Informaation haku ja asuntomarkkinat: Ennustaako Google markkinoiden suunnan?

Timo Aleksi Järvenpää

Helsingin yliopisto
Valtiotieteellinen tiedekunta, taloustiede

Pro gradu -tutkielma

Huhtikuu 2017

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta	Laitos
Valtiotieteellinen tiedekunta	Politiikan ja talouden tutkimuksen laitos

Tekijä
Timo Järvenpää

Työn nimi
Informaation haku ja asuntomarkkinat: Ennustaako Google markkinoiden suunnan?

Oppiaine
Taloustiede

Työn laji	Aika	Sivumäärä
Pro gradu -tutkielma	Huhtikuu 2017	67

Tiivistelmä

Asunnot muodostavat merkittävän osan suomalaisten varallisuudesta. Tässä tutkielmassa esitellään asuntomarkkinoiden informaationhakumalli, jonka avulla muodostetaan intuitio siitä, miten asuntoa hankkivien hakuaktiivisuus vaikuttaa kolmeen asuntomarkkinoiden muuttujaan: asuntojen hintoihin, asuntokauppojen lukumääriin sekä asuntojen myyntiaikoihin. Mallin tuomaa intuitiota hyödynnetään selvittämällä, auttaako Google Trends -hakuindeksi ennustamaan edellä mainittuja muuttujia Suomen asuntomarkkinoilla. Aiemmassa tutkimuksessa Google-hakujen on havaittu auttavan ennustamaan moninaisia talouden ilmiöitä. Hakuaktiivisuuden nousun havaitaan teoreettisen mallin perusteella lisäävän asuntokauppojen lukumäärää ja lyhentävän myyntiaikoja, mutta vaikutus asuntojen hintoihin on epävarma.

Tutkielman empiirisessä osiossa tutkitaan Granger-kausaalisuuden avulla, sisältävätkö Google-haut ennustamisen kannalta hyödyllistä informaatiota asuntomarkkinamuuttujista. Kustakin muuttujasta muodostetaan myös yksinkertainen, koko saatavilla olevaan historiaan sovitettava autoregressiivinen vertailumalli, josta tehdään Google-indeksillä laajennettu versio. Vertailumalleja ja Google-indeksillä laajennettuja malleja verrataan korjatun selitysasteen sekä Akaiken ja Schwarzin informaatiokriteereiden avulla. Google-hakujen ennustekykyä arvioidaan jakamalla data estimointiperiodiin ja ennusteperiodiin sekä simuloimalla reaaliaikaista ennustamista. Tutkielmassa analysoidaan seitsemän erilaista Google-haut sisältävää ennustespesifikaatiota.

Google-hakujen havaitaan Granger-aiheuttavan hintoja ja markkinointiaikoja. Koko historiaan sovitettujen autoregressiivisten mallien perusteella Google-hakutermien kertoimet eivät noudata johdonmukaisesti teoreettisen mallin mukaisia merkkejä. Sekä markkinointiaika- että lukumäärämalleissa Google-termit saavat sekä negatiivisia että positiivisia arvoja. Google-hakujen havaitaan parantavan nykyisyyden hintaennusteita absoluuttisella keskivirheellä mitattuna yhtä lukuun ottamatta kaikilla spesifikaatioilla, mutta ennustevirheiden erot eivät Diebold-Mariano-testin perusteella pääsääntöisesti kuitenkaan eroa tilastollisesti merkitsevästi nollasta. Lukumäärien nykyisen arvon ennusteissa Google-haut tuottavat useassa spesifikaatiossa merkittävästi suurempia ennustevirheitä kuin vertailumallit. Yhden kuukauden päähän ennustettaessa internethaut kuitenkin vaikuttavat pienentävän lukumäärien ja hintojen ennustevirheitä. Paneelidataspesifikaatiolla sekä hinta- että lukumääräennusteet ovat tarkempia internethakuja hyödyntämällä. Tulosten perusteella Google-hakujen hyödyllisyys asuntomarkkinoiden ennustamisessa on altis mallin spesifikaatiolle eivätkä Google-haut pysty johdonmukaisesti parantamaan ennusteita kaikilla muuttujilla.

Avainsanat
Google Trends, big data, aikasarja-analyysi, ennustaminen, paneelidata, asuntomarkkinat

Sisältö

Luku 1

Johdanto

Asuntovarallisuutta on noin 70 prosentilla suomalaisista kotitalouksista, ja asunnot muodostavat noin 69 prosenttia suomalaisten kotitalouksien kokonaisvarallisuudesta (Tilastokeskus 2013). Asunnon ostaminen tai myyminen koskettaa siis hyvin suurta osaa suomalaisista paitsi henkilökohtaisesti, myös laajemmin makrotalouden tasolla. Asunnot toimivat esimerkiksi vakuutena uudelle yritystoiminnalle (esim. Black, Meza & Jeffreys 1996; Schmalz, Sraer & Thesmar 2017) ja toisaalta omistusasunnot voivat osaltaan vaikuttaa työvoiman liikkuvuuteen (esim. Tervo 2000). Ajantasaisten ja tarkkojen tietojen saamisella asuntomarkkinoiden nykytilanteesta ja lähitulevaisuudesta on siten suurta merkitystä monille talouden toimijoille.

Asunnot ovat monimutkaisia hyödykkeitä, minkä vuoksi asunnon ostoa luultavasti edeltää pitkä informaationhakuprosessi. Esimerkiksi asuntoja välittävä oikotie.fi-palvelulla oli vuoden 2017 viikolla 12 yli 700 000 kävijää[1], joten internethakudatan potentiaali asuntomarkkinoiden ennakoijana on ilmeinen. Aiemmin esimerkiksi Wu ja Brynjolfsson (2014) ja McLaren ja Shanbhogue (2011) ovat havainneet internethakudatan parantavan lyhyen aikavälin asuntomarkkinaennusteita.

Esittelen tässä tutkielmassa Wheatonin (1990) kehittämän asuntomarkkinoiden informaationhakumallin, joka auttaa muodostamaan intuition siitä, miten asuntoa etsivien hakuaktiivisuus vaikuttaa keskeisiin asuntomarkkinoiden muuttujiin. Teoria muodostaa suunnan tutkielman empiiriselle osiolle, jossa tutkin, auttavatko Google-haut ennustamaan Suomen asuntomarkkinoita. Internethakudatan kautta tutkielma kytkeytyy osaksi jo varsin laajaa kirjallisuutta, jossa etsitään yhä uusia ennustekohteita ajantasaiselle hakudatalle.

Tutkielman rakenne on seuraava. Toisessa luvussa johdan asuntomarkkinoiden informaationhakumallin. Kolmannessa luvussa kuvailen aiempaa internethakuihin pohjaavaa ennustekirjallisuutta. Neljännessä luvussa esittelen tutkielmassa käytetyn datan. Viidennessä luvussa ovat empiirisen tutkimuksen metodit. Kuudes luku

[1] `http://tnsmetrix.tns-gallup.fi/public/`

esittelee tulokset, joiden robustisuutta tarkastellaan seitsemännessä luvussa. Viimeisenä luvussa kahdeksan tulevat johtopäätökset.

Luku 2

Informaatio, asuntomarkkinat ja Google

Tämä luku jakautuu kolmeen osaan. Ensimmäisessä luvun osassa kuvailen asuntomarkkinoiden erityispiirteitä. Toisessa osassa käsittelen lyhyesti informaation taloustiedettä yleisellä tasolla sekä teoriaa soveltavaa empiiristä tutkimusta Suomessa. Kolmannessa osassa esittelen Wheatonin (1990) informaationhakumallin asuntomarkkinoille. Kolmannen osan tarkoitus on tarjota intuitio sille, miksi Google-haut mahdollisesti voivat auttaa ennustamaan asuntomarkkinoita.

Asuntomarkkinoiden erityispiirteitä

Asunnoilla on kulutus- ja investointihyödykkeinä useita piirteitä, jotka tekevät asuntomarkkinoista epätäydelliset. Asunnot ovat esimerkiksi heterogeenisiä, erittäin kestäviä ja asuntojen sijainnilla on suuri merkitys. Lisäksi julkinen sektori puuttuu usein voimallisesti asuntomarkkinoihin esimerkiksi verotuksella, tuilla ja sääntelyllä. (Smith, Rosen & Fallis 1988) Asuntojen hintoihin vaikuttavia seikkoja ovat kysyntäpuolella tulotaso, väestörakenteen muutokset sekä rahoituksen saatavuus. Tarjontapuoleen vaikuttavat maan hinta, rakennuskustannukset ja sääntely. (Malpezzi 1996.)

Oikarinen (2005) esittää tärkeimmiksi Suomen pääkaupunkiseudun asuntohintoihin vaikuttaviksi tekijöiksi reaalikoron, tulotason ja tulo-odotukset. Holappa et al. (2015) puolestaan käyttävät ennusteessaan muuttujina korkotasoa ja kotitalouksien tulokehitystä sekä epävarmuutta kuvaavia muuttujia, kuten työttömyysasteen ja pörssin yleisindeksin volatiliteetin muutosta.

Oikarinen (2012) havaitsee, että asuntojen myyntimäärät reagoivat kysyntäshokkeihin huomattavasti hintoja nopeammin. Ilmiö viitannee siihen, että myyjät reagoivat ostajia hitaammin markkinatilanteen muutoksiin. Myyntimäärät voivat siis

mahdollisesti ennustaa tulevaa hintakehitystä.

Wu ja Brynjolfsson (2014) jaottelevat asuntomarkkinoiden ennustemallit fundamenttipohjaisiin ja teknisiin ennusteisiin. Näistä ensimmäinen pyrkii ennustamaan edellä mainittujen muuttujien (esimerkiksi tulotaso ja rakennuskustannukset) avulla, kun taas jälkimmäinen keskittyy tilastollisiin säännönmukaisuuksiin.

Informaation ja tiedon etsimisen merkitys kuluttajamarkkinoille

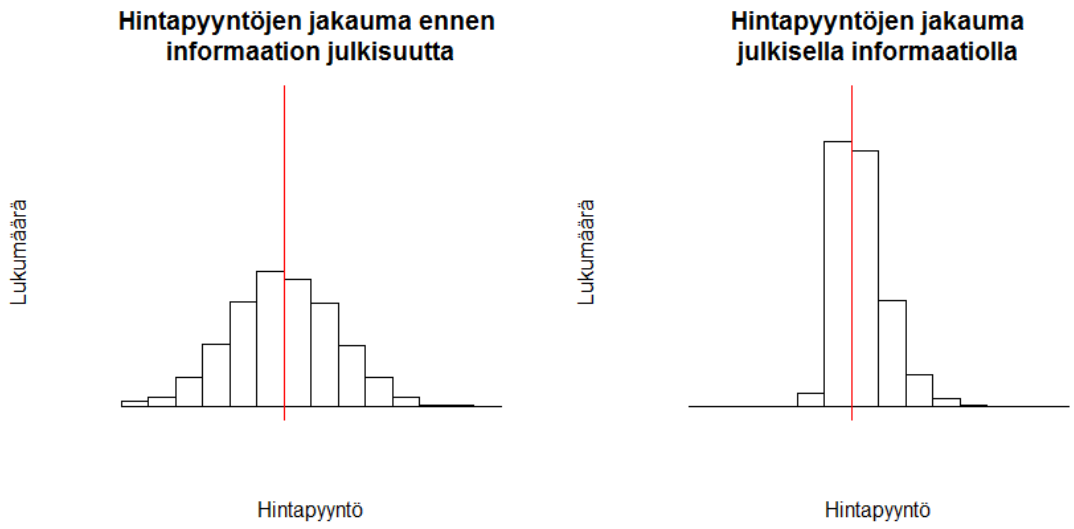
Taloustieteessä informaation merkitystä markkinoiden toiminnalle on tutkittu 1960-luvun alusta lähtien, jolloin George Stigler julkaisi artikkelinsa The Economics of Information (1961). Artikkeli tarkastelee homogeenisen tuotteen hintaeroja, ja käyttää termiä etsiminen toiminnasta, jolla ostaja (myyjä) käy läpi eri myyjiä (ostajia) etsiessään itsellensä edullisinta kauppaa. Etsiminen tuottaa kustannuksia. Mitä suuremmat hintaerot ovat, sitä kannattavampaa etsiminen on markkinaosapuolelle. Kuluttaja etsii informaatiota, kunnes etsimisprosessin rajakustannus vastaa etsimisen odotettua rajahyötyä.

Nelson (1970) soveltaa informaation ja tiedon etsimisen merkitystä tilanteessa, jossa tuotteilla on laatueroja. Hän erottelee tuotteet sen mukaan, voiko tuotteen laadun määrittää kokemuksen kautta vai täytyykö tuotteen laadun määrittämiseen käyttää tiedon etsimistä. Hänen hypoteesinsa mukaan tiedon etsiminen tuottaa kustannuksia, joten joidenkin tuotteiden kohdalla laadun määrittäminen voi olla järkevämpää kokeilemalla eri tuotemerkkejä. Melko edullisten ja helposti saatavilla olevien tuotteiden kohdalla kokemuspohjainen päätöksenteko voi olla järkevämpää, kun taas kalliimpien ja harvemmin hankittavien hyödykkeiden hankinta nojaa luultavammin tiedonhankintaan.

Eerola ja Lyytikäinen (2015) selvittävät yksityiskohtaisen hintainformaation julkisuuden vaikutusta kerrostaloasuntojen kauppahintoihin ja myyntiaikoihin Helsingin seudulla. He havaitsevat, että informaation julkisuus on johtanut hintojen kallistumiseen ja toisaalta myyntiaikojen lyhenemiseen. Teoreettinen pohja ilmiölle perustuu jaksottaiseen kaupankäyntimalliin, jossa myyjäosapuoli asettaa asunnollensa hinnan ja ostajaehdokas joko hyväksyy tai hylkää tarjouksen. Jos hinta on alle ostajaehdokkaan maksuvalmiuden, kauppa syntyy. Muussa tapauksessa myyntiprosessi etenee seuraavaan periodiin, jolloin myyjä voi halutessaan muuttaa hintapyyntöään. Liian alhaiseksi hinnan asettavat myyjät eivät siis saa mahdollisuutta oppia todellista markkinahintaa, sillä asunto myydään jo ensimmäisellä kierroksella.

Asuntohintojen julkisuus vähentää Eerolan ja Lyytikäisen (2015) mallissa myyjän kohtaamaa epävarmuutta, sillä myyjä voi verrata asuntoaan aiemmin myytyi-

hin verrokkiasuntoihin. Asuntonsa hinnan aliarvioineet myyjät nostavat julkisuuden seurauksena hintojaan, kun taas hinnan yliarvioineet laskevat hintoja. Koska asuntoja ei juurikaan enää myydä alle oikean markkinahinnan, toteutuneiden kauppojen keskihinta nousee. Kuva 2.1 havainnollistaa muutosta. On tärkeää huomioida, että mallissa informaation keskeinen käyttäytymisvaikutus kohdistuu ensisijaisesti myyjäosapuoleen eli asuntojen tarjontaan. Kysyntäosapuolen hallussa oleva informaatio saattaa tuottaa erilaisia vaikutuksia markkinaolosuhteisiin.



Kuva 2.1: Vasen kuvio kuvaa tilannetta ennen informaation julkistusta. Markkinahinta on punaisen pystyviivan kohdalla, ja hintapyynnöt sen alle johtavat toteutuneeseen kauppaan. Markkinahintaa kalliimmat pyynnöt eivät johda kaupan toteutumiseen. Oikea kuvio kuvaa tilannetta informaation julkistamisen jälkeen. Nyt alihintaa pyytäviä on huomattavasti vähemmän. Koska hintapyyntöjen jakauman massa siirtyy alihinnoittelusta kohti markkinahintaa, toteutuneiden kauppojen keskihinta nousee.

Informaationhakumalli asuntomarkkinoilla

Tämän osion tarkoituksena on muodostaa intuitio siitä, miksi Google-haut mahdollisesti voivat auttaa ennustamaan asuntomarkkinoita. Seuraan tässä osiossa Wheatonin (1990) kehittämää asuntomarkkinoiden matching-mallia sekä esitän kyseisen mallin implikaatioita hakuaktiviteetin kasvusta asuntomarkkinoille. Mallissa on kahdenlaisia kotitalouksia, esimerkiksi yksin asuvia ja perheitä, sekä kahdenlaisia asuntoja, esimerkiksi pieniä ja suuria. Alaindeksit $i = 1, 2$ ja $j \neq i$ viittaavat kotitalouden tyyppiin. Osa kotitalouksista muuttuu yhden tarkasteluperiodin aikana tyypistä i

tyyppiin j. Merkitsemme tätä osuutta β_i:llä. Yhtäpitävästi β_i voidaan tulkita myös yksittäisen kotitalouden tyypin muuttumisen keskimääräiseksi todennäköisyydeksi annetun ajanjakson aikana.

Kaikki kotitaloudet omistavat vähintään yhden asunnon. Kukin tyypin i kotitalous voi olla yhdessä kolmesta eri tilasta: asua sopivassa asunnossa, asua epäsopivassa asunnossa tai asua sopivassa asunnossa, mutta omistaa edelleen vanhan, epäsopivan asunnon. Merkitsemme kussakin tilassa olevien kotitalouksien määrää lyhentein HM_i, HS_i ja HD_i. Siten tyypin i kotitalouksien kokonaismäärä on $H_i = HM_i + HS_i + HD_i$.

Muutokset tilasta toiseen etenevät seuraavasti. Itselleen sopivassa asunnossa asuvan, eli HM_j:hin kuuluvan kotitalouden tyyppi voi muuttua j:stä i:ksi todennäköisyydellä β_j. Muutoksen toteutuessa aiemmin sopiva asunto muuttuu epäsopivaksi ja siten kotitalous siirtyy HS_i:hin. Kotitalous alkaa etsiä ja lopulta ostaa uuden itselleen sopivan tyypin asunnon, muuttaa sekä laittaa vanhan asuntonsa myyntiin. Kotitalous omistaa siis myyntiaikana kaksi asuntoa, eli lukeutuu HD_i:hin. Kun vanha asunto on myyty, kotitalous on jälleen itselleen sopivassa asunnossa ja siten lasketaan HM_i:hin. Malli sallii myös epäsopivassa asunnossa asuvan tai kaksi asuntoa omistavan kotitalouden tyypin muuttumisen takaisin alkuperäiseen tyyppiin. Kotitalous siis voi myös siirtyä HS_i:stä takaisin HM_j:hin tai HD_i:stä HD_j:hin todennäköisyydellä β_i.

Koska kaikilla kotitalouksilla on vähintään yksi asunto, vapaana olevien asuntojen määrä eli asuntovarauma tyypin i kotitalouksille on $V_i = S_i - H_i$, jossa S_i on tyypin i kotitalouksille sopivien asuntojen kokonaismäärä. Oletamme, että vapaita asuntoja on aina ei-negatiivinen määrä. Oletamme myös, että epäsopivassa asunnossa asuvat kotitaloudet eivät välittömästi löydä uutta asuntoa, vaan se joudutaan etsimään asuntovaraumasta. Vapaiden asuntojen ja tilaan HS_i kuuluvien kotitalouksien kohtaaminen tapahtuu Poisson-prosessia noudattaen. Prosessin parametrina on

$$m_i \equiv m_i(E);\ m_i(0) = 0,\ m_i'(E) > 0,\ m_i''(E) < 0, \tag{2.1}$$

jossa E on hakemiseen käytetty panos. Funktion $m_i(E)$ muoto voidaan tulkita "etsimisteknologiaksi".[1] Toisin sanoen se kuvaa, kuinka tehokkaasti lisäpanostus etsimiseen lisää oikean asunnon löytymisen todennäköisyyttä. Koska m_i:n ensimmäinen derivaatta E:n suhteen on positiivinen, suurempi hakupanos johtaa suurempaan löytyneiden asuntojen määrään. Tyypin i ostettujen asuntojen kokonaismäärä on siten $Q_i = HS_i \times m_i(E)$. Asunnon myynti noudattaa myös Poisson-prosessia. Koska ostettujen ja myytyjen asuntojen määrän on vastattava toisiaan, kyseisen prosessin

[1] Wheaton (1990) esittää funktion muodossa $m_i(E, \frac{V_i}{S_i})$, eli parametrin arvo riippuu myös siitä, kuinka suuri osa asunnoista on vapaana. Mallin yksinkertaistamiseksi oletan kuitenkin m:n riippuvan vain hakuaktiivisuudesta.

parametri on

$$q_i(E) = \frac{HS_i \times m_i(E)}{V_i}. \tag{2.2}$$

Seuraavaksi määrittelemme, kuinka kotitaloudet muuttuvat tilasta toiseen yli ajan. Saamme seuraavan systeemin:

$$\dot{HS}_i = -m_i(E)HS_i - \beta_i HS_i + \beta_j HM_j, \tag{2.3}$$

$$\dot{HD}_i = -q_j(E)HD_i + m_i(E)HS_i + \beta_j HD_j - \beta_i HD_i, \tag{2.4}$$

$$\dot{HM}_i = -\dot{HS}_i - \dot{HD}_i,\ i = 1, 2,\ j \neq i. \tag{2.5}$$

Epäsopivassa asunnossa asuvien kotitalouksien määrän muutos muodostuu kolmesta osasta. Määrä pienenee, kun kotitaloudet joko löytävät sopivan asunnon tai muuttavat tyyppiään. Määrä puolestaan kasvaa, kun osa j-tyypin kotitalouksista muuttaa tyyppiään ja siirtyy HM_j:stä HS_i:hin. Kahden asunnon kotitaloudet vähenevät onnistuneen myyntiprosessin seurauksena ja lisääntyvät onnistuneen ostoprosessin seurauksena. Yhtälön kaksi viimeistä termiä kuvaavat kahden asunnon kotitalouksien muuttumista tyypistä toiseen. Koska kaikki kotitaloudet kuuluvat yhteen näistä kolmesta ryhmästä, saamme HM_i:n muutoksen laskemalla yhteen kahden edellisen muutoksen vastaluvut.

Olettaen, että erilaisten kotitaloustyyppien lukumäärä ja käyttäytyminen ovat samanlaisia, saamme yksinkertaistettua yllä olevaa kuuden yhtälön systeemiä. Jos $\beta_1 = \beta_2$, $V_1 = V_2$, $H_1 = H_2$ ja $m_1(E) \equiv m_2(E)$, malli on täydellisen symmetrinen. Voimme siten jättää huomiotta kotitalouden tyyppiä ilmaisevat alaindeksit ja päädymme kahden yhtälön systeemiin:

$$\dot{HS} = -HS(2\beta + m(E)) + \beta H - \beta HD, \tag{2.6}$$

$$\dot{HD} = m(E)HS(1 - \frac{HD}{V}). \tag{2.7}$$

Systeemi on tasapainossa, kun $\dot{HS} = 0$ ja $\dot{HD} = 0$. Tasapainoksi tulee siten

$$HS = \frac{\beta(H - V)}{2\beta + m(E)}, \tag{2.8}$$

$$HD = V. \tag{2.9}$$

Mallin tasapainossa kahden asunnon kotitalouksien määrä vastaa asuntovarauman asuntoja, eli yksikään asunto ei ole ilman omistajaa. Korkea täsmäysaste alentaa epäsopivassa asunnossa asuvien kotitalouksien määrää. Toisaalta pieni vapaiden asuntojen määrä suhteessa asuntojen kokonaismäärään nostaa epäsopivassa asunnossa asuvien määrää.

