Liike- ja toimistotonttien hintojen mallintaminen:

Esimerkkinä Suomi 2015–2019

Saul Hentunen
Helsingin yliopisto
Valtiotieteellinen tiedekunta
Taloustiede
Maisterintutkielma
Huhtikuu 2022

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiivistelmä

Tiedekunta: Valtiotieteellinen tiedekunta

Koulutusohjelma: Taloustieteen maisteriohjelma

Opintosuunta: Tutkimuksen opintosuunta

Tekijä: Saul Hentunen

Työn nimi: Liike- ja toimistotonttien hintojen mallintaminen: Esimerkkinä Suomi 2015–2019

Työn laji: Maisterintutkielma

Kuukausi ja vuosi: Huhtikuu 2022

Sivumäärä: 52

Avainsanat: liike- ja toimistotontit, tontin hinta, lineaarinen malli, regressiopuu, Gradient Boosting

Ohjaaja tai ohjaajat: Niku Määttänen

Säilytyspaikka: Helsingin yliopiston kirjasto

Muita tietoja:

Tiivistelmä:

Tonttien tilastollisilla hinta-arvioilla on käyttöä arvostuspohjaisen hintaindeksin rakentamisessa sekä suurien tonttikauppojen hintojen jaottelemisessa kohteilleen. Tämä tutkimus laajentaa aikaisempaa tutkimusta asuintonttien hinnoista tutkimalla liike- ja toimistotonttien hintoja. Tutkimuksessa selvitetään, poikkeaako toimitilatonttien hinnat asuintonttien hinnoista. Lisäksi selvitetään mallien hinta-arvioiden tarkkuutta tonttien hintojen mallintamisessa.

Tutkimus toteutetaan Maamittauslaitoksen kauppahintarekisterillä, joka sisältää tietoja Suomessa tehdyistä kiinteistö- ja tonttikaupoista. Tutkimuksessa tuodaan esille rekisteriaineiston rajauksessa käytetyt ehdot sekä aineiston tietojen täydentämiseen käytetyt aineistot ja menetelmät. Tutkimuksessa esitellään yksityiskohtaisesti tonttien hinta-arvioiden laskemiseen käytettävät mallit. Tonttien hintoja mallinnetaan lineaarisella mallilla sekä koneoppimismetodilla tehostetulla regressiopuu-mallilla. Malleissa käytetyt selittävät muuttujat on valittu rekisteriaineistosta aikaisempaa tutkimusta apuna käyttäen.

Rekisteriaineiston pohjalta on mahdollista koota useita tekijöitä, joilla voidaan arvioida tontin neliöhintaa. Mallien pohjalta ei voida kuitenkaan yksiselitteisesti sanoa, että liike- ja toimistotontit olisivat lähtökohtaisesti arvokkaampia kuin asuintontit. Poikkeavien tonttikauppojen poistamisen jälkeen koneoppimismetodilla tehostetulla regressiopuu-mallilla voidaan arvioida asuintonttien hintoja 15 prosentin tarkkuudella noin kolmannekselle tonteista. Liike- ja toimistotonteille vastaava tarkkuus saadaan noin kuudennekselle toimitilatonteista.

Tutkimuksen tuloksena suositellaan, että tonttien hintoja mallinnetaan koneoppismetodein tehostetuilla regressiopuilla lineaarisen mallin sijasta. Mallin hinta-arvioiden tarkkuuden parantamiseksi suositellaan aineiston kasvattamista aikaväliä laajentamalla ja erityisesti liike- ja toimistotonttien määrän lisäämistä tutkimusaineistoon. Lisäksi suositellaan maapohjan laatutekijöiden tarkempaa tutkimista tutkimusaineiston tonteille.

Sisällys

1 Johdanto

Yksittäisten kiinteistöjen hintojen arviointi on perinteisesti kiinteistövälittäjien ja alan konsulttifirmojen tarjoama palvelu. Kun tarvitaan kiinteistöjen hinta-arvioita koko maan rakennuskannalle, ei ole käytännöllistä hankkia ja käyttää ammattihenkilöiden arvioita kiinteistöjen hinnoista. Kiinteistöjen hintojen arvioiminen empiirisellä mallilla on edellä mainitussa tilanteessa realistinen vaihtoehto.

Maapohja ja rakennus voidaan nähdä kahtena toisistaan poikkeavina hyödykkeinä, vaikka ne yleensä myydään yhdessä (Davies ja Heathcote, 2007). Tästä johtuen on mielekästä mallintaa maapohjan ja rakennuksen hinnat erikseen. Tässä tutkimuksessa keskityn liikekiinteistöjen maapohjan hinnan mallintamiseen ja oletan, että maapohjalla olevien rakennusten hinnat ovat ennalta tiedossa. Hyödykkeenä maapohja poikkeaa monesta muusta hyödykkeestä, koska sitä ei voida tuottaa lisää ja se ei kulu käytössä. Lisäksi tonttien tarjonta on joustamatonta suhteessa kysyntään, koska maapohjan tarjontaa korkean kysynnän alueilla ei voi suoraan lisätä kysynnän mukaan. Tarjonnan joustamattomuuden seurauksena korkean kysynnän alueilla tonttien hinnat voivat nousta merkittävästi. Toisaalta alhaisen kysynnän alueilla tontin myyntihinta on usein sen vaihtoehtoiskäytön arvo, esimerkiksi käyttö viljelysmaana tai maa- ja metsätalousmaana.

Tontin sijainti, käyttötarkoitus ja maapohjan fyysiset ominaisuudet ovat merkittäviä tekijöitä tontin arvon muodostuksessa. Hyödykkeiden arvon muodostumista on käsitelty taloustieteessä varhaisimmista teoksista asti, kuten Smith (1776) ja Ricardo (1817). Rakennettavien tonttien arvoon vaikuttaa tontin sijainnin lisäksi tontin rakennuskustannukset. Tonttien arvojen määrittäminen havaittujen kauppahintojen perusteella ei ole täysin yksiselitteistä, sillä tontista maksettu hinta ei välttämättä vastaa tontin arvoa. Tontti voidaan myydä tontin arvoa alemmalla kauppahinnalla esimerkiksi tilanteessa, jossa myyjä on halukas kauppaamaan tontin nopeasti ja on valmis hyväksymään tontin arvoa alhaisemmat tarjoukset. Vastaavasti tontti voidaan myydä sen arvoa korkeampaan hintaan tilanteessa, jossa tontin kysyntä on korkea ja tarjouskilpailu nostaa kauppahintaa tontin arvoa korkeammaksi. Lisäksi eri toimijoiden näkemykset tontin arvosta voivat vaihdella. Keskityn tässä tutkimuksessa tontin arvon sijasta mahdollisen kauppahinnan arvioimiseen.

Tutkimuksessa selvitän kauppahintarekisterin soveltuvuutta maapohjan hinnan mallintamiseen sekä mallien soveltuvuutta tutkimusaineiston ulkopuolisten tonttien hintojen arvioimiseen. Maapohjan hinnan mallintamisella tarkoitetaan empiirisen mallin kykyä arvioida tontin hintaa tontista havaittujen ominaisuuksien perusteella. Osana maapohjan hinnan mallintamista selvitän, vaikuttaako tontin käyttötarkoitus liike- tai toimistokäytössä tontin hinta-arvioon. Toimitiloja koskevalle hintamallille on käyttöä arvostuspohjaisen toimitilojen hintaindeksin laatimisessa sekä suurempien kiinteistökauppojen kauppahinnan jakamisessa kohteille. Toimitilahintaindeksillä on hyödyllinen käyttötarkoitus talouden mittarina julkisen hallinnon edustajille politiikkapäätöksissä sekä yksityiselle kiinteistösijoittajalle työkaluna sijoituksen arvioinnissa (Eurostat, 2017).

Tässä tutkimuksessa toimitiloilla tarkoitan toimistokäytössä olevia rakennuksia, kuten yhden vuokralaisen tai omistajan käyttämät pääkonttorit, useamman vuokralaisen toimistorakennukset sekä toimistokampukset. Toimitilat kattavat myös liikekäytössä olevat kiinteistöt, kuten ostoskeskukset, hyper- ja supermarketit ja liikekeskukset. Toimitiloista rajaan tarkoituksen mukaisesti pois varastot ja tuotannolliset laitokset niiden käyttötarkoitusten poikkeavuuden johdosta. Tässä tutkimuksessa tontti tarkoittaa kohteen maapohjaa ja kiinteistö tarkoittaa maapohjaa sekä sille rakennettuja rakennuksia. Käytän termejä tontti ja maapohja toistensa synonyymeinä ja niillä tarkoitan kaupan kohteena olevaa maa-aluetta.

Tutkimus etenee seuraavasti. Luvussa kaksi esittelen tutkimukseen liittyvää kirjallisuutta aihealueittain. Kolmannessa luvussa esittelen mallinnuksessa käytettävän rekisteriaineiston sekä aineiston rajaamiseen ja käsittelyyn käyttämäni menetelmät. Neljännessä luvussa tuon esille tonttien hintojen mallintamisessa käyttämäni empiiriset menetelmät. Viidennessä luvussa raportoin tutkimuksen tulokset. Päätän tutkimuksen kommentoimalla tutkimuksen tuloksia ja annan ehdotukseni jatkotutkimukselle.

2 Aikaisempi tutkimus

Maapohjan hinnan empiirisen mallintamisen menetelmiä on aikaisemmissa tutkimuksissa käytetty asuintonttien hintojen mallintamiseen ilman tarkempaa selvitystä toimitilatonttien hinnoista. Mallintamisongelman lähtökohdat ovat kuitenkin yhteneviä, joten aiempaa tutkimusta hintojen empiirisestä mallintamisesta voidaan käyttää soveltaen tässä tutkimuksessa. Erityisesti aiemmassa tutkimuksessa käytettyjen tonttien ominaisuuksien pohjalta otan mallia maapohjan arvon mallintamiseen tässä tutkimuksessa. Usein tutkimusten lähtökohtana on hintojen muutosten laajempi tarkastelu ja hintaindeksien muodostaminen. Tässä tutkimuksessa keskityn yksinomaan tonttien hintojen mallintamiseen, tonttien hintojen ennustettavuuteen niiden ominaisuuksien pohjalta ja yleisen hintamallin kehittämiseen. Tutkimuksen tuloksia voidaan vastaavasti hyödyntää laajemmassa käytössä esimerkiksi arvostuspohjaisen hintaindeksin perustamisessa.

2.1 Hedoninen regressiomalli

Hedonisen, eli hyödykkeen ominaisuuksien tuottaman hyödyn, hinnoittelumallin esitteli alunperin Rosen (1974). Hedoniset mallit pohjautuvat oletukseen, että hyödykkeen tasapainohinta voidaan määrittää sen havaittavien ominaisuuksien pohjalta. Hedoninen menetelmä on käyttökelpoinen hyödykkeen hinnan arvioinnissa, kun hyödykkeen havaittavat ominaisuudet tunnetaan. Tontti on lähtökohtaisesti heterogeeninen hyödyke, jonka hinnan muodostukseen vaikuttavat tontin sijainti, käyttötarkoitus sekä maapohjan fyysiset ominaisuudet. Hedonisessa mallissa voidaan ottaa huomioon tontin ominaisuuksiin vaikuttavat tekijät, mikä tekee mallista hyödyllisen lähestymistavan maapohjan hinnan empiirisessä mallintamisessa.

Haughwout et al. (2008) ja Glumac et al. (2019) käyttävät hedonista regressiomallia kaupunkialueen maapohjan hinnan empiiriseen mallintamiseen. Molemmissa tutkimuksissa aineistona käytetään tyhjillään olevien tonttien kauppahintatietoja. Aineistoon on laskettu mukaan myös tontit, joilla olevat rakennukset ovat arvottomia ja odottavat purkua. Molemmat tutkimukset käyttävät mallintamisessa paikkakohtaisia ja maapohjan laadullisia tekijöitä mallin muuttujina. Haughwout et al. (2008) käyttämä aineisto sisältää

tonttikauppoja New Yorkin alueelta ja Glumac et al. (2019) käyttämä aineisto sisältää tonttikauppoja Luxemburgin alueelta vuosien 2010 ja 2014 väliltä. Kaupan toteutumisen ajankohta on Glumac et al. (2019) mallissa otettu huomioon siten, että mallin parametrien arvot estimoidaan jokaiselle vuodelle erikseen. Haughwout et al. (2008) sisällyttää malliinsa ajankohdan vaikutuksen siten, että ajalle sekä ajan ja tonttityypin väliselle yhteisvaikutukselle annettaan omat indikaattorimuuttujat.

2.2 Automatisoitu arvon mallintaminen

Useassa kiinteistöjen hintojen mallintamista käsittelevässä kirjallisuudessa hintamallista käytetään termiä automatisoitu arvon mallintaminen (AVM). AVM:t käyttävät hintarekisteritietoja ja tilastollisia menetelmiä kiinteistön hinnan arvioimisessa koulutetun ammattilaisen arvion sijasta. AVM ovat puhtaasti matemaattisen mallintamisen tuloksena saatu kiinteistön markkinahinta (IAAO, 2018). AVM:t ovat erityisesti hyödyllisiä tahoille, jotka tarvitsevat tehokkaasti ja vähillä kuluilla hinta-arviot usealle kohteelle. Ammattihenkilön arvio yksittäisen kiinteistön hinnasta on useissa tapauksissa tarkempi kuin tilastollisen mallin, mutta tilastollisella mallilla tuotetut hinta-arviot voidaan kustannustehokkaasti skaalata suureen määrään tontteja. Schulz et al. (2014) muodostavat hintamallin aineiston ulkopuolisten kohteiden hintojen mallintamiseen Berliinin alueella Saksassa. Tutkimus keskittyy AVM:n rakentamiseen ja mallin parametrien määrittelemiseen liittyviin haasteisiin. Pohjimmiltaan heidän luoma AVM perustuu hedoniseen regressiomalliin.

Toista lähestymistapaa käyttävät Kok et al. (2017). He esittävät, että tilastollisella mallilla saadut hinta-arviot kiinteistöstä olisivat tarkempia kuin ammattihenkilöiden tekemät hinta-arviot. He perustavat AVM:n lineaarisen mallin lisäksi koneoppimismenetelmällä tehostettuihin regressiopuihin. Regressiopuu-menetelmässä aineisto jaetaan peräkkäisessä järjestyksessä osajoukkoihin, ja jokaiseen osajoukkoon sovelletaan regressiomenetelmää parhaimman aineiston jaotteluehdon löytämiseksi. Menetelmän tavoitteena on löytää parhaimmat ennustavat muuttujat minimoimalla selittävän ja selitettävien muuttujien välistä varianssia. Koneoppimisen metodilla Kok et al. (2017) saavat regressiomallia tarkemmat tulokset selitysasteella ja prosentuaalisella mediaanivirhehajonnalla mitattuna. Empiirisiä menetelmiä, joita Kok et al. (2017) käyttävät, sovelletaan tässä tutkimuksessa tonttien

neliöhintojen mallintamisessa.

2.3 Kiinteistön sijainti

Kiinteistön sijainti on yksi oleellisista maapohjan arvoa tuottavista ominaisuuksista. Cheshire ja Sheppard (1995) käsittelevät tontin sijainnin ja ympäristötekijöiden vaikutusta tontin hintaan. He huomauttavat, että myyntihintaan voi vaikuttaa tontin ympäristötekijät, vaikka tontilla ei ole rakennusta. Tontin arvoon heijastuvat etäisyys kaupungin keskustasta ja tontin läheisen ympäristön ominaisuuksien ulkoisvaikutukset. Vaikka ympäristön ulkoisvaikutukset ovat erilaisia toimitilojen ja asuintonttien välillä, ympäristötekijöitä ei voida kuitenkaan kokonaan sivuuttaa toimitilatonttien hintojen mallintamisessa.

Suomalaisessa tutkimuksessa Loikkanen ja Laakso (2019) käsittelevät maapohjan arvon muodostusta politiikkanäkökulmasta. Kaavoitus vaikuttaa kiinteistöjen arvoihin välillisesti, sillä lähistöllä olevien julkisten palveluiden ulkoisvaikutukset välittyvät kiinteistöjen arvoihin. He vahvistavat näkemystä siitä, että kiinteistöjen hinnat ovat korkeimpia parhaan saavutettavuuden ja palveluiden läheisyydessä. Heidän tutkimuksensa pohjalta on tonttien hintojen mallintamisessa otettava mukaan muuttujia, jotka kuvaavat tontin lähistöllä olevia palveluita ja tontin saavutettavuutta.

2.4 Indeksit ja makrotalous

Aiemmassa kirjallisuudessa tutkimuksen kohteena on usein maapohjan hintojen kokonaisvaltainen tarkastelu yksittäisten tonttien hintojen arvioinnin sijaan. Hintojen aggregointi ja hintaindeksin muodostus on usein tavoitteena, kun tarkastellaan kiinteistömarkkinoita kokonaisuutena. Rakennusten ja tonttien hintojen kokonaiskehitys on taloustieteellisesti kiinnostava aihe, koska suuri osa kansan säästöistä on sijoitettu kiinteistöjen hintaan. Kiinteistömarkkinoilla on myös suuria instituutionaalisia sijoittajia, jotka tarjoavat asuntojen ja toimitilojen vuokrauspalveluita yrityksille ja henkilöasiakkaille. Kiinteistöjen hintojen muutoksilla on merkittäviä vaikutuksia kansantalouteen, koska kiinteistöjen omistaminen sitoo suuren osan henkilöiden ja yritysten varallisuudesta.

Davis ja Heathcote (2007) tarkastelevat maapohjan aggregaattihintojen kehitystä ja rakentavat maapohjan määrä- ja hintaindeksin. Maapohjan hintojen tarkastelua erikseen rakennusten hinnoista perustellaan sillä, että maapohjan ja rakennusten hinnat muodostuvat toisistaan poikkeavilla tavoilla. Oletuksena on, että rakennuksen hinnan määräytyminen ei poikkea muiden kulutushyödykkeiden hinnan määräytymisestä. Rakennuksen hintaan vaikuttavat yksinomaan rakentajan tehokkuus ja rakennusmateriaalien hinnat. Maapohjan hintaan sen sijaan heijastuu kiinteistön sijainnin arvo. Naapuruston tarjoamat julkiset palvelut konkretisoituvat maapohjan hinnassa.

Kiinteistöjen hinta-arvioiden käyttöä osana hintaindeksin laatimista esitteli Bourassa et al. (2006). SPAR-mallilla (Sales Price Appreciation Ratio) tuotettu hintaindeksi rakentuu kiinteistöjen kauppahintojen ja hinta-arvioiden suhteelle. Indeksi seuraa kauppahintojen muutoksia suhteutettuna alkuperäiseen hinta-arvioon. SPAR-mallista on hyötyä hintaindeksiä laativalle taholle, sillä se tarjoaa yksinkertaisen lähestymistavan hintaindeksin laatimisessa ilman epäluotettavuuden lisäystä muihin indeksointimenetelmiin verrattuna. Toisin kuin hedoniseen malliin perustuva indeksi, SPAR-malliin pohjaava hintaindeksi ei vaadi kattavaa yksityiskohtaista tietoa kiinteistöistä. Tämän tutkimuksen hinta-arvioita voidaan käyttää osana SPAR-mallilla rakennettua hintaindeksiä.

2000-luvun alkupuolella moni kiinteistömarkkinoita koskeva tutkimus oli keskittynyt asuntojen hintakuplan tutkimiseen (Case ja Shiller, 2003; McCarthy ja Peach, 2004; Del Negro ja Otrok, 2006). Case ja Shiller (2003) määrittävät hintakuplan tarkoittavan tilannetta, jossa liialliset yleisesti vallitsevat odotukset tulevasta hinnan noususta aiheuttavat hintojen hetkittäisen kohoamisen. Kiinteistömarkkinoiden ylikuumentumisen Yhdysvalloissa katsotaan olleen yhtenä vaikuttavana tekijänä vuoden 2007 talouskriisille. Viisi vuotta kriisin alkamisen jälkeen Liu et al. (2013) saivat näyttöä, että tonttien hinnat seuraavat suhdannesykleissä makrotaloudellisia muuttujia, kuten yritysten investointeja.

