ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY

ETLA Raportit
ETLA Reports

14.12.2016

No 63

Google-haut Suomen asuntojen hintojen ennustajana

Joona Widgrén*

* ETLA – Elinkeinoelämän tutkimuslaitos, joona.widgren@helsinki.fi

Suositeltava lähdeviittaus: Widgrén, Joona (14.12.2016). "Google-haut Suomen asuntojen hintojen ennustajana". ETLA Raportit No 63. https://pub.etla.fi/ETLA-Raportit-Reports-63.pdf

Kiitän seuraavia henkilöitä käydyistä keskusteluista sekä kommenteista: Markku Lehmus (Etla), Markku Kotilainen (Etla), Petri Rouvinen (Etlatieto/Etla), Joonas Tuhkuri (MIT/Etla) sekä seminaariosallistujia Etlassa.

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

Sisällysluettelo

Google-haut Suomen asuntojen hintojen ennustajana

Tiivistelmä

Tässä raportissa selvitetään, voidaanko Google-hauilla ennustaa asuntojen hintoja Suomessa. Google-hauista kertovan muuttujan lisääminen yksinkertaiseen perusmalliin parantaa ennustetta asuntojen nykyhetken hinnoista 7,5 prosentilla. Google-hakujen lisääminen parantaa ennustetta myös lähitulevaisuuden asuntohinnoista. Asuntojen hintojen nykyhetken sekä lähitulevaisuuden mahdollisimman tarkka ennakointi antaa monille toimijoille, kiinteistönvälittäjistä poliittisiin päätöksentekijöihin, paremmat eväät päätöksentekoa varten.

Asiasanat: Google Trends, Internet, nykyhetken ennustaminen, ennustaminen, asuntomarkkinat, aikasarja-analyysi

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

Predicting housing prices with Google searches in Finland

Abstract

This report examines whether Google search queries can be used to predict the present and the near future house prices in Finland. Compared to a simple benchmark model, Google searches improve the prediction of the present house price index by 7.5 % measured by mean absolute error. In addition, search queries improve the forecast of near future house prices. Predicting the present and near future house prices is relevant information to many agents, such as realtors and political decision makers.

Key words: Google Trends, Internet, nowcasting, forecasting, housing market, time series

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

1 Johdanto

Asunnon ostaminen on useimmille elämän suurin yksittäinen ostos, ja asunnot muodostavatkin kotitalouksien suurimman yksittäisen varallisuusesineluokan (Juntto, 2008). Asuntojen hinnat kiinnostavat taloustieteilijöiden lisäksi esimerkiksi yksittäisiä kuluttajia, asuntosijoittajia ja kiinteistönvälittäjiä. Kärjistäen emme kuitenkaan tiedä asuntojen tämän hetkisiä hintoja, sillä viralliset tilastot asuntohinnoista julkaistaan noin kuukausi jokaisen tarkasteluperiodin päättymisen jälkeen. Tilastot asuntokaupasta julkaistaan hieman aineistosta riippuen kuukausittain tai neljästi vuodessa, joten ajankohtaisen tiedon saaminen asuntomarkkinoista on usein hankalaa. Tietojen julkaisusta aiheutuvan viiveen takia asuntomarkkinoiden nykyhetken ja lähitulevaisuuden ennakoiminen onkin tärkeää monille päätöksentekijöille. Voisiko ihmisten Internetissä tekemistä hauista olla apua asuntomarkkinoiden ennakoimisessa?

Asuntokauppoja tehdään harvoin hetken mielijohteesta; tarkka asuntojen hintojen ja ominaisuuksien vertailu kuuluu usein ostoa edeltävään prosessiin (Horrigan & Vitak, 2008). Ottaen huomioon Internetin käytön räjähdysmäisen kasvun 2000-luvulla, on perusteltua olettaa, että nykyään suuri osa tästä asunnon ostoa edeltävästä prosessista tapahtuu Internetissä. Aina kun ihminen tekee haun hakukoneella, hän paljastaa jotain itsestään, esimerkiksi asunnon ostoa edeltävät haut voivat kertoa kiinnostuksesta ostaa asunto.Nämä haut toimivat signaaleina ihmisten kiinnostuksen kohteista tai vaikka asunnon ostoaikeista. Tämän tyylisiä signaaleja käytetään jatkuvasti hyväksi esimerkiksi käyttäjäkohtaisesti kohdennetussa mainonnassa. Mutta onko näistä signaaleista muutakin hyötyä kuin markkinoinnin kohdentaminen?

Google on maailman käytetyin hakukone ja sen kautta tehdään maailmanlaajuisesti päivittäin 3,5 miljardia hakua ja Suomessakin noin 30 miljoonaa [1]. Näiden hakujen tutkiminen voi tarjota mahdollisuuksia havaita muutoksia ihmisten käyttäytymisessä ja siten myös taloudessa. Aikaisempien tutkimusten mukaan, Internetissä tehdyt haut todella auttavat talouden ennustamisessa. Esimerkiksi ennuste työttömyysasteesta paranee, kun siihen lisätään Internet-hakuja kuvaava muuttuja (kts. esim. Choi & Varian, 2012; McLaren & Shanbhogue, 2011; Tuhkuri, 2014), toisaalta Internet-hakutilastojen on myös huomattu parantavan asuntomarkkinaennusteita

[1] Lähde: Google inc.

(McLaren & Shanbhogue, 2011; Wu & Brynjolfsson, 2013).

Tarkastelen tässä raportissa, onko uusista massadatan lähteistä apua Suomen asuntomarkkinoiden nykyhetken ennakoimisessa ja lyhyen aikavälin ennustamisessa. Massadatan lähteenä käytän tässä yhteydessä Google-hakukoneella tehtyjä hakuja, mutta asuntomarkkinoille sopivaa massadataa on varmasti olemassa muistakin lähteistä, kuten esimerkiksi asunnonvälitys-sivustoilta tai vaikka Internetin keskustelupalstoilta. Googlen aineisto on kuitenkin erittäin laaja ja tällä hetkellä se on helpoiten saatavissa.

Tarkastelen Google-hakujen informaatioarvoa asuntomarkkinoiden ennustamisessa muodostamalla kaksi mallia. Ensimmäinen malleista on yksinkertainen aikasarjamalli, joka toimii perusmallina. Toinen malli muodostetaan laajentamalla perusmallia Google-hauista kertovalla muuttujalla. Vertailemalla näita kahta mallia, pystytään selvittämään parantaako Google-hakujen lisääminen mallin ennustetarkkuutta perusmalliin verrattuna. Vaikka asuntomarkkinoiden ennustamisessa käytetyt mallit ovat usein suhteellisen monimutkaisia Wu & Brynjolfsson (2013), käytän tässä yhteydessä mahdollisimman yksinkertaista mallia. Koska raportin tarkoituksena on tarkastella massadatan käytettävyyttä asuntomarkkinoiden ennakoimisessa, on yksinkertaisella mallilla hyvä aloittaa.

Raportti on jäsennelty siten, että seuraavassa luvussa esitellään raportin kannalta relevanttia aikaisempaa tutkimusta Internet-hakujen käytöstä talouden ennustamisessa ja ennenkaikkea asuntomarkkinoiden ennustamisessa. Kolmannessa luvussa esitellään käytettävä aineisto sekä *Google Trends* -palvelu, josta Google-hakuja koskevaa aineistoa saa ilmaiseksi ja hyvin käyttäjäystävällisessä muodossa. Neljännessä luvussa esittelen raportissa käytettävät empiiriset menetelmät. Viidennessä luvussa esitellään keskeisimmät tulokset ja pohditaan hieman niiden tulkintaa. Viimeinen luku vetää raportin keskeisimmät havainnot yhteen.

2 Kirjallisuus

2.1 Internet-haut talouden ennustamisessa

On sanottu, että datan määrä maailmassa kaksinkertaistuu 18 kuukauden välein. Lisäksi yritysdatan määrä tulee todennäköisesti kasvamaan muutaman seuraavan vuoden aikana jopa 650 % (Chang et al., 2014). Valtavien datamäärien käsittely nopeutuu jatkuvasti, joten yhä enemmän

uutta ja entistä reaaliaikaisempaa dataa saadaan niin tutkimuksen kuin yritystoiminnankin tueksi. Yksi osa tästä kasvavasta aineistomääräästä koostuu ihmisten Internetissä tekemistä hauista. Jokaisella haulla Internetin käyttäjä antaa signaalin omista mieltymyksistään tai suunnitelmistaan. Mutta voidaanko tätä aineistoa käyttää talouden trendien ennakoimisessa? Arrow (1987) kutsuu yksittäisten osto- ja myyntipäätösten tutkimista ”nanotaloudeksi”[2]. Internet-hakujen tutkimisen voidaan ajatella vastaavan käytännössä yksittäisten osto- ja myyntipäätösten tutkimista (Wu & Brynjolfsson, 2013).

Vaikka Internet-hakujen käyttö ennustamisessa onkin suhteellisen tuore ilmiö, on tutkimusta ehditty tästä huolimatta julkaista jonkin verran. Näistä tutkimuksista ensimmäinen lienee Ettredge et al. (2005), jossa löydetään yhteys työttömyyden ja Internet-hakujen välillä. Yhteyden havaitseminen on merkittävää, sillä aineiston kasvaessa on mahdollista tarkastella Internet-hakujen käyttöä myös varsinaisessa ennustamisessa. Ennustamistarkoituksessa Internet-aineistojen käyttö on yleistynyt viime vuosina, etenkin helposti saatavissa olevien aineistojen ansiosta.

Ensimmäisenä varsinaista ennustamista Internet-hakujen avulla tarkastellaan Ginsberg et al. (2009) artikkelissa ”Detecting influenza epidemics using search engine query data”, jossa hakuaineistoa käytetään influenssa-epidemioiden ennustamiseen. Myös Polgreen et al. (2008) tarkastelevat influenssa-epidemioiden ennustamista Internet-hakujen avulla. Tieto influenssa-epidemian leviämisestä julkaistaan viiveellä, joten nopean reagoimisen kannalta on tärkeää saada ajantasaista tietoa epidemian leviämisestä. Sekä Ginsberg et al. (2009) että Polgreen et al. (2008) osoittavat Internet-hakujen parantavan influenssa-epidemioiden nykyhetken ennustamista. Vaikka kumpikaan artikkeleista ei käsittele talouden ennustamista, ovat niissä käytetyt menetelmät hyvin samankaltaisia usean talouden ennustamista käsittelevän artikkelin kanssa.

Taloustieteessä Choi & Varian (2009a,b) esittelevät Google-hakuaineiston mahdollisuuksia talouden nykyhetken ennustamisessa[3]. Jälkimmäinen artikkeleista toimii ikään kuin johdatusartikkelina *Google Trends*in käyttöön ennustamistarkoituksessa, kun taas ensimmäinen artikkeleista keskittyy työttömyysasteen ennustamiseen Yhdysvalloissa. Choi & Varian (2012) on

[2] engl. nanoeconomics
[3] engl. nowcasting

käytännössä jalostettu versio artikkelista Choi & Varian (2009b). Kaikissa kolmessa artikkelissa esitellään Internet-hakujen mahdollisuuksia talouden indikaattorien ennustamisessa. Työttömyyden lisäksi artikkeleissa ennustetaan automyyntiä, asuntomyyntiä, matkustajamääriä eri matkakohteisiin sekä kuluttajaluottamusta.

Wu & Brynjolfsson (2013) käyttävät Google-hakuja neljännesvuosittaisten asuntohintojen ja myyntimäärien ennustamiseen. He takastelevat artikkelissaan Internet-hakujen käyttöä asuntohintojen sekä asuntojen myyntimäärien ennustamisessa. Wu ja Brynjolfsson huomaavat, että Google-hakujen lisääminen ennusteeseen parantaa etenkin myyntimäärien ennusteita, mutta myös hintaennustetta. Absoluuttisella keskivirheellä (MAE) mitattuna Google-hakuindeksin sisältävä ennuste myyntimääristä on keskimäärin 7,1% parempi kuin ennuste ilman hakuindeksiä. Wu & Brynjolfsson (2013) menetelmät noudattelevat pitkälti Choi & Varian (2012) esittelemää menetelmää. Erotuksena näissä kahdessa artikkelissa on kuitenkin se, että Wu & Brynjolfsson (2013) ennustavat nykyhetken lisäksi myös lähitulevaisuutta. He myös huomaavat, että heidän mallinsa ennustaa jopa paremmin seuraavaa neljännestä kuin nykyhetkeä.

McLaren & Shanbhogue (2011) käyttävät niin ikään Google-hauista muodostettua indeksiä asuntohintojen sekä työttömyyden ennustamiseen Isossa-Britanniassa. He havaitsevat artikkelissaan asuntohintojen korreloivan positiivisesti Google-indeksin kanssa, joka viittaa siihen, että Google-hauissa asunnon ostajien haut olisivat dominoivia. Myös heidän tuloksensa osoittavat, että Google-indeksin lisääminen parantaa ennustetta perusmalliin nähden. Kyseisessä artikkelissa McLaren ja Shanbhogue tutkivat myös työttömyyden ennustamista Google-hauilla. Heidän havaintonsa mukaan Google-indeksi parantaa myös työttömyysennustetta Isossa-Britanniassa.