Yhdistämällä systeemin ratkaisusta yhtälö 2.8 Q:n määritelmään saamme asuntokauppojen lukumääräksi tasapainotilassa

$$Q(E) = \frac{\beta(H-V)m(E)}{(2\beta + m(E))}, \tag{2.10}$$

jonka ensimmäinen derivaatta E:n suhteen on

$$\frac{\partial Q}{\partial E} = \frac{2\beta^2(H-V)m'(E)}{(2\beta + m(E))^2} \geq 0. \tag{2.11}$$

Näemme, että suurempi hakupanostus kasvattaa ostettujen asuntojen määrää. Olettaen, että Google-hakuindeksin suuremmat arvot voisivat indikoida lisääntynyttä hakupanosta (suurempaa E:tä), saamme mallin ensimmäisen implikaation:

1. *Asunnon ostajien hakupanoksen suureneminen (suurempi Google-hakuindeksin arvo) johtaa suurempaan asuntokauppojen määrään.*

Lisäksi, koska myyntimäärä noudattaa Poisson-prosessia parametrilla q, aika asuntojen myyntien välillä noudattaa eksponenttijakaumaa parametrilla $1/q$. Eksponenttijakauman ominaisuuksien nojalla myyntien välisen ajan odotusarvo on $L = 1/q$ ja siten

$$\frac{\partial L}{\partial E} = -\frac{2Vm'(E)}{(H-V)m^2(E)} < 0, \tag{2.12}$$

eli suurempi hakupanostus lyhentää myyntien välisen ajan odotusarvoa. Saamme tästä toisen implikaation:

2. *Asunnon ostajien hakupanoksen suureneminen johtaa asuntojen lyhempiin myyntiaikoihin.*

Asuntojen hinnan määrittelemiseksi tarvitsemme vielä hakuun liittyvät kustannukset. Määrittelemme kustannukset hakuaktiivisuuden E funktiona seuraavasti:

$$c(E);\ c(0) = 0,\ c'(E) > 0,\ c''(E) > 0. \tag{2.13}$$

Seuraavaksi määrittelemme kunkin tilan arvon kotitalouksille. Olkoon UM kotitalouden sopivasta asunnosta yhden periodin aikana saama rahallinen hyöty sekä US kotitaloudelle sopimattomasta asunnosta saama hyöty. Seuraavat kolme yhtälöä kuvaavat nyt kunkin tilan tuottamaa arvoa yhden periodin aikana:

$$rWM = UM - \beta(WM - WS), \tag{2.14}$$

$$rWD = UM + q(E)(WM - WD + R), \tag{2.15}$$

$$rWS = US - c(E) + \beta(WM - WS) + m(E)(WD - WS - R). \tag{2.16}$$

Yhtälöissä WM, WD ja WS ovat tilojen nykyarvot (sopiva asunto, kaksi asuntoa, epäsopiva asunto), r on diskonttokorko, R on asunnon markkinahinta ja c, m, q ja E ovat kuten määritelty edellä.

Yhtälön 2.14 mukaan kotitalouden yhden periodin aikana saama tuotto sopivassa asunnossa asumisesta vastaa asunnosta koettua hyötyä. Hyödystä vähennetään arvon menetyksen odotusarvo, joka kuvaa mahdollisuutta, että kotitalous muuttaa tyyppiään sopivassa asunnossa asuessaan. Vastaavasti yhtälö 2.15 kuvaa kahden asunnon kotitalouden saamaa yhden periodin tuottoa. Se muodostuu oikeanlaisessa asunnossa asumisesta koetusta hyödystä sekä mahdollisesta pääoman lisääntymisestä, mikäli toisen asunnon myynti onnistuu. Epäsopivassa asunnossa asuvan kotitalouden saama tuotto muodostuu neljästä osasta. Ensimmäinen on epäsopivasta asunnosta koettu hyöty. Toinen, arvoa alentava osa on sopivan asunnon etsimisestä aiheutuva kustannus. Kolmas osa kuvaa arvon lisäyksen odotusarvoa, jos kotitalous vaihtaa tyyppiään nykyisessä asunnossa asuessaan. Viimeinen termi kuvaa ostoprosessin tuottaman arvon odotusarvoa, eli uuden asunnon löytymisen todennäköisyyttä kerrottuna arvon lisäyksellä, jonka kotitalous saa siirtymällä sopimattomasta asunnosta kahden asunnon kotitaloudeksi.

Asunnon markkinahinnan on asetuttava siten, että ostaja suostuu ostamaan asunnon ($R \leq WD - WS$) ja toisaalta myyjä suostuu myymään asunnon ($R \geq WD - WM$). Oletamme, että kummankin osapuolen neuvotteluvoima on yhtä suuri, jolloin kaupasta saatu hyöty jakautuu tasan sekä ostajalle että myyjälle. Nyt voimme ratkaista hinnan R yhtälöstä

$$\begin{aligned} WM - WD + R &= WD - WS - R \\ R &= \frac{WD - WS + WD - WM}{2}. \end{aligned} \tag{2.17}$$

Vähentämällä yhtälön 2.16 yhtälöstä 2.14 saamme

$$WM - WS = \frac{UM - US + c(E) - m(E)(WD - WS - R)}{2\beta + r} \tag{2.18}$$

ja käyttämällä hyväksi hinnoitteluehtoa 2.17 edelleen

$$WM - WS = \frac{UM - US + c(E) - m(E)(WM - WD + R)}{2\beta + r}. \tag{2.19}$$

Vastaavasti vähentämällä yhtälö 2.15 yhtälöstä 2.16 ja yhdistämällä yhtälöön 2.17 ja 2.19 saamme

$$WD - WS - R = \frac{[UM - US + c(E)](\beta + r)}{Z - \beta m(E)} - \frac{rR(2\beta + r)}{Z - \beta m(E)}, \tag{2.20}$$

jossa $Z = [r + m(E) - q](2\beta + r)$. Yhtälö 2.20 kuvaa siis kodinostajan saamaa nettohyötyä. Vastaavasti myyjän nettohyöty on

$$WM - WD + R = \frac{-\beta[UM - US + c(E)]}{X - \beta m(E)} + \frac{rR(2\beta + r)}{X - \beta m(E)}, \tag{2.21}$$

jossa $X = (r + q)(2\beta + r)$.

Yhtälöt 2.20 ja 2.21 ovat lineaarisia hinnan R suhteen, joten saamme ratkaistua hinnan R hakuaktiivisuuden E ja parametrien suhteen asettamalla yhtälöiden oikeat puolet yhtäsuuriksi yhtälön 2.17 mukaisesti:

$$R = [UM - US + c(E)]\frac{2\beta + r + q(E)}{r[4\beta + 2r + m(E)]}. \tag{2.22}$$

R:n derivaatta E:n suhteen on siten

$$\begin{aligned}\frac{\partial R}{\partial E} = {} & c'(E)\frac{2\beta + r + \frac{\beta m(E)(H-V)}{V(2\beta+m(E))}}{r[4\beta + 2r + m(E)]} - \frac{m'(E)[UM - US + c(E)]}{r[4\beta + 2r + m(E)]^2} \times \\ & [\frac{2\beta^2(H - V)}{V(2\beta + m(E))^2}(4\beta + 2r + m(E)) + 2\beta + r - \frac{\beta m(E)(H - V)}{V(2\beta + m(E))}].\end{aligned} \tag{2.23}$$

Derivaatan ensimmäisen termin merkki on aina negatiivinen, kun taas toisen termin merkki on epävarma. Emme voi siten päätellä derivaatan merkkiä suoraan olettamatta ensin eksplisiittisesti funktiomuotoja $c(E)$:lle ja $m(E)$:lle.

Voimme tiivistää tuloksen kolmanteen implikaatioon:

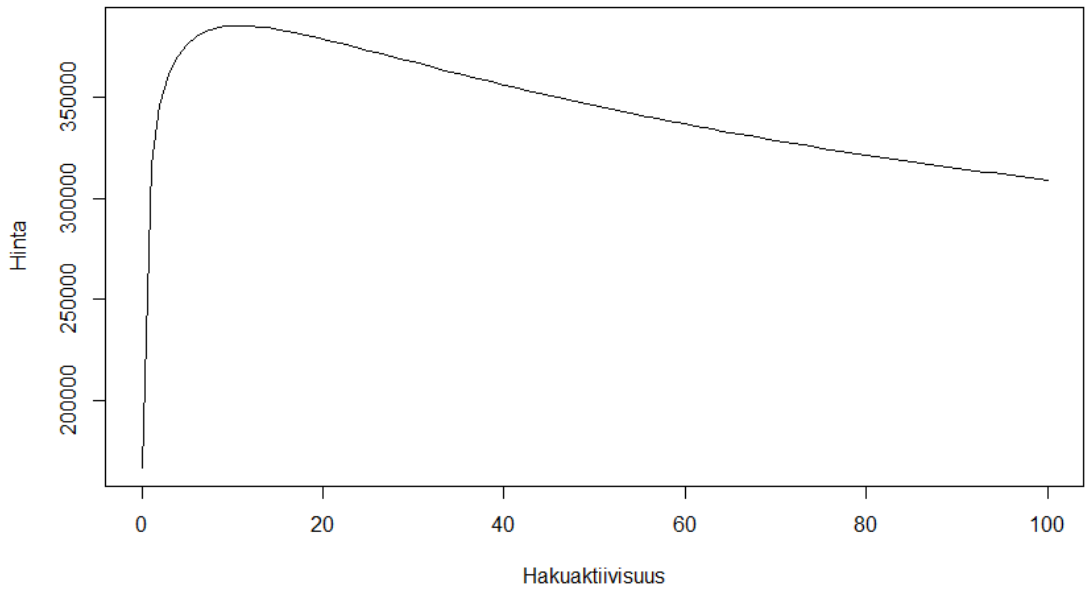
3. *Asunnon ostajien hakuaktiivisuuden vaikutus asuntojen hintoihin voi olla positiivinen tai negatiivinen.*

Esimerkiksi Wu ja Brynjolfsson (2014) ja Hohenstatt, Käsbauer ja Schäfers (2012) havaitsevat Google-muuttujan viiveen merkin olevan negatiivinen tekemissään asuntojen hintojen ennusteissa. Wu ja Brynjolfsson (2014) arvioivat, että merkki johtuu tarjontapuolen tulosta ensin markkinoille. Tällöin tehdyt haut indikoisivat, että asuntoaan myyvät kotitaloudet kartoittavat ensin asuntomarkkinoiden tilanteen. Runsas hakuaktiivisuus viittaa siten runsaaseen tarjontaan tulevaisuudessa. Tarjonnan noustessa hinnoilla on puolestaan taipumusta laskea. Hohenstatt, Käsbauer ja Schäfers (2012) tarjoavat merkkien vaihtumiselle selitykseksi voimakkaasti muuttuvat markkinatilanteet. Jos Google-hakujen voidaan olettaa indikoivan hakuaktiivisuutta Wheatonin informaationhakumallin mielessä, saamme merkeille kolmannen selityksen.

Wheatonin mallin perusteella ei ole syytä olettaa ex ante, että hakuaktiivisuuden lisääntyminen näkyisi välttämättä asuntojen hintojen kasvuna tulevaisuudessa, vaikka vain kysyntäpuoli olisi aktiivinen hakuprosessissa. Intuition saamiseksi siitä, miksi hinta saattaa laskea lisääntyneen hakuaktiviisuuden seurauksena, tarkastelemme yhtälöitä 2.15 ja 2.16. Yhtälön 2.15 mukaan kahden asunnon kotitalouden

tilan arvoon vaikuttaa, millä todennäköisyydellä kotitalous saa myytyä epäsopivan asuntonsa. Mitä todennäköisempää kohtaaminen on, sitä suurempi nykytilan arvo. Asunnon myyjät voivat pyytää korkean hakuaktiivisuuden ympäristössä asunnosta alempaa hintaa, sillä riski, että asunto jää myymättä, on pienempi.

Hakuaktiivisuus vaikuttaa epäsopivassa asunnossa asuvan kotitalouden tilan arvoon kahta kautta yhtälössä 2.16. Ensimmäinen vaikutus tulee haun kustannusten kautta. Mitä enemmän kotitalous käyttää resurssejaan hakuun, sitä pienempi nykytilan arvo on. Toinen vaikutuskanava on sopivan asunnon löytämisen todennäköisyyden kasvu. Vaikutuksista ensimmäinen luo laskupainetta hinnoille, kun taas toinen pyrkii nostamaan hintoja. Kuva 2.2 havainnollistaa simulaation avulla hakuaktiivisuuden vaikutusta hintoihin. Simulaatiossa käytettyjen funktiomuotojen tapauksessa matalalla hakuaktiivisuuden tasolla hinnat nousevat hakuaktiivisuuden noustessa, kun taas korkeammilla tasoilla hinnat laskevat.



Kuva 2.2: Hakuaktiivisuuden simuloitu vaikutus hintoihin. Käytetyt parametrit ovat $UM = 10000$, $US = 5000$, $r = 0,015$, $\beta = 0,15$, $H = 2800000$ ja $V = 252000$. Lukujen pohjana on Vainio, Belloni ja Jaakkonen (2012). Hakuaktiivisuus E saa Google-indeksin hengessä arvoja väliltä $0 - 100$ ja funktiomuodoksi m:lle olen valinnut $m(E) = \frac{\sqrt{E}}{10}$, joka täyttää määritelmän (2.1) kriteerit. Haun kustannus on $c(E) = (\frac{E}{10})^2$, joka täyttää määritelmän 2.13 kriteerit. Google-indeksin ominaisuuksista kerron tarkemmin luvussa 4.

Luku 3

Internethaut ennustajina

Ensimmäisiä internethakudataa hyödyntäviä tieteellisiä artikkeleita on tiettävästi Ettredge, Gerdes ja Karuga (2005). Historiallisen datan rajoitetun saatavuuden vuoksi kyseisessä tutkimuksessa ei vielä käytetä aikasarjamalleja eikä luoda varsinaisia ennusteita. Tutkimuksessa kuitenkin havaitaan työttömyyteen liittyvien internethakujen olevan yhteydessä työttömyysasteeseen Yhdysvalloissa.

Ensimmäisiä Google-hakudataa hyödyntäviä ennustemalleja on Ginsbergin et al (2008) influenssaepidemiaennustemalli. Google-dataa on käytetty myös elokuvien lipputulojen, videopelimyyntien ja musiikin listasijoitusten (Goel et al 2010), osakkeiden hintojen (esim. Vlastakis & Markellos 2012), Yhdysvaltain yksityisen kulutuksen kehittymisen (Kholodilin et al. 2010) ja työttömyysasteen (esim. Zimmerman & Askitas 2009; Tuhkuri 2014) ennustamiseen. Kenties tunnetuimpana esimerkkinä Choi ja Varian (2009; 2012) esittelevät Google Trends -datan hyödyntämistä ”nowcastingiin”, eli nykyisyyden ennustamiseen. He esittelevät esimerkinomaisesti Google-hakuindeksin soveltamista autojen ja autotarvikkeiden, työttömyyskorvaushakemusten, matkustuskohteiden ja kuluttajaluottamuksen nykyisyyden ennustamiseen. Google-hakuindeksin käyttö malleissa parantaa ennusteita absoluuttisella keskivirheellä mitattuna 5-20 % kausittaisiin AR-malleihin verrattuna. Vaikka Choi ja Varian (2009) eivät väitä Google-hakudatan auttavan tulevaisuuden ennustamisessa, he nostavat esille mahdollisuuden, että pitkäkestoista suunnittelua ja harkintaa vaativien ostosten osalta myös tulevaisuuden tarkempi ennustaminen on mahdollista hakuaineistojen avulla.

Google-hakudataa kiinteistömarkkinoiden ennustamiseen käyttävät esimerkiksi liiketilakiinteistöjen tapauksessa Dietzel et al (2014) ja asuntomarkkinoilla esimerkiksi Das et al (2015), Kulkarni et al (2009), McLaren ja Shanbhogue (2011) sekä Brynjolfsson ja Wu (2009 ja 2013). Perehdyn seuraavassa hieman tarkemmin Brynjolfssonin ja Wun (2009 ja 2013), Kulkarnin et al (2009) sekä Mclarenin ja Shanbhoguen (2011) tuloksiin.

Wu ja Brynjolfsson (2014) ennustavat Yhdysvaltojen asuntomarkkinoiden kehitystä useassa eri osavaltiossa. Heidän mallinsa rakentuu kausitasoitetun autoregressiomallin päälle. Selittävinä muuttujina mallissa on edellisen periodin asuntojen myyntimäärä, edellisen periodin asuntojen hintaindeksi, nykyisen ja edellisen periodin Google-hakuindeksi, osavaltiokohtainen dummy-muuttuja, ja kausimuuttuja. Nykyisen periodin lisäksi tutkijat käyttävät myös mallia ennustamaan periodin $t+1$ asuntojen myyntiä. Kirjoittajat rakentavat vastaavat mallit myös hintaindeksille. Hakuindekseinä Brynjolfsson ja Wu käyttävät Googlen valmiita kiinteistöteemaisia hakuja sisältäviä indeksejä.

Wu ja Brynjolfsson (2014) havaitsevat, että myyntimäärien osalta Google-indeksin sisältävä ennustemalli parantaa absoluuttisella keskivirheellä (MAE) mitattuna nykyhetken ennustetarkkuutta 2,3 % vertailumalliin verrattuna. Vertailumalli on muilta osin sama kuin Google-ennustemalli, mutta se ei sisällä hakuindeksimuuttujaa. Tulevan periodin myyntimäärissä Google-mallin ennuste puolestaan on MAE:llä mitattuna 7,1 % tarkempi.

Kulkarni et al (2009) selvittävät Google-hakujen ja asuntojen hintojen välistä Granger-kausaalisuutta kahdellekymmenellä suurella metropolialueella Yhdysvalloissa. He tutkivat Granger-kausaalisuuden avulla, auttavatko Google-haut ennustamaan asuntojen hintakehitystä, sekä toisinpäin, auttaako hintakehityksen aikasarja ennustamaan Google-hakuintensiteetin kehittymistä. Heidän keskeisenä tuloksenaan on, että Google-haut auttavat ennustamaan hintojen kehittymistä, mutta kääntäen vastaavaa yhteyttä ei löydy.

McLaren ja Shanbhogue (2011) soveltavat Google-dataa asuntomarkkinoiden ennustamiseen Iso-Britanniassa. He havaitsevat käyttämissään malleissa Google-hakuindeksin kertoimen poikkeavan tilastollisesti merkitsevästi nollasta. Hakuindeksin sisällyttäminen parantaa myös malleja Akaiken informaatiokriteerillä mitattuna.

Tarvonen (2016) puolestaan havaitsee Granger-kausaalisuuden Suomen asuntojen hinnoista Google-hakuihin, mutta ei päinvastoin. Estimoitu parametri on positiivinen, eli korkeammat hinnat indikoivat suurempaa hakuaktiivisuutta seuraavalla periodilla. Tämä tulos poikkeaa Kulkarnin et al. (2009) havainnosta Yhdysvaltain metropolialueiden asuntomarkkinoilla.

Google-indeksiin liittyy kuitenkin ongelmia. Brynjolfsson et al. (2015) pyrkivät ohjaamaan huomiota siihen, miten käytettävät hakutermit valitaan useissa Google-hakuindeksiä hyödyntävissä tutkimuksissa. Vaihtoehtoisia menetelmiä ovat esitelleet esimerkiksi Scott ja Varian 2013 sekä Tierney ja Pan 2012). McLaren ja Shanbhogue (2011) nostavat esiin myös hakukäyttäytymiseen liittyviä ongelmia: osa ihmisistä voi hakea tietoa samasta asiasta hyvinkin erilaisilla hakusanoilla. Toisaalta hyvin erilaista asiaa etsivä voi käyttää samaa hakusanaa kuin tutkija on ajatellut.

Hakuaktiivisuus saattaa myös olla seurausta muuttuneesta markkinatilanteesta.

Aiemmassa informaationhakua tutkivassa kirjallisuudessa on havaittu, että esimerkiksi polttoaineen hintojen muutokset voivat näkyä lisääntyneenä tiedonhakuaktiivisuutena. Hypoteesina on, että kohonnut hinta saa kuluttajat etsimään edullisinta vaihtoehtoa. (Esim. Byrne et al 2014; Byrne ja Roos 2015.)

Olson et al. (2013) arvioivat kriittisesti Google Flu Trends -mallia (GFT), joka pyrkii ennustamaan influenssaepidemioiden ajankohtaa, sijaintia ja vakavuutta. He huomaavat, että GFT epäonnistuu usein epidemioiden ennustamisessa. Malli esimerkiksi yliarvioi vuosien 2012–2013 influenssaepidemian vakavuuden ja toisaalta ei onnistu lainkaan ennustamaan vuoden 2009 A/H1N1-epidemiaa. Mahdollisia syitä epäonnistumiselle voisivat heidän mukaansa olla esimerkiksi internethakukäyttäytymisen muutokset tai epidemian maantieteellinen tai ikärakenteellinen kohdentuminen.

Luku 4

Data

Kuvailen tässä luvussa käyttämäni datan. Keskityn etenkin Google-indeksiin, jonka muodostamisen haasteet tuodaan esille useassa tutkimuksessa (esim. Tierney ja Pan 2012; Smith 2016).

Google-indeksi

Google on tarjonnut vapaata pääsyä Google Trends -hakuindeksidataan vuodesta 2008 alkaen, mutta hakuhistoria ulottuu vuoden 2004 alkuun saakka. Hakuindeksi muodostuu Googlessa kuukausittain tehdyistä hauista siten, että käyttäjän antamalla hakutermillä tehdyt haut suhteutetaan muihin hakuihin, joita kyseisen kuukauden aikana tehtiin. Tuhkuria (2016) mukaileva formaali muotoilu Google-hakuindeksin arvolle ajanhetkellä t on

$$GI(K_t) = \frac{\frac{K_t}{G_t}}{max(\frac{K}{G})} \times 100, \tag{4.1}$$

jossa $GI(K_t)$ on Google-indeksin arvo ajanjaksolla t, K_t on kiinnostustermeillä tehtyjen hakujen absoluuttinen määrä jaksolla t, G_t on kaikkien tehtyjen hakujen määrä jaksolla t ja $max(\frac{K}{G})$ on kiinnostustermin suurin hakuosuus koko tarkasteltavan historian aikana. Jos hakujen määrä kuitenkin on liian pieni, hakuindeksi saa arvon 0. Google pyrkii näin suojelemaan käyttäjien yksityisyyttä.

Google Trends tarjoaa myös mahdollisuuden aluerajaukseen: esimerkiksi Yhdysvalloissa osavaltio- ja jopa metropolialuetason rajaaminen on mahdollista, kun taas Suomessa rajaus ulottuu maakuntatasolle asti. Aluerajaus suhteuttaa hakumäärät vain kyseisellä alueella tehtyihin hakuihin.

Määritelmästä 4.1 voi nähdä, että Google-indeksin historialliset arvot eivät välttämättä kestä muuttumattomana yli ajan. Hakuintensiteetin uusi maksimiarvo muuttaa koko olemassaolevan historian arvoja, sillä maksimiarvo normalisoidaan aina arvoon 100. Indeksin arvot perustuvat myös satunnaisotantaan kaikista hauista, mikä

voi myös aiheuttaa heiluntaa eri ajanhetkinä noudetuissa Google-indeksin arvoissa.

Ensimmäinen vaihe Google-indeksin aikasarjaa luodessa on valita sopiva hakutermi tai sopivat hakutermit, joiden hakuaktiivisuutta halutaan seurata. Useissa Google-datan ennustekykyä arvioivissa tutkimuksissa sopivien hakutermien valinta nojaa intuitioon (Brynjolfsson, Geva & Reichman 2015). Esimerkiksi McLaren ja Shanbhogue (2011) hyödyntävät työttömyyttä ennustaessaan vain yhtä hakutermiä. Useat tutkimukset, kuten Choi ja Varian (2012) sekä Wu ja Brynjolfsson (2014), käyttävät myös Googlen valmiiksi tarjoamia kategorioita. Näissä kategorioissa on Googlen valmiiksi luokittelemia hakutermejä, jotka liittyvät tiettyyn aihealueeseen.

Käytin Koopin ja Onoranten (2013) sekä Smithin (2016) hyödyntämää lähestymistapaa sopivien hakutermien valintaan. Valintaprosessi oli kuusivaiheinen. Aloitin valitsemalla omaan intuitiooni nojaten yhden termin, jota uskon ihmisten käyttävän asuntoja etsiessään. Toisessa vaiheessa hyödynsin Google Trends -palvelun listaamia syötettyyn hakutermiin liittyviä suosituimpia hakuja. Suosituimmat haut kertovat IP-osoitteen perusteella, mitä hakusanoja hakijat ovat käyttäneet alkuperäisen hakusanan kanssa samassa yhteydessä. Listan avulla on mahdollista muodostaa mielikuvaa siitä, millaisilla termeillä ihmiset hakevat kiinnostuksen kohteena olevia asioita. Toisaalta liittyvät haut auttavat havaitsemaan, onko alkuperäinen hakutermi välttämättä sopiva mielenkiinnon kohteena olevan asian tutkimiseksi. Mikäli liittyvät haut poikkeavat suuresti mielenkiinnon kohteena olevasta asiasta, on mahdollista, että alkuperäinen termi ei välttämättä kuvaa kiinnostuksen kohteena olevaa ilmiötä.