3 Aineisto

Tutkimuksessa käyttämäni aineisto pohjautuu eri viranomaislähteiden ylläpitämiin rekistereihin, jotka on toimitettu Tilastokeskukselle. Rekisterit poikkeavat sisällöltään toisistaan, sillä rekisteriä ylläpitävä taho kerää tietoa kyseisen viraston tarpeiden pohjalta. Aineistojen yhteensopivuutta on kartoitettu Tilastokeskuksessa (Tilastokeskus, 2021). Rajaan rekisteriaineistoa tutkimuskysymyksen mukaisesti. Tarkasteltavana olevat tontit voidaan tunnistaa eri rekisteriaineistosta kiinteistötunnuksen pohjalta. Tässä tutkimuksessa rajaan tarkasteltavan aineiston koskemaan viiden vuoden ajanjaksoa vuosien 2015 ja 2019 välillä.

3.1 Kauppahintarekisteri

Kauppahintarekisteri (KHR) on Maanmittauslaitoksen ylläpitämä rekisteri Suomessa toteutuneista kiinteistökaupoista ja luovutuksista. Rekisterissä näkyvät kaikki kiinteistöjen luovutukset eli kaupat, vaihdot, lahjat, jakosopimukset, kiinteistökaupan esisopimukset ja apportit. Luovutuksen kohteena voi olla koko kiinteistö, sen määräosa tai määräala sekä myös osuus yhteisiin maa- ja vesialueisiin tai yhteismetsään. Kauppahintarekisteriin merkitään kaupanvahvistajien laatimien kiinteistöluovutusilmoitusten perusteella tietoja luovutusten kohteista, kauppahinnoista sekä luovuttajista ja luovutuksen saajista. Rekisteriin sisältyvät myös kiinteistövaihdannan palvelussa tehtyjen luovutusten tiedot, jotka tehdään ilman kaupanvahvistajaa. Maanmittauslaitoksessa tarkistetaan kohteeseen liittyvät tiedot, ja rekisteriin lisätään sijainti sekä täydennetään luovutuksen tiedot kunnasta saatavilla kaavatiedoilla. Rekisterin lisätään myös kiinteistöä yksilöivä kiinteistötunnus, joka on yhtenevä muiden rekisterinpitäjien kanssa. KHR sisältää perustietoja luovutuksen kohteesta, kuten kohteen käyttötarkoituksesta, kauppahinnasta, maapohjan pinta-alasta ja luovutukseen sisältyvistä rakennuksista. (Tilastokeskus, 2021.)

Maanmittauslaitos on koonnut kauppahintarekisterin luovutustiedot vuosilta 2015– 2019 neljännesvuosittain. Yhdistän nämä neljännesvuosittaiset aineistot yhdeksi viisi vuotta kattavaksi aineistoksi. Kaikki luovutukset kattavan aineiston suurus on n. 360 000 tapahtumaa, joista noin puolet sijoittuu maa- ja metsätalousalueille. Empiiristä hintamallia

varten rajaan aineistoa niin, että se edustaa toteutuneita kauppahintoja liike- ja toimistokäyttöön kaavoitetuille tonteille. Rajaan aineistosta pois luovutukset, joiden sijaintia ei tunneta, kauppahinta jää alle 10 000 euron tai kokonaispinta-ala on vähemmän kuin 5 neliömetriä. Näin aineistosta rajautuu pois luovutukset, joilla ei ole kauppahintaa tai joissa kauppahinnan suuruus on varsin pieni. Myös pinta-alaltaan pienet kaupat saattavat vääristää aineistoa, kun mallinnettavana muuttujana käytetään tontin neliöhintaa, sillä vain muutaman neliön tonttikauppojen neliöhinnat kohoavat pienillä kokonaiskauppahinnoilla poikkeuksellisen korkeiksi.

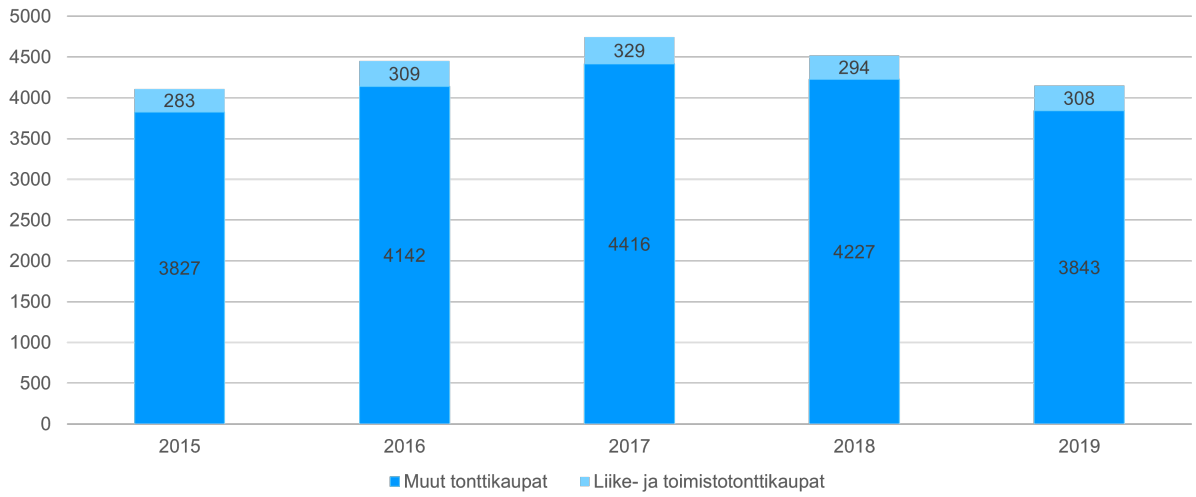
Rajaan tutkimusaineistoa edelleen käyttötarkoituksen mukaan. Valitsen tonttikaupoista kaikki kaupat, joiden käyttötarkoitus on asuin-, liike- tai toimistorakennuspaikka. Yksittäinen luovutus voi sisältää myös useamman erillisen tontin. Rajaan tutkittavaan aineistoon mukaan vain yhtä tonttia koskevat luovutukset, sillä useamman kohteen luovutukselle on ilmoitettu vain kokonaiskauppahinta. Useamman tontin kauppahintaa ei pystytä luotettavasti jakamaan kaupan kohteille, joten rajaan kyseiset kaupat mallin tarkkuuden vuoksi pois tutkittavasta aineistosta. Tutkimusaineiston käsittelyn tarkoituksena on rajata pois luovutukset, joista voidaan suurella varmuudella sanoa, että ne eivät edusta mallin kannalta oikeaa markkinahintaa. Markkinahinnalla tarkoitan rahallista summaa, jonka hyvin informoitu ostaja on valmis maksamaan ja jonka myyjä voi perustellusti hyväksyä vapailla markkinoilla. Rajausten jälkeen tutkittavan aineiston suuruus on n. 22 000 kauppaa, joista n. 500 edustaa liike- ja toimistotonttikauppoja.

3.2 Rakennushintojen erittely kiinteistökaupoista

Rakennuksia sisältävät kaupat on mielekästä ottaa osaksi käsiteltävää aineistoa, sillä ainoastaan tonttikauppoja sisältävä aineisto ei antaisi tarkkaa kuvaa tiheästi rakennettujen alueiden tonttihinnoista. Voidaan olettaa, että rakennetun alueen tontit ovat olleet ominaisuuksiltaan halutumpia kuin vielä vapaana olevat tontit ja ovat siten lähtökohtaisesti arvokkaampia. Kiinteistökohtaisista kaupoista on eroteltava rakennuksen ja tontin hinnat kokonaishinnasta. Voidaan ajatella, että rakennuksen hinta koostuu rakentamisen kustannuksista, kun taas tontin hintaan vaikuttaa aluekohtainen kysyntä sekä tontin läheisyydessä olevat palvelut (Davis ja Heathcote, 2007). Maapohjan hinta saadaan

eriteltyä kauppahinnasta jäännösarvomenetelmällä, jossa kauppasummasta vähennetään rakennuskustannukset, ja jäljelle jäävä osa edustaa kiinteistön maapohjan osuutta kokonaiskauppasummasta. Rakennuskustannusten määrittämiseen on käytetty Laukkasen ja Mäkelän (2021) tutkimusta rakennusten ikäalennusten määrittämisestä, Tilastokeskuksen ryhmiteltyä uudishintalaskuria ja Haahtela Oy:n selvitystä uudisrakentamisen hinnoista. Jäännösarvomenetelmällä tuotettuja tonttien kauppahintoja käytetään vastaamaan muiden tonttikauppojen kauppasummia.

Tonttikauppojen lisäksi valitsen tutkittavaan aineistoon mukaan rakennuksia sisältävät kiinteistökaupat, joiden rakennukset on tarkoitettu toimisto- tai liiketoimintaan. Verohallinnon ylläpitämän kiinteistörekisterin avulla tunnistan vielä tarkemmin KHR:stä luovutukset, joissa kyseessä on toimisto- tai liikerakennus. Tässä tutkimuksessa oletetaan, että rakennusten rakentamiskustannukset ikäalentumiset huomioiden ovat entuudestaan tunnettuja ja sovellettavissa KHR:n toimitiloja koskeviin kiinteistökauppoihin. Entuudestaan tunnettuja rakennusten hintoja käytetään jäännösarvomenetelmän soveltamiseen niille kiinteistökaupoille, jotka sisältävät liike- tai toimistorakennuksia. Tutkimusaineistoon lisätään 468 rakentamattoman toimitilatontin lisäksi 797 liiketonttia ja 258 toimistotonttia, joista rakennusten arvot on poistettu jäännösarvomenetelmällä.



Kuva 1: Tonttikauppojen lukumäärä vuosittain.

3.3 Aineiston muuttujat

Tonttien neliöhintojen mallintamisessa käyttämäni muuttujat voidaan jakaa kolmeen pääkategoriaan: sijaintitekijöihin, tontin ominaisuuksiin ja indikaattorimuuttujiin. Tontin sijaintitekijöihin lukeutuvat etäisyys lähimmästä kaupungista, matkustusaika lähimmälle keskuspaikalle ja tontin naapuruston väestötiedot. Sijainnin lisäksi rekisteriaineistosta tiedämme myytyjen tonttien pinta-alat ja rakennusoikeudet. Muita tontin yksilöiviä tekijöitä, kuten maapohjan muoto, maaperä ja tontin rakennettavuus, ei rekisteritiedoista ole saatavilla. Lisään aineistoon indikaattorimuuttujat vuosille 2016−2019 ja liike- ja toimistotonteille.

Aineiston muuttujien logaritmointi on mahdollista, kun muuttujien arvot ovat suurempia kuin nolla. Aineiston rajauksesta johtuen neliöhinnat ja pinta-alat ovat arvoiltaan suurempia kuin nolla. Rakennusoikeus ja naapuruston väestömuuttujat on ilmoitettu kokonaisluvuissa, ja osa niiden arvoista on nolla. Logaritmista muunnosta sovelletaan näihin muuttujiin siten, että muuttujat saavat logaritmisen arvon, jos sen arvo on yksi tai enemmän, ja muussa tapauksessa niiden arvoksi asetetaan nolla eli:

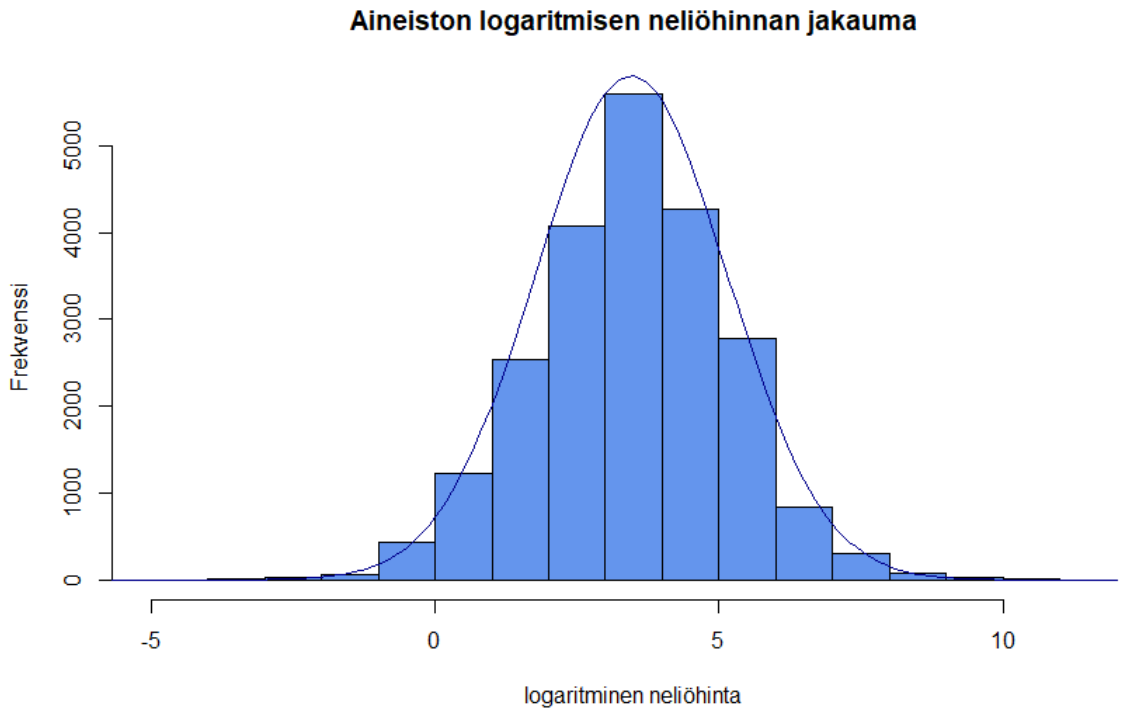
$$\ln(x_i) = \begin{cases} 0 & \text{jos } x_i < 1 \\ \ln(x_i) & \text{muulloin.} \end{cases}$$

Tällä muunnoksella muuttujat, joiden arvot ovat nolla tai yksi, kuvautuvat muunnoksen tuloksena nollaksi. Tämä ei lisää mallin virhettä merkittävästi, sillä yhden neliömetrin rakennusoikeuksia ja yhden henkilön sisältäviä väestötietoruutuja on varsin pieni osa aineistosta.

3.3.1 Selitettävä muuttuja

Maapohjan hinnan mallintamiseen valitaan selitettäväksi muuttujaksi tontin neliöhinnan logaritmi. Selitettävä muuttuja voidaan valita kauppahinnan, neliöhinnan ja niiden logaritmimuunnosten joukosta. Alustavien mallinnusten mukaisesti neliöhinnan logaritmi tuottaa tarkimmat ennusteet muihin muuttujavaihtoehtoihin verrattuna. Muuttujien logaritmimuunnoksella on suotuisa ominaisuus lineaaristen mallien kannalta, sillä muuttu-

jien välinen eksponentiaalinen suhde muuttuu lineaariseksi logaritmisen muunnoksen tuloksena. Jakaumaltaan neliöhinnan logaritmi noudattaa kutakuinkin normaalijakaumaa: $\ln(\boldsymbol{p}) \sim \mathcal{N}(\bar{p}, s^2)$, missä otoskeskiarvo on $\bar{p} = 3,44$ ja otoskeskihajonta $s = 1,69$.



Kuva 2: Selitettävän muuttujan jakauma.

Toimitilatontteihin keskittyvän empiirisen hintamallin tarkkuuden kannalta suotuisinta olisi, että liike- ja toimistotonttikauppoja olisi runsaasti ja kattavasti saatavilla. Mallin kannalta on päätettävä, valitaanko suppea aineisto, jossa kaikki kaupat edustavat toimitilakauppoja, vai otetaanko mukaan myös kauppoja, jotka eivät sisällä liike- tai toimistotontteja. Tässä tutkimuksessa päädyn jälkimmäiseen ratkaisuun. Sisällyttämällä muita tonttikauppoja tutkimusaineistoon voin vertailla, eroaako toimitilatonttien hinnat keskimäärin muusta tonttikannasta. Toteutuneiden toimitilatonttikauppojen pienen määrän takia kasvatan tutkittavaa aineistoa tonttikaupoilla, jotka hintajakaumaltaan vastaavat liike- ja toimistotonttien logaritmisen neliöhinnan jakaumaa. Lisään mallinnuksessa käytettävään aineistoon asuinrakennuspaikalla sijaitsevat tonttikaupat, sillä niiden logaritminen neliöhinnan jakauma ei merkittävästi poikkea tutkittavan aineiston jakaumasta. Lisään tutkittavaan aineistoon myös KHR:n muu rakennuspaikka -tunnuksella sijaitsevat tonttikaupat, vaikka niiden hintajakauma poikkeaa tutkittavan aineiston hintajakaumas-

ta. Osa muulla rakennuspaikalla sijaitsevista tonteista voi sisältää toimitilatarkoitukseen soveltuvia tontteja. Esitän taulukossa 1 tonttilajeittain neliöhinnan logaritmin keskiarvot, keskihajonnat, pienin ja suurin arvo sekä tonttikauppojen lukumäärä ja onko tonttilaji otettu mukaan tutkimusaineistoon. Tästä edespäin, kun viittaan tutkimuksessa hintaan, tarkoitan sillä tontin neliöhintaa, eli kauppahintaa jaettuna tontin pinta-alalla.

Tonttilaji	$\bar{p}$	s	min	max	N	Aineistossa
Liike ja toimisto	3,609	1,838594	-2,951	10,946	1 523	Kyllä
Asuinrakennuspaikka	3,530	1,631677	-4,605	11,092	18 978	Kyllä
Muu rakennuspaikka	2,1168	1,64955	-3,2189	8,3016	1 477	Kyllä
Maa- ja metsätalous	-0,9047	1,108036	-4,6052	9,9474	56 312	Ei
Lomarakennuspaikka	1,924	1,316395	-3,219	6,693	5 694	Ei

Taulukko 1: Tonttilajien yhteenveto.

