RETAX

Het schatten van de verkoopprijs van residentiële onroerende goederen

Sven Damen & Stef Schildermans

KU LEUVEN

UHASSELT

Voor meer informatie over deze publicatie: sven.damen@kuleuven.be en stef.schildermans@kuleuven.be

RETAX is een Strategisch Basis Onderzoek (SBO, S005718N) gefinancierd door het Fonds voor Wetenschappelijk Onderzoek (FWO).

Deze publicatie is ook beschikbaar via www.retax.be

INHOUD

INLEIDING

Het kadastraal inkomen (KI) is een schatting van het jaarlijkse netto huurinkomen van het onroerend goed. Het wordt gebruikt als de belastbare basis in zowel de onroerende voorheffing (OV) als de personenbelasting. De schattingen waarop het KI gebaseerd is, dateren echter van 1975. De huurwaarde van een onroerend goed kan doorheen de tijd drastisch veranderen. Eerdere onderzoeken tonen aan dat woningen met een vergelijkbare verkoop- of huurprijs sterk verschillende KI's kunnen hebben (Mahieu et al., 2012; Boogaerts et al., 2020). Met andere woorden, woningen die de dag van vandaag aan dezelfde prijs gekocht of verhuurd kunnen worden, worden niet altijd even zwaar belast. Het gebruik van het KI als belastbare basis in de OV en personenbelasting zorgt voor een ongelijke behandeling van belastingplichtigen met vergelijkbare inkomens. Deze zogenaamde horizontale ongelijkheid is één van de belangrijke redenen waarom een vervanging van het kadastrale inkomen als belastbare basis door een actuele waarde zich opdringt.

Het doel van dit onderzoek is om na te gaan welke statistische modellen het meest accuraat zijn in het schatten van de verkoopprijs van onroerende goederen op basis van de data die momenteel voor de overheid beschikbaar is.[1] Met de verkoopprijs doelen we op de prijs die een verkoper redelijkerwijze kan verwachten wanneer men het onroerend goed verkoopt op de markt. We zullen de verkoopprijs voor woonhuizen, appartementen en bouwgronden apart behandelen. Voor elk van de drie categorieën zijn er namelijk verschillende data beschikbaar en verschillende predictoren zijn van minder of meer belang in verschillende categorieën of helemaal niet beschikbaar. Intuïtief zal de predictie dus ook beter zijn indien we de onroerende goederen opsplitsen in de drie genoemde categorieën.[2] Uit deze analyse zal dan ook duidelijk worden in welke mate het mogelijk is de prijs te schatten en of dit al dan niet verschilt tussen de categorieën. Belangrijk om op te merken is dat de modellen echter niet exclusief op de categorie kunnen toegepast worden waarmee ze geconstrueerd zijn. Zo is het bijvoorbeeld mogelijk om het predictiemodel verkregen via de transacties van bouwgrond uiteindelijk ook toe te passen op woonhuizen. Aan de hand van deze methode is het dan mogelijk om de voorspelde landprijs te verkrijgen voor een bouwgrond waarop reeds een woonhuis staat.

Om de prijzen te schatten werden diverse databronnen met betrekking tot de woningstock, verkopen en EPC verzameld die we bespreken in hoofdstuk 1 en hoofdstuk 2. Hierbij zal er speciale aandacht zijn voor de relatie tussen de variabelen en de verkoopprijs. In hoofdstuk 2 wordt bovendien ook de methode van predictie in detail uitgelegd alsook de manier waarop we de performantie kunnen meten. In hoofdstuk 3 bestuderen we dan welke predictiemodellen het best presteren met betrekking tot de schatting van de prijs van woonhuizen. Daarnaast bekijken we ook welke variabelen het best opgenomen kunnen worden en welke niet, alsook de optimale constructie van extra variabelen die we zelf opstellen. Ook zal er speciale aandacht zijn voor de impact van het toevoegen van de EPC-score, aangezien deze uit een extra databank komt die niet beschikbaar is voor elk onroerend goed. Als laatste interpreteren we de coëfficiënten van de OLS regressie om zo een duidelijk beeld te krijgen over het effect van woningkarakteristieken op de prijs. Hoofdstuk 4 en 5 zullen hetzelfde doen voor respectievelijk appartementen en bouwgronden.

Op het einde van dit rapport interpreteren we de predictie in het licht van het beoogde doel, namelijk het schatten van een belastbare basis. We geven aan welke valkuilen kunnen bestaan bij het praktisch implementeren van deze methode en lichten toe hoe de huidige data en bijhorende problemen de performantie kunnen beïnvloeden.

[1] In een later rapport zullen we de mogelijkheden nagaan om huurprijzen te schatten van de volledige woningstock.
[2] Na een test bleek dit inderdaad het geval te zijn.

1. DATA

Om de prijzen te schatten gebruiken we data uit drie bronnen: 1) de woningstock – alle gebouwen en gronden in België, 2) de transacties – het geheel van transacties in notariële overeenkomsten voor heel België en 3) de energieprestatie certificaten (EPC) voor Vlaanderen. Voor een beschrijving van de woningstock en de transacties verwijzen we naar een voorgaand verslag getiteld “De staat van het KI” (Boogaerts et al., 2020). In dit rapport zullen we voornamelijk focussen op de opdeling van de transacties in huizen, appartementen en bouwgronden alsook welke observaties we verwijderden. Ook zullen we in deze sectie de belangrijkste beschrijvende statistieken tonen en de relatie met de verkoopprijs. Verder zullen we de algemene beschrijving van de EPC databank hier wel behandelen, aangezien deze in voorgaande rapporten nog niet aan bod kwam. De woningstock is enkel gebruikt om omgevingsvariabelen te creëren voor de schatting van landprijzen, een onderwerp dat in bod komt in hoofdstuk **Error! Reference source not found.**.

1.1 Transacties

1.1.1 Selectie van observaties

De dataset met de transacties bestaat uit 4,62 miljoen observaties. Een typische observatie is een woning, appartement of grond die verkocht is. Echter, niet alle observaties zijn individuele transacties. Een deel van deze observaties behoren tot één transactie. Met andere woorden, ze worden samen verkocht. Het gaat dan bijvoorbeeld om de gezamenlijke aankoop van een appartement en een garage. Aangezien we tot op heden niet over een variabele beschikken waarmee we de verschillende observaties die samen één transactie vormen kunnen identificeren, hebben we een alternatieve methode moeten hanteren om hier enigszins zicht op te krijgen. Zo veronderstellen we dat observaties die zowel dezelfde prijs, datum van verkoop en in dezelfde kadastrale sectie gelegen zijn tot één en dezelfde transactie behoren. Deze observaties worden verwijderd aangezien we de prijs niet exact kunnen opsplitsen over de verschillende goederen die verkocht zijn. Tenslotte verwijderen we ook observaties die dezelfde capakey en datum van verkoop hebben, aangezien ook hier niet duidelijk is welke prijs correct is. Deze werkwijze doet de volledige dataset slinken van 4,62 miljoen observaties naar 2,01 miljoen.

De dataset bestaat niet enkel uit verkopen uit de hand, maar bevat ook openbare verkopen, schenkingen en erfenissen. We verwijderen deze observaties - 74.091 stuks - aangezien de prijs van deze transacties mogelijks geen goede weerspiegeling is van de marktprijs. Daarna splitsen we de dataset op in drie delen: woonhuizen, appartementen en bouwgronden. We filteren telkens op akte en constructie indicatie alvorens de uitschieters te verwijderen die hoogstwaarschijnlijk een (meet)fout zijn in de data. De observaties moeten dus duidelijk tot één van de drie categorieën behoren op basis van zowel de akte als de constructie indicatie alvorens we deze opnemen.

Voor woonhuizen selecteren we de observaties die in de akte omschreven worden als ‘woonhuis’, ‘villa’, ‘bungalow’ of ‘hoevegebouw’ én over een constructie indicatie beschikken waarin ‘huis’, ‘hoeve’, ‘villa’, ‘bungalow’ of ‘fermette’ voorkomt. Het opleggen van beide selectiecriteria verkleint de data set verder tot 911.899 unieke observaties. Appartementen zijn de observaties met een akte die vermelding maakt van een ‘appartement’ of ‘studio’ en een constructie indicatie waarin ‘wooneenheid’ voorkomt. Deze set bestaat uit 258.240 unieke observaties. Voor bouwgronden selecteren we de observaties waarin men in de akte vermelding maakt van een ‘bouwgrond’ of een ‘perceel grond’ en een constructie indicatie die niet ingevuld is. Deze set bestaat uit 223.622 unieke observaties.

Voor zowel woonhuizen als appartementen verwijderen we alle observaties met een prijs lager dan EUR 50.000 of hoger dan EUR 2.000.000. Ook verwijderen we observaties met een kadastraal inkomen lager dan EUR 50 of hoger dan EUR 3000 en huizen met een perceeloppervlakte kleiner dan 20 m² of hoger dan 4000 m² (niet van toepassing voor appartementen). Voor nuttige oppervlakte en bebouwde oppervlakte hanteren we minimumwaarden van respectievelijk 15 en 10 m² (laatste opnieuw niet van toepassing voor appartementen), terwijl we maximumwaarden hanteren van 2000 m² voor beide variabelen. Als laatste verwijderen we ook observaties met meer dan 5 garages, meer dan 5 bovenverdiepingen, meer dan 3 badkamers, meer dan 10 slaapkamers en meer dan 10 aparte wooneenheden. Uiteraard verwijderen we hiermee helaas niet alleen foutieve observaties maar we geloven dat het nadeel verkregen door het verwijderen van de correcte observaties niet opweegt tegen het voordeel verkregen door het verwijderen van de incorrecte observaties. We verkrijgen uiteindelijk twee datasets bestaande uit respectievelijk 826.314 en 244.208 transacties.

Voor bouwgronden moet de prijs van de bouwgrond tussen de EUR 20.000 en EUR 800.000 liggen, het kadastraal inkomen moet lager zijn dan EUR 300 zijn en de oppervlakte van het perceel moet opnieuw minstens 20 m² zijn en maximaal 2000 m². Om zeker te zijn dat er nog geen gebouw staat controleren we ook eens of het constructiejaar niet ingevuld is en de wooneenheden gelijk zijn aan nul. Deze uiteindelijke dataset bestaat uit 128.422 observaties.

1.1.2 Aanpassingen en toevoegingen

Enkele waarden van bepaalde variabelen moesten ook aangepast worden vooraleer we naar de predictie konden overgaan. Zo bevat de transactie dataset twee verschillende variabelen omtrent de oppervlakte van het perceel, namelijk de oppervlakte zoals die genoteerd staan in het kadaster en de oppervlakte zoals die opgegeven is in de akte. Beide variabelen hebben meestal dezelfde waarde, maar indien dit niet het geval is gebruiken we de oppervlakte uit de akte, tenzij deze niet ingevuld is en oppervlakte uit het kadaster wel vermeld is. De oppervlakte uit de akte geniet onze voorkeur omdat we veronderstellen dat deze meer up-to-date is.

Verder bevat de dataset niet het exacte constructiejaar voor oudere huizen, maar enkel een interval. Indien dit het geval is nemen we de middelste waarde van dit interval, aangezien we één specifiek bouwjaar nodig hebben voor de predictie. Ook bevat het bouwjaar voor bepaalde observaties enkel de laatste twee cijfers van het jaar, indien dit het geval was vervolledigde we het jaar. Ook dit is noodzakelijk voor de predictie die hieronder aan bod komt.

Overigens bestaat de transactie dataset uit twee verschillende bronnen, namelijk “cadnet” en “stipad”. De dataverzameling en codering gebeurt op dezelfde manier voor zo ver wij weten met uitzondering van het aantal bovenverdiepingen.[3] Cadnet telt het aantal bovenverdiepingen zonder het gelijkvloers, terwijl stipad het gelijkvloers hierbij telt. Om dit op te lossen coderen we het aantal bovenverdiepingen van cadnet zoals stipad dit doet, dit wil zeggen dat we de initiële waarde sommeren met één indien cadnet de bron is.