Kolmannessa vaiheessa poistin oman intuitioni perusteella Suosituimmat haut -listasta hakutermit, jotka eivät liity asuntojen ostamiseen.[1] Toistin näin saamilleni uusille termeille vaiheet kaksi ja kolme. Viidennessä vaiheessa järjestin saadut hakutermit järjestykseen odotettavissa olevan hakumäärän perusteella. Käytin järjestämisessä hyväkseni Gooogle Adwords -palvelua, joka antaa arvioita hakutermien odotettavissa olevista hakumääristä seuraavan kuukauden aikana. Lopullisiksi termeiksi valitsin yhdeksän suosituinta hakutermiä. Käytin indeksin juuriterminä sanaa *"asunnot"*. Prosessin tuloksena saadut termit sekä niillä tehtyjen hakujen määrä ovat taulukossa 4.1. Hakumäärissä huomionarvoista on, että listan kolmella suosituimmalla termillä tehdään kullakin 1 000 000 hakua kuukaudessa, mikä on enemmän hakuja kuin lopulla kuudella termillä yhteensä.

Kaupunkikohtaiset indeksit ovat muotoa *"asunnot **kaupunki**"*, jossa kaupunkisanan kohdalle tulee kunkin tarkasteltavan kaupungin nimi. Suurimmalla osalla kaupungeista liittyvät haut eivät sisältäneet laajemman indeksin muodostamiseen oleel-

[1]Perusteina poistamiselle olivat esimerkiksi suoraan vuokra-asuntoihin liittyvät termit, kuten *"vuokra asunnot"* , liian yleiset tai aiheeseen liittymättömät termit, kuten *"iltalehti.fi"* tai yhdistelmätermit, jotka sisältävät samoja termejä kuin jo listallani olevat termit.

Hakutermi	Arvioitu hakujen määrä
asunnot	1 000 000
asuntoja	1 000 000
etuovi	1 000 000
oikotie	673 000
jokakoti	110 000
skv	74 000
etuovi.com	60 500
kiinteistömaailma	40 500
habita	27 100

Taulukko 4.1: Yleiseen hakuindeksiin käytetyt hakutermit ja niiden arvioidut hakumäärät Google Adwords-palvelun mukaan, joulukuu 2016.

lisia hakutermejä, joten päätin olla hyödyntämättä yllä mainittua prosessia kaupunkikohtaisten indeksien rakentamisessa. Google Adwords -palvelun perusteella muotoa *"asunnot **kaupunki**"* olevilla hauilla on odotettavissa noin 3000 - 33 000 kuukausittaista hakua. Hakumäärät ovat huomattavasti pienempiä kuin koko maan kattavilla hauilla.

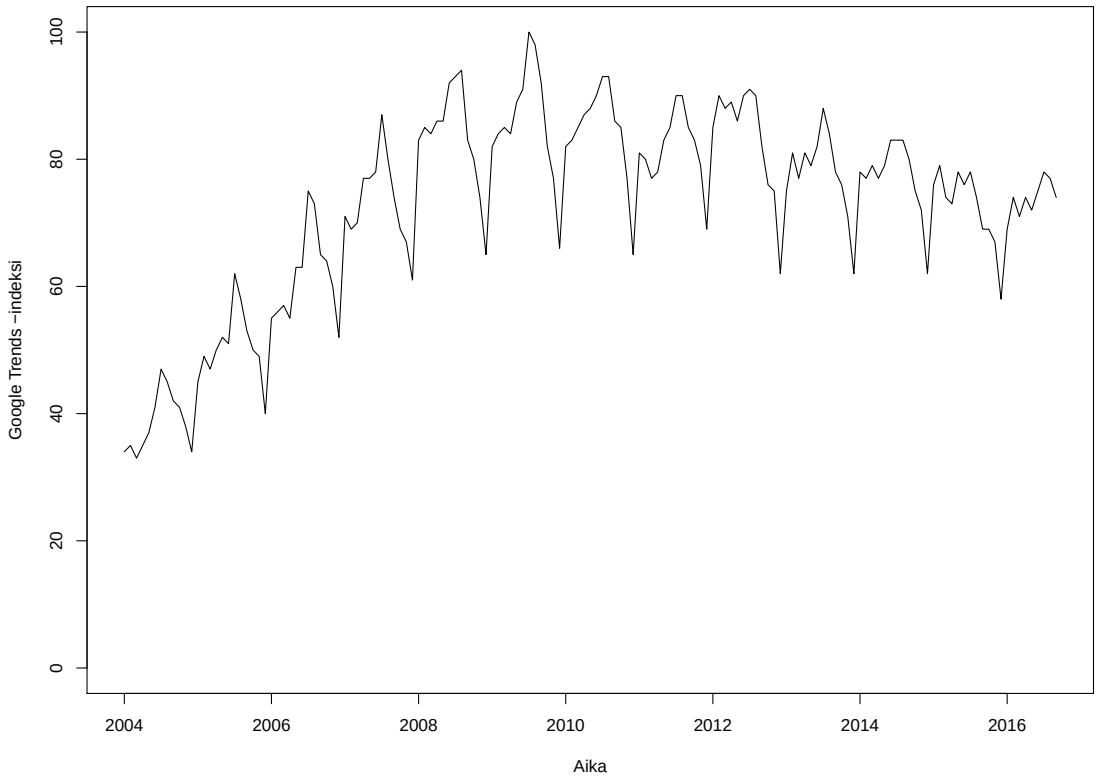
Kuvassa 4.1 on esitetty ensimmäisen indeksin kehitys koko tarkasteltavalta ajanjaksolta. Visuaalisella tarkastelulla hakuaktiivisuudessa on havaittavissa kausittaista vaihtelua siten, että hakuaktiivisuus on vuoden aikana alimmillaan aina joulukuussa ja suurimmillaan loppukesästä ja alkusyksystä. Autokorrelaatioiden tarkastelu (liite A.1) vahvistaa kausittaisuuden, minkä lisäksi aikasarjassa on havaittavissa persistenssiä.

Asuntomarkkinadata

Asuntojen hinnat ja kauppamäärät

Tilastokeskus julkaisee neljännesvuosittaista dataa asuntojen hinnoista. Tiedot ovat saatavilla sekä keskiarvoina velattomista neliöhinnoista että indeksilukuna. Käytän tutkielmassani reaalihintaindeksiä, joka muodostuu hedonista menetelmää käyttäen, eli se sisältää laatukorjauksia esimerkiksi asuntojen huoneluvun ja sijainnin mukaan. Hintatiedot on mahdollista saada kaupunkikohtaisesti. Kuva A.2 esittää hintojen kehityksen yhdeksässä suomalaisessa kaupungissa vuoden 2004 ensimmäisestä vuosineljänneksestä alkaen vuoden 2016 kolmanteen vuosineljännekseen saakka. Asuntojen mahdollisimman hyvän vertailtavuuden varmistamiseksi käytän vanhojen kerrostalo-osakeasuntojen tietoja.

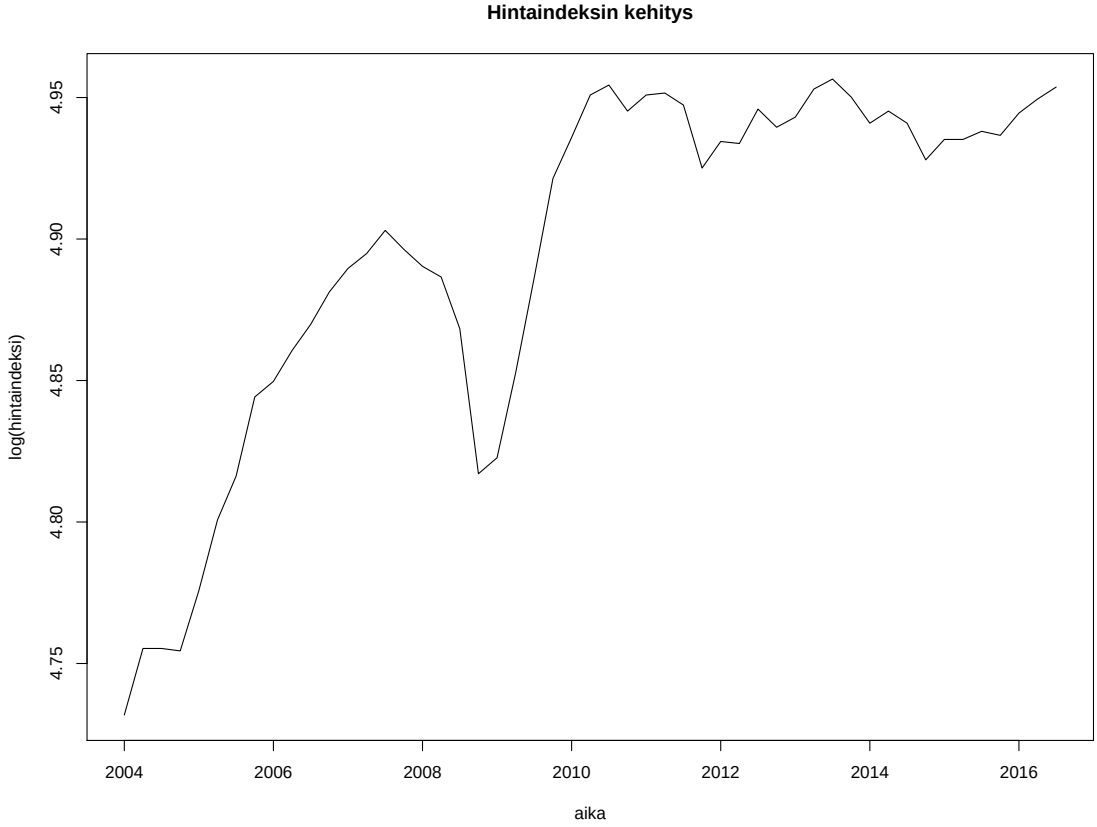
Tilastokeskus julkaisee asuntojen kauppamääristä kuukausittaista dataa. Kysei-



Kuva 4.1: Google Trends -indeksin kehitys vuoden 2004 alusta vuoden 2016 syyskuuhun. Indeksiluvut muodostettu taulukon 4.1 mukaisista hakutermeistä. Lähde: Google Trends

set tiedot ovat saatavilla vuoden 2010 tammikuusta alkaen. Aineisto on hintoja hieman suppeampi, ja se on saatavilla kaupunkikohtaisena vain Helsingistä, Espoosta, Vantaalta, Tampereelta, Turusta ja Oulusta. Hinnoista ja markkinointiajoista poiketen käytän kauppojen lukumäärissä sekä kerrostalo- että rivitaloasuntojen tietoja, sillä kauppojen lukumäärät eivät ole saatavilla vain kerrostaloihin rajattuina. Kuvassa 4.3 on asuntojen kauppamäärien logaritmi koko maan tasolla vuoden 2010 tammikuusta lähtien. Visuaalisen tarkastelun perusteella kauppamäärien tasossa vaikuttaisi tapahtuvan rakenteellinen muutos vuoden 2013 alussa, ensin poikkeavan suurena kauppamääränä, minkä jälkeen kauppamäärät putoavat jyrkästi ja jäävät vaihtelemaan alemmalle tasolle. Rakenteellinen muutos saattaa olla mahdollinen, sillä varainhoitovero nousi vuoden 2013 maaliskuun alusta lukien 1,6 prosentista 2 prosenttiin (Verohallinto, 2013). Testaan visuaalisen tarkastelun lisäksi rakennemuutoksen olemassaoloa myös regressiolla

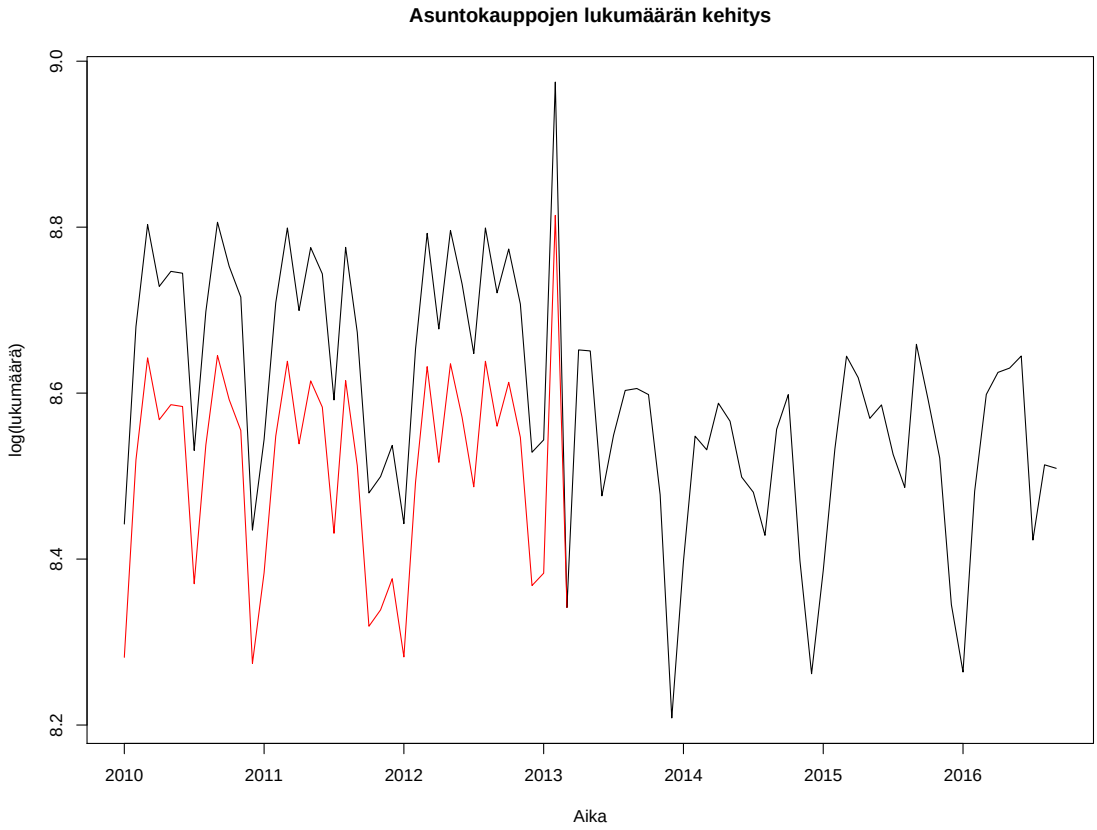
$$log(y_t) = \alpha + Vero_t, \tag{4.2}$$



Kuva 4.2: Asuntohintaindeksin logaritmi. Vuoden 2000 ensimmäinen vuosineljännes = 100. Lähde: Tilastokeskus

jossa y_t on asuntokauppojen lukumäärä kuukautena t. $Vero_t$-dummy saa arvon 1 aikasarjan alusta helmikuuhun 2013 saakka ja muulloin arvon 0. OLS-menetelmällä estimoitu Vero-dummyn kerroin on 0,1606, ja se on heteroskedastisuuden ja autokorrelaation sallivien Newey-West-keskivirheiden perusteella tilastollisesti merkitsevä 0,1 prosentin merkitsevyystasolla. Korjaan koko saatavilla olevaa historiaa koskevassa analyysissä muutoksen lisäämällä regressiomalleihin $Vero$-dummyn. Lyhemmillä ajanjaksoilla jätän muutoksen huomiotta.

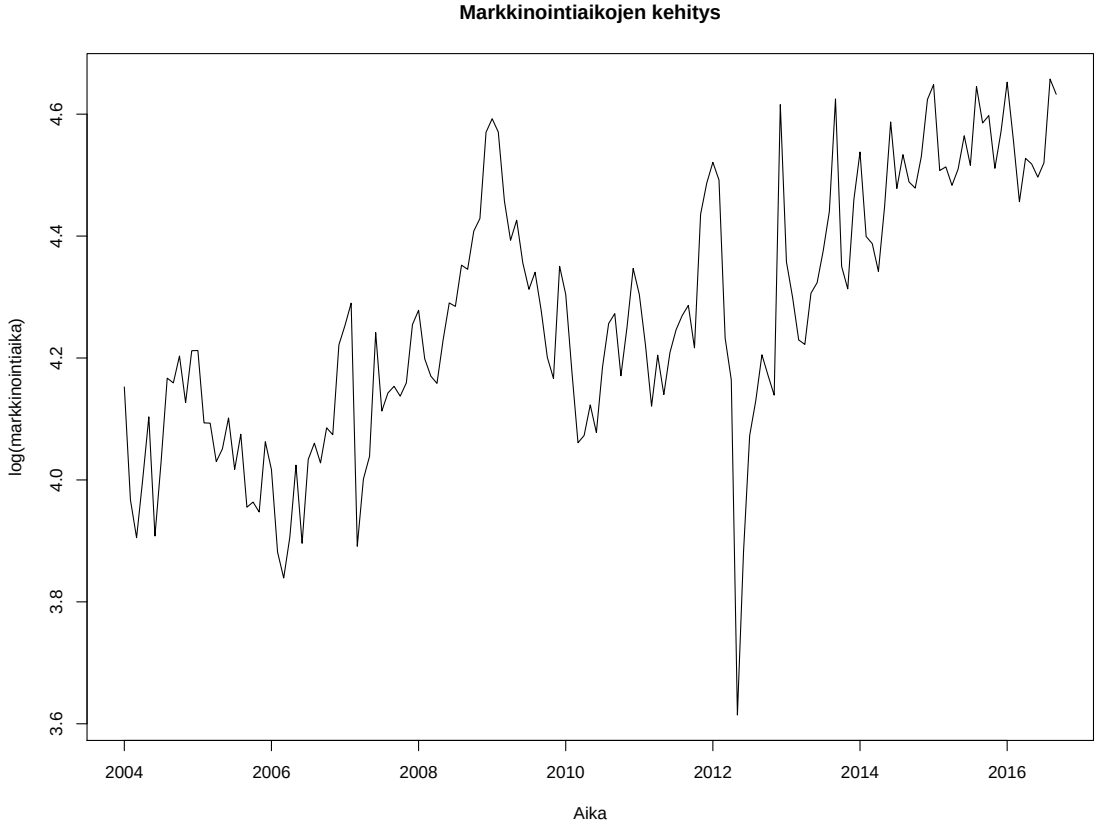
Tilastokeskus julkaisee tilastot kuukauden viiveellä (Tilastokeskus 2016), kun taas Google-indeksi ja markkinointiaikatiedot ovat saatavilla välittömästi kuukauden loputtua. Choin ja Varianin (2012) argumentoinnin mukaan Google-datalla on sen reaaliaikaisuudesta johtuen etua muuttujien nykytilan ennustamiseen, sillä tiedot ovat virallisia tilastoja nopeammin saatavilla.



Kuva 4.3: Asuntojen kauppamäärien logaritmi. Rakenteellisesta muutoksesta korjattu aikasarja punaisella, alkuperäinen aikasarja mustalla. Lähde: Tilastokeskus

Markkinointiajat

Asuntojen markkinointiajoissa käytän etuovi.com-sivuston Markkinapuntari-palvelusta saatavilla olevaa kuukausittaista dataa asuntojen markkinointiajoista. Data on saatavilla käytännössä lähes reaaliaikaisesti, sillä kunkin kuukauden tieto päivittyy välittömästi kyseisen kuukauden loputtua. Markkinointiaika ilmaisee, kuinka monta päivää asunnonmyynti-ilmoitus on sivustolla. On syytä huomauttaa, että markkinointiaika ei siis ole välttämättä sama kuin asunnon toteutunut myyntiaika. Asunnon myyjä on voinut yrittää myydä asuntoaan esimerkiksi välittäjän avulla ennen ilmoituksen laittamista etuovi.com-sivustolle. Toisaalta myyjä voi poistaa ilmoituksen sivustolta, vaikka ei olisi saanut asuntoaan myytyä. Poistaminen voi johtua siis oikeasta toteutuneesta asuntokaupasta, mutta myös esimerkiksi peruuntuneesta myyntipäätöksestä, palveluntarjoajan vaihtamisesta tai asuntoilmoituksen "tuoreuden" päivittämisestä, mikäli myyjä pelkää vanhan ilmoituksen karkottavan potentiaalisia ostajia. Markkinointiajat voivat myös olla alttiita muutoksille sivuston suosiossa, ja eri sivustojen suosiossa voi olla myös alueellisia eroja.



Kuva 4.4: Asuntojen markkinointiajat, päivää. Etuovi.com

Oletan kuitenkin tutkielmassani, että markkinointiajan muutokset voivat heijastella myös todellisia myyntiaikoja. Koska tarkoitukseni ei kuitenkaan ole tehdä kattavaa mallia asuntomarkkinoiden dynamiikasta, vaan tutkia Google-hakujen kykyä parantaa asuntomarkkinoihin liittyviä ennusteita, markkinointiaikadata on tarkoitukseeni riittävän tarkkaa.

Data markkinointiajoista on saatavilla kaupunkitasolla suurimpien kaupunkien osalta, ja olen tehnyt rajauksen käytettyihin kerrostaloasuntoihin. Luvut ovat keskiarvoja asunnon tyypin ja alueen myynti-ilmoitusten markkinointiajoista. Google-indeksin tavoin aikasarja on saatavilla vuoden 2004 tammikuusta lähtien. Kuva 4.4 havainnollistaa Suomen keskiarvoisten markkinointiaikojen kehittymistä.

Datan muokkaus analyysiä varten

Käytän analyysissäni logaritmista muunnosta selitettävistä muuttujista, mutta Google-indeksi on malleissa tasoina. Logaritmin ottaminen selitettävistä muuttujista tekee mallin tulkinnasta intuitiivisempaa, sillä selittävänä muuttujana olevan Google-indeksin kertoimen arvo on helppoa tulkita prosentuaalisina muutoksina se-

litettävissä muuttujissa. Lütkepohl ja Xu (2012) kuitenkin huomauttavat, että logaritmimuunnos auttaa vain, mikäli se pienentää selitettävän muuttujan aikasarjan varianssia.