3.3.2 Tontin pinta-ala

Pinta-ala on osana neliöhinnan määritelmää oletusarvoisesti, joten oleellinen kysymys on, onko pinta-alan sisällyttäminen osaksi tontin neliöhintaa arvioivaa mallia perusteltua. Tontin neliöhinta on:

$$p = \frac{P}{A},$$

missä P on tontin kauppahinta ja A on tontin pinta-ala. Neliöhinnan logaritminen muunnos voidaan esittää muodossa:

$$\ln(p) = \ln(\frac{P}{A}) = \ln(P) - \ln(A).$$

Kun pinta-ala on osana neliöhinnan regressiomallia, pinta-alan vaikutus kokonaiskauppahintaan on:

$$\ln(P) = \ln(A) + \beta \ln(A) = (1 + \beta) \ln(A),$$

missä β on neliöhinnan regressiomallin kerroin pinta-alalle. Tontin pinta-alan ottaminen mukaan regressiomalliin kertoo kaupan koon vaikutuksesta sen yksikköhintaan. Mikäli pinta-alan kerroin poikkeaa merkittävästi nollasta, regressiomalli tuloksena on, että pinta-alalla on vaikutusta kaupan yksikköhintaan. Jos pinta-ala vaikuttaa maapohjan yksikköhintaan, seuraa siitä, että kauppahinta ei ole vain neliöhinta kerrottuna maapohjan

alalla eli:

$$P \neq pA$$

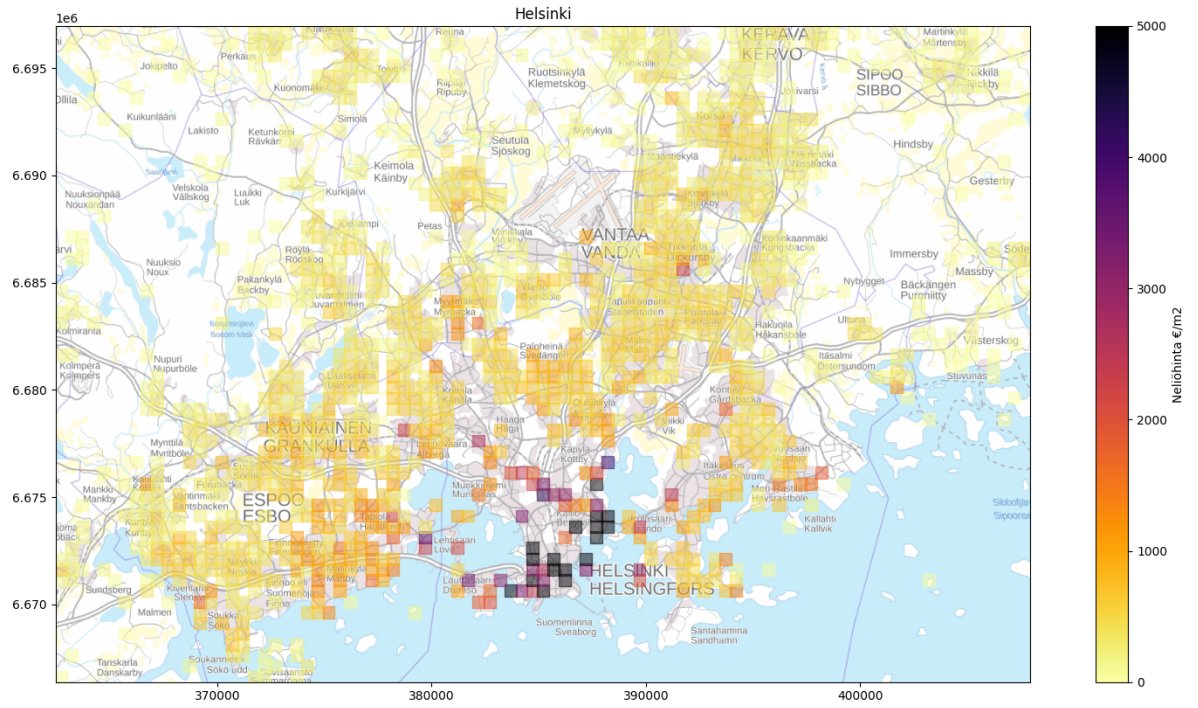
On hyvin mahdollista, että tontin pinta-alan vaikutus kauppahintaan ei ole rajattu vain määräämään kaupan kokoa, mutta se voi vaikuttaa myös kaupan yksikköhintaan. Esimerkiksi tontin ostajan halukkuus maksaa samaa yksikköhintaa rakennustarpeiden ylittävästä maapohjasta voi varsin oletettavasti laskea, mutta ostaja voi olla valmis maksamaan sovitun kokonaiskauppahinnan suuremmasta maapohjan pinta-alasta, mikäli kokonaishinta ei ylitä haluttuja kustannuksia. Ostajan maapohjan kysynnän jouston pohjalta on perusteltua, että tontin pinta-alaa käytetään neliöhintaa selittävän lineaarisen regressiomallin muuttujana.

3.3.3 Etäisyys lähimmän kaupungin keskustasta

Rekisteriaineiston sisältämien pohjois- ja itäkoordinaattien pohjalta luokittelen tonttikaupat lähimmän kaupungin mukaisesti. Havaitsen useamman tonttikaupan keskittymät kuvaamalla tutkimusaineiston tonttikaupat Suomen kartalle. Kaupungeille on myös tyypillistä, että tonttien neliöhinnat ovat korkeampia kaupungin keskustojen läheisyydessä. Tonttikauppojen keskittymien ja korkeiden neliöhintojen pohjalta valitsen kaupunkimuuttujaan 29 kaupunkia ympäri Suomea. Toisiaan hyvin lähellä olevat kaupungit tai lähiöalueet nimeän alueen suurimman kaupungin mukaan kaupunkimuuttujaan, esimerkiksi Espoossa ja Vantaalla tapahtuneet tonttikaupat merkitsen kaupunkimuuttujaan tunnuksella 'Helsinki'.

Kuvassa 3 Helsingissä toteutuneet kiinteistökaupat vuosilta 2016– 2019 kuvataan kartalle siten, että havaintoruudun väri tummenee havaintoruudun tonttikauppojen neliöhinnan keskiarvojen kasvaessa. Kartalle kuvatuista kaupoista voidaan selvästi havaita, että tonttikaupan neliöhinta kasvaa keskusta-aluetta lähestyttäessä. Sama ilmiö pätee myös kaikissa suurimmissa kaupungeissa, kuten Tampereella, Oulussa ja Turussa.

Jokainen tonttikauppa i sisältää ETRS-TM35FIN-koordinaatiston itä- ja pohjoiskoordinaatit. Koordinaatistossa pohjoiskoordinaatin arvo on päiväntasaajalla nolla ja itäkoordinaatin arvo on asetettu siten, että Suomen keskimediaanin arvo on 500 000, jotta välty-



Kuva 3: Kiinteistöjen neliöhinnat vuosilta 2015–2019 Helsingin alueella.

tään negatiivisilta itäkoordinaatin arvoilta. Yhden asteen muutos koordinaatistossa vastaa yhden metrin muutosta maastossa. Koordinaattipisteiden välinen etäisyys voidaan johtaa kahden pisteen välisestä mitasta euklidisessa avaruudessa. Etäisyysfunktio kaupungin keskustasta on:

$$\Delta(x_i, y_i, x_k, y_k) = \frac{1}{1000}\sqrt{(x_i - x_k)^2 + (y_i - y_k)^2}, \tag{1}$$

missä x_k on lähimmän kaupungin keskustan itäkoordinaatti ja y_k on pohjoiskoordinaatti. x_i ja y_i ovat puolestaan tonttikaupan i vastaavat koordinaatit. Pisteiden välinen etäisyys jaetaan vielä luvulla 1 000, jolloin etäisyyden yksiköksi saadaan kilometri. Tonttikaupan i lähin kaupunki k määritetään ehdosta:

$$k_i = \{k \mid \min_{k \in K} \Delta(x_i, y_i, x_k, y_k)\}, \tag{2}$$

missä K on kokoelma aineiston kannalta merkittävimmistä kaupungeista ja (x_k, y_k) ovat vastaavan kaupungin sijaintikoordinaatit.

3.3.4 Väestömuuttujat

Tilastokeskuksen ruututietokannan pohjalta yhdistän tonttikauppoihin sijainnin perusteella tonttia ympäröivän alueen väestötiedot. Ruututietokannassa Suomen asutut alueet on jaoteltu 250 m × 250 m ruudukkoihin. Jokainen ruudukko sisältää väestötilastoja vuodelta 2019. Tietokantaan on koottu asukastilastot asukasmäärästä, 18 vuotta täyttäneistä, työllisistä, työttömistä, lapsista, opiskelijoista ja eläkeläisistä. Tietokanta sisältää myös koulutus- ja taloustilastoja ylemmän korkeakoulututkinnon suorittaneista, talouksien keskituloista, mediaanituloista ja ostovoimakertymästä. Asumistilastoista tietokannassa on tiedot omistusasunnoissa asuvista sekä vuokra- ja asumisoikeusasunnoissa asuvista talouksista. Ruututietokannan tilastojen lisäksi väestötilastoista voidaan johtaa kullekin väestömuuttujalle prosentuaaliset osuudet ruudun koko väestöön suhteutettuna.

Ruututietokannan väestötiedot antavat paremman kuvan tonttikauppaa ympäröivästä alueesta, vaikka ne eivät suoraan kerro tontin laadusta. Käytän ruututietoja empiirisissä malleissa kuvaamaan ympäröivän alueen haluttavuutta ja laatua. Esimerkiksi ruudun sisältämä väestömäärä ja koulutustaso voivat viestittää ympäröivän alueen haluttavuudesta ja siten korreloida tonttien neliöhintojen kanssa.

3.3.5 Matkustusaika

Etäisyyden lisäksi toinen kiinnostava sijaintiin liittyvä muuttuja on matkustusaika lähimpään keskustaan tai kauppakeskittymään. Suomen ympäristökeskuksen kokoaman kauppaja keskusta-aineiston pohjalta voidaan kauppahintarekisterin aineisto yhdistää lähimpiin keskuksiin. Kauppakeskittymät ovat alueita, joilla on vähintään 50 kaupan työpaikkaa. Keskusta-alueet on määritetty kunnan toiminnalliselle keskustalle, jossa sijaitsee tiiviisti palvelutoimintoja, kuten päivittäistavarakauppoja, erikoiskauppoja, vapaa-ajan palveluja ja julkisia palveluja.

Matka-ajat kullekin tontille haetaan lähimpään kauppa- tai keskusta-alueelle henkilöautolla klo 8 arkiaamuna vuonna 2019 Digitransitin matka-aikalaskurin rajapinnasta. Matkustusajan kestoon vaikuttavat etäisyys lähimmästä keskustasta tai kauppapaikasta sekä tonttia yhdistävät liikenneyhteydet. Matka-ajan tarkasteleminen etäisyyden kanssa voi

tuoda tontin hintaan vaikuttavia tekijöitä, joita etäisyys lähimmän kaupungin keskustasta ei tuo esille. Samalla etäisyydellä keskustasta olevat tontit voivat erota toisistaan merkittävästi saavutettavuudessa, jonka vaikutus välittyy matkustusaikaan. Matka-aika korreloi tonttia yhdistävien liikenneyhteyksien kanssa, mikä oletettavasti vaikuttaa tontin hintaan. Matka-ajan kesto kuvaa myös lähellä olevien palveluiden saavutettavuutta.

3.4 Arvostusfunktio

Tontin sijainnilla on merkittävä vaikutus tontin hintaan. Kuten kuvassa 3 havaitaan, tontin neliöhinta kasvaa, mitä lähempänä keskustaa tontti sijaitsee. Tontin sijaintikohtainen hinta koostuu monesta eri tekijästä, kuten ympäröivistä palveluista, kulkuyhteyksistä ja saavutettavuudesta. Esimerkiksi suuri asukasluku, hyvät kulkuväylät ja julkisen liikenteen yhteydet ovat eduksi monelle liikeyritykselle laajan asiakaskunnan ja tehokkaan toimitusketjun mahdollistamisessa. Myös monelle yritykselle toimiston sijainti hyvien palveluiden lähellä voi lisätä yrityksen houkuttelevuutta, ja sijainti arvostetulla alueella vahvistaa yrityksen mainetta. Tonttien hintoja voi myös nostaa maapohjan ja rakennusoikeuden määrittävän kaavoituksen tarjonnan joustamattomuus kysynnän kasvaessa.

Kysynnän ja tarjonnan mallintaminen ja monien yksityiskohtaisten ominaisuuksien huomioiminen osana empiiristä hintamallia vaatii runsaasti tietoa paikallisesta kiinteistömarkkinasta ja sijainnista. Näiden tietojen puutteesta johtuen kaupunkikohtaisen käytän neliöhintojen mallintamisessa etäisyyspohjaista arvostusfunktiota. Aikaisemmassa tutkimuksessa sijaintiin liittyviä muuttujia on otettu malleissa huomioon erilaisin keinoin. Glumac et al. (2019) ovat valinneet mallissaan sijaintimuuttujiksi etäisyyden tielle, valtatielle, bussi-, juna- ja lentoasemalle sekä matkustusajan keskustaan. Haughwout et.al (2008) malli keskittyy New Yorkin tonttien neliöhinnan mallintamiseen, ja sijainti on otettu huomioon lineaarisessa mallissa tonttikaupan etäisyytenä Empire State Buildingista.

Osana tonttien hintojen mallintamista kehitän arvostusfunktion, jolla mallinnan tonttien neliöhintaa etäisyyden funktiona. Arvostusfunktion etäisyys saadaan yhtälöstä (1). Tontin neliöhinta on käänteisesti riippuvainen etäisyydestä, eli neliöhinta pienenee etäisyyden kasvaessa. Arvostusfunktion ensimmäisen asteen ehtona on: $\frac{\partial V(\delta)}{\partial \delta} < 0$. Valitsen

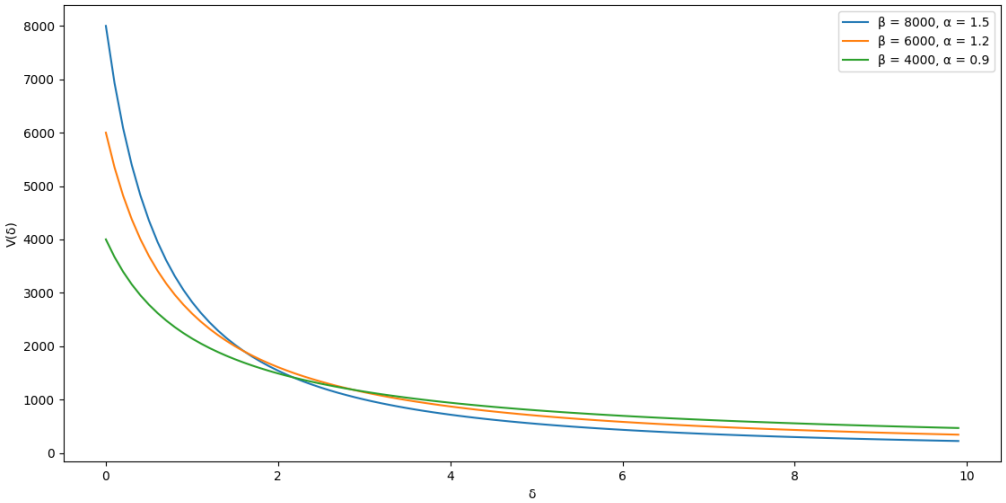
arvostusfunktion muodoksi:

$$V(\delta) = \frac{\beta}{(\delta + 1)^{\alpha}}, \tag{3}$$

missä α ja β ovat mallin estimoitavat parametrit ja $\delta = \Delta(x_i, y_i, x_k, y_k)$ on kohteen etäisyys lähimmän kaupungin keskustasta kilometreissä. Rajaan parametriarvoja siten että, $\alpha > 0$, $\beta > 0$ ja $\delta \geq 0$. Valitun arvostusfunktion ensimmäisen asteen ehto on:

$$\frac{\partial V(\delta)}{\partial \delta} = -\frac{\alpha\beta}{(\delta + 1)^{\alpha - 1}} < 0,$$

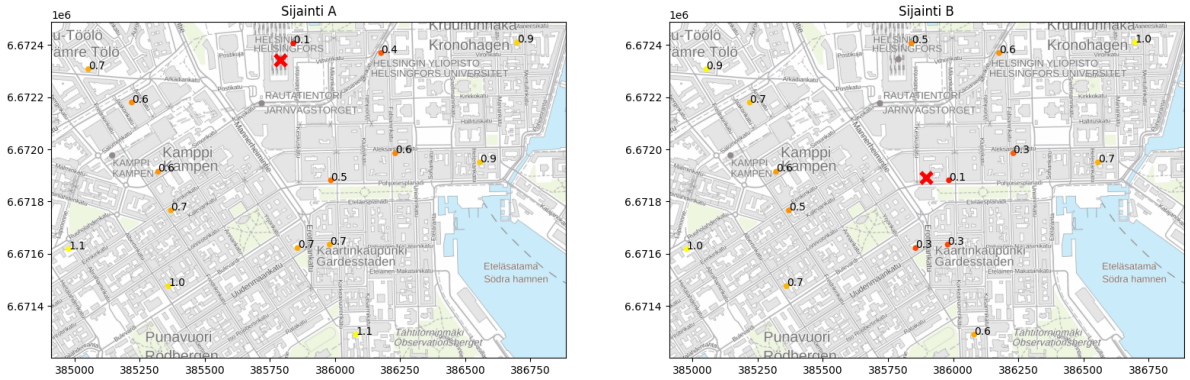
kaikilla $\alpha > 0$, $\beta > 0$ ja $\delta \geq 0$. Arvostusfunktio saa suurimman arvonsa $V(0) = \beta$, kun $\delta = 0$. Arvostusfunktiossa parametri β on arvio neliöhinnasta kaupungin keskustassa ja α-parametri kertoo, kuinka voimakkaasti neliöhinnat laskevat etäisyyden kasvaessa keskustasta. Estimoin arvostusfunktion parametriarvot jokaiselle kaupungille erikseen. Kaupungit poikkeavat toisistaan neliöhinnan suuruudessa ja alueen laajuudessa. Kaupunkikohtaiset arvostusfunktiot ottavat nämä yksilölliset vaihtelut huomioon estimoiduilla parametriarvoilla.



Kuva 4: Arvostusfunktion kuvaajat kolmella eri α- ja β-parametrien arvoilla.

Kaupungin keskustan sijaintikoordinaatit (x_k, y_k) määrittävät tonttikauppojen etäisyyden lähimmän kaupungin keskustasta. Kaupungin keskustan sijainti voidaan ottaa annettuna tai estimoida aineiston pohjalta. Ennalta määrättynä kaupungin keskipisteenä voidaan käyttää esimerkiksi maantieteellistä keskustaa. Arvostusfunktion estimointitarkkuu-

den kannalta parempi tulos saavutetaan, kun kaupungin sijaintikoordinaatit estimoidaan mallinnettavan aineiston pohjalta. Kuvassa 5 esitän, kuinka kaupungin keskustan sijainnin siirtäminen vaikuttaa lähellä tehtyjen kiinteistökauppojen etäisyyteen keskustan keskipisteestä. Kuvan sijainnissa A keskustan sijaintikoordinaatit on asetettu sijaitsemaan rautatieasemalle. Keskustan sijainti B on valittu sijaitsemaan Keskuskadun ja Pohjoisesplanadin risteykseen. Keskustan sijainti vaikuttaa kaikista eniten kiinteistökauppoihin, jotka on tehty ydinkeskustassa. Suhteellinen etäisyyden muutos on varsin pientä mitä kauemmas keskustasta tonttikauppa on tehty.



Kuva 5: Keskustapisteen sijainnin muutoksen vaikutus viereisten tonttien etäisyyteen keskustasta.

3.4.1 Arvostusfunktion parametrien estimointi

Arvostusfunktion tuottamat hinnat määräytyvät etäisyyden lisäksi funktion parametriarvojen mukaan. Parametrien estimointi tehdään jokaiselle kaupungille erikseen, jolloin kokonaisuudessaan estimoitavien parametrien lukumäärä riippuu kaupunkien lukumäärästä mallissa. Kaupungin k parametriestimaatit ovat:

$$\hat{\beta}_k = [\alpha_k, \beta_k, x_k, y_k],$$

missä β_k on estimaatti kaupungin k hinnasta ydinkeskustassa, α_k on arvio neliöhinnan muutoksesta etäisyyden kasvaessa, x_k ja y_k ovat keskustan koordinaatit.

Määritetään vielä tavoitefunktio $\Psi(x)$, jonka arvot määräytyvät arvostusfunktion estimaattien ja tonttikauppojen neliöhinnan välisestä erotuksesta. Arvostusfunktion estimaatin ja rekisteröidyn tonttikaupan i neliöhinnan välinen estimointivirhe kaupungissa k on:

$$e_i = p_i - V_k(\delta_i) = p_i - \frac{\beta_k}{(\delta_i + 1)^{\alpha_k}},$$

missä tonttikaupan i neliöhinta on p_i. Olkoon tavoitefunktio virheiden neliöiden summa:

$$\Psi(\hat{\beta}_k) = \sum_{i=1}^{n} e_i^2 \tag{4}$$

Valittu tavoitefunktio pohjaa lineaarisissa malleissa yleisesti käytettyyn pienimmän neliösumman menetelmään. Virheiden neliöinnillä vältytään negatiivisten ja positiivisten virheiden summaamiselta yhteen ja siten päätymällä virheellisesti pieneen virheiden summaan. Virheiden neliöinnin vaikutus on, että yksittäisten virheiden suuret arvot korostuvat tavoitefunktiossa. Kaupungin k arvostusfunktiolle valitut parametriestimaatit on valikoitu parametriarvojen joukosta, jotka minimoivat tavoitefunktion arvoa:

$$\mathbf{b}_k \equiv \arg\min_{\hat{\beta}_k} \Psi(\hat{\beta}_k). \tag{5}$$

Selvitän parametriarvot $\mathbf{b}_k$ kokoelmalle kaupunkeja K numeerisia menetelmiä käyttäen. Ensimmäisessä vaiheessa teen karkean haun mahdollisille parametriarvoille $\hat{\beta}_k$ toteuttamalla pisteruudukkohaku kaupungin keskustassa. Kokeilen itä- ja pohjoiskoordinaatin arvoja karttakeskustasta 100 metrin välein 500 metrin säteeltä. Kartalle muodostuu siis 1000 m × 1000 m pisteruudukko. Kokeilen α-parametrin arvoja väliltä [1,2] 0,1 välein ja β-parametrin arvot valitaan arvojen [1000, 8000] joukosta 500 välein. Parametrikombinaatioita on yhteensä n. 20 000. Syötän parametriarvot arvostusfunktioon ja sovitan arvostusfunktion aineistoon. Järjestän parametriarvot paremmuusjärjestykseen tavoitefunktion arvojen mukaan. Ensimmäiseen vaiheen paras parametrivalikoima on se, jonka tavoitefunktion arvo on pienin.