Variabelen die we zelf construeren en toevoegen aan de dataset zijn i) de leeftijd van het gebouw, ii) een variabele die aangeeft of er wel of geen renovatie heeft plaatsgevonden, iii) een variabele die de leeftijd van de renovatie aangeeft indien er een renovatie heeft plaatsgevonden en iv) het geïndexeerd kadastraal inkomen. De leeftijd van het gebouw verkrijgen we door het jaar van verkoop te verminderen met het constructiejaar. De variabele die aangeeft of er geen renovatie heeft plaatsgevonden is gelijk aan één indien het renovatiejaar niet is ingevuld en gelijk aan nul indien er een renovatiejaar is toegevoegd. Indien er een renovatie heeft plaatsgevonden verminderen we het jaar van verkoop met het renovatiejaar om aan te geven hoe lang de renovatie geleden is. Als laatste

[3] Er is ook een belangrijk verschil in de oppervlakte uit het kadaster, maar daar wij voornamelijk de oppervlakte uit de akte gebruiken vormt dit geen probleem.

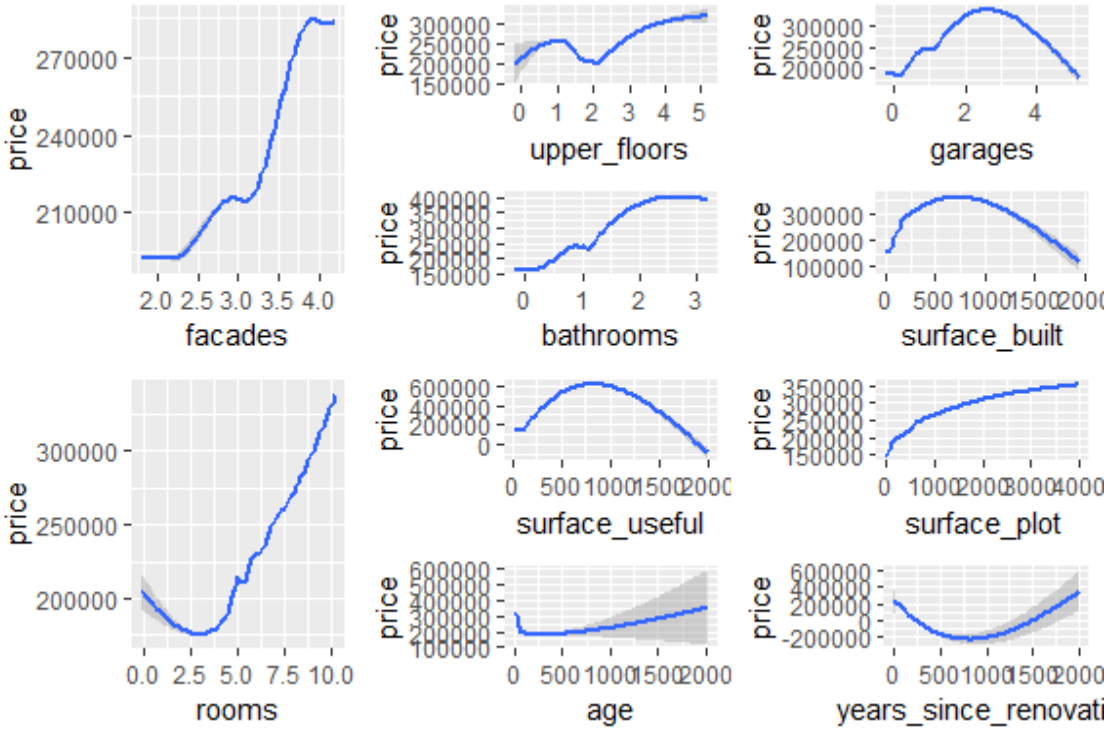
vermenigvuldigen we het kadastraal inkomen met de index van het jaar van verkoop om het geïndexeerd kadastraal inkomen te bekomen.

1.1.3 Relatie met de verkoopprijs

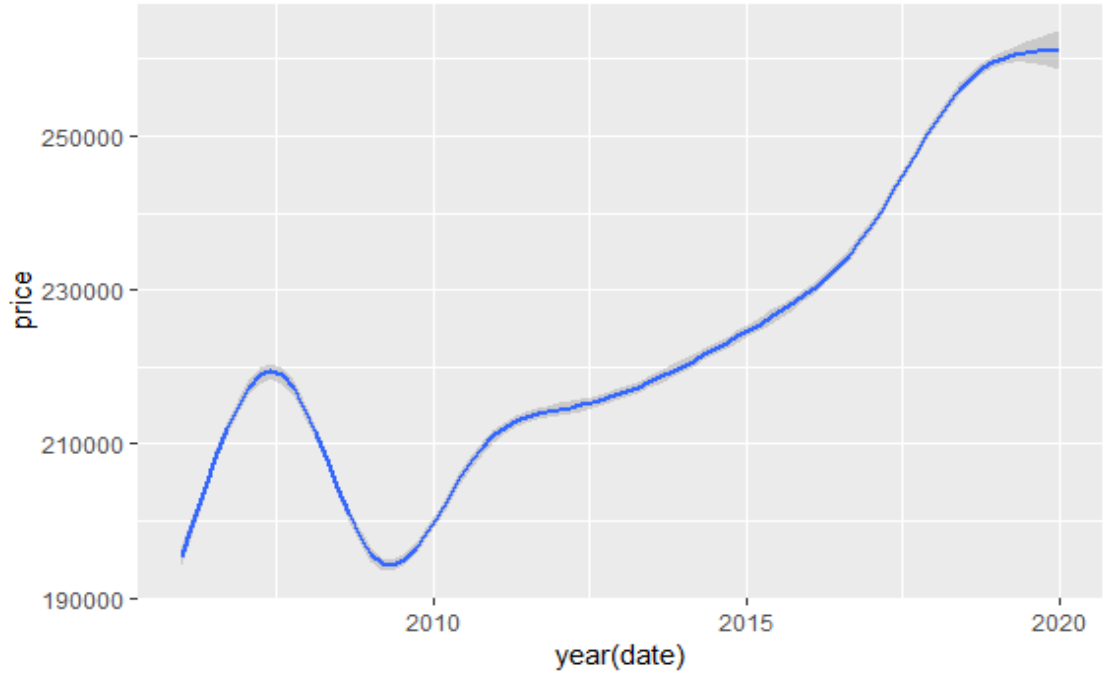
Woonhuizen

In deze sectie bekijken we de relatie tussen de verkoopprijs van woonhuizen en de belangrijke variabelen die later gebruikt zullen worden in de predictie. Voor appartementen is de relatie zeer gelijkaardig, daarom zullen we dit niet nogmaals bespreken in het rapport. Figuur 1 toont deze relatie voor de continue variabelen. Alle variabelen vertonen een relatie met de verkoopprijs in lijn van de verwachtingen. De relatie tussen de oppervlakte en de prijs is kwadratisch van aard voor alle drie de maatstaven van oppervlakte waarover we beschikken. De prijs daalt wel sterk naarmate de nuttige en bebouwde oppervlakte toeneemt. Dit kan verklaard worden doordat de grootte van de woning vaak sterk samen hangt met de locatie. Woningen zijn doorgaans groter op plaatsen waar de grond goedkoper is. Hetzelfde geldt voor de relatie tussen aantal garages en de prijs. Voor de schatting van de verkoopprijzen vormt dit geen probleem aangezien we verschillende variabelen met betrekking tot de locatie opnemen. De leeftijd van het gebouw vertoont ook een kwadratische relatie met verkoopprijs, maar dan convex in plaats van concaaf. Een nieuwe woning is over het algemeen duurder, maar voor zeer oude woningen is het effect omgekeerd. Dit kan verklaard worden door een bepaalde voorkeur voor oude architectuur, waardoor de zeer oude woningen net populair zijn. De resterende variabelen in figuur 1 vertonen een min of meer lineaire relatie met de verkoopprijs.

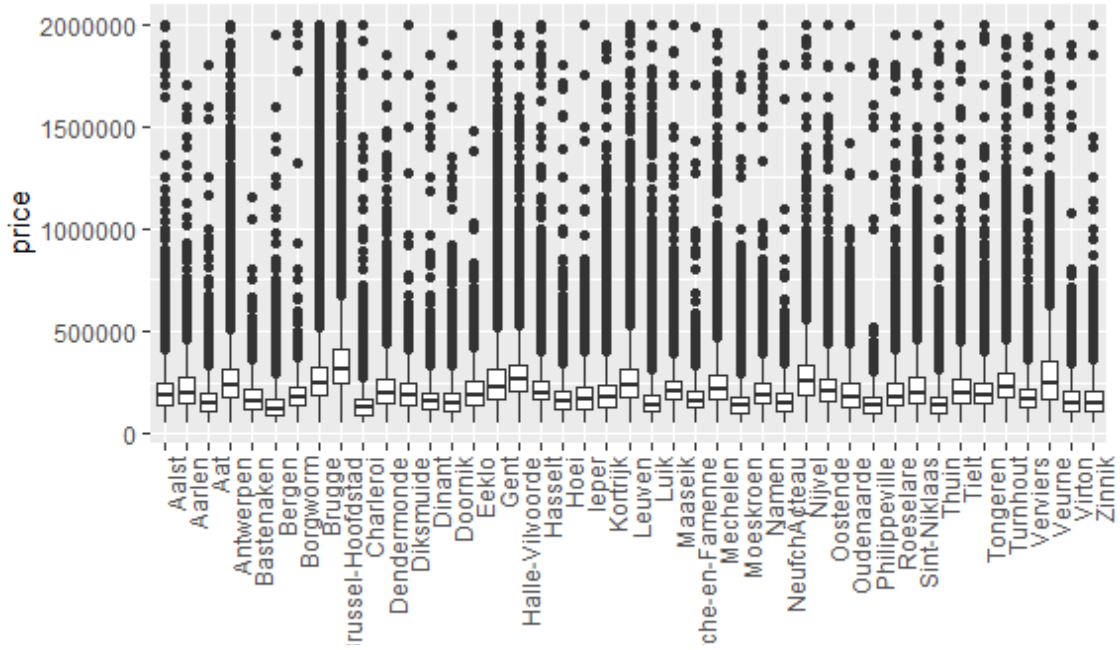
Figuren 2 en 3 tonen de relatie tussen de verkoopprijs en respectievelijk jaar van verkoop en arrondissement. De eerst voornoemde figuur toont een duidelijk positieve relatie tussen het jaar van verkoop en de prijs zoals verwacht. De financiële crisis is op deze figuur ook duidelijk zichtbaar, al was het effect op de Belgische huizenmarkt toch relatief klein. De sterk zichtbare daling is daarom opmerkelijk. De volgende figuur toont aan dat de woningprijzen in onder andere arrondissement Brussel en nabije arrondissementen, Gent en Leuven hoog zijn, terwijl deze laag zijn in bijvoorbeeld Bergen en Charleroi. Dit ligt opnieuw in lijn met de verwachtingen, het is algemeen bekend dat dit inderdaad de dure en goedkope regio’s zijn. Alle arrondissementen hebben gemeen met elkaar dat er veel positieve uitschieters zijn. Het is duidelijk dat woningprijzen scheef verdeeld zijn: de staart langs rechterzijde is veel langer dan de staart die een standaard normaalverdeling karakteriseert. Ook is er veel overlap tussen de verschillende arrondissementen wat betreft huisprijzen. Dit is normaal, in elk arrondissement kan men zowel goedkope als dure woningen vinden, in bepaalde arrondissementen zal men echter meer moeite moeten doen om goedkope al dan niet dure woningen ook daadwerkelijk te vinden.



Figuur 1: De relatie tussen continue variabelen en de verkoopprijs van woonhuizen



Figuur 2: De relatie tussen jaar van verkoop en de verkoopprijs van woonhuizen



Figuur 3: De relatie tussen arrondissement en de verkoopprijs van woonhuizen

Bouwgronden

In deze sectie bekijken we de relatie tussen de belangrijke variabelen en de verkoopprijs van bouwgronden. Figuur 1: De relatie tussen continue variabelen en de verkoopprijs 4 toont opnieuw de relatie tussen de continue variabelen die we verder in de predictie zullen gebruiken en de verkoopprijs. Met uitzondering van een initieel negatieve relatie tussen de oppervlakte van het perceel en de prijs voor zeer kleine bouwgronden vertoont deze relatie opnieuw een kwadratisch verband. De verkoopprijs stijgt naarmate de oppervlakte van het perceel toeneemt zoals we ook verwachten. Vanaf een bepaalde oppervlakte (ongeveer 2700 m²) begint de curve echter opnieuw te dalen. Dit zou verklaard kunnen worden doordat deze gronden zich bevinden op locaties waar er minder vraag is en dus ook een lagere prijs gehanteerd wordt. Een andere mogelijke verklaring is dat deze gronden zich in woonuitbreidingsgebied bevinden en het daarom onzeker is of deze gronden ooit bebouwd mogen worden. Maar even goed kan dit een gevolg zijn van fouten in de data. Zo is het mogelijk dat gronden die in de akte geregistreerd staan als bouwgrond in werkelijkheid landbouwgrond zijn. Dit zou een probleem kunnen vormen voor de predictie en daarom is het ook belangrijk om over een correcte categorisering te beschikken in de mate van het mogelijke. Let op dat de observaties met extreme waarden reeds verwijderd zijn waardoor dit effect in mindere mate een rol zal spelen dan in de ruwe dataset.

