mentioned pitfalls, we decided to choose MAE as the error measure.

Choi & Varian (2012), Tuhkuri (2014) and Widgren (2016) also use MAE for evaluating the predictive ability of Google searches while Tuhkuri (2015) utilizes MAPE because of its advantage of being scale-independent.

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

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

$$d_t = g\left(e_{1,t}\right) - g(e_{2,t}) \tag{8}$$

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

$$H_0: E(d_t) = 0 \tag{9}$$

against a two-sided alternative that the expectation is non-zero.

However, there are potential issues when using the Diebold-Mariano test in a (pseudo) out-of-sample environment. For example, the test comes with a potential power loss compared to full-sample alternatives, as we will notice in later in this thesis. For this reason, the results should be interpreted with caution. On the other hand, the Diebold-Mariano test is used by several studies (e.g. D'Amuri and Marcucci, 2012; Tuhkuri, 2014 & 2015; Widgrén, 2016) to describe if the incremental predictive ability from the internet search data is statistically significant.

This chapter presented the data and explained the methods for answering the research question. Next, we will look at the results.

4 RESULTS

This chapter presents the prediction results of the models described in the previous section. We summarize the performance of the models with and without the Google search volumes for forecasting the consumer confidence indicator in Finland using a pseudo out-of-sample forecast comparison methodology.

First we will analyse the performance of predicting the present (nowcast) and then forecasting near future (six months ahead). After that we will look at the performance in specific time periods.

4.1 Nowcasting

The estimation results of Models (0) and (1) are presented in Table 4.1. The coefficients for the two Google searches, Trucks & Suv and Hybrid & Alternative Vehicles, are statistically significant at 5% level. The positive sign of the coefficients means that the Google searches are positively connected to the CCI and the negative sign indicate the opposite. For example, in case of Trucks & Suv, the coefficient 0.0435 means that 1 percent increase in current search intensity is associated with 4.35 percent increase in the CCI. Similarly, in case of Hybrid & Alternative Vehicles, 1 percent increase current search intensity is associated with 5.88 percent decrease in the CCI.

These results are in line with Choi & Varian (2012) who reported also positive sign for the coefficient of Trucks & Suv and negative sign for the coefficient of Hybrid & Alternative Vehicles, but they found only the latter one as statistically significant at 5% level. The increase in oil prices might motivate the people to look more for hybrid and alternative vehicles thus increasing the related internet searches. On the other hand, the decrease in oil prices can increase the demand for Trucks & Suv type of vehicles and subsequently increase the related internet searches.

Based on R^2, which is 0.866 for Model (0), we can determine that benchmark model itself explains most of the variation in the CCI. When including the Google searches, the R^2 increases to 0.871 that can be considered a minor improvement. Similarly, extending the benchmark model (0) with the Google searches decreases the Akaike information criteria but increases Bayesian information criteria. Both criteria are based on various assumptions and asymptotic approximations, but in practice their main difference is the size of the penalty. BIC penalizes more complex models, which might explain our results since we add three Google searches to the benchmark model. The outcome of an F-test of joint significance of the Google variables is statistically significant at 1% level for Model (1). These results suggests that the Google searches offer useful information in explaining variation of the CCI within the estimation sample.

Figure 4.1 presents the fitted values for both models (0) and (1). As can be seen in the figure, both models seem to predict the CCI well. However, precise fitted values do not necessary indicate accurate forecasts. Next, let's assess the nowcasts. Results from one-step-ahead (pseudo) out-of-sample predictions using a rolling window of 48 months are illustrated in Figure 4.2. The mean absolute errors for nowcasts are presented in Table 4.2.

Table 4.1. The estimation results of the benchmark model (0) and the extended model (1).

Model	(0)	(1)
Variables		
y_{t-1}	0.9439**	0.9140**
	(0.030)	(0.034)
y_{t-12}	-0.0368	-0.0578
	(0.032)	(0.034)
$x_{1,t}$		0.0435*
		(0.020)
$x_{2,t}$		-0.0588*
		(0.027)
$x_{3,t}$		-0.0206
		(0.015)
Constant	1.1213*	1.6461
	(0.460)	(1.346)
Summary		
R^2	0.866	0.871
F-statistic	546.5	224.8
Prob (F-statistic)	1.65e-74	5.82e-72
AIC	826.5	825.7
BIC	836.0	844.5
n	184	184

y = CCI, x_1=Trucks & Suv, x_2=Hybrid & Alternative Vehicles, x_3=Unemployment benefit.
* and ** denote statistical significance at 5% and 1% levels. The standard errors of the estimated coefficients are given in parentheses. The sample period is Jan 2004 – Apr 2019.

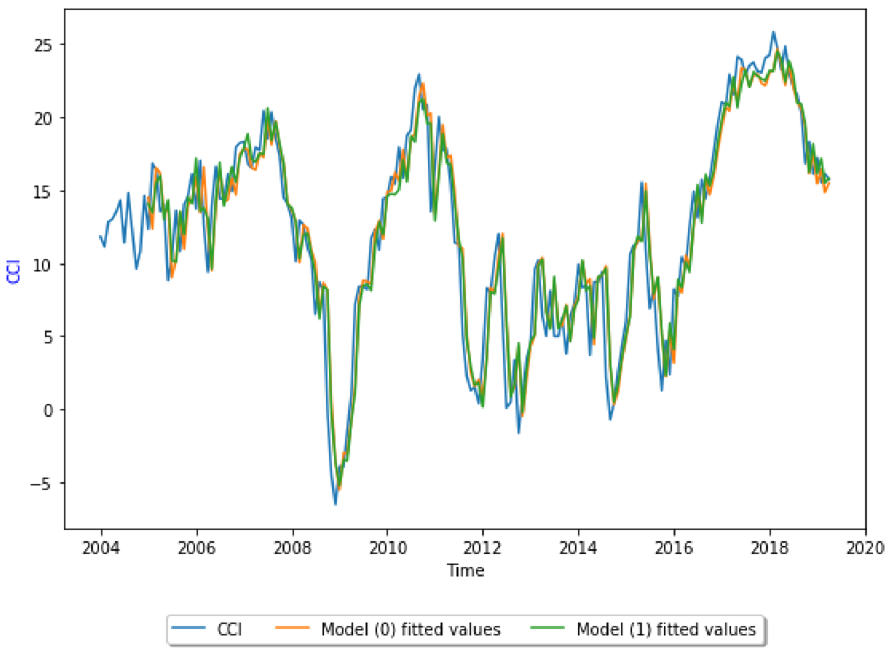

Figure 4.1.CCI with the fitted values for the benchmark model (0) and extended model (1).