Myös Kulkarni et al. (2009) tarkastelee Yhdysvaltojen asuntomarkkinoita. Tutkijat havaitsevat Google-indeksin Granger-aiheuttavan[4] asuntohintoja tutkimuksessaan, joka käsittää 20 metropolialuetta Standard & Poor'sin Case & Shiller indeksin mukaan. He havaitsevat myös, että Granger-kausaalisuus ei ole voimassa toiseen suuntaan, eli asuntojen hinnat eivät Granger-aiheuta Google-indeksiä. Myös heidän tuloksensa antavat viitteitä Google-hakujen käyttökelpoisuudesta ennusteessa.

[4]Granger-kausaalisuudella tarkoitetaan, että yksi aikasarja auttaa ennustamaan toista kts. (Granger, 1969).

Internetin hakutuloksia on käytetty jonkin verran myös Yhdysvaltojen ja Isossa-Britannian ulkopuolella, vaikka valtaosa tutkimuksesta onki tehty nimenomaan yhdysvaltalaisella aineistolla. Esimerkiksi Askitas & Zimmermann (2009) käyttävät Google-hakuja työttömyyden ennustamiseen Saksassa, lisäksi Vicente et al. (2015) huomaavat Google-hakujen auttavan työttömyyden ennustamisessa Espanjassa.

Tarvonen (2015) tarkastelee asuntomarkkinoiden ennustamista Suomessa Google-hakujen avulla. Hän tarkastelee kerrostaloasuntojen hintojen ja myyntimäärien ennustamista yksinkertaisella VAR-mallilla[5] ja havaitsee, että Google-haut eivät paranna ennusteita tilastollisesti merkistevällä tasolla. Lisäksi hän huomaa, että Suomen aineistolla näyttää siltä, että Google-haut seuraavat asuntomarkkinoita, eikä päinvastoin, kuten aiemmat kansainväliset tutkimukset osoittavat.

Suomessa Tuhkuri (2014) käyttää Google-dataa Suomen työttömyysasteen ennustamisessa. Hän havaitsee, että Google-indeksin lisääminen työttömyysennusteeseen parantaa nykyhetken ennustetta keskimäärin 10% ja käännekohdissa 15% absoluuttisella keskivirheellä mitattuna. Lisäksi Tuhkuri (2016) laajentaa saman ennustemallin käsittämään koko EU28-alueen.

2.2 Asuntomarkkinoiden ennustaminen

Case & Shiller (1989) havaitsevat asuntomarkkinoiden olevan ennustettavia, sillä ne eivät toimi täysin tehokkaasti. Toisin kuin esimerkiksi osakemarkkinoilla, asuntomarkkinoilla valtaosa kauppaakäyvistä on yksityishenkilöitä ammattimaisten sijoittajien sijasta, lisäksi transaktiokustannukset ovat huomattavasti osakekauppaa korkeammat, joten esimerkiksi lyhyeksi myyminen on haastavampaa. Asuntomarkkinoiden epätehokkuus mahdollistaa niiden ennustamisen, sillä asuntohinnat eivät noudattele täysin satunnaiskulkua[6].

Asuntomarkkinoita on pyritty ennustamaan monella tavalla. Wu & Brynjolfsson (2013) jakavat nämä tavat karkeasti kahteen ryhmään. Ensimmäinen tapa on keskittyä asuntojen hintoihin vaikuttaviin tekijöihin, kuten rakennuskustannuksiin, korkoihin sekä käytettävissä oleviin tuloihin.

[5] VAR = Vector Autoregressive

[6] engl. random walk

He kuitenkin huomauttavat, että asuntomarkkinoita selittävät muuttujat eivät täysin selitä asuntomarkkinoiden muutoksia. Toinen paljon käytetty tapa on enemmän tekninen, jossa keskitytään asuntohintojen tilastollisiin ominaisuuksiin. Tämän tyylisissä malleissa oletetaan, että asuntohinnat hakeutuvat pitkällä aikavälillä keskiarvoaan kohti[7]. Asuntohinnoilla voi kuitenkin olla lyhyellä aikavälillä trendikäyttäytymistä.

Esimerkiksi Glaeser & Gyourko (2006) löytävät todisteita asuntohintojen pitkänaikavälin keskiarvoon hakeutumisesta. Heidän mukaansa ceteris paribus viiden vuoden aikana tapahtunut yhden dollarin nousu johtaa keskimäärin 32 sentin hintojen laskuun seuraavat viisi vuotta. Myös Case & Shiller (1989) ja Case & Shiller (1990) havaitsevat pitkän aikavälin keskiarvoon palautumista sekä lyhyen aikavälin trendikäyttäytymistä.

Tässä raportissa käytettävä ennustemalli on yksinkertainen aikasarjamalli. Vaikka useimmat asuntojen hintoja ennustavista malleista ovat suhteellisen monimutkaisia, ei se välttämättä takaa niiden paremmuutta. Esimerkiksi Wu & Brynjolfsson (2013) huomauttavat, että yksinkertaisten autoregressiivisten mallien on havaittu ennustavan asuntomarkkinoita yhtä tarkaisti kun monimutkaisempienkin mallien. Tärkein syy yksinkertaisen mallin valitsemiselle on kuitenkin raportin varsinainen tarkoitus testata, auttavatko Google-haut asuntohintojen ennustamisessa Suomessa. Valitsemalla mahdollisimman yksinkertainen malli, voimme keskittyä uuden datan toimivuuden tarkasteluun.

Tarkastelen, lisääkö Google-hakujen lisääminen yksinkertaiseen aikasarjamalliin ennusteen tarkkuutta. Kuten todettu, asuntoja ostavat sekä myyvät pääasiassa yksityishenkilöt. Asunnon osto on suuri päätös, joten sitä edeltää todennäköisesti selvitystyö liittyen mahdolliseen ostopäätökseen. Toisaalta asunnon myynnissä käytetään suhteellisen usein välittäjän palveluita. Tämä todennäköisesti vähentää myyjäosapuolen tekemiä Internet-hakuja. Näin ollen Google-hakujen intensiteetin muutoksen voisi olettaa johtuvan kysynnän tai kiinnostuksen muutoksesta.

[7] engl. mean reverting

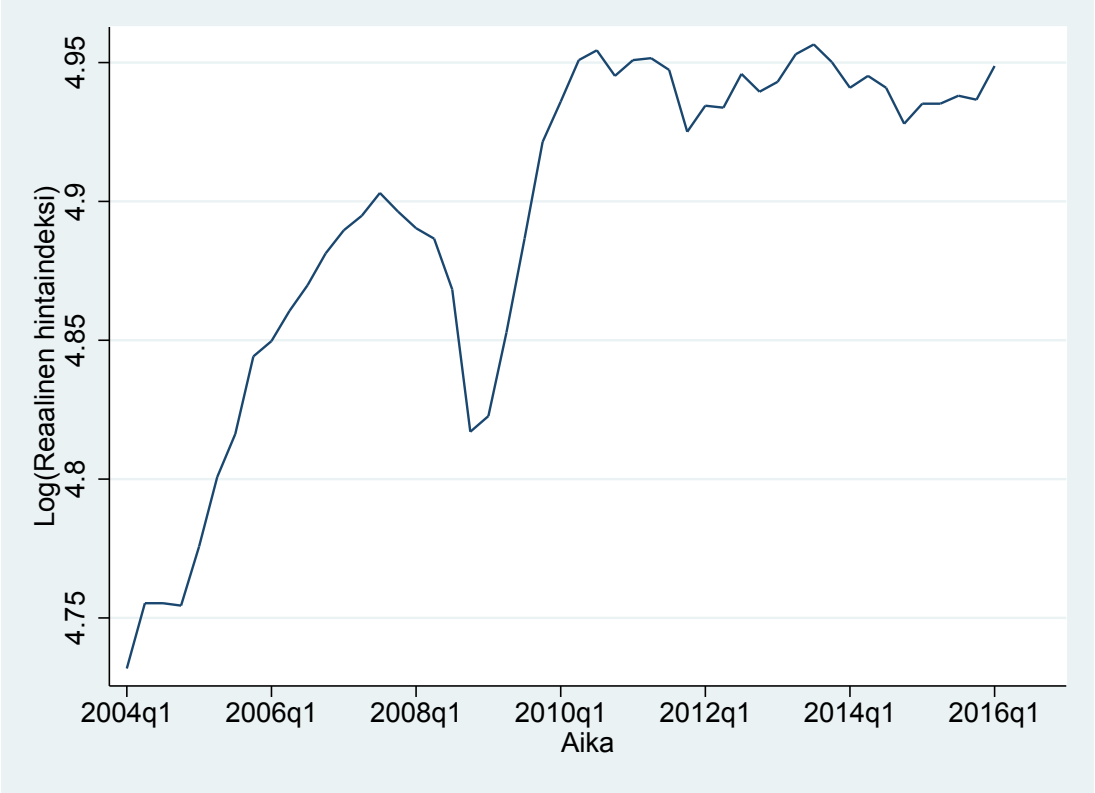
3 Aineisto

3.1 Asuntomarkkina-aineisto

Asuntomarkkinoita kuvaava aineisto on kerätty Tilastokeskuksen avoimesta StatFin -tietokannasta. Asuntohintojen kuvaamiseen olen valinnut reaalisen asuntohintaindeksin, jonka kehitys on esitetty kuvassa 1. Asuntohintaindeksin tiedot on kerätty alunperin Verohallinnon varainsiirtoveroaineistosta, ja neljännesvuosittainen tilasto sisältää noin 2/3 kaikista tehdyista asuntokaupoista (Tilastokeskus). Reaalisessa asuntohintaindeksissä on huomioitu kuluttajahintojen muutokset. Lisäksi hintaindeksiä on laatukorjattu siten, että asuntojen ominaisuuksien erot tasoittuvat. Tässä raportissa tarkastellaan asuntohintaindeksistä ainoastaan vuoden 2004 jälkeistä aikaa, sillä Internet-hakuja koskevaa aineistoa ei ole saatavilla ennen vuotta 2004. Lisäksi, koska Tilastokeskus ei ole julkaissut ennen vuotta 2010 kuukausittaisia asuntohintaindeksejä, täytyy pidemmän aikavälin saavuttamiseksi käyttää neljännesvuosittaista aineistoa. Käytettävässä asuntohintaindeksissä perusvuotena on vuosi 2000, jolloin indeksi saa arvon 100.

Käytettävä asuntomarkkina-aineisto sisältää ainoastaan vanhat osakemuotoiset kerrostaloasunnot. Vanhojen kerrostaloasuntojen voidaan ajatella antavat riittävän hyvän kuvan asuntomarkkinoista, sillä niiden osuus kaikista markkinoilla olevista asunnoista on 45% (Tilastokeskus). Lisäksi rajaamalla tarkasteltava aineisto ainoastaan vanhoihin kerrostaloasuntoihin, saadaan käsiteltävästä asuntokannasta huomattavasti homogeenisempi kuin tarkasteltaessa kaikkia asuntoja (Oikarinen & Engblom, 2014).

2000-luvulla asunto-osakkeiden hinnat ovat olleet pääasiassa nousujohteisia. Globaali finanssikriisi näkyy kuitenkin myös asuntojen hinnoissa, kuten kuvasta 1 huomataan. Asuntojen hinnoissa on nähtävissä selkeä kuoppa, joka alkaa vuonna 2008. Vuodesta 2010 alkaen asuntojen hinnat ovat pysytelleet melko vakaina. Kuitenkin esimerkiksi Helsingin asuntohinnat ovat jatkaneet kasvamistaan myös 2010-luvulla.



Kuva 1: Logaritmi reaalisesta hintaindeksistä, koko maa. Lähde: Tilastokeskus

3.2 Google Trends

Internet-hakuja koskevan aineisto on kerätty hakukoneyhtiö Googlen ylläpitämästä *Google Trends* -palvelusta. Palvelun avulla käyttäjä pystyy vertailemaan eri hakusanojen tai niiden yhdistelmien suosiota verrattuna muihin käytettyihin hakusanoihin. Kaikista hakukoneista nimenomaan Googlen hakuaineiston valinta perustuu ennen kaikkea sen helppoon saatavuuteen, mutta samalla Google on selvästi maailman käytetyin hakukone noin kahden kolmasosan maailmanlaajuisella markkinaosuudella. Lisäksi Suomessa Googlen osuus kaikista Internet-hauista on arviolta noin 95%[8], joten sillä on selvästi dominoiva asema Suomessa käytössä olevista hakukoneista. Googlen aineisto tarjoaa näin ollen hyvin kattavan kuvan suomalaisten tekemistä Internet-hauista.