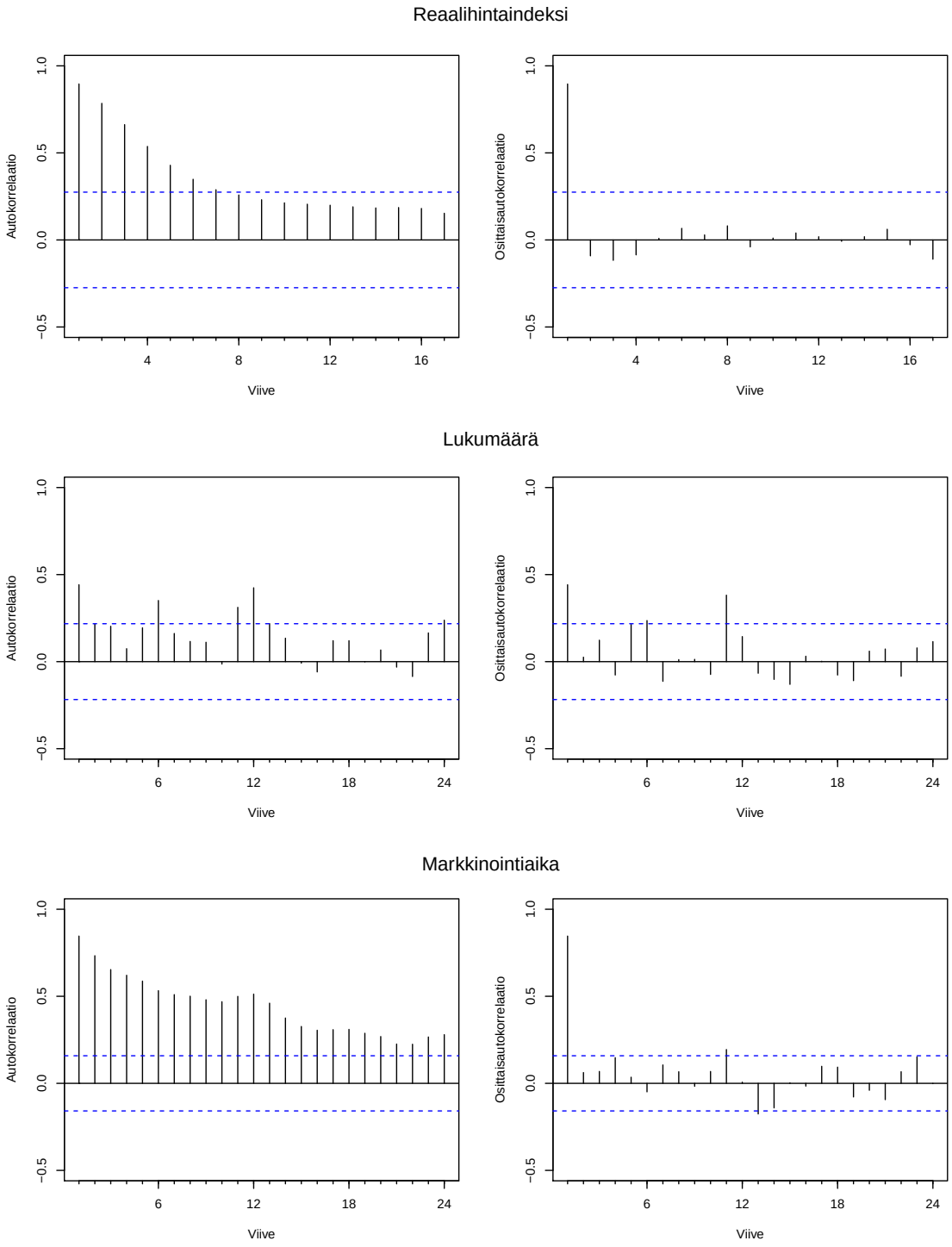
Seuraavaksi tutkin muuttujien stationaarisuutta. Kuvassa 4.5 on yhteenveto muuttujien autokorrelaatioista ja osittaisautokorrelaatioista. Autokorrelaatioiden perusteella sekä asuntojen hinnoissa että markkinointiajoissa vaikuttaa olevan persistenssiä. Markkinointiaikojen autokorrelaatiot viittaavat myös mahdolliseen kausittaiseen vaihteluun, mutta hintojen osalta vastaavaa vaikutusta ei näyttäisi olevan. Kauppojen lukumäärissä puolestaan vaikuttaa olevan selvää kausittaisuutta, mutta ei kuitenkaan persistenssiä.

Kwiatkoskin et al. (1991) kehittämä KPSS-stationaarisuustesti hylkää nollahypoteesin markkinointiaikojen taso- ja trendistationaarisuudesta, ensimmäisen viiden ja jälkimmäisen yhden prosentin merkitsevyystasoilla. Hinnoissa molemmat stationaarisuushypoteesit hylätään yhden prosentin merkitsevyystasolla. Lukumäärissä KPSS-testi ei hylkää nollahypoteesia trendistationaarisuudesta kymmenen prosentin merkitsevyystasolla. Kuvan 4.3 visuaalisen tarkastelun perusteella lukumäärillä ei vaikuta olevan ilmeistä trendiä, joten teen testin myös veromuutoksesta korjatulla aikasarjalla. KPSS-testi ei hylkää nollahypoteesia lukumäärien tasostationaarisuudesta.

Dickeyn ja Fullerin (1979) kehittämä laajennettu Dickey-Fuller-yksikköjuuritesti[2] ei hylkää nollahypoteesia yksikköjuuresta Google-muuttujassa. Myöskään minkään kiinnostusmuuttujan osalta nollahypoteesia ei voida hylätä kymmenen prosentin merkitsevyystasolla. Ensimmäisen differenssin jälkeen nollahypoteesi yksikköjuuresta hylätään viiden prosentin merkitsevyystasolla sekä markkinointiajoissa että hintaindekissä. Lukumäärissä nollahypoteesi hylätään kymmenen prosentin merkitsevyystasolla. Aiemmin myös esimerkiksi Oikarinen ja Engblom (2016) ovat havainneet, että Suomen asuntojen hinnoissa esiintyy mahdollisesti yksikköjuurta.

Vaikka selitettävissä muuttujissa vaikuttaa ADF-testien perusteella esiintyvän yksikköjuuria, käytän analyysissäni kaikista niistä tasoja differenssien sijaan. Swanson ja White (1997) huomauttavat, että etenkin ennustamiskykyä mitattaessa ja ennustemalleja tehdessä muuttujien stationaarisuus ei välttämättä edes ole suuri huolenaihe. Käytän myös Google-muuttujaa tasoina, vaikka nollahypoteesia yksikköjuuren olemassaolosta ei voida hylätä. Syitä tähän on kaksi. Ensimmäinen liittyy jo mainittuun mallin tulkintaan. Google-indeksin dataa generoivan prosessin luonne on epävarma, ja indeksiin on tehty jo normalisoinnin ja skaalauksen kaltaisia muutoksia ennen kuin se on yleisölle saatavissa. Lisätransformaatiot entisestään hankaloittaisivat Google-muuttujan tulkitsemista. Toisaalta aiemmin mainittu käyttäjien

[2]jatkossa ADF-testi



Kuva 4.5: Selitettävien muuttujien logaritmien autokorrelaatiot ja osittaisautokorrelaatiot.

yksityisyyden suojaamiseksi asetettu hakumäärien raja voi aiheuttaa ongelmia ja vaikeuttaa differenssin tulkintaa, mikäli Google-muuttuja saa arvon nolla joissakin tarkasteluajanjakson pisteissä. Toiseksi, Tuhkuri (2015) huomauttaa, että Google-

indeksi on määritelmällisesti rajattu välille 0 - 100, joten globaalin yksikköjuuri ei siten periaatteessa ole mahdollinen. Aiemmassa Google-ennustekirjallisuudessa asuntomarkkinamuuttujista ja Google-indeksistä on käytetty sekä ensimmäisiä differenssejä (mm. McLaren & Shanboghue 2011) että tasoja (mm. Wu & Brynjolfsson 2014).

Yksikköjuuren lisäksi myös kausivaihtelu voi tehdä muuttujista epästationaarisia. Kauppojen lukumäärissä, markkinointiajoissa ja Google-indeksissä vaikuttaa olevan kausittaista vaihtelua. Lähestymistavat selitettävien muuttujien kausitasoittamiseen vaihtelevat aiemmassa kirjallisuudessa. Wu ja Brynjolfsson (2014) sekä Choi ja Varian (2009; 2012) eivät itse kausitasoita muuttujia. Sen sijaan mahdollinen kausittaisuus huomioidaan lisäämällä regressiomalleihin kiinnostusmuuttujan vuoden takainen arvo. Toisaalta esimerkiksi Tarvonen (2016) kausitasoittaa muuttujat X11 Census -menetelmällä ennen analyysiä. Choi ja Varian (2009; 2012) puolestaan tasoittavat Google-muuttujasta kausivaihtelun, mikäli selitettävä muuttuja on myös kausitasoitettu. Käytin analyysissäni muuttujista sekä kausitasoittamattomia että -tasoitettuja versioita, lukuun ottamatta asuntojen hintaindeksiä, josta käytin vain kausitasoittamatonta versiota. Tasoitin muuttujat X-13-ARIMA-SEATS-menetelmällä.

Google-indeksin havaintotiheys on kuukausittaisena aikasarjana tiheämpi kuin hintaindeksin. Siksi muodostin hintojen ennustamista varten neljännesvuosittaisen Google-indeksin ottamalla kunkin vuosineljänneksen kuukausien Google-indeksin arvoista aritmeettisen keskiarvon. Havaintotiheyden vähentäminen aritmeettisella keskiarvolla saattaa kuitenkin johtaa informaation menettämiseen (Ghysels, Santa-Clara & Valkanov 2010). Andreou, Ghysels ja Kourtellos (2010) huomauttavat myös, että a priori -oletukseen yhtä suurista painoista ei kaikissa malleissa välttämättä ole perusteita. Ghysels et al. (2004) esittävät ratkaisuna niin kutsuttua MIDAS-regressiota[3], joka hyödyntää kaikkia saatavilla olevia arvoja ilman aggregointia pienemmän tiheyden muuttujaksi. Tämän tutkielman tarkoitukseen Google-muuttujan aggregointi aritmeettisella keskiarvolla on kuitenkin riittävä.

Kausitasoitetussa tapauksessa tasoitin Google-muuttujan ensin, minkä jälkeen aggregoin aikasarjan neljännesvuosittaiseksi. Taulukko 4.2 esittää yhteenvedon kunkin muuttujan jakauman keskeisimmistä tunnusluvuista.

[3]MIxed DAta Sampling

Muuttuja	N	μ	σ	min	max	vinous	huipukkuus
Reaalihintaindeksi	51	134,0	8,2	113,5	142,1	-1,0	2,8
Kauppojen lukumäärä	81	5437,5	771,8	3672	7902	0,2	3,1
Markkinointiaika	153	73,0	15,5	37,1	105,4	0,3	2,2
Google-indeksi (kk)	153	72,7	15,4	33	100	-0,9	3,0
Google-indeksi (aggregoitu)	51	72,7	15,1	34	96,7	-0,9	3,1

Taulukko 4.2: Muuttujien jakaumien tunnusluvut. Google-indeksin tunnusluvut ovat yleisestä indeksistä.

Luku 5

Metodit ja käytettävät mallit

Luvun 2 perusteella Google-haut voivat sisältää mielenkiintoista informaatiota kolmesta asuntomarkkinoiden muuttujasta: kauppojen lukumääristä, asuntojen myyntiajoista sekä hinnoista. Tässä luvussa esittelen metodit, joilla pyrin selvittämään Google-muuttujan sisältämää informaatiota sekä ennustekykyä näistä kolmesta muuttujasta.

Luku koostuu neljästä osasta. Ensimmäiseksi tarkastelen Google-muuttujan ja kiinnostusmuuttujien ristikorrelaatioita. Toisessa osassa esittelen menetelmät, joilla tutkin Google-muuttujien ja kiinnostusmuuttujien välistä Granger-kausaalisuutta. Kolmannessa kuvaan mallinvalintaprosessit koko otokset kattaville malleille ja viimeiseksi esitän menetelmät eri mallien tuottamien ennusteiden vertaamiseksi.

Notationaalisena huomiona merkitsen tutkielmassani alaindeksillä t sellaista periodia, jolloin selitettävästä muuttujasta saatavilla olevat havainnot yltävät periodiin $t-1$ saakka. Tilastokeskuksen tilastojen viiveen vuoksi Google-indeksin saatavilla olevat havainnot yltävät tällöin periodiin t. Markkinointiajoissa Google-indeksillä ei ole reaaliaikaisuuden etua, joten markkinointiaikoja koskevassa analyysissä sekä markkinointiaikojen että Google-indeksin tuorein havainto periodilla t on periodilta $t-1$.

Ristikorrelaatio

Laskin kullekin muuttujalle otosristikorrelaation luomani Google-indeksin kanssa. Otosristikorrelaatio kuvaa havaintojen x_t korrelaatioita havaintojen y_{t+h} kanssa. Ristikorrelaatioiden laskeminen auttaa havaitsemaan, mitkä Google-indeksin viiveet voivat olla merkityksellisiä asuntomarkkinoiden ennustamisen kanssa. Toisaalta ristikorrelaatiot antavat viitteitä myös siitä, mikäli kiinnostusmuuttujat edeltävät Google-hakuja eikä toisinpäin.

Granger-kausaalisuus

Tarkastelin Google-hakujen mahdollista ennustekykyä Granger-kausaalisuuden avulla. Granger-kausaalisuustestin avulla voidaan selvittää, auttaako muuttujan x_t historia ennustamaan toisen muuttujan y_t arvoja paremmin kuin vain muuttujan y_t oma historia (Granger, 1969). Aiemmin esimerkiksi Kulkarni et al. (2009) ja Dietzel, Braun ja Schäfers (2014) ovat hyödyntäneet Granger-kausaalisuustestejä selvittääkseen Google-hakujen ennustekykyä kiinteistömarkkinamuuttujiin.

Granger-kausaalisuustesti tehdään vektoriautoregressiomallin (VAR) kertoimille. Olkoon kahden muuttujan VAR(p)-malli muotoa

$$\begin{aligned} x_t &= \alpha_1 + \sum_{i=1}^{p} \beta_{1,i} x_{t-i} + \sum_{i=1}^{p} \theta_{1,i} y_{t-i} + \epsilon_{1,t} \\ y_t &= \alpha_2 + \sum_{i=1}^{p} \beta_{2,i} y_{t-i} + \sum_{i=1}^{p} \theta_{2,i} x_{t-i} + \epsilon_{2,t}, \end{aligned} \tag{5.1}$$

jossa p on mallin viivemäärä, $\epsilon_{j,t}$ on IID virhetermi, $\beta_{j,i}$ on muuttujan oman viiveen kerroin viiveellä i ja $\theta_{j,i}$ on toisen muuttujan kerroin viiveellä i. Granger-testi on F-testi nollahypoteesille $H_0 : \theta_{1,1} = \theta_{1,2} = ... = \theta_{1,p} = 0$ vaihtoehtoisen hypoteesin ollessa H_1 : ei H_0. Nollahypoteesina siis on, että y_t ei Granger-aiheuta x_t:tä. Vastaavalla tavalla nollahypoteesi voidaan asettaa niin, että x_t ei Granger-aiheuta y_t:tä. Tein analyysissäni Granger-testit pareittain Google-muuttujalle ja kullekin kiinnostusmuuttujalle. Koska Google-indeksi on saatavilla Tilastokeskuksen tilastoja nopeammin, tein asuntojen hintaindeksin ja asuntokauppojen lukumäärän tapauksessa Granger-kausaalisuustestin myös muotoa

$$\begin{aligned} x_t &= \alpha_1 + \sum_{i=1}^{p} \beta_{1,i} x_{t-i} + \sum_{i=1}^{p} \theta_{1,i} y_{t-i+1} + \epsilon_{1,t} \\ y_{t+1} &= \alpha_2 + \sum_{i=1}^{p} \beta_{2,i} y_{t-i+1} + \sum_{i=1}^{p} \theta_{2,i} x_{t-i} + \epsilon_{2,t} \end{aligned} \tag{5.2}$$

olevasta mallista, jossa x_t on kiinnostusmuuttuja ja y_t on Google-indeksi. Aiemmin Tuhkuri (2015) on käyttänyt vastaavaa spesifikaatiota, joka hyödyntää Google-indeksin ajantasaisuutta.

Granger-kausaalisuutta testattaessa tulee kiinnittää huomiota useisiin tuloksiin vaikuttaviin seikkoihin. Ensiksi, Thornton ja Batten (1985) huomauttavat, että Granger-kausaalisuustestin tulokset ovat herkkiä malliin valittujen viiveiden määrälle. He suosittelevat informaatiokriteereiden hyödyntämistä mallin sopivan rakenteen valitsemisessa. Jones (1989) havaitsee kuitenkin, että viiveiden valinta *ad hoc* suoriutuu kausaalisuuden havaitsemisessa tilastollista testaamista paremmin.

Toiseksi, F-testin asymptotiikka ei ole voimassa, jos VAR-mallin molemmat muuttujat eivät ole stationaarisia. Esimerkiksi yhteisintegroituvuus saattaa aiheuttaa näennäisen yhteyden tarkasteltavien muuttujien välille (Toda & Phillips 1993). Tämän potentiaalisen ongelman voi korjata käyttämällä Todan ja Yamamoton (1995) kehittämää mallin ylisovittamiseen perustuvaa metodia. Menetelmän avulla on mahdollista tehdä tilastollisia testejä VAR-mallin parametrien rajoitteille, vaikka mallin muuttujat ovatkin integroituneita tai yhteisintegroituneita. Metodin etuna on myös, että sitä voi käyttää, vaikka integroituneisuudesta ei ole täyttä varmuutta. Toisaalta, jos mallin muuttujat eivät ole integroituneita tai yhteisintegroituneita, Granger-testin voima kärsii metodia hyödynnettäessä. (Dolado & Lütkepohl 1996.)

Kolmanneksi, Granger (1979) huomauttaa, että muuttujien kausivaihtelu saattaa tuottaa ongelmia Granger-testin validiteettiin. Muuttujien kausittaisten komponenttien tyypistä riippuen sekä kausitasoittaminen että tasoittamatta jättäminen voivat johtaa näennäisiin yhteyksiin muuttujien välillä.

Tein Granger-kausaalisuustestit kahden muuttujan VAR-malleille, eli muodostin mallit kiinnostusmuuttujasta sekä Google-indeksistä. En ottanut muuttujista differenssejä, sillä Todan ja Yamamoton (1995) metodi edellyttää muuttujien käyttämistä tasoina. Hyödynsin kunkin mallin viivemäärän valinnassa Schwarzin informaatiokriteeriä. Schwarzin informaatiokriteeri on muotoa

$$BIC = log\hat{\sigma}^2 + \frac{p+1}{T} logT, \tag{5.3}$$

jossa $\hat{\sigma}^2$ on ϵ_t:n estimoitu varianssi, p on selittävien muuttujien määrä ja T havaintojen kokonaismäärä (Schwarz 1978). Informaatiokriteeri kuvaa valintaa mallin hyvän sovitteen ja toisaalta mallin yksinkertaisuuden välillä. Informaatiokriteerin toinen termi rankaisee malleja, joilla on enemmän selittäviä muuttujia. Informaatiokriteerin suosittelema malli on se, jolla informaatiokriteeri saa pienimmän arvonsa.

Tarkistin valitun mallin residuaalien autokorrelaatiot sekä Breusch-Godfreyn (Breusch 1978; Godfrey 1978) että Edgerton-Shukur-testillä (Edgerton & Shukur 1999). Testeistä ensimmäinen on Lagrangen kertojatesti, jonka nollahypoteesina on, että residuaaleissa ei esiinny autokorrelaatiota viiveeseen h saakka. Jälkimmäinen puolestaan sisältää pienen otoksen korjauksen. Molempien testien tarkastelemisen perusteena on, että yleisesti käytetyllä BG-testillä tyypin 1 virheen todennäköisyys on pienillä otoksilla suurempi kuin testin nimellinen merkitsevyystaso (Kiviet 1986). Mikäli toinen testeistä ei hylännyt nollahypoteesia, mutta toinen hylkäsi, tarkastelin residuaalien autokorrelaatioita myös visuaalisesti.

Autokorrelaatiotesteissä käyttämäni viiveiden määrä oli $h = max\{p, s\}$, jossa p on mallin aste ja s on kausittainen frekvenssi, eli kuukausittaisessa tapauksessa $s = 12$ ja neljännesvuosittaisessa $s = 4$. Residuaalien tarkastelu vähintään kausittaiseen frekvenssiin saakka auttaa havaitsemaan mahdollisesti jäljellä olevan kausittaisen

autokorrelaation. Mikäli Schwarzin informaatiokriteerin valitseman mallin virheissä oli vielä autokorrelaatioita, lisäsin malliin viiveitä, kunnes autokorrelaatioita ei enää ollut.

Seuraavaksi tutkin mahdollisten yksikköjuurten olemassaoloa mallin muuttujissa ja selvitin aikasarjojen integroitumisen asteen. Testaan muuttujien integroitumisen asteen vain kiinnostusmuuttujilta. Kuten edellä mainittu, Google-indeksillä ei määritelmällisesti ole mahdollista olla globaalia yksikköjuurta, sillä sen vaihteluväli on rajattu. Tästä syystä en myöskään tee yhteisintegraatiotestejä Google-muuttujan ja kiinnostusmuuttujien välillä.

Merkitsemme näin saatua kiinnostusmuuttujan integroitumisen astetta d:llä. Todan ja Yamamoton (1995) metodin perusteella estimoitava malli on nyt VAR($p+d$), josta Granger-kausaalisuustesti tehdään viiveeseen p saakka. Jos todellinen dataa generoiva prosessi on VAR(p) ja sovitettu malli on VAR($p+d$), Granger-kausaalisuustestin kaltaista Wald-testiä voidaan soveltaa mallin ensimmäisiin p kerroinmatriisiin. (Dolado ja Lütkepohl 1996.)

Mallit

Valitsin kullekin muuttujalle vertailumalliksi yksinkertaisen autoregressiivisen mallin. Aiemmassa Google-hakujen ennustekykyä arvioivassa kirjallisuudessa vertailumallina on usein kausittainen AR(1)-malli, joka sisältää ensimmäisen viiveen lisäksi vuoden takaisen viiveen. Mallia käyttävät esimerkiksi Choi ja Varian (2009; 2012), Wu ja Brynjolfsson (2014) sekä Tuhkuri (2015).

Autoregressiivisiin malleihin rajoittumiseen on useita syitä. Ensimmäinen on mallin rakenteen yksinkertaisuus sekä estimoitavien parametrien pieni määrä. Monimutkaiset mallit saattaisivat törmätä ongelmiin etenkin asuntojen hintoja mallinnettaessa, sillä havaintoja on vain 51. Toiseksi, useissa tutkimuksissa yksinkertaisten lineaaristen ennustemallien on havaittu onnistuvan hyvin ennustamisessa, toisinaan jopa hienostuneempia malleja paremmin (esim. Wu ja Brynjolfsson 2014). Myös Crawford ja Fratantoni (2003) havaitsevat, että yksinkertaiset AR- tai ARMA-mallit suoriutuvat asuntomarkkinoille tehdyissä ennusteissa varsin hyvin verrattuna monimutkaisempiin GARCH-, AR-GARCH- tai regiimimuutosmalleihin[1].

Valitsin vertailumallien sopivan viiveiden määrän Schwarzin informaatiokriteerin perusteella. Ilmoitan malleista myös Akaiken (1974) informaatiokriteerin, joka on muotoa

$$AIC = log\hat{\sigma}^2 + 2\frac{p+1}{T}. \tag{5.4}$$

Merkinnät ovat samat kuin määritelmässä 5.3. Akaiken informaatiokriteeri eroaa

[1]engl. regime switching model, käännös kirjoittajan

Schwarzin informaatiokriteeristä rankaisutermin osalta. Schwarzin informaatiokriteeri rankaisee voimakkaammin mallin monimutkaisuudesta. Mallin valinnassa käytin Schwarzin informaatiokriteeriä Akaiken informaatiokriteerin sijaan, sillä asymptoottisilta ominaisuuksiltaan ensin mainittu on yleensä parempi: Schwarzin informaatiokriteeri lähes varmasti valitsee oikean mallin, kun $T \to \infty$. (Verbeek 2004.)

Estimoin autoregressiiviset mallit OLS-, eli pienimmän neliösumman menetelmällä. Harkitsin kaikista muuttujista sekä kausittaisia autoregressiivisiä malleja että malleja ilman kausittaista termiä. Schwarzin informaatiokriteerin perusteella vertailumalleiksi valikoituivat

$$\begin{aligned} \text{Malli } 0_{HI} &: log(y_t) = \alpha + \beta_1 log(y_{t-1}) + \beta_2 log(y_{t-2}) + \epsilon_t, \\ \text{Malli } 0_{LKM} &: log(y_t) = \alpha + \beta_1 log(y_{t-1}) + \beta_{12} log(y_{t-12}) + \theta Vero_t + \epsilon_t \\ \text{Malli } 0_{MA} &: log(y_t) = \alpha + \beta_1 log(y_{t-1}) + \beta_{12} log(y_{t-12}) + \epsilon_t. \end{aligned} \tag{5.5}$$

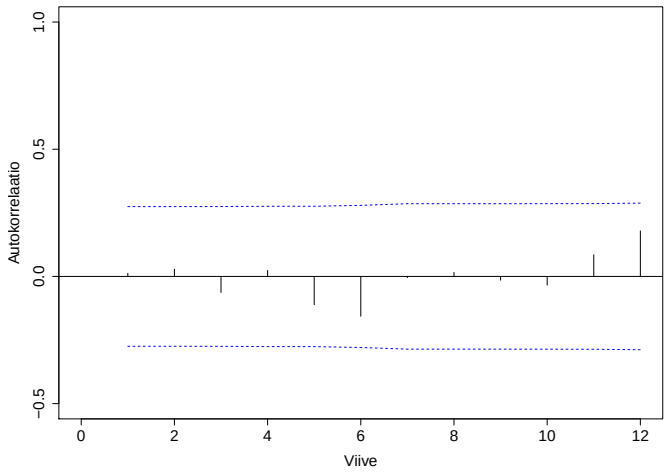
Alaindekseistä HI viittaa asuntojen reaalihintaindeksiin, LKM asuntokauppojen lukumääriin ja MA markkinointiaikoihin. Lukumäärissä huomioin dummy-muuttujalla $Vero_t$ luvussa 4 esiintuodun varallisuusveron muutoksen maaliskuun 2013 alusta lukien. Malleista 0_{LKM} ja 0_{MA} ovat kausittaisia AR(1)-malleja, eli ensimmäisen viiveen lisäksi selittävänä muuttujana on myös vuoden takainen viive. Hintaindeksimalli on puolestaan AR(2) ilman kausitermiä.