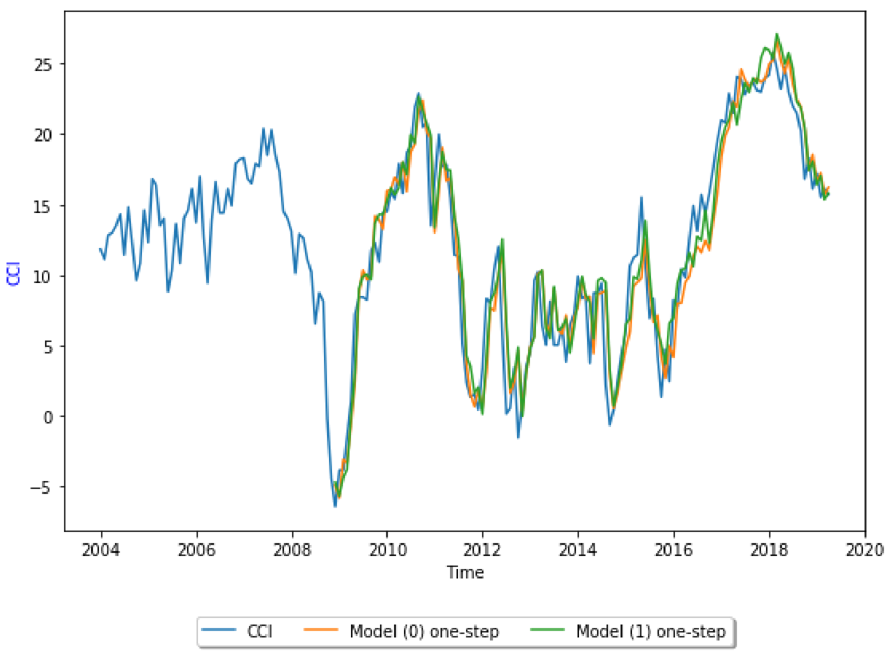

Figure 4.2.CCI with the one-step-ahead nowcasts for the benchmark model (0) and the extended model (1).

Table 4.2. Nowcasting (one-step-ahead) accuracy of the benchmark model (0) and the extended model (1).

Model	MAE	Δ
(0)	2.03	-5.04%
(1)	1.92	

MAE = mean absolute percentage error,
Δ = improvement in forecasting accuracy. The evaluation period is Dec 2008 – Apr 2019. Rolling window of 48 months is employed.

The mean absolute error for forecasts computed from Model (0) without Google data is 2.03 and for Model (1) with Google data is 1.92. These figures suggest that using Model (1) yields a 5.04 percent improvement in nowcasting the CCI. Based on this, we can say that Google searches help to predict consumer confidence when compared to the benchmark.

However, the Diebold-Mariano test for predictive accuracy results in no statistical significance at the 10% level on the difference between the forecasts of model (0) and (1). This might be caused by the fact that the observation period is relatively short (184 months) and the power of the test is low, as noted by Tuhkuri (2015). According to Diebold (2015) the low power combined with finite samples may result in the test failing to reject the null hypothesis even if the alternative were true.

4.2 Forecasting Near Future

Can Google searches also predict near future? Table 4.3 summarizes the mean absolute errors of (pseudo) out-of-sample forecasts up to the horizon $h = 6$. In general, we can observe that increasing the horizon decreases the forecasting accuracy of both models, which is logical. However, in case of the model extended with Google searches, the forecasting accuracy stays nearly on the same level in between horizons $h = 4$ and $h = 6$. Actually, the accuracy seems to increase a bit when moving from $h = 4$ towards the end.

When comparing the models, it can be seen that the forecasts computed from the model with Google data (1) outperform the benchmark model (0) in each step. As the horizon increases the difference between models' performance spread wider in favour the for the model with Google searches although, the difference somewhat varies between months. For example, the two-step ahead forecasts improve 3.63 percent on average when we add Google data, compared to 5.04 percent improvement in the one-step ahead forecasts (nowcast). Towards the end of the horizon at h = 6 the model with Google searches generates 12.43 percent better results than the benchmark model.

The Diebold-Mariano test delivers no statistically significant differences at 10% level in the forecasts between horizons h = 1 and h = 3. From then on, however, the test reports are statistically significant at 5% level in horizons h = 4 and h = 5, and at 1% level at h = 6. The results indicate that Google data might help forecasting CCI multiple steps ahead. This is in line with the cross-correlation analysis presented in Section 3.2 that suggest that the correlation is strongest between the current searches and the CCI six months ahead.

Figure 4.3 illustrates the series of (pseudo) three-steps-ahead out-of-sample plotted with the CCI. From the figure, Google searches appear to provide an early signal (marked with red circle) before of the turning point in the CCI in October 2010 ahead of the decrease caused by emerging negative news of the Greek debt issues.

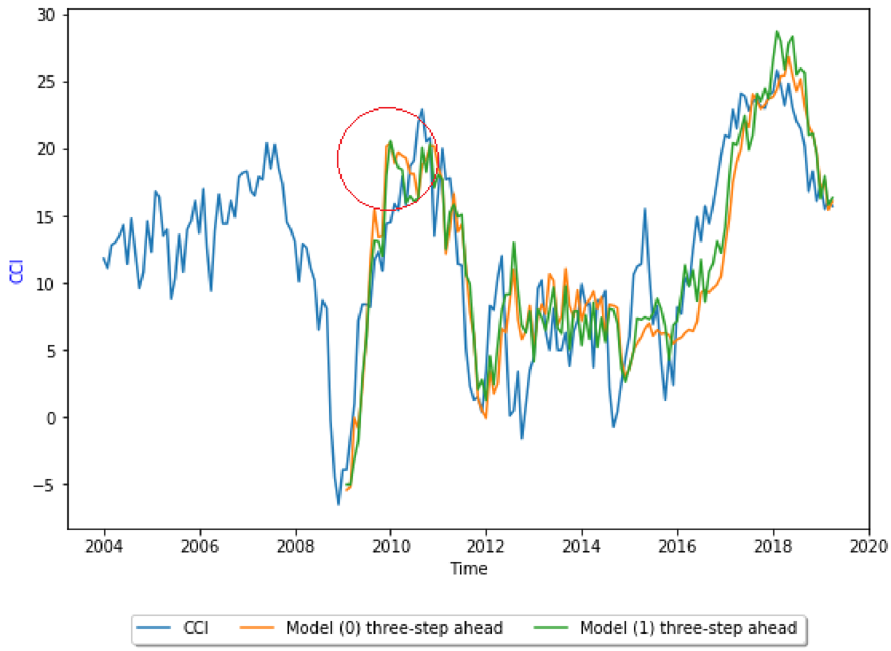

Figure 4.3. CCI with the two-steps-ahead forecasts for the benchmark model (0) and extended model (1).