Google Trends ei kerro todellisista hakumääristä mitään, sen sijaan se kertoo

[8] http://gs.statcounter.com/#search_engine-FI-monthly-201507-201607

hakusanan tai niiden yhdistelmän osuudesta kaikkiin samalla alueella tehtyihin hakuihin nähden. Jokainen hakusana saa arvon 0 ja 100 välillä, joka kertoo kyseisen hakusana hakuintensiteetin valitulla maantieteellisellä alueella. Riippuen käytettävästä aikavälistä aineistoa on saatavilla päivä-, viikko- tai kuukausitasolla. Hakuintensiteetti normalisoidaan nollan ja sadan välille siten, että hetkellä, jolloin kyseisen hakutermin suosio on korkeimmillaan, se saa arvon 100. Näin ollen esimerkiksi 20% pudotus hakuintensiteetissä tarkoittaa todellisten hakumäärien pudonneen 20%.

Tuhkurin (2016) esitystä mukaillen hakuintensiteetti voidaan esittää seuraavasti: Jos hakusanalla k tehdään yhteensä K_t hakua ajanhetkellä t ja valitulla maantieteellisellä alueella tehdään kaikilla hakutermeillä yhteensä G_t hakua ajanhetkellä t, voidaan hakuintensiteetti esittää kaavalla

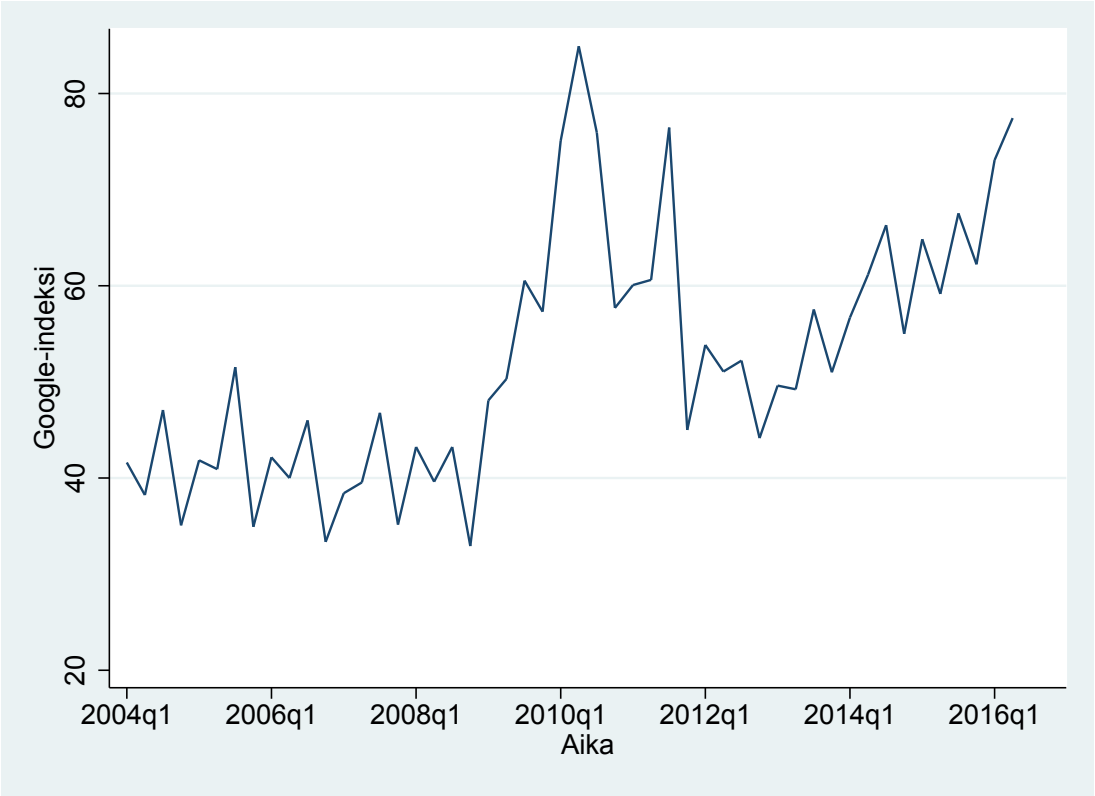
$$I(K_t) = \left\{ \frac{\frac{K_t}{G_t}}{max\left(\frac{K}{G}\right)} \right\} \times 100 \qquad (1)$$

*Google Trends*in ensimmäiset havainnot ovat vuoden 2004 tammikuulta, jonka jälkeen termien intensiteetille on saatavissa sekä viikko- että kuukausitasoinen aikasarja. Koska käyttämämme asuntohintoja kuvaava asuntohointaindeksi on neljännesvuositasolla, päätän aggregoida myös Google-hakuintensiteettiä kuvaavan muuttujan neljännesvuositasolle laskemalla kuukausittaisesta aikasarjasta keskiarvon. Aggregoitua Google-indeksiä kuvaava aikasarja on esitettu kuvassa 1.

Google Trends -aineiston suurin hyöty on sen reaaliaikaisuus. Siinä missä asuntohintaindeksi julkaistaan kuukausi kunkin neljänneksen loppumisen jälkeen, on Google-aineistosta saatavilla käytännössä reaaliaikaisesti. Google-hakujen käyttö ennustamisessa perustuukin juuri näihin julkaisuviiveiden erotuksiin; tämän hetken Google-haut voivat antaa signaalin siitä, mihin suuntaan asuntomarkkinat kehittyvät.

Google-hakuihin liittyy kuitenkin myös joitain ongelmia. Kuten todettua, *Google Trends* ei kerro hakusanojen todellisista hakumääristä mitään. Lisäksi hakuintensiteetin normalisointi johtaa aikaisempien arvojen muuttumiseen uusien arvojen tullessa, eli käytännössä elokuun arvo tietylle hakusanalle saattaa olla eri tänään kuin viikon päästä. Normalisoinnin hyvä puoli on se, että se poistaa käytännössä jokaiseen hakutermiin liittyvän todellisten hakumäärien kasvavan trendin. Toisaalta normalisointi voi aiheuttaa useimmille hakutermeille hieman laskevan trendin, sillä uusien hakusanojen

määrä kasvaa jatkuvasti. Tierney & Pan (2012) havaitsee Google-aineiston aggregoinnin johtavan tiedon osittaiseen häviämiseen.



Kuva 2: Google Indeksi

4 Menetelmät

Tässä luvussa esiteltävien menetelmien avulla selvitän, onko Google-hauista apua asuntojen hintojen ennustamisessa Suomessa. Tässä raportissa käytettävä menetelmä on samankaltainen esimerkiksi Wun ja Brynjolfssonin (2013) käyttämän kanssa. Lisäksi esimerkiksi menetelmät, joita Tuhkuri (2016) ja Choi & Varian (2009b) käyttävät ovat hyvin samantyylisiä. Karkeasti käytettävä menetelmä voidaan jakaa kahteen osaan: muuttujien, ennenkaikkea Google-indeksin, valintaan sekä tilastolliseen testaamiseen.

Asunnon ostoon liittyvät Google-haut kertovan ihmisten mahdollisista asunnonostoaikeista tai kiinnostuksesta asunnon ostoa kohtaan. Todennäköisesti potentiaalinen asunnonostaja selvittää asunnon ostoon liittyviä asioita ennen varsinaista asuntokauppaa. Näitä asioita voivat olla esimerkiksi

asuntonäyttöjen etsiminen tai asuntolainojen vertailu ja kilpailutus. Nykyään suuri osa tästä selvitystyöstä tehdään Internetin välityksellä, ja näin ollen asunnon ostoon liittyvä haku saattaa toimia signaalina esimerkiksi kiinnostuksesta ostaa asunto. Lisäksi koska Internet-hakuihin ei liity strategista käyttäytymistä, voidaan hakujen olettaa olevan rehellinen signaali kiinnostuksesta tai kysynnästä jotain asiaa kohtaan (Wu & Brynjolfsson, 2013).

Asunnon ostoprosessiin liittyy siis usein erilaisia aiheeseen liittyviä Internet-hakuja. Aina ei kuitenkaan ole täysin selvää, minkä tyylisiä hakusanoja ihmiset käyttävät tietoa etsiessään. Toisaalta asuntomarkkinoilla hakuja voi tehdä niin asunnon ostoa suunnitteleva kuin asunnon myynnistä kiinnostunut henkilö. On kuitenkin todennäköistä, että asunnon ostamista suunnittelevat tekevät enemmän aiheeseen liittyviä hakuja (McLaren & Shanbhogue, 2011). Tarkastellaan ensin, miten tässä raportissa käytettävä Google-indeksi on muodostettu.

Toisin kuin esimerkiksi Yhdysvalloissa, Suomessa *Google Trends* ei tarjoa valmiita kategorioita[9], joten käytettävä Google-indeksi on muodostettava yhdistämällä yksittäisiä hakusanoja. Kirjallisuudessa on esitelty muutamia strategioita käytettävien hakutermien valitsemiseksi. Esimerkiksi Choi & Varian (2009a,b, 2012) käyttävät valmiita kategorioita Google-indeksin muodostamiseksi. Toisaalta esimerkiksi Tuhkuri (2016) valitsee käytettävät hakusanat omaan asiantuntemukseensa perustuen. Tämän tavan hyvä puoli on siinä, että tutkija tietää jokaisen käytettävän hakusanan ja pystyy muuttamaan niitä tarvittaessa. Scott & Varian (2013) esittelevät automaattisen hakutermien valintamenetelmän, joka hyödyntää Googlen Correlate-palvelua ja valitsee muuttujan kanssa eniten korreloivia termejä Google-indeksiin. Eniten korreloivista hakusanoista täytyy kuitenkin poistaa sanat, jotka korreloivat ainoastaan sattumalta ennustettavan aikasarjan kanssa. Brynjolfsson et al. (2015) muodostaa Google-indeksin kyselyyn pohjautuvalla menetelmällä, jossa vastaajia pyydetään listaamaan sanoja, jotka tulevat ensimmäisenä mieleen esimerkiksi asunnon ostamisesta. Näin pyritään selvittämään, mitkä voisivat olla hakusanoja, joita ihmiset todella käyttävät tietoa etsiessään.

Tässä raportissa käyettävä Google-indeksi on muodostettu omaan harkintaan pohjautuen. Google-indeksi on rakennettu siten, että olen ensin

[9]Valmiit kategoriat sisältävät suuren määrän tiettyyn aihepiiriin liittyviä hakusanoja.

listannut kaikki mieleeni tulleet mahdollisesti asuntomarkkinoihin liittyvät hakusanat. Näistä hakusanoista lopulliseen Google-indeksiin valikoidaan hakusanat niiden todellisten hakumäärien mukaan Google Adwordsin avulla. Google Adwords on mainostamiseen tarkoitettu palvelu, jonka avulla on mahdollista vertailla todellisia hakumääriä viimeisen kahden vuoden ajalta kuukausittaisina keskiarvoina.

Lopullinen Google-indeksi muodostetaan 7 hakutermistä, jotka ovat: myytävät asunnot, asunnot, asunto, asuntolaina, kiinteistönvälittäjä, asuntolainan korko, asunnon osto. Google-indeksin kehitystä on kuvattu yhdessä reaalisen asuntohintaindeksin kanssa kuvassa 1. Kuvan molemmat muuttujat on esitetty neljännesvuositasolla.