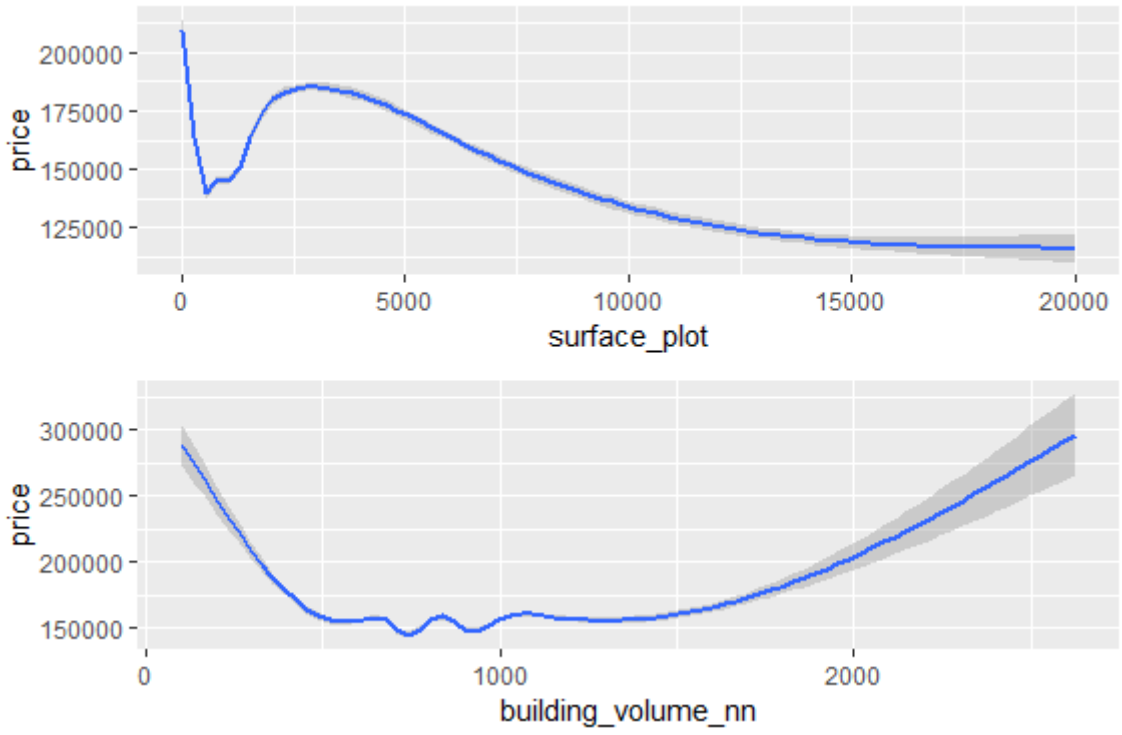
Het typische kenmerk van bouwgronden, namelijk dat er nog niets op gebouwd is, vormt een grote uitdaging voor het schatten van de verkoopprijs. Men mag niet vrij kiezen wat men op een bepaalde grond bouwt. Hetgeen je mag bouwen wordt bepaald door regelgeving die grotendeels door gemeentelijke overheden bepaald wordt, we verwijzen naar deze regelgeving als bouwnormen. Jammer genoeg zijn deze bouwnormen niet gedigitaliseerd. We observeren dus niet wat er op een perceel gebouwd mag worden. Hoewel er dus geen harde data bestaat van bouwnormen, weten we wel dat ze niet alleen van gemeente tot gemeente verschillen, maar dat ze ook binnen gemeenten kunnen verschillen van wijk tot wijk en zelfs van straat tot straat. Meer nog, zelfs binnen een bepaalde straat kan bijvoorbeeld de maximale hoogte van een gebouw verschillen van perceel tot perceel omdat gemeenten deze afhankelijk maken van de hoogte van het bestaande gebouw op de naburige

percelen. Met andere woorden, het volstaat zeker niet te weten in welke gemeente een bouwgrond gelegen is. Nochtans bepalen deze bouwnormen, of anders gezegd het “bouwpotentieel”, in belangrijke mate de waarde van een perceel. De waarde van een perceel op een bepaalde locatie kan bijvoorbeeld veel groter zijn wanneer men er een gebouw mag zetten met meerdere appartementen dan wel één eengezinswoning.

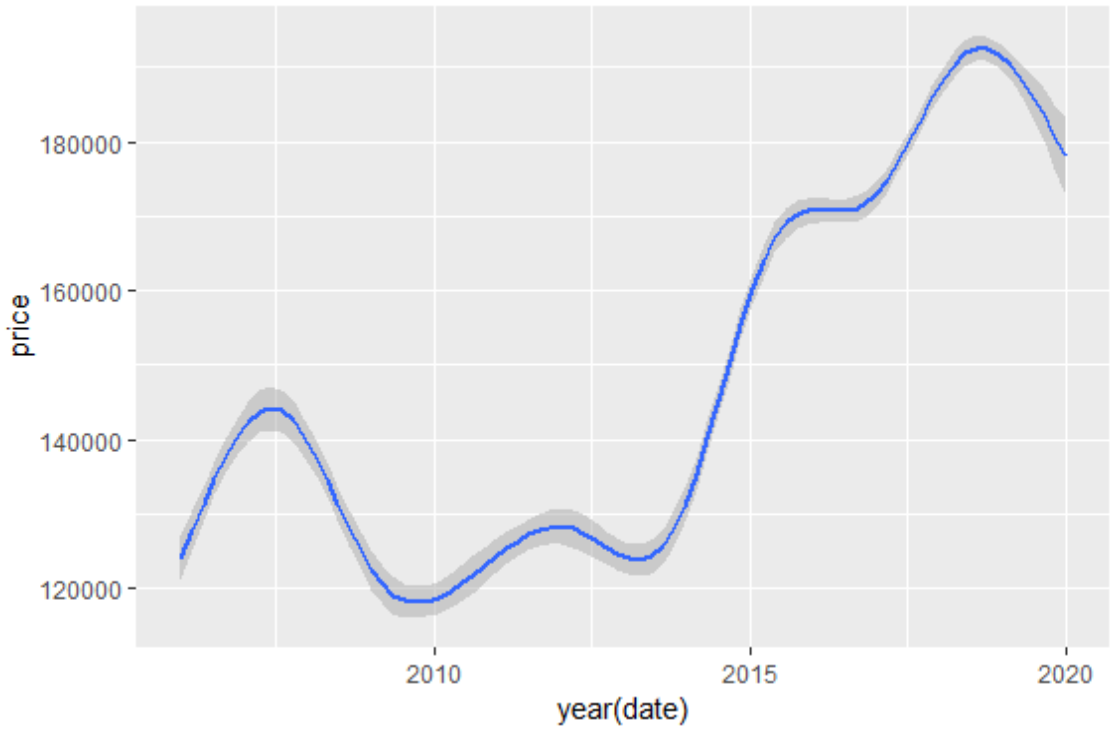
Gezien het belang van bouwnormen voor het schatten van de verkoopprijs van grond, hebben we geprobeerd om deze zo goed mogelijk in kaart te brengen met behulp van de data waarover we momenteel beschikken. Concreet zullen we de gebouwen in de nabije omgeving gebruiken als indicator voor het bouwpotentieel van het desbetreffende perceel. De veronderstelling is dat de gebouwen in de nabije omgeving hun bouwpotentieel hebben verwezenlijkt en daarom aangeven wat de bouwnormen zijn.

We hebben daarom voor alle gebouwen naar de tien dichtstbijzijnde gebouwen gekeken en van deze gebouwen het totale volume en het type bebouwing in rekening genomen. Zo krijgen we voor elke bouwgrond een maatstaf die enigszins aangeeft hoe groot het gebouw mag worden op die desbetreffende locatie en de kans dat het om een bouwgrond gaat voor open bebouwing, halfopen en gesloten bebouwing.

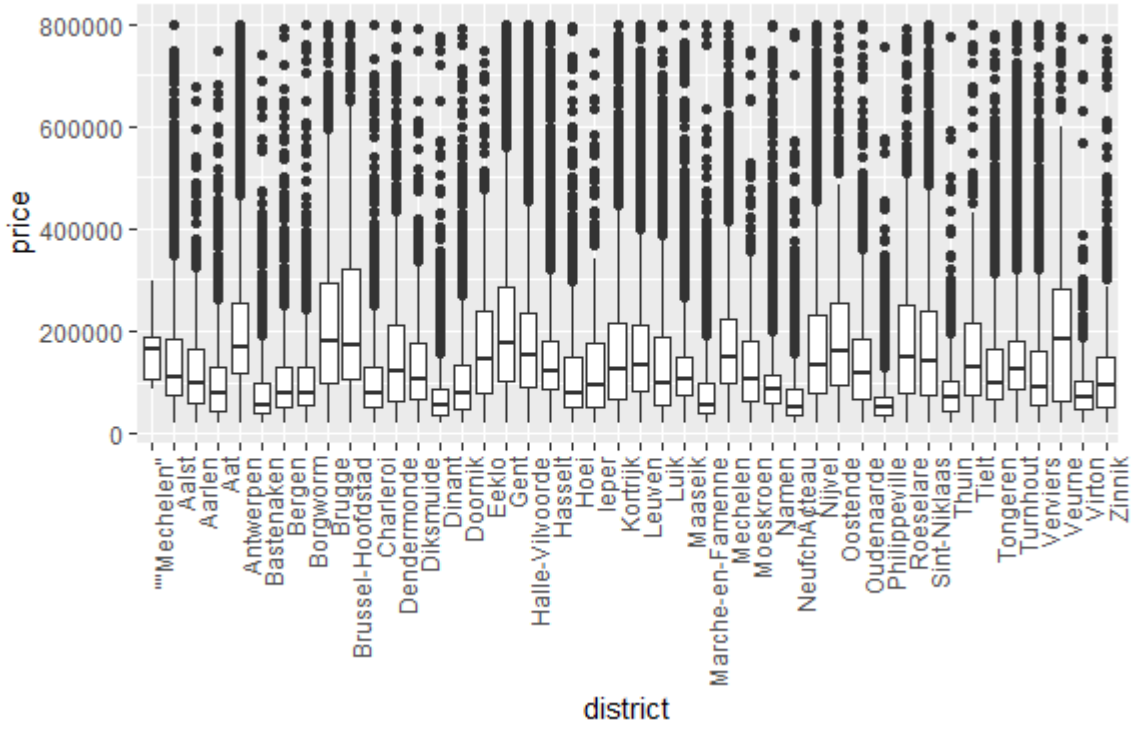
De initiële negatieve relatie tussen het bouwvolume van omliggende gebouwen (weergegeven door de variabele “building_volume_nn”) en de prijs zou op zijn beurt verklaard kunnen worden door de aanwezigheid van kleinere huizen in duurdere locaties. Figuren 5 en 6 tonen gelijkaardige fenomenen als de equivalente relaties bij woonhuizen. Opmerkelijk is wel dat de financiële crisis een langdurigere dip veroorzaakte in de prijs van bouwgronden vergeleken met de prijs van woonhuizen.