Table 4.3. Forecasting accuracy of the benchmark model (0) and the extended model (1).

Horizon	Model	MAE	Δ
h=0	(0)	2.03	-5.04%
	(1)	1.92	
h=1	(0)	2.73	-3.63%
	(1)	2.62	
h=2	(0)	3.46	-6.58%
	(1)	3.24	
h=3	(0)	3.89	-6.13%
	(1)	3.65	
h=4	(0)	4.24	-7.72%*
	(1)	3.91	
h=5	(0)	4.34	-11.45%*
	(1)	3.84	
h=6	(0)	4.37	-12.43**
	(1)	3.83	

MAE = mean absolute percentage error,
Δ = improvement in forecasting accuracy,
* and ** denote statistical significance at 5% and 1% levels.
The evaluation period is Dec 2008 – Apr 2019. Rolling window of 48 months is employed.

Similarly to Table 4.3, we can assess the predictive ability of Google data by extending the benchmark model with lagged Google search volumes and estimating the model fit. We create a new model (2) that contain the Google variables with lags 1 and 2 in addition to the present values of those and run estimation. The results are shown in Table 4.4. Introducing the lagged variables increase the both R^2 and Akaike information criteria when compared against Models (0) and (1). The increase in R^2 indicates an improved regression fit but the increase in Akaike information criteria suggest that the models (0) and (1) would perform better than Model (2). AIC penalizes because we added six new variables into the model. However, the outcome of an F-test of joint significance of the Google variables is statistically significant at 1% level for Model (2). Thus, these results support the idea that the current Internet searches are likely to offer information on the future CCI.

To summarize the results of this section, it appears that Google data can be useful when forecasting coming months' CCI, not only on the present. Although the accuracy of both models decrease as the horizon increases, but the accuracy of the model with Google searches decrease less than in the benchmark model.

In the next section, we will analyse the prediction performance in specific time periods.

Table 4.4. The estimation results of the benchmark model (0) and extended models (1) and (2) that include current and lagged Google searches in 2004–2019.

Model	(0)	(1)	(2)
Variables			
y_{t-1}	0.9439**	0.9140**	0.9034**
	(0.030)	(0.034)	(0.035)
y_{t-12}	-0.0368	-0.0578	-0.0753*
	(0.032)	(0.034)	(0.036)
$x_{1,t}$		0.0435*	0.039
		(0.020)	(0.039)
$x_{2,t}$		-0.0588*	-0.0628*
		(0.027)	(0.029)
$x_{3,t}$		-0.0206	-0.0187
		(0.015)	(0.016)
$x_{1,t-1}$			0.0388
			(0.045)
$x_{2,t-1}$			-0.0525
			(0.030)
$x_{3,t-1}$			-0.0110
			(0.017)
$x_{1,t-2}$			0.0101
			(0.040)
$x_{2,t-2}$			0.0279
			(0.029)
$x_{3,t-2}$			-0.0082
			(0.016)
Constant	1.1213*	1.6461	2.5483
	(0.460)	(1.346)	(1.810)
Summary			
R^2	0.866	0.871	0.876
F-statistic	546.5	224.8	102.4
Prob (F-statistic)	1.65e-74	5.82e-72	1.75e-66
AIC	826.5	825.7	831.9
BIC	836.0	844.5	869.6
n	184	184	182

y = CCI, x_1=Trucks & SUV, x_2=Hybrid & Alternative Vehicles, x_3=Unemployment benefit.
* and ** denote statistical significance at 5% and 1% levels. The standard errors of the estimated coefficients are given in parentheses. The sample period is Jan 2004 – Apr 2019.

4.3 Forecasting Over Time

Since the CCI has fluctuated greatly over the last 10 years, the recent economic history provides a good opportunity to test if the Google data will help to predict sudden changes in consumer confidence. As Figure 4.4 illustrates, the CCI in Finland reached the lowest point since 2004 in winter 2008-2009 due to increased oil prices and emerging financial crisis in U.S. However, this was only a short-term decline and the confidence bounced back quickly and reached the pre-crisis level in summer 2010, even though the Finnish economy had not recovered entirely. Soon after this, in autumn 2011, the confidence decreased again because of the negative news of the Greek debt issues. This was followed by a short-term increase before the CCI went back to negative towards end of 2012 due to growing pessimism caused by the debt problems in other eurozone countries like Italy and Spain.[5]

Most of the professional forecasts fail to identify so called "turning points" in the economic cycle like e.g. recession (Tuhkuri, 2015). However, for example Choi & Varian (2012) reckon internet search data could help to overcome this shortcoming. Next, we will test this with the Finnish CCI. Table 4.5 presents the mean absolute errors of the forecasts up to h = 6 from January 2009 until December 2012. The period is selected because the CCI has six turning points during this time that we want to include in the analysis. In the one-step-ahead forecasts, we can observe that the mean absolute error reduces from 2.12 to 1.99 when including the Google data, which is a 5.89 percent improvement in prediction accuracy. In the two-steps-ahead (pseudo) out-of-sample forecasts, there is only a minor 0.12 percent improvement, but at the three-steps-ahead horizon, there is a gain of 2.93 percent. However, after that the benchmark model performs better towards later horizons.

The Diebold-Mariano test results no statistically significant differences between the forecasts at the 10% level. As mentioned earlier, a possible reason for this is a rather short time series available from Google data.

[5] Hyvinvointikatsaus 1/2013 – Teema: Eurooppalaisten elinolot ja talouskriisi, Statistics Finland.