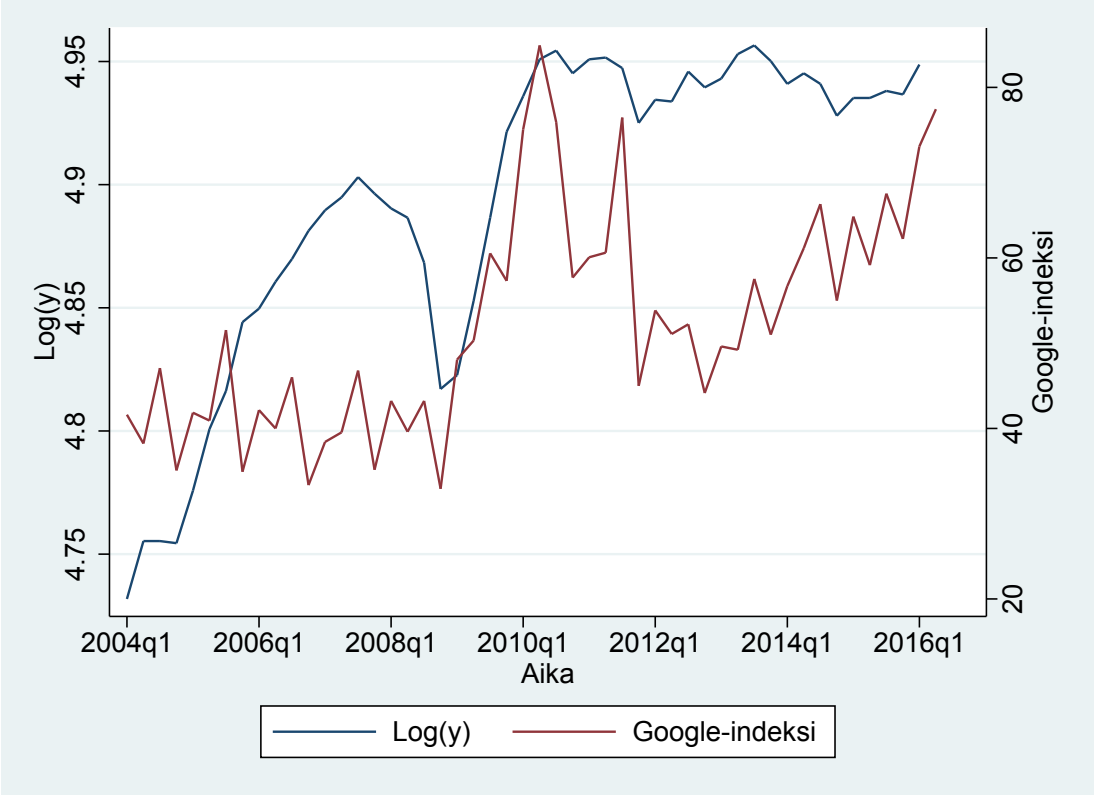
Maantieteellisen alueen rajaaminen *Google Trends* -palvelussa on mahdollista. Suomessa maakuntatasoinen aluerajaus on otettu käyttöön vasta vuonna 2013. Luotettavien tulosten takaamiseksi 12 neljännestä (2013-2016) on liian lyhyt, joten maantieteellistä rajausta ei voida tässä yhteydessä käyttää. Koska asuntomarkkinoiden alueelliset erot ovat kuitenkin valtavat, on syytä muodostaa alueellisia Google-indeksejä. Maantieteellisen rajauksen puuttuessa muodostan indeksit käyttäen aluekohtaisia hakutermejä. Esimerkiksi pääkaupunkiseutua kuvaavaan Google-indeksiin sisällytän termit ”Asunnot *Kaupunki*”ja ”Asunto *Kaupunki*”, jossa *Kaupunki* kohdalla on kukin pääkaupunkiseudun kaupungeista.

Muuttuja	N	μ	σ	*min*	*max*
Reaalinen asuntohintaindeksi	49	113,5	8,2	133,7	142,1
Google-indeksi	50	52,3	13,0	32,9	84,9

Taulukko 1: Muuttujien tilastollisia tunnuslukuja. Lähde: *Google Trends*, Tilastokeskus ja omat laskelmat

Taulukossa 1 on esitetty perustilastoja molemmista muuttujista. Kuten kuvasta 1 huomataan muuttujien välinen korrelaatio näyttää olevan hyvin pientä etenkin tarkastelujakson alussa. Molemmat indeksit ovat hyvin matalalla vuonna 2009, jonka jälkeen molemmat lähtevät kasvuun. Sarjojen välinen korrelaatio näyttäisi olevan hieman parempi tarkasteluperiodin keskellä ja loppupäässä. Alun korreloimattomuudelle voi olla monia selityk-

siä. Esimerkiksi ihmisten Internet-käyttäytymisen muutos voi olla osittain syynä muuttuneeseen korrelaatioon.



Kuva 3: Logaritmi asuntohintaindeksistä sekä Google-indeksi

Tarkastelen muuttujien kovarianssi-stationaarisuutta Dickey & Fuller (1979) esittelemällä laajennetulla Dickey–Fuller -testillä. Kummankin muuttujan kohdalla huomataan, että nollahypoteesia yksikköjuuresta ei pystytä hylkäämään. Lisäksi testaan ADF-testillä, sisältävätkö ensimmäiset differenssit yksikköjuurta. Testi hylkää nollahypoteesin yksikköjuurista molempien muuttujien osalta, joka viittaa ensimmäisten differenssien stationaarisuuteen. Testaan muuttujien mahdollista trendistationaarisuurtta KPSS-testillä[10] (Kwiatkowski et al., 1992). Testin nollahypoteesin mukaan muuttuja on stationaarinen deterministisen trendin ympärillä. Testi ei pysty hylkäämään logaritmisen asuntohintaindeksin trendistationaarisuutta 5% tasolla. Testien perusteella olisi perusteltua käyttää muuttujista ensimmäisiä differenssejä tasojen sijasta.

[10] Kwiatkowski–Phillips–Schmidt–Shin -testi

Kuitenkin esimerkiksi Stockin (2001) mukaan differenssin ottaminen ennustemallissa ei ole välttämätöntä, jos ennustehorisontti on riittävän lyhyt sekä tarkasteltava periodi riittävän pitkä suhteessa ennustehorisonttiin. Tässä yhteydessä ennustehorisontti on lyhyt, ja toisaalta myös sisällytettävien periodien määrä on ennustehoirisonttiin nähden suuri. Lisäksi Swanson & White (1997) huomauttavat, että stationaarisuuden pohtiminen ei ole ensisijainen ongelma, jos tarkoituksena on tarkastella mallin aineiston ulkopuolista ennustekykyä. Näistä syistä johtuen päätän käyttää molemmista muuttujista tasoja differenssien sijaan. Aikaisemmin esimerkiksi Wu & Brynjolfsson (2013) ja Kulkarni et al. (2009) tarkastelevat asuntohintoja tasomuuttujina. Toisaalta esimerkiksi McLaren & Shanbhogue (2011) tarkastelee asuntohintojen ensimmäisiä differenssejä.

Aloitan sopivan perusmallin valitsemisen tarkastelemalla reaalisen hintaindeksin autokorrelaatio- sekä osittaisautokorrelaatiofunktioita. Näiden kuvaajat on esitetty liitteessä 1. Autokorrelaatiofunktio näyttää laskevan hitaasti, kun taas osittaisautokorrelaatio on tilastollisesti merkitsevä ainoastaan kaksi ensimmäistä viivettä, jonka jälkeen se ei ole enään merkitsevä. Autokorrelaatiofunktioiden visuaalinen tarkastelu antaisi tukea AR(2)-mallin käytölle (Verbeek, 2008). Autokorrelaatiofunktiot eivät myöskään osoita vahvaa kausittaista käyttäytymistä.

Myös yleisimmin käytetyt informaatiokriteerit, Akaiken infromaatiokriteeri (AIC) sekä Schwarzin informaatiokriteeri (BIC), tukevat toisen asteen autoregressiivisen mallin valintaa. Tarkastelen vielä lopuksi valitun mallin virhetermejä. Liitteessä 2 on esitetty residuaalien autokorrelaatifunktio, joka ei osoita selvää autokorrelaatiota millään viivepituudella.Tarkastelen vielä residuaaleja Ljung–Box (1978) -testin avulla. Nollahypoteesia virhetermien riippumattomuudesta ei pystytä hylkäämään, joka tukee mallin valintaa.

Edellä esitettyjen syiden johdosta valitsen perusmalliksi toisen asteen autoregressiivisen mallin. Käytännössä tämä tarkoittaa, että tämän hetkistä asuntohintaindeksiä selitetään kahden edellisen neljänneksen asuntohintaindekseillä. Aikaisemmin McLaren & Shanbhogue (2011) käyttää niin ikään toisen asteen autoregressiivistä mallia. Käytettävä perusmalli on esitetty formaalisti kaavassa 2a.

Google-haut sisältävä malli muodostetaan lisäämällä perusmalliin muodostettu Google-indeksi. Koska Google-indeksi julkaistaan lähestulkoon

reaaliajassa, voimme käyttää Google-indeksin[11] nykyhetken arvoa x_t. Google-hauista muodostetun indeksin tarkoitus on reaaliaikaisuutensa ansiosta antaa ennakkoon signaaleja asuntomarkkinoilla tapahtuvista muutoksista. Google-indeksillä laajennettu malli on esitetty formaalimmin kaavassa 2b. Google-indeksin lisäämisen jälkeen mallit estimoidaan QML (Quasi Maximum Likelhood)-menetelmällä.

$$\text{Malli (0):} \quad log(y_t) \sim log(y_{t-1}) + log(y_{t-2}) + \varepsilon_t \tag{2a}$$
$$\text{Malli (1):} \quad log(y_t) \sim log(y_{t-1}) + log(y_{t-2}) + x_t + \varepsilon_t, \tag{2b}$$

jossa y_t on reaalinen asuntohintaindeksi, x_t on asuntoihin liittyvistä Google-hauista muodostettu indeksi ja ε on virhetermi.

Mallien estimoinnin jälkeen tarkastelen mallien selitysasteita (R^2) sekä informaatiokriteereitä. Keskityn kuitenkin pääasiassa mallien ennustetarkkuuksien vertailuun.

Ennusteiden tekemiseen käytetään ainoastaan sillä hetkellä käytettävissä olevia arvoja. Käytännössä tämä tarkoitttaa, että nykyhetken ennusteessa voidaan käyttää edellisen periodin hintaindeksiä sekä tämän hetken Google-indeksiä. Ennusteita vertaillaan muodostamalla nykyhetken ennuste niin sanotulla *"rolling window"* -menetelmällä. Käytännössä mallia opetetaan ensin 20 periodia, jonka jälkeen ennustetaan 21. periodin hintaindeksiä. Seuraavaksi malli estimoidaan uudelleen, tällä kertaa käyttäen havaintoja toisesta periodista 21. periodiin. Näin ollen mallin estimointiin käytettävä ikkuna siirtyy jokaisen ennusteen jälkeen yhden askeleen eteenpäin, joten ikkuna pysyy saman suuruisena koko ajan.

Nykyhetken ennustamisen lisäksi tarkastelen myös lähitulevaisuuden ennustamista. Käytännössä menetelmä on täysin yhtenevä nykyhetken ennustamisen kanssa, ainoastaan käytettävissä oleva aineisto on nyt yhden periodin vanhempaa ennustettavaan periodiin nähden. Ennustettaessa yhden periodin tulevaisuuteen käytettävissä on nykyhetken Google-indeksi ja edellisen periodin asuntohintaindeksi. Eli jos ennustettava periodi on y_{t+1}, käytetään ennusteen tekemiseen x_t:tä sekä y_{t-1}. Tarkastelen lisäksi Google-indeksin ja asuntohintaindeksin välistä ristikorrelaatiota sekä

[11] Google-indeksi on koottu hakusanoista: myytävät asunnot, asunnot, asunto, asuntolaina, kiinteistönvälittäjä, asuntolainan korko, asunnon osto.

Granger-kausaalisuutta, joiden avulla pyrin selvittämään Google indeksin nykyisten ja menneiden arvojen yhteyttä asuntohintaindeksiin.

Mallin ulkopuolisten ennusteiden tarkkuutta verrataan absoluuttisen keskivirheen avulla. Absoluuttinen keskivirhe (MAE) kertoo keskiarvon ennusteen ja todellisen arvon välisestä etäisyydestä. Absoluuttinen keskivirhe voidaan esittää kaavalla

$$\mathrm{MAE} = \frac{1}{T}\sum_{i=1}^{T}\left|\frac{\hat{y}_t - y_t}{y_t}\right|, \tag{3}$$

jossa $\hat{y}_t$ on ennustettu arvo ja y_t on todellinen, havaittu arvo.

Mallien ennustevirheiden erotusten tilastollista merkitsevyyttä tarkastellaan Diebold–Mariano -testillä (Diebold & Mariano, 1995). Testin nollahypoteesi on $H_0 : d_t = 0$, jossa $d_t = e_{1,t} - e_{0,t}$ tarkoittaa kahden mallin virheiden erotusta. Nollahypoteesin mukaan ennustevirheiden erotus ei eroa merkitsevästi nollasta. Testillä voidaan arvioida parantaako Google-indeksin lisääminen mallin tarkkuutta tilastollisesti merkitsevällä tavalla.

Vaikka kummassakaan muuttujassa ei esiinny selvästi havaittavaa kausivaihtelua, tarkastelen AR(2)-mallin lisäksi myös ensimmäisen asteen kausivaihtelu autoregressiivisellä mallilla, jossa asuntohintaindeksiä selitetään edellisen neljänneksen asuntohintaindeksillä sekä vuodentakaisella asuntohintaindeksillä. Kausivaihtelumalli ottaa huomioon myös mahdolliset kausivaihtelut. On mahdollista, että Google-indeksi selittää asuntohintoja ainoastaan yhteisen kausivaihtelun takia.

5 Empiiriset tulokset

Taulukossa 2 on esitetty mallien (0) ja (1) tulokset. Ensimmäisessä sarakkeessa on perusmallin estimoidut tulokset ja toisessa Google-indeksillä laajennetun mallin tulokset. Kuten taulukoista huomataan Google-indeksi on molemmissa malleissa tilastollisesti merkitsevä selittäjä 1% tasolla. Google-indeksin kerroin on positiivinen, joka tarkoittaa, että hakujen kasvu on yhteydessä hinnan nousuun. Käytännössä kerroin tulkitaan siten, että mallissa (1) 1% kasvu Google-indeksissä johtaa noin 0,036% nousuun reaalisessa hintaindeksissä. Perusmallin selitysaste, R^2, on jo itsessään hyvin

korkea. Google-indeksin lisääminen malliin parantaa selitysastetta hieman, mutta tämä ei ole sinällään yllätys, sillä selitysaste paranee käytännössä aina kun malliin lisätään uusia selittäjiä. Selitysasteen kasvaminen kertoo, että Google-indeksillä laajennettu malli selittää asuntohintaindeksin variaatiota paremmin kuin perusmalli.