Käytän parametriarvojen estimoinnin toisessa vaiheessa Python-ohjelmiston pakettia *scipy optimize* ja paketin *least squares*-metodia (The SciPy community). *Least squares*-metodi tarvitsee lähtöarvauksen parametriarvoille sekä tavoitefunktion. Käytän ensim-

mäisen vaiheen pisteruudukkohaun parasta parametrivalikoimaa toisessa vaiheessa parametriarvojen lähtöarvauksena ja tavoitefunktiona käytän funktiota, joka on esitetty yhtälössä (4). *Least square*-metodi löytää tavoitefunktion paikallisen minimin parametriarvojen joukosta. Tarvittaessa vaiheet yksi ja kaksi voidaan toistaa, jotta varmistutaan, että lopputulos ei ole vain paikallinen minimi, vaan on myös kaupunki alueen k globaali minimipiste parametriarvoille $\hat{\beta}_k$ ja siten täyttää yhtälön (5) ehdon.

Kuva 6: Kaupunkien keskustapisteet ja arvostusfunktion keskustapisteiden estimaatit.

Kuvassa 6 esitän neljän suurimman kaupungin keskustojen estimoidut sijainnit punaisella rastilla. Arvostusfunktion estimaattivirheen kannalta sijoitetut keskustat eivät merkittävästi poikkea kaupunkien maantieteellisistä keskustoista. Maantieteelliset keskustat on merkitty kuvaan 6 sinisellä rastilla.

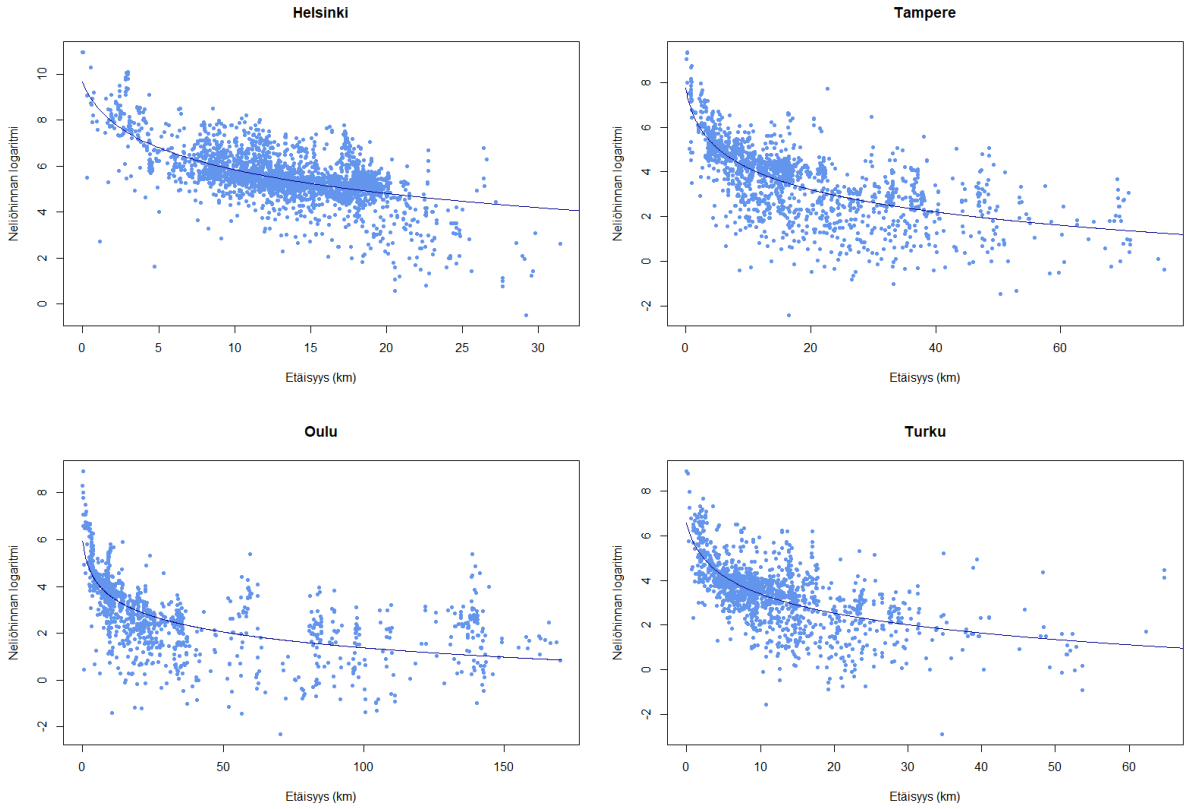
3.4.2 Arvostusfunktion logaritmi

Arvostusfunktio voidaan muuntaa lineaariseksi ottamalla funktiosta logaritminen muunnos. Lineaarisen muunnoksen etuna on se, että mallia voidaan käyttää osana lineaarista mallia. α- ja β-parametrejä ei lineaarisessa mallissa tarvitse etukäteen estimoida, sillä kyseiset parametrit estimoidaan osana lineaarista mallia. Lineaarisen mallin tarkkuutta

voidaan kuitenkin parantaa estimoimalla etukäteen kaupunkien keskustojen sijaintikoordinaatit. Sijaintikoordinaattien estimoimisella etäisyysmuuttujan δ arvoja voidaan käyttää osana lineaarista mallia. Arvostusfunktion logaritminen muunnos on:

$$\ln(V(\delta_i)) = \ln(\frac{\beta}{(\delta_i + 1)^{\alpha}}) = \ln(\beta) - \alpha \ln(\delta_i + 1) \tag{6}$$

Kuvassa 7 havainnollistan arvostusfunktion logaritmisen muunnoksen sopivuutta aineiston logaritmiseen neliöhintaan neljässä suurimmassa kaupungissa.



Kuva 7: Arvostusfunktion sovite neljän suurimman kaupungin tonttikauppa-aineistoon.

4 Mallit

Tässä kappaleessa esittelen tonttien hintojen mallintamiseen sovellettavat menetelmät, jotka ovat lineaarinen regressiomalli, regressiopuu-malli sekä regressiopuu-malliin sovellettava Gradient Boosting -koneoppimismenetelmä. Mallien esittelyssä korostan niiden toimintaa ja soveltuvuutta tonttien neliöhintojen mallintamisessa. Mallit, joita käsittelen tässä tutkimuksessa, on valittu aikaisemman tutkimuksen pohjalta (Glumac et al., 2019; Haughwout et al., 2008; Kok et al., 2017).

4.1 Lineaarinen malli

Nimensä mukaisesti lineaarisen mallin oletuksena on, että selitettävän ja selittävien muuttujien välinen yhteys on lineaarinen. Lineaarista regressiomallia kutsutaan joskus myös ekonometrian perusmalliksi (Hayashi, 2011). Sen etuina on, että malli on yksinkertaista estimoida ja suurin osa tilastollisista ohjelmistoista kykenee estimoimaan lineaarisen mallin sisäänrakennetuilla toiminnoilla. Tässä tutkimuksessa lineaarista regressiomallia käytetään mittapuuna, johon muiden mallien suorituskykyä aineiston mallinnuksessa ja mallin ennusteiden tarkkuutta verrataan. Lineaarisen mallin selittävien muuttujien kertoimet valitaan pienemmän neliösumman menetelmällä, joka vastaa yhtälön (4) tavoitefunktion minimoimisongelmaa, ja yhtälön (5) parametrien estimaatit vastaavat lineaarisen mallin kertoimien estimaatteja.

Lineaariseksi regressiomalliksi valitaan tutkimusaineiston muuttujien pohjalta malli, joka on muotoa:

$$\ln(p_i) = \beta_0 + \boldsymbol{\beta_1} \ln(\boldsymbol{x}_i) + \beta_2 \boldsymbol{I}_i + \beta_3 \boldsymbol{r}_i + \ln(V(\delta_i)) + \epsilon_i \quad (i = 1, 2, \ldots, n), \tag{7}$$

missä $\boldsymbol{x}$ on muuttujien rakennusoikeuden, pinta-alan ja matkustusajan muuttujavektori, $\boldsymbol{I}$ on toimitilan ja vuosien 2016-2019 indikaattorivektori, $\boldsymbol{r}$ on ruututietokannan muuttujavektori ja $\ln(V(\delta_i))$ on arvostusfunktion logaritmi kuten yhtälössä (6).

Käytän lineaarista mallia tässä tutkimuksessa kahteen tarkoitukseen. Ensimmäiseksi estimoin lineaarisen mallin kertoimet käyttäen koko tutkimusaineistoa, jotta saadaan pa-

rempi ymmärrys, miten selittävät muuttujat vaikuttavat neliöhintaan. Mallin yksittäisten muuttujien tilastollisen merkittävyyden testaus t-testillä kertoo, kuinka oletettavaa on, että muuttujan vaikutus poikkeaa merkittävästi nollasta mallin perusteella. Toiseksi jaan tutkimusaineiston koulutus- ja testiaineistoon. Estimoin lineaarisen mallin kertoimet koulutusaineistolla ja testiaineistolla selvitän, kuinka tarkasti se ennustaa mallin ulkopuolisten tonttikauppojen hintoja. Sen lisäksi, että malli istuu tutkimusaineistoon, selvitän, miten malli soveltuu koulutusaineiston ulkopuolisten tonttien hintojen arvioimiseen. Tämä ominaisuus on oleellinen, kun mallia käytetään maapohjan hintojen estimoimiseen.

Lineaarisessa regressiomallissa otan ajankohdan vaikutuksen mallissa huomioon indikaattorimuuttujalla. Indikaattorimuuttujavalikoima voidaan koota edustamaan vuositasolla tai tiheämmällä aikavälillä kuten neljännesvuosittain tai kuukausittain. Mallin kannalta valitsen aikavälin, joka parantaa mallin tarkkuutta eniten. Aluksi kokeilen malliin indikaattorimuuttujia vuositasolla ja neljännesvuositasolla. Neljännesvuositason indikaattorimuuttujien lisääminen vain marginaalisesti parantaa mallin tarkkuutta lisäämällä malliin 15 indikaattorimuuttujaa enemmän verrattuna vuositasolla tarkasteltavaan ajanjaksoon. Neljännesvuositason indikaattorit ovat mallin kannalta parempi jättää pois ja keskittyä vuositason muutosten tarkkailuun.

Empiirisen mallin yksi tutkimuskohde on toimitilatonttien hintojen poikkeavuus asuinrakennuskäyttöön tarkoitetuista tonteista. Lisään malliin kolme indikaattorimuuttujaa kuvaamaan rakennettuja tontteja liike- ja toimistotiloille sekä rakentamattomia tontteja, jotka on kaavoitettu liike- tai toimistokäyttöön. Rakennettujen tonttien luokittelun teen liike- ja toimistotiloihin verohallinnon kiinteistörekisterin mukaan. Kiinteistörekisteristä löytyy tarkemmat käyttötarkoitustiedot toimitilakäytössä oleville kiinteistöille, jonka mukaan jaottelu liike- ja toimistotiloihin on mahdollista. Tontit, joilla rakennusta ei vielä ole, mutta jotka on kaavoitettu toimitilakäyttöön, tunnistan aineistosta KHR:n käyttötarkoituslajin pohjalta. Liiketiloja aineistossa on yhteensä 797, toimistoja on 258 ja rakentamattomia toimitilatontteja on yhteensä 468. Indikaattorimuuttujan käyttäminen toimitilatonttien hintojen erittelemiseen muusta aineistosta on karkea tapa tutkia tonttilajin vaikutusta neliöhintaan. Indikaattorimuuttujat toimitilatonteille kertoo, miten toimitilatonttien hinnat poikkeavat keskimäärin muusta tonttikannasta.

	Ei muunnosta	%-osuus	logaritminen
Asukkaat yhteensä	0,008***	0,006***	0,841***
	(0,0003)	(0,0001)	(0,019)
18 vuotta täyttäneet	0,015***	0,243**	−0,391***
	(0,002)	(0,114)	(0,051)
Työlliset	−0,010***	−0,386***	0,352***
	(0,001)	(0,076)	(0,035)
Lapset	0,008***	0,061	−0,357***
	(0,001)	(0,126)	(0,015)
Eläkeläiset	−0,0002	−0,556***	0,184***
	(0,001)	(0,087)	(0,014)
Koulutetut	0,009***	0,111*	−0,153***
	(0,001)	(0,062)	(0,049)
Ylempi tutkinto	−0,008***	2,876***	0,782***
	(0,001)	(0,062)	(0,013)
Omistusasunto	−0,022***	−0,335***	−0,379***
	(0,002)	(0,090)	(0,021)
Vuokra-asunto	−0,029***	0,612***	−0,011
	(0,001)	(0,098)	(0,009)
Vakio	2,655***	2,801***	1,333***
	(0,013)	(0,044)	(0,028)
Havainnot	21 978	21 978	21 978
R^2	0,350	0,377	0,558
Tarkistettu R^2	0,350	0,377	0,558
Virheiden keskihajonta (va = 21 968)	1,359	1,331	1,121
F-tunnusluku (va = 9; 21 968)	1 315***	1 478***	3 085***
Note:			*p<0,1; **p<0,05; ***p<0,01

Taulukko 2: Väestömuuttjien ja neliöhinnan regressio.

Ruututietokanta sisältää tilastotiedot väestö- ja kotitalouksien määrästä sekä ikä- ja koulutusjakaumasta ruuduittan. Tietokanta sisältää myös taloustietoja ja tietoja asumismuodosta. Ruututietokannassa on useita muuttujia, ja empiirisen mallin kannalta ongelmana

on valita neliöhinnan muutoksia parhaiten kuvaavat muuttujat. Tietokannan muuttujista voidaan tarkastella prosentuaalista osuutta tai logaritmista muunnosta. Prosentuaalisia osuuksia verrataan koko väestömäärään tai talouksien lukumäärään yhteensä. Taulukossa 2 esitän ruututietokannan muuttujien ja logaritmisen neliöhinnan välisen lineaarisen regression tuloksia. Taulukossa kertoimien keskivirheet esitän suluissa kertoimen estimaatin alapuolella. Ruututietojen muunnosten kykyä selittää tonttien neliöhintoja vertailen selitysasteen ja mallin virheiden keskihajonnan mukaan. Logaritminen muunnos ruututietokannan muuttujista on selvästi muita muunnoksia kykenevämpi selittämään neliöhinnan muutosta. Ruututietokannan muuttujien tilastollinen merkittävyys kuitenkin muuttuu, kun malliin lisätään muut muuttujat. Lineaarisen regressiomalliin muuttujavalikoimaan vaikuttaa lopulta, mitkä muuttujat lisäävät mallin tarkkuutta. Osa tarjolla olevista muuttujista tulee karsia pois joko multikollineaarisuuden tai tilastollisen merkitsevyyden puuttumisen takia.

Jotta lineaarista mallia voidaan käyttää tonttikauppojen neliöhintojen mallintamiseen, tulee lineaarisen mallin oletusten olla voimassa. Lineaarisen mallin oletukset ovat, että selitettävän ja selittävien muuttujien välinen suhde on lineaarinen, selittävät muuttujat eivät korreloi mallin virheen kanssa, muuttujat eivät ole multikollineaarisia, eli selittävät muuttujat ovat arvoiltaan yksilöllisiä, ja mallin virheet ovat homoskedastisia, eli virhetermien varianssi pysyy vakiona mallissa. Aineiston logaritmoinnin seurauksena neliöhinnan ja selittävien muuttujien välinen suhde on lineaarinen.

Tontin pinta-alan sekä rakennusoikeuden sisällyttäminen osaksi mallin selittäviä tekijöitä voi rikkoa mallin oletusta muuttujien välisestä multikollineaarisuudesta. Kaavoitus määrittää rakennusoikeuden, ja rakennushankkeelle myönnetään rakennuslupa, mikäli rakennus vastaa alueen rakentamista määrittävää kaavoitusta. Rakennuslupa voi silti vahvasti korreloida tontin pinta-alan kanssa, sillä tontin pinta-ala määrittää, kuinka suuri rakennus on mahdollista rakentaa yhdessä rakennusoikeuden kanssa. Muuttujien välinen kollineaarisuus heikentää muuttujien vaikutusten estimoinnin tarkkuutta ja lisää muuttujien varianssia mallissa. Muuttujan korkea varianssi heikentää muuttujan tilastollista merkitsevyyttä.

Testaan muuttujien välistä kollineaarisuutta varianssin paisuttamiskertoimella VIF.

$$VIF = \frac{1}{1 - R^2},$$

missä R^2 on tutkittavien muuttujien välisen regression selitysaste. Korkea VIF:n arvo kertoo muuttujien välisestä vahvasta multikollineaarisuudesta. Tarkan rajan asettaminen muuttujien multikollineaarisuudelle VIF:n perusteella ei ole yksiselitteistä, ja VIF:n arvoja on tarkasteltava itse kontekstin yhteydessä (O'brien, 2007). Valitsen testin ylärajaksi $VIF < 5$. Mikäli VIF-testin tulos on pienempi kuin viisi, oletus muuttujien välisestä multikollinearisuudesta hylätään. Toteutan VIF-testin regressoimalla pinta-alan logaritmin rakennusoikeuden logaritmilla. Rakennusoikeuden ja pinta-alan välinen VIF on 1,52, mikä tarkoittaa, että rakennusoikeus ja pinta-ala ovat heikosti kollineaarisia. Molempien muuttujien käyttäminen lineaarisessa mallissa ei multikollineaarisuuden kannalta tuota ongelmia.

Lineaarisen mallin oletusten kannalta tulee vielä tarkistaa mallin virheiden homoskedastisuus. Testaan virhetermien skedastisuutta Breusch ja Pagan (1979) kehittämän heteroskedastisuustestin avulla. Lineaarisen mallin heteroskedastisuustestin tulos osoittaa, että malli on vahvasti heteroskedastinen. Mallin tulosten raportoinnissa tulee käyttää heteroskedastisesti korjattuja keskivirheitä. Heteroskedastisesti korjattujen keskivirheiden menetelmän esitteli White (1980). Heteroskedastisesti korjatut keskivirheet eivät muuta mallin kertoimien estimaatteja, mutta muuttavat muuttujien kertoimien keskivirheitä, mikä vaikuttaa muuttujien tilastolliseen merkittävyyteen.

4.2 Regressiopuu

Lineaarisen regressiomallin lisäksi tonttien neliöhintoja mallinnetaan regressiopuu-mallilla. Regressiopuu on päätöspuumenetelmä, missä selitettävänä muuttujana on reaaliarvoinen muuttuja. Luokittelu- ja regressiopuu metodien varhaisimpia esittelijöitä olivat Breiman et al. (1984). Regressiopuumenetelmässä aineisto jaotellaan rekursiivisesti osiin selittäviin muuttujiin pohjautuvien ehtojen perusteella. Puu-mallissa ei oleteta selittävien ja selitettävän muuttujan välistä lineaarisuutta. Regressiopuu-malli on epälineaarinen luo-

kittelumalli.