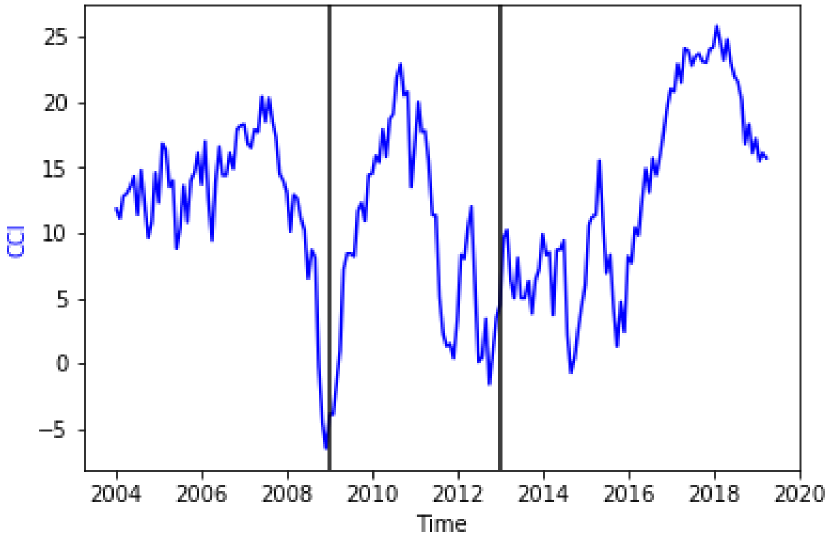

Figure 4.4. The fluctuations in the CCI in 2009-2012.

Table 4.5. Turning points. Forecasting accuracy of the benchmark model (0) and the extended model (1).

Horizon	Model	MAE	Δ
h=0	(0)	2.12	-5.89%
	(1)	1.99	
h=1	(0)	2.94	-0.12%
	(1)	2.93	
h=2	(0)	3.88	-2.93%
	(1)	3.77	
h=3	(0)	4.41	0.60%
	(1)	4.44	
h=4	(0)	4.62	0.55%
	(1)	4.64	
h=5	(0)	4.32	8.30%
	(1)	4.68	
h=6	(0)	4.08	8.47%
	(1)	4.42	

MAE = mean absolute percentage error,
Δ = improvement in forecasting accuracy,
* and ** denote statistical significance at 5% and 1% levels.
The evaluation period is Jan 2009 – Dec 2012. Rolling window of 48 months is employed.

On the other hand, when comparing the forecasts in Table 4.5 to those in Table 4.3 in the previous section, we can see that the improvement in one-step ahead nowcasts with Google data is even greater during the economic turning points. This observation suggests that Google search queries tend to improve the prediction accuracy, especially during the sudden economic changes. However, the models using Google data improve predictions only until h = 2 while the both models also generate less accurate predictions in terms of mean absolute error in 2009-2012 than on average. The results are in line with Tuhkuri (2015) that reports gain in the prediction accuracy up to three-steps-ahead horizon during the recession period when using Google data.

Next, let's analyse when the Google data is particularly useful in predicting CCI. Figures 4.5 and 4.6 present the results of one-step-ahead and two-steps-ahead forecasts using a rolling window of 48 months. Figure 4.5 illustrates the difference in one-step-ahead forecast errors for the baseline model and the extended model with the Google searches for each month. The difference is positive when the model with the Google searches produces more accurate predictions and negative when the benchmark is more accurate. Similarly, Figure 4.6 illustrates the difference in two-steps-ahead foresting errors for the baseline model and the extended model with the Google searches for each month.

From the figures, the main observation is that the model extended with Google searches (1) generates more accurate forecasts when the changes in the CCI are steeper i.e. the confidence is changing more suddenly. For example, this is shown in Figure 4.5 when the CCI climbs up from value 1.3 to 25.8 between September 2015 - February 2018. During this period, the difference in forecasts errors between Models (0) and Model (1) is more often positive than negative. On the other hand, the benchmark model (0) seem to perform better the when fluctuations are modest.

To summarize the findings of this section, it appears that the predictive ability of Google data varies over time. Google searches improves the accuracy of nowcasts during the 2009–2012 economic turning points in Finland. However, the observation period is short which might limit its ability to answer when Google searches are especially useful.

The next chapter discusses the results, limitations and future research in more detail.

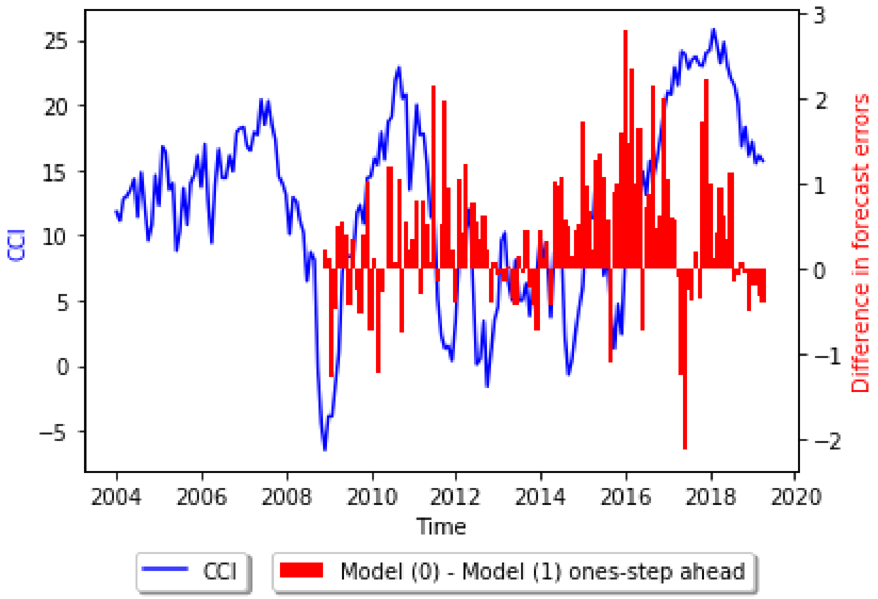

Figure 4.5. The model performance measured by the difference in absolute mean errors for one-step-ahead nowcasts of the benchmark model (0) and the extended. The positive vertical bar indicates that the extended model performs better, and vice versa.