Malli	(0)	(1)
Selittäjät		
$log(y_{t-1})$	1.483^{***}	1.500^{***}
	(0.1846)	(0.1826)
$log(y_{t-2})$	-0.5046^{***}	-0.521^{***}
	(0.1921)	(0.1924)
x_t		0.0003632^{**}
		(0.0001)
Vakio	4.866^{***}	4.845^{***}
	(0.0926)	(0.0939)
Yhteenveto		
R^2	0.867	0.880
AIC	−271.03	−275.20
BIC	−263.47	−265.75
N	49	49

Semirobustit keskivirheet suluissa

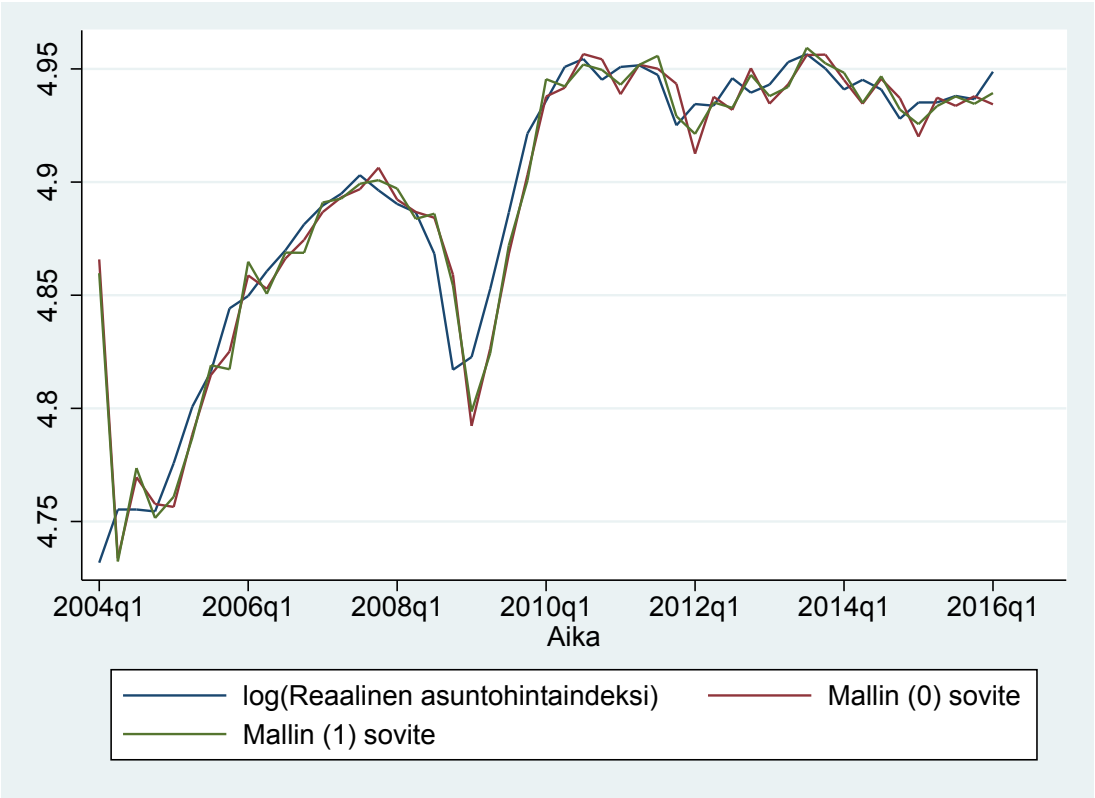
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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

Sekä Akaiken (AIC) että Schwarzin (BIC) informaatiokriteerien arvot pienenevät Google-indeksin lisäämisen myötä. Tästä voidaan päätellä, että Google-indeksi sisältää asuntohintaindeksin selittämisen kannalta merkittävää informaatiota.

Kuvassa 4 on esitetty perusmallin (0) ja Google-indeksillä laajennetun mallin (1) sovitteet, lisäksi kuvassa on toteutunut reaalinen asuntohintaindeksi. Kuten kuvasta huomataan, molemmat mallit näyttäisivät selittävän asuntohintaindeksiä melko hyvin. Google-indeksin lisääminen näyttäisi hieman

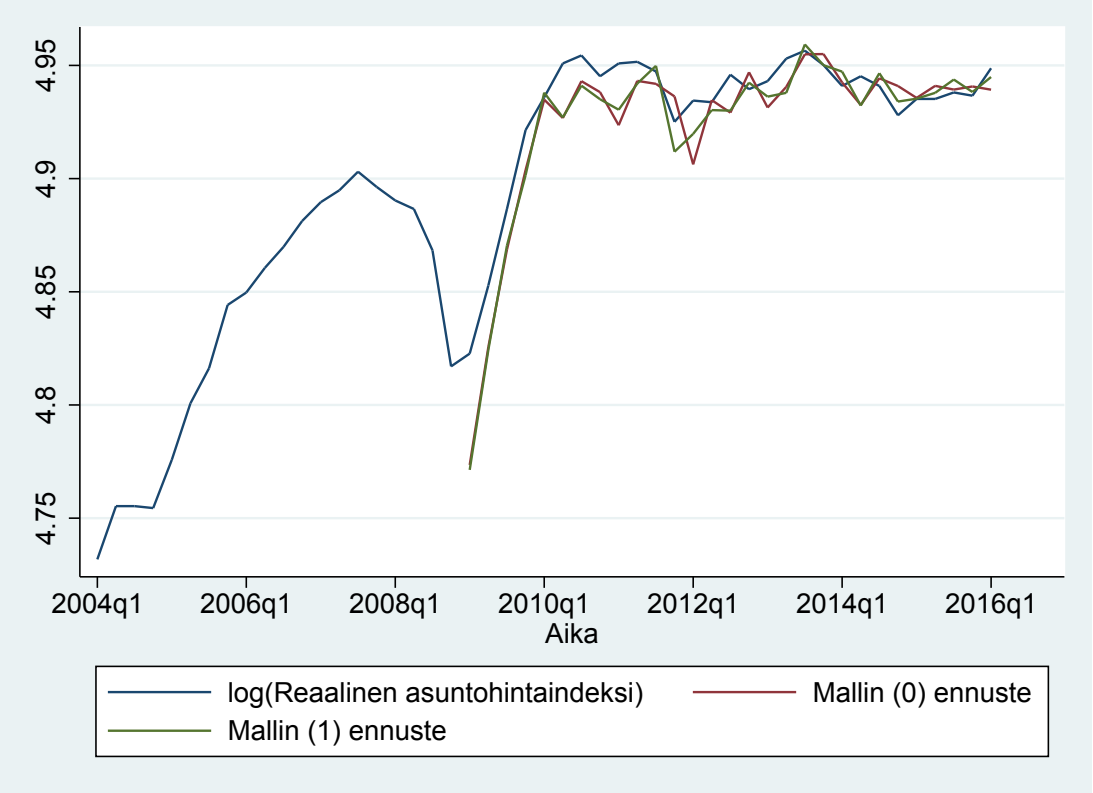
parantavan mallin sovitetta, joskaan parannus ei ole silmämääräisesti kovin suurta.



Kuva 4: Mallien sovitteet ja reaalinen asuntohintaindeksi

Tarkkakaan sovite ei välttämättä takaa hyvää ennustetta, joten tarkastellaan seuraavaksi mallin ulkopuolista ennustekykyä. Mallien ulkopuolisen ennustekyvyn tutkimiseksi käytämme aikaisemmin esiteltyä *"rolling window"* -menetelmää. Ennusteet on tehty jokaiselle periodille alkaen 13. periodista, eli vuoden 2007 ensimmäisestä periodista. Aloitetaan tarkastelemalla, onko Google-indeksin lisäämisestä hyötyä nykyhetken ennustamisessa.

Kuvassa 5 on esitetty mallien yhden askeleen ennusteet. Kuten kuvasta huomataan, molempien mallien yhden askeleen ennustekyky näyttää olevan melko hyvä. Erot ennustetarkkuudessa ovat silmämääräisesti jälleen melko pieniä, joten tarkastellaan seuraavaksi ennustetarkkuutta hieman formaalimmin.



Kuva 5: Mallien ennusteet ja reaalinen asuntohintaindeksi

Vertaillaan mallien ennustetarkkuuksia absoluuttisen keskivirheen avulla[12]. Koko ennustejaksolla Google-indeksillä laajennetun mallin absoluuttinen keskivirhe on pienempi kuin perusmallin. Käytännössä tämä tarkoittaa, että keskimäärin Google-indeksin lisääminen tarkentaa ennustetta[13]. Taulukossa 3 on esitetty ennustejakson absoluuttiset keskivirheet ja niiden erotus. Keskivirheet on mitattu kaikkien ennusteperiodien virheiden keskiarvona.

Malli	(0)	(1)	Δ%
MAE	1.12	1,01	-7.5%**

** 5% merkitsevä Diebold–Mariano -testillä

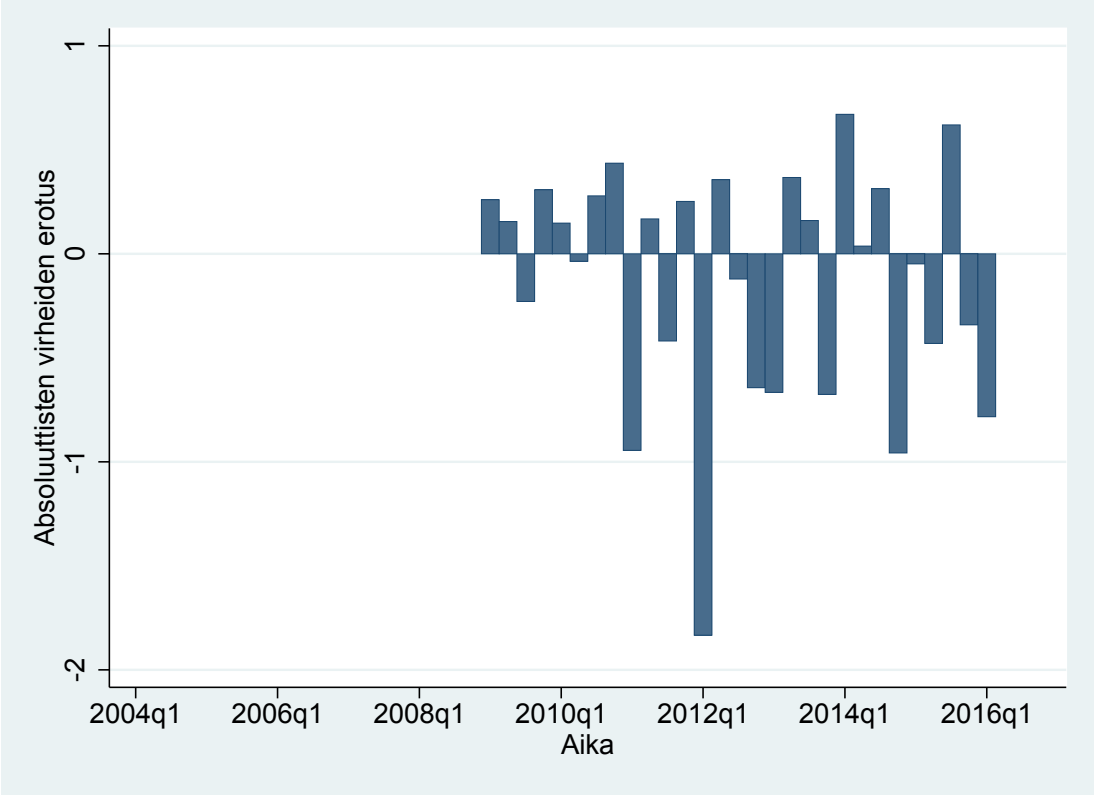
Taulukko 3: Otoksen ulkopuolinen ennustetarkkuus

Vaikka Google-indeksillä laajennetun mallin absoluuttinen virhe onkin kes-

[12]MAE, Mean Absolute Error

[13]Johtopäätökset ovat samansuuntaiset, jos käytetään keskineliövirhettä (MSE)

kimäärin pienempi kuin perusmallin, ei se tarkoita, että Google-indeksillä laajennettu malli olisi parempi jokaisella periodilla. Kuvassa 6 on jokaisen periodin absoluuttisten virheiden erotus, siten että negatiiviset arvot tarkoittavat mallin (1) virheen olevan pienempi kuin mallin (0). Kuten huomataan, Google-indeksin lisääminen ei paranna ennustetta jokaisella ennustetulla periodilla.