Regressiopuun etuina lineaariseen malliin verrattuna on, että malli ei vaadi aineiston lineaarisointia tai muiden lineaarisen mallin oletusten täyttämistä mallin toiminnan kannalta. Regressiopuu pystyy jaottelemaan aineiston myös luokkamuuttujien, kuten kaupungin tai vuoden perusteella, joten indikaattorimuuttujien muodostaminen ei mallin ennusteiden kannalta ole tarpeen. Regressiopuun tulosten tulkinta on varsin yksinkertaista, sillä mallin jaotteluehdot kuvaavat suoraan, miten selittävät muuttujat vaikuttavat mallin arvioon tontin neliöhinnasta.

Regressiopuun haarakohtia, eli jaotteluehtoja, kutsutaan puun oksiksi. Puun ensimmäistä haarakohtaa kutsutaan juureksi ja puun oksien loppukohtia, eli mallin tuloksena jaettuja ryhmiä, kutsutaan lehdiksi. Puun syvyys kertoo, kuinka monta oksaa kannon lisäksi puussa on, eli kuinka monella ehdolla tutkimusaineisto jaotellaan. Esimerkiksi regressiopuumallilla, jonka syvyys on kolme, aineisto jaotellaan kolmen jaotteluehdon pohjalta neljään hintaryhmään. Esimerkin puu muodostuu kannosta, kahdesta oksasta ja neljästä lehdestä. Olkoon $\boldsymbol{A}$ tutkittava aineisto ja L lehtien lukumäärä puussa. Tutkittava aineisto jakautuu regressiopuun oksien mukaan L-määrään ryhmiä:

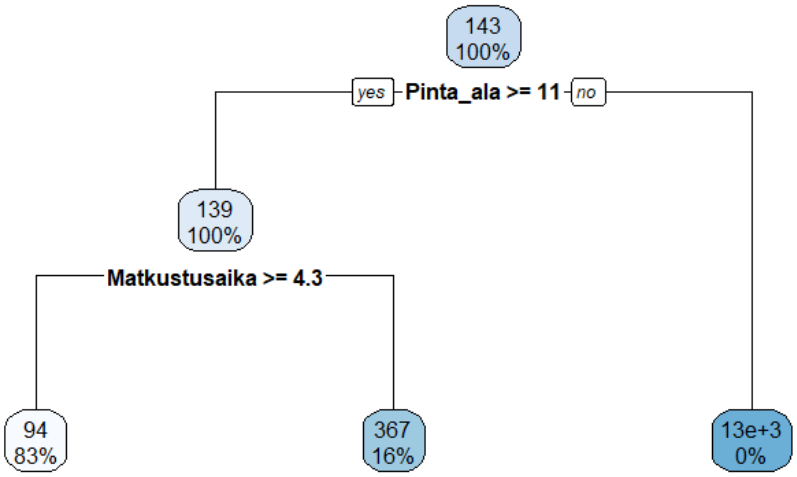
$$\boldsymbol{A} = (R_1, R_2, \ldots, R_L | \boldsymbol{X}),$$

missä R_1 on järjestyksessä ensimmäinen ryhmä ja $\boldsymbol{X}$ on selittävien muuttujien matriisi. Regressiopuu jakaa aineiston ryhmiin selittävien muuttujien perusteella. Mallin estimaatit neliöhinnalle ovat ryhmän sisäisten kauppojen neliöhintojen otoskeskiarvo. Regressiopuun estimaatit neliöhinnoille ovat siis kokoelma ryhmien neliöhintojen otoskeskiarvoja.

$$\hat{p}_l = \frac{1}{n_l} \sum_{i \in R_l} p_i$$

missä $\hat{p}_l$ on ryhmän R_l neliöhinnan keskiarvo ja estimaatti ryhmän tonttikaupoille.

Regrssiopuun kanto eli ensimmäinen luokitteluehto valitaan selittävien muuttujien joukosta siten, että mallin ennusteen ja aineiston neliöhintojen välisten virheiden neliöiden summa on pienin kaikkien muuttujien ja mahdollisten luokitteluehtojen joukosta. Luo-



Kuva 8: Regressiopuu, jonka syvyys on kaksi.

kitteluehdot valitaan siten, että malli istuu sovitettavaan aineistoon mahdollisen tarkasti. Luokitteluehto täyttää siis ehdon:

$$\min_R \{\sum_{i \in R_1} (p_i - \hat{p}_1)^2 + \sum_{i \in R_2} (p_i - \hat{p}_2)^2\}$$

Aineiston ensimmäinen haarakohta, eli kanto, jakaa aineiston kahteen osaan R_1 ja R_2. Kannosta voidaan kasvattaa oksia regressiopuulle. Oksien jaotteluehdot valitaan aineiston selittävien muuttujien joukosta samalla tavoin kuin kannon jaotteluehto, mutta sitä sovelletaan ryhmille R_1 ja R_2. Uusien oksien kasvattaminen puulle ei muuta aikaisemmin sovellettuja jaotteluehtoja. Puun kasvattamista jatketaan siitä haarasta, joka eniten vähentää puun ennusteiden virheiden neliöiden summaa, joka on:

$$SSR = \sum_{i \in R_1} (p_i - \hat{p}_1)^2 + \sum_{i \in R_2} (p_i - \hat{p}_2)^2 + \cdots + \sum_{i \in R_L} (p_i - \hat{p}_L)^2$$

Kuvassa 8 havainnollistan regressiopuuta, jonka syvyys on 2. Kuvan puussa aineiston neliöhinta on jaoteltu tontin pinta-alan ja matkustusajan mukaisesti kolmeen hintaryhmään. Ilman rajoitteita regressiopuuta voidaan kasvattaa rajattomasti niin, että se istuu täydellisesti aineistoon. Ilman rajoitteita mallintamisessa voidaan helposti sortua mallin ylisovittamiseen. Mallin ylisovittaminen tarkoittaa, että malli kuvaa liian tarkasti mallinnettavaa aineistoa ja siten heikentää mallin ennustustarkkuutta mallia sovitettaessa uuteen aineistoon. Regressiopuun kasvattamista on siis hyvä rajoittaa. Regressiopuun jaotteluehtoja voidaan määrittää siten, että mallin ylisovittaminen aineistoon vältetään. Puun kasvattamista voidaan rajoittaa niin, että puun lehtien tulee sisältää vähintää tietyn määrän tonttikauppoja. Tällä ehdolla vältytään siltä, että aineisto jaettaisiin yksittäisten kauppojen tarkkuudella. Toinen vaihtoehto ylisovittamisen välttämiseksi on asettaa mallille kokoa rajoittava parametri siten, että pienimpään neliösummaan lisätään puun lehtien määrä:

$$\min\{SSR + \alpha|L|\},$$

missä α on kasvunrajoitusparametri. Suurilla α:n arvoilla puun kasvua rajoitetaan vahvasti ja pienillä arvoilla puu saa kasvaa vapaammin. Liian lyhyellä puulla malli voi olla alisovitettu eli malli istuu huonosti sovitettavaan aineistoon sekä mallin ulkopuoleiseen aineistoon. Puuta kasvatettaessa vältellään molempia estimointivirheitä samalla pyrkien mahdollisimman tarkkaan malliin.

4.3 Gradient Boosting

Regressiopuun istuvuus aineistoon voi vaatia varsin suuren puun rakentamista, mikä heikentää puun kykyä arvioida tarkasti koulutusaineiston ulkopuolisten tonttikauppojen hintoja. Mallin tarkkuutta voidaan tehostaa koneoppimismetodilla nimeltä ”Gradient Boosting” tai ”Gradient Boosting Machine” (GBM). Yksi ensimmäisistä GBM-metodin esittelijöistä oli Friedman (2001). GBM-metodin periaatteena on, että yhden syvän regressiopuun sijasta aineistoa mallinnetaan usealla yksinkertaisella regressiopuulla.

GBM-malli perii regressiopuun suotuisat ominaisuudet verrattuna lineaariseen malliin. Nämä suotuisat ominaisuudet ovat, että malli ei vaadi oletusta selitettävän ja selittä-

vien muuttujien lineaarisuudesta ja tutkimusaineistolle ei ole välttämätöntä tehdä logaritmista muunnosta tai muodostaa luokkamuuttujille omia indikaattorimuuttujia. GBM-mallilla voidaan tehostaa regressiopuun ennustustarkkuutta. Mallin heikkous verrattuna regressiopuu- ja lineaariseen malliin on, että mallin tulokset eivät ole niin yksiselitteisiä ja helposti tulkittavia. Useissa käyttötarkoituksissa, kuten maapohjan arvostusmallina, GBM-mallin haitat jäävät varsin vähäisiksi verrattuna sen tuomiin hyötyihin.

4.3.1 GBM-mallin algoritmi

Friedmanin (2001) Gradient Boosting Machine toimii seuraavalla periaatteella. Olkoon $\Psi(y, f)$ mallin tappiofunktio. Tappiofunktioksi voidaan valita mikä tahansa differentioituva funktio, joka kuvaa tonttien hintojen ja mallin ennusteen välistä virhettä. Tämän tutkimuksen GBM-malleissa tappiofunktiona käytän pienimmän neliösumman menetelmää. Valitsen tappiofunktioksi funktion:

$$\Psi(y, f) = \sum_{i=1}^{n} (p_i - f(\boldsymbol{x}_i))^2,$$

joka vastaa menetelmää, jolla lineaarisen mallin muuttujien kertoimet estimoidaan.

GBM-menetelmän tavoitteena on estimoida funktio, joka pienentää tappiofunktion suuruutta. Tappiofunktion arvon pienentäminen, eli mallin istuvuuden parantaminen, saavutetaan kokoamalla malli useasta lyhyestä regressiopuusta, missä seuraava mallia täydentävä puu saa uuden painoarvon mallissa. Malliin lisättävät puut sovitetaan mallin ja aineiston väliselle residuaalille. GBM noudattaa seuraavaa algoritmia:

Annetaan mallille aluksi vakioarvoinen alkuarvo, joka on ratkaisuna tappiofunktion minimoimiseksi

$$F_0(\boldsymbol{x}) = \arg\min_{\rho} \sum_{i=1}^{N} \Psi(y_i, \rho) \tag{A0}$$

Iteroidaan algoritmia luvusta $m = 1$ valittuun määrään M asti:

1. Lasketaan algoritmin pseudo-residuaalit, joihin uusi puu sovitetaan

$$\hat{y}_i = -[\frac{\partial \Psi(y_i, F(\boldsymbol{x}_i))}{\partial F(\boldsymbol{x}_i)}]_{F(\boldsymbol{x})=F_{m-1}(\boldsymbol{x})}, \quad i = 1, \ldots, N \tag{A1}$$

2. Sovitetaan regressiopuu laskettuihin pseudo-residuaaleihin käyttäen koulutusaineistoa

$$\boldsymbol{a}_m = \arg\min_{\boldsymbol{a},\beta} \sum_{i=1}^{N} [\hat{y}_i - \beta h(\boldsymbol{x}_i; \boldsymbol{a})]^2 \tag{A2}$$

3. Lasketaan gradienttiaskeleen suuruus ratkaisemalla yksiulotteinen minimointiongelma

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(\boldsymbol{x}_i) + \rho h(\boldsymbol{x}_i; \boldsymbol{a}_m)) \tag{A3}$$

4. Päivitetään mallia uudella gradienttiaskeleella painotetulla regressiopuulla

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \rho_m h(\boldsymbol{x}; \boldsymbol{a}_m) \tag{A4}$$

Algoritmin lähtöaskeleena (A0) estimoidaan funktiolle vakioarvo, joka määrittää funktion lähtötason, jolle loput mallin lisäosat rakennetaan. Algoritmin askeleet 1—4 toistetaan valitulla määrällä M toistoja. Iteroinnin ensimmäisessä vaiheessa (A1) lasketaan algoritmille pseudo-residuaalit, joihin regressiopuu sovitetaan. Pseudo-residuaalit ovat mallin ennusteiden ja tonttikauppojen neliöhintojen välisten virheiden muunnos. Pseudo-residuaaleja käyttämällä varsin hankala funktion optimoimishaaste voidaan korvata pienimmän neliösumman minimoimisella. Kohdassa (A2) ratkaistaan pseudo-residuaalien ja regressiopuun välisen pienimmän neliösumman minimoimisongelma, minkä tuloksena saadaan puun parametrivektori $\boldsymbol{a}_m$. Regressiopuulle parametrit $\boldsymbol{a}_m$ ovat jaotteluehtojen muuttujat, muuttujien jaottelukohdat ja päätehaarojen otoskeskiarvot jokaiselle yksittäiselle puulle. Kohdassa (A3) lisättävälle regressiopuulle lasketaan vielä sopiva painotus, joka antaa tappiofunktiolle kaikista pienimmän arvon, kun kohdan (A2) estimoitu puu lisätään aiemmin estimoituun funktioon. Lopuksi kohdassa (A4) malli päivitetään lisäämällä siihen kohdan (A2) estimoidulla regressiopuulla painotettuna kohdan (A3) lasketulla gradienttiaskeleella.

Yksinkertaistaen GBM-malli voidaan esittää useamman regressiopuun summana:

$$F(x) = \sum_{m=1}^{M} f_m(\boldsymbol{x}) + c_0,$$

missä

$$f_m(\boldsymbol{x}) = \rho_m h(\boldsymbol{x}; \boldsymbol{a}_m).$$

GBM-mallissa $\boldsymbol{x}$ on mallin selittävät muuttujat, $\boldsymbol{a}_m$ on m:nen regressiopuun aineiston jaotteluehdot sekä päätehaarojen otoskeskiarvot, $h(\boldsymbol{x}; \boldsymbol{a}_m)$ on yksittäinen regressiopuu, ρ_m on yksittäiselle puulle valittu painoarvo mallissa ja c_0 on mallin alustava vakioarvoinen alkuarvo. GBM on siis kokoelma regressiopuita, jotka yhdessä määrittävät selittävien muuttujien pohjalta tontin neliöhintaa.

4.3.2 Virittäminen

GBM-malli estimoidaan R-ohjelmiston gbm-paketin funktiolla *gbm* (Greenwell et al., 2020). GBM-mallia on mahdollista virittää mallin toimintoihin vaikuttavilla viritysominaisuuksilla. Koska malli koostuu useammasta perusoppijasta, jotka tässä GBM-mallissa ovat lyhyitä regressiopuita, voidaan mallin ominaisuuksia säätää muuttamalla yksittäisten regressiopuiden pituutta ja lukumäärää mallissa. Kaksi oleellista ominaisuutta mallissa ovat puiden lukumäärä ja puiden syvyys. Mallin keskeinen viritysominaisuus ylisovittamisen ehkäisemiseksi on lehtien vähimmäiskoon asettaminen, eli kuinka monta havaintoa regressiopuun uloimmissa päätyhaaroissa tulee vähintään olla. Päätyhaaralle ei tehdä uutta jaottelua, jos tuloksena olevien haarojen havaintomäärä alittaa asetetun minimitason.

GBM-mallin gradienttiaskeleen toimintaa voidaan säätää mallin viritysparametreillä. Gradienttiaskeleen säätöparametrit ovat rajoitusparametri ja gradienttiaskeleen stokastisuusparametri. Rajoitusparametri nimensä mukaisesti rajoittaa malliin lisättävän puun gradienttiaskeleen suuruutta. Kun malliin lisäätään rajoitusparametri, algoritmin viimeisestä yhtälöstä (A4) tulee:

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \lambda \rho_m h(\boldsymbol{x}; \boldsymbol{a}_m),$$

missä λ on mallin rajoitusparametri, joka saa arvonsa puoliavoimelta väliltä $\lambda \in (0, 1]$. Rajoitusparametri määrittää mallin oppimisnopeuden, eli tappiofunktion arvojen pienentymisen tappiofunktiota pienentävän gradientin suuntaan. Rajoitusparametriarvon pienentäminen yleisesti vähentää mallin ennustamisvirhettä, mutta kasvattaa mallissa estimoitavien puiden lukumäärää.

Gradienttiaskeleen stokastisuus toteutetaan siten, että mallin estimointiin käytettävästä koulutusaineistosta valitaan satunnainen osajoukko. Parametriarvolla määritetään koulutusaineistosta otettavan osajoukon osuus koulutusaineistosta. Koulutusaineistosta otetulla osajoukko korvataan koulutusaineiston käyttö pseudo-residuaalien (A1) ja regressiopuun estimoimisessa (A2). Koulutusaineiston osajoukolla estimoidut puut muuttavat mallin tuottamia tuloksia verrattuna koulutusaineistolla estimoituun malliin. Stokastisen gradienttiaskeleen käyttö mallissa on kannattavaa silloin, kun koulutusaineistolla estimoitu malli ei tuota tappiofunktion pienintä mahdollista arvoa eli parhainta istuvuutta aineistoon. Stokastisella gradienttiaskeleella on tällöin mahdollista saada pienempi tappiofunktion arvo, mikä parantaa mallin hinta-arvioiden tarkkuutta.

Mallin virittämisen tavoitteena on löytää GBM-malli, joka tuottaa tarkimpia hinta-arvioita annettujen ominaisuuksien perusteella. Kuvaan mallin hinta-arvioiden tarkkuutta virheiden keskihajonnalla. Mallin estimointivirhe tonttikaupalle i on:

$$e_i = p_i - F(\boldsymbol{x}_i) \quad (i = 1, \ldots, n),$$

missä p_i on toteutuneen kaupan neliöhinta ja $F(\boldsymbol{x}_i)$ on mallin ennustama neliöhinta. Mallin estimointivirheiden keskihajonta on:

$$RMSE = \sqrt{\frac{e_1^2 + e_2^2 + \cdots + e_n^2}{n}}$$

Mittaan mallien estimointitarkkuutta $RMSE$-tunnusluvun pohjalta. Sopivien viritysarvojen löytämiseksi toteutan GBM-mallille ruudukkoetsinnän viiden eri virittämisparametrin muodostamilla kombinaatioilla. Valitsen GBM-mallien puiden maksimimääräksi 5 000 puuta per malli sekä puiden syvyydet joukosta $\{1;\ 3;\ 5\}$. Valitsen gradienttiaskeleen rajoitusparametrin joukosta $\{0,01;\ 0,1;\ 0,3\}$, havaintojen määrä lehdessä rajoitetaan arvoilla $\{5;\ 10;\ 15\}$, ja stokastisen gradienttiaskeleen säädän siten, että otan koulutusainestosta osajoukot kooltaan $\{0,65;\ 0,80;\ 1\}$. Suoritan ruudukkoetsinnän sopivimman virittämisparametri-yhdistelmän löytämiseksi. Sopivimmalla virittämisparametri-yhdistelmällä saadaan tarkimmat GBM-mallin hinta-arviot, mikä tarkoittaa pienintä $RMSE$:n arvoa. Taulukossa 3 esittelen kahdeksan pienintä $RMSE$-arvoa saanutta GBM-mallia. Taulokos-

ta 3 havaitsen, että puun syvyydellä 5 saadaan muita syvyyksiä paremmat tulokset. Taulukon 3 GBM-mallit on sovitettu logaritmoituun rekisteriaineistoon. Mallin logaritmiset hinta-arviot muutetaan takaisin normaalimuotoon, jota verrataan tontin kauppahintaan. Taulukon $RMSE$:iden yksikkö on euroa neliömetriltä.

Mallin nro	Grdientin rajoitus	Puiden syvyys	Lehden minimi-koko	Osuus aineis-tosta	Puiden lkm	RMSE
1	0,1	5	15	1	888	341,7886
2	0,3	5	15	1	468	347,9724
3	0,1	5	15	0,8	709	358,9121
4	0,01	5	10	0,65	4 947	363,359
5	0,1	3	15	0,8	1 385	363,6782
6	0,01	5	15	1	4 999	366,5587
7	0,01	5	10	0,8	4 245	370,8766
8	0,01	5	10	1	4 999	371,6267

Taulukko 3: GBM-mallin ensimmäisen ruudukkoetsinnän tarkimpien mallien parametriarvot.

