ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY

ETLA Raportit
ETLA Reports

14.8.2014

No 31

Big Data: Google-haut ennustavat työttömyyttä Suomessa

Joonas Tuhkuri*

* ETLA – Elinkeinoelämän tutkimuslaitos, joonas.tuhkuri@etla.fi ja
Helsingin yliopisto, joonas.tuhkuri@helsinki.fi

Suositeltava lähdeviittaus: Tuhkuri, Joonas (14.8.2014). "Big Data: Google-haut ennustavat työttömyyttä Suomessa".
ETLA Raportit No 31. http://pub.etla.fi/ETLA-Raportit-Reports-31.pdf

Kiitän seuraavia henkilöitä kommenteista, keskusteluista, tiedoista ja avusta: Kimmo Aaltonen (ETLA), Kaija Hyvönen-Rajecki (ETLA), Eveliina Kuittinen (Google), Harri Laine (ETLA), Petteri Larjos (ETLA), Heini Lehikoinen (Kela), Jani Luoto (HY), Aleksandr Peussa (ETLA), Laila Riekkinen, Antti Ripatti (HY), Petri Roponen (Kela), Petri Rouvinen (Etlatieto/ETLA), Antti Ukkonen (Aalto), Roope Uusitalo (HECER/HY), Pekka Vanhala (ETLA), Vesa Vihriälä (ETLA), Johanna Wahlroos (Google) sekä seminaariosallistujia ETLAssa.

ISSN-L 2323-2447
ISSN 2323-2447 (print)
ISSN 2323-2455 (online)

Sisällysluettelo

Big Data: Google-haut ennustavat työttömyyttä Suomessa

Tiivistelmä
Suomessa tehdään päivittäin 30 miljoonaa Google-hakua. Tämä raportti selvittää, voiko Google-hauilla ennustaa nykyhetken ja lähitulevaisuuden työttömyyttä Suomessa. Nykyhetken ja lähitulevaisuuden ennustaminen on kiinnostavaa, sillä viralliset tiedot talouden tilasta julkaistaan viiveellä. Google-hakujen sisältämän informaation arvioimiseksi raportissa vertaillaan yksinkertaista työttömyyttä kuvaavaa mallia sellaiseen malliin, johon on lisätty Google-aineistosta muodostettu muuttuja, Google Index. Tämän lisäksi tarkastellaan muuttujien välisiä ristikorrelaatioita ja suoritetaan Granger-kausaalisuustesti. Yksinkertaiseen vertailukohtaan nähden Google-haut tarkentavat nykyhetken ennustetta 10 % absoluuttisella keskivirheellä mitattuna. Lisäksi Google-hakujen lisääminen malliin parantaa kolme kuukautta eteenpäin tehtyä ennustetta 39 %. Havaitaan myös, että Google-haut tarkentavat ennustetta erityisesti käännekohdissa. Tulokset viittaavat siihen, että Google-haut sisältävät hyödyllistä informaatiota nykyhetken ja lähitulevaisuuden työttömyydestä Suomessa.

Asiasanat: Big Data, Google, Internet, nykyhetken ennustaminen, ennustaminen, työttömyys, aikasarja-analyysi

JEL: C1, C22, C43, C53, C82, E27

Big Data: Google Searches Predict Unemployment in Finland

Abstract
There are over 3 billion searches globally on Google every day. This report examines whether Google search queries can be used to predict the present and the near future unemployment rate in Finland. Predicting the present and the near future is of interest, as the official records of the state of the economy are published with a delay. To assess the information contained in Google search queries, the report compares a simple predictive model of unemployment to a model that contains a variable, Google Index, formed from Google data. In addition, cross-correlation analysis and Granger-causality tests are performed. Compared to a simple benchmark, Google search queries improve the prediction of the present by 10 % measured by mean absolute error. Moreover, predictions using search terms perform 39 % better over the benchmark for near future unemployment 3 months ahead. Google search queries also tend to improve the prediction accuracy around turning points. The results suggest that Google searches contain useful information of the present and the near future unemployment rate in Finland.

Key words: Big Data, Google, Internet, nowcasting, forecasting, unemployment, time-series analysis

JEL: C1, C22, C43, C53, C82, E27

1 Johdanto

Voivatko tilastot Internet-hauista auttaa ennustamaan taloudellista toimintaa Suomessa?

Viralliset tiedot talouden tilasta julkaistaan tyypillisesti kuukausittain tai neljännesvuosittain. Nämä tiedot ovat kuitenkin käytettävissä vain viiveellä. Tietyn kuukauden tiedot julkaistaan useimmiten aikaisintaan seuraavan kuukauden lopulla, ja tyypillisesti tietoja korjataan useita kuukausia myöhemmin. Tieto talouden nykytilasta ei ole tarkka.

Talouden tilasta on kuitenkin tarjolla myös reaaliaikaista informaatiota. Tilastot tehdyistä Internet-hauista ovat esimerkki tällaisesta informaatiosta. Esimerkiksi *Google Trends* -tietokanta esittää päivittäin ja viikoittain tietoa Google-hakujen yleisyydestä eri toimialoilla. Aineisto on julkisesti saatavilla.

Tämä raportti selvittää, voisivatko Google-haut olla yhteydessä *nykyhetken* taloudellisen toiminnan kanssa tietyllä toimialalla Suomessa, ja siten olla avuksi ennustettaessa nykyhetkeä.

Tavoite on nykyhetken ennustaminen,[1] ei pitkälle tulevaisuuden ennustaminen.[2] *Google Trends* voi kuitenkin potentiaalisesti myös auttaa ennustamaan nykyhetken lisäksi lähitulevaisuutta. Raportissa ei esitetä, että Google-haut *vaikuttaisivat* talouden tilaan, vaan että ne saattavat olla yhteydessä talouden nykytilaan ja lähitulevaisuuteen.

Selkeyden vuoksi raportissa käsitellään vain yhtä ennustettavaa ilmiötä, työttömyyttä. Työttömyys on valittu esimerkiksi neljästä syystä. Ensimmäiseksi työttömyys on kuluttajailmiö, jossa Internet-haut potentiaalisesti ovat yhteydessä vallitsevaan tilanteeseen työmarkkinoilla. Toisin sanoen Google-haut eivät välttämättä tarjoa tietoa kaikista ilmiöistä, mutta työttömyyden kehityksestä ne saattavat kertoa. Työttömyys on lisäksi yhteiskunnallisesti merkittävä ilmiö, työttömyystilastot ovat käytettävissä viiveellä ja Internetin merkitys työmarkkinoilla on suuri (Kuhn & Skuterud 2004, Stevenson 2008, Kroft & Pope 2014). Esimerkiksi työnhaku tapahtuu laajalti Internetissä (Kuhn & Mansour 2014). Raportin menetelmillä on kuitenkin mahdollista tarkastella Google-hakujen kykyä ennustaa muitakin ilmiöitä.

Esimerkiksi työttömyyskorvaukseen liittyvillä hakusanoilla tehtyjen hakujen määrä toukokuussa voi olla hyödyllinen tieto ennustettaessa touko-

[1] engl. *nowcasting*
[2] engl. *forecasting*

kuun työttömyysastetta, joka julkaistaan vasta kesäkuun lopussa. Voi myös olla, että tehtyjen hakujen määrä toukokuussa auttaa ennustamaan kesäkuun tai jopa elokuun työttömyyttä. Se kuinka kauas ennuste kantaa, riippuu ennustettavasta ilmiöstä.

Tämän raportin tavoite on esitellä Big Datan, tässä tapauksessa Google-hakuaineiston, mahdollisuuksia talouden ennustamisessa. Se rakentuu pitkälti aikaisemman kirjallisuuden varaan erityisesti Choin ja Varianin (2009a, 2009b, 2012) esittämään lähestymistapaan ja menetelmään Google-hakuaineiston hyödyllisyyden testaamiseksi ennustamisessa. Vaikka Internet-hakuja käsittelevien tilastojen hyödyllisyyttä talouden tilan hahmottamisessa on tutkittu kansainvälisesti, selvitystä Internet-hakujen käyttämisestä Suomen talouden ennustamisessa ei ole aiemmin esitetty.

Google-hakuaineiston sisältämän informaation arvioimiseksi tässä raportissa vertaillaan yksinkertaista työttömyyttä kuvaavaa mallia sellaiseen malliin, johon on lisätty Google-aineistosta muodostettu muuttuja, Google Index. Tämän lisäksi tarkastellaan muuttujien välisiä ristikorrelaatioita ja suoritetaan Granger-kausaalisuustesti.

Internet-hakuja koskevia tietoja hyödyntämällä on epäilyksettä mahdollista rakentaa kehittyneempiä malleja nykyhetken sekä tulevaisuuden ennustamiseksi. Yksinkertaisesta on kuitenkin hyvä aloittaa. Raportti ei kuvaa ETLAn muita menetelmiä työttömyyden ennustamiseksi.

Nykyhetken ennustaminen on jo käytännön syistä kiinnostavaa. Erityisesti talouskriisin aikana tarkka kuva talouden nykytilasta on arvokas. Lisäksi nykyhetken ennustamiseen liittyy myös monia taloustieteellisesti kiinnostavia tutkimuskysymyksiä (Castle et al. 2009). Näitä ovat esimerkiksi muuttujan valinta jopa miljoonien mahdollisten selittäjien joukosta, eri aikajänteillä julkaistavien aineistojen yhteensovittaminen sekä aineiston tarkistusten huomioiminen (Giannonne et al. 2008).

2 Kirjallisuus

Ensimmäinen julkaistu artikkeli, joka tarkastelee Internet-hakuaineiston käyttöä ennustamisessa, on Ettredge et al. (2005). Artikkeli esittelee tiettyjen hakusanojen ja Yhdysvaltojen työttömyysasteen välistä yhteyttä. Vuodesta 2005 lähtien lukuisat tutkimukset ovat tarkastelleen Internet-hakuaineiston käyttöä eri konteksteissa.