Kuva 6: Absoluuttisten keskivirheiden erotus

Ennustetarkkuuksien erotusta tarkastellaan Diebold–Mariano -testillä (Diebold & Mariano, 1995). Ennustetarkkuuksien erotusten havaitaan olevan tilastollisesti merkitseviä 5%-tasolla. Erotus on näin ollen selkeästi tilastollisesti merkitsevä, ja tarkasteltavalla periodilla Google-indeksin voidaan todeta parantavan ennustetta. Parannukset ovat kuitenkin tilastollisesta merkitsevyydestä huolimatta suhteellisen pieniä. Lisäksi Diebold–Mariano -testin voima heikkenee, kun otoskoko on pieni.

Aikaisemmin kirjallisuudessa esimerkiksi Kulkarni et al. (2009) huomaavat Google-indeksin Granger-aiheuttavan asuntohintoja. Tarkastellaan seuraa-

vaksi ensimmäisen asteen vektori autoregressiivisen mallin (VAR) avulla, auttaako Google-indeksi seuraavan periodin asuntohintaindeksin ennustamisessa. Lisäksi tarkastelen yhden periodin edellä olevan Google-indeksin ja asuntohintaindeksin välistä Granger-kausaalisuutta kuten Tuhkuri (2016). Tällöin tarkastellaan todellisuudessa käytettävissä olevaa aineistoa, sillä Google-aineisto on käytettävissä aikaisemmin kuin hintaindeksi. Grangerin (Granger, 1969) ei-kausaalisuustestin tulokset on raportoitu taulukossa 4. Tulokset näyttävät siltä, että Google-indeksin ensimmäinen viive ei auttaisi asuntohintaindeksin ennustamisessa. Sen sijaan nykyhetken Google-indeksi auttaa nykyhetken asuntohintaindeksin ennustamisessa. Asuntohintaindeksi ei näytä Granger-aiheuttavan Google-indeksiä kummassakaan tapauksessa, joten asuntohintaindeksistä ei ole apua Google-indeksin ennakoimisessa.

VAR(1)				VAR(1) viivästetyllä Google-indeksillä			
$y \nrightarrow x$		$x \nrightarrow y$		$y \nrightarrow x$		$x \nrightarrow y$	
χ^2	p-arvo	χ^2	p-arvo	χ^2	p-arvo	χ^2	p-arvo
2,50	0,286	1,29	0,524	0,15	0,927	10,86	0,004***

Taulukko 4: Muuttujien tilastollisia tunnuslukuja. Lähde: *Google Trends*, Tilastokeskus ja omat laskelmat

Wu & Brynjolfsson (2013) huomaavat asuntohintaindeksin nykyhetken korreloivan aikaisempien Google-indeksien kanssa jopa vahvemmin kuin nykyhetken Google-indeksien kanssa. Saman ilmiö havaitaan myös työttömyyttä ennustettaessa (Tuhkuri, 2014). Taulukossa 5 raportoidut ristikorrelaatiokertoimet kertovat, että myös tämän raportin aineistolla nykyhetken asuntohintaindeksin kanssa korreloi parhaiten kahden periodin takainen Google-indeksi. Tämän mielenkiintoisen havainnon taustalla voi olla muun muassa ihmisten käyttäytyminen ennen asunnon ostoa: Asunnon ostoa suunnitteleva todennäköisesti alkaa hankkimaan tietoa hyvissä ajoin ennen varsinaisen ostopäätöksen tekemistä. Tämä viittaa siihen, että Google-indeksistä voisi olla apua myös tulevaisuuden ennustamisessa (Tuhkuri, 2014).

Lag	-4	-3	-2	-1	0	1	2	3	4
CCF	0.62	0.61	0.65	0.63	0.62	0.51	0.46	0.33	0.25

n=40, CCF = ristikorrelaatiofunktion arvo

Taulukko 5: Google-indeksin ja logaritmisen reaalisen asuntohintaindeksin ristikorrelaatio

Ristikorrelaatiofunktion arvot antavat viitteitä siitä, että Google-indeksin lisääminen malliin voisi parantaa myös lähitulevaisuuden ennustetta. Tarkastellaan seuraavaksi lähitulevaisuuden ennustamista. Taulukossa 6 on esitetty mallien ennustevirheet ja niiden erotukset nykyhetken lisäksi seuraaville kolmelle neljännekselle. Ensimmäinen rivi kertoo jo aikaisemmin tarkastellun nykyhetken ennusteen. Seuraavilla riveillä on esitetty lähitulevaisuuden ennustevirheet. Google-indeksin lisääminen näyttäisi parantavan tulevaisuuden ennusteita jopa enemmän kuin nykyhetken enustetta. Molempien mallien ennustetarkkuus näyttää heikkenevän mitä pidemmälle tulevaisuuteen ennuste tehdään. Google-indeksillä laajennetun mallin ennusteet kuitenkin heikkenevät huomattavasti vähemmän kuin perusmallin ennustevirheet.

	Malli	MAPE	Δ%
t	(0)	1.17	-7.5%**
	(1)	1.09	
$t+1$	(0)	1.96	-7.42%
	(1)	1.81	
$t+2$	(0)	3.04	-24.5%*
	(1)	2.29	
$t+3$	(0)	4.01	-34.7%**
	(1)	2.62	

Δ= absoluuttisten keskivirheiden (MAE) erotus
*,** ja *** tarkoittavat tilastollista merkitsevyyttä 10%, 5% ja 1% tasolla Diebold–Mariano -testillä mitattuna

Taulukko 6: Nykyhetken ja lähitulevaisuuden ennusteiden prosentuaaliset absoluuttiset keskivirheet (MAPE) ja niiden erotukset

Tulokset ovat samansuuntaisia, jos perusmallina käytetään ensimmäisen asteen kausivaihtelu-autoregressiivistä mallia. Mallin estimoidut tulokset sekä ennustevirheiden erotukset on kuvattu liitteessä 3. Tarkastelemalla mahdollisen kausivaihtelun huomioivaa mallia voimme poissulkea mahdollisuuden yhteisen kausivaihtelun aiheuttamasta korrelaatiosta.

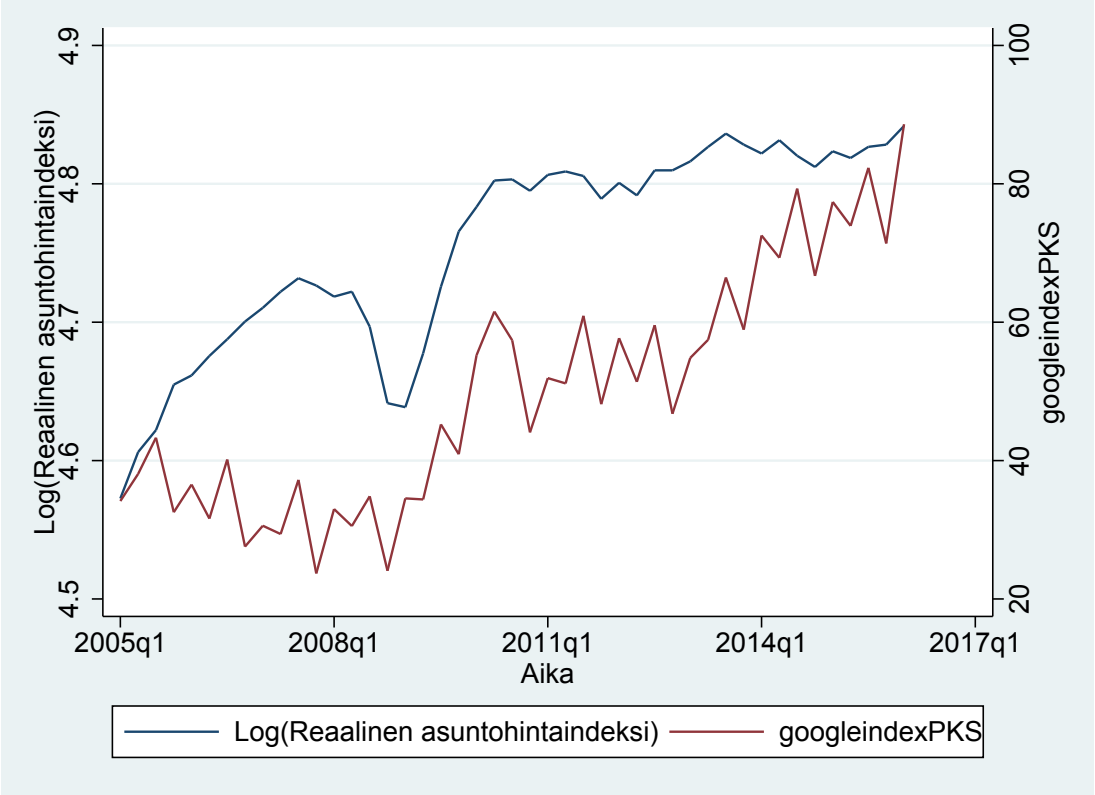
6 Pääkaupunkiseutu ja muu Suomi

Asuntomarkkinat ovat hyvin erilaiset esimerkiksi pääkaupunkiseudulla verrattuna muuhun Suomeen. Asuntomarkkinoissa on toki suuria eroja esimerkiksi kaupunkimaisen ja maaseutumaisen alueen välillä, toisaalta myös kaupunkialueiden välillä on suuria eroja. Olisi mielenkiintoista tarkastella mallin toimivuutta esimerkiksi kuntatasolla tai suurimpien kaupunkien tasolla. Kuten aiemmin todettiin Google-aineistoa ei ole saatavilla maantieteellisesti rajattuna pienempiin yksiköihin kuin koko Suomi ennen vuotta 2013.

Alueellisten Google-indeksien muodostaminen on kuitenkin mahdollista esimerkiksi hakusanojen valinnalla. Tekemällä alueellinen jaottelu hakusanojen valinnalla, tulee huomioitua myös muualta kyseiseen kaupunkiin

suuntautuva kiinnostus. Näin ei tapahdu Googlen aluerajausta käytettäessä, sillä aluerajaus rajaa haut niiden tekopaikan mukaan. Todellisuudessa helsinkiläisiä asuntoja etsitään myös Helsingin ulkopuolelta ja toisaalta helsinkiläiset etsivät asuntoja myös muualta kuin Helsingistä.

Pääkaupunkiseudun Google-indeksi muodostetaan aikaisemmin kuvatulla tavalla valitsemalla yksittäisiä hakusanoja. Asuntomarkkinoita kuvaava aineisto pääkaupunkiseudulle on valittu samoin kuin aikasemmassa koko maan tapauksessa, kuitenkin siten, että mukana on vain pääkaupunkiseutu. Lisäksi otos alkaa vuodesta 2005 vuoden 2004 sijaan. Pääkaupunkiseudun Google-indeksi sekä reaalinen asuntohintaindeksi on esitetty kuvassa 7. Muun Suomen Google-indeksi muodostetaan käyttämällä aikaisempaa koko maan Google-indeksiä, siten että siitä poistetaan kaikki haut, jotka sisältävät jonkin sanoista Helsinki, Espoo, Vantaa tai Kauniainen. Pääkaupunkiseudun ja muun Suomen aineistolla käytän mallina liitteessä 3 esitettyä kausivaihtelu-AR(1)-mallia.



Kuva 7: Pääkaupunkiseudun Google-indeksi sekä reaalinen asuntohintaindeksi

Pääkaupunkiseudun aineistolla estimoiduissa malleissa Google-indeksi näyttää olevan tilastollisesti merkitsevä muuttuja 10% tasolla. Lisäksi Google-indeksin lisääminen parantaa selitysastetta sekä informaatiokriteereitä myös Pääkaupunkiseudun mallissa. Pääkaupunkiseudulle muodostettu Google-indeksi näyttäisi siis parantavan mallin sovitetta.