Ensimmäisen kierroksen ruudukkoetsinnän tuloksista voidaan selvästi havaita, että syvemmät puut tuovat tarkemman estimointituloksen kuin lyhyemmät ja lehtien vähittäiskoko on hyvä rajata vähintään kymmeneen havaintoon. Tehdään toisen kierroksen ruudukkoetsintä lisäämällä puiden syvyyttä. Toisen kierroksessa mallien puun syvyydeksi valitaan $\{5;\ 7;\ 9\}$. Taulukosta 4 voidaan havaita, että syvemmät puut tuottavat tarkemmat ennusteet kuin muut GBM-mallit matalimmilla puun syvyyksillä. Ruudukkoetsinnän pohjalta GBM-malli on hyvä virittää siten, että puiden syvyys on 7, gradienttiaskelta rajoitetaan parametriarvolla 0,1, lehden minimikoko asetetaan 15 havaintoon ja mallin estimoinnissa käytetään koko koulutusaineistoa. Näyttäisi myös siltä, että parhain tulos saadaan, kun puiden lukumäärä mallissa on vähemmän kuin 2 500.

Mallin virittämisen kannalta on tärkeää vielä selvittää, että mallin suorituskyky ei ole ainoastaan satunnaisesti valittuun koulutusaineistoon sidonnainen, vaan pärjää myös muilla koulutusaineiston jaotteluilla. Testataan kymmenellä eri koulutus- ja testiaineiston jaottelulla mallin virittämisen toimintoja. Kokeilen mallin asetuksina puiden lukumäärää 2 500 kpl, rajoitusparametrejä joukosta $\{0,05;\ 0,1\}$, syvyyksiä joukosta $\{5;\ 7\}$ ja pienimmällä määrällä havaintoja lehdessä arvoilla $\{10;\ 15\}$. Selvitän, vaikuttaako koulutusaineiston

Mallin nro	Grdientin rajoitus	Puiden syvyys	Lehden minimi-koko	Osuus aineis-tosta	Puiden lkm	RMSE
1	0,1	7	15	1	1 011	333,6487
2	0,05	7	15	1	2 233	337,6478
3	0,1	9	15	1	1 048	339,4073
4	0,1	5	15	1	888	341,7886
5	0,05	9	15	1	2 257	343,2244
6	0,05	5	15	1	2 375	350,1414
7	0,1	5	15	0,8	709	358,9121
8	0,05	7	10	0,8	598	362,6166

Taulukko 4: GBM-mallin toisen ruudukkoetsinnän tarkimpien mallien parametriarvot.

valinta samalla tavoin kaikkiin malleihin. Toimivin malli on se, jonka $RMSE$-keskiarvo ja $RMSE$-keskihajonta ovat pienimmät kaikissa kymmenessä aineiston satunnaisessa jaottelussa. Esittelen taulukossa 5 mallien $RMSE$:n keskiarvot ja keskihajonnat aineiston jaottelun pohjalta. Taulukon 5 tulosten perusteella GBM-mallin paras suorituskyky kymmenellä eri aineistonjaolla saadaan, kun malli viritetään siten, että puun syvyys on 7, rajoitusparametrin arvo on 0,05 ja lehden pienin koko on 10. Mallissa käytettävien viritämisarvojen tulokset muuttuvat hieman, kun GBM-malli sovitetaan uuteen koulutusaineistoon.

Mallin nro	Grdientin rajoitus	Puiden syvyys	Lehden minimi-koko	RMSE keskiar-vo	RMSE keskiha-jonta
1	0,05	7	10	556	296
2	0,1	5	15	557	301
3	0,1	7	15	558	300
4	0,05	7	15	562	302
5	0,05	5	15	568	322
6	0,1	7	10	569	322
7	0,05	5	10	575	295
8	0,1	5	10	577	317

Taulukko 5: GBM-mallin tarkimmat parametriarvot 10 satunnaiselle koulutusaineistojaolle.

5 Tulokset

Esittelen tulokset kahdessa osassa. Ensimmäisessä osassa tutkin lineaarisen mallin ja GBM-mallin selittävien muuttujien vaikutusta tonttien neliöhintoihin. Sovitan molemmat mallit logaritmoituun rekisteriaineistoon ja käytän mallien kertoimien estimoinnissa koko rekisteriaineistoa. Toisessa osassa keskityn mallien ennustetarkkuuden tutkimiseen ristiinvalidoinnilla. Toteutan ristiinvalidoinnin jakamalla tutkimusaineiston koulutus- ja testiaineistoon ja lasken testiaineistolle hinta-arviot koulutusaineistolla estimoiduilla malleilla. Osiossa selvitän mallien tarkkuutta usealla eri koulutusaineiston jaottelulla. Toisessa osiossa selvitetän, mikä aikaisemmassa kappaleessa esitetyistä malleista kykenee tarkimmin arvioimaan tonttien arvoja ja kykeneekö malli tekemään sen johdonmukaisesti riippumatta aineistonjaosta.

5.1 Lineaarisen mallin tulokset

Taulukossa 6 esittelen lineaarisen regressiomallin estimaatit indikaattorimuuttujille, pinta-alalle, rakennusoikeudelle, matka-ajalle ja ruututietokannan muuttujille. Logaritmisen arvostusfunktion parametrien estimaattien tulokset löytyvät liitteestä A. Indikaattorimuuttujien tuloksista voidaan vetää karkea yhteenveto ajan ja toimitilakäytön yleisestä vaikutuksesta mallin antamiin hinta-arvioihin. Vuoteen 2015 verrattuna neliöhinnat eivät kasva merkittävästi vuosina 2016 ja 2017. Vuotena 2018 on havaittavissa 4,6 prosentin kasvua neliöhinnassa ja vuotena 2019 neliöhinnassa on havaittavissa 3,6 prosentin eroa verrattuna vuoteen 2015.

Liike- ja toimistoindikaattorit käsittävät tontit, joilla on kyseisessä käytössä oleva rakennus. Toimitilatontti-indikaattori sisältää tontit, joille ei ole vielä rakennettu, mutta jotka on kaavoitettu toimitilakäyttöön. Keskimäärin tonttien, joissa on liikerakennus, mallin logaritminen hinta on 0,82 enemmän kuin asuintonteilla. Toimistotonteilla vastaava indikaattorimuuttujan arvo on 0,72. Kun logaritminen hinta muunnetaan euroa neliömetriltä -hinnaksi, liike- ja toimistotonttien hinnat ovat kaksinkertaisia vastaavaan rakentamattomaan asuintonttiin verrattuna. Rakennetun alueen tontit ovat keskimääräisesti arvokkaampia kuin tontit, joille ei vielä ole rakennettu mitään. Mallissa ei ole muuttujia

kuvaamaan maapohjan laatua, mutta olemassa olevasta rakennuksesta voi päätellä, että maapohja on ollut rakennettavissa tai sitä on muokattu siten, että sille voidaan rakentaa. Tontin laatu on yksi tontin hintaan vaikuttavista tekijöistä rakennustoiminnan kannalta, sillä heikko laatu lisää rakentamisen kustannuksia maapohjan muokkauksen myötä. Liike- ja toimistoindikaattorin implikoima hinnan nousu tontissa ei siis hyvin luultavasti ole käyttötarkoituksen ansiota vaan tontin laadun ansiota. Toimitilatonteille, jotka eivät sisällä rakennusta, keskimääräinen hinnan lisäys mallissa on noin 23 prosenttia verrattuna vastaavaan tonttiin eri käyttötarkoituksella. Indikaattorimuuttuja antaa kuitenkin vain karkean arvion käyttötarkoituksen vaikutuksesta eikä kerro tarkempaa dynamiikkaa, kuinka toimitilakäyttö vaikuttaa hinnan kehittymiseen eri kaupungeissa ja ajankohtana.

	β	keskivirhe	muuttuja
β_0	7,719103***	0,099681	
2016	-0,003863	0,015522	I
2017	0,013773	0,015348	I
2018	0,045626**	0,015759	I
2019	0,036157*	0,016762	I
Liike	0,827184***	0,039775	I
Toimisto	0,724513***	0,072894	I
Toimitilatontti	0,236926***	0,049718	I
Pinta-ala	-0,532951***	0,007779	ln(x)
Rakennusoikeus	0,0829***	0,002356	ln(x)
Matka-aika	-0,108776***	0,008405	ln(x)
Työttömät	-0,037233***	0,010695	ln(r)
Lapset	-0,156753***	0,008595	ln(r)
Korkeakoulutetut	0,1929***	0,008904	ln(r)
Omistusasunnot	-0,200806***	0,013462	ln(r)
Vuokra-asunnot	0,040557***	0,007301	ln(r)
Muut asunnot	0,08721***	0,013012	ln(r)
Taloudet yhteensä	0,306364***	0,014161	ln(r)
Asumisväljyys	-0,054889***	0,010877	ln(r)
Havaintoja	21 978		
R^2	0,816		
Korjattu R^2	0,816		
Residuaalien keskivirhe	0,724 (va = 21 914)		
F-tunnusluku	1 543,890*** (va = 63; 21 914)		
Note:	*p<0,1; **p<0,05; ***p<0,01		

Taulukko 6: Lineaarisen mallin tulokset

Mallin mukaan pinta-alan vaikutus tontin neliöhintaan on vähenevä pinta-alan kasvaessa. Maapohjan yksikköhinta vähenee pinta-alan lisääntyessä. Mallin ennuste hinnan ja määrän suhteelle on, että yksikköhinta vähenee 0,5 prosenttia, kun määrää lisätään 1 prosentti. Syynä voi myös olla, että tiheämmin rakennetuilla alueilla ei ole suuria vapaita tontteja markkinoilla tarjolla, kun taas kauempana kaupungin keskustoista on suurempia tonttialueita saatavilla. Mallissa rakennusoikeudella on tontin hintaa lisäävä vaikutus. Rakennusoikeus määrittää, kuinka suuren rakennuksen tontille voi rakentaa, mikä suoraan viestittää tontin rakennettavuudesta. Keskimäärin rakennusoikeuden lisääminen 1 prosentilla lisää tontin hintaa 0,08 prosentilla. Mallin arvio matkustusajan vaikutuksesta vastaa oletusta, että matka-ajan kasvaessa lähimmälle keskusta- tai kauppapaikalle, tontin hinta alenee. Matkustusajan kasvaessa yhdellä prosentilla tontin hinta pienenee keskimäärin prosentin kymmenyksellä.

Ruututietokannanmuuttujat otetaan lineaariseen malliin mukaan logaritmisina. Muuttujien määrää mallissa rajaavat pois muuttujien tilastollinen merkittävyys sekä ruututietokannan muuttujien välinen multikollineaarisuus. Esimerkiksi talouksien lukumäärä on vahvasti kollineaarinen asukasmäärän kanssa. Merkittävimmät neliöhintaan vaikuttavat muuttujat ovat talouksien, korkeakoulutettujen ja omistusasuntojen logaritminen lukumäärä. Talouksien lukumäärä näyttää korreloivan positiivisesti tonttien neliöhintojen kanssa. Tämä mahdollisesti indikoi, että alueella, jossa on jo useita talouksia, vapaita tontteja on vähemmän tarjolla sekä vapaiden tonttien kysyntä on suurempaa. Lisäksi tiheämmin asutetut alueet rakentuvat hyvien palveluiden lähelle tai tiheästi asutetut alueet voivat houkutella palveluita lähelle, jos palveluita ei ole ennestään ollut olemassa alueen läheisyydessä. Mallin antama tulos lähialueen omistusasumiselle on, että omistusasuntojen lukumäärä vähentää alueen neliöhintoja. Tähän syynä voi olla se, että valtaosa omistusasunnoista sijoittuu harvemmin asutetuille aluille ja vuokra-asunnot keskittyvät enemmän tiheämmin asutuille kaupunkialueille.

Lineaarisen mallin estimoimat arvostusfunktion parametriarvot vastaavat oletusta siitä, miten tonttien hinnat vaihtelevat kaupungeittain ja miten tontin etäisyys vaikuttaa tontin hintaan. Kaupungin ja etäisyyden yhteisvaikutukset ovat myös arvostusfunktion asettamien ehtojen rajoissa. Odotetusti Helsingin keskusta on arvokkainta aluetta. Mui-

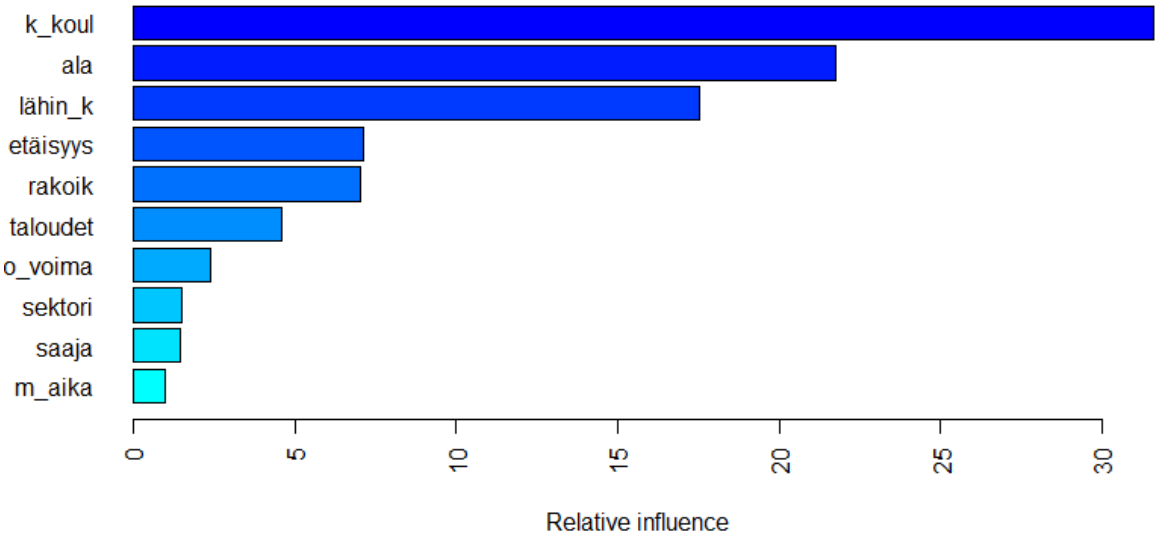
ta arvokkaita kaupunkialueita on mm. Tampere, Oulu, Turku sekä Hyvinkää. Keskustan etäisyyden vaikutus tontin neliöhinnan alenemiseen on Helsingissä muita kaupunkeja suurempaa. Lineaarisessa mallissa on kaupungille ja etäisyydelle annettu omat hinta-arviot, jos ne poikkeavat merkittävästi mallin lähtötasosta. Etäisyyden vaikutuksen lähtötaso määräytyy 12 eri kaupungin mukaan, joilla on varsin samanlainen etäisyyden vaikutus neliöhintaan. Keskustan hinta-arvion pohjatason muodostaa 8 kaupunkia, joiden tontin hinta-arvio kaupungin keskustassa on samansuuruinen.

5.2 GBM-mallin tulokset

Saan GBM-mallin virittämisen tuloksena, että logaritmoidulle rekisteriaineistolle sovitettu malli viritetään puiden syvyydellä 7, lukumäärällä 2500 kpl, havaintojen minimimäärällä 10 ja rajoittamisparametrin arvolla 0,05. Tässä osiossa kuvaan, miten selittävät muuttujat vaikuttavat tontin neliöhinnan muodostumisessa GBM-mallissa. Raportoin mallin tulokset GBM-mallille, joka on viritetty edellä mainituin parametriarvoin ja on sovitettu logaritmiseen rekisteriaineistoon. Tämän osion tuloksissa olen käyttänyt koko aineistoa kuten lineaarisen mallin tulosten raportoinnissa, jotta selittävien muuttujien vaikutukset olisivat vertailukelpoisia lineaarisen mallin tulosten kanssa.

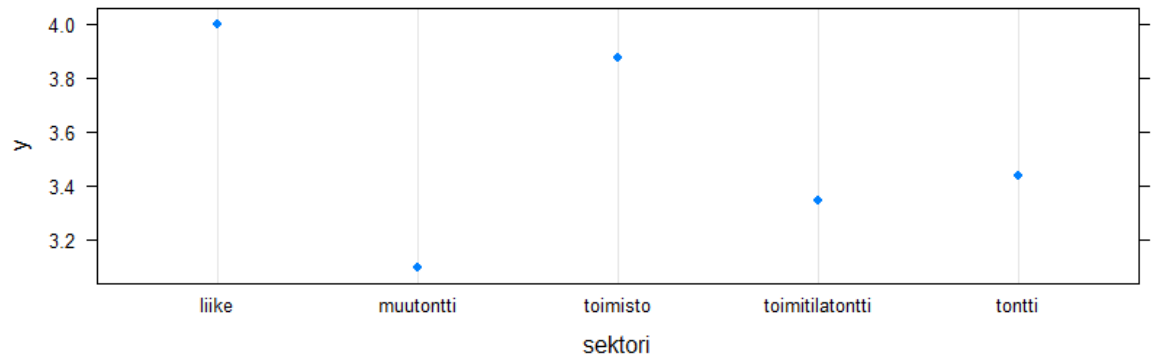
Kuvassa 9 esitän mallin muuttujien suhteellisen vaikutuksen ja keskinäisen tärkeysjärjestyksen neliöhinnan mallintamisessa. GBM-mallin muuttujien suhteellisen vaikutuksen kuvauksen kehitti Friedman (2001). Muuttujien suhteellinen vaikutus mittaa, kuinka paljon muuttuja vähentää mallin ennusteen keskivirhettä suhteutettuna muihin muuttujiin. Mallin keskivirhettä eniten vähentävät, eli vaikuttavimmat, muuttujat ovat korkeasti koulutettujen määrä, tontin pinta-ala, lähin kaupunki, etäisyys keskustasta ja rakennusoikeus. Varsin yllättävä tulos on, että korkeasti koulutettujen määrä vaikuttaa lähes kolmanneksella mallin estimointitarkkuuteen. Korkeasti koulutettujen määrän käyttäminen regressiopuiden jaotteluehtoina vastaa siis mallin ennustusvirheen suuruuden pienentämisestä kolmasosalla.