Taloustieteessä Choi ja Varian (2012) esittelevät ensimmäisten joukos-

sa Google-hakuaineiston käyttöä ennustamisessa käyttäen esimerkeinä nykyhetken työnvälitystilastojen, autokaupan, kuluttajien luottamusindeksin ja matkailun ennustamista. Choi ja Varian (2012) on yhteenveto kahdesta aiemmasta Google-aineiston mahdollisuuksia käsittelevästä artikkelista (Choi & Varian 2009a, 2009b).

Ginsberg et al. (2009) esittää, että Internet-haut voivat ennustaa influenssatapauksia tarkemmin kuin tunnetut perinteiset menetelmät. Tämä tutkimus on saanut paljon huomiota, ja sitä on seurannut useita tutkimuksia Internet-hakujen yhteydestä epidemioihin. Tunnetuimpia näistä ovat Browstein et al. (2009) sekä Hulth et al. (2009). Internet-hakuaineiston hyödyllisyys influenssan ennustamisessa on kuitenkin myös kyseenalaistettu useaan otteeseen (Lazer et al. 2014).

Työttömyyden ennustaminen on ollut Internet-hakuja käsittelevässä *"Googlenomics"*-kirjallisuudessa[3] keskeinen teema. Esimerkiksi Askitas ja Zimmermann (2009), D'Amuri ja Marcucci (2012), D'Amuri (2009) ja Suhoy (2009) tarkastelevat työttömyyden ennustamista ja Internet-hakuja Saksassa, Yhdysvalloissa, Italiassa ja Israelissa. Guzman (2011) tarkastelee Google-aineiston käyttökelpoisuutta inflaation ennustamisessa.

Wu ja Brynjolfsson (2014) esittelevät asuntomarkkinoiden ennustamista Yhdysvalloissa Google-hakujen avulla. He havaitsevat, että asunnon hankintaan liittyvät Google-haut hahmottavat nykyhetken asuntokaupan lisäksi myös tulevaisuutta. Google-hakusanojen yleisyyttä kuvaavien tilastojen on esitetty auttavan ennustaettaessa osakekursseja (Preis et al. 2013), osakekauppaa (Bordino et. al 2012), yksityistä kulutusta (Vosen & Schmidt 2011) ja maksuvaikeuksia (Askitas & Zimmermann 2011).

Goel et al. (2010) tarjoaa selkeän katsauksen Internet-hakulokien käyttämiseen ennustamisessa ja kuvailee hakudatan käytön rajoituksia. Goel et. al (2010) huomauttaa, että vaikka hakuaineisto on hyvin saatavilla ja se parantaa usein ennusteita, parannukset jäävät kovin pieniksi. Vaikka yleisesti ottaen Goel et al. (2010) tarjoama näkemys on järkevä, parannukset tosiaan ovat usein pieniä, joissakin tapauksissa pienistäkin parannuksista voi olla hyötyä (Choi & Varian 2012). Tästä tunnettu esimerkki on osakemarkkinat, jossa hyvin pienellä ennustetarkkuuden parannuksella voi olla hyvin suuri taloudellinen arvo.

Koop ja Onorante (2013) ennustavat onnistuneesti yhdeksää keskeistä

[3]Levy, S. (2009). Secret of Googlenomics: Data-Fueled Recipe Brews Profitability. *Wired Magazine*, 17(06).

makromuuttujaa Yhdysvalloissa käyttäen bayesilaista lähestymistapaa, jossa Google-muuttujien valintaan ei käytetä lainkaan talousteoriaa tai ennakkotietämystä sopivista hakusanoista. Bayesilaista muuttujanvalintaa Google-hakuaineiston hyödyntämisessä käsittelee myös Scott ja Varian (2013, 2014). Sopivien muuttujien valinta useiden potentiaalisten selittäjien joukosta on keskeisiä haasteita Google-hakuaineiston sekä muun Big Datan hyödyntämisessä. Ginsberg et al. (2009) sekä Curme et. al (2014) esittelevät kaksi erilaista automaattista menetelmää sopivien hakusanojen valitsemiseksi.

Käytännön soveltajaa varten McLaren ja Shanbhoge (2011) esittävät selkeän yhteenvedon Internet-hakuja koskevien tilastojen hyödyntämisestä Englannin keskuspankin ennustetyössä. McLaren ja Shanbhoge (2011) tarjoavat myös hyödyllisen vertailun Google-hakuaineiston ja perinteisten ennakoivien indikaattoreiden kuten kuluttajien luottamusindeksin välillä.

Tämä raportti käsittelee ensimmäisenä Suomessa Google-hakujen käyttöä ennustamisessa.

3 Aineisto

Raportin pääasiallisena aineistona käytetään Google-hakujen yleisyyttä kuvaavaa *Google Trends* -aineistoa sekä Tilastokeskuksen työvoimatutkimuksen aineistoa. Raportti keskittyy Suomen talouteen.

Tilastokeskus julkaisee Suomen työttömyysasteen kuukausittain. Julkaisu tapahtuu noin kuukauden viiveellä, mikä tarkoittaa sitä, että esimerkiksi toukokuun työttömyysaste selviää kesäkuun lopussa. Kärjistäen: emme tiedä mikä on työttömyysaste juuri nyt. Joskus tiedämme viime kuun työttömyyden, mutta useimmiten tiedämme vain toissa kuun työttömyysasteen. Työttömyys selvitetään kyselytutkimuksena, ja saatua tulosta ei muuteta tilastojulkistuksen jälkeen.

Google Trends -tietokanta raportoi minkä tahansa Google-hakusanan yleisyyden verrattuna muihin hakusanoihin päivittäin. Aineisto on laaja. Esimerkiksi Suomessa tehtiin vuonna 2013 keskimäärin 30 miljoonaa Google-hakua päivässä;[4] maailmanlaajuisesti Google-hakuja tehtiin vuoden 2014 alussa yli 3 miljardia päivässä.[5] Tilastokeskuksen mukaan Internetiä

[4]Googlen sisäinen data 2013.

[5]Googlen sisäinen data 2014.

käytti vuonna 2013 16–74 vuotiaista suomalaisista 92 %.[6] Myös Googlen käyttäjiä on paljon. Googlea käytti huhtikuussa 2014 suomalaisista internetinkäyttäjistä 96 %,[7] ja vastaavasti vuonna 2013 Googlen osuuden kaikista Suomessa tehdyistä Internet-hauista on arvioitu olleen 96 %.[8] Seuraavaksi yleisin hakukone oli vuonna 2013 *Microsoftin Bing*, jonka osuus tehdyistä hauista oli arviolta 2 %.[9] Tästä syystä Internet-hakuja tutkittaessa Google-haut ovat luonteva valinta aineistoksi. Google on myös tiettävästi ainoa hakukone, joka julkaisee tilastoja kaikista tehdyistä hauista.

Google Trends palvelu selvittää, kuinka monta hakua tietyillä termeillä on tehty verrattuna samana aikana tehtyjen Google-hakujen kokonaismäärään. Tämä tehdään analysoimalla osa Googlessa tehdyistä verkkohauista.

Google ei kuitenkaan raportoi tietyllä hakusanalla tehtyjen hakujen tarkkaa lukumäärää vaan indeksin, joka kuvaa hakuintensiteettiä. Indeksi on rakennettu siten, että valitulla hakusanalla tehtyjen hakujen lukumäärä jaetaan ensin samana ajankohtana ja samalla maantieteellisellä alueella tehtyjen hakujen kokonaismäärällä. Tällä tavalla tuotettu luku on valitulla hakusanalla tehtyjen hakujen osuus kaikista Google-hauista.[10] Lopuksi tiedot normalisoidaan asteikolle 0–100. Normalisointi tehdään siten, että jokainen arvo jaetaan sen hetken arvolla, jolla valitulla hakusanalla tehtyjen hakujen osuus kaikista Google-hauista oli suurin. Tämä hetki saa indeksin arvon 100, sillä tulokseksi saatu osuus kerrotaan vielä luvulla 100. Jos tietoja ei ole tarpeeksi eli vähintään 50 hakua (Choi & Varian 2011), lukemana näytetään 0.

Yhteenvetona tämä tarkoittaa sitä, että jos aikavälillä $\{1, \ldots, f\}$ hakusanalla k tehdään K_t hakua hetkellä t ja samana ajankohtana kaikkien tehtyjen hakujen lukumäärä on G_t, niin käytetty hakuintensiteetin mittayksikkö on

[6] Väestön tieto- ja viestintätekniikan käyttö -tutkimus 2013, Tilastokeskus, "Käyttänyt internetiä viimeisten 3 kk aikana".

[7] comScore MMX, Finland, Age 15+, April 2014.

[8] StatCounter Global Stats, Top 5 Desktop, Tablet & Console Search Engines in Finland from Jan to Dec 2013.

[9] StatCounter Global Stats, Top 5 Desktop, Tablet & Console Search Engines in Finland from Jan to Dec 2013.