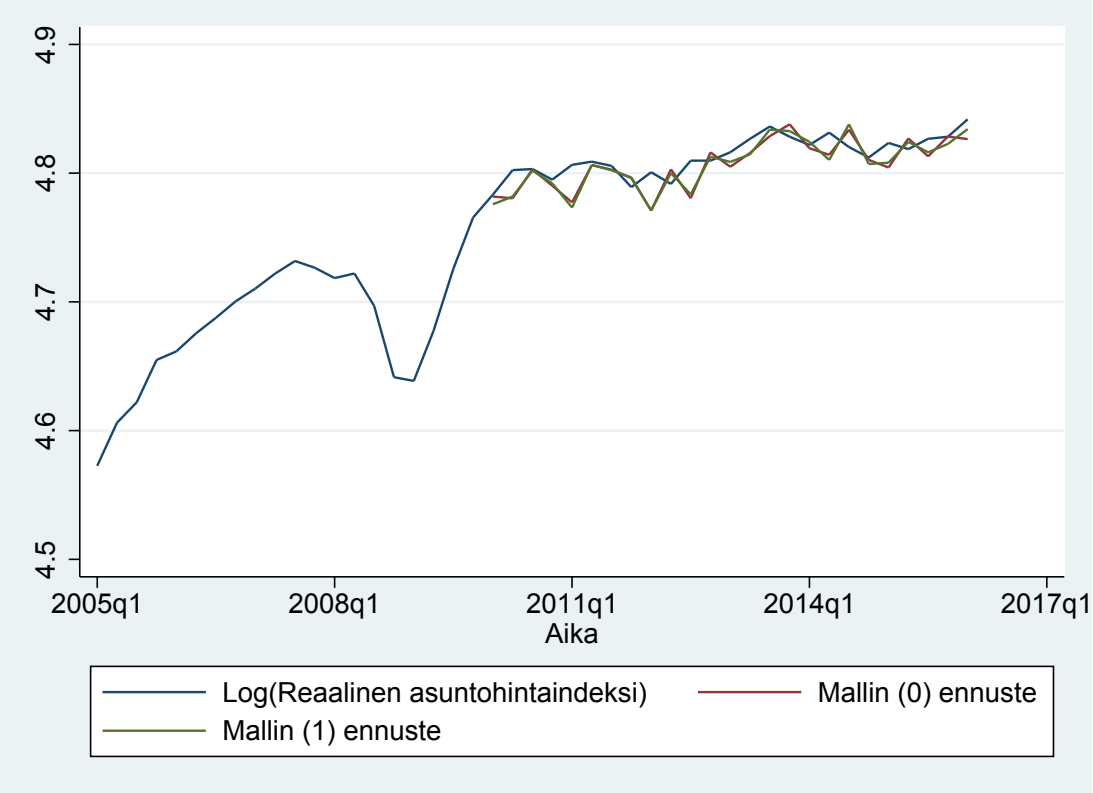
Malli	(00)	(01)
Selittäjät		
$log(y_{t-1})$	1.123^{***}	1.125^{***}
	(0.065)	(0.063)
$log(y_{t-4})$	-0.141^{**}	-0.145^{***}
	(0.0588)	(0.052)
x_t		0.00051^{*}
		(0.00030)
Vakio	4.92^{***}	4.89^{***}
	(0.105)	(0.098)
Yhteenveto		
R^2	0.878	0.885
AIC	−233.57	−236.230
BIC	−224.65	−225.53
N	45	45

Semirobustit keskvirheet suluissa

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Taulukko 7: Mallien (00) ja (01) tulokset Pääkaupunkiseudulle

Aivan kuten koko aikaisemmin esitetyssä koko maan mallissa, myös pääkaupunkiseudun mallissa mallin ulkopuolinen ennustekyky paranee Google-indeksin lisäämisen myötä. Absoluuttisella keskivirheellä mitattuna nykyhetken ennuste tarkentuu noin 4,8% Google-indeksin lisäämisen myötä. Toisin kuin aikaisemmin esitetyssä koko maan mallissa, nyt parannus ei ole Diebold–Mariano -testillä mitattuna tilastollisesti merkitsevä.



Kuva 8: Mallien ennusteet ja reaalinen asuntohintaindeksi pääkaupunkiseudulla

Muulle suomelle estimoidun mallin tulokset ovat samankaltaiset kuin Pääkaupunkiseudun mallissa. Google-indeksin lisääminen tarkentaa mallin sovitetta sekä parantaa selitysastetta ja informaatiokriteereitä. Mallin ulkopuolinen nykyhetken ennuste tarkentuu lähes 24%, joka on enemmän kuin Pääkaupunkiseudun mallissa. Ennusteen tarkentuminen ei kuitenkaan ole tilastollisesti merkitsevä Diebold–Mariano -testillä.

7 Johtopäätökset

Tarkka asuntohintojen nykyhetken tuntemus on tärkeää esimerkiksi päätöksentekijöille ja kiinteistönvälittäjille. Tässä raportissa on selvitetty, auttavatko Google-haut ennustamaan asuntojen hintojen nykyhetkeä ja lähitulevaisuutta. Aikaisemmin Suomen aineistolla Tuhkuri (2014) on osoittanut Google-aineiston parantavan työttömyyden nykyhetken ja lähitulevaisuuden ennustetta. Tämä on ensimmäinen tutkimus, joka osoit-

taa että Google-aineiston käyttö parantaa myös asuntohintojen nykyhetken ennustetta verrattaessa yksinkertaiseen autoregressiiviseen malliin.

Yksinkertaiseen perusmalliin nähden Google-hauilla laajennetun mallin ennustetarkkuus näyttää paranevan hieman. Vaikka parannus on tilastollisesti merkitsevä, on se silti melko pieni. Tulokset kuitenkin osoittavat Google-aineiston olevan käyttökelpoinen selittäjä makrotaloudellisen indikaattorin ennustamisessa. Tässä raportissa on selvitetty, onko uudentyylisestä datasta apua ennustetarkoituksessa, ja näyttää siltä, että Google-hauista on todella apua asuntojen hintojen ennakoimisessa. On kuitenkin otettava huomioon, että käytetyt mallit ovat vielä hyvin yksinkertaisia eikä kovin pitkälle meneviä johtopäätöksiä voida vielä vetää. Tulokset ovat kuitenkin rohkaisevia jatkotutkimusta ajatellen.

Tutkimukset (kts. esim. Kulkarni et al., 2009; McLaren & Shanbhogue, 2011; Wu & Brynjolfsson, 2013) ovat osoittaneet, että Google-haut voivat olla hyödyllisiä asuntohintoja ennustettaessa. Tässä raportissa on esitetty vastaavanlaisin menetelmin Google-hakujen auttavan asuntohintojen ennustamisessa myös Suomessa. Raportissa on osoitettu Google-aineiston parantavan sekä nykyhetken että lähitulevaisuuden ennustetta. Aikaisempien tutkimusten tapaan huomataan, että Google-haut auttavat jopa enemmän lähitulevaisuuden kuin nykyhetken ennustamisessa.

Vaikka Google-aineiston voidaan todeta auttavan talouden ennustamisessa, liittyy siihen silti myös haasteita. Tässä tutkimuksessa ei ole esimerkiksi pohdittu, miten ihmisten Internet-käyttäytyminen muuttuu ajassa. Varmaa on, että Internetin käytön yleistyttyä ihmisten tapa hakea asioita ja toisaalta käytetyt hakusanat ovat muuttuneet. Voikin olla, että Google-indeksi, joka toimii nyt, ei toimi samalla tavalla viiden vuoden päästä.

Kirjallisuus

Arrow, K. J. (1987). Reflections on the essays. Teoksessa *Arrow and the foundations of the theory of economic policy* (s. 727–734). Springer.

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, *55*(2), 107–120.

Brynjolfsson, E., Geva, T., & Reichman, S. (2015). Crowd-squared: amplifying the predictive power of search trend data. *Brynjolfsson, E., Geva, T., & Reichman, S., Crowd-Squared: Amplifying the Predictive Power of Search Trend Data. MIS Quarterly (Forthcoming)*.

Case, K. E., & Shiller, R. J. (1989, March). The Efficiency of the Market for Single-Family Homes. *American Economic Review*, *79*(1), 125-37.

Case, K. E., & Shiller, R. J. (1990). Forecasting prices and excess returns in the housing market. *Real Estate Economics*, *18*(3), 253–273.

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, *63*, 67–80.

Choi, H., & Varian, H. (2009a). *Predicting initial claims for unemployment benefits.* Citeseer.

Choi, H., & Varian, H. (2009b). *Predicting the present through google search queries.* April.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, *88*(s1), 2–9.

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*(366a), 427–431.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & economic statistics*.

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

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), 1012–1014.

Glaeser, E. L., & Gyourko, J. (2006). Housing dynamics. *NBER Working Paper Series*, 12787.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.

Horrigan, J. B., & Vitak, J. (2008). *The internet and consumer choice: online americans use different search and purchase strategies for different goods*. Pew Internet & American Life Project.

Juntto, A. (2008). Asumisen muutos ja tulevaisuus. rakennetarkastelu: erilaistuva asuminen. osaprojekti 1.

Kulkarni, R., Haynes, K. E., Stough, R. R., & Paelinck, J. H. (2009). Forecasting housing prices with google econometrics. *GMU School of Public Policy Research Paper*(2009-10).

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*(1-3), 159–178.

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

McLaren, N., & Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*(2011), Q2.

Oikarinen, E., & Engblom, J. (2014). Regional differences in housing price dynamics: panel data evidence. *Aboa Centre for Economics, Discussion Paper No. 94*.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., & Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical infectious diseases*, *47*(11), 1443–1448.

Scott, S. L., & Varian, H. R. (2013). Bayesian variable selection for nowcasting economic time series. *NBER Working Paper*(w19567).

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

Swanson, N. R., & White, H. (1997). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics*, *79*(4), 540–550.

Tierney, H. L., & Pan, B. (2012). A poisson regression examination of the relationship between website traffic and search engine queries. *NETNOMICS: Economic Research and Electronic Networking*, *13*(3), 155–189.

Tuhkuri, J. (2014). Big data: Google-haut ennustavat työttömyyttä suomessa. *Etlan raportit*(31).

Tuhkuri, J. (2016). Etlanow: A model for forecasting with big data–forecasting unemployment with google searches in europe. *Etlan raportit*(54).

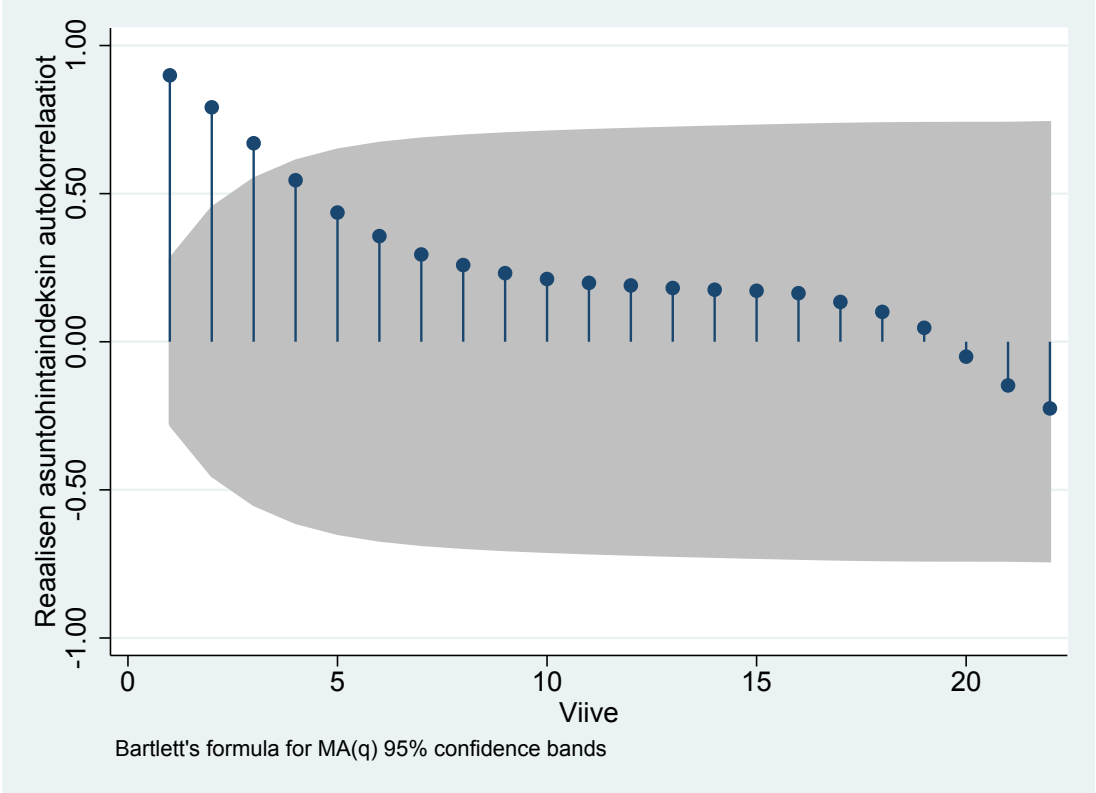
Verbeek, M. (2008). *A guide to modern econometrics*. John Wiley & Sons.

Vicente, M. R., López-Menéndez, A. J., & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change*, *92*, 132–139.