Kuvassa 10 esitän toimitilatonttien keskimääräisen vaikutuksen mallin estimaatteihin. Kuvassa muuttuja tontti sisältää kaikki rakennuskäyttöön tarkoitetut tontit ja muu tont-



Kuva 9: Suhteellinen vaikutus

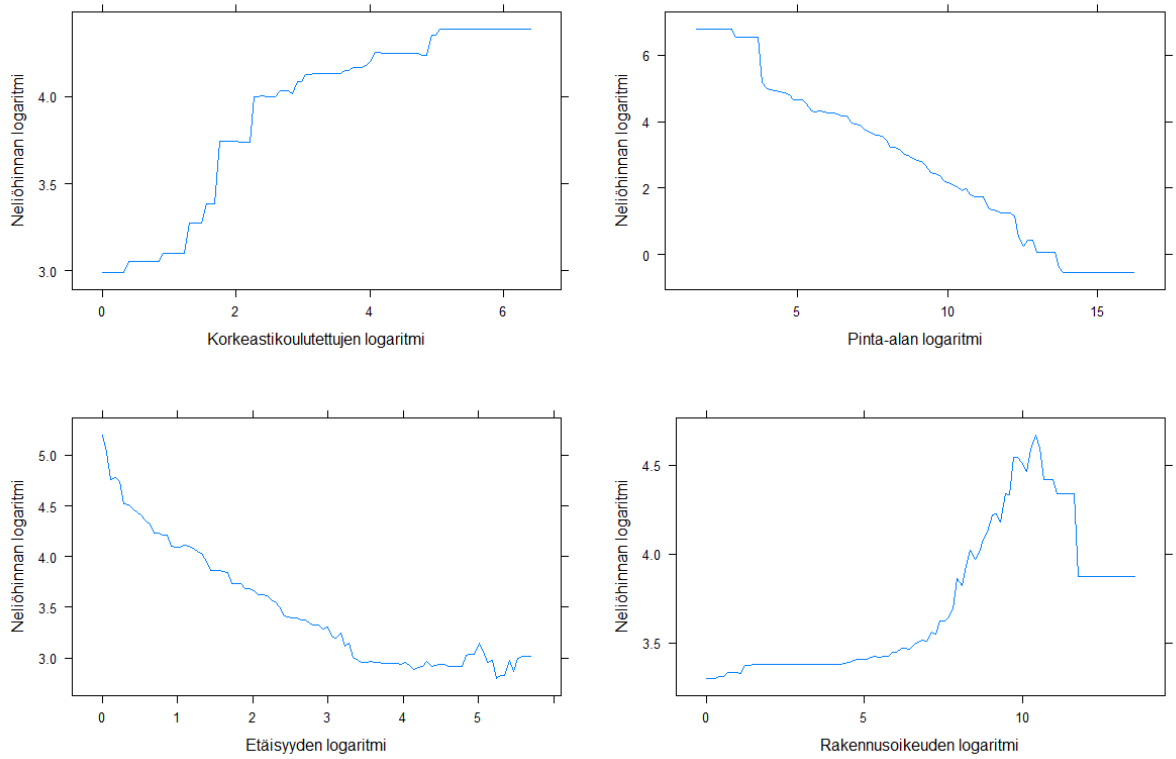
ti sisältää tontit, jotka on tarkoitettu muuhun rakennuskäyttöön. Mallin tuloksena vapaat toimitilatontit eivät merkittävästi poikkea muista rakennettavista tonteista. GBM-mallin liike- ja toimistotontit, jotka sisältävät rakennuksia, saavat samansuuntaisia tuloksia kuin lineaarisessa mallissa. Tontin käyttötarkoitus vastaa mallissa 1,5 prosentin mallin keskivirheen pienentämisestä ja on kahdeksanneksi merkittävin muuttuja. Käyttötarkoituksella on siis merkitystä mallin kannalta, vaikka se ei ole kovinkaan suuri.



Kuva 10: Tontin käyttötarkoituksen vaikutus GBM-mallissa

Havainnollistan GBM-mallin selittävien muuttujien vaikutusta neliöhintaan osittaisriippuvuuskuvaajalla. Osittaisriippuvuuskuvaaja esittää, kuinka selittävä muuttuja keskimää-

räisesti vaikuttaa logaritmisen neliöhinnan muutokseen, kun muiden muuttujien vaikutukset pidetään vakiona. Kuvassa 11 esitän mallin selitysasteeltaan tärkeimpien muuttujien suhdetta tonttien neliöhintoihin. GBM-mallissa selittävät muuttujat vaikuttavat samansuuntaisesti kuin lineaarisessa mallissakin. Mallin etuna on, ettei selittävien muuttujien vaikutus selitettävään muuttujaan ole lineaarisesti sidottu. Osittaisriippuvuuskuvaajat havainnollistavat GBM-mallin epälineaarisuutta. Tonttien neliöhintojen muodostumisessa on paljon yksilöllisiä eroja, jota yleistävä lineaarinen malli ei kykene ottamaan huomioon. GBM-malli asettuu lineaarista mallia paremmin aineistossa esiintyviin yksilöllisiin eroihin.



Kuva 11: GBM-mallin muuttujien marginaalikuvio

Mallin selitysaste R^2 on 0,939 ja residuaalien keskivirhe on 0,416. Nämä tunnusluvut vastaavat lineaarisen mallille ilmoitettuja tunnuslukuja. Ne eivät kuitenkaan kerro, kuinka hyvin mallilla voidaan arvioida aineiston ulkopuolisia tonttikauppoja. Periaatteessa GBM-mallin residuaalien keskivirhe voidaan saada hyvin lähelle nollaa ja selitysaste lähelle numeroa yksi, sillä GBM-malli voidaan halutessa sovittaa miltei täydellisesti aineistoon. Aineistoon täysin sovitettu malli ei kuitenkaan ole kovin hyvä aineiston ulkopuo-

listen tonttikauppojen hintojen selittämiseen. Seuraavassa kappaleessa selvitän aiemmin esiteltyjen mallien kykyä arvioida mallin kehittämiseen käytetyn aineiston ulkopuolisten tonttien neliöhintoja.

5.3 Mallien ristiinvalidointi

Mallin käyttökelpoisuutta tonttien hintojen arvioimisessa voidaan testata ristiinvalidoinnilla, kuten Schulz et al. (2014) ja Kok et al. (2017) tekevät heidän tutkimuksissaan. Toteutan ristiinvalidoinnin siten, että jaan tutkimusaineistosta sattumanvaraisesti 70 prosenttia koulutusaineistoon ja loput 30 prosenttia testiaineistoon. Käytän testiaineistoa koulutusaineistolla estimoidun mallin tarkkuuden testaamiseen. Mallin antamat hinta-arviot testiaineiston tonteille ja testinaineiston toteutuneiden kauppahintojen välinen ero kertoo, kuinka paljon mallin antamat hinta-arviot poikkeavat toteutuneista hinnoista. Ristiinvalidoinnilla testaan mallin hinta-arvioiden tarkkuutta mallin estimoinnissa käytettyjen tonttikauppojen ulkopuolisten tonttikauppojen hintojen arvioimiseen. Ristiinvalidoinnin tulokset ovat oleellista tietoa mallin käyttämisessä KHR:n ulkopuolisten tonttien hintojen arvioimisessa ja hinta-arvioiden mahdollisesta tarkkuudesta. Jos mallin hinta-arviot ovat hyvin poikkeavia testiaineiston kauppahintojen kanssa, asettaa se mallin käyttökelpoisuuden KHR:n ulkopuolisten tonttien hintojen arvioimisessa kyseenalaiseksi.

Tutkimusaineisto voidaan jakaa lukuisilla eri tavalla koulutus ja tutkimusaineistoon. Erilaisia kombinaatioita satunnaiselle aineistojaolle on yhteensä noin 10^{5834} kappaletta, kun jaettavan aineiston koko on 22 000 havaintoa. Toteutan ristiinvalidoinnin 100 sattumanvaraisella aineistojaolla lineaariselle-, regressiopuu- ja GBM-mallille. Selvitän useammalla eri aineistojaon ristiinvalidoinnilla mallien estimointitarkkuuden yhteneväisyyttä aineistojaosta riippumatta. Mallien estimointitarkkuuden ristiinvalidointi useamman sattumanvaraisen aineistojaon estimaateille kertoo, onko tarkkuus riippuvainen jaetusta koulutus- ja testiaineistosta.

Mallien testiaineiston ennustetarkkuus antaa kuvan siitä, miten aineisto kykenee mallintamaan tonttien hintoja, joita ei ole käytetty mallin estimoinnissa. Estimointitarkkuus mallin ulkopuoleiseen aineistoon kertoo mallin yleistettävyydestä tonttien hintojen ar-

vioimiseen, joilla ei ole käyty kauppaa havaintoperiodin aikana. Estimoinnin tarkkuutta käytetään mittarina mallin kyvystä tehdä täsmällisiä hinta-arvioita tonttikannalle. Koulutusaineiston ulkopuolisten tonttikauppojen hinta-arvioiden tarkkuus on koko tonttikannan hinta-arvioinnissa tärkeä, sillä tonttikannan tonteista käydään kauppaa vain pienellä osalla koko havaintoperiodin aikana. Mallien estimoinnin tarkkuutta ristiinvalidoinnissa käytetään $RMSE$ -tunnuslukua, joka kertoo mallin hinta-arvion ja toteutuneen kauppahinnan välisen virheen keskihajonnan.

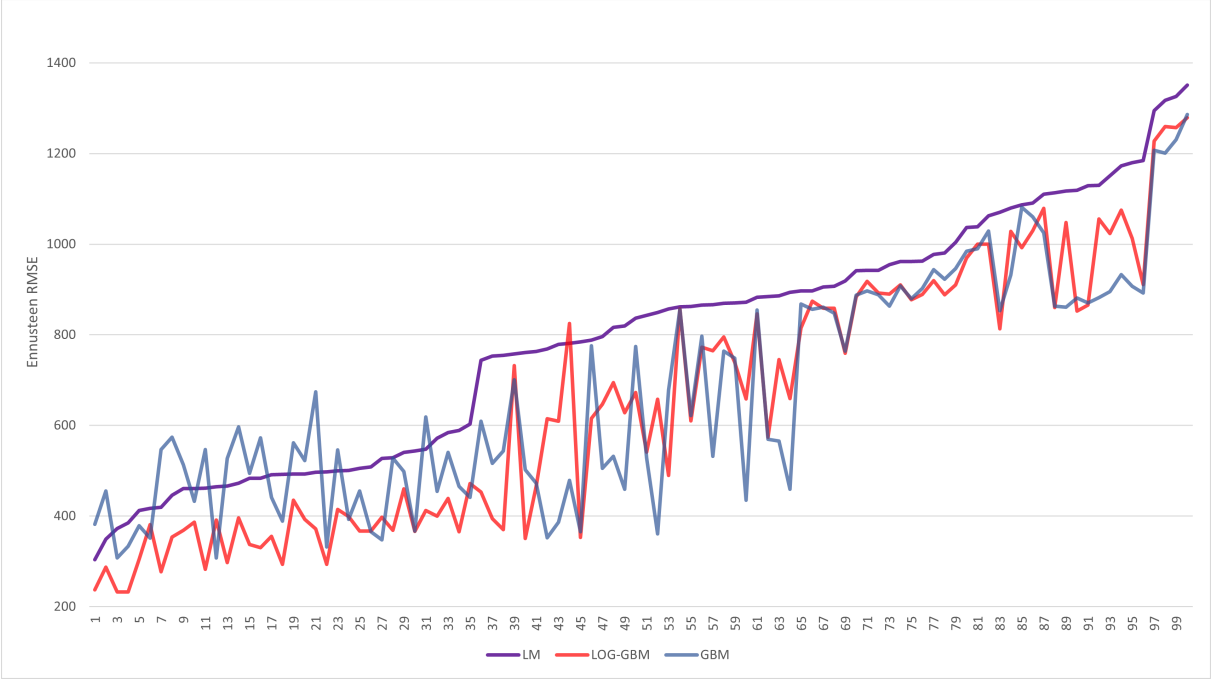
	LM	LOG-GBM	GBM	PUU	LOG-PUU
keskiarvo	790,41	653,78	670,99	812,74	830,67
keskihajonta	263,03	283,97	245,44	222,15	275,50
minimi	303,83	232,62	307,35	479,65	374,28
maksimi	1 350,66	1 278,85	1 285,80	1 329,61	1 384,84

Taulukko 7: Ennusteiden RMSE:den tunnusluvut 100 satunnaiselle koulutus- ja testiaineiston jaolle.

Esitän taulukossa 7 lineaarisen mallin (LM), logaritmoidulla aineistolla estimoidun GBM-mallin (LOG-GBM), rekisteriaineistoon sovitetun GBM-mallin (GBM), regressiopuu-mallin (PUU) ja logaritmoituun aineistoon sovitetun yksittäisen puu-mallin (LOG-PUU) ristiinvalidoinnin tulokset. Taulukossa esitän mallien ennusteiden $RMSE$:n keskiarvot, keskihajonnat sekä minimi- ja maksimiarvot 100 eri aineistojaolle. Taulukon keskiarvo kertoo, mikä on mallin keskimääräinen virheen keskihajonta, esimerkiksi lineaarisen mallin virheiden keskihajonnan keskiarvo 100 satunnaiselle aineistojaolle on 790,41 euroa neliömetriltä. Taulukon keskihajonta kertoo, mikä on mallien virheen keskihajonnan keskihajonta aineistojakojen välillä. Taulukon keskihajonnan arvot kertovat mallien virheiden keskihajontojen vaihtelevuutta aineistojakojen välillä. Taulukon minimi- ja maksimiarvot kertovat mallien pienimmän ja suurimman virheen keskihajonnan. Esimerkiksi LOG-GBM mallilla suurin virheen keskihajonta on 1278,85 euroa neliömetriltä ja pienin virheen keskihajonta on 232,62 euroa neliömetriltä.

Mallien virheiden hajonnasta aineistojakojen välillä havaitsen kolme tulosta. Ensimmäiseksi GBM-mallit ovat keskimäärin tarkempia kuin lineaarinen malli tai yksittäiset puumallit riippumatta aineiston jaosta. Toiseksi testiaineistojen välillä on suuria eroja siinä, kuinka hyvin malli istuu koulutusaineiston ulkopuoleiseen aineistoon. Tulos osoittaa, että

mallien hinta-arvioiden tarkkuus on riippuvainen mallin estimointiin käytetystä koulutusaineistosta ja ristiinvalidointiin käytetystä testiaineistosta. Kolmanneksi mallien estimointitarkkuus on varsin heikko virheen keskihajonnan perusteella. Aineiston keskiarvo tontin neliöhinnalle on 142 euroa neliömetriltä. LOG-GBM-mallin tarkimmalla aineistonjaolla virheiden keskihajonta on 1,6-kertainen neliöhinnan keskiarvoon verrattuna ja vastaavasti epätarkimmalla aineistojaolla virheen keskihajonta on yhdeksänkertainen.



Kuva 12: Mallien RMSE otoksittain

Mallien virheiden keskihajonnoissa on paljon vaihtelua aineiston jaosta riippuen. Vastaan seuraavaksi kysymykseen, vaikuttaako aineistonjako kaikkiin malleihin samalla tavoin, eli onko mallit epätarkkoja samalla aineistojaolla vai onko mallien välillä poikkeuksia. Kuvassa 12 esitän lineaarisen-, logaritmisen GBM-, ja GBM-mallin virheiden keskihajonnat 100 eri aineistojaolla siten, että ne on järjestetty lineaarisen mallin virheiden keskihajonnan mukaan pienimmästä suurimpaan. Kaikki kolme mallia on estimoitu samalla koulutusaineistolla ja ristiinvalidoitu vastaavalla testiaineistolla, ja siten mallien virheiden keskihajonnat ovat vertailukelpoisia toisiinsa nähden. Kuvasta havaitsen, että LOG-GBM- ja lineaarisen mallin ennusteiden virheiden keskihajontojen vaihtelevuudet ovat yhteneviä eri aineistojaoilla. LOG-GBM-malli on kaikissa paitsi yhdessä tapauksessa tarkempi kuin lineaarinen malli. GBM-malli näyttää tuottavan epätarkempia tuloksia kuin LOG-GBM-

ja lineaarinen malli silloin kun lineaarisen mallin $RMSE < 600$ euroa neliömetriltä. Toisaalta GBM-malli on keskimääräisesti tarkempi kuin LOG-GBM- ja lineaarinen malli, kun lineaarisen mallin $RMSE > 600$ euroa neliömetriltä. LOG-GBM-malli on GBM-mallia tarkempi 60 tapauksessa 100:sta ja lineaarista mallia tarkempi 99 tapauksessa 100:sta. Kuva 12 osoittaa taulukon 7 kanssa, että LOG-GBM-malli on keskimääräisesti tarkempi 100 satunnaisella aineistonjaolla kuin muut mallit. Kuvasta 12 tulkitsen, että mallien ristiinvalidoinnin tarkkuuksiin vaikuttaa koulutus- ja testiaineiston jaottelu huolimatta siitä, että mallien tarkkuudet eroavat samalla aineistojaolla.

5.4 Poikkeavat havainnot

Koska aineiston satunnaisella jaolla on merkittävä vaikutus mallien ristiinvalidoinnin tarkkuuteen, eli mallien kykyyn arvioida tonttien hintoja lähelle toteutunutta kauppahintaa, selvitän, mitkä tekijät vaikuttavat tarkkuuden vaihtelevuuteen. Jatkossa käsittelen yksinomaan LOG-GBM-mallia, koska toteutetun ristiinvalidoinnin perusteella se tuottaa keskimäärin muita malleja tarkemmat hinta-arviot. Vastedes kun käytän termiä ”malli”, tarkoittaa se LOG-GBM-mallia, ja termi ”mallit” tarkoittavat eri koulutusaineistoilla estimoituja LOG-GBM-malleja.

Mallin tarkkuuden vaihtelevuudelle on ainakin kaksi syytä. Koulutusaineiston pohjalta estimoitujen mallien erilaisuus tuottaa vaihtelevat ristiinvalidoinnin tulokset, tai aineisto sisältää havaintoja, joiden hintoja malli ei pysty arvioimaan tarkasti. Testaan oletusta mallin vaihtelevuudesta siten, että estimoin mallin tarkimman aineistojaon koulutusaineistolla ja testaan mallin tarkkuutta epätarkimman aineistojaon testiaineistoon. Toteutan saman myös toisinpäin, eli testaan epätarkimman aineistojaon koulutusaineistolla estimoitua mallia tarkimman aineistojaon testiaineistolla. Tarkimmalla aineistolla koulutetun mallin virheiden keskihajonta on 935,39 euroa neliömetriltä sovitettuna epätarkimman aineistojaon testiaineistoon. Epätarkimmalla aineistolla koulutetun mallin virheiden keskihajonta on 230,70 euroa neliömetriltä sovitettuna tarkimman aineistojaon testiaineistoon. Näistä tuloksista havaitsen, ettei mallien selittävien muuttujien pohjalta tehtyihin neliöhinnan arvioihin ole merkittävää eroa mallien tarkkuuden vaihtelevuuteen aineistojaon suhteen. Tulosten parempi tarkkuus johtuu siitä, että osa uudesta testiaineistosta on

käytetty mallin kouluttamiseen.

Mitä luultavimmin mallin ristiinvalidointien virheiden keskihajontojen vaihtelu johtuu koulutus- ja testiaineiston satunnaisesta jaosta. Tutkimusaineistolle ei ole suoritettu poikkeavien havaintojen karsintaa kappaleen kolme aineiston rajauksen lisäksi. Tutkimusaineistossa voi olla tonttikauppoja, joiden selittävien muuttujien arvot ja neliöhinta poikkeavat huomattavasti muusta aineistosta. Selvitän, onko tutkimusaineistossa tonttikauppoja, joille LOG-GBM-malli antaa jokaisella ristiinvalidoinnilla merkittävästi suurempia tai merkittävästi pienempiä hinta-arvioita. Toteutan poikkeavien havaintojen tunnistamisen seuraavalla tavalla. Ensiksi yhdistän kaikki 100 ristiinvalidoinnissa käytetyt testiaineistot yhdeksi 659 400 havaintoa kattavaksi aineistoksi. Seuraavaksi yhdistän tonttikaupat samalla kiinteistötunnuksella ryhmiksi, joiden lukumäärä on 20 149. Yksittäisiä kiinteistötunnuksia on alkuperäisessä tutkimusaineistossa yhteensä 20 149 kappaletta, joten kaikista tonteista on päätynyt havaintoja testiaineistoon. Pienimmän ryhmän koko on 14 havaintoa, eli alimmillaan tonttikauppa on päätynyt 14 eri testiaineistoon. Jokaisesta tonttikaupasta on riittävästi edustusta testiaineistoissa.

Tunnistan poikkeavat havainnot seuraavilla ehdoilla. Lasken jokaiselle ryhmälle mallin hinta-arvion ja neliöhinnan välisen virheprosentin. Tässä tapauksessa virheprosentti kuvaa paremmin mallin tarkkuutta kuin virheen neliö, sillä se suhteuttaa hinta-arvion virheen alkuperäiseen neliöhintaan. Lasken jokaiselle kiinteistötunnukselle virheprosentin keskiarvon sekä pienimmän ja suurimman virheprosentin arvon. Tunnistan poikkeavat tonttikaupat testiaineistosta koottujen tonttiryhmien pienimmän ja suurimman virheprosentin avulla. Tarkoituksenani on rajata aineistosta pois sellaiset tonttikaupat, jotka selvästi poikkeavat hinta-arvioiltaan kaikilla aineistonjaoilla. Poistan aineistosta tonttikaupat, joiden ryhmän pienin virheprosentti on yli 100 tai suurin virheprosentti on vähemmän kuin -60. Poistan siis aineistosta tonttikaupat, joille mallin hinta-arviot ovat toistuvasti yli kaksi kertaa neliöhintaa suuremmat tai vähemmän kuin kaksi viidesosaa toteutuneesta hinnasta. Rajaan tutkimusaineistosta yhteensä 1 044 poikkeavaa tonttikauppaa, joista 177 on toimitilatontteja. Suhteessa muihin tonttikauppoihin toimitilatontteja rajataan aineistosta pois suurempi suhteellinen osa.