[10] Ajankohta ja maantieteellinen alue ovat käyttäjän valittavissa.

$$I(K_t) = \left\{ \frac{\frac{K_t}{G_t}}{max\left(\frac{K}{G}\right)} \right\} \times 100 \quad (1)$$

jossa

$$K = \{K_1, K_2, \ldots, K_t, \ldots, K_f\}$$
$$G = \{G_1, G_2, \ldots, G_t, \ldots, G_f\}$$

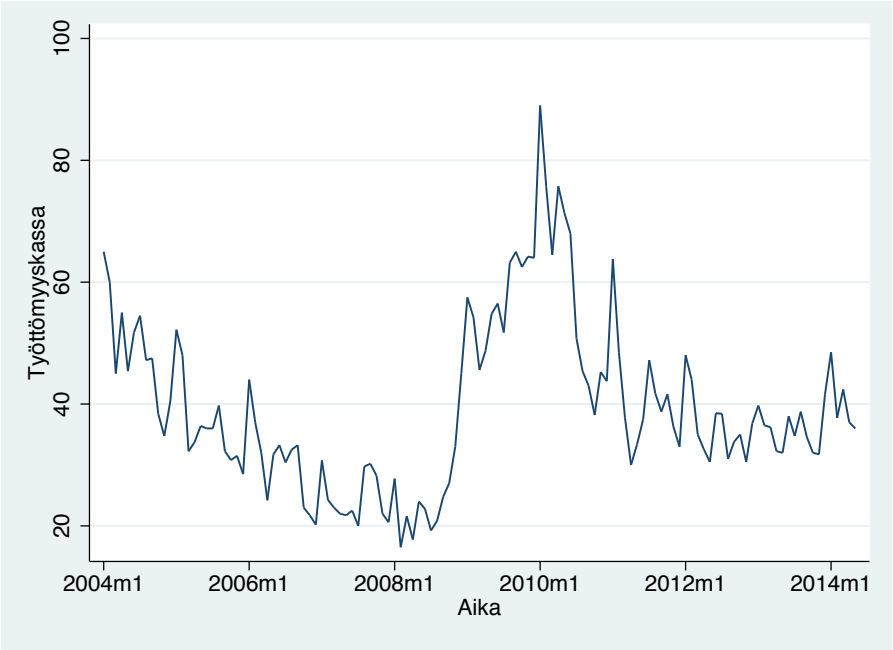
Kuvio 1 esittää esimerkiksi hakusanalla "työttömyyskassa" tehtyjen hakujen kehityksen vuoden 2004 tammikuusta vuoden 2014 huhtikuulle. Hakusanan yleisyys oli laskussa vuoden 2008 puoliväliin saakka, kunnes sen suosio kasvoi voimakkaasti saavuttaen huippunsa vuoden 2009 lopussa. Vuoden 2011 loppuun mennessä hakusanan "työttömyyskassa" hakuintensiteetti oli kuitenkin laskenut jo vuoden 2006 tasolle. Hakusanalla "työttömyyskassa" tehtyjen hakujen suhteellisen osuuden voimakas kasvu vuonna 2008 saattoi olla yhteydessä vuonna 2008 alkaneeseen talouskriisiin.

Google Trends -aineisto kattaa maailmanlaajuisesti aikajänteen tammikuun alusta 2004 nykyhetkeen.

Internet-hakuaineistolla on monia etuja talouden ennustamista ajatellen. Kiinnostavinta Google-aineistossa on sen reaaliaikaisuus. Useimmat tiedot ovat saatavissa viikkotasolla, mutta uusimman viikon saamaa arvoa tarkistetaan päivittäin. Google Trendsin tiedot ovat saatavissa usein selvästi ennen virallisia tietoja. Esimerkiksi 23.6.2014 viimeisin tieto työttömyysasteesta oli huhtikuulta, sillä toukokuun työttömyysaste julkistettiin 24.6.2014. Google-aineistoa oli kuitenkin saatavilla 7 viikon verran pidemmälle, 23.6.2014 asti.

Toisin kuin monet kyselyaineistot, tilastot tehdyistä Google-hauista syntyvät Internetin käytön sivutuotteena (Varian 2010). Tämä saattaa vähentää kyselytutkimuksiin tavallisesti liittyvän kadon ja epätarkkuuden aiheuttamia ongelmia. *Google Trends* -aineisto on myös ilmainen. Lisäksi kysymysten ei tarvitse olla etukäteen päätettyjä, vaan informaatiota kerätään jatkuvasti laajalta alueelta. Tämän seurauksena Internet-hakuaineisto voi auttaa analysoimaan yllättäen esiin nousevia ilmiöitä.

Aineiston eduista huolimatta Internet-hakuja koskevan aineiston käyttöön liittyy myös ongelmia. Internet on vielä suhteellisen uusi ilmiö, joten tilastoja tehdyistä Google-hauista on saatavilla vain vuodesta 2004 lähtien. Tämä on lyhyt ajanjakso verrattuna muihin taloudellisiin indikaat-



Kuva 1: Hakusanalla ”työttömyyskassa” tehtyjen hakujen intensiteetti 2004–2014. Lähde: *Google Trends*

toreihin, sekä lyhyt ajanjakso moniin ekonometrisiin sovelluksiin. Vaikka aineisto on laaja, se ei ole välttämättä kattava. Internetin käyttö on yhä voimakkaasti yhteydessä henkilön koulutukseen ja asuinpaikkaan. Esimerkiksi Suomessa vuonna 2013 korkeakoulutetuista 97 % käytti Internetiä, kun taas ylimpänä tutkintonaan perusasteen suorittaneista vain 68 % oli käyttänyt Internetiä viimeisen 3 kuukauden aikana.[11] On myös paljon taloudellista toimintaa, jossa Internet-hakujen rooli on vähäinen. On epätodennäköistä, että yritysten tulevat investoinnit näkyisivät selkeästi havaittavina hakuina Googlen hakutilastoissa. Tästä syystä Google-haut ovat erityisen hyödyllisiä kuluttajalähtöisissä sovelluksissa, kuten esimerkiksi kulutuspäätösten (Goel et al. 2010, Vosen & Schmidt 2011), asuntokaupan (Wu & Brynjolfsson 2014, McLaren & Shanbhoge 2011, Choi & Varian 2009b) tai työttömyyden ennustamisessa (Choi & Varian 2012, Askitas & Zimmermann 2009).

Keskeinen ongelma on myös se, että aineistoa tuottava mekanismi ei ole täysin selvillä. Eri käyttäjät saattavat käyttää eri hakusanoja, vaikka heidän hakunsa tarkoitus olisi sama. Myös päinvastoin, käyttäjät joilla

[11] Väestön tieto- ja viestintätekniikan käyttö -tutkimus 2013, Tilastokeskus.

on täysin eri aikomukset saattavat tehdä hyvinkin samankaltaisia Internet-hakuja. Esimerkiksi hakusanalla "työttömyyskassa" hakevista henkilöistä osa on mahdollisesti kiinnostunut liittymään kassan jäseneksi, osa ilmoittamaan jääneensä työttömäksi ja osa vain kiinnostunut selvittämään, mistä työttömyyskassassa on kyse. Vastaavasti tiettyyn yritykseen liittyvät haut voivat liittyä yhtä lailla positiiviseen tai negatiiviseen tuloskehitykseen (Preis et al. 2010). Tämän seurauksena hakutilastojen tulkinnassa on oltava varovainen.

Yksi huomioitava vaikeus on *Google Trends* -aineiston mittayksikön tulkinta. Hakuintensiteettiä kullakin hetkellä kuvataan osamääränä $^{K_t}/_{G_t}$, jossa K_t on valitulla hakusanalla hetkellä t tehtyjen hakujen lukumäärä ja G_t on samalla ajanhetkellä tehtyjen hakujen yhteenlaskettu lukumäärä. Aineisto ei siten kerro hakujen todellista lukumäärää. Hakuintensiteetin saama arvo voi käytetyn mittayksikön seurauksena muuttua, kun kiinnostuksen kohteena olevalla hakusanalla tehtyjen hakujen määrä muuttuu tai kun Google-hakujen yhteenlaskettu määrä muuttuu. Hakusanan yleisyyden kuvaamisessa on hyvää se, että se poistaa lähes kaikkiin hakusanoihin liittyvän nousevan trendin, joka johtuu siitä, että tehtyjen hakujen lukumäärät ovat kasvaneet moninkertaiseksi vuodesta 2004. Toisaalta menetelmä saattaa tuottaa uuden, laskevan trendin. Monen suositunkin hakusanan osuus tehdyistä hauista laskee jatkuvasti, sillä Internetissä tehtävien asioiden määrä kasvaa.

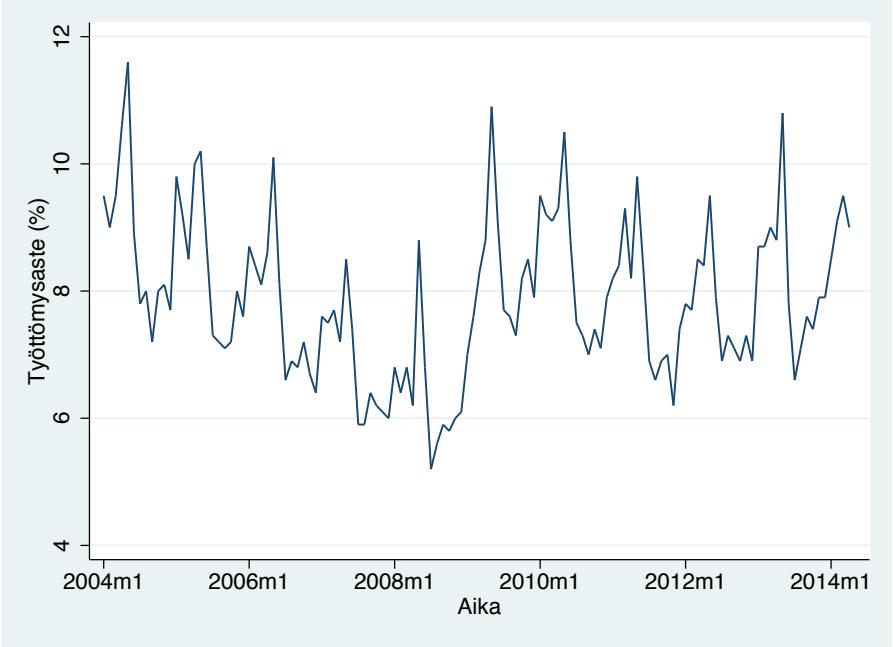
Yhtälössä (1) kuvatusta hakuintensiteetin laskentatavasta johtuen uusien arvojen mukaan ottaminen muuttaa indeksin saamia vanhojakin arvoja.

Googlen aineisto perustuu otokseen kaikista Googlessa tehdyistä verkkohauista. Tästä syystä se vaihtelee hieman päivästä toiseen. Mitä harvinaisempi hakusana on kyseessä, sitä suurempaa vaihtelu on.

Yhteenvetona: Google-hakuaineiston suurin etu on sen reaaliaikaisuus ja suurin heikkous se, että hakuja tuottava mekanismi ei ole täysin selvillä. Toisin sanoen käytössä on reaaliaikainen korrelaatio. Tästä syystä yksi luontevimmista käyttötarkoituksista Google-aineistolle on nykyhetken ennustaminen.