Figuur 4: De relatie tussen continue variabelen en de verkoopprijs van bouwgronden



Figuur 5: De relatie tussen jaar van verkoop en de verkoopprijs van bouwgronden



Figuur 6: De relatie tussen arrondissement en de verkoopprijs van bouwgronden

1.2 EPC

1.2.1 Selectie van variabelen

Om de verkoopprijs te schatten zullen we ook gebruik maken van gegevens uit de energieprestatie certificaten, kortweg EPC. We zullen zowel de finale EPC-score overwegen als het totaal energieverbruik zoals die geschat is per woning.[4] Het EPC-kengetal - of de EPC-score - geeft de energieprestatie per m² weer. Deze laatste maatstaf bestaat uit de som van het geschatte energieverbruik voor ruimteverwarming, energieverbruik voor sanitair warm water, energieverbruik voor hulpenergie en energieverbruik voor koeling. Hiervan wordt dan energiebijdrage door PV panelen en energiebijdrage door elektriciteitsproductie van een warmtekrachtkoppeling afgetrokken. We spreken in het vervolg van de energieprestatie van de woning om de kenmerken van de energetische prestaties in de EPC databank aan te geven. Belangrijke opmerking is dat de EPC databank enkel betrekking heeft op de Vlaamse woningstock. Indien we de EPC databank gebruiken voor predictie zal het daarom altijd uitsluitend betrekking hebben op de Vlaamse onroerende goederen.

Zowel de EPC-score als het totaal energieverbruik worden berekend op basis van een uitgebreide set van energetische kenmerken van de woning die in de EPC-databank zijn opgeslagen. De EPC-databank bevat dus meer informatie dan die men terugvindt op een EPC-fiche.[5] In een volgende stap zullen we nagaan of deze extra informatie gebruikt kan worden om de schattingen te verbeteren. Hierbij zal dan onder andere duidelijk worden of er een differentieel effect is van de verschillende vormen van energieverbruik of –bijdrage. In dit rapport beperken we ons echter tot de twee belangrijkste samenvattende maatstaven zoals hierboven beschreven. Deze zijn volgens ons het meest zichtbaar voor de koper en daarom verwachten we ook dat deze variabelen de schattingen van de verkoopprijzen het meest zullen verbeteren.

1.2.2 Extra informatie om scheeftrekkingen in schattingen te verminderen

Het doel van dit onderzoek is om te proberen de verkoopprijs van residentieel onroerende goederen zo accuraat mogelijk te schatten. Het toevoegen van informatie met betrekking tot de energie-efficiëntie, zoals de EPC-score, zal de schatting niet enkel accurater maken, het zal er ook voor zorgen dat de coëfficiënten van de modellen niet scheefgetrokken worden door de afwezigheid van deze informatie.

Verschillende karakteristieken van woningen hangen namelijk sterk samen. Neem nu de leeftijd van de woning. Figuur 7 toont de relatie tussen het bouwjaar van de woning en de EPC-score waarbij het jaar 2000 als referentiejaar wordt genomen. De EPC-score van woningen begint gradueel te verbeteren voor woningen die gebouwd zijn na 1970.[6] Dit is wat we verwachten gegeven de evolutie van de bouwnormen overheen deze periode. Kort gezegd, recentere woningen hebben gemiddeld gezien een betere EPC-score. Heel belangrijk is dat we hier over *gemiddeld* spreken. Want niet alle woningen van een bepaald bouwjaar zullen identiek dezelfde EPC-score hebben.

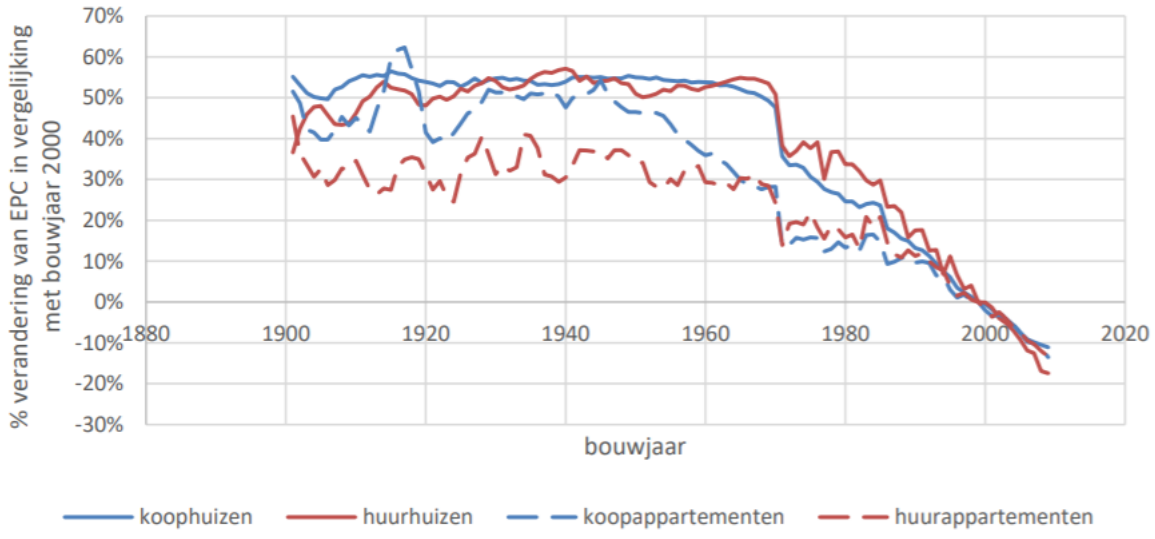
Indien we nu in het model de EPC-score niet opnemen, zal de afwezigheid van deze variabele de coëfficiënt van de leeftijd van het onroerend goed beïnvloeden. Door het sterke verband tussen de EPC score en de leeftijd van de woning, zal de coëfficiënt zal namelijk het effect van de verbetering in de EPC-score opvangen. Dus ook als we een schatting zouden willen maken die geen rekening houdt met de EPC-score van een woning, bv. om te vermijden dat energie efficiëntere woningen zwaarder worden belast, moeten we deze variabele opnemen tijdens het trainen van het model om

[4] Zie ook Damen (2019) voor onderzoek naar de relatie tussen de EPC-score en verkoopprijzen.

[5] Voor verdere info verwijzen we naar Verbeeck, G. & Ceulemans, W. (2015) en Vastmans, F. (2020).

[6] Het bouwjaar wordt bij de berekening van de EPC-score gebruikt om standaardwaarden te bepalen in het geval sommige gegevens niet ingevuld of niet gekend zijn. Een deel van de relatie tussen bouwjaar en EPC-score zou dus mechanisch tot stand kunnen komen door deze methode.

coëfficiënten te verkrijgen die niet beïnvloedt worden door de onderliggende relatie met de EPC-score. Achteraf, tijdens het schatten van prijzen, moeten we de EPC-score dan een gelijk getal toewijzen voor elk huis. Enkel op die manier kunnen we een schatting verkrijgen waarop de EPC-score geen invloed heeft.



Figuur 7: Procentuele verandering in EPC-score volgens bouwjaar (referentie is 2000), Vlaams gewest 2009 – 2016

Bron: Vastmans, F. (2020)

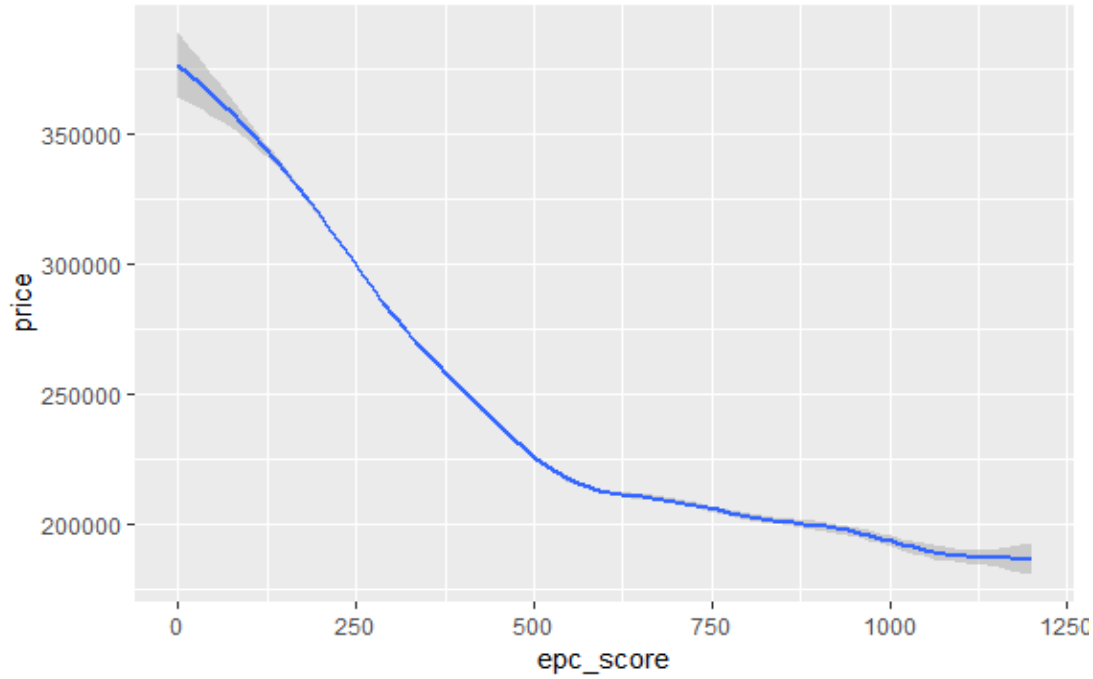
De cijfers in figuur 7 bevatten uiteraard enkel de woningen waarvoor een EPC-fiche beschikbaar is. De woningen waarvoor geen EPC-fiche beschikbaar is hebben dus geen invloed op de cijfers weergegeven in de figuur. Een belangrijke kanttekening daarbij is dat een EPC-fiche vaak aangevraagd wordt voor de verkoop van een woning in het geval van eigenaar-bewoners.[7] Daarom zullen bepaalde woningen ondervertegenwoordigd zijn in deze dataset en andere oververtegenwoordigd. Het is dus niet verstandig om deze cijfers zo maar te extrapoleren naar de volledige Vlaamse woningstock, want dit zou een vertekend beeld kunnen geven. Dit zal echter geen probleem vormen voor de schattingen in dit rapport, gegeven dat elk type woning wel vertegenwoordigd is in de dataset.

De EPC-score is echter niet enkel gecorreleerd met het bouwjaar, maar ook met tal van andere karakteristieken van een woning. Het effect van de energie efficiëntie op de prijs zal dan ook door al deze coëfficiënten worden opgevangen indien we de EPC-score niet toevoegen aan het model. Kort gezegd, indien we energie efficiëntie niet wensen te belasten via de onroerende voorheffing zullen we de EPC-score net wel moeten opnemen tijdens het trainen van het model om achteraf de mogelijkheid te hebben een correctie voor de EPC-score door te voeren.

[7] Het is in de meeste gevallen zelfs verplicht om een EPC-fiche voor te leggen bij de verkoop.

1.2.3 Relatie met verkoopprijs

Om de sectie over de EPC databank af te sluiten tonen we het verband tussen de EPC-score en de verkoopprijs van een woning in figuur 8. Zoals verwacht is er een negatief verband tussen beide variabelen, hoe hoger de EPC-score hoe lager de prijs. Bovendien is het verband niet volledig lineair, de curve is steiler in het geval de EPC-score lager is. Dit betekent dat een daling van één EPC-punt voor een woning met een reeds lage EPC-score gerelateerd is aan een sterkere stijging in de verkoopprijs vergeleken met een woning met een reeds hoge EPC-score. Het is wel belangrijk op te merken dat dit verband niet noodzakelijk causaal is (i.e. oorzaak en gevolg). Wel valt het op dat er een sterk verband is, wat suggereert dat de toevoeging van de EPC-score de schattingen nauwkeuriger zal maken. Verderop in dit rapport zullen we dit ook expliciet empirisch testen.



Figuur 8: De relatie tussen de EPC-score en de verkoopprijs van woonhuizen

1.3 De locatie van onroerende goederen

In de immosector zijn leuzen zoals “locatie is alles” en “locatie locatie locatie” nooit ver weg. Locatie is dan ook een belangrijke factor bij onroerende goederen. Het is echter geen gemakkelijke taak om de locatie-effecten op prijzen van onroerende goederen goed te capteren. Er is geen kant-en-klare variabele beschikbaar die alle facetten van de locatie weet te capteren en die we kunnen gebruiken in een model. Coördinaten of de combinatie van de gemeente en straatnaam zijn genoeg om een specifieke locatie aan te duiden, maar zijn helaas onbruikbaar in de modellen die we hier gebruiken. De combinatie van de gemeente en straatnaam zal namelijk veel te veel mogelijke waarden kunnen aannemen en dat vormt doorgaans een probleem voor de meeste modellen. Het model de coördinaten laten verdelen in bepaalde regio’s zal een gelijkaardig probleem kennen: om realistische regio’s te verkrijgen moeten we de coördinaten zo vaak opsplitsen dat het model veel te complex zou worden om nog goed te kunnen functioneren.