Kuvissa 5.1, 5.2 ja 5.3 esitettyjen autokorrelaatiofunktioiden perusteella vaikuttaa, että yhtälöiden 5.5 mukaisten mallien residuaaleissa ei esiinny autokorrelaatiota. Myöskään Breusch-Godfrey-testi ei hylkää nollahypoteesia autokorreloimattomuudesta minkään mallin kohdalla. Mahdollisen jäljellä olevan heteroskedastisuuden ratkaisin käyttämällä heteroskedastisuuden ja autokorrelaation sallivia Newey-West-keskivirheitä, jonka kehittivät Newey ja West (1986).

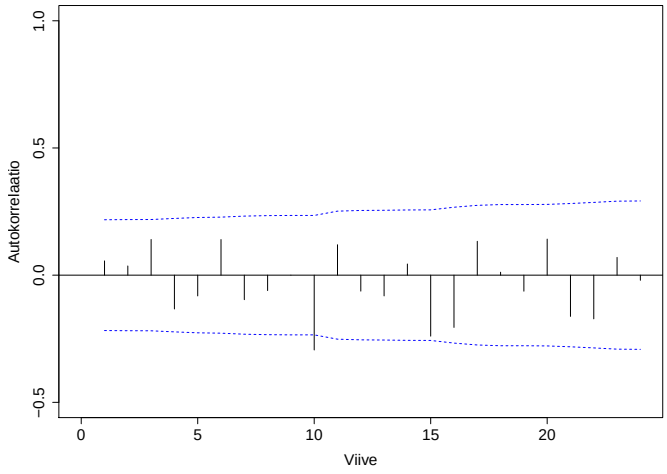
Selvittääkseni, parantavatko Google-haut mallien sovitetta, laajensin mallit Google-indeksillä ja sen viiveillä. Vertaan referenssimallien ja Google-mallien sopivuutta tutkimalla mallien korjattua selitysastetta, Google-termien merkitsevyyttä sekä Schwarzin ja Akaiken informaatiokriteereitä.

Hakumuuttujan sopivien viiveiden valinta ei ole intuitiivisesti täysin selvää. Anglin (1997) havaitsee kanadalaisen aineiston pohjalta, että asunnon etsiminen kestää ostajalla keskimäärin kymmenen viikkoa. Elder, Zumpano ja Baryla (1999) erottavat etsimisestä keston lisäksi myös toisen ulottuvuuden, intensiteetin. Google suodattaa otosta tehdessään hauista pois saman henkilön samoilla hakusanoilla lyhyen ajan sisällä tehtyjä hakuja. Toistuvat haut samalla termillä lasketaan siis indeksilukuun mukaan vain kerran. Google ei kuitenkaan täsmennä, kuinka pitkä hakujen välisen ajan täytyy olla, että haut tulkitaan erillisiksi hauiksi.[2] Samojen henkilöiden toistuvat haut kuitenkin voivat kertoa juuri kohonneesta hakuintensiteetistä ja siten

[2] `https://support.google.com/trends/answer/4355213?hl=en&ref_topic=6248052`



Kuva 5.1: Mallin 0_{HI} residuaalien autokorrelaatiot. Luottamusvälinä Bartlettin MA(q) 95%.

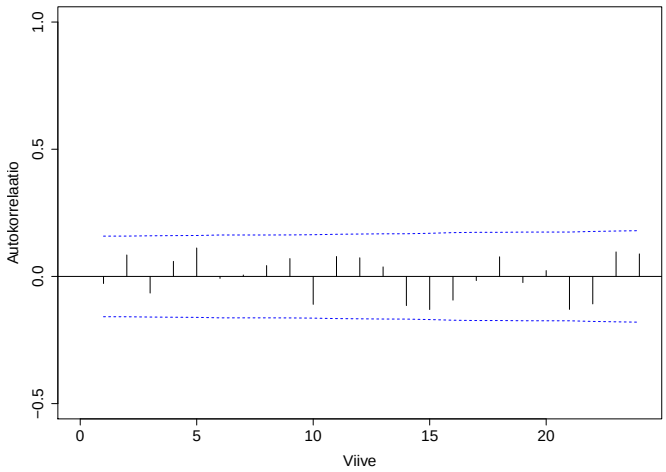


Kuva 5.2: Mallin 0_{LKM} residuaalien autokorrelaatiot. Luottamusvälinä Bartlettin MA(q) 95%.

suuremmasta hakuaktiivisuudesta eli E:stä. Tämä suodatusmekanismi voi osaltaan heikentää Google Trends -datan kykyä ennustaa asuntomarkkinoita.

Aiemmassa empiirisessä tutkimuksessa Wu ja Brynjolfsson (2014) käyttävät Google-indeksin sisältävissä malleissa enimmillään kahta Google-muuttujan neljännesvuosittaista viivettä. He perustelevat valintaa sillä oletuksella, että yhdeksän kuukauden tai vuoden takainen hakuaktiivisuus tuskin enää vaikuttaa nykyiseen asuntomarkkinoiden tilanteeseen. Vastaavin perustein harkitsin enintään kuuden kuukauden takaista Google-muuttujan viivettä. Neljännesvuosittaisessa Google-muuttujassa harkitsin enimmillään kahden vuosineljänneksen takaista viivettä.

Vastaavasti harkitsin myös malleja sen mukaan, onko niissä mukana Google-



Kuva 5.3: Mallin 0_{MA} residuaalien autokorrelaatiot. Luottamusvälinä Bartlettin MA(q) 95%.

muuttujan viiveet i:hin asti vai vain i:s viive. Etenkään kuukausittaisissa malleissa ei välttämättä ole syytä olettaa, että Google-indeksin ajantasaisella arvolla on informaatiota nykyisestä kuukaudesta. Arvioitavia malleja on siis kuukausittaisten muuttujien osalta yhteensä 20 ja hintaindeksin osalta viisi. Pitääkseni analyysin ja arvioitavat mallit yksinkertaisena jätän huomiotta muut mahdolliset permutaatiot eri viiveistä. Valitsen sopivimman Google-mallin referenssimallien tavoin Schwarzin informaatiokriteerin perusteella. Google-mallit ovat

$$\begin{aligned}
&\text{Malli } 1_{HI} : log(y_t) = \alpha + \beta_1 log(y_{t-1}) + \beta_2 log(y_{t-2}) + \gamma_t x_t + \gamma_{t-1} x_{t-1} + \epsilon_t \\
&\text{Malli } 1_{LKM} : log(y_t) = \alpha + \beta_1 log(y_{t-1}) + \beta_{12} log(y_{t-12}) + \theta Vero_t + \gamma_t x_t + \gamma_{t-5} x_{t-5} + \epsilon_t \\
&\text{Malli } 1_{MA} : log(y_t) = \alpha + \beta_1 log(y_{t-1}) + \beta_{12} log(y_{t-12}) + \gamma_{t-1} x_{t-1} + \gamma_{t-3} x_{t-3} + \epsilon_t.
\end{aligned} \tag{5.6}$$

Malli 1_{HI} sisältää Google-indeksin ajantasaisen ja yhdellä viivästetyn arvon. Mallissa 1_{LKM} puolestaan on Google-indeksin ajantasainen ja viidellä viivästetty arvo.

Ennustaminen

Simuloin todellista ennustamista sekä vertailu- että Google-indeksin sisältävillä malleilla ja vertasin saatuja ennusteita toisiinsa. Valitsin kullekin muuttujalle aikasarjojen alusta estimointiperiodin, jonka perusteella valitsin mallin. Käytin estimointijakson pituutena hintojen osalta 20 vuosineljännestä, lukumäärien osalta 36 kuukautta ja markkinointiaikojen osalta 48 kuukautta. Eripituiset estimointijaksot johtuvat eripituisista saatavilla olevista muuttujien historioista. Valitsin sovitettavan mallin Schwarzin informaatiokriteerin perusteella. Ennustin näin estimoidun mallin pe-

rusteella kiinnostusmuuttujan seuraavan arvon, eli esimerkiksi hinta-aikasarjan 21. vuosineljänneksen arvon. Seuraavaksi siirsin estimointijaksoa yhden periodin verran eteenpäin siten, että estimointijakson pituus säilyi samana. Hintaindeksin tapauksessa estimointijakso käsitti siis aikasarjan vuosineljännekset toisesta havainnosta 21. havaintoon. Toistin mallin valinnan uudelle estimointijaksolle, mikä kuvaa saatavilla olevan informaatiojoukon muuttumista yli ajan. Vastaavaa menetelmää, jossa ennustemallit voivat muuttua yli ajan, käyttävät esimerkiksi Pesaran ja Timmermann (1995). Tein uudella mallilla seuraavan ennusteen ja siirsin jälleen estimointijaksoa eteenpäin. Toistin prosessia kunkin aikasarjan loppuun saakka, jolloin testijaksojen pituuksiksi muodostuivat hintaindeksille 27, lukumäärille 33 ja markkinointiajoille 93 periodia.

Estimointijaksojen pituuden pitämistä kiinteänä kutsutaan *rolling window*-menetelmäksi, joka on tyypillinen etenkin finanssimarkkinoiden aikasarjojen ennustamisessa. Vastaavasti makromuuttujia ennustavien mallien suoriutumista arvioidaan usein pitenevän estimointijakson menetelmällä. Tällöin jokaista ennustetta varten estimoituun malliin huomioidaan koko ennustehetkeen asti saatavilla oleva historia. Kiinteänpituisen estimointijakson menetelmän etuna on, että taustalla olevan dataa generoivan prosessin muuttuessa vanhan prosessin muodostama data häviää estimoinnista. Kääntöpuolena menetelmä rajaa estimointijakson pituutta ja siten käytettävän datan määrää, mikä voi lisätä parametrien estimaattien varianssia. (Clark ja McCracken, 2009.) Koska etenkin hintojen osalta estimointijakso on erittäin lyhyt, esitän myös pitenevän estimointijakson menetelmällä tehdyt ennusteet.

Loin ennusteet sekä nykyisille että tuleville muuttujien arvoille. Esimerkiksi Choi ja Varian (2012) käyttävät Google-hakuja vain nykyisyyden ennustamiseen, mutta huomauttavat, että etenkin pidempää informaationhakuperiodia vaativat prosessit voivat mahdollistaa myös tulevaisuuden ennustamisen. Tulevaisuuden ennustamista hyödyntävätkin mm. Wu ja Brynjolfsson (2014). AR-mallien avulla tulevaisuutta voidaan ennustaa karkeasti jaotellen iteratiivisesti tai suoraan. Iteratiivisessa ennusteessa ennusteet tehdään yksi periodi kerrallaan eteenpäin. Ensimmäinen ennuste tehdään periodille t, minkä jälkeen ennustettua arvoa käytetään "datana" ajanhetken $t+1$ ennustamiseen. Prosessia toistetaan, kunnes saavutaan halutun periodin $t+h$ ennusteeseen. Suorassa ennustamisessa puolestaan y_{t+h} ennustetaan vain ajanhetkellä t saatavilla olevan datan perusteella. (Chevillon & Hendry 2005; Marcellino, Stock & Watson 2006.)

Marcellino, Stock ja Watson (2006) havaitsevat, että alle vuoden pituisella ennustehorisontilla suora ennuste tuottaa iteroituja ennusteita pienemmän keskineliövirheen, jos malli on valittu Schwarzin informaatiokriteerin perusteella. Lisäksi he suosittelevat suoraa ennustamista, jos data on epästationaarista ja otoskoko on pie-

ni. Siksi tein ennusteet hintaindeksille 0-2 periodia eteenpäin ja kuukausittaisille muuttujille 0-6 periodia eteenpäin hyödyntäen suoraa ennustamista.

Referenssiennustemallit perustuvat regressiomalliin

$$log(y_{t+h}) = \alpha + \sum_{i=1}^{p} \beta_i log(y_{t-i}) + \epsilon_{t+h} \tag{5.7}$$

jonka parametrit estimoin pienimmän neliösumman menetelmällä. Periodin $t + h$ suora ennuste on siten

$$\widehat{log(y_{t+h})} = \hat{\alpha} + \sum_{i=1}^{p} \hat{\beta}_i log(y_{t-i}). \tag{5.8}$$

Google-muuttujan sisältävät mallit ovat muotoa

$$log(y_{t+h}) = \alpha + \sum_{i=1}^{p} \beta_i log(y_{t-i}) + \rho \mathbf{x} + \epsilon_{t+h}, \tag{5.9}$$

jossa $\mathbf{x}$ on Schwarzin informaatiokriteerin perusteella valitut Google-muuttujan viiveet.

Vertaan mallien ennustekykyä keskineliövirheen neliöjuurella sekä absoluuttisella keskivirheellä mitattuna. Ilmoitan vertailu- ja Google-mallien absoluuttisen prosenttikeskivirheen prosentuaalisen erotuksen, eli

$$\Delta MAPE = \frac{MAPE_{GI} - MAPE_{REF}}{MAPE_{REF}} \times 100, \tag{5.10}$$

jossa $MAPE_{GI}$ on Google-mallin ja $MAPE_{REF}$ on referenssimallin absoluuttinen virhe. $\Delta MAPE$:n negatiiviset arvot viittaavat siis pienempään absoluuttiseen virheeseen Google-mallissa ja positiiviset referenssimallissa. Teen ennustevirheille myös Diebold-Mariano-testin, joka on tilastollinen testi ennustetarkkuuksien eroista (Diebold & Mariano 1995; 2002).

Luku 6

Tulokset

Ristikorrelaatio

Hintaindeksin ja Google-indeksin ristikorrelaatio on suurimmillaan on viiveellä nolla. Google-indeksin ensimmäinen viive on voimakkaimmin korreloitunut lukumäärien nykyisen arvon kanssa. Markkinointiajoissa suurimmat korrelaatiot ovat viiveillä 4-6. Google-indeksin viiveet ovat yleisesti ottaen hieman voimakkaammin korreloituneita sekä hintojen että kauppojen lukumääriin kuin Google-indeksin tulevat arvot. Huomionarvoista on, että Google-indeksin viiveiden ja lukumäärien välillä korrelaatiokerroin on negatiivinen viiveillä 3-6. Taulukossa 6.1 on yhteenveto muuttujien ristikorrelaatioista Google-indeksin kanssa.

h	-6	-5	-4	-3	-2	-1	**0**	1	2	3	4	5	6
HI	0,41	0,46	0,52	0,55	0,60	0,66	0,74	0,64	0,56	0,51	0,44	0,35	0,28
LKM	-0,22	-0,36	-0,23	-0,05	0,09	0,35	0,31	0,01	-0,03	-0,07	-0,04	-0,05	-0,30
MA	0,43	0,46	0,42	0,38	0,34	0,32	0,32	0,34	0,32	0,28	0,28	0,32	0,35

Taulukko 6.1: Muuttujien ristikorrelaatiot Google-indeksin kanssa. h:n negatiiviset arvot viittaavat Google-indeksin viiveiden ja kiinnostuksen kohteena olevan muuttujan viivästämättömän arvon väliseen korrelaatioon.

Granger-kausaalisuus

Asuntojen hinnoissa Schwarzin informaatiokriteeri valitsee VAR(5)-mallin. Breusch-Godfrey-testi hylkää viiden prosentin merkitsevyystasolla nollahypoteesin autokorreloimattomuudesta, mutta Edgerton-Shukur-testi ei. Visuaalisen tarkastelun perusteella residuaalien autokorrelaatio ei kuitenkaan vaikuta ongelmalta, joten käytän lähtökohtana VAR(5)-mallia. Luvun 4 tarkastelun perusteella asuntojen hinnoissa esiintyy yksikköjuuri. Yhden differenssin jälkeen ADF-testi hylkää nollahypoteesin yksikköjuuresta, joten lopullinen spesifikaatio on VAR(5+1). Nollahypoteesi

siitä, että Google-haut eivät Granger-aiheuta hintoja, hylätään yhden prosentin merkitsevyystasolla. Toiseen suuntaan nollahypoteesia ei hylätä. Google-indeksin ajantasaista arvoa hyödyntämällä malliksi valikoituu jälleen VAR(5+1). Tulokset vastaavat mallia, jossa hyödynnetään vain Google-indeksin viiveitä: Granger-kausaalisuutta vaikuttaa olevan hauista hintoihin, mutta ei toiseen suuntaan.

Schwarzin informaatiokriteerin perusteella paras sovite kausitasoittamattomissa lukumäärissä on mallilla VAR(15). Breusch-Godfreyn testi hylkää nollahypoteesin residuaalien autokorreloimattomuudesta yhden prosentin merkitsevyystasolla, mutta Edgerton-Shukur-testi puolestaan ei hylkää nollahypoteesia. Visuaalisen tarkastelun perusteella vaikuttaa kuitenkin, että autokorrelaatiota ei olisi, joten käytän lukumäärien Granger-kausaalisuuden analysointiin VAR(15)-spesifikaatiota.

Yhden differenssin jälkeen ADF-testi hylkää nollahypoteesin, joten $d = 1$ ja siten teen Granger-kausaalisuustestin VAR(15+1)-mallin 15 ensimmäiselle kerroinmatriisille. Granger-kausaalisuustesti ei hylkää nollahypoteesia kumpaankaan suuntaan. Google-indeksin ajantasaista arvoa käyttäen malli on myös VAR(15+1). Tässäkään mallissa Granger-kausaalisuustesti ei hylkää nollahypoteesia kumpaankaan suuntaan.

Markkinointiajoissa paras malli on Schwarzin informaatiokriteerin perusteella VAR(14). Kumpikaan testi residuaalien autokorrelaatioista ei hylkää nollahypoteesia autokorreloimattomuudesta kyseisellä mallilla. ADF-testin perusteella markkinointiajoissa esiintyy yksikköjuuri, mutta yhden differenssin jälkeen aikasarja vaikuttaa olevan stationaarinen. Teen Granger-kausaalisuustestin siten mallin VAR(14+1) 14 ensimmäiselle kerroinmatriisille. Google-haut vaikuttavat valitussa mallissa Granger-aiheuttavan markkinointiaikoja 0,1 prosentin merkitsevyystasolla. Toiseen suuntaan nollahypoteesia Granger ei-kausaalisuudesta ei voida hylätä.

Kausitasoitetuilla versioilla kaikkien mallien aste on tasoittamatonta huomattavasti pienempi. Suuri ero viitannee siihen, että kausittaisuudella on suuri merkitys muuttujien dataa generoiviin prosesseihin. Hinnoissa käytetty malli on VAR(1+1), lukumäärissä VAR(3+1) ja markkinointiajoissa VAR(2+1). Hintaindeksimallissa vain Google-muuttuja on kausitasoitettu, sillä hintaindeksillä itsellään ei autokorrelaatioiden perusteella vaikuta olevan kausittaisuutta. Kahdella muulla muuttujalla sekä kiinnostusmuuttuja että Google-indeksi on kausitasoitettu. Kausitasoituksen jälkeen Granger-ei-kausaalisuuden nollahypoteesia ei voida hylätä millään muuttujista kumpaankaan suuntaan. Yhteenveto Granger-kausaalisuustestien tuloksista on taulukossa 6.2.

Yhteenvetona Granger-kausaalisuustestien tulokset antavat viitteitä siitä, että Google-haut voivat sisältää hyödyllistä informaatiota asuntojen hinnoista. Kausitasoittamattomat Google-haut vaikuttivat Granger-aiheuttavan hintoja sekä Google-indeksin ajantasaista että viivästettyä aikasarjaa hyödyntävissä malleissa. Lukumää-

		Nollahypoteesi			
		$GI \not\to y$		$y \not\to GI$	
y	Malli	GI_t	GI_{t+1}	GI_t	GI_{t+1}
HPI	VAR(5+1)	p = 0,009***	p = 0,005***	p = 0,396	p = 0,331
LKM	VAR(15+1)	p = 0,794	p = 0,249	p = 0,646	p = 0,408
MA	VAR(14+1)	p < 0,001***	-	p = 0,512	-
	Kausi-tasoitettu				
HPI	VAR(1+1)	p = 0,261	p = 0,867	p = 0,334	p = 0,612
LKM	VAR(3+1)	p = 0,842	p = 0,839	p = 0,636	p = 0,760
MA	VAR(2+1)	p = 0,720	-	p = 0,249	-

* p < 0,1; ** p < 0,05; *** p < 0,01

Taulukko 6.2: Granger-kausaalisuustestien tulokset.

rien osalta tulokset eivät ole erityisen rohkaisevia Google-hakujen ennustekyvyn kannalta. Toisaalta käytetyissä VAR-spesifikaatioissa veromuutoksen vaikutusta ei ole huomioitu, mikä voi vaikuttaa testin tuloksiin. Kausitasoittamattomilla markkinointiajoilla Google-haut vaikuttavat Granger-aiheuttavan markkinointiaikoja. Kausitasoitetussa versiossa Granger-kausaalisuutta ei kuitenkaan havaita millään muuttujalla. Ero tasoitetun ja tasoittamattoman mallin välillä voi viitata siihen, että aikasarjoilla on yhteinen kausikomponentti. Havaittu Granger-kausaalisuus tasoittamattomissa malleissa saattaa siis johtua Granger-kausaalisuudesta yhteisten kausikomponenttien välillä.

Pyrin Todan ja Yamamoton metodia hyödyntämällä välttämään muuttujien epästationaarisuudesta johtuvia ongelmia Granger-kausaalisuustestissä. Osassa näin valituista spesifikaatioista viiveiden määrä on kuitenkin huomattavan suuri verrattuna tarkasteluperiodin pituuteen. Toisaalta mallien spesifikaatioissa saattaa olla myös muita virheitä, joten liian vahvojen tulkintojen tekemistä pelkkien Granger-kausaalisuustestien perusteella tulee välttää.

Mallit

Seuraavaksi estimoin luvussa 5 kuvatut mallit sekä vertaan referenssimallien suoriutumista Google-haut sisältäviin malleihin.

Hintaindeksin vertailumallin korjattu selitysaste on 0,939, joten jo referenssimalli selittää suuren osan hintaindeksin vaihtelusta. Google-indeksin sisältävällä mallilla

Malli	(0_{HPI})	(1_{HPI})
Muuttuja		
$log(y_{t-1})$	1,311***	1,365***
	(0,168)	(0,155)
$log(y_{t-2})$	-0,417***	-0,467***
	(0,152)	(0,130)
x_t		0,0005*
		(0,0003)
x_{t-1}		-0,0004
		(0,0004)
Vakio	0,523***	0,495*
	(0,143)	(0,1825)
Havaintoja	47	47
R^2	0,939	0,946
AIC	-276,25	-278,33
BIC	-268,85	-267,23

* p < 0,1; ** p < 0,05; *** p < 0,01

Taulukko 6.3: Koko otokseen sovitetut mallit, hintaindeksi. Suluissa ilmoitettu heteroskedastisuuden ja autokorrelaation sallivat Newey-West-keskivirheet.

korjattu selitysaste on kuitenkin vielä hieman korkeampi 0,946. Google-termin nykyisen arvon kerroin on positiivinen ja eroaa nollasta kymmenen prosentin merkitsevyystasolla, mutta yhden periodin takaisen arvon kerroin on negatiivinen eikä eroa tilastollisesti merkitsevästi nollasta. Nykyisen arvon kerroin on 0,0005, mikä viittaa siihen, että nykyisen hakuintensiteetin nousu yhdellä yksiköllä näkyy asuntojen hintaindeksin 0,05 prosentin nousuna. Informaatiokriteerit eivät anna yksiselittäistä kuvaa siitä, kumpi malleista on parempi. Akaiken informaatiokriteerin arvo on pienempi Google-mallilla, mutta lisäparametreista ankarammin rankaiseva Schwarzin informaatiokriteeri suosii mallia 0_{HPI}. Yhteenveto on taulukossa 6.3.