4 Menetelmä

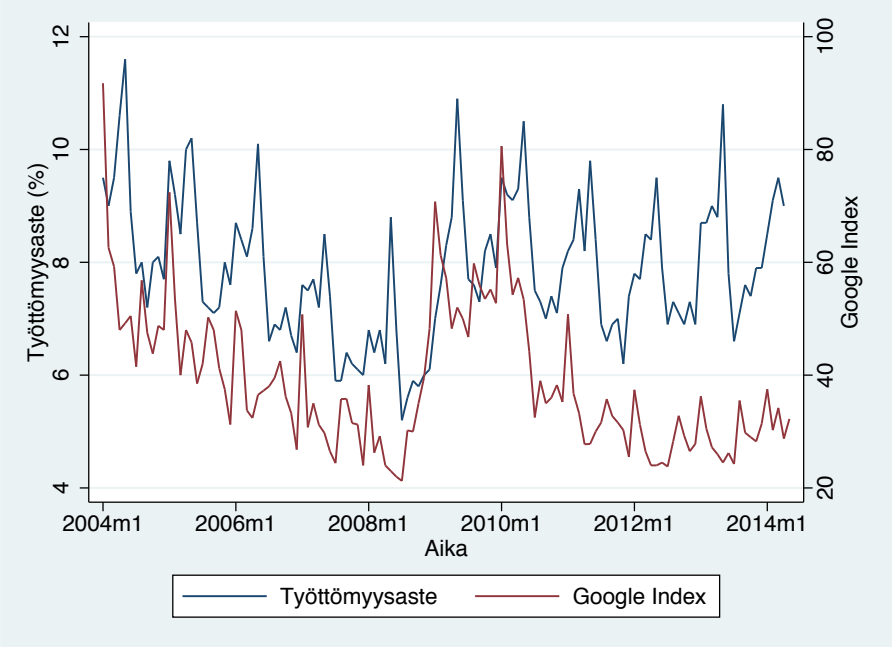
Tässä luvussa esitellään Choin ja Varianin (2012) esittämä menetelmä, jolla tarkastellaan, auttavatko Google-haut ennustamaan työttömyyttä Suomessa. Menetelmässä on kaksi vaihetta: muuttujan valinta ja tilastollinen testaus.



Kuva 2: Työttömyysaste 2004–2014. Lähde: Tilastokeskus

Tarkastellaan ensin muuttujan valintaa. Kuviossa 2 esitetään kuukausittaisen työttömyysasteen kehitys vuoden 2004 alusta vuoden 2014 huhtikuuhun. Tyypillisesti työttömyys vaihtelee vuodenaikojen mukaan, toisin sanoen työttömyyden kausivaihtelu on voimakasta. Valitulla ajanjaksolla 2004–2014 työttömyysasteessa erottuu myös laskeva trendi vuoteen 2008 asti, jonka jälkeen työttömyys nousi korkeammalle tasolle. Työttömyyden nousu liittyi vuonna 2008 alkaneeseen kriisiin.

Tämän raportin tavoite on selvittää, voiko Google-hakutilastoilla ennustaa työttömyyttä Suomessa. Käytettyjä Google-hakusanoja on kuitenkin lukemattomia. Haasteena on siten valita sopiva äärellinen määrä hakusanoja, joiden yhteyttä työttömyystilastoihin voidaan tarkastella. Selkeyden vuoksi tässä raportissa muodostetaan yksi Google Index, joka kuvaa yhtäaikaisesti useiden työttömyyteen liittyvien hakusanojen yleisyyttä muihin hakusanoihin nähden. Muodostettu Google Index toimii muuttujana, jonka ennustekykyä voidaan testata raportin esittelemässä yksinkertaisessa ekonometrisessa aikasarjamallissa. Muuttujan valintaan on monia tekniikoita. Tässä raportissa muuttujan valinta perustuu ennakkotietämykseen työnhausta sekä harkintaan.



Kuva 3: Työttömyysaste ja Google Index 2004–2014. Lähde: Tilastokeskus ja *Google Trends*

Kantava ajatus on se, että työttömyys ja työttömyyden uhka vaikuttaa henkilön tekemiin Internet-hakuihin. Tyypillisesti työttömäksi jäävä ottaa selvää työttömyyskorvauksesta ja hakee töitä Internetin hakukoneen avulla. Esimerkiksi kesäkuussa 2014 työttömyysturvaan liittyvistä hakemuksista 72 % tehtiin Internetissä.[12] Tämä tarkoittaa myös aiheeseen liittyviä Internet-hakuja, sillä sivustoille siirrytään usein hakukoneen avulla (Broder 2002). On siten luontevaa ajatella, että työtön tai työttömäksi jäävä hakee keskimääräistä useammin työttömyyteen liittyviä hakuja.

Tässä raportissa käytetty Google Index on muodostettu yhdistämällä kuusi työttömäksi jäämiseen liittyvää hakusanaa. Nämä hakusanat ovat työttömyyskorvaus, työttömyyspäiväraha, ansiopäiväraha, peruspäiväraha, ansiosidonnainen päiväraha ja työttömyyskassa. *Google Trends* -aineistoa käytetään tammikuusta 2004 toukokuun loppuun 2014. Työttömyysastetta on käytössä yksi kuukausi vähemmän, huhtikuuhun 2014 asti.

Kuvio 3 havainnollistaa muodostetun Google Indexin sekä työttömyysasteen kehitystä vuosina 2004–2014. Työttömyysaste ja Google Index näyttävät

[12]Kela 2014.

käyttäytyvän samankaltaisesti vuoteen 2011, jonka jälkeen ne erkanevat. Google Index ei näytä sisältävän selkeää kausivaihtelua, toisin kuin työttömyysaste. Muuttujien välinen korrelaatiokerroin on 0.38.

Muuttujan valinnan perusidea on seuraava. Yksittäiseen työttömyyteen yhteydessä olevaan hakusanaan liittyy työttömyyden lisäksi myös muihin syihin liittyvää vaihtelua. Voidaan kuitenkin olettaa, että usean eri työttömyyteen liittyvien hakusanojen *yhteinen* vaihtelu liittyy suurelta osin työttömyyteen. Siten käyttämällä samanaikaisesti useita eri työttömyyteen liittyviä hakusanoja esiin nousee pääosin työttömyyteen liittyvä vaihtelu hakuintensiteetissä. Yhteisen vaihtelun mallintamiseen on olemassa useita eri tekniikoita, mutta yksinkertaisuuden vuoksi tässä raportissa käytetään yhteenlaskettua hakuintensiteettiä. Käyttämällä samanaikaisesti useita sanoja pienennetään myös riskiä siitä, että mahdollisesti havaittu yhteys työttömyyden ja Internet-hakujen välillä liittyisi ainoastaan sopivasti valittuun muuttujaan.

Raportin valitut hakusanat liittyvät työttömyysturvaan, sillä nämä Internet-haut ovat luultavimmin ensimmäisten hakujen joukossa, joita työttömäksi jäävä tekee Googlella. Lisäksi voidaan olettaa, että työttömyyskorvaukseen liittyviä hakuja tekevät erityisesti työttömäksi jäävät. Toisin kuin työttömyyskorvaukseen liittyvät Google-haut, esimerkiksi työnhakuun liittyvät haut saattavat lisääntyä muistakin syistä kuin työttömyyden kasvuun liittyen. Myös aikaisempi Internet-hakuja käsittelevä kirjallisuus Englannista (McLaren & Shanbhogue 2011), Yhdysvalloista (Choi & Varian 2012) ja Saksasta (Askitas & Zimmermann 2009) ehdottaa, että työttömyyskorvauksen hakemiseen liittyvät hakusanat ovat potentiaalisesti hyviä työttömyyden ennustamisessa.

Käyttämällä useita hakusanoja saadaan myös kasvatettua otoskokoa. Vaikka Google ei raportoi tarkkaa hakujen lukumäärää vaan hakuintensiteetin, *Google Adwords* -palvelu tarjoaa arvion annetuilla hakusanoilla tehtyjen hakujen lukumäärästä. Toukokuun 2012 ja huhtikuun 2014 välisenä aikana valituilla työttömyyteen liittyvillä hakusanoilla tehtiin keskimäärin 18 900 hakua kuukaudessa. Tämä tarkoittaa, että tarkasteltavan otoksen taustalla on noin 2 300 000 Google-hakua.

Hakuaineistosta muodostettu Google Index on saatavissa viikoittain. Suomen virallinen työttömyysaste julkaistaan kuitenkin kuukausittain, ja tästä syystä Google Index on aggregoitu kuukausitasolle. Aggregointi on tehty laskemalla tietyn kuukauden hakuintensiteetti siihen kuuluvien viikkojen ha-

Muuttuja	n	μ	σ	min	max
Työttömyys (%)	124	7.883	1.250	5.2	11.6
Google Index	125	39.035	12.904	21.25	91.75

n = otoskoko, μ= keskiarvo, σ= keskihajonta, min = pienin arvo, max = suurin arvo

Taulukko 1: Työttömyysasteen ja Google Indexin tunnuslukuja 2004–2014. Lähde: Tilastokeskus ja *Google Trends*

kuintensiteettien keskiarvona. Viikko on yhdistetty kuukauteen sen alkamispäivän kuukauden perusteella.

Taulukkoon 1 on koottu joitakin kuvailevia tilastoja työttömyysasteesta ja muodostetusta Google Indexistä vuosina 2004–2014.

Työttömyydestä ja Google Indexistä käytetään kausitasoittamattomia arvoja kolmesta syystä. Ensimmäinen ja tärkein syy on kausitasoituksesta johtuva aikasarjan päätepisteisiin liittyvä epävarmuus. Kausitasoitetun aikasarjan viimeisimmät luvut perustuvat osittain ennusteisiin, ja ne voivat muuttua voimakkaastikin seuraavien arvojen ilmestyttyä. Koska raportin tavoite on selvittää, voiko Google-hakudatalla ennustaa nykyhetkeä, on perusteltua käyttää mahdollisimman ajankohtaisia ja tarkkoja arvoja. Toiseksi kuvion 3 perusteella Google Index ei vaikuta sisältävän selkeää kausivaihtelua, ja tästä syystä sen kausitasoittaminen ei välttämättä ole järkevää. On kuitenkin perusteltua joko tasoittaa molemmat sarjat tai ei kumpaakaan. Kolmas syy perustuu käytäntöön. Lyhyen aikavälin hahmottamisessa Suomessa käytetään usein kausitasoittamatonta työttömyysastetta, jota perinteisesti verrataan edellisen vuoden saman kuukauden arvoon.