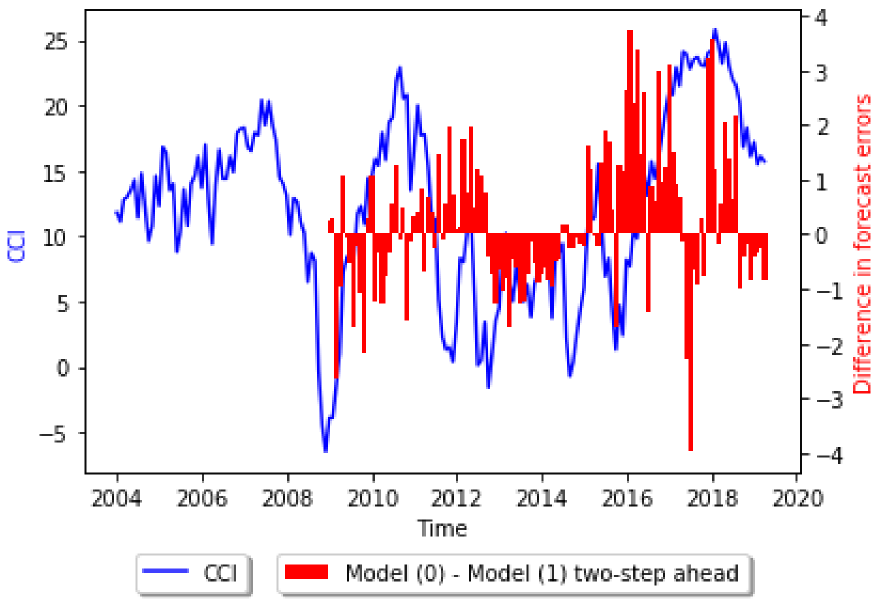

Figure 4.6. The model performance measured by the difference in absolute mean errors for two-step-ahead nowcasts of the benchmark model (0) and the extended. The positive vertical bar indicates that the extended model performs better, and vice versa.

5 DISCUSSION

5.1 Results

In an in-sample setting, we found that the coefficients of the Google categories “Trucks & Suv” and “Hybrid & Alternative Vehicles” are statistically significant at 5% level. This is in line with the previous studies by Choi & Varian (2012) and Della Penna (2009) that also report the connection of these categories with consumer confidence in U.S. On the other hand, our results show that the unemployment related searches have no statistical significance at 10% level. This is a bit unexpected result since one of the main drivers of CCI is unemployment as mentioned in Section 2.4. Also, previous research has shown positive results when predicting unemployment with Google data (e.g. Tuhkuri, 2015), which could suggest the same in case of predicting consumer confidence.

The benchmark model can alone explain a large part of the variation in the CCI. Including the Google searches to the model increases the R^2 and decreases Akaike information criteria, which indicate that Google data can be useful predictor for CCI. Changes in the Bayesian information criteria, on the other hand, appears to be in contrast with the previous conclusion. A possible reason why BIC penalizes more than AIC is that we use multiple Google search variables in the model. To improve the model further, one could construct a single Google variable from multiple search categories and terms as it has been done in e.g. Tuhkuri (2014, 2015) and Widgrén (2016). This could be a possible direction for further development.

Our results from (pseudo) out-of-sample predictions show five percent improvement for the nowcasts when including the Google data. When forecasting the future, the accuracy of the both models decrease generally, which is logical. However, when comparing the models, we found that the forecasts computed from the model with Google data outperform the benchmark model for each length of the horizon we tried. The difference between models’ performance grows as the forecast horizon increases.

When assessing the output of the both models in more detail, the Diebold-Mariano test reports that there is statistical difference at 5% level or 1% level for the forecasts between horizons h = 4 and h = 6. This indicates that Google data might help forecasting the CCI multiple steps ahead, which is in line with our cross-correlation analysis in Section 3.2. and the model estimates with lagged search queries in Section 4.2. The link appears to be strongest between current search activity and the CCI six months ahead.

Interestingly, we also observed that the improvement in one-step ahead nowcasts with Google data is even greater during the economic turning points than in the flat time series, although the Diebold-Mariano test reports no statistical difference at 10% level in this scenario. However, the visual inspection of the figures 4.5 and 4.6 supports the idea that Google data can be useful in forecasting when consume confidence is changing more suddenly, which aligns with the findings of Choi & Varian (2012) and Goel et al. (2010).

To summarize the results, the improvements in predictions in most cases are only modest. It is however not necessary discouraging that the improvements are not large. Our results suggest that Google data can boost the prediction accuracy especially in the turning points of the economic time series, which can be hard to identify when using only autoregressive baseline models.

Next, let’s discuss about the limitations of using internet search data in forecasting.

5.2 Limitations

One of the main difficulties lies in identifying the relevant search queries. Particularly, this is a challenge in the consumer confidence forecasting, because there can be several diverse factors affecting as we presented in Section 3.2. In this thesis, we generate the predictions with a very limited set of search volumes as our objective is to answer whether Google data can predict consumer confidence, not to identify the optimal set of queries (that can be a research topic of its own). We reviewed the existing literature of consumer confidence determinants and used those in identifying the Google search categories and search terms that could represent consumer confidence. In order to find the optimal set of Google search volumes, one could extract a larger amount of CCI-related search queries from Google Trend and test the models with multiple combinations of them.

Also, the signalling effect of the internet searches can vary among the query domains. Niesert et al. (2018) mention that Google search data appear to be most helpful when the series under investigation directly relates to an individual’s personal situation and is closely linked with specific search behaviour (such as employment status), but can be less reliable when it comes to macroeconomic measures that are unknown to the individual or too general to be linked to specific search terms. For example, many unemployed people may have known in advance that they were at risk of becoming unemployed, knowledge that would have generated specific and predictable online search behaviour. Conversely, in case of consumer confidence, the similar type of knowledge can be insufficient to generate specific and predictable search behaviour.

More generally, technological development can cause major changes in online search behaviour over time. For example, a move from personal computers to mobile devices changes the way people search and explore the data. In mobile devices, the interactions and data lie more within native applications if compared against PC users that rely more on browsers. Furthermore, an increased use of voice search might generate longer and more unique search queries that might make the search query selection for the predictions more complex.

One disadvantage is the unstructured nature of Google data, which is not originally generated for analysis purposes. Sawaengsuksant (2019) points out that data from Google applications provide searching frequency which is actually a bundle of numerous signals: both the informative signals and unrelated noises. Sources of noises are also different from the traditional structured data, raising a challenge in further applications' validity. For examples, noises could be generated simply from technical issues of the program interface, or from human behaviours unrelated to the topic of interest. Furthermore, Bortoli and Combes (2016) note that shortness of series and lack of transparency about treatments and sampling processes are weakness of Google Trend.