Een iets ruimere territoriale eenheid dan de straat, zoals de regio, provincie, arrondissement, gemeente en statistische sector kan wel werken aangezien het minder verschillende waarden kan aannemen. De keuze voor een bepaalde territoriale eenheid is echter niet voor de hand liggend.

Statistische sectoren zijn nauwkeuriger dan de andere vernoemde eenheden maar omwille van het probleem van over-fitting, dat we later nog zullen bespreken, niet noodzakelijk beter. Later in dit verslag zullen we voor de predictiemodellen analyseren welke territoriale eenheid het best presteert.

Daarnaast zullen we kijken of de afstand tot het centrum een meerwaarde vormt voor de schattingen. Deze afstand is een maatstaf die voor veel mensen een belangrijke factor is en daarom is het ook een logische overweging in een predictiemodel. Ook hier is het cruciaal om de "juiste" afstand te gebruiken. Dit kan zowel de afstand zijn tot het centrum van de gemeente als de afstand tot het centrum van de dichtstbijzijnde centrumstad bijvoorbeeld. We gebruiken hiervoor altijd de afstand in vogelvlucht, niet de afstand die men realistisch zou moeten afleggen door het bestaande wegennetwerk.

Tenslotte gebruiken we ook nog de schattingsfout van een lineair regressiemodel dat verder in de tekst toegelicht zal worden. Specifiek schatten we een lineair model met daarin de gemeente waarin het onroerend goed gelegen is maar zonder andere locatie variabelen. Aan de hand van dit model berekenen we dan de schattingsfout voor iedere transactie. Vervolgens nemen we de schattingsfout van x aantal dichtstbijzijnde onroerende goederen in rekening als we de finale modellen schatten.[8] Via deze schattingsfouten hopen we verschillen in de waarde van locaties binnen een bepaalde gemeente nog beter te capteren. Indien het eerste lineair model de huizen in de nabije omgeving bijvoorbeeld structureel onderschat is het waarschijnlijk zo dat de huizen zich in een goede locatie bevinden en omgekeerd. Deze x aantal schattingsfouten zullen gewogen worden op basis van afstand tot het desbetreffende onroerende goed.

1.4 Regulering en bouwnormen

Een probleem bij bouwgronden is dat we niet beschikken over data omtrent regulering en bouwnormen. Dit is echter wel een belangrijke factor bij het bepalen van de verkoopprijs van een bouwgrond. Een bouwgrond waar een grote villa op gebouwd mag worden zal logischerwijs duurder zijn dan een bouwgrond waar enkel een chalet op mag komen. Daarom proberen we in dit verslag een benadering te creëren van deze bouwnormen door de gebouwen in de nabije omgeving in rekening te nemen. Opnieuw zullen we naar de tien dichtstbijzijnde gebouwen kijken en van deze gebouwen het totale volume en het type bebouwing in rekening nemen. Zo krijgen we voor elke bouwgrond een maatstaf die (hopelijk) aangeeft hoe groot het gebouw mag worden op die desbetreffende locatie en de kans dat het om een bouwgrond gaat voor open bebouwing, halfopen en gesloten bebouwing.

2. METHODE

2.1 Predictiemodellen

Er bestaan verschillende predictiemodellen die men kan gebruiken om de verkoopprijs van onroerende goederen te schatten. Alle mogelijke modellen behoren tot de categorie van regressiemodellen, aangezien we een continue numerieke variabele willen schatten, namelijk de verkoopprijs. Deze groep staat in contrast met classificatiemodellen, die men gebruikt om een categorische variabele te schatten. In alles wat volgt zullen we daarom in dit rapport ook focussen op regressiemodellen.

[8] We overwegen 1 tot en met 20 buren in de analyse hieronder en kunnen via de accuraatheid van de schattingen zien welk aantal het best presteert.

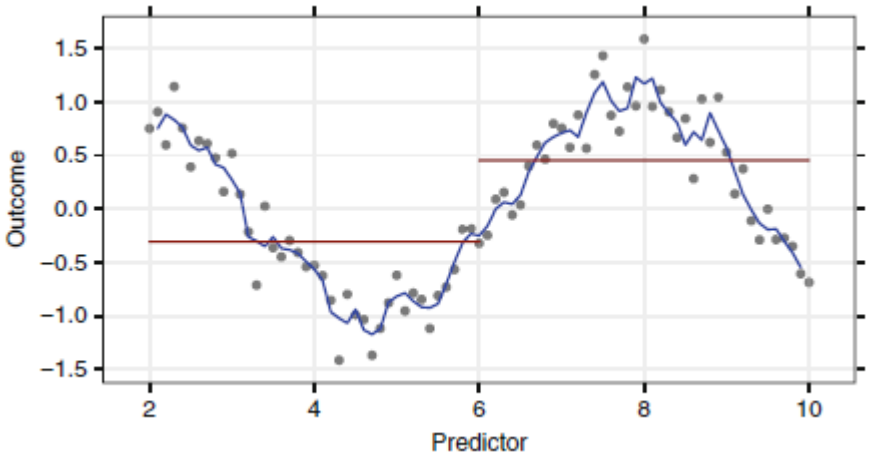
2.1.1 Een één-twee-drie over predictie

Het voornaamste verschil tussen de verschillende regressiemodellen is de plaats die een model inneemt op het zogenaamde “variance-bias trade-off” continuüm. De schattingsfout van een model – namelijk het verschil tussen de geschatte uitkomst en de werkelijke uitkomst - bestaat enerzijds uit ruis en anderzijds uit de bias en variantie van het model. De fout die ontstaat door ruis in de data kan men niet verminderen door het model aan te passen. Zo zullen er huishoudens zijn die meer betalen voor een woning dan de gangbare verkoopprijs gegeven de karakteristieken en locatie van de woning, bijvoorbeeld omwille van persoonlijke smaak of een speciale band met een bepaalde locatie. Een model kan deze persoonlijke smaak van een huishouden niet capteren. Net zo min een professionele schatter dit zou kunnen. Dit is het soort ruis dat altijd in de verkoopprijzen zal vervat zitten. We gaan niet proberen om deze ruis zo goed mogelijk in te schatten met onze modellen. Integendeel, we willen er juist voor zorgen dat deze ruis zoveel als mogelijk niet in de belastbare basis terecht komt. Het lijkt ons immers niet wenselijk dat een enkele uitschieter in de verkoopprijzen – omwille van een zeer specifieke smaak of band met de locatie - de belastbare basis van alle omwonenden zou doen stijgen.

De fout die ontstaat door enerzijds de bias en anderzijds de variantie van het model kan men wel verkleinen of vergroten door het model aan te passen. Het zijn echter tot op zeker hoogte communicerende vaten. Indien men de bias wilt verminderen, zal de variantie van het model stijgen en omgekeerd. Het voornaamste verschil tussen de vele modellen is de plaats die ze innemen op het zogenaamde “*variance-bias trade-off*” continuüm. De *bias* van een model reflecteert hoe goed de functionele vorm van het model de daadwerkelijke relatie tussen de verklarende variabelen, of ‘predictors’, en de verkoopprijs kan capteren. Hoe hoger de *bias* van een model, hoe slechter het model de relatie tussen predictors en verkoopprijs capteert. De variantie geeft aan hoe gevoelig het model is voor afwijkingen in de data. Een model heeft een hogere variantie wanneer het kleinere afwijkingen in de relatie capteert.

Figuur 9 geeft deze trade-off weer aan de hand van gesimuleerde data afkomstig van een sinus-functie. De rode rechte lijnen weerspiegelt een model met een hoge bias maar een lage variantie, terwijl het model weergegeven door de blauwe lijn een lage bias heeft maar een hoge variantie. Simpele modellen, zoals die weergegeven door de rode lijn, hebben een hogere *bias*. dit noemt men ook wel *under-fitting* als het model niet flexibel genoeg is om de daadwerkelijke relatie te capteren.

Complexe modellen, zoals het blauwe model in figuur 9, hebben doorgaans een hogere variantie. De hoge variantie kan leiden tot *over-fitting*. De schattingsfout van een model dat *over-fit* zal zeer klein zijn voor de steekproef op basis waarvan men de parameters van het model geschat heeft, de zogenaamde *in-sample* performantie; ze zal echter sterk toenemen wanneer men het model gebruikt om de prijzen te schatten voor een andere steekproef, de *out-of-sample* performantie. Het model is namelijk zeer sterk beïnvloed door de specifieke afwijkingen in de data van de *training sample*. Afwijkingen die mogelijks niet aanwezig zijn in een andere steekproef. Men dient dan ook altijd een andere steekproef te gebruiken om de performantie van het model te testen dan degene op basis waarvan het geschat is. De exacte methode leggen we hieronder nog in detail uit, maar momenteel is het belangrijk om op te merken dat de geschatte performantie van het model door deze methode niet artificieel hoog kan zijn door over-fitting.



Figuur 9: Twee modellen geschat op gesimuleerde data van een sinus-functie. De rode lijn weerspiegelt een model met een hoge bias en lage variantie, terwijl het model weergegeven door de blauwe lijn een lage bias en hoge variantie heeft.

2.1.2 Verschillende modellen

De modellen die hieronder besproken worden zullen doorgaans eerder neigen naar een relatief hogere bias dan wel een hoge variantie. We beginnen met de lineaire regressiemodellen, die voornamelijk een hoge bias hebben. Vervolgens bespreken we de niet-lineaire modellen en de "regression tree" modellen. Deze modellen omvatten een brede waaier van verschillende specifiekere modellen, en zijn daarom niet in zijn geheel te klasseren op het spectrum van de variance-bias trade-off. Zoals hierboven al vermeldt hangt dit voornamelijk af van de complexiteit van het specifieke model.

Beide type modellen die we hier zullen bespreken hebben ieder een belangrijk voor- en nadeel. Het eerste verschil heeft te maken met de relatie tussen de uitkomst- en de verklarende variabelen, de zogenaamde functionele vorm van het model. Men kan in beide type modellen een niet-lineaire relatie tussen de uitkomst en de verklarende variabelen capteren, maar in de lineaire modellen zal de onderzoeker deze niet-lineaire relatie zelf moeten toevoegen. Niet-lineaire modellen daarentegen zullen automatisch op zoek gaan naar de beste vorm om de niet-lineaire relaties te capteren. Men moet als onderzoeker hier niet zelf naar op zoek gaan.

Het nadeel is dan weer dat de parameters van de niet-lineaire modellen niet of moeilijk te interpreteren zijn. De parameters van lineaire modellen kan men wel interpreteren. Deze tonen namelijk het effect op de uitkomstvariabele van een verandering in een bepaalde predictor. Men kan dan gaan kijken of deze effecten in lijn liggen van de verwachtingen. In het geval van een woning zal men bijvoorbeeld kunnen zien hoeveel de geschatte verkoopprijs stijgt indien men een extra badkamer zou toevoegen. Dit zal niet mogelijk zijn voor de niet-lineaire modellen zonder een simulatie uit te voeren.

Lineaire regressiemodellen

Globaal gezien bestaan de regressiemodellen voornamelijk uit lineaire modellen. De lineaire modellen kan men allemaal schrijven in volgende vorm:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + e_i$$

Aan de linkerzijde van deze vergelijking staat $y_{i,}$ die de numerieke variabele weergeeft die we gaan trachten de schatten. Het subscript i wijst erop dat het in dit geval gaat over de uitkomst voor observatie of in de context van dit rapport beter de woning i. Aan de rechterzijde van de vergelijking vinden we de predictoren terug. De b_0 is het zogenaamde intercept. Deze variabele wordt geschat en is gelijk voor alle observaties, vandaar wordt het ook wel de constante genoemd. Daarna volgen een hele reeks combinaties van b's en x's. De x's zijn hier de verschillende verklarende variabelen die gebruikt zullen worden om de uitkomst te schatten. Denk in dit geval aan de oppervlakte van de woning, het aantal badkamers, enz… Deze verschillen natuurlijk overheen woningen en de specifieke waarde voor een woning wordt opnieuw aangegeven met het subscript i.