Muuttujista käytetään tasoja peräkkäisten arvojen välisen erotuksen sijaan Choin ja Varianin (2012) esimerkin mukaisesti.

Tarkastellaan seuraavaksi menetelmän jälkimmäistä vaihetta, tilastollista testausta. Perusidea on muodostaa vertailukohdaksi yksinkertainen malli työttömyyden ennustamiseksi, lisätä tähän malliin Google-muuttuja ja verrata uuden ja vanhan mallin kykyä ennustaa työttömyyttä (Choi & Varian 2012).

Vertailukohdaksi asetetaan seuraava malli: tämän kuukauden työttömyyttä y_t ennustetaan viime kuun työttömyydellä y_{t-1} sekä

työttömyydellä 12 kuukautta sitten y_{t-12}. Kaikista arvoista on otettu logaritmit. Vertailukohtainen malli (0) esitetään yhtälössä (2). Muuttuja e_t on virhetermi. Tämän tyyppinen malli tunnetaan kirjallisuudessa kausivaihtelu-autoregressiivisenä mallina tai kausivaihtelu-AR-mallina.[13] Malliin on lisätty selittäväksi tekijäksi 12 kuukauden takainen työttömyysaste, jotta mahdollisesti havaittu yhteys Google-muuttujan ja työttömyyden välissä ei liittyisi vain yhteiseen kausivaihteluun.

Vertailukohtainen malli on selvyyden vuoksi mahdollisimman yksinkertainen. Akaiken ja Schwarzin informaatiokriteerien nojalla yhden muuttujan malleista AR(15)-malli olisi selittänyt parhaiten työttömyyden vaihtelua vuosina 2004–2014. Selkeyden ja lyhyen tarkasteluvälin vuoksi käytetään kuitenkin yksinkertaisempaa mallia. Alle 13. asteen autoregressivisten mallien joukossa vertailukohdaksi asetettu malli on informaatiokriteerien nojalla paras. Yksinkertainen vertailukohta on sikäli myös järkevä, että AR(1)-malli on pärjännyt hyvin myös verrattuna suomalaisten ennustelaitosten ennusteisiin. Tarkasteltaessa yleisestä ennustetarkkuutta vuonna 2013 vain kaksi ennustelaitosta tuotti tarkemman ennusteen Suomessa taloudelle kuin AR(1)-malli olisi tuottanut.[14]

Seuraavaksi malliin lisätään kuukausitasolle muodostettu Google Index, joka kuvaa tiettyjen työttömyysturvaan liittyvien hakusanojen yleisyyttä kullakin ajanhetkellä. Merkitään Google Indexiä muuttajalla x_t. Google Indexistä on käytettävissä aina nykyhetken t arvo, kun taas työttömyydestä on saatavilla vain aikaisintaan kuukauden takaisia arvoja hetkeltä $t-1$. Tämä ero julkaisuviiveessä on keskeisin syy, miksi Google Indexistä voisi olla hyötyä nykyhetken ennustamisessa. Google Indexillä täydennetty malli (1) esitetään yhtälössä (3).

$$\text{Malli (0):} \quad log(y_t) \sim \quad log(y_{t-1}) + log(y_{t-12}) + e_t \tag{2}$$

$$\text{Malli (1):} \quad log(y_t) \sim \quad log(y_{t-1}) + log(y_{t-12}) + x_t + e_t \tag{3}$$

Mallien estimoinnin jälkeen vertaillaan niiden selitysasteita ja informaatiokriteereitä, parametrien merkitsevyyttä sekä erityisesti vertaillaan mallien tuottamia nykyhetken ja lähitulevaisuuden ennusteita käyttäen mittarina absoluuttista keskivirhettä.

[13] engl. *seasonal AR model*

[14] Talousennustajakilpailu, Jyväskylän yliopiston kauppakorkeakoulu, 2014

Nykyhetken ennusteita verrataan muodostamalla molempien mallien ennusteet askeleeksi eteenpäin, toisin sanoen nykyhetkeksi t, *"rolling window"* -menetelmällä. Mallia opetetaan ensin 48 kuukauden aineistolla, minkä jälkeen ennustetaan 49. havainnon saama arvo käyttäen 49. havaintoa Google Indexistä ja 48. sekä 36. havaintoa työttömyydestä. 48 havainnon ikkunaa siirretään aina askel eteenpäin ja mallin parametrit estimoidaan joka kerta uudelleen. Ennusteita vertaamalla voidaan arvioida, auttaako Google aineisto ennustamaan työttömyyttä. Ennusteen ja toteutuneen arvon välisen erotuksen keskiarvo tunnetaan absoluuttisena keskivirheenä.[15] Absoluuttinen keskivirhe on määritelty yhtälön (4) mukaisesti.

$$MAE = \frac{1}{T}\sum_{t=1}^{T} \mid E_t \mid \tag{4}$$

jossa

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

Lähitulevaisuuden ennusteita verrataan muodostamalla ennusteet vain saatavilla olevien arvojen perusteella. Tarkasti ottaen muodostetaan edellä kuvatulla *"rolling window"*-menetemällä dynaamiset ennusteet (Hamilton 1994) n periodiksi eteenpäin. Mallin (1) tapauksessa malli estimoidaan käyttäen lisäksi Google Indexistä ennusteperiodin n lähintä saatavilla olevaa arvoa x_{t-n}. Esimerkiksi ennuste kahdeksi kuukaudeksi eteenpäin, eli hetkeksi $t+2$, muodostetaan käyttämällä hetken $t-1$ työttömyysastetta, $t-12$ työttömyysastetta sekä hetken t Google Indexin arvoa. Estimointi tapahtuu suurimman uskottavuuden menetelmällä käyttäen Kalman-suodinta.

Lopuksi suoritaan Granger-kausaalisuustesti ja tarkastellaan muuttujien välisiä ristikorrelaatioita.

5 Tulokset

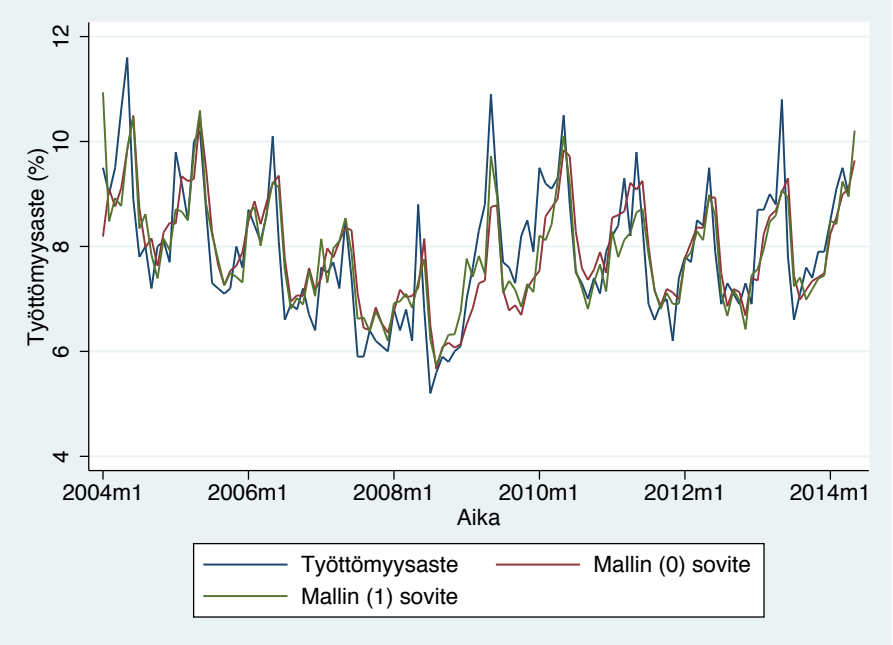
Mallin (0) ja mallin (1) tulokset esitetään taulukossa 2. Google Indeksin kerroin on merkitsevä 1 prosentin merkitsevyystasolla, ja sen positiivinen kerroin tarkoittaa sitä, että työttömyysturvaan liittyvät haut ovat positiivisesti yhteydessä työttömyysasteeseen. Tarkemmin, Google Indexin kerroin 0.00564

[15] engl. *mean absolute error* [MAE]

Malli	(0)	(1)
Selittäjät		
$log(y_{t-1})$	0.513**	0.371**
	(0.0614)	(0.0584)
$log(y_{t-12})$	0.413**	0.564**
	(0.0553)	(0.0558)
x_t		0.00564**
		(.000877)
Vakio	2.102**	1.874**
	(0.0821)	(0.0798)
Yhteenveto		
R^2	0.61	0.70
AIC	-223	-254
BIC	-211	-240
n	124	124

Taulukon yksittäinen kerroin on merkitsevä **1 % merkitsevyystasolla käyttäen kaksisuuntaista testiä.

Taulukko 2: Mallien tulokset



Kuva 4: Työttömyysaste ja mallien sovitteet 2004–2014

tarkoittaa sitä, että 1 prosentin kasvu hakuintensiteetissä on yhteydessä noin 0.5 prosentin kasvuun työttömyydessä.