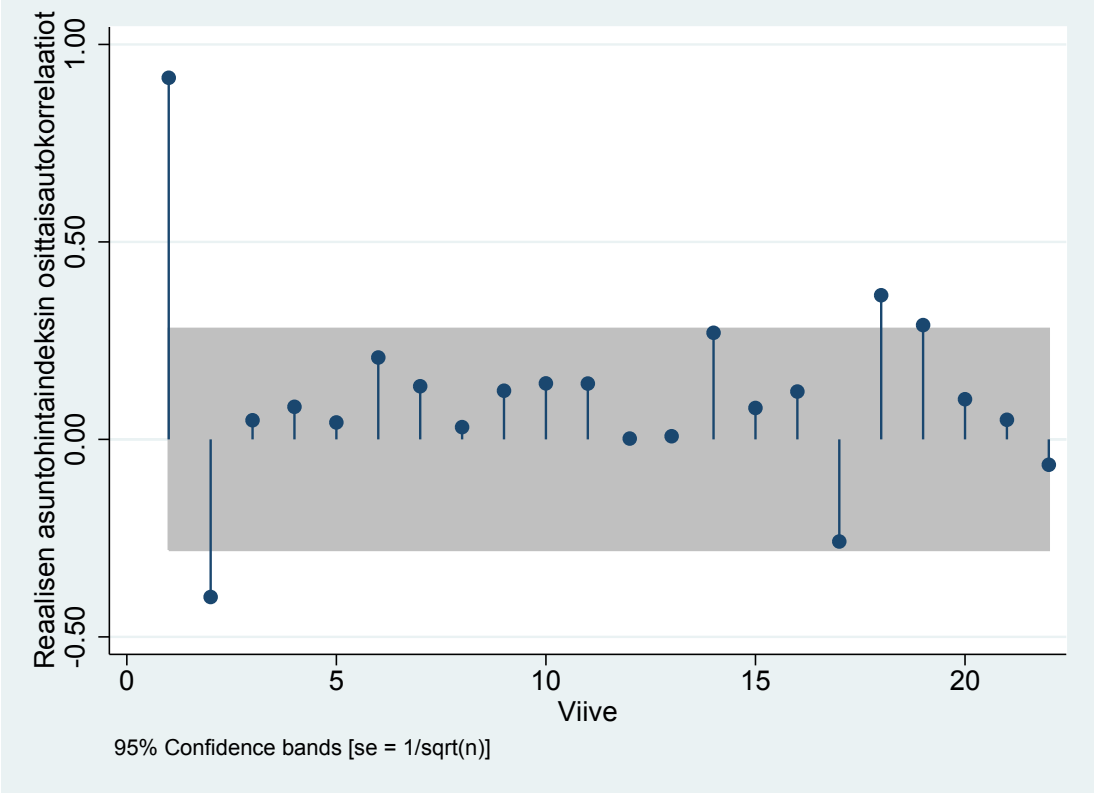
Wu, L., & Brynjolfsson, E. (2013). The future of prediction: How google searches foreshadow housing prices and sales. *Available at SSRN 2022293*.

Liitteet

Liite 1: Reaalisen asuntohintaindeksin autokorrelaatio- ja osittaisautokorrelaatiofunktio

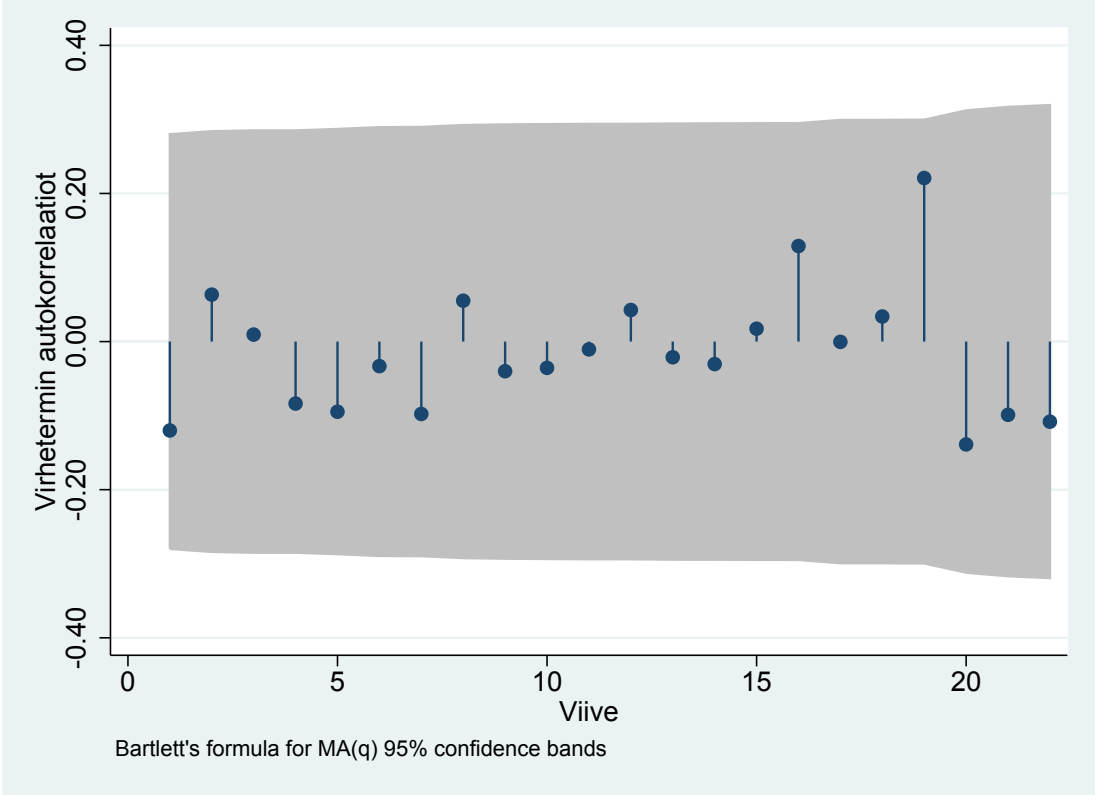


Kuva A1: Reaalisen asuntohintaindeksin autokorrelaatiofunktio

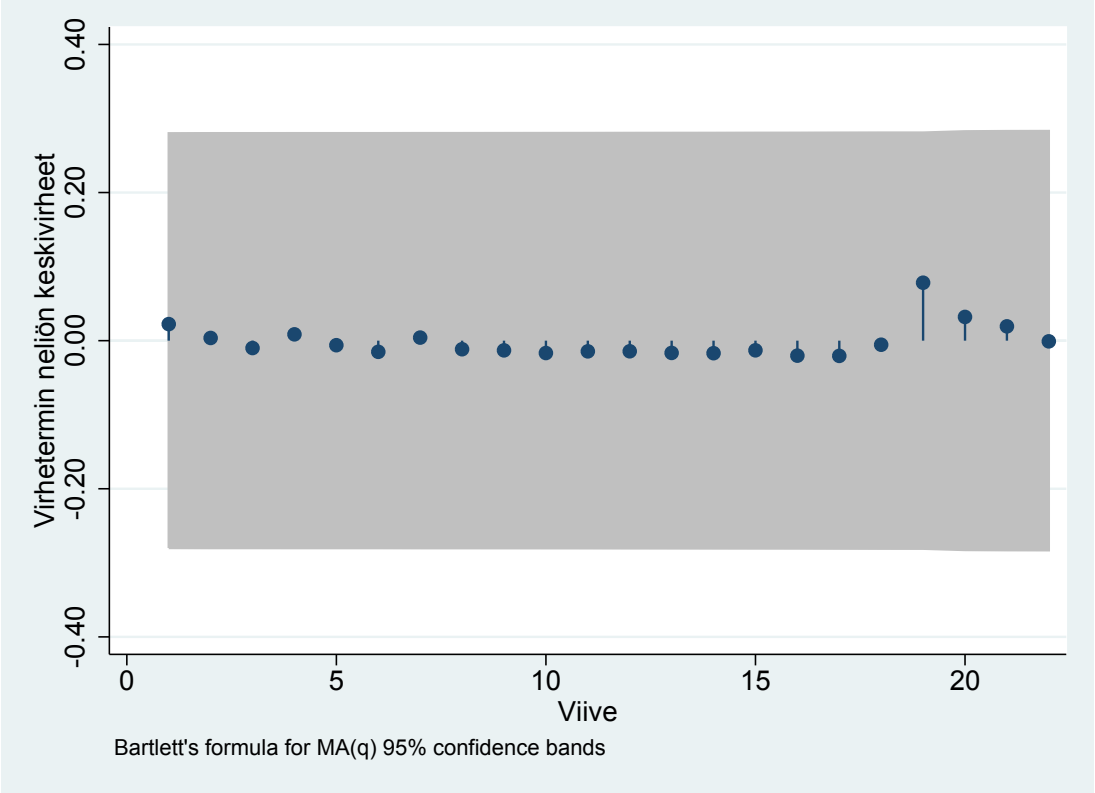


Kuva A2: Reaalisen asuntohintaindeksin osittaisautokorrelaatiofunktio

Liite 2: Estimoidun perusmallin residuaalien autokorrelaatio ja residuaalien neliöiden autokorrelaatio



Kuva A3: Perusmallin virhetermin autokorrelaatiofunktio



Kuva A4: Perusmallin virhetermin neliön autokorrelaatiofunktio

Liite 3: Kausivaihtelu AR(1)-mallin estimoidut tulokset

Tässä liitteessä on esitelty estimoidut tulokset kausivaihtelutermin sisätäville malleille koko maan aineistolla.

$$\text{Malli (00):} \quad log(y_t) \sim log(y_{t-1}) + log(y_{t-4}) + \varepsilon_t \tag{1a}$$

$$\text{Malli (01):} \quad log(y_t) \sim log(y_{t-1}) + log(y_{t-4}) + x_t + \varepsilon_t, \tag{1b}$$

Malli	(0)	(1)
Selittäjät		
$log(y_{t-1})$	1.132^{***}	1.138^{***}
	(0.0602)	(0.0687)
$log(y_{t-4})$	-0.161^{**}	-0.168^{**}
	(0.071)	(0.0841)
x_t		0.000473^{***}
		(0.000166)
Vakio	4.870^{***}	4.844^{***}
	(0.0896)	(0.0900)
Yhteenveto		
R^2	0.856	0.870
AIC	−264.20	268.63
BIC	−256.63	−259.17
N	49	49

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Taulukko A1: Mallien (00) ja (01) tulokset koko maan aineistolla

Aikaisemmin ilmestynyt ETLA Raportit-sarjassa (ennen ETLA Keskusteluaiheita)

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

No 48 *Jesper Bagger – Mika Maliranta – Niku Määttänen – Mika Pajarinen*, Innovator Mobility in Finland and Denmark. 13.1.2016. 20 p.

No 49 *Paavo Suni – Vesa Vihriälä*, Finland and Its Northern Peers in the Great Recession. 15.1.2016. 33 p.

No 50 *Antti Kauhanen – Vesa Vihriälä*, Työn määrä: Miksi Suomessa pitäisi tehdä enemmän työtä? 18.2.2016. 29 s.

No 51 *Tero Kuusi – Mika Pajarinen – Petri Rouvinen – Tarmo Valkonen*, Arvio t&k-verokannusteen vaikutuksista yritysten toimintaan Suomessa. 11.3.2016. 55 s.

No 52 *Joonas Tuhkuri – Hans Lööf – Ali Mohammadi – Petri Rouvinen*, Offshoring R&D. 4.5.2016. 13 p.

No 53 *Jyrki Ali-Yrkkö – Timo Seppälä – Juri Mattila*, Suurten yritysten ja niiden arvoketjujen rooli taloudessa. 18.5.2016. 37 s.

No 54 *Joonas Tuhkuri*, ETLAnow: A Model for Forecasting with Big Data: Forecasting Unemployment with Google Searches in Europe. 25.5.2016. 16 p.

No 55 *Klaus Castren – Aleksi Kortelainen – Timo Seppälä*, Rajaresurssien puute hidastaa teollisen internetin alustaekosysteemien syntyä. 26.8.2016. 12 s.

No 56 *Niku Määttänen – Olli Ropponen*, Listaamattomien yhtiöiden osinkoverotus, tuotantopanosten allokaatio ja tuottavuus. 26.8.2016. 16 s.

No 57 *Kristian Lauslahti – Juri Mattila – Timo Seppälä*, Älykäs sopimus – Miten blockchain muuttaa sopimuskäytäntöjä? 12.9.2016. 29 s.

No 58 *Antti Tahvanainen – Peter Adriaens – Annu Kotiranta*, Growing Pains of Industrial Renewal: Case Nordic Cleantech. 26.9.2016. 59 p.

No 59 *Hannu Karhunen – Niku Määttänen – Roope Uusitalo*, Opintotukijärjestelmän uudistaminen: Rakenteelliseen malliin perustuvia vaikutuslaskelmia. 10.10.2016. 26 s.

No 60 *Mika Maliranta – Niku Määttänen – Mika Pajarinen*, Firm Subsidies, Wages and Labor Mobility. 13.10.2016. 18 p.

No 61 *John Zysman – Martin Kenney*, The Next Phase in the Digital Revolution: Platforms, Abundant Computing, Growth and Employment. 17.10.2016. 21 p.

No 62 *Jyrki Ali-Yrkkö – Petri Rouvinen – Pekka Sinko – Joonas Tuhkuri*, Suomi globaaleissa arvoketjuissa. 30.11.2016. 41 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
Arkadiankatu 23 B
00100 Helsinki

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

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