The search algorithms are not a static entity either, but merely the subject of constant testing and improvements. Google makes hundreds of changes to its search algorithm each year.[6] While most of these changes are minor, Google occasionally rolls out a major algorithmic update that can affect search results in significant ways. For example, Lazer et al. (2014) reports that several changes in Google's search algorithm have likely affected negatively Google Flu Trend's (GFT) tracking, which is reported to be performing unreliably at times. The modifications in e.g. recommended searches that usually are based on what others have searched, will increase the relative magnitude of certain searches. Because GFT uses the

[6] The official Google blog, https://blog.google/products/search/

relative prevalence of search terms in its model, improvements in the search algorithm probably had affected GFT's estimates.

As discussed above, there can be several challenges in using internet search data in predictions. Search patterns are the result of thousands of decisions made by programmers and by millions of consumers worldwide. Thus, the search behaviour is not only exogenously determined, but also endogenously related to the search engine. Consequently, it is relevant to understand the data and algorithms in order to produce robust economic forecasts.

Finally, the methods we have used in this paper are relatively simple and might not represent the optimal way generate forecasts from this data. Also, the observation period is short and there is only two major increase and one decrease in the CCI. Based these few events, it is not clear whether the patterns detected in this thesis would hold in the future. Thus, longer time series would be needed to make more accurate conclusions of the predictive ability of internet search data.

Next, we will provide potential directions for future research from the results of this thesis.

5.3 Future Research

As we mentioned in the previous section, it would be interesting to extract a larger set of the CCI-related search queries from Google Trends and test if they improve the predictive ability of our model. In the defining phase of the relevant queries, one could utilize more extensively the official CCI questionnaire[7] used by Statistics Finland as the basis for search term identification. This might reveal more potential areas for tracking consumer search behaviour that could benefit in predicting the CCI.

Furthermore, as this thesis uses simple autoregressive models, one could test the predictive ability of the Google data with more advanced models. For example, it would be interesting to try how e.g. neural network-based models would perform. Xie et al. (2014) has predicted the Chinese consumer confidence index with a radial basis function (RBF) neural network. They compared the results of RBF against autoregressive models and reported that the latter

[7] Official Statistics of Finland (OSF): Consumer Confidence [e-publication].
ISSN=2669-8889. Helsinki: Statistics Finland [referred: 25.12.2019].
Access method: http://www.stat.fi/til/kbar/kbar_2017-05-05_men_001_en.html

one provided more accurate predictions and better fitting effects. This result encourages one to test if the same approach would improve the predictions also in case of the Finnish CCI.

To study the robustness of the results, one could consider extending the methods used in this thesis to cover more countries than Finland. A natural choice would be to select other EU countries to conduct a panel data exercise. Panel data methods provide an opportunity to control for unobserved factors in the relationship between Google searches and consumer confidence. To achieve comparability between countries, the consumer confidence data collection has been harmonized among the EU member states by using similar questionnaires and by conducting the national surveys with transmission of the results according to a common timetable. Also, the gap in Internet penetration among the EU countries is decreasing as almost 85 percent of European households have an internet connection[8]. To include all 27 EU member states would increase the number of observations from 184 to 5192, which would help to compensate for the relatively short time series that is available from Google Trends from 2004 onwards.

Incorporating a higher number of Google search queries, new forecasting models and other countries to scope would increase the amount of effort needed in e.g. model training and maintenance. This can be an exhaustive task to do manually. For example, automation would be useful in selecting the best performing search query and model pair among the hundreds or thousands of combinations. In the bigger picture, it would be interesting to build a similar tool than ETLAnow[9], which utilizes Google search data to predict the official unemployment rate the EU-27 countries. The ETLAnow model automatically predicts the unemployment rate for three months ahead using data from Google Trends database and Eurostat, and publishes the updated forecasts every morning (Tuhkuri, 2016). The same can be done for consumer confidence predictions.

In addition to Google Trends, there are several other sources of data on real time economic activity from e.g. private sector companies like MasterCard, UPS, Twitter and many others. In Finland, for example, ETLA has recently studied if truck traffic data could improve the gross domestic product forecasts[10]. The traffic flows are publicly available on the website

[8] https://www.statista.com/topics/3853/internet-usage-in-europe/
[9] ETLAnow, https://www.etla.fi/etlanow/
[10] Kuorma-autoliikennedatan käyttö talouden nykyhetken ennustamisessa, ETLA Suhdanne, https://www.suhdanne.fi/artikkelit/kuorma-autoliikenne-datan-kaytto-talouden-nykyhetken-ennustamisessa/

of The Finnish Transport Infrastructure Agency[11]. The measurement of the traffic flows is accomplished with automatic cameras that are scattered along the Finnish roads. The available time series are longer than Google data as they start from 1997, which is a benefit for prediction models. Furthermore, the data is available almost in real time basis as publication delay is one day. This could be a potential data source also for the consumer confidence predictions since the traffic flows might reveal useful information of e.g. the demand of durable goods.

[11] The Finnish Transport Infrastructure Agency, https://vayla.fi/

6 CONCLUSIONS

Government agencies periodically release indicators of the level of economic activity in various sectors including, for example, consumer confidence. However, these releases are typically only available with a reporting lag of several weeks and are often revised a few months later. It would clearly be helpful to have more timely forecasts of these economic indicators.

In this thesis we study whether Google search volumes could help to predict the consumer confidence indicator in Finland. We found that autoregressive models with relevant Google variables tend to generate, on average, more accurate forecast than the same models without those predictors. The result of (pseudo) out-of-sample predictions indicate that changes in Google searches often precede changes in consumer confidence, which suggest Google data could help to predict the present and near future.

There are few of findings we like to emphasize. First, improvements in the predictive accuracy of using Google data appear to be limited for short term predictions. Our results show that current search activity provides useful information for the CCI predictions up to six months ahead. Second, the informative value of search data tends to be time specific. We found that in 2009-2012, when the CCI fluctuated greatly in Finland due to e.g. financial crisis, Google data improves the accuracy of nowcasts over the benchmark, on average, more than on the periods when the fluctuations are modest. This aligns with Goel et al. (2010) and Choi & Varian (2012) who reckon that sudden changes in search intensity could help to identify the sudden changes in economic time series.

As an answer to the research question, we conclude that Google searches do predict consumer confidence. More generally, this thesis illustrates the potential of Internet searches in predicting economic indicators. The results also demonstrate that big data can be utilized to forecast official statistics. However, it is important to emphasize that the predictive power of Google searches seems to be limited to relatively short-term predictions, and the improvements are modest.