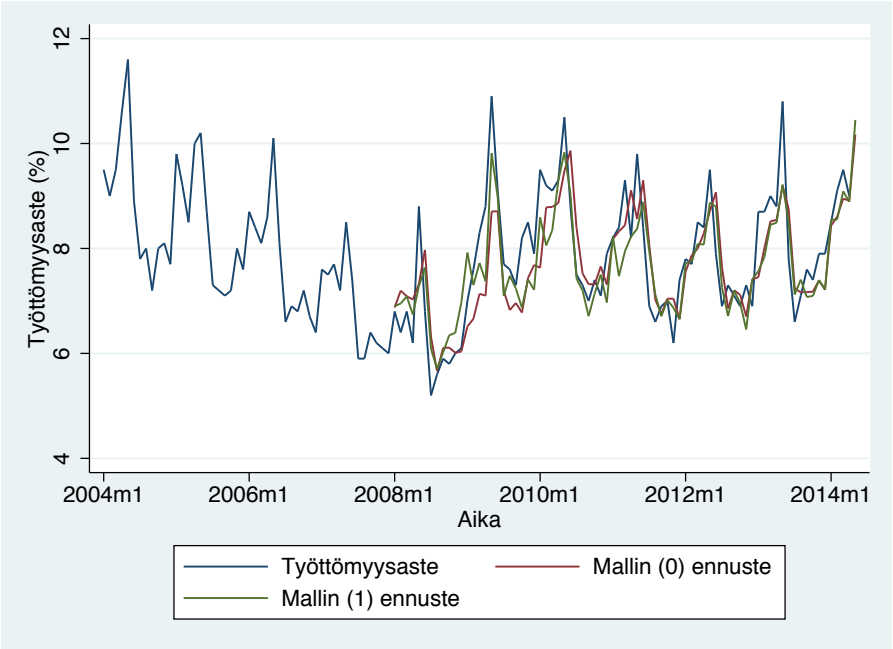
Mallin (0) selitysaste on 0.61, mikä tarkoittaa sitä, että vertailukohtana toimiva malli pystyy selittämään jo yksinään melko suuren osan työttömyyden vaihtelusta. Google Indexin sisällyttäminen malliin parantaa kuitenkin mallin selitysastetta 15 prosenttia, 0.61:stä 0.70:neen. Selitysasteen parantumisen tulkintaan liittyvät riskit ovat kuitenkin hyvin tunnettuja.

Google Indexin sisällyttäminen malliin (0) pienentää kuitenkin myös mallin sekä Akaiken [AIC] että Schwarzin [BIC] informaatiokriteerien saamia arvoja. Tämä viittaa siihen, että Google-haut sisältävät hyödyllistä informaatiota työttömyysasteen vaihtelun selittämiseksi.

Kuviossa 4 esitetään mallien (0) ja (1) sovitteet sekä työttömyysasteen kehitys Suomessa 2004–2014. Google-hakujen lisääminen malliin vaikuttaa tarkentavan sovitetta, kuten parantunut selitysaste antaa ymmärtää. Sovite on kuitenkin eri asia kuin ennuste. Tarkka sovite ei välttämättä tarkoita tarkkaa ennustetta.

Auttaako Google-aineisto ennustamaan työttömyyttä?

Selvitetään aluksi, auttaako Google-aineisto ennustamaan *nykyhetken*



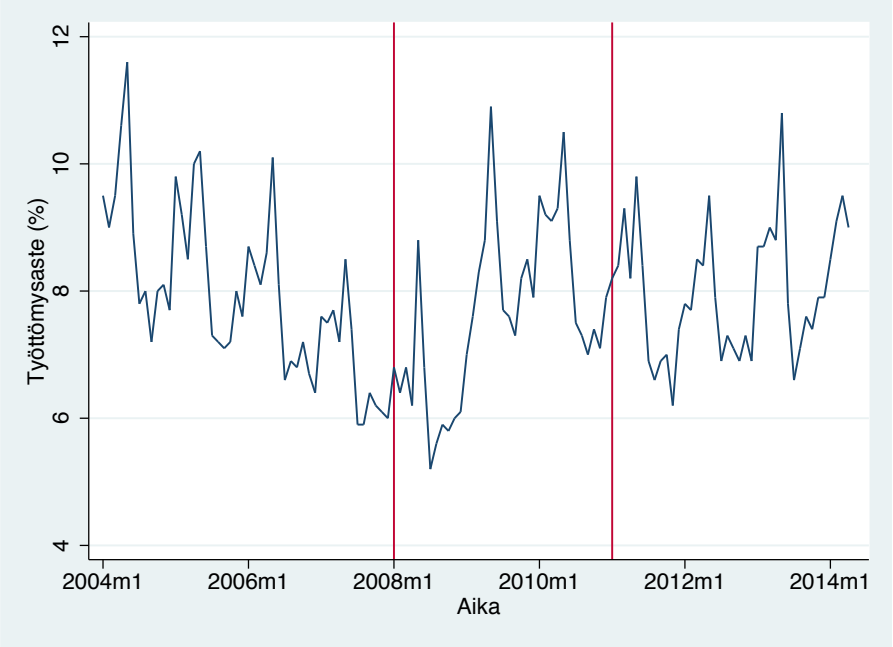
Kuva 5: Työttömyysaste ja mallien ennusteet 2004–2014

työttömyyttä, sillä Google-haut ovat tyypillisesti vain yhteydessä talouden nykytilaan, eivät välttämättä tulevaisuuteen. Google-aineiston reaaliaikaisuuden ansiosta tästäkin voi olla hyötyä, sillä viralliset työttömyystilastot julkaistaan kuukauden viiveellä. Ennustekyvyn tarkastelemiseksi käytetään edellisessä luvussa esitettyä *"rolling window"* -menetelmää. Menetelmän avulla saadut ennusteet esitetään kuviossa 5. Mallien (0) ja (1) tuottamien nykyhetken ennusteiden absoluuttiset keskivirheet esitetään taulukossa 3. Mallin (0) tuottaman nykyhetken absoluuttinen keskivirhe vuosina 2008–2014 on 7.8 prosenttia, kun taas mallin (1) absoluuttinen keskivirhe

Malli	(0)	(1)
MAE	7.8 %	7.0 %

MAE = absoluuttinen keskivirhe

Taulukko 3: Mallin (0) ja (1) nykyhetken ennustetarkkuus



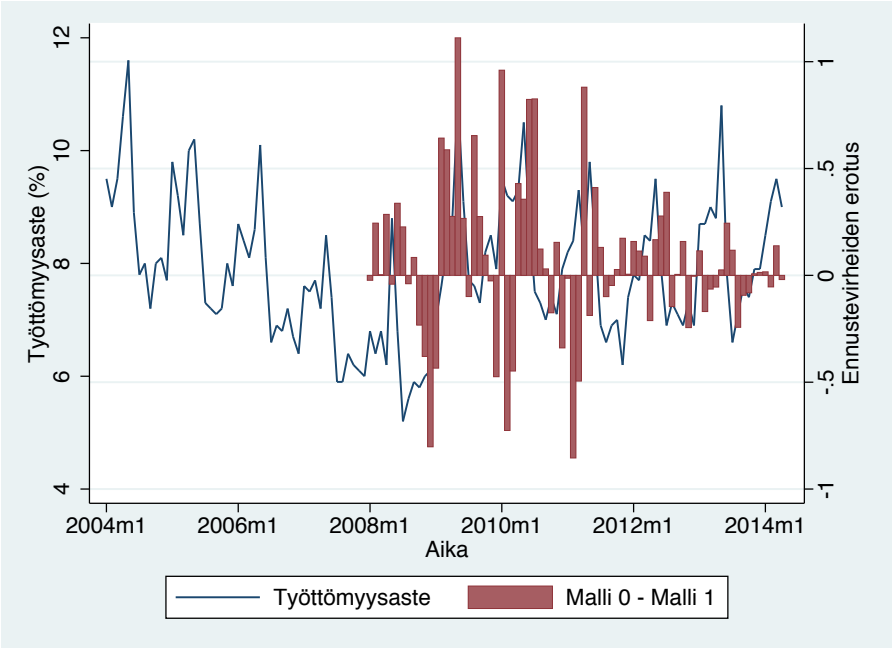
Kuva 6: Käännekohta

on 7.0 prosenttia. Google Indexin sisällyttäminen malliin parantaa siten absoluuttisella keskivirheellä mitattuna ennustetarkkuutta 10 prosenttia.

Käännekohtien ennustaminen on selvästi vaikeampaa kuin tasaisen kehityksen ennustaminen. Kansantalouden ennustamisessa käännekohdat ovat kuitenkin erityisen kiinnostavia. Onko Google-aineistosta erityistä hyötyä käännekohtien löytämisessä?

Käsiteltävällä ajanjaksolla selkein käännekohta on kuviossa 6 näkyvä työttömyyden kasvu ja supistuminen vuosien 2008 ja 2011 välillä. Jos tarkastellaan vain käännekohdan sisältäviä vuosia 2008–2011, Google Indexin sisällyttäminen malliin pienentää nykyhetken ennusteen absoluuttista keskivirhettä 15 prosenttia, 8.9 prosentista 7.5 prosenttiin. Google-aineistosta vaikuttaa siis olevan erityistä hyötyä käännekohtien löytämisessä.

Google Indexin kuukauden mittaisen etumatkan vuoksi tulos on arkijärjen mukainen. Pelkästään edellisten kuukausien informaatiota käyttäen tulevan käännekohdan ennustaminen on hyvin epävarmaa. Jos kuitenkin työttömyyden ja Internet-hakujen välillä on yhteys, *nykyhetken* Google Index voi auttaa paljastamaan, onko työttömyysaste juuri nyt huomattavasti erilainen kuin viime kuussa. Esimerkiksi se, että hakusanalla ”työttömyyskorvaus”



Kuva 7: Mallin (0) ja mallin (1) ennusteiden erotus 2008–2014 sekä työttömyysaste 2004–2014

on haettu selvästi enemmän kuin edellisenä vuonna samaan aikaan, saattaa paljastaa uuden käänteen, jota pelkästään menneisyyden informaatio, esimerkiksi edellisten kuukausien työttömyysluvut eivät kerro.