De b's zijn de coëfficiënten of parameters van het model die geschat moeten worden. Ze geven de verandering weer in de uitkomstvariabele y voor een zeer kleine stijging in de predictor x. De b's hebben geen subscript i omdat zij verondersteld worden gelijk te zijn voor alle woningen. Het toevoegen van een extra badkamer zal de geschatte verkoopprijs voor alle woningen in dezelfde mate doen stijgen of dalen. Indien we als uitkomstvariabele het logaritme van de prijs gebruiken zal de coëfficient in plaats van de stijging in absolute waarde de stijging in procentuele waarde weergeven. De parameters beschikken wel over een subscript j om aan te geven over welke variabele het nu juist gaat. Helemaal achteraan hebben we e_i. Deze geeft de schattingsfout weer voor observatie i: het verschil tussen de werkelijke waarde van y_i en de schatting van y_i.

De bepaling van het KI is te vergelijken met een lineair regressiemodel. De lokale schatters creëerden een "model" dat bijvoorbeeld aangaf hoe hard de geschatte verkoop- of huurprijs van een woonhuis steeg naarmate het één extra kamer had. Zoals men kan zien is dit conform de bovenstaande vergelijking: het aantal kamers is in dit geval weergegeven door x_{ij} en het effect op de prijs door b_j. Het grote verschil is dat men tijdens de perequatie in de jaren '70 voor elke gemeente, of zelfs voor kleinere geografische eenheden zoals de kadastrale sectie, een afzonderlijk model heeft gemaakt met een eigen reeks van coëfficiënten. In tegenstelling tot het lineaire regressiemodel dat wij zullen hanteren, variëren de b's wel voor de berekening van het kadastraal inkomen. Dit is met de lineaire regressiemodellen ook mogelijk, maar de accuraatheid zal sterk dalen voor gemeenten waar weinig transacties plaatsvinden.

Niet-lineaire regressiemodellen

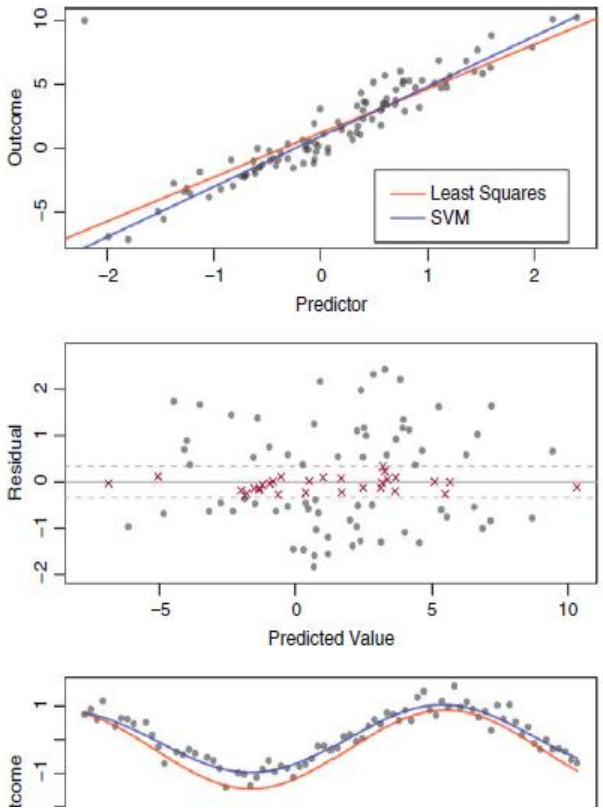
In tegenstelling tot lineaire regressiemodellen moet de onderzoeker bij de niet-lineaire regressiemodellen de vorm van het verband niet a priori specifiëren en daarom ook geen voorkennis hebben over dit verband. In dit rapport behandelen we drie types van niet-lineaire regressiemodellen: i) k-nearest neighbors (kNN), ii) support vector machines (SVM) en iii) regression tree modellen. We zullen een aparte sectie wijden aan het laatste type aangezien deze sterk presteren in het schatten van verkoopprijzen van onroerende goederen.

Een kNN model voorspelt de uitkomstvariabele van een bepaalde observatie simpelweg door de uitkomsten van de k dichtstbijzijnde observaties in rekening te nemen. Bij huizen kan men hierbij enkel de fysieke locatie in rekening nemen zodat het enkel de omliggende huizen in rekening neemt, maar men kan ook alle predictoren in rekening nemen zodat men eigenlijk kijkt naar huizen die in de buurt liggen en gelijkaardige kenmerken hebben. Meestal neemt men het gemiddelde van deze k dichtstbijzijnde observaties, maar andere statistieken zoals de mediaan zijn ook mogelijk.

SVMs worden dan weer gekenmerkt door het feit dat niet alle observaties in de data het model beïnvloeden. De mate waarin een bepaalde observatie de schatting beïnvloedt is afhankelijk van de grootte van de schattingsfout. Observaties met een hoge schattingsfout zullen het model beïnvloeden, terwijl observaties met een lage schattingsfout geen enkel effect hebben. De observaties die het model beïnvloeden noemt men ook wel de "support vectors", vandaar de naam van dit model.

Twee aspecten zijn van groot belang voor een SVM model. Het eerste aspect is de keuze van de "kernel". De kernel bepaalt welke observaties wel en niet het model beïnvloeden. Men kan een lineaire kernel specifiëren, maar men kan ook niet-lineaire kernels opgeven waardoor dit model niet-lineaire relaties kan capteren. Naast de kernel is er ook de "cost parameter", deze parameter bepaalt de flexibiliteit van het model. Wanneer men een hoge cost parameter ingeeft zal het model heel flexibel zijn aangezien het effect van fouten vergroot wordt, wanneer men een lage cost parameter ingeeft zal het omgekeerde gebeuren. Dit heeft als grote voordeel dat men het effect van "outliers" kan verminderen indien nodig. Een outlier is een observatie die ver verwijderd ligt van alle andere observaties. In ons geval zal dat bijvoorbeeld een huis zijn met gemiddelde karakteristieken maar wel een heel hoge verkoopprijs. Indien men het effect van outliers vermindert, vermindert men in feite de variantie van het model ten koste van een hogere bias. De keuze van de cost parameters en kernel gebeurt echter volledig automatisch door de methode die we hieronder nog zullen uitleggen.

Het bovenste en onderste paneel in figuur 10 tonen gesimuleerde data (de grijze bollen) en de geschatte waarden van een lineair regressiemodel (rode lijn) en een SVM model (rode lijn). In beide panelen is het duidelijk dat het SVM model minder gevoelig is dan het lineair regressiemodel voor observaties die verder verwijderd zijn van de "puntenwolk", de outliers. Doordat het minder gevoelig is voor outliers is het SVM model beter in staat om de onderliggende relatie te capteren. De grijze bollen op de middelste figuur zijn de observaties die een invloed hebben op de schatting van het SVM model, de support vectors, de rode kruizen geven de observaties die de schatting niet beïnvloeden. Deze figuur toont dus duidelijk het voordeel aan van een cost parameter indien er outliers in de data zijn.



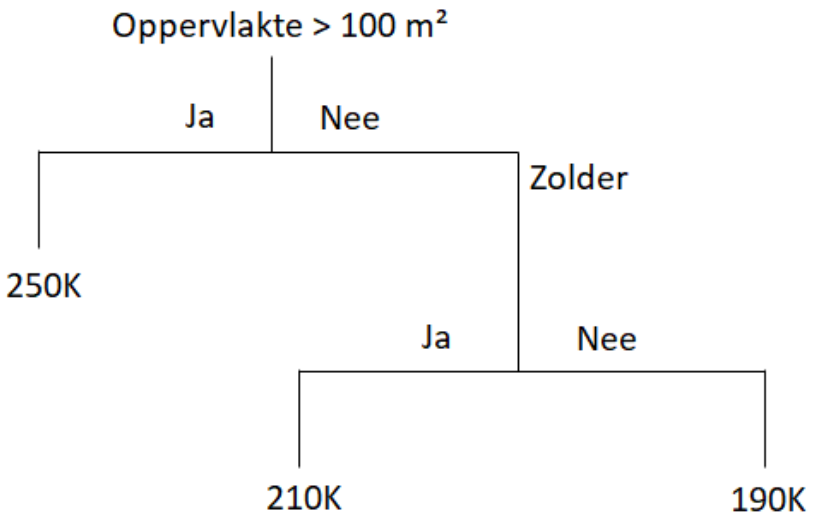
Figuur 10: De bovenste figuur toont aan hoe een SVM model minder sensitief kan zijn ten opzichte van één grote outlier dan het OLS model. De middelste figuur toont de support vectors van een SVM model aan de hand van grijze bollen en de andere observaties die het model niet beïnvloeden aan de hand van rode kruizen. De onderste figuur toont hetzelfde als de bovenste figuur maar dan voor een niet-lineaire relatie waarbij we de relatie kennen en daarom ook een OLS model kunnen schatten.

Regression tree modellen

Een regression tree model verdeelt de data in kleinere groepen die homogener zijn in termen van de uitkomstvariabele, in het geval van dit rapport de verkoopprijs. Teneinde dit te bereiken beslist het model in elk punt van de regression tree i) op basis van welke predictor het de data zal opsplitsen en ii) voor welke waarde van die bepaalde predictor. Het model zal stoppen met de boom te laten groeien wanneer het gewenste niveau van diepte (aantal punten) of accuraatheid bereikt is.

Figuur 11 geeft een voorbeeld van een zeer eenvoudige regression tree. In dit voorbeeld splitst het transacties op in woningen met een oppervlakte groter dan 100m² en woningen met een oppervlakte kleiner of gelijk aan 100 m². De huizen met een oppervlakte groter dan 100m² zullen een prijs van EUR 250.000 toegewezen krijgen, terwijl de kleinere huizen verder opgesplitst worden in huizen met en zonder zolder. De kleinere huizen met zolder zullen daarna een prijs van EUR 210.000 toegewezen krijgen, terwijl de huizen zonder zolder een prijs van EUR 190.000 toegewezen krijgen.

Net zoals een lineair regressiemodel is een regression tree vrij gemakkelijk te interpreteren en intuïtief. Het is echter veel moeilijker om te weten te komen in welke mate de verkoopprijs stijgt of daalt als een bepaalde karakteristiek verandert. Dit is namelijk afhankelijk van de vele interactie-effecten. In figuur 11 is er een interactie-effect tussen de oppervlakte en de zolder: een zolder speelt enkel een rol bij de verkoopprijs indien het huis over een kleinere oppervlakte dan 100 m² beschikt. Een groter huis zal daarentegen niet meer of minder kosten indien het een zolder heeft. Het effect van een zolder is dus afhankelijk van de oppervlakte en niet uniform voor elk huis, dit noemt men een interactie-effect. Een realistische regression tree zal bovendien veel groter zijn dan onderstaand voorbeeld. Dit model verliest daarom de aantrekkelijke kenmerken van het lineair model, ook al is de denkwijze nog steeds zeer intuïtief.



Figuur 11: Voorbeeld van een regression tree.

Een standaard regression tree is doorgaans echter een veel te simplistisch model. Het verdeelt de verklarende variabelen in verschillende groepen en zal voor elke groep slechts één bepaalde verkoopprijs opgeven zoals in Figuur 11. In het voorbeeld van hierboven met twee verschillende predictoren verdeelt het model de ruimte van de predictoren daarom in rechthoekige regio's waarbij het model voor elke regio een verschillende verkoopprijs schat. In realiteit kan de relatie tussen de verklarende variabelen en de uitkomstvariabele meestal niet adequaat beschreven worden door zo'n rechthoek. Bovendien zijn het aantal mogelijke geschatte waarden eindig en bepaald door het aantal eindpunten van de boom. In de praktijk zijn er echter oneindig veel verschillende verkoopprijzen mogelijk. Deze twee beperkingen van een regression tree vormen dus een groot nadeel in deze toepassing.