In summary, as the Internet becomes more and more an integral part of daily life, people use it to search and share information about issues and items that interest or concern them. This generates an increasing amount of Big Data, such as Google's search volumes, that will contain new information that can be useful in near future. The information, that previously was unmeasurable, have now become available for measure. This can create a vast amount of new opportunities in variety of domains, including predictions of economic indicators.

REFERENCES

Akaike, H. 1973. *Information theory and an extension of the maximum likelihood principle*. In Petrov, B. N. and Cszaki, F., editors, Second International Symposium on Information Theory, p. 267–281. Akadémiai Kiadó, Budapest

Adams, F. & Green, E. 1965. Explaining and Predicting Aggregative Consumer Attitudes. *International Economic Review, 6*, p. 275.

Askitas, N. & Zimmermann, K. F. 2009. Google Econometrics and Unemployment Forecasting. *SSRN Electronic Journal*. doi:10.2139/ssrn.1480251

Beltran, H. 2003. The determinants of consumer confidence: The case of United States and Belgium. *IDEAS Working Paper Series from RePEc*.

Combes, S. and Bortoli, C., 2016, October. Nowcasting with Google Trends, the more is not always the better. In CARMA 2016: 1st International Conference on Advanced Research Methods in Analytics (pp. 15-22). Editorial Universitat Politècnica de València.

Bram, J. 1998. Does consumer confidence forecast household expenditure? A sentiment index horse race. *Economic Policy Review - Federal Reserve Bank of New York, 4*(2), pp. 59-78.

Campbell, J. Y. 1989. Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence. *NBER Macroeconomics Annual, 4*, pp. 185-216. doi:10.1086/654107

Choi, H., & Varian, H. 2009a. Predicting initial claims for unemployment benefits. Technical Report, Google Inc, 1-5.

Choi, H., & Varian, H. 2009b. Predicting the present with google trends. Technical Report, Google Inc, 1-5.

Choi, H. & Varian, H. 2012. Predicting the Present with Google Trends. *Economic Record, 88*(Supp.1), pp. 2-9. doi:10.1111/j.1475-4932.2012.00809.x

Clar, M., Duque, J. & Moreno, R. 2007. Forecasting business and consumer surveys indicators-a time-series models competition. *Applied Economics, 39*(20), pp. 2565-2580. doi:10.1080/00036840600690272

Cooper, C., Mallon, K., Leadbetter, S., Pollack, L. & Peipins, L. 2005. Cancer Internet search activity on a major search engine, United States 2001-2003. *Journal of Medical Internet Research, 7*(3), . doi:10.2196/jmir.7.3.e36

D'Amuri, F. & Marcucci, J. 2012. The predictive power of Google searches in forecasting unemployment. *IDEAS Working Paper Series from RePEc*.

Della Penna, N. 2009. *Constructing consumer sentiment index for US using Google searches*. No 2009-26, Working Papers from University of Alberta, Department of Economics

Dickey, D. A. & Fuller, W. A. 1979. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association, 74*(366), pp. 427-431. doi:10.2307/2286348

Diebold, F. & Mariano, R. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics,* p. 253.

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

Ekici, T. 2016. Subjective Financial Distress in the Formation of Consumer Confidence: Evidence from Novel Household Data *. *Boğaziçi Journal, 30*(2), pp. 11-36. doi:10.21773/boun.30.2.2

Ettredge, M., Gerdes, J. & Karuga, G. 2005. Using web-based search data to predict macroeconomic statistics. *Communications of the ACM, 48*(11), pp. 87-92. doi:10.1145/1096000.1096010

Fuhrer, J. C. 1993. What role does consumer sentiment play in the U.S. Macroeconomy? *New England Economic Review,* p. 32

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. 2008. Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), p. 1012. doi:10.1038/nature07634

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. & Watts, D. J. 2010. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences, 107*(41), p. 17486. doi:10.1073/pnas.1005962107

Guzman, G. 2011. Internet search behavior as an economic forecasting tool: The case of inflation expectations.(Report)(Author abstract). *Journal of Economic and Social Measurement, 36*(3), p. 119. doi:10.3233/JEM-2011-0342

Hollanders, D. & Vliegenthart, R. 2011. The influence of negative newspaper coverage on consumer confidence: The Dutch case. *Journal of Economic Psychology, 32*(3), pp. 367-373. doi:10.1016/j.joep.2011.01.003

Jennings, E. & Mcgrath, P. 1994. The influence of consumer sentiment on the sales of durables. *The Journal of Business Forecasting Methods & Systems, 13*(3), p. 17

Juster, F. T. & Wachtel, P. 1972a. Anticipatory and Objective Models of Durable Goods Demand. *The American Economic Review, 62*(4), pp. 564-579.

Juster, F. T. & Wachtel, P. 1972b. Uncertainty, Expectations, and Durable Goods Demand Models. In Human Behavior in Economic Affairs: Essays in Honor of George Katona, eds. Strumpel, Morgan, and Zahn. Amsterdam: Elsevier Scientific Publishing Co.

Keynes, J. M. 1936. *The general theory of employment interest and money*. London: McMillan.

Kwiatkowski, D., Phillips, P. C., Schmidt, P. & Shin, Y. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics, 54*(1), pp. 159-178. doi:10.1016/0304-4076(92)90104-Y

Lazer, D. M., Kennedy, R., King, G. & Vespignani, A. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science, 343*(6176), . doi:10.1126/science.1248506

Lindberg, F., 2011. *Nowcasting Swedish Retail Sales with Google Search Query Data*. Master's Thesis, Stockholm University

Ljung, G. M. & Box, G. E. P. 1978. On a Measure of Lack of Fit in Time Series Models. *Biometrika, 65*(2), pp. 297-303. doi:10.1093/biomet/65.2.297

Lovell, M. C. 1975. Why was the Consumer Feeling so Sad? *Brookings Papers on Economic Activity, 1975*(2), pp. 473-479. doi:10.2307/2534109

Ludvigson, S. C. 2004. Consumer Confidence and Consumer Spending. *Journal of Economic Perspectives, 18*(2), pp. 29-50. doi:10.1257/0895330041371222

Mandal A. & McCollum J. 2013. Consumer Confidence and the Unemployment Rate in New York State: A Panel Study. New York Economic Review, New York State Economics Association (NYSEA), vol. 44(1), pages 3-19.

Mankiw, N. G. 2004. *Principles of economics*. 3rd ed. Mason, OH: Thomson/South-Western.