Vaikka malli (1) tuottaa keskimäärin vertailukohtaa (0) paremman ennusteen, näin ei ole kuitenkaan aina. Kuviossa 7 esitetään vertailukohtana toimivan mallin (0) ja Google Indexillä täydennetyn mallin (1) tuottamien nykyhetken ennusteiden välinen erotus. Kuvion arvo on positiivinen niinä hetkinä, joina malli (1) tuotti tarkemman ennusteen, ja negatiivinen silloin, kun malli (0) tuotti tarkemman ennusteen. Kuvioon on lisätty selkeyden vuoksi myös työttömyysasteen kehitys vuosina 2004–2014. Virheet alkavat kuvion 5 tapaan vuodesta 2008, sillä ensimmäistä ennustetta varten mallia opetetaan ensimmäiset 48 kuukautta. Kuviosta nähdään, että Google Indexillä täydennetty malli (1) tuotti tarkemman ennusteen usein työttömyyden huippujen aikana ja heti niiden jälkeen, kun taas vertailukohtana toimiva malli (0) tuotti tarkemman nykyhetken ennusteen ennen työttömyyshuippuja. Mallien yksinkertaisuuden vuoksi tästä ei tule välttämättä tehdä kovin pitkälle vieviä johtopäätöksiä.

Tähän mennessä olemme käsitelleet vain nykyhetken ennustamista. Tarkastellaan seuraavaksi kuinka hyvin Google-haut ennustavat *lähitulevaisuutta.* Taulukkoon 4 on koottu mallin (0) ja mallin (1) tuottamien nykyhetken ja tulevaisuuden ennusteiden absoluuttiset keskivirheet sekä ennustetarkkuuden muutos joka kuukaudelle puoleksi vuodeksi eteenpäin. Nykyhetken t ennusteiden absoluuttiset keskivirheet on jo aiemmin esitetty taulukossa 3. Taulukosta 4 nähdään, että vertailukohtaisen mallin (0) ennustetarkkuus pienenee ymmärrettävästi, mitä kauemmas tulevaisuuteen se tehdään. Huomionarvoista on kuitenkin, että Google Indexillä täydennetyn mallin (1) tuottamat ennusteet pysyvät lähes yhtä tarkkoina jopa neljä kuukautta eteenpäin. Mallin (1) ennuste kolmeksi kuukaudeksi eteenpäin on jopa tarkempi kuin nykyhetkeksi tehty ennuste. Google Indexin sisällyttäminen vertailukohtana toimivaan malliin parantaa ennustetarkkuutta kolmen kuukauden päähän 39 % absoluuttisella keskivirheellä mitattuna.

Samankaltaisesti kuin edellä taulukossa 4 Google-hakujen kykyä ennustaa tulevaisuutta voi tarkastella myös tutkimalla, kuinka monen kuukauden takaiset Google Indexin arvot tarjoavat hyödyllistä informaatiota nykyhetken työttömyysasteesta malliin (0). Mikäli esimerkiksi kahden kuukauden takainen Google Index tarjoaa hyödyllistä informaatiota nykyhetkestä, nykyhetken Google Index tarjoaa hyödyllistä informaatiota työttömyysasteesta kahden kuukauden päästä. Sekä Akaiken että Schwarzin informaatiokriteerit suosittelevat ottamaan vertailukohtaiseen malliin (0) nykyhetken sekä neljän edellisen kuukauden Google Indexin arvot. Tulos on samansuuntainen taulukon 3 havainnon kanssa. Laajennetun mallin (2) tulokset on koottu taulukkoon 5 edellisten tulosten rinnalle.

Autoregressiivisten mallien vertailuun perustuvan Choin ja Varianin (2012) esittämän menetelmän lisäksi Google-hakujen sisältämän informaation hyödyllisyyttä voidaan tarkastella kahden menetelmän, ristikorrelaatiofunktion sekä Granger-kausaalisuustestin, avulla.

Tarkastellaan ensin työttömyysasteen ja Google Indexin eri viiveiden välisiä korrelaatioita. Taulukossa 6 esitetään työttömyysasteen ja Google Indexin välisen ristikorrelaationfunktion[16] arvoja.

Silmiinpistävin havainto on, että tulevaisuuden työttömyyden ja nykyhetken Google Indexin väliset korrelaatiokertoimet näyttävät suuremmilta kuin päinvastaisessa tapauksessa. Vastaava ilmiö havaitaan myös osakemark-

[16] engl. *cross correlation function* [CCF]

	Malli	MAE	Δ
t	(0)	7.8 %	10.0 %
	(1)	7.0 %	
$t+1$	(0)	9.3 %	16.9 %
	(1)	7.7 %	
$t+2$	(0)	10.5 %	32.9 %
	(1)	7.0 %	
$t+3$	(0)	11.1 %	39.2 %
	(1)	6.7 %	
$t+4$	(0)	11.3 %	30.5 %
	(1)	7.7 %	
$t+5$	(0)	11.3 %	25.3 %
	(1)	8.4 %	
$t+6$	(0)	11.4 %	20.5 %
	(1)	9.0 %	

MAE = absoluuttinen keskivirhe

Δ = ennustetarkkuuden muutos

Taulukko 4: Mallin (0) ja (1) nykyhetken ja tulevaisuuden ennustetarkkuus

Malli	(0)	(1)	(2)
Selittäjät			
$log(y_{t-1})$	0.513**	0.371**	0.319**
	(0.0614)	(0.0584)	(0.0630)
$log(y_{t-12})$	0.413**	0.564**	0.632**
	(0.0553)	(0.0558)	(0.0.628)
x_t		0.00564**	0.00233
		(.000877)	(0.0013)
x_{t-1}			-0.0011
			(0.0013)
x_{t-2}			0.0028*
			(0.0013)
x_{t-3}			0.0014
			(0.0013)
x_{t-4}			0.0034*
			(0.0014)
Vakio	2.102**	1.874**	1.753**
	(0.0821)	(0.0798)	(0.089)
Yhteenveto			
R^2	0.61	0.70	0.79
AIC	-223	-254	-279
BIC	-211	-240	-254
n	124	124	118

Taulukon yksittäinen kerroin on merkitsevä **1 %- ja *5 % -merkitsevyystasolla käyttäen kaksisuuntaista testiä.

Taulukko 5: Mallien (0), (1) ja (2) tulokset

δ	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
CCF	0.20	0.38	0.59	0.51	0.46	0.36	0.35	0.15	0.09	0.17	0.09	0.05	0.01

n = 113, δ = viive, CCF = ristikorrelaatiofunktion arvo

Taulukko 6: Työttömyysasteen ja Google Indexin välinen ristikorrelaatiofunktio

kinoilla (Bordino et al. 2012) sekä asuntomarkkinoilla (Wu & Brynjolfsson 2014). Huomioitavaa on myös se, että Google Indexin ja työttömyysasteen välinen korrelaatio kasvaa Google Indexin neljänteen viivästettyyn arvoon asti. Tämä tarkoittaa sitä, että esimerkiksi kesäkuun työttömyyden kanssa ei korreloi parhaiten kesäkuun Google Index vaan pikemminkin helmikuun Google Index. Laajemmin tämä viittaa siihen, että nykyhetken ennustamisen lisäksi Google-haut myös ennakoivat *tulevaa* työttömyyttä. Ennakoivuutta ei tule kuitenkaan sekoittaa kausaalisuuteen. Taulukon 4 korrelaatiofunktion arvo CCF = 0.35 kun $\delta = 0$ poikkeaa hieman aikaisemmin esitetystä muuttujien välisestä korrelaatiosta, sillä taulukko on laskettu käyttäen vain niitä arvoja, joita on voitu käyttää kaikkien arvojen laskemiseen.

Seuraavaksi tarkastellaan Granger-kausaalisuustestin avulla Google Indexin ja työttömyysasteen välistä yhteyttä. Taulukko 7 tiivistää Granger-kausaalisuustestin tulokset. Granger-kausaalisuustesti on toteutettu käyttäen ensimmäisen asteen VAR-mallia sekä neljännen asteen VAR-mallia, joka on Schwarzin informaationkriteerin nojalla melko hyvä yksinkertaistus korkeamman asteen mallista. Ensimmäisen asteen VAR-malliin pohjautuvan Granger-kausaalisuustestin mukaan 5% merkitsevyystasolla Google Indexin edellisen kuukauden arvot tarjoavat hyödyllistä informaatiota nykyhetken työttömyydestä. Ensimmäisen asteen mallia käyttäen edellisen kuukauden työttömyysaste ei kuitenkaan välttämättä sisällä hyödyllistä informaatiota nykyhetken Google-hakuintensiteetin ennustamiseksi. Tulos on samansuuntainen kuin ristikorrelaatioita tarkasteltaessa. Työttömyyskorvaukseen liittyvien Google-hakujen yleisyyden voi tulkita olevan työttömyyttä ennakoiva indikaattori. Kuitenkin käyttäen neljännen asteen VAR-mallia, emme voi jättää voimaan kumpaakaan 0-hypoteesia Granger-ei-kausaalisuudesta. Tämä on luontevaa, sillä muuttujat ovat pohjimmiltaan endogeenisia: samat talouden ilmiöt ovat yhteydessä työttömyyteen sekä työttömyyteen liitty-

0-hypoteesi							
1. asteen VAR				4. asteen VAR			
$y \nrightarrow x$		$x \nrightarrow y$		$y \nrightarrow x$		$x \nrightarrow y$	
χ^2	p-arvo	χ^2	p-arvo	χ^2	p-arvo	χ^2	p-arvo
3.33	0.068	3.91	0.048*	30.6	<0.001***	30.9	<0.001***

y = työttömyysaste, x = Google Index

Taulukon symbolit * ja ** tarkoittavat merkitsevyyttä *5 %- ja ***0.1 %-tasolla käyttäen kaksisuuntaista testiä, tarkoittaen että Granger-ei-kausaalisuus hylätään.

Taulukko 7: Granger-kausaalisuustesti

viin Google-hakuihin. Vielä kerran, endogeenisuudesta huolimatta Google-hakuintensiteetistä kannattaa olla kiinnostunut, sillä se on käytettävissä miltei kuukauden aikaisemmin kuin työttömyysaste ja lisäksi se vaikuttaa ennakoivan työttömyyttä.

6 Pohdinta

Tulokset viittaavat siihen, että Google-haut sisältävät hyödyllistä informaatiota työttömyyden ennustamiseksi. Miten tätä tietoa tulisi hyödyntää?