Lukumäärämalleissa mallien korjatut selitysasteet ovat hintamalleja alempia. Referenssimallin korjattu selitysaste on 0,439 ja Google-mallin selitysaste 0,518. Google-indeksin sisältämä malli selittää siis referenssimallia paremmin kauppojen lukumäärän vaihtelua. Google-indeksin nykyisen arvon kerroin on 0,002, mutta estimaatti ei poikkea tilastollisesti merkitsevästi nollasta. Viiden kuukauden takaisen Google-muuttujan kerroin on puolestaan -0,006, eli menneisyyden yhden Google-indeksipisteen kohoaminen näkyy 0,6 prosenttia pienempänä asuntokauppojen lu-

Malli	(0_{LKM})	(1_{LKM})
Muuttuja		
$log(y_{t-1})$	0,106	0,050
	(0,226)	(0,204)
$log(y_{t-12})$	0,358***	0,223**
	(0,108)	(0,112)
x_t		0,002
		(0,002)
x_{t-5}		-0,006***
		(0,002)
$Vero$	0,120**	0,176***
	(0,050)	(0,046)
Vakio	4,540***	6,473***
	(1,629)	(1,808)
Havaintoja	69	69
Korjattu R^2	0,439	0,518
AIC	-108,13	-116,83
BIC	-96,96	-101,19

* p < 0,1; ** p < 0,05; *** p < 0,01

Taulukko 6.4: Koko otokseen sovitetut mallit, lukumäärät. Suluissa ilmoitettu heteroskedastisuuden ja autokorrelaation sallivat Newey-West-keskivirheet.

kumääränä. Viivästetyn Google-indeksin kertoimen arvo eroaa tilastollisesti merkitsevästi nollasta yhden prosentin merkitsevyystasolla. Kummassakaan mallissa lukumäärien ensimmäinen viive ei poikkea tilastollisesti merkitsevästi nollasta. Molemmat informaatiokriteerit suosivat Google-mallia. Yhteenveto on taulukossa 6.4.

Markkinointiaikojen osalta Schwarzin informaatiokriteerin perusteella paras Google-indeksillä laajennettu malli sisältää indeksin ajantasaisen arvon sekä kolmannen viiveen. Ajantasaisen Google-indeksin kerroin on negatiivinen ja poikkeaa tilastollisesti merkitsevästi nollasta viiden prosentin merkitsevyystasolla. Kolmannen viiveen kerroin on positiivinen ja eroaa nollasta tilastollisesti merkitsevästi yhden prosentin merkitsevyystasolla. Akaiken informaatiokriteeri suosii Google-indeksin sisältävää mallia, kun taas Schwarzin informaatiokriteeri suosii vertailumallia. Jälkimmäisessä ero on kuitenkin erittäin pieni. Google-indeksin sisällyttäminen malliin parantaa myös korjattua selitysastetta. Taulukossa 6.5 on yhteenveto tuloksista.

Malli	(0_{MA})	(1_{MA})
Muuttuja		
$log(y_{t-1})$	0,765***	0,765***
	(0,061)	(0,057)
$log(y_{t-12})$	0,176***	0,135*
	(0,070)	(0,067)
x_{t-1}		-0,002*
		(0,001)
x_{t-3}		0,003***
		(0,001)
Vakio	0,262	0,360**
	(0,181)	(0,179)
Havaintoja	141	141
R^2	0,747	0,764
AIC	-223,89	-229,52
BIC	-212,09	-211,83

* p < 0,1; ** p < 0,05; *** p < 0,01

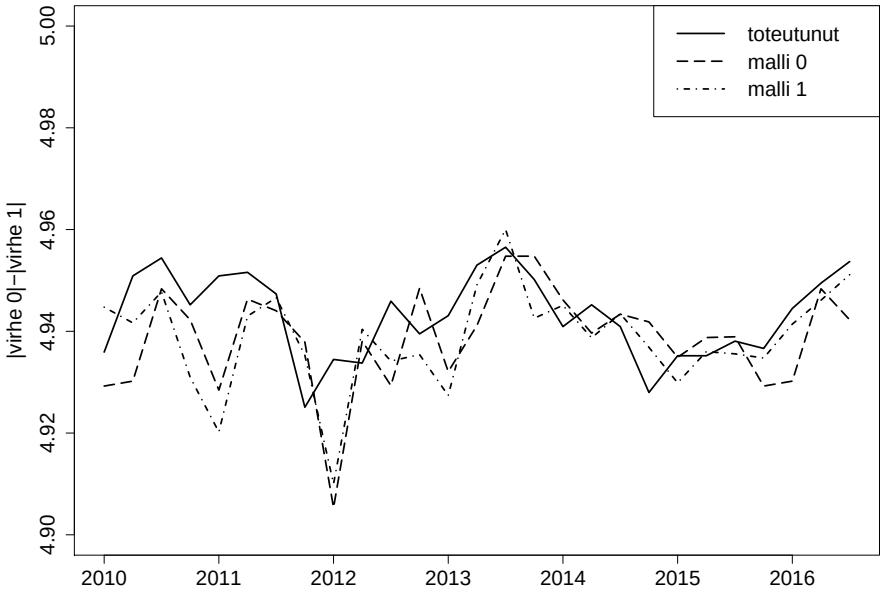
Taulukko 6.5: Koko otokseen sovitetut mallit, markkinointiajat. Suluissa ilmoitettu heteroskedastisuuden ja autokorrelaation sallivat Newey-West-keskivirheet.

Ennustaminen

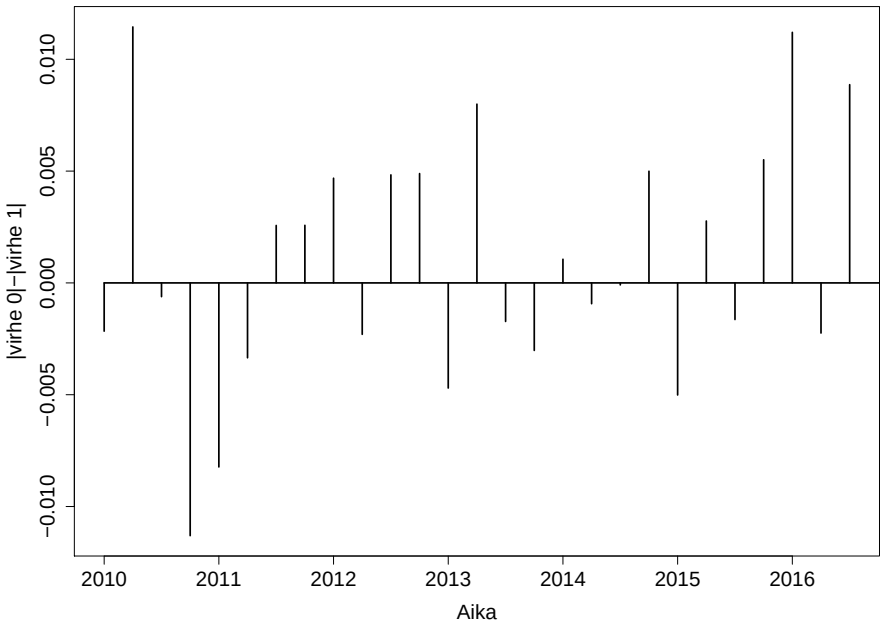
Nykyisyyden ennustaminen

Google-hakujen sisällyttäminen kiinteänpituiseen estimointiperiodiin perustuviin hintaennusteisiin pienentää absoluuttista keskivirhettä 11,1 prosentilla. Myös keskineliövirheen neliöjuuri on Google-ennusteilla pienempi. Diebold-Mariano-testi ei kuitenkaan hylkää nollahypoteesia keskivirheiden yhtäsuuruudesta. Kuva 6.1 havainnollistaa asuntokauppojen logaritmin toteutuneet arvot sekä molempien mallien antamat ennusteet. Kuvassa 6.2 on esitettynä vertailumallin absoluuttisten virheiden ja Google-hakuja hyödyntävien ennusteiden absoluuttisten virheiden erotus yli ajan.

Kahdella muulla muuttujilla tilanne näyttää varsin toisenlaiselta. Lukumääräennusteissa Google-hakujen sisällyttäminen kasvattaa ennusteiden absoluuttista keskivirhettä 15,6 prosentilla ja markkinointiajoissa 2,7 prosentilla. Nämäkään erot eivät kuitenkaan ole Diebold-Mariano-testin perusteella tilastollisesti merkitseviä. Muuttujien toteumat sekä ennusteet ovat kuvissa 6.3 ja 6.4.

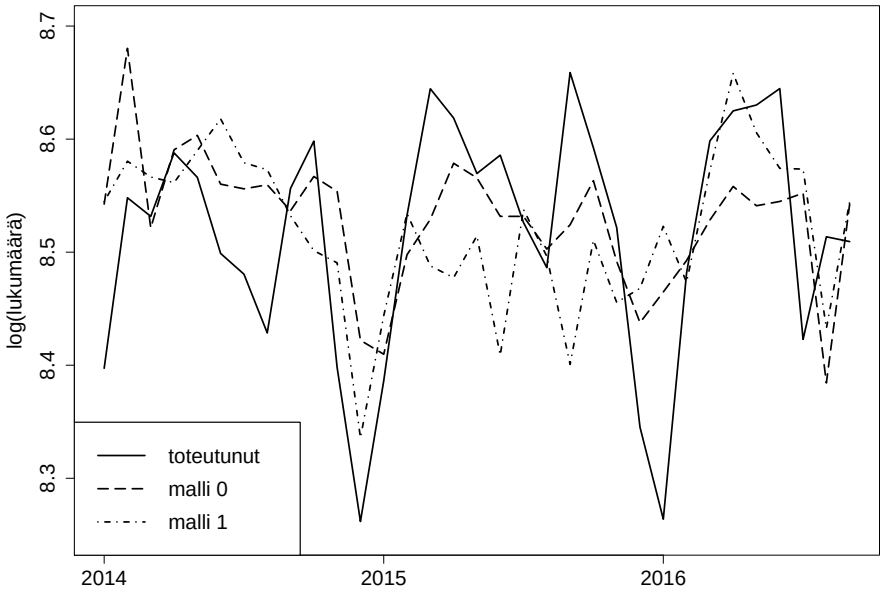


Kuva 6.1: Ennustettu sekä totetunut hintaindeksin kehitys.



Kuva 6.2: Hintaindeksin absoluuttisten ennustevirheiden erotus. Positiiviset arvot viittaavat Google-hakuja hyödyntävien ennusteiden pienempiin virheisiin.

Pidentyvään estimointiperiodiin perustuva estimointi tuottaa ennustevirhe-eroissa kvalitatiivisesti vastaavia tuloksia kuin kiinteänpituiseen estimointiperiodiin perustuvat ennusteet. Jälleen hintaennusteissa Google-indeksin hyödyntäminen parantaa ennusteita, tosin kiinteänpituiseen estimointiperiodiin verrattuna vähemmän. Vaikka



Kuva 6.3: Ennustettu sekä totetunut asuntokauppojen lukumäärän kehitys.

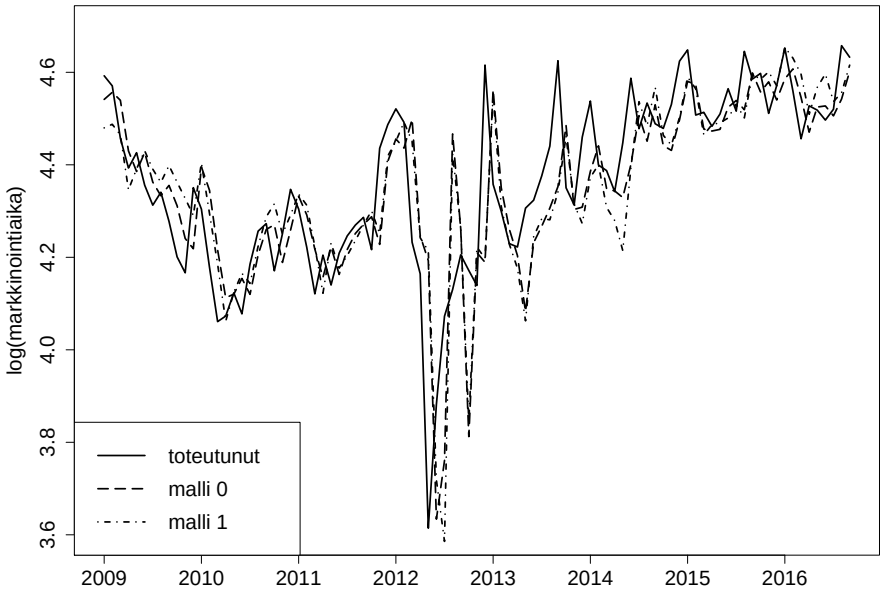
absoluuttinen keskivirhe onkin Google-hakuja käyttävällä ennusteella pienempi, keskineliövirheen neliöjuuri on niukasti suurempi. Muilla muuttujilla Google-ennusteet tuottavat lukumäärillä 18,2 prosenttia ja markkinointiajoilla 2 prosenttia suuremman absoluuttisen keskivirheen.

Lukumäärillä ja markkinointiajoilla pitenevä estimointiperiodi tuottaa pienempiä keskivirheitä kuin kiinteänpituisella estimointiperiodilla tehdyt ennusteet. Tarkkuus parantuu sekä vertailu- että Google-hakuennusteilla. Hinnoissa puolestaan kiinteänpituinen estimointiperiodi tuottaa tarkempia ennusteita.

Taulukko 6.6 tiivistää kaikille kolmelle muuttujalle tehtyjen ennusteiden erot kiinteällä estimointiperiodilla ja taulukko 6.7 pidentyvällä estimointiperiodilla. Merkitsen kiinteänpituisen estimointiperiodin vertailuennusteita 0_y:llä ja Google-hakuja hyödyntäviä ennusteita 1_y:llä. Vastaavasti pidentyvällä estimointiperiodilla merkinnät ovat vertailuennusteille 2_y ja Google-ennusteille 3_y. Merkinnöissä alaindeksi y viittaa kiinnostusmuuttujaan.

Tulevaisuuden ennustaminen

Edellisen alaluvun tulosten perusteella Google-hakujen kyky ennustaa asuntomarkkinoiden nykytilannetta ei vaikuta hintoja lukuun ottamatta erityisen vahvalta. Asuntomarkkinoilla informaationhaun ja asuntokaupan toteutumisen välillä voi kuitenkin kulua huomattavan pitkä aika, joten nykyhetken Google-haut eivät välttämättä ole erityisen hyödyllisiä asuntomarkkinoiden nykytilanteen ennustamisessa.



Kuva 6.4: Ennustettu sekä totetunut markkinointiaikojen kehitys.

Koko otokselle valikoituneet Google-mallitkin viittaisivat lukumäärien ja markkinointiaikojen osalta siihen, että menneet Google-haut voivat sisältää hyödyllistä informaatiota asuntomarkkinoiden nykyisestä tilanteesta. Lukumäärissä Scwarzin informaatiokriteeri valitsi mallin, joka sisältää Google-indeksin nykyisen sekä viiden kuukauden takaisen arvon ja markkinointiajoissa puolestaan nykyisen ja kolmen kuukauden takaisen arvon. Tässä alaluvussa teen ennusteet 1-6 periodille eteenpäin kuukausittaisille muuttujille sekä 1-2 periodille eteenpäin hintaindeksille.

Hinnoissa yhden periodin ennustehorisontilla Google-mallin ennustevirheet ovat absoluuttisella keskivirheellä mitattuna 1,1 prosenttia suuremmat kuin referenssimallilla. Kahden periodin päähän ennustettaessa absoluuttinen keskivirhe on Google-mallilla kuitenkin pienempi, mutta ero on alle 0,1 prosenttia. Erot ennustevirheissä eivät ole Diebold-Mariano-testin perusteella tilastollisesti merkitseviä.

Lukumäärillä Google-mallien ennustevirheet ovat pienempiä yhden, viiden ja kuuden kuukauden ennustehorisonteilla. Diebold-Mariano-testi ei kuitenkaan hylkää nollahypoteesia millään näistä horisonteista. Absoluuttisella keskivirheellä mitattuna suurin ero syntyy yhden kuukauden ennustehorisontilla. Google-muuttujan sisällyttäminen ennustemalliin pienentää virhettä 9,5 prosenttia ja keskineliövirheen neliöjuurta 13,2 prosenttia. Pidemmillä horisonteilla parannukset ovat alle prosentin.

Markkinointiajoissa Google-hakuja hyödyntävillä malleilla ennustevirheet ovat hieman pienempiä, kun periodin t informaatiolla ennustetaan periodin $t + 1$ arvo-

Ennuste	RMSE	MAE	MAPE	Δ MAPE[1]
(0_{HPI})	0,0108	0,0083	0,1679	-11,1%
(1_{HPI})	**0,0103**	**0,0077**	**0,1562**	
(0_{LKM})	**0,0902**	**0,0719**	**0,8476**	15,6%
(1_{LKM})	0,1065	0,0831	0,9783	
(0_{MA})	**0,1345**	**0,0937**	**2,1867**	2,7%
(1_{MA})	0,1409	0,0961	2,2441	

1: H_0: Mallin 1 virhe = Mallin 0 virhe

* p $<$ 0,1; ** p $<$ 0,05; *** p $<$ 0,01

Taulukko 6.6: Kiinteänpituisen estimointiperiodin ennusteiden keskivirheet. Estimointiperiodien pituudet ovat 20 vuosineljännestä hintaindeksille, 36 kuukautta lukumäärille ja 48 kuukautta markkinointiajoille. Ennusteperiodien pituudet ovat hintaindeksille 27 vuosineljännestä, lukumäärille 33 kuukautta ja markkinointiajoille 93 kuukautta. Pienempi ennustevirhe on lihavoitu. Δ MAE:n negatiiviset arvot viittaavat Google-hakuja hyödyntävien ennusteiden pienempään absoluuttiseen keskivirheeseen.

Ennuste	RMSE	MAE	MAPE	Δ MAPE[1]
(2_{HPI})	**0,0108**	0,0086	0,1741	-1,9%
(3_{HPI})	0,0109	**0,0084**	**0,1708**	
(2_{LKM})	**0,0823**	**0,0623**	**0,7361**	18,2%
(3_{LKM})	0,0896	0,0736	0,8648	
(2_{MA})	**0,1155**	**0,0826**	**1,9176**	2,0%
(3_{MA})	0,1189	0,0843	1,9456	

1: H_0: Mallin 1 virhe = Mallin 0 virhe

* p $<$ 0,1; ** p $<$ 0,05; *** p $<$ 0,01

Taulukko 6.7: Pitenevän estimointiperiodin ennusteiden keskivirheet. Ennusteperiodien pituudet ovat hintaindeksille 27 vuosineljännestä, lukumäärille 33 kuukautta ja markkinointiajoille 93 kuukautta. Pienempi ennustevirhe on lihavoitu. Δ MAE:n negatiiviset arvot viittaavat Google-hakuja hyödyntävien ennusteiden pienempään absoluuttiseen keskivirheeseen.

ja. Diebold-Mariano ei kuitenkaan hylkää kaksisuuntaista nollahypoteesia ennustevirheiden yhtäsuuruudesta. Pidemmälle ennustettaessa puolestaan referenssimallien ennustevirheet ovat pienempiä. Periodia $t+3$ ennustettaessa Diebold-Mariano-testin kaksisuuntainen nollahypoteesi hylätään 10 prosentin merkitsevyystasolla, mutta

muilla ennustehorisonteilla nollahypoteesia ei voida hylätä. Yhteenveto ennusteiden absoluuttisen keskivirheiden eroista on taulukossa 6.8.

	Δ MAPE					
h	1	2	3	4	5	6
HPI	1,1	-0,0				
LKM	-9,5	5,2	5,3	1,7	-2,8	-1,0
MA	-2,2	0,9	5,6*	4,0	0,8	-0,4

1: H_0: Mallin 1 virhe = Mallin 0 virhe

* p < 0,1; ** p < 0,05; *** p < 0,01

Taulukko 6.8: Absoluuttisen keskivirheen erotus, prosenttia. Luku h ilmaisee ennustetta periodille $t + h$. Negatiiviset luvut viittaavat pienempiin ennustevirheisiin Google-indeksin sisältävissä malleissa.

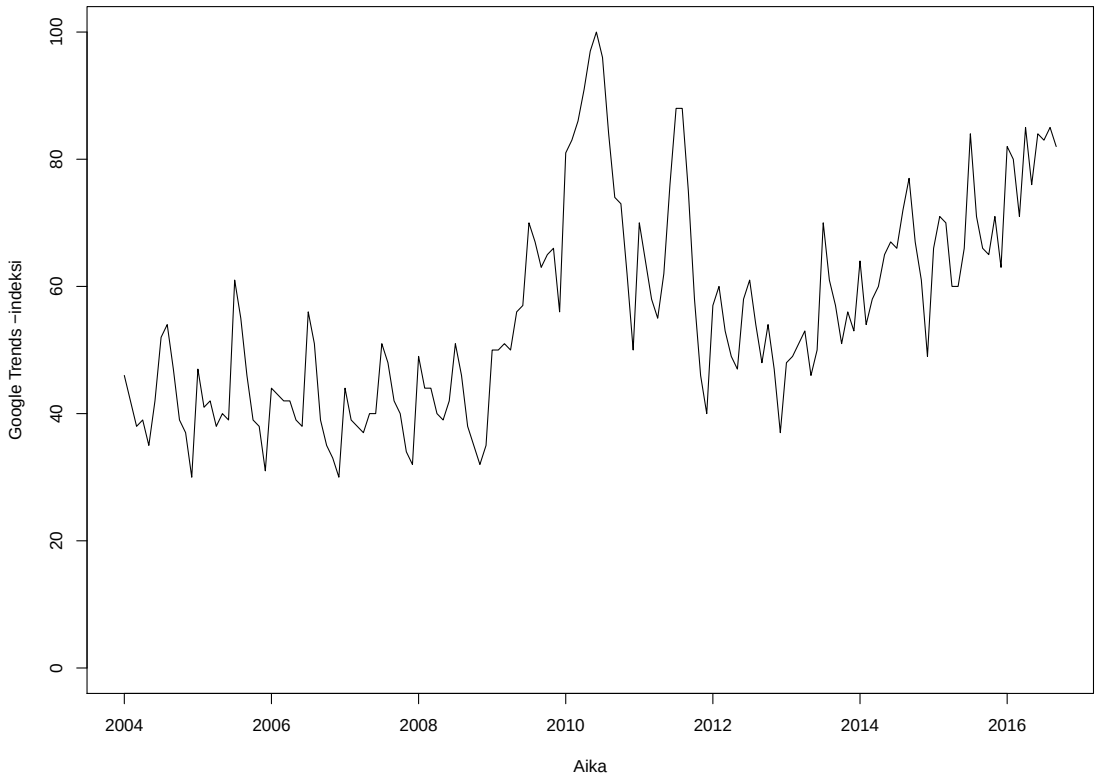
Luku 7

Robustisuus

Tarkastelen tässä luvussa saatujen tulosten robustisuutta. Hyödynnän kolmea erilaista muunnosta edellisen luvun metodeihin. Ensimmäinen liittyy Google-indeksiä rakennettaessa käytettävien hakutermien valintaan, toinen muuttujien kausitasoittamiseen ja kolmas mallien asteiden valintaan. Lisäksi analysoin myös paneelidataa hyödyntävää mallin versiota, jossa koko Suomen tasolle aggregoitujen muuttujien sijaan hyödynnän kaupunkikohtaista dataa yhdeksästä suomalaisesta kaupungista.