Om bovenstaande problemen op te lossen kan men een "bootstrap" procedure volgen. Bootstrappen is het proces waarbij er verschillende steekproeven willekeurig getrokken worden uit de originele dataset. Vervolgens schat men een regression tree op ieder van deze steekproeven en gebruikt men alle verkregen modellen om een schatting te maken voor dezelfde observatie. Uiteindelijk neemt men het gemiddelde van deze waarden als finale schatting. Deze methode levert een *bagged tree model* op en levert doorgaans veel kleinere schattingsfouten op dan het standaard *regression tree* model.

De verschillende regression trees in een bagged tree model zullen echter sterk gecorreleerd zijn met elkaar door de relatie tussen de verklarende variabelen en de uitkomstvariabele. Hierdoor is de variantie in een bagged tree model hoger dan wat wenselijk is. Een simpele oplossing is om slechts een beperkt aantal willekeurig gekozen predictors in overweging te nemen in elk punt van de boom, in plaats van de volledige set van predictors. Doordat de predictors willekeurig gekozen zijn, zullen de verschillende regression trees anders opgebouwd zijn. Hierdoor zullen deze minder sterk gecorreleerd zijn, wat de variantie van het gehele model ten goede komt. Het bagged regression tree model met willekeurig gekozen predictors heet een random forest model.

Uit verschillende competities op de website "Kaggle" blijkt dat doorgaans een random forest model, en gelijkaardige modellen, zeer goed zijn in het schatten van verkoopprijzen van woningen. Eén van die gelijkaardige modellen is de gradient boosting machine (GBM). De GBM is ook gebaseerd op meerdere regression trees, maar anders dan bij het random forest model zijn de trees niet onafhankelijk van elkaar. De verschillende trees bouwen voort op elkaar. Men schat telkens een eenvoudige tree en met de predictiefout die men verkrijgt bouwt men dan de volgende tree op totdat men een bepaald niveau van accuraatheid of een bepaald aantal trees heeft bereikt.

Alle regression tree modellen die we tot nu hebben besproken eindigen in een punt met één bepaalde waarde voor de uitkomstvariabele. Men kan echter ook een lineair regressiemodel toepassen op elk eindpunt in plaats van één bepaalde waarde. Het lineair model zal dan geschat worden aan de hand van de beperkte dataset die men verkrijgt op elk eindpunt van de boom. Het *cubist* model implementeert dit idee. Er zijn echter nog andere zaken waarin het cubist model met voorgaande modellen verschilt, maar om de lengte en complexiteit van deze tekst te beperken zullen we de andere verschillen hier niet toelichten.

In het vervolg van dit rapport vergelijken we voor zowel woonhuizen, appartementen als bouwgronden elk hierboven beschreven model. Of we het één model boven het andere verkiezen hangt onder andere af van de performantie van elk model. Dit concept beschrijven we in het volgend hoofdstuk.

2.2 Performantie

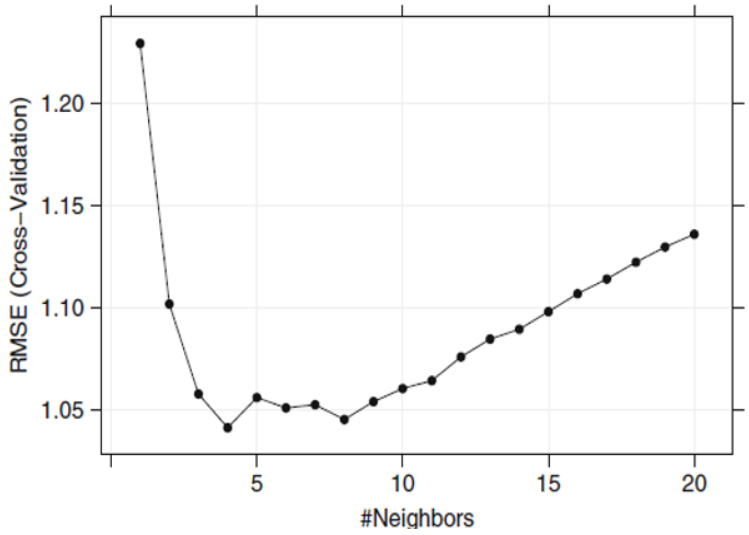
2.2.1 Cross-validatie

Om de performantie van het model te testen maken we gebruik van een methode genaamd cross-validatie. Hierbij splitsen we de volledige transactie dataset op in meerdere gelijke delen, in ons geval exact vijf. Daarna starten we een procedure waarbij we de modellen schatten, of "trainen" om het in *Artificial Intelligence* jargon te zeggen, met behulp van de data. Hierbij laten we telkens één van die vijf delen eruit. Het deel van de data dat we niet gebruiken om het model mee te schatten, zullen we gebruiken voor de zogenaamde out-of-sample test. We berekenen een schatting van de prijs van de onroerende goederen in deze "test" steekproef aan de hand van de geschatte modellen. Het verschil tussen deze schatting en de werkelijke verkoopprijs – de schattingsfout – is de maatstaf voor de accuraatheid van het model. Deze stap herhalen we daarna tot elk van de vijf datasets éénmaal gebruikt is om het model te testen. Bijgevolg zal elke dataset dan ook vier keer gebruikt worden om het model te trainen.

Belangrijk is dat we de performantie van de modellen testen op basis van een steekproef die niet gebruikt is geweest om de modellen mee te trainen, vandaar de naam *out-of-sample*. De out-of-sample test komt namelijk overeen met de manier waarop men deze modellen zou gebruiken om een nieuwe belastbare basis te schatten. Men zal de modellen eerst trainen op de data van de transacties en ze daarna gebruiken om de prijs van alle woningen die niet verkocht werden te schatten.

Een out-of-sample test is ook de enige manier om te zien of *over-fitting* een probleem vormt. Dit kan men niet nagaan met *een in-sample* test aangezien over-fitting net gaat zorgen voor een betere performantie in de training dataset. Bovendien kunnen we aan de hand van cross-validatie objectief waarnemen of een gegeven variabele de performantie daadwerkelijk beïnvloedt, iets wat in-sample ook niet kan aangezien elke toegevoegde variabele de performantie in dat geval zal verhogen of op zijn minst constant houden.

Dit proces is ook belangrijk om na te gaan welke parameters het best werken voor modellen die één of meerdere parameters vereisen, zoals bijvoorbeeld de cost parameter in een SVM model. We doorlopen dit hele proces dan voor verschillende waarden voor deze parameters en kijken voor welke waarden het model de meest accurate schattingen oplevert. Figuur 12 geeft bijvoorbeeld aan hoe de accuraatheid van een kNN model, weergegeven door de *Root Mean Squared Error* (RMSE), varieert met het aantal *neighbours*. In onderstaand geval is het model het meest accuraat wanneer men vier neighbours in rekening neemt.

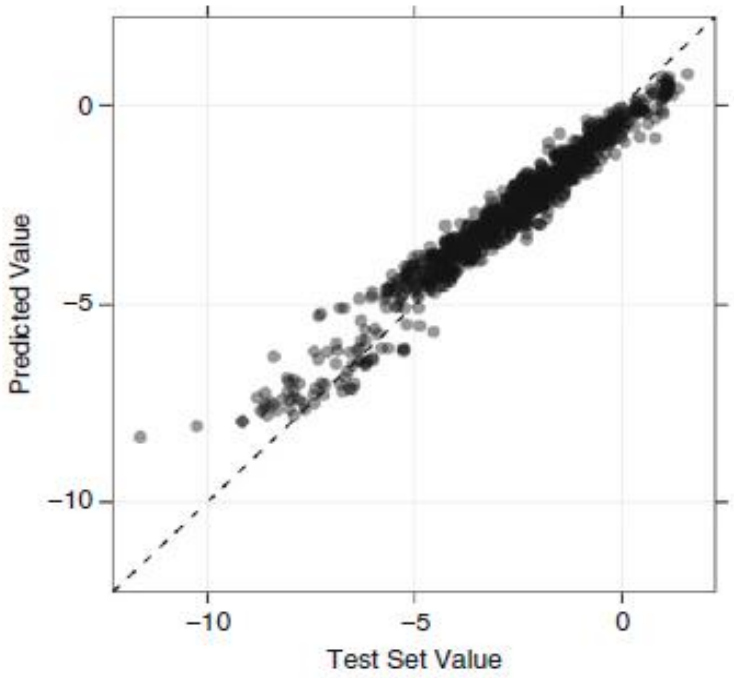


Figuur 12: Cross-validatie bij een kNN model

2.2.2 Kwantitatieve maatstaven van performantie

In dit verslag maken we gebruik van twee populaire maatstaven van accuraatheid, namelijk de *R-squared* (R^2) en de *mean absolute error* (MAE). De R^2-waarde geeft aan hoeveel variatie in de verkoopprijzen verklaard kan worden door het model. Een R^2-waarde van 0.50 betekent bijvoorbeeld dat het model 50% van de variatie in de uitkomstvariabele kan verklaren. Deze maatstaf wordt verkregen door het kwadraat te nemen van de correlatie tussen de geschatte uitkomst en de werkelijk geobserveerde uitkomst. De R^2 is makkelijk te interpreteren, maar aangezien het gebaseerd is op een correlatie kan het ook enkele problemen met zich meebrengen. Figuur 13 toont hier een voorbeeld van. Het model overschat lage waarden en onderschat hoge waarden, maar vertoont toch nog een hoge correlatie tussen de geschatte waarde en de werkelijke uitkomst en dus een hoge R^2-waarde. Bovendien hangt de waarde af van de variatie in de uitkomstvariabele. Als verkoopprijzen bijvoorbeeld variëren tussen EUR 50,000 en EUR 10,000,000 zal een model met een hoge R^2-waarde nog altijd grote schattingsfouten kunnen voortbrengen. Indien verkoopprijzen variëren tussen EUR 50,000 en EUR 500,000 zullen de schattingsfouten kleiner zijn voor eenzelfde R^2-waarde.

Omwille van deze problemen met de R^2 gebruiken we ook de MAE als maatstaf voor de performantie van de modellen. Deze waarde is afhankelijk van de schattingsfouten of "*residuals"*. Ze is gelijk aan het gemiddelde van de absolute waarde van de schattingsfouten. Aangezien we verder in het verslag het logaritme van de verkoopprijs zullen gebruiken is deze waarde moeilijker te interpreteren dan de R^2-waarde, want dan geeft deze maatstaf de absolute schattingsfout van het logaritme van de verkoopprijs weer. Het logaritme van de verkoopprijs is helaas moeilijker te interpreteren dan de verkoopprijs in euro-waarde, maar voor bepaalde modellen is het makkelijker om deze getransformeerde uitkomstvariabele te schatten.[9] De MAE geeft ons wel een maatstaf die niet onderhevig is aan de problemen van de R^2-waarde.



Figuur 13: Verband tussen de uitkomstvariabele en de voorspelde waarde, waarbij er overschatting is bij lage waarden en onderschatting bij hoge waarden

[9] We hebben ook de niet-getransformeerde verkoopprijs gebruikt om te modellen te trainen ter controle.

Als maatstaven van de performantie van het model lijken de R^2-waarde en de MAE ons van de belangrijkste indicatoren voor de selectie van het model dat uiteindelijk gebruikt zal worden om de prijzen voor de volledige woningstock te schatten. Maar daarnaast zijn de kenmerken van het model volgens ons ook van tel. Men kan bijvoorbeeld het lineair regressiemodel verkiezen ten koste van een nauwkeuriger model omwille van de makkelijk te interpreteren coëfficiënten en de intuïtieve opbouw van het model. De andere kenmerken mogen echter enkel in rekening worden gebracht indien verschillende modellen relatief gelijkaardig presteren op vlak van performantie. In andere woorden, we menen dat de performantie van het model de belangrijkste maatstaf dient te zijn in de selectie, enkel als er op basis van deze indicator geen eenduidige keuze gemaakt kan worden kunnen andere kenmerken de doorslag geven.