Poikkeavien havaintojen poiston jälkeen LOG-GBM-mallin tulokset ristiinvalidoinnissa 100 satunnaiselle aineistonjaolle ovat seuraavat: mallien keskiarvo virheen keskihajonnalle on 505,67 euroa neliömetriltä, keskihajonta on 214,04 euroa neliömetriltä, pienin virheen keskihajonta on 236,92 euroa neliömetriltä ja suurin on 1065,05 euroa neliömetriltä. Poikkeavien havaintojen poistaminen tutkimusaineistosta selvästi parantaa mallin keskimääräistä tarkkuutta testiaineiston neliöhintojen arvioinnissa.

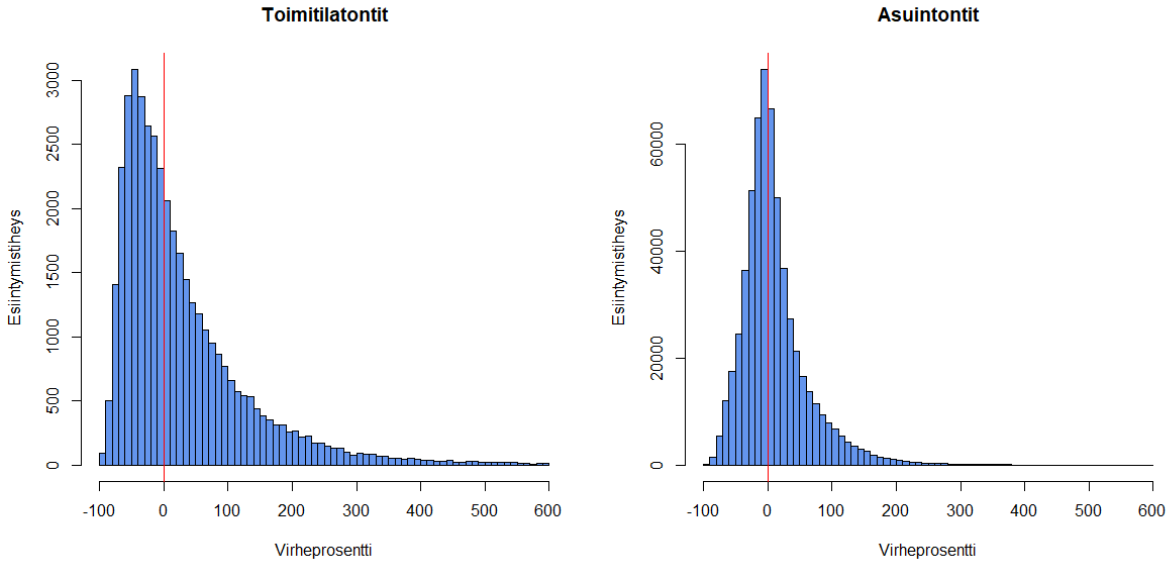
5.5 Soveltuvuus toimitilatonttien hintojen mallintamiseen

Ristiinvalidoinnin $RMSE$-tunnusluku kuvaa mallin estimointitarkkuutta koko aineistolle. Erottelen vielä toimitilatontit muista tonteista erilleen tutkiakseni, kuinka tarkasti malli tuottaa hinta-arvioita toimitilatonteille. Tutkin mallin tarkkuutta edellä suoritetun ristiinvalidoinnin pohjalta. Kuvaan hinta-arvioiden tarkkuutta kauppahinnan ja hinta-arvion välisellä virheprosentilla. Kun virheprosentti on nolla, hinta-arvio on yhtä suuri kuin tontin kauppahinta. Negatiivinen virheprosentti tarkoittaa, että hinta-arvio on pienempi kuin kauppahinta. Positiivinen virheprosentti tarkoittaa vastaavasti, että hinta-arvio on suurempi kuin kauppahinta. Kuvassa 13 esitän virheprosenttien jakaumat toimitilatonteille ja asuintonteille. 100 satunnaisesti jaetussa testiaineistossa on yhteensä 41 167 toimitilatonttia ja 586 933 asuintonttia.

Kuva 13 osoittaa, että mallin estimointitarkkuus toimitilatonttien hintojen estimoinnissa on selvästi epätarkempi verrattuna asuintonttien hinta-arvioihin. Mallin hinta-arviot toimitilatonttien hinnoille ovat keskimäärin kauppahintaa alhaisemmat. Mallin antamien yli kaksinkertaisten hinta-arvioiden osuus kaikista hinta-arvioista on 18,0 prosenttia. Alle puolet kauppahinnasta olevien hinta-arvioiden osuus toimitilatonttien hinta-arvioista on 17,5 prosenttia. Hinta-arvioiden, jotka poikkeavat kauppahinnasta vähemmän kuin 15 prosenttia, osuus on 16,0 prosenttia kaikista toimitilatonttien hinta-arvioista. Vastaava osuus asuintonteille on 34,3 prosenttia. Toimitilatonttien hinta-arvioiden, joiden poikkeavuus on enintään 50 prosenttia kauppahinnasta, osuus on 52,8 prosenttia testiaineistojen toimitilatonttien hinta-arvioista. Asuintonteille vastaava osuus on 77,1 prosenttia.

GBM-malli antaa selvästi tarkempia arvioita asuintoteille kuin toimitilatonteille. Yksi

selvä syy siihen on, että asuintonttien määrä tutkimusaineistossa on merkittävästi suurempi kuin toimitilatonttien. Lisäksi on mahdollista, että mallissa käyttämäni selittävät muuttujat eivät sovellu liike- ja toimistotonttien hintojen selittämiseen. Toimitilatonttien hinnanmuodostukseen voi vaikuttaa eri tekijät, sillä toimitilatonttien ostajat ovat pääsääntöisesti yrityksiä, kun taas asuintonttien ostajista valtaosa on henkilöasiakkaita. Toimitilatonttien virheprosenttien jakaumasta havaitsen, että mallin hinta-arviot ovat usein pienemmät kuin kauppahinta. Vahvasti asuintontti painottein koulutusaineisto arvio toimitilatonttien hintoja alakanttiin. Tämä voi merkitä sitä, että toimitilatonttien hinnat ovat lähtökohtaisesti korkeammat kuin asuintonttien. Toimitilatonttien hinta-arvioiden tarkkuudessa on kuitenkin suurta vaihtelua, joten toimitilatonttien korkeammasta hintatasosta asuintontteihin verrattuna ei voida tehdä yksiselitteistä tulkintaa.



Kuva 13: Ristiinvalidoitujen tonttihintojen ennusteiden virheprosenttien jakauma toimitila- ja asuintonteille

6 Yhteenveto

Tutkimuksen tulosten pohjalta totean, että rekisteriaineiston tietoihin perustuen on mahdollista rakentaa empiirinen malli, joka laskee tonteille hinta-arvion sijaintitekijöiden, tontin koon, rakennusoikeuden ja lähiseudun tilastotietojen pohjalta. Tonttien hinnan muodostukseen vaikuttavat useat eri tekijät, joista kaikista ei löydy tietoja kauppahintarekisteristä. Merkittävimmät puutteet aineistossa ovat maapohjan laatua ja rakennettavuutta tarkasti kuvaavat tiedot. Mallien ristiinvalidoinnin perusteella havaitsin, että empiirisen mallin sovittaminen rekisteriaineistoon onnistuu parhaiten GBM-mallilla. Ristiinvalidoinnin tulos on, että GBM-mallilla on mahdollista päästä tarkempiin hinta-arvioihin kuin lineaarisella regressiomallilla tai regressiopuu-mallilla riippumatta siitä, mitä koulutus- ja testiaineiston jaottelua käytetään.

Ristiinvalidointiin perustuen GBM-mallilla saadaan muita malleja tarkemmat hinta-arviot. En kuitenkaan pystynyt osoittamaan, että mallilla voidaan tehdä tarkkoja ennusteita toimitilatonttien neliöhinnoille. Mallin ennustustarkkuus painottuu asuintonttien neliöhintojen arvioimiseen. Yksi selvä syy tähän on, että asuintonttikaupoista on rekisterissä enemmän tietoja kuin toimitilatonttien kaupoista. Toinen syy voi olla, että mallintamisessa käyttämäni selittävät muuttujat eivät kuvaa tarkasti toimitilatonttien hinnan muodostusta. Kolmas mahdollinen syy on, että toimitilatonttien kauppahinnan määräytymisessä on suurempaa vaihtelua kuin asuintonttien, mikä hankaloittaa toimitilatonttien neliöhintojen tarkkaa arviointia empiirisellä mallilla.

Tutkimuksen tuloksena ei täysin voida väittää, että toimitilatontit olisivat lähtökohtaisesti arvokkaampia kuin muut rakennettavat tontit. Liike- ja toimistotonttien poikkeavaan hintaan voi vaikuttaa se, että ne sisältävät rakennuksen, joka on merkki maapohjan rakennettavuudesta. Tarkemmat tiedot tontin laadusta ja rakennettavuudesta voi parantaa mallin tarkkuutta, mikäli tonttien hintojen keskinäinen vaihtelu johtuu yksinomaan laadusta ja rakennettavuudesta. Aineiston heterogeenisuuden takia aineistossa on tonttikauppoja, jotka jakavat samat havaitut ominaisuudet, mutta joiden neliöhinnat ovat varsin erilaisia. Tarkempi kuvaus maapohjan laadusta voi mahdollisesti selittää neliöhinnan eroja, joita en tutkimuksessani käyttämien muuttujien perusteella havainnut.

Tutkimuksen tuloksena suosittelen, että tonttien hintoja mallinnetaan GBM-mallilla. Kauppahintarekisteriin sovitetulla GBM-mallilla pystytään laskemaan hinta-arvioita, jotka poikkeavat kauppahinnasta enintään 15 prosentilla, kolmannekselle asuintonteista ja kuudennekselle toimitilatonteista. Mallin hinta-arvioiden tarkkuuden parantamiseksi ehdotan maapohjan laatua ja rakennettavuutta kuvaavien muuttujien lisäämistä tutkimusaineistoon, sekä laajentamalla tutkimusaikaväliä ja kasvattamalla toimitiliatonttien määrää tutkimusaineistossa.

Lähteet

Bourassa, Steven C., Martin Hoesli, and Jian Sun. "A simple alternative house price index method." Journal of Housing Economics 15.1 (2006): 80-97.

Bourassa, Steven C., Eva Cantoni, and Martin Hoesli. "Spatial dependence, housing submarkets, and house price prediction." The Journal of Real Estate Finance and Economics 35.2 (2007): 143-160.

Breiman, Leo. Classification and Regression Trees. Belmont (CA): Wadsworth, 1984. Print.

Breusch, Trevor S., and Adrian R. Pagan. "A simple test for heteroscedasticity and random coefficient variation." Econometrica: Journal of the econometric society (1979): 1287-1294.

Case, Karl E., and Robert J. Shiller. "Is there a bubble in the housing market?." Brookings papers on economic activity 2003.2 (2003): 299-362.

Cheshire, Paul, and Stephen Sheppard. "On the price of land and the value of amenities." Economica (1995): 247-267.

Davis, Morris A., and Jonathan Heathcote. "The price and quantity of residential land in the United States." Journal of Monetary Economics 54.8 (2007): 2595-2620.

Del Negro, Marco, and Christopher Otrok. "99 Luftballons: Monetary policy and the house price boom across US states." Journal of Monetary Economics 54.7 (2007): 1962-1985.

Eurostat, Commercial Property Price Indicators: Sources, Methods and Issues, Eurostat, Luxembourg, 2017.

Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.

Glumac, Brano, Marcos Herrera-Gomez, and Julien Licheron. "A hedonic urban land price index." Land Use Policy 81 (2019): 802-812.

Greenwell et al., 2020, `https://CRAN.R-project.org/package=gbm`, (9.3.2022).

Haughwout, Andrew, James Orr, and David Bedoll. "The price of land in the New York metropolitan area." Current Issues in Economics and Finance 14.3 (2008).

Hayashi, Fumio. Econometrics. Princeton University Press, 2011.

IAAO, Standard on Automated Valuation Models (AVMs). International Association of Assessing Officers, Kansas City, Missouri, USA, 2018.

Kok, Nils, Eija-Leena Koponen, and Carmen Adriana Martínez-Barbosa. "Big data in

real estate? From manual appraisal to automated valuation." The Journal of Portfolio Management 43.6 (2017): 202-211.

Laukkanen, Kalle, ja Mäkelä Pekka. "Tutkimusraportti rakennusten ikäalennusten määrittämisestä". (2021). `https://vm.fi/kiinteistoverouudistus`, (12.4.2022).

Liu, Zheng, Pengfei Wang, and Tao Zha. "Land-price dynamics and macroeconomic fluctuations." Econometrica 81.3 (2013): 1147-1184.

Loikkanen, Heikki A., ja Seppo Laakso. "Kaupunkialueen maankäyttö." Kansantaloudellinen Aikakauskirja 115.2 (2019): 219-237.

McCarthy, Jonathan, and Richard W. Peach. "Are home prices the next bubble?." Available at SSRN 634265 (2004).

O'brien, Robert M. "A caution regarding rules of thumb for variance inflation factors." Quality & quantity 41.5 (2007): 673-690.

Ricardo, David. "From the principles of political economy and taxation." Readings in the economics of the division of labor: The classical tradition. [1817], 2005. 127-130.

Rosen, Sherwin. "Hedonic prices and implicit markets: product differentiation in pure competition." Journal of political economy 82.1 (1974): 34-55.

Schulz, Rainer, Martin Wersing, and Axel Werwatz. "Automated valuation modelling: a specification exercise." Journal of Property Research 31.2 (2014): 131-153.

Smith, Adam. The wealth of nations [1776]. Vol. 11937. na, 1937.

Tilastokeskus. "Rakennetun ympäristön rekisteritietojen mahdollisuudet ja karikot." Tilastokeskus, Helsinki, (2021).

The SciPy community, `https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.least_squares.html`, (24.2.2022)

White, Halbert. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." Econometrica: journal of the Econometric Society (1980): 817-838.

Liite A

Arvostusfunktion parametrien estimaatit						
k	Kaupunki	$\ln(\beta)$	α	x_k	y_k	N
1	Helsinki	9,652038	1,5899967	385895	6671892	3173
2	Tampere	7,758901	1,4953763	327531	6822638	1937
3	Oulu	5,942932	0,9886558	428201	7210565	1597
4	Turku	6,570794	1,3273604	239687	6710979	1908
5	Jyvaskyla	5,999994	1,133924	435354	6901469	1389
6	Kuopio	5,00993	0,7604975	534496	6973814	695
7	Lahti	5,47552	1,0061264	427570	6761120	646
8	Pori	4,419107	0,6986492	222727	6827715	536
9	Joensuu	4,923678	0,9492078	642110	6943565	429
10	Lappeenranta	4,993835	0,8587015	563844	6769676	326
11	Vaasa	5,020022	0,9943965	228501	7007649	585
12	Rovaniemi	3,763285	0,3778496	443026	7376412	663
13	Hameenlinna	5,083551	0,9471235	363714	6764965	414
14	Mikkeli	3,254127	0,2639543	514220	6839712	262
15	Porvoo	4,889476	0,7164286	425847	6695470	596
16	Hyvinkaa	7,036569	1,5815436	379098	6724451	521
17	Rauma	3,888397	0,6184933	203831	6789992	147
18	Kouvola	4,21297	0,9048819	484164	6748025	91
19	Seinajoki	4,897132	0,8888688	287518	6968940	1172
20	Salo	3,218216	0,3279828	286577	6700791	234
21	Kokkola	3,816359	0,4918872	310103	7085043	712
22	Lohja	4,331262	0,4238928	337909	6682977	853
23	Jarvenpaa	4,776845	0,2582085	395184	6706033	1832
24	Hanko	4,096752	0,5543815	273907	6639187	136
25	Pietarsaari	3,575573	0,6230726	287550	7068160	312
26	Uusikaupunki	2,613935	0,300376	196062	6753581	132
27	Kotka	4,461759	0,9052857	496388	6703196	166
28	Kemi	2,914718	0,2694594	388674	7292658	176
29	Ahvenanmaa	4,902544	1,0574099	107878	6683435	338

Taulukko 8: Arvostusfunktion parametrit.

	β	keskivirhe	muuttuja
Etäisyys	-0,192188***	0,009652	ln(V(δ))
Etäisyys Helsinki	-0,607392***	0,047497	ln(V(δ))
Etäisyys Tampere	-0,59178***	0,028252	ln(V(δ))
Etäisyys Oulu	-0,365772***	0,021578	ln(V(δ))
Etäisyys Turku	-0,320403***	0,037178	ln(V(δ))
Etäisyys Jyväskylä	-0,237071***	0,025736	ln(V(δ))
Etäisyys Kuopio	-0,239937***	0,033752	ln(V(δ))
Etäisyys Lahti	-0,290611***	0,042516	ln(V(δ))
Etäisyys Joensuu	-0,15613***	0,041793	ln(V(δ))
Etäisyys Lappeenranta	-0,037671*	0,022338	ln(V(δ))
Etäisyys Vaasa	-0,282864***	0,040938	ln(V(δ))
Etäisyys Hämeenlinna	-0,223155***	0,033935	ln(V(δ))
Etäisyys Hyvinkää	-0,405809***	0,06939	ln(V(δ))
Etäisyys Seinäjoki	-0,201621***	0,025031	ln(V(δ))
Etäisyys Lohja	0,158277***	0,011802	ln(V(δ))
Etäisyys Järvenpää	0,163969***	0,025002	ln(V(δ))
Etäisyys Uusikaupunki	0,18792**	0,074858	ln(V(δ))
Etäisyys Kotka	-0,11961***	0,027478	ln(V(δ))
Helsinki	3,225564***	0,133151	ln(V(δ))
Tampere	2,273426***	0,082299	ln(V(δ))
Oulu	1,382606***	0,06892	ln(V(δ))
Turku	1,048076***	0,094146	ln(V(δ))
Jyvaskyla	0,776562***	0,068352	ln(V(δ))
Kuopio	0,934461***	0,108769	ln(V(δ))
Lahti	0,811751***	0,108892	ln(V(δ))
Pori	-0,278398***	0,039234	ln(V(δ))
Joensuu	0,284905**	0,10832	ln(V(δ))
Vaasa	0,659135***	0,106133	ln(V(δ))
Hameenlinna	0,463168***	0,085416	ln(V(δ))
Porvoo	0,408183***	0,036646	ln(V(δ))
Hyvinkaa	1,184755***	0,172454	ln(V(δ))
Kouvola	-0,445136***	0,089837	ln(V(δ))
Seinajoki	0,353369***	0,068684	ln(V(δ))
Salo	-0,318097***	0,057764	ln(V(δ))
Kokkola	-0,162127***	0,032058	ln(V(δ))
Jarvenpaa	0,4422***	0,067052	ln(V(δ))
Pietarsaari	-0,277371***	0,046929	ln(V(δ))
Uusikaupunki	-0,925044***	0,217165	ln(V(δ))
Kemi	-0,353262***	0,065428	ln(V(δ))
Observations	21978		
R^2	0,816		
Adjusted R^2	0,816		
Residual Std. Error	0,724 (df = 21914)		
F Statistic	1543,890*** (df = 63; 21914)		
Note:	*p<0,1; **p<0,05; ***p<0,01		

Taulukko 9: Regression tulokset etäisyydelle ja kaupungin keskuspisteelle.