Hakutermien valinta

Luvun 5 mallissa hyödynsin taulukon 4.1 hakutermeistä muodostettua Google-hakuindeksiä. Näistä yhdeksästä termistä seitsemän viittaa kaupallisiin toimijoihin. Kaupallisten toimijoiden sisällyttäminen hakuindeksiin voi muodostaa haasteita, sillä palveluiden suosio alueellisesti tai yli ajan saattaa muuttua. Lisäksi esimerkiksi jokakoti-palvelua ei ole enää olemassa, vaikka se onkin yksi suosituimmista asuntoihin liittyvistä hauista hakumäärien perusteella. Haasteiden vuoksi muodostin myös Google-indeksin, joka sisältää vain sellaiset taulukon 4.1 hakutermit, jotka eivät suoraan liity kaupallisiin toimijoihin, eli *asunnot* ja *asuntoja*. Vaihtoehtoisen hakuindeksin kehitys tarkastelujaksolla on kuvassa 7.1. Visuaalisen tarkastelun perusteella vaihtoehtoisessa indeksissä vaikuttaa olevan kausittaista vaihtelua samaan tapaan kuin laajemmassa indeksissäkin. Toisaalta laajempi indeksi vaikutti nousevan trendinomaisesti vuoteen 2010, minkä jälkeen laaja indeksi laski tarkastelujakson loppuun saakka. Vaihtoehtoisessa indeksissä kehitys vaikuttaa varsin toisenlaiselta. Vuoteen 2009 saakka indeksi vaikuttaa kausivaihtelua lukuun ottamatta pysyvän lähes paikallaan, minkä jälkeen vuosina 2009-2011 hakuintensiteetti kohoaa huomattavasti aiempaa korkeammaksi. Vuonna 2012 hakuintensiteetti vaikuttaa olleen lähes samalla tasolla kuin ennen kohoamista, mutta vuodesta 2012 eteenpäin indeksissä on jälleen trendinomaista nousua.



Kuva 7.1: Google-hakuindeksin kehitys yli ajan. Käytetyt hakutermit *asunnot + asuntoja*.

Tein sekä koko otoksen kattavat mallit että ennusteet vaihtoehtoisella indeksillä. Käytin Google-indeksistä kausitasoittamatonta versiota. Kiinteänpituisella estimointiperiodilla Google-malli vaikuttaa onnistuvan hieman referenssimallia paremmin ennustamaan asuntojen hintoja, kun käytössä on laajempi hakuindeksi. Suppeammalla hakuindeksillä virheitä mittaavat suureet ovat kuitenkin suurempia Google-mallilla kuin vertailumallilla. Diebold-Mariano-testi ei kuitenkaan hylkää kaksisuuntaista nollahypoteesia virheiden yhtäsuuruudesta. Kasvavalla estimointiperiodilla suppeaa indeksiä hyödyntävä Google-malli on MAE:llä ja MAPE:lla mitattuna marginaalisesti referenssimallia parempi. Diebold-Mariano-testi ei kuitenkaan hylkää nollahypoteesia ennustevirheiden yhtäsuuruudesta.

Lukumäärissä Google-ennustemallin virheet ovat referenssimallia suurempia. Diebold-Mariano-testi hylkää nollahypoteesin virheiden yhtäsuuruudesta.

Markkinointiajoissa ennusteiden virheet vastaavat laajan indeksin tuloksia. Diebold-Mariano-testi hylkäsi kaksisuuntaisen nollahypoteesin siitä, että virheet ovat yhtäsuuret. Tämä indikoi, että Google-hauilla laajennettu malli ennustaa referenssimallia huonommin markkinointiaikojen kehitystä. Yhteenveto estimointiperiodin

Ennuste	RMSE	MAE	MAPE	Δ MAPE
(4_{HPI})	0,0120	0,0086	0,1743	3,8%
(4_{LKM})	0,1012	0,0860	1,0122	19,6%*
(4_{MA})	0,1394	0,0979	2,2824	4,5%
(5_{HPI})	0,0108	**0,0078**	**0,1575**	-5,1%
(5_{LKM})	0,0938	0,0792	0,9302	6,7%
(5_{MA})	0,1189	0,0859	1,9938	3,9%

1: H_0: Mallin 1 virhe = Mallin 0 virhe
* p < 0,1; ** p < 0,05; *** p < 0,01

Taulukko 7.1: Suppeaa Google-indeksiä hyödyntävien mallien keskivirheet. Mallien 1 ennusteet tehty kiinteänpituisen estimointiperiodin menetelmällä ja mallien 2 ennusteet pitenevän estimointiperiodin menetelmällä. Ennusteperiodien pituudet ovat hintaindeksille 27 vuosineljännestä, lukumäärille 33 kuukautta ja markkinointiajoille 93 kuukautta. Ennustevirhe on lihavoitu, jos parempi kuin referenssimallin virhe (ks. referenssimallin virhe taulukko 6.7).

ennustevirheistä on taulukossa 7.1. Merkintä 4_y viittaa kiinteänpituisella estimointiperiodilla tehtyihin Google-hakuja hyödyntäviin ennusteisiin ja 5_y pidentyvällä estimointiperiodilla tehtyihin Google-hakuja hyödyntäviin ennusteisiin.

Kausitasoitus

Lukumäärissä, markkinointiajoissa ja Google-indeksissä esiintyy aikasarjojen visuaalisen tarkastelun perusteella ja autokorrelaatioita tutkimalla kausittaisuutta. Teen kausitasoituksen kullekin edellä mainituista muuttujista X13-ARIMA SEATS-metodilla. Kausitasoituksen jälkeen aggregoin Google-muuttujan myös neljännesvuosittaiseksi hintamalleja varten. Koska reaalihintaindeksissä ei vaikuta olevan selvää kausittaisuutta, en tasoita sitä analyysiä varten.

Koko historiaan perustuvalla datalla Google-haut eivät paranna hintamallin sovitetta. Schwarzin informaatiokriteeri valitsee Google-muuttujan ajantasaisen arvon, mutta sen kerroin ei poikkea tilastollisesti merkitsevästi nollasta. Myöskään mallin selitysaste ei parane referenssimalliin verrattuna. Molemmat informaatiokriteerit ovat sekä referenssimallia että kausitasoittamatonta Google-mallia huonompia.

Kauppojen lukumäärien tasoittaminen tekee aikasarjasta lähes puhtaan satunnaisprosessin. Schwarzin informaatiokriteeri valitsee referenssimalliksi AR(1)-mallin, mutta viiveen kertoimen arvo ei poikkea tilastollisesti merkitsevästi nollasta. Lisäksi mallin selitysaste on hyvin lähellä nollaa. Schwarzin informaatiokriteerin perusteella

paras Google-malli sisältää Google-muuttujan ajantasaisen arvon ja kuuden kuukauden takaisen viiveen. Ajantasaisen Google-indeksin kohoaminen yhdellä pisteellä indikoi kauppojen lukumäärien kasvua 0,7 prosentilla, kun taas kuuden kuukauden takaisen Google-indeksin yhden pisteen muutos alentaa lukumääriä 0,7 prosentilla. Molempien kertoimet eroavat nollasta yhden prosentin merkitsevyystasolla. Myös mallin selitysaste on huomattavasti parempi ja molemmat informaatiokriteerit saavat pienemmän arvon Google-mallilla.

Markkinointiajoissa referenssimallina on AR(2), jonka molempien viivetermien kertoimet ovat nollasta poikkeavia yhden prosentin merkitsevyystasolla. Paras Google-malli sisältää Google-muuttujan yhdellä viivästetyn arvon, jonka kerroin ei kuitenkaan poikkea tilastollisesti merkitsevästi nollasta. Mallien selitysasteet ovat lähes samat, ja Google-malli on molempien informaatiokriteereiden perusteella referenssimallia huonompi. Taulukossa 7.2 on esitetty lukumäärille estimoidut mallit. Kahden muun muuttujan mallit löytyvät liitteestä ...

Huomioin periodin t ennustetta varten tehdyssä kausitasoituksessa datan vain periodiin $t-1$ asti. Muodostin siis esimerkiksi lukumääristä 34 eri kausitasoitettua aikasarjaa, joista ensimmäisen sisälsi vuosien 2010-2013 datan eli 48 havaintoa ja viimeinen koko saatavilla olevan historian syyskuuhun 2016 saakka. Tein lukumäärien aikasarjalle lisäksi veromuutoskorjauksen vuoden 2014 alusta eteenpäin ennen kausitasoitusta. Perusteluna verokorjauksen aloittamiseen vasta vuodenvaihteessa eikä välittömästi veromuutoksen jälkeen oli korjauksen suuruuden estimointiin tarvittavien havaintojen kerääminen.

Ennusteiden keskivirheiden perusteella kaikki vertailuennusteet voittavat Google-hakuja hyödyntävät ennusteet. Suurin ero on lukumäärissä, jossa vertailumallin absoluuttinen keskivirhe on peräti 36,6% Google-mallia pienempi. On tosin syytä huomauttaa, että kausitasoitetut lukumäärät vaikuttavat noudattavan lähes puhdasta satunnaisprosessia. Vertailumalli sisältää tyypillisesti vain ensimmäisen viiveen, jonka kertoimen estimoitu arvo on hyvin lähellä nollaa ja harvoin poikkeaa siitä tilastollisesti merkitsevästi. Diebold-Mariano-testi hylkää lukumäärissä nollahypoteesin ennustevirheiden yhtäsuuruudesta viiden prosentin merkitsevyystasolla, mutta muilla muuttujilla nollahypoteesia ei voida hylätä. Taulukossa 7.3 on yhteenveto kausitasoitettujen muuttujien keskivirheistä.

Viiveiden valinta *ad hoc*

Ennustemalleja tehdään myös valitsemalla sopiva malli ad hoc. Usein näin valitun mallin viivemäärä vastaa yhtä tai useampaa kausifrekvenssiä. Neljännesvuosittaiselle datalle tehdyissä malleissa käytetään siten esimerkiksi neljää tai kahdeksaa

Malli	(6_{LKM})	(7_{LKM})
Muuttuja		
$log(y_{t-1})$	-0,116	-0,254
	(1,669)	(0,154)
x_t		-0,007***
		(0,002)
x_{t-6}		0,007***
		(0,002)
Vakio	9,502***	10,630***
	(1,669)	(1,292)
Havaintoja	69	69
R^2	0,013	0,116
AIC	-156,92	-163,58
BIC	-150,22	-152,41

* $p < 0,1$; ** $p < 0,05$; *** $p < 0,01$

Taulukko 7.2: Otoksen sisäiset estimoidut mallit kausitasoitetuilla muuttujilla. Suluissa ilmoitettu heteroskedastisuuden ja autokorrelaation sallivat Newey-West-keskivirheet.

viivettä. Valitsen vertailumalleiksi kuukausittaisille muuttujille AR(12)- mallin ja hintaindeksille AR(4)-mallin. Google-malli sisältää referenssimallien AR-termien lisäksi lukumäärämalleissa Google-indeksin viiveet 0-6, markkinointiajoissa viiveet 1-6 ja hintaindeksissä 0-2. Teen ennusteet kiinteän estimointiperiodin mentelmällä. Merkitsen vertailuennusteita 8_y:llä ja Google-indeksiä hyödyntäviä ennusteita 9_y:llä.

Lukumäärissä sekä referenssi- että Google-malli suoriutuvat heikommin kuin Schwarzin informaatiokriteeriä hyödyntänyt mallin valinta. Hinnoissa Google-malli on hieman parempi absoluuttisella keskivirheellä mitattuna, mutta häviää niukasti absoluuttisella keskineliövirheellä mitattuna. Yksikään eroista ei ole Diebold-Mariano-testin perusteella tilastollisesti merkitsevä kymmenen prosentin merkitsevyystasolla.

Dynaaminen paneelidatamalli

Suomen asuntomarkkinat ovat alueellisesti eriytyneet. Kasvukeskusten ja erityisesti pääkaupunkiseudun asuntojen hintojen nousu on muuta maata nopeampaa. Helsingissä asuntojen hintojen nousu on myös ollut yleisen tulotason nousua nopeam-

Ennuste	RMSE	MAE	MAPE	Δ MAPE
(6_{HPI})	**0,0108**	0,0086	0,1741	-1,9%
(7_{HPI})	0,0109	**0,0084**	**0,1708**	
(6_{LKM})	**0,0550**	**0,0460**	**0,5412**	36,6%**
(7_{LKM})	0,0741	0,0629	0,7390	
(6_{MA})	**0,1103**	**0,0750**	**1,7393**	1,1%
(7_{MA})	0,1120	0,0758	1,7557	

1: H_0: Mallin 1 virhe = Mallin 0 virhe

* p < 0,1; ** p < 0,05; *** p < 0,01

Taulukko 7.3: Kiinteänpituisen estimointiperiodin ennusteiden keskivirheet, kausitasoitetut muuttujat. Ennusteperiodien pituudet ovat hintaindeksille 27 vuosineljännestä, lukumäärille 33 kuukautta ja markkinointiajoille 93 kuukautta. Pienempi ennustevirhe on lihavoitu.

paa. Koska myös luotonannon kriteerit ovat tiukentuneet, asuntosijoittajat lienevät keskeisessä asemassa hintakehityksen taustalla. Asuntosijoittajat ostavat erityisesti keskusta-alueen pieniä asuntoja. (Holappa et al. 2015.) Eriytymisen vuoksi Suomen asuntomarkkinoita on mielekästä tarkastella myös hienojakoisemmalla datalla kuin koko maan tasolle aggregoiduilla tiedoilla.

Analysoin Google-indeksin ennustekykyä paneelidatalla, joka hyödyntää saatavilla olevaa dataa yhdeksästä suomalaisesta kaupungista.[1] Paneelidatan hyödyntäminen mahdollistaa analyysin suuremmalla määrällä havaintoja. Esimerkiksi hinnoista on nyt kaikkiaan 459 havaintoa koko maan tasolle aggregoidun aikasarjan 51 havainnon sijaan. Käytän referenssimallina autoregressiivistä dynaamista paneelidatamallia, joka on muotoa

$$y_{it} = \sum_{j=1}^{p} \gamma_j y_{i,t-j} + \alpha_i + \epsilon_{it}. \tag{7.1}$$

Yhtälössä alaindeksi $i = 1, 2, ..., N$ viittaa poikittaisiin havaintoihin eli eri kaupunkeihin ja $t = 1, 2, ..., T$ on aikaindeksi. α_i on mahdollisesti erisuuri vakio kullekin kaupungille i. Valitsen mallien asteiksi kausittaisuuden mukaisesti 12 kuukausittaisille muuttujille ja neljä hintaindeksille. Google-indeksistä hyödynnän kaikki viiveet 0-2 hintaindeksille, 0-6 lukumäärille ja 1-6 markkinointiajoille.

Yhtälö 7.1 kuvaa niin kutsuttua kiinteiden vaikutusten mallia[2]. Kiinteiden vaikutusten mallin käyttäminen satunnaisten vaikutusten mallin sijaan on perusteltua, jos tutkittavaa aineistoa ei voi pitää satunnaisotoksena tutkittavasta populaatios-

[1] Kaupungit ovat Espoo, Helsinki, Jyväskylä, Kuopio, Lahti, Oulu, Tampere, Turku ja Vantaa.

[2] engl. fixed effects

Ennuste	RMSE	MAE	MAPE	Δ MAPE[1]
(8_{HPI})	**0,0111**	0,0086	0,1732	-10,2%
(9_{HPI})	0,0111	**0,0078**	**0,1556**	
(8_{LKM})	**0,1435**	**0,1009**	**1,1889**	29,3%
(9_{LKM})	0,1710	0,1304	1,5325	
(8_{MA})	**0,1859**	**0,1176**	**2,7347**	7,7%
(9_{MA})	0,2003	0,1273	2,9453	

1: H_0: Mallin 1 virhe $\neq$ Mallin 0 virhe

* p $<$ 0,1; ** p $<$ 0,05; *** p $<$ 0,01

Taulukko 7.4: Kiinteänpituisen estimointiperiodin ennusteiden keskivirheet, *ad hoc* valitut viiveet. Ennusteperiodien pituudet ovat hintaindeksille 27 vuosineljännestä, lukumäärille 33 kuukautta ja markkinointiajoille 93 kuukautta. Pienempi ennustevirhe on lihavoitu.

ta (Judson & Owen 1999). Myös Hausman-testi (Hausman 1978) hylkää hintojen osalta 0,1 prosentin merkitsevyystasolla nollahypoteesin siitä, että sekä satunnaisten että kiinteiden vaikutusten mallin estimaattorit ovat tarkentuvia. Hylkääminen tukee kiinteiden vaikutusten mallin valintaa.

Lyhyillä ja leveillä paneeleilla, eli pienellä T ja suurella N, käytetään usein yleistettyyn momenttimenetelmään[3] perustuvia estimaattoreita, jotka tuottavat konsistentteja estimaatteja (Baltagi & Breitung 2015). GMM-estimaattoreiden, kuten Arellano-Bond-estimaattorin (Arellano & Bond 1991), ongelmana on kuitenkin suuri instrumenttimuuttujien määrä, jos paneelin aikaulottuvuus on havainnoitavia yksilöitä suurempi (Kiviet 1995). Esimerkiksi Judson ja Owen (1999) huomaavat, että äärellisen otoskoon simulaatioissa pienimmän neliösumman dummy-muuttujaestimaattori[4] tuottaa varsin hyviä estimaatteja, kun käytettävissä olevan paneelin aikaulottuvuus on pitkä. R. P. Smith (2000) toteaa myös, että FE-tyyppiset yhdistetyt OLS-estimaattorit voivat toimia hyvin esimerkiksi ennustamisessa. Estimoin mallit myös GMM-menetelmällä, jossa rajoitan käytettävien instrumenttien määrää. Judson ja Owen (1999) havaitsevat, että instrumenttien määrän rajoittaminen ei juurikaan heikennä ennustevirheitä. Valitsen instrumenttien enimmäismääräksi seitsemän.

Taulukossa 7 on sekä kiinteiden vaikutusten LSDV- että GMM-estimaattoreita hyödyntävien mallien muuttujien estimaatit. Yleistettyyn momenttimenetelmään perustuvilla estimaattoreilla saadut estimaatit ovat samansuuntaisia kuin LSDV-

[3]engl. generalized method of moments, lyhennettynä GMM

[4]engl. least squares dummy variable, lyhennettynä LSDV

Malli	(FE_{HPI})	(GMM_{HPI})
Muuttuja		
$log(y_{t-1})$	0,797***	0,738***
	(0,073)	(0,073)
$log(y_{t-2})$	0,102	0,105
	(0,085)	(0,028)
$log(y_{t-3})$	0,035	0,038
	(0,083)	(0,028)
$log(y_{t-4})$	-0,117*	-0,123*
	(0,058)	(0,028)
x_t	0,0004**	0,0004***
	(0,0001)	(0,0001)
x_{t-1}	-0,0002**	-0,0002**
	(0,0001)	(0,0001)
x_{t-2}	0,0000	-0,0000
	(0,0001)	(0,0001)
Havaintoja	459	459

* p < 0,1; ** p < 0,05; *** p < 0,01

Taulukko 7.5: Otoksen sisäiset estimoidut mallit. Suluissa ilmoitettu heteroskedastisuuden ja autokorrelaation sallivat Arellano-keskivirheet.

estimaattoreilla saadut. Ensimmäisen AR-termin kerroin on kuitenkin pienempi GMM-estimoinnilla. Muiden parametrien osalta sekä estimoidut suuruusluokat että merkit ovat yhteneviä.

Ajantasaisen Google-muuttujan kerroin on hintamallissa 0,0004 ja poikkeaa 0,1 prosentin merkitsevyystasolla nollasta. Vastaavasti edellisen vuosineljänneksen Google-muuttujan kerroin on -0,0002 ja eroaa nollasta viiden prosentin merkitsevyystasolla. Ajantasaisen hakuaktiivisuuden nousu yhdellä indeksipisteellä näkyy siis hintaindeksin 0,04 prosentin nousuna ja vastaavasti seuraavalle vuosineljännekselle 0,02 prosentin laskuna.

Taulukko 7.6 esittää kaikkien kiinnostusmuuttujien osalta referenssimallien ja Google-mallien ennusteiden absoluuttiset keskivirheet. Ennusteissa käytetyt mallit hyödyntävät LSDV-estimaattoreita. Google-indeksin sisältävien sekä hinta- että lukumäärämallien ennustevirheet ovat vertailumallia pienempiä, ensimmäisen 5,5 prosenttia ja jälkimmäisen 8,4 prosenttia. Markkinointiajoissa Google-mallin virhe puolestaan on 0,9 prosenttia vertailumallia suurempi.

MAE	HPI	LKM	MA
Vertailu	0,0160	0,1545	0,1338
GI	0,0152	0,1415	0,1350
Δ	-5,5%	-8,4%	0,9%

Taulukko 7.6: Paneelidatamallien tuottamien ennusteiden absoluuttiset keskivirheet. Δ ilmaisee vertailumallin ja Google-mallin absoluuttisen keskivirheen prosentuaalista eroa. Prosenttiluvun negatiivinen arvo viittaa Google-mallin ennustevirheen olevan vertailumallia pienempi.

Tulosten perusteella Google-haut voivat parantaa asuntohintaindeksin nykyisen arvon ennusteita. Kuudessa spesifikaatiossa seitsemästä Google-haut pienentävät absoluuttista keskivirhettä 1,9-11,1% spesifikaatiosta riippuen. Yhdessä tapauksessa Google-hakuja hyödyntävä ennustevirhe on 3,8% vertailumallia suurempi. Lukumäärissä ennustevirheet puolestaan ovat 6,7-36,6% suurempia kuudessa spesifikaatiossa ja yhdessä 8,4% pienempi. Markkinointiajoissa kaikki Google-hakuja hyödyntävät ennusteet ovat 0,9-7,7% vertailuennusteita heikompia.

Yhtä spesifikaatiota lukuun ottamatta Google-haut paransivat hintaennusteiden tarkkuutta, mikä indikoi, että Google-haut voivat olla hyödyllisiä asuntohintaindeksin nykyarvoa ennustettaessa. Muiden muuttujien osalta Google-haut vaikuttivat pystyvän ennustamaan yhden kuukauden eteenpäin vertailumallia paremmin. Lukumäärissä ennusteet paranivat myös 5-6 kuukautta tulevaisuuteen ja markkinointiajoissa kuusi kuukautta tulevaisuuteen. Pidemmille horisonteille parannukset tosin olivat varsin pieniä, eivätkä erot tarkkuuksissa eronneet Diebold-Mariano-testin perusteella merkitsevästi nollasta. Saadut tulokset eivät puhu yhtä vahvasti Google-hakujen ennustekyvyn puolesta kuin useat aiemmin saadut empiiriset tulokset.

Luku 8

Johtopäätökset

Tarjosin tässä tutkielmassa teoreettisen intuition sille, kuinka hakuaktiivisuus voi vaikuttaa asuntomarkkinoilla asuntojen hintoihin, kauppamääriin ja myyntiaikoihin. Teoreettisen mallin perusteella kohonnut hakuaktiivisuus kasvattaa asuntokauppojen lukumäärää ja lyhentää asuntojen myyntiaikoja, mutta vaikutus hintoihin on epävarma. Tutkielman perusteella tehdyt empiiriset havainnot eivät vaikuta vahvistan yhteyttä Wheatonin mallin ja Google-hakujen välillä. Lukumäärä- ja markkinointiaikamalleissa Google-muuttujan kertoimissa havaittiin sekä positiivisia että negatiivisia arvoja, vaikka Wheatonin mallin tuoman intuition perusteella hakuaktiivisuuden lisääntymisen tulisi näkyä suurempana asuntokauppojen lukumääränä ja lyhyempinä myyntiaikoina. Google-haut eivät siten välttämättä ole oikea muuttuja kuvaamaan Wheatonin mallin mukaista hakuaktiivisuutta asuntomarkkinoilla.

Tässä tutkielmassa saadut tulokset eivät varauksetta vahvista aiemmin saatuja tuloksia Google-hakujen ennustekyvystä asuntomarkkinoille. Ennusteissa Google-hakujen havaittiin pienentävän hintaindeksin nykyisen arvon ennusteen absoluuttista prosenttikeskivirhettä kuudessa spesifikaatiossa seitsemästä. Vastaavasti lukumääräennusteissa Google-haut paransivat ennustetarkkuutta vain yhdessä spesifikaatiossa seitsemästä. Markkinointiajoissa hakudata ei parantanut tarkkuutta millään ennustespesifikaatiolla. Lähitulevaisuuden ennustaminen vaikuttaa kuitenkin hieman lupaavammalta, sillä sekä lukumäärillä että markkinointiajoilla yhden kuukauden päähän tehdyn ennusteen absoluuttinen keskivirhe on vertailumallia pienempi. Myös perusteellisempi paneelidata-analyysi voisi tarjota mielenkiintoisia tuloksia hakujen ennustekyvystä.