Mclaren, N. & Shanbhogue, R. 2011. Using internet search data as economic indicators. *Bank of England. Quarterly Bulletin, 51*(2), pp. 134-140.

Mueller, E. 1966. The Impact of Unemployment on Consumer Confidence. *The Public Opinion Quarterly, 30*(1), p. 19. doi:10.1086/267379

Niesert, R., Oorschot, J., Veldhuisen, C., Brons, K. & Lange, R. 2018. Can Google Search Data Help Predict Macroeconomic Series? *SSRN Electronic Journal*. doi:10.2139/ssrn.3353123

Polgreen, P. M., Chen, Y., Pennock, D. M. & Nelson, F. D. 2008. Using internet searches for influenza surveillance. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 47*(11), pp. 1443-1448. doi:10.1086/593098

Praet, P. & Vuchelen, J. 1989. The contribution of consumer confidence indexes in forecasting the effects of oil prices on private consumption. *International Journal of Forecasting, 5*(3), pp. 393-397. doi:10.1016/0169-2070(89)90042-3

Sawaengsuksant, P., 2019. Standardised approach in developing economic indicators using internet searching applications. IFC Bulletins chapters, 50.

Schwarz, G. 1978. Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), pp. 461-464. doi:10.1214/aos/1176344136

Stock, J. H. 2001. Forecasting economic time series. *A Companion to Theoretical Econometrics, Blackwell Publishers*, 562–84.

Throop, A. W. 1992. Consumer sentiment its causes and effects. *Economic review : an annual report of the Economic Research Department,* pp. 35-59.

Tuhkuri, J. 2014. *Big Data: Google-haut ennustavat työttömyyttä Suomessa*. ETLA Raportit No 31. http://pub.etla.fi/ETLA-Raportit-Reports-31.pdf

Tuhkuri, J. 2015. *Big Data: Do Google Searches Predict Unemployment?* Master's thesis. University of Helsinki, Faculty of Social Sciences, Department of Political and Economic Studies.

Tuhkuri, J. 2016. *ETLAnow: A Model for Forecasting with Big Data – Forecasting Unemployment with Google Searches in Europe*.

Varian, H. R. 2003. *Intermediate microeconomics: A modern approach*. 6th ed. New York: Norton.

Vosen, S. & Schmidt, T. 2011. Forecasting private consumption: Survey-based indicators vs. Google trends. *Journal of Forecasting, 30*(6), pp. 565-578. doi:10.1002/for.1213

Widgrén, J. 2016. ”Google-haut Suomen asuntojen hintojen ennustajana”. ETLA Raportit No 63. https://pub.etla.fi/ETLA-Raportit-Reports-63.pdf

Wu, L. & Brynjolfsson, E. 2009. The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. *SSRN Electronic Journal*. doi:10.2139/ssrn.2022293

Xie, H. Y., Chen, G. Y., He, Z. & Zhao, D. P. 2014. Forecasting Consumer Confidence Index Based on Radial Basis Function Neural Network. *Applied Mechanics and Materials, 596*(Mechatronics and Industrial Informatics II), pp. 200-203. doi:10.4028/www.scientific.net/AMM.596.200

APPENDIX 1

Table A.1. The correlations between the CCI and the Google searches.

	The lag order *h*												
h	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Google category/search term													
Trucks & Suv [C]	0.505	0.503	0.484	0.476	0.477	0.471	0.450	0.443	0.431	0.416	0.407	0.402	0.393
Hybrid & Alternative Vehicles [C]	0.280	0.276	0.269	0.269	0.239	0.255	0.214	0.189	0.200	0.198	0.177	0.176	0.168
Welfare and Unemployment [C]	0.073	0.067	0.040	0.037	0.033	0.030	0.037	0.041	0.042	0.064	0.069	0.063	0.051
Bankcrupcy [C]	0.029	0.044	0.078	0.068	0.098	0.100	0.106	0.082	0.114	0.092	0.080	0.122	0.139
tyottomyyskorvaus	-0.391	-0.391	-0.387	-0.358	-0.332	-0.315	-0.314	-0.309	-0.284	-0.235	-0.190	-0.161	-0.120
ansiosidonnainen paivaraha	-0.369	-0.361	-0.376	-0.393	-0.370	-0.367	-0.335	-0.285	-0.227	-0.169	-0.111	-0.041	-0.041
kela tyottomyyspaivaraha	-0.197	-0.120	-0.104	-0.091	-0.066	-0.020	-0.001	0.033	0.069	0.065	0.064	0.079	0.126
vaihtoautot	0.214	0.234	0.230	0.225	0.257	0.259	0.239	0.245	0.229	0.201	0.222	0.217	0.205
pikavippi	-0.229	-0.234	-0.233	-0.237	-0.235	-0.249	-0.262	-0.269	-0.285	-0.294	-0.311	-0.318	-0.319
osakesijoittaminen	0.104	0.120	0.091	0.103	0.091	0.095	0.113	0.121	0.116	0.122	0.134	0.098	0.119
ytk	-0.271	-0.301	-0.311	-0.324	-0.283	-0.228	-0.150	-0.086	-0.029	0.005	0.052	0.053	0.040
tyottomyyskassa	-0.055	-0.027	-0.026	-0.010	0.032	0.067	0.118	0.155	0.204	0.243	0.281	0.301	0.317
velkaneuvonta	-0.023	-0.026	-0.029	-0.025	-0.005	-0.005	0.000	-0.014	0.021	0.033	0.076	0.078	0.061
ostaa	-0.012	-0.005	0.003	0.014	0.008	0.021	0.029	0.029	0.034	0.039	0.040	0.058	0.071
perintatoimisto	-0.005	0.014	-0.023	-0.051	-0.051	-0.075	-0.069	-0.084	-0.073	-0.049	-0.023	-0.051	-0.083
lennot	-0.090	-0.071	-0.056	-0.061	-0.024	-0.003	0.007	0.009	0.003	-0.015	-0.011	-0.014	-0.025
perinta	-0.044	-0.050	-0.069	-0.037	-0.045	-0.055	-0.045	-0.056	-0.036	-0.020	-0.010	0.018	0.019

n = 184. The label [C] indicates Google category (the search terms have no labels assigned).
The values associated with negative lag order *h* tell the correlation coefficients between the past Google search volumes and the present CCI. Similarly, the values associated with positive lag order *h* tell the correlation coefficients between future Google search volumes the present CCI. The lag order *h* zero tells the correlation coefficients between present Google search volumes and the present CCI.