Olemme ETLAssa rakentaneet automaattisen ennustemallin ETLAnow. ETLAnow-malli tuottaa automaattisesti päivittäin ennusteen nykyhetken ja lähitulevaisuuden työttömyysasteesta. Se hyödyntää reaaliaikaista tietoa Suomessa tehdyistä työttömyyteen liittyvistä Google-hauista sekä tuoreimpia Tilastokeskuksen tietoja työttömyydestä.

ETLAnow pyrkii hyödyntäen Big Dataa, tässä tapauksessa nykyhetken Google-hakuja, tarjoamaan nopean ja ajankohtaisen tilannekuvan työttömyydestä päätöksenteon tueksi. Yksinkertaisesti se vastaa kysymykseen: Mikä on työttömyysaste juuri nyt? Lisäksi ETLAnow arvioi Suomessa tehtyjen Google-hakujen perusteella työttömyysasteen kehitystä lähikuukausina.

Tarkasta tilannekuvasta on erityisesti hyötyä, kun kuva nykyhetkestä ja lähitulevaisuusta ei ole täysin selvä. Tällöin ETLAnow voi antaa signaalin

tulevista käännekohdista. Käytännön työkalun lisäksi ETLAnow on myös kokeilu siitä, kuinka nykyhetken ja lähitulevaisuuden ennusteita voidaan automatisoida.

Internet-hakuja koskevat tilastot voivat potentiaalisesti auttaa hahmottamaan uusilla tavoilla työmarkkinoiden toimintaa. Lyhyesti: Internet-haut kertovat ihmisten käyttäytymisestä. Esimerkiksi tieto työnhakua koskevien Google-hakujen yleisyydestä saattaa tarjota hyödyllistä tietoa työnhakuintensiteetistä. Tästä tiedosta taas voi olla hyötyä tarkasteltaessa esimerkiksi työmarkkinapolitiikan toimivuutta. Myös hallinnollisten uudistusten ennakointivaikutusten selvittämisessä Internet-hakutilastoista saattaa olla hyötyä.

Google Trends on valtava reaaliaikainen tietokanta, jota ei kannata jättää käyttämättä.

Mallien yksinkertaisuuden vuoksi turhan pitkälle menevien päätelmien tekemistä tulee välttää. Pääviesti on kuitenkin se, että Google-haut sisältävät hyödyllistä informaatiota työttömyyden ennustamiseksi.

7 Yhteenveto

Yksinkertainen kausivaihtelun huomioivalla muuttujalla täydennetty AR(1)-malli, johon on liitetty Google-hakujen yleisyyttä kuvaava muuttuja, ennustaa paremmin nykyhetken sekä lähitulevaisuuden työttömyyttä Suomessa kuin sama malli ilman Google-muuttujaa. Havaitaan myös, että reaaliaikaisesta Google Indexistä on erityisesti hyötyä käännekohtien ennustamisessa. Lisäksi täydennetyn mallin informaatiokriteerien arvot myös ovat pienemmät kuin vertailukohdan. Mallien vertailun lisäksi tarkastellaan muuttujien välisiä ristikorrelaatioita ja suoritetaan Granger-kausaalisuustesti. Molemmat menetelmät viittaavat siihen, että Google-haut sisältävät hyödyllistä informaatiota myös lähitulevaisuuden työttömyysasteesta.

Raportti perustuu Choin ja Varianin (2012) esittämään menetelmään.

Nykyhetken ennustetarkkuuden parantumista koskeva tulos on yhdenmukainen aikaisemman kirjallisuuden kanssa. Esimerkiksi Askitas ja Zimmermann (2009) sekä Suhoy (2009) havaitsevat, että Google-haut sisältävät hyödyllistä informaatiota työttömyyden ennustamiseksi Saksassa (Askitas & Zimmermann 2009) ja Israelissa (Suhoy 2009). Tulevaisuuden ennustamista koskeva tulos työttömyyden yhteydessä on tiettävästi uusi. Raportti käsittelee ensimmäisenä Suomessa Google-hakujen käyttöä ennustamisessa.

Google-hakuja koskevaan aineistoon liittyy vielä haasteita. Tulosten mukaan nykyisessäkin muodossa Google-hakuaineistosta on hyötyä nykyhetken ja potentiaalisesti lähitulevaisuuden työttömyyden ennustamisessa.

Yhteenvetona todettakoon, että yhä suurempi osa ihmisten toiminnasta on Internetissä (Edelman 2012, Einav & Levin 2013). Tästä syystä on tärkeää tarkastella Internet-aineiston tarjoamia mahdollisuuksia. Big Data, esimerkiksi Google-hakuaineisto, sisältää yhä paljon uutta informaatiota, josta on hyötyä lähitulevaisuudessa.

Raportin tulos on, että Google-haut ennustavat työttömyyttä Suomessa. Muiden taloudellisten indikaattorien ohella Google-hakujen hyödyntäminen tarkentaa potentiaalisesti nykyhetken ja lähitulevaisuuden ennustetta muutamia prosentteja. Se ei ehkä ole paljon, mutta mahdollisuutta ei kuitenkaan kannata jättää käyttämättä.

Viitteet

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

Askitas, N. & Zimmermann, K. F. (2011). Detecting Mortgage Delinquencies. *IZA Discussion Paper 5895*.

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

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

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

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

Choi, H. & Varian, H. R. (2009a). Predicting Initial Claims for Unemployment Benefits. *Technical Report, Google*.

Choi, H. & Varian, H. R. (2009b). Predicting the Present with Google Trends. *Technical Report, Google*.

Choi, H. & Varian, H. R. (2011). Predicting the Present with Google Trends. *Presentation at SF Fed*, March 18.

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

Curme, C., Preis, T., Stanley, H. E., & Moat, H. S. (2014). Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences of the United States of America*, (Early edition, doi: 10.1073/pnas.1324054111).

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

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

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

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

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4), 665–676.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012–14.

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

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

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

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

Koop, G. & Onorante, L. (2013). Macroeconomic Nowcasting Using Google Probabilities. *Working Paper*.

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

Kuhn, P. & Mansour, H. (2014). Is Internet Job Search Still Ineffective? *forthcoming in The Economic Journal*.

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

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

McLaren, N. & Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, 2011(1), 134–140.

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

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

Scott, S. L. & Varian, H. R. (2014a). Bayesian Variable Selection for Nowcasting Economic Time Series. In A. Goldfarb, S. Greenstein, & C. Tucker (Eds.), *forthcoming in Economics of Digitization*. University of Chicago Press.

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

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

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

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

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

Wu, L. & Brynjolfsson, E. (2014). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. In A. Goldfarb, S. Greenstein, & C. Tucker (Eds.), *forthcoming in Economics of Digitization*. University of Chicago Press.

Aikaisemmin ilmestynyt ETLA Raportit-sarjassa (ennen ETLA Keskusteluaiheita)

Previously published in the ETLA Reports series (formerly ETLA Discussion Papers)

No 15 *Olavi Rantala*, Kilpailukyvyn mittaamisen teoriaa ja käytäntöä. 14.8.2013. 29 s.

No 16 *Jyrki Ali-Yrkkö – Petri Rouvinen*, Implications of Value Creation and Capture in Global Value Chains. Lessons from 39 Grassroots Cases. 19.8.2013. 20 p.

No 17 *Martti Kulvik – Marja Tähtinen – Pekka Ylä-Anttila*, Business and Intellectual Capital Development in Financial Riptide. Case Studies of Finnish Biotechnology and Pharmaceutical Companies Dispersing into Global Value Chains. 15.10.2013. 82 p.

No 18 *Olavi Rantala*, Postitoiminnan kehitys vuoteen 2020. 18.11.2013. 22 s.

No 19 *Heli Koski*, Yleispalveluvelvoitteen merkitys postin kannattavuudelle. 18.11.2013. 15 s.

No 20 *Tuomo Virkola*, Exchange Rate Regime, Fiscal Foresight and the Effectiveness of Fiscal Policy in a Small Open Economy. 3.3.2014. 62 p.

No 21 *Ville Kaitila – Tuomo Virkola*, Openness, Specialisation and Vulnerability of the Nordic Countries. 27.3.2014. 25 p.

No 22 *Mika Maliranta – Niku Määttänen*, Innovation, Firm Risk and Industry Productivity. 1.4.2014. 14 p.

No 23 *Olavi Rantala*, Saksan ja muun euroalueen kilpailukyvyn ero eurokriisin taustalla. 1.4.2014. 23 s.

No 25 *Cinzia Alcidi – Daniel Gros*, Implications of EU Governance Reforms: Rationale and Practical Application. 6.5.2014. 26 p.

No 26 *Antti Suvanto – Kimmo Virolainen*, Mihin pankkiunionia tarvitaan? 7.5.2014. 21 s.

No 27 *Topias Leino – Jyrki Ali-Yrkkö*, How Does Foreign Direct Investment Measure Real Investment by Foreign-owned Companies? Firm-level Analysis. 15.5.2014. 25 p.

No 28 *Timo Nikinmaa*, Kone- ja metallituoteteollisuuden visio 2025. 23.5.2014. 52 s.

No 29 *Antti Pelkonen – Duncan A. Thomas – Terttu Luukkonen*, Project-based Funding and Novelty in University Research – Findings from Finland and the UK. 12.6.2014. 18 p.

No 30 *Antti Kauhanen*, Tulevaisuuden työmarkkinat. 6.8.2014. 16 s.

Sarjan julkaisut ovat raportteja tutkimustuloksista ja väliraportteja tekeillä olevista tutkimuksista.

Julkaisut ovat ladattavissa pdf-muodossa osoitteessa: www.etla.fi » julkaisut » raportit

Papers in this series are reports on research results and on studies in progress.

Publications in pdf can be downloaded at www.etla.fi » publications » reports

ETLA
Elinkeinoelämän tutkimuslaitos
The Research Institute of the Finnish Economy
Lönnrotinkatu 4 B
00120 Helsinki

Puh. 09-609 900
Fax 09-601 753
www.etla.fi
etunimi.sukunimi@etla.fi

ISSN-L 2323-2447, ISSN 2323-2447, ISSN 2323-2455 (Pdf)