Toistaiseksi hakudataa hyödyntävässä ennustamisessa ongelmana on vielä muuttujien lyhyt historia. Etenkin neljännesvuosittaiseksi aggregoituna havaintojen määrä on hintaindeksimalleissa erittäin pieni. Lyhyen historian lisäksi aggregoinnissa saattaa hävitä myös tärkeää informaatiota hakujen dynamiikasta (esim. Ghysels, Santa-Clara & Valkanov 2004). Tulevaisuudessa asuntomarkkinoiden ennustamiseen

voisi käyttää esimerkiksi niin kutsuttua MIDAS-regressiota, joka pystyy hyödyntämään tiheämmän havaintotiheyden muuttujien sisältämän aggregointia paremmin (emt.). Vastaavaa menetelmää on jo käyttänyt Smith (2016) ennustaessaan Iso-Britannian työttömyyttä Google-hakujen avulla.

Kirjallisuus

Akaike, Hirotugu (1974). "A new look at the statistical model identification". *IEEE transactions on automatic control* 19(6), s. 716–723.

Andreou, Elena, Eric Ghysels ja Andros Kourtellos (2010). "Regression models with mixed sampling frequencies". *Journal of Econometrics* 158(2), s. 246–261.

Anglin, Paul M. (1997). "Determinants of buyer search in a housing market". *Real estate economics* 25(4), s. 567–589.

Arellano, Manuel ja Stephen Bond (1991). "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations". *The review of economic studies* 58(2), s. 277–297.

Askitas, Nikos ja Klaus F. Zimmermann (2009). "Google econometrics and unemployment forecasting". *German Council for Social and Economic Data (RatSWD) Research Notes*(41).

Baltagi, Badi H. ja Joerg Breitung (2015). *The Analysis of Macroeconomic Panel Data*.

Black, Jane, David de Meza ja David Jeffreys (1996). "House Prices, The Supply of Collateral and the Enterprise Economy". *The Economic Journal* 106(434), s. 60–75. ISSN: 00130133, 14680297. DOI: 10.2307/2234931.

Breusch, T. S. (1978). "Testing for autocorrelation in dynamic linear models". *Australian Economic Papers* 17(31), s. 334–355. ISSN: 1467-8454.

Brynjolfsson, Erik, Tomer Geva ja Shachar Reichman (2015). "Crowd-Squared: Amplifying the Predictive Power of Search Trend Data". *MIS Quarterly (Forthcoming)*.

Byrne, David P. ja Nicolas De Roos (2015). "Consumer Search in Retail Gasoline Markets". *Available at SSRN 2427556*.

Byrne, David P., Nicolas De Roos ja Daniel Tiong (2014). "The Internet, Search, and Asymmetric Pricing: A Natural Experiment in Retail Gasoline". *Search, and Asymmetric Pricing: A Natural Experiment in Retail Gasoline (November 16, 2014)*.

Chevillon, Guillaume ja David F. Hendry (2005). "Non-parametric direct multi-step estimation for forecasting economic processes". *International Journal of Forecasting* 21(2), s. 201–218.

Choi, Hyunyoung ja Hal Varian (2009). *Predicting initial claims for unemployment benefits*.

Choi, Hyunyoung ja Hal Varian (2012). "Predicting the Present with Google Trends". *Economic Record* 88, s. 2–9. ISSN: 1475-4932. DOI: 10.1111/j.1475-4932.2012.00809.x.

Clark, Todd E. ja Michael W. McCracken (2009). "Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts". *International Economic Review* 50(2), s. 363–395. ISSN: 00206598, 14682354.

Crawford, Gordon W. ja Michael C. Fratantoni (2003). "Assessing the forecasting performance of regime-switching, ARIMA and GARCH models of house prices". *Real Estate Economics* 31(2), s. 223.

Das, Prashant, Alan Ziobrowski ja N. Edward Coulson (2015). "Online Information Search, Market Fundamentals and Apartment Real Estate". *The Journal of Real Estate Finance and Economics* 51(4), s. 480–502.

Dickey, David A. ja Wayne A. Fuller (1979). "Distribution of the estimators for autoregressive time series with a unit root". *Journal of the American statistical association* 74(366a), s. 427–431.

Diebold, Francis X. ja Robert S. Mariano (2002). "Comparing predictive accuracy". *Journal of Business & economic statistics* 20(1), s. 134–144.

Diebold, Francis X. ja Roberto S. Mariano (1995). "Comparing Predictive Accuracy". *Journal of Business & Economic Statistics* 13(3), s. 253–263. DOI: 10.1080/07350015.1995.10524599.

Dietzel, Marian Alexander, Nicole Braun ja Wolfgang Schäfers (2014). "Sentiment-based commercial real estate forecasting with Google search volume data". *Journal of Property Investment & Finance* 32(6), s. 540–569.

Dolado, Juan J. ja Helmut Lütkepohl (1996). "Making wald tests work for cointegrated VAR systems". *Econometric Reviews* 15(4). doi: 10.1080/07474939608800362, s. 369–386. ISSN: 0747-4938. DOI: 10.1080/07474939608800362.

Edgerton, David ja Ghazi Shukur (1999). "Testing autocorrelation in a system perspective testing autocorrelation". *Econometric Reviews* 18(4), s. 343–386. ISSN: 0747-4938. DOI: 10.1080/07474939908800351.

Eerola, Essi ja Teemu Lyytikäinen (2015). "On the role of public price information in housing markets". *Regional Science and Urban Economics* 53, s. 74–84.

Elder, Harold W., Leonard V. Zumpano ja Edward A. Baryla (1999). "Buyer search intensity and the role of the residential real estate broker". *The Journal of Real Estate Finance and Economics* 18(3), s. 351–368.

Ettredge, Michael, John Gerdes ja Gilbert Karuga (2005). "Using web-based search data to predict macroeconomic statistics". *Communications of the ACM* 48(11), s. 87–92.

Ghysels, Eric, Pedro Santa-Clara ja Rossen Valkanov (2004). "The MIDAS touch: Mixed data sampling regression models". *Finance*.

Ginsberg, Jeremy et al. (2009). "Detecting influenza epidemics using search engine query data". *Nature* 457(7232), s. 1012–1014.

Godfrey, L. G. (1978). "Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables". *Econometrica* 46(6), s. 1293–1301. ISSN: 00129682, 14680262. DOI: 10.2307/1913829.

Goel, Sharad et al. (2010). "Predicting consumer behavior with Web search". *Proceedings of the National Academy of Sciences of the United States of America* 107(41). J1: Proc Natl Acad Sci U S A, s. 17486–17490. ISSN: 0027-8424; 1091-6490. DOI: 10.1073/pnas.1005962107.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". *Econometrica* 37(3), s. 424–438. ISSN: 00129682, 14680262. DOI: 10.2307/1912791.

Granger, Clive WJ (1979). "Seasonality: causation, interpretation, and implications". Teoksessa: Seasonal analysis of economic time series. NBER, s. 33–56.

Hausman, J. A. (1978). "Specification Tests in Econometrics". *Econometrica* 46(6), s. 1251–1271. ISSN: 00129682, 14680262. DOI: 10.2307/1913827.

Hohenstatt, Ralf, Manuel Käsbauer ja Wolfgang Schäfers (2012). ""Geco"and its potential for real estate research: Evidence from the US housing market". *Journal of Real Estate Research* 33(4), s. 471–506.

Holappa, Veera et al. (2015). *Alueellisten asuntomarkkinoiden kehitys vuoteen 2017*. Tekninen raportti 169. PTT työpapereita.

Jones, Jonathan D. (1989). "A comparison of lag–length selection techniques in tests of Granger causality between money growth and inflation: evidence for the US, 1959–86". *Applied Economics* 21(6), s. 809–822.

Judson, Ruth A. ja Ann L. Owen (1999). "Estimating dynamic panel data models: a guide for macroeconomists". *Economics letters* 65(1), s. 9–15.

Kholodilin, Konstantin A., Maximilian Podstawski ja Boriss Siliverstovs (2010). "Do Google searches help in nowcasting private consumption? A real-time evidence for the US". *KOF Swiss Economic Institute Working Paper No. 256*.

Kiviet, Jan F. (1995). "On bias, inconsistency, and efficiency of various estimators in dynamic panel data models". *Journal of Econometrics* 68(1), s. 53–78. ISSN: 0304-4076. DOI: http://dx.doi.org/10.1016/0304-4076(94)01643-E.

Kiviet, Jan F. (1986). "On the Rigour of Some Misspecification Tests for Modelling Dynamic Relationships". *The Review of Economic Studies* 53(2), s. 241–261. ISSN: 00346527, 1467937X. DOI: 10.2307/2297649.

Koop, Gary ja Luca Onorante (2013). "Macroeconomic nowcasting using Google probabilities". *University of Strathclyde*.

Kulkarni, Rajendra et al. (2009). "Forecasting housing prices with Google econometrics: A demand oriented approach". *GMU School of Public Policy Research Paper No. 2009-10*.

Kwiatkowski, Denis et al. (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), s. 159–178.

Lütkepohl, Helmut ja Fang Xu (2012). "The role of the log transformation in forecasting economic variables". *Empirical Economics* 42(3), s. 619–638.

Malpezzi, Stephen (1996). "Housing prices, externalities, and regulation in US metropolitan areas". *Journal of Housing Research* 7(2), s. 209.

Marcellino, Massimiliano, James H. Stock ja Mark W. Watson (2006). "A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series". *Journal of Econometrics* 135(1), s. 499–526.

McLaren, Nick ja Rachana Shanbhogue (2011). "Using internet search data as economic indicators". *Bank of England Quarterly Bulletin*(2011), Q2.

Nelson, Phillip (1970). "Information and Consumer Behavior". *Journal of Political Economy* 78(2), s. 311–329. ISSN: 00223808, 1537534X.

Newey, Whitney K. ja Kenneth D. West (1986). "A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix". *Econometrica* 55(3), s. 703–708.

Oikarinen, Elias (2012). "Empirical evidence on the reaction speeds of housing prices and sales to demand shocks". *Journal of Housing Economics* 21(1), s. 41–54.

Oikarinen, Elias (2005). "Is housing overvalued in the Helsinki metropolitan area?" *ETLA Discussion Papers, The Research Institute of the Finnish Economy (ETLA), No. 992*.

Oikarinen, Elias ja Janne Engblom (2016). "Differences in housing price dynamics across cities: A comparison of different panel model specifications". *Urban Studies* 53(11), s. 2312–2329. ISSN: 0042-0980. DOI: 10.1177/0042098015589883.

Olson, Donald R. et al. (2013). "Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales". *PLoS Comput Biol* 9(10), e1003256.

Pesaran, M. H. ja Allan Timmermann (1995). "Predictability of Stock Returns: Robustness and Economic Significance". *The Journal of Finance* 50(4), s. 1201–1228. ISSN: 00221082, 15406261. DOI: 10.2307/2329349.

Schmalz, Martin C., David A. Sraer ja David Thesmar (2017). "Housing Collateral and Entrepreneurship". *The Journal of Finance* 72(1), s. 99–132. ISSN: 1540-6261. DOI: 10.1111/jofi.12468.

Schwarz, Gideon (1978). "Estimating the dimension of a model". *The annals of statistics* 6(2), s. 461–464.

Scott, Steven L. ja Hal Varian (2013). "Bayesian variable selection for nowcasting economic time series". *NBER working papers*.

Smith, Lawrence B., Kenneth T. Rosen ja George Fallis (1988). "Recent developments in economic models of housing markets". *Journal of economic literature* 26(1), s. 29–64.

Smith, Paul (2016). "Google's MIDAS Touch: Predicting UK Unemployment with Internet Search Data". *Journal of Forecasting*.

Smith, Ron P. (2000). "Estimation and inference with non-stationary panel time-series data".

Stigler, George J. (1961). "The economics of information". *The journal of political economy*, s. 213–225.

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

Tarvonen, Ville (2016). "Googlen trendit asuntomarkkinoiden indikaattorina".

Tervo, Hannu (2000). "Migration and Labour Market Adjustment: Empirical evidence from Finland 198590". *International Review of Applied Economics* 14(3), s. 343–360. ISSN: 0269-2171. DOI: 10.1080/02692170050084079.

Thornton, Daniel L. ja Dallas S. Batten (1985). "Lag-Length Selection and Tests of Granger Causality Between Money and Income". *Journal of Money, Credit and Banking* 17(2), s. 164–178. ISSN: 00222879, 15384616. DOI: 10.2307/1992331.

Tierney, Heather LR ja Bing Pan (2012). "A poisson regression examination of the relationship between website traffic and search engine queries". *NETNOMICS: Economic Research and Electronic Networking* 13(3), s. 155–189.

Tilastokeskus (2013). *Kotitalouksien varallisuus [verkkojulkaisu]*. URL: http://www.stat.fi/til/vtutk/2013/vtutk_2013_2015-04-01_kat_002_fi.html (viitattu 18. 03. 2017).

Tilastokeskus (2016). *Osakeasuntojen hinnat [verkkojulkaisu]*. URL: http://www.stat.fi/til/ashi/2016/12/ashi_2016_12_2017-01-27_laa_001_fi.html.

Toda, Hiro Y. ja C. B. Phillips Peter (1993). "Vector Autoregressions and Causality". *Econometrica* 61(6), s. 1367–1393. ISSN: 00129682, 14680262. DOI: 10.2307/2951647.

Toda, Hiro Y. ja Taku Yamamoto (1995). "Statistical inference in vector autoregressions with possibly integrated processes". *Journal of Econometrics* 66(1), s. 225–250.

Tuhkuri, Joonas (2015). *Big Data: Do Google Searches Predict Unemployment?*

Tuhkuri, Joonas (2014). "Big Data: Google-haut ennustavat työttömyyttä Suomessa". *Etlan raportit*(31).

Vainio, Terttu, Kaisa Belloni ja Liisa Jaakkonen (2012). ”Asuntotuotanto 2030. Asuntotuotantotarpeeseen vaikuttavia tekijöitä”. VTT Technology 2.

Verbeek, Marno (2004). *A Guide to Modern Econometrics*. 2nd [rev. and updated] ed. John Wiley & Sons Ltd.

Verohallinto (2013). *Asunto- ja kiinteistöyhtiöiden osakkeiden varainsiirtoverotus*. URL: https://www.vero.fi/fi-FI/Syventavat_veroohjeet/Varainsiirtoverotus/Asunto_ja_kiinteistoyhtioiden_osakkeiden (viitattu 02. 04. 2017).

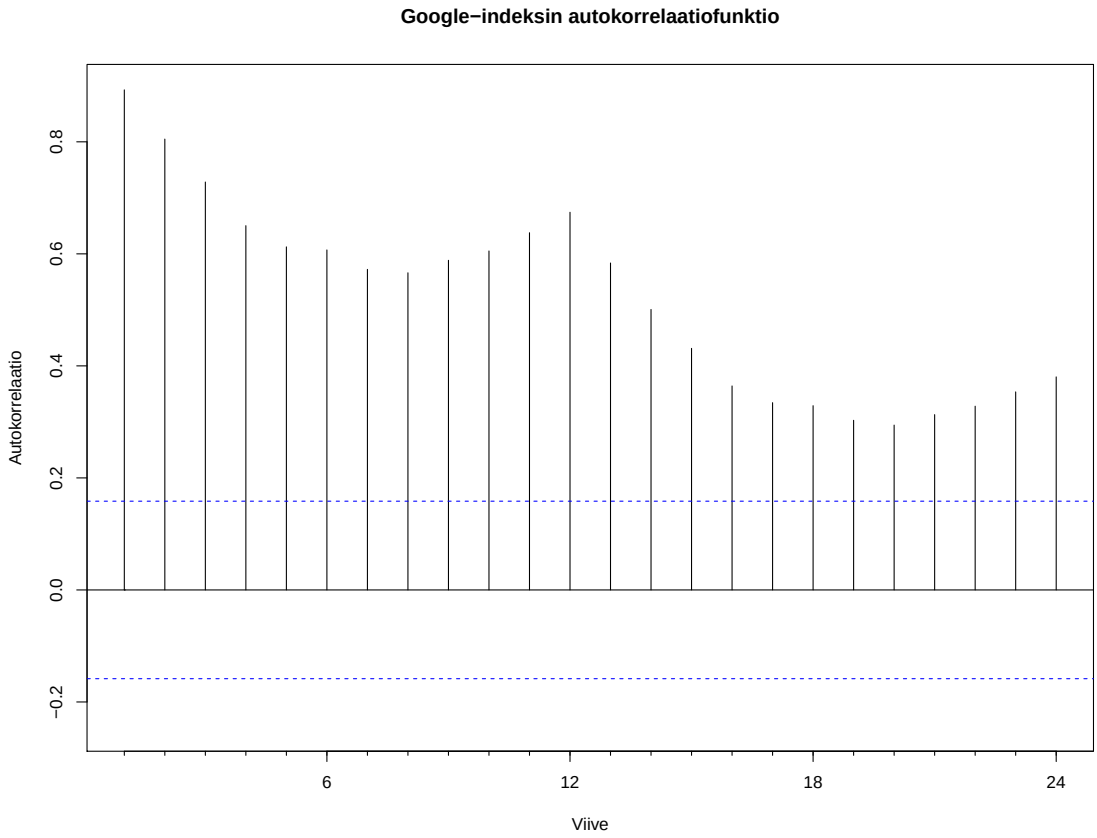
Wheaton, William C. (1990). ”Vacancy, Search, and Prices in a Housing Market Matching Model”. *Journal of Political Economy* 98(6), s. 1270–1292. ISSN: 00223808, 1537534X.

Vlastakis, Nikolaos ja Raphael N. Markellos (2012). ”Information demand and stock market volatility”. *Journal of Banking & Finance* 36(6), s. 1808–1821. ISSN: 0378-4266. DOI: http://dx.doi.org/10.1016/j.jbankfin.2012.02.007.

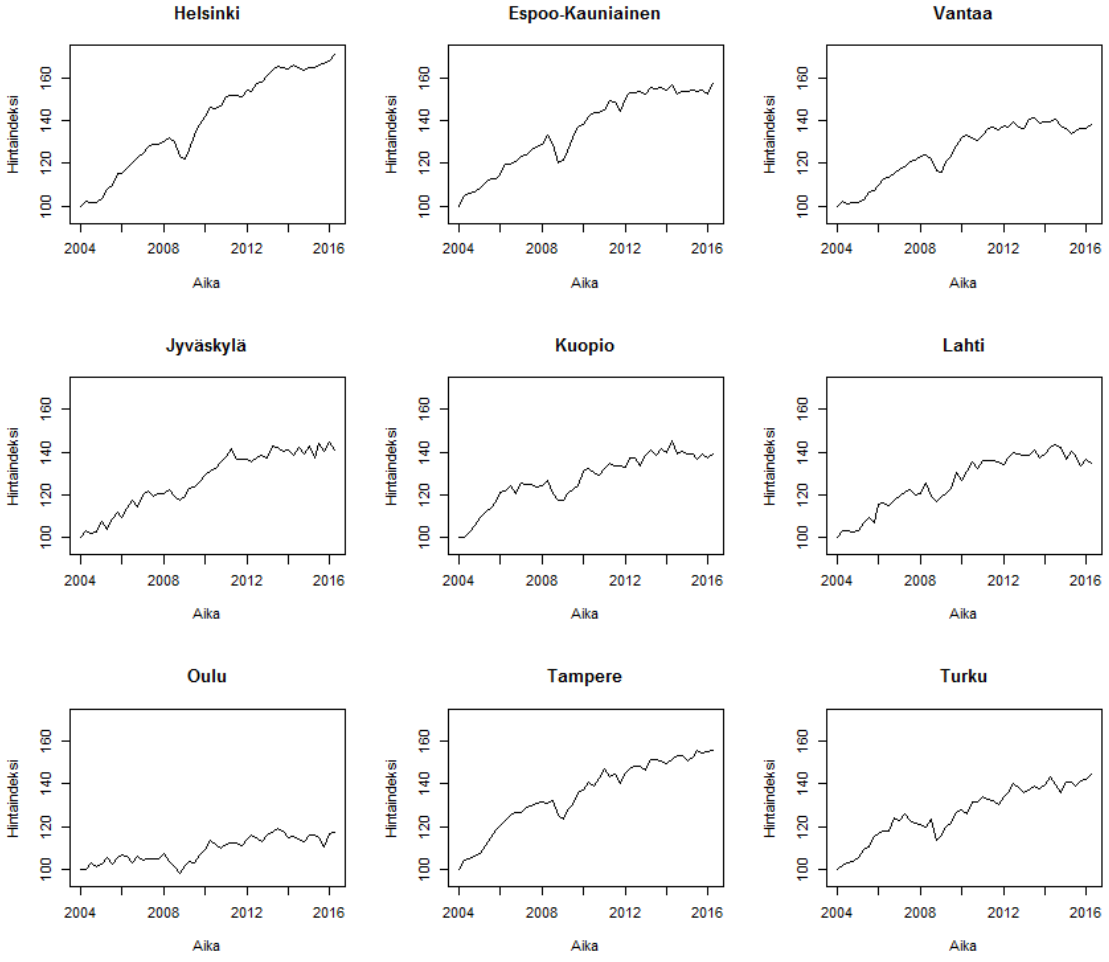
Wu, Lynn ja Erik Brynjolfsson (2014). ”The future of prediction: How Google searches foreshadow housing prices and sales”. Teoksessa: Economic analysis of the digital economy. University of Chicago Press, s. 89–118.

Liite A

Liitteet



Kuva A.1: Google-indeksin autokorrelaatiot



Kuva A.2: Vanhojen osakeasuntojen indeksoitu hintakehitys yhdeksässä suomalaisessa kaupungissa vuosineljänneksittäin. Indeksoitu, vuoden 2004 ensimmäinen vuosineljännes = 100.

Malli	(8_{HPI})	(9_{HPI})
Muuttuja		
$log(y_{t-1})$	1,318***	1,357***
	(0,212)	(0,000)
$log(y_{t-2})$	-0,413	-0,407
	(0,168)	(0,257)
$log(y_{t-3})$	-0,085	-0,133
	(0,698)	(0,469)
$log(y_{t-4})$	0,074	0,086
	(0,493)	(0,229)
x_t		0,0005**
		(0,0002)
x_{t-1}		0,0004
		(0,0004)
x_{t-2}		0,0000
		(0,918)
Vakio	0,524***	0,471
	(0,001)	(0,104)
Havaintoja	47	47
R^2	0,938	0,938
AIC	-272,62	-272,78
BIC	-261,52	-256,13

* p < 0,1; ** p < 0,05; *** p < 0,01

Taulukko A.1: Otoksen sisäiset estimoidut mallit. Suluissa ilmoitettu heteroskedastisuuden ja autokorrelaation sallivat Newey-West-keskivirheet.

Malli	(7_{HPI})
Muuttuja	
$log(y_{t-1})$	1,314***
	(0,162)
$log(y_{t-2})$	-0,431***
	(0,139)
x_t	0,0001
	(0,0003)
Vakio	0,569**
	(0,245)
Havaintoja	69
R^2	$0,939$
AIC	-274,46
BIC	-265,21

* p < 0,1; ** p < 0,05; *** p < 0,01

Taulukko A.2: Otoksen sisäiset estimoidut mallit kausitasoitetulla Google-indeksillä. Suluissa ilmoitettu heteroskedastisuuden ja autokorrelaation sallivat Newey-West-keskivirheet.

Malli	(6_{MA})	(7_{MA})
Muuttuja		
$log(y_{t-1})$	0,637***	0,631***
	(0,079)	(0,083)
$log(y_{t-2})$	0,285***	0,280***
	(0,096)	(0,093)
x_{t-1}		0,0005
		(0,0008)
Vakio	0,337*	0,342*
	(0,187)	(0,182)
Havaintoja	141	141
R^2	0,789	0,790
AIC	-262,60	-261,13
BIC	-250,81	-246,38

* p < 0,1; ** p < 0,05; *** p < 0,01

Taulukko A.3: Otoksen sisäiset estimoidut mallit kausitasoitetuilla muuttujilla. Suluissa ilmoitettu heteroskedastisuuden ja autokorrelaation sallivat Newey-West-keskivirheet.