2.2.3 Impact van een grote sample

Doorgaans is het beter om over een grotere sample te beschikken. We kunnen in dat geval gebruik maken van meer informatie om een accurater model te produceren. Dit komt echter wel met computationele problemen, aangezien de tijd om een model te produceren sterk toeneemt met het aantal samples. Ook is er sprake van afnemende meeropbrengt: hoe meer samples we reeds hebben, hoe minder nuttig elke bijkomende sample zal zijn. Om de computationele problemen te verminderen zullen we bij het vergelijken van verschillende modellen daarom gebruik maken van een random subsample van 100.000 observaties indien de volledige sample zeer groot is. Dit zal het geval zijn voor huizen en appartementen, maar niet voor bouwgronden. We zullen echter altijd ook de volledige sample gebruiken bij het beste model om de uiteindelijke performantie zo accuraat mogelijk in te schatten. In het vervolg van dit rapport zal u zien dat dit de performantie eigenlijk niet beïnvloedt aangezien we al over 100.000 observaties beschikken in de eerste fase, een hoeveelheid die genoeg blijkt te zijn.

3. PRIJSSCHATTING VAN WOONHUIZEN

3.1 Vergelijking van modellen

In deze sectie vergelijken we de verschillende predictiemodellen die hierboven beschreven zijn, voor het schatten van de verkoopprijs van woonhuizen. Voor elk model gebruiken we cross-validatie om de MAE en R^2 te berekenen. We beperken de sample tot 100.000 observaties aangezien het trainen van modellen op de volledige sample te veel tijd in beslag neemt.[10]

Figuur 14 toont de MAE voor alle modellen. Elke lijn geeft de performantie weer voor één van onze vijf steekproeven (zie hierboven voor een uitleg van cross-validatie). Het is duidelijk dat 'simpelere' modellen veel slechter presteren in vergelijking met de meer 'gesofisticeerde' modellen. De vijf beste modellen presteren zeer gelijkaardig, met een MAE tussen 0.20 en 0.22. Deze groep van modellen bestaat uit GBM, SVM, cubist, lineaire regressie en random forest. Hun niveau van performantie correspondeert met een R^2 net onder 0.70 zoals weergegeven in figuur 15. Wat betekent dat deze modellen ongeveer 70% van de variantie in de verkoopprijzen kunnen verklaren. Hoe we de MAE, die nu weergegeven is in het natuurlijk logaritme van de prijs, kunnen vertalen naar een monetaire waarde zal later in dit hoofdstuk besproken worden.

Tenslotte willen we opmerken dat we bij het lineaire regressiemodel zelf bepaalde termen hebben toegevoegd, zoals kwadratische en logaritmische variabelen, om de performantie te verbeteren. Indien we alle termen simpelweg lineair zouden toevoegen zou de performantie van dit

[10] We hebben een aantal modellen op de volledige sample geschat en op basis van deze resultaten bleek dat er geen significante verschillen optreden in de performantie.

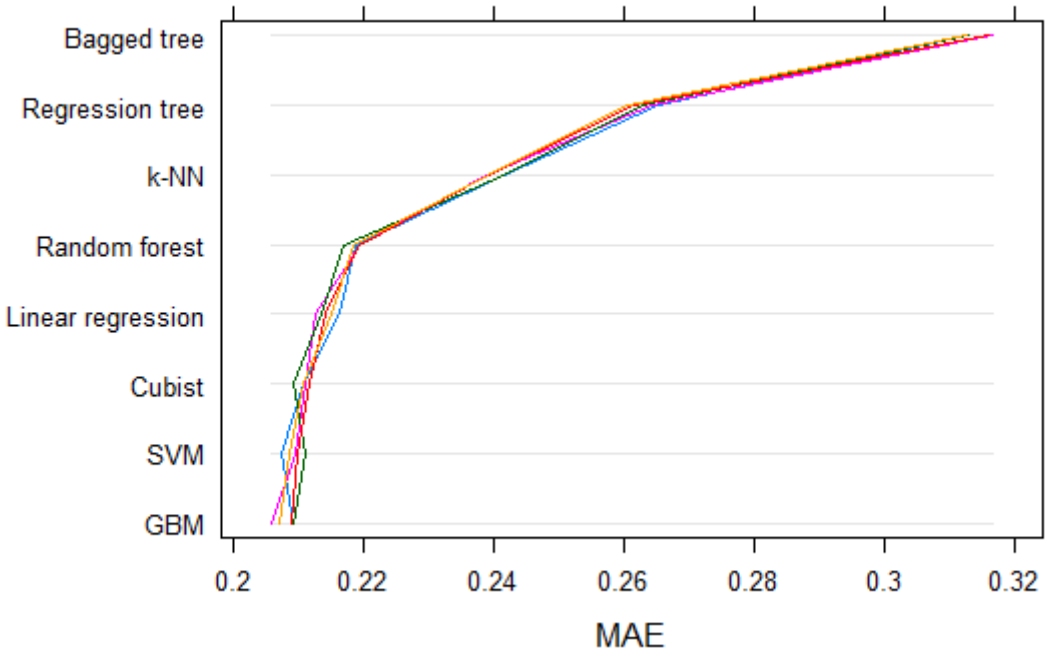
regressiemodel lager zijn. Dit is het voordeel van kennis te bezitten over het domein als onderzoeker, men kan een zeer makkelijk te interpreteren model verbeteren door bepaalde termen toe te voegen waarvan men weet uit de literatuur dat deze de schattingen kunnen verbeteren.[11]

De regressie specificatie die we hier gebruikten voor het OLS model ziet eruit als volgt:

$$\begin{aligned}\ln(prijs) = {} & \beta_0 + \beta_1 \ln(oppervlakte\ perceel) + \beta_2 \ln(oppervlakte\ bebouwd) \\ & + \beta_3 \ln(nuttige\ oppervlakte) + \beta_4\ drie\ gevels + \beta_5\ vier\ gevels \\ & + \beta_6\ lage\ kwaliteit + \beta_7\ gemiddelde\ kwaliteit + \beta_8\ leeftijd + \beta_9\ leeftijd^2 \\ & + \beta_{10}\ geen\ renovatie + \beta_{11}\ jaren\ sinds\ renovatie \\ & + \beta_{12}\ jaren\ sinds\ renovatie^2 + \beta_{13}\ \#\ verdiepingen + \beta_{14}\ \#\ verdiepingen^2 \\ & + \beta_{15}\ \#\ kamers + \beta_{16}\ \#\ kamers^2 + \beta_{17}\ \#\ garages + \beta_{18}\ \#\ garages^2 \\ & + \beta_{19}\ \#\ badkamers + \beta_{20}\ \#\ badkamers^2 + \beta_{21}\ zolder + \beta_{22}\ cv \\ & + \beta_{23}\ schattingsfout\ buren + A + J + \varepsilon\end{aligned}$$

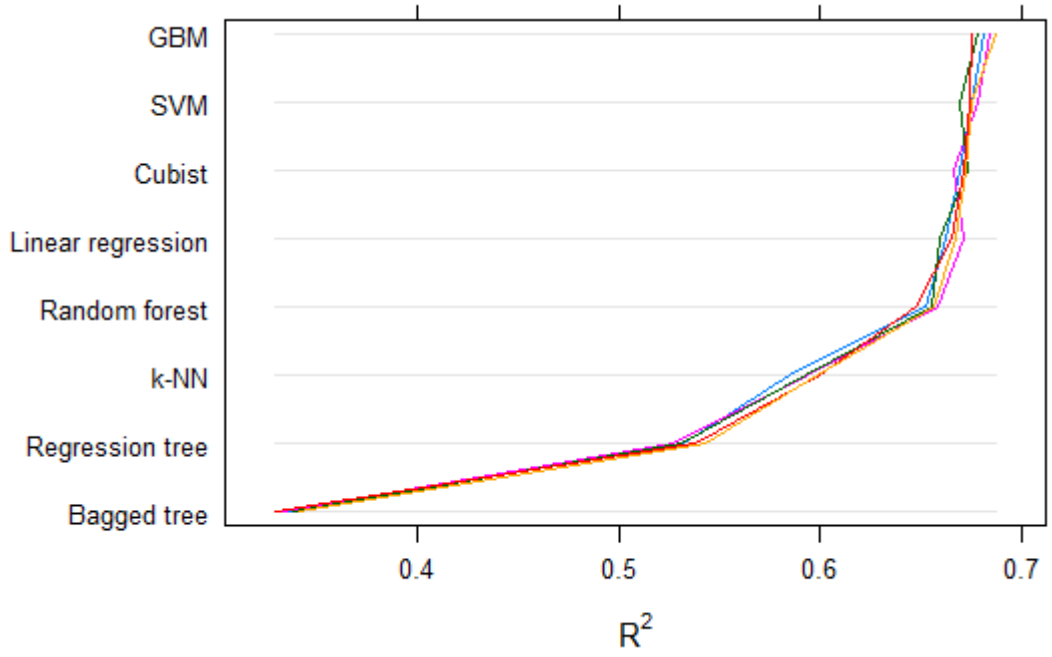
In bovenstaande formule geven A en J de fixed effects voor locatie en tijd aan. Meer specifiek maken we gebruik van arrondissement en jaar FE. We houden bovendien rekening met de tien dichtstbijzijnde buren, een aantal dat we in het vervolg van dit hoofdstuk nog zullen onderwerpen aan een test om uit te maken of dit inderdaad het meest optimaal is.

Aangezien het lineaire regressiemodel ongeveer even goed presteert als meer complexe modellen zoals het GBM model hebben we beslist we om hieronder verder te gaan met het lineaire regressiemodel. Dit heeft als grote voordeel dat de coëfficienten interpreteerbaar zijn en het systeem werkt bovendien gelijkaardig aan de huidige berekening van het KI.



Figuur 14: De MAE van de geteste modellen voor elke fold van de 5-fold cross-validation wanneer deze als test dataset gebruikt werd

[11] Bij de niet-lineaire modellen hebben we dit niet moeten doen aangezien dit via de algoritmes automatisch gebeurd.



Figuur 15: De R^2 van de geteste modellen voor elke fold van de 5-fold cross-validation wanneer deze als test dataset gebruikt werd

3.2 Vergelijking van specificaties van het lineair model

Als we het lineaire regressiemodel zoals hierboven beschreven toepassen verkrijgen we een MAE van 0.21. In dit hoofdstuk zullen we naar deze performantie als de 'baseline' performantie verwijzen en gaan we kijken of we dit kunnen verbeteren door de specificatie verder te verfijnen. Ook geven we aan hoe belangrijk bepaalde variabelen zijn door de baseline performantie te vergelijken met de verkregen performantie zonder die bepaalde variabele.

Ten eerste daalt de baseline performantie wanneer we uitsluitend lineaire termen gebruiken, de MAE stijgt namelijk naar 0.22. Let op, als de MAE stijgt betekent dit dat de performantie daalt, gezien de MAE de grootte van de fout weergeeft. Bijgevolg zijn de kwadratische en logaritmische termen hierboven besproken dus niet zo heel belangrijk voor de performantie.

Als we nu terug het baseline model nemen zien we dat de MAE opnieuw stijgt naar 0.22 indien we geen jaar FE opnemen. Indien we in plaats van fixed effects een lineaire tijdstrend gebruiken blijft de MAE op 0.21 steken.[12] Het verschil tussen fixed effects en een lineaire trend is in feite dat bij fixed effects elk jaar een differentieel effect op de prijs kan hebben, terwijl bij een lineaire trend elk jaar hetzelfde effect zal hebben op de prijs. De beperkte bijdrage van het jaar van verkoop aan de performantie kan verklaard worden door de relatief korte tijdsperiode van de steekproef, ongeveer 10 jaar, die we hanteren. In een steekproef over een langere termijn kunnen deze wel belangrijk worden.

Indien we geen arrondissement FE opnemen stijgt de MAE naar 0.28. Dit is dus een zeer belangrijke variabele, want het weglaten van deze variabele heeft een enorm effect op de performantie. De

[12] Dit is goed nieuws, indien men het model zou willen gebruiken om een schatting te maken voor een jaar waarover er nog geen data beschikbaar zijn. In de praktijk kan het bijvoorbeeld nodig zijn om ooit een schatting te maken voor een jaar waarvoor we nog niet over de verkopen beschikken. Fixed effects zijn in dat geval niet mogelijk, maar een lineaire tijdstrend is dat wel.

schattingsfout van de buren brengt dan weer eerder een beperkte verbetering met zich mee. de MAE stijgt slechts naar 0.22 wanneer we deze variabele laten vallen. Belangrijk om op te merken is dat we pas hieronder het aantal optimale buren gaan testen. We werken momenteel daarom nog met (mogelijks) een suboptimale variabele, waardoor deze daling in performantie een onderschatting kan geven van het belang van deze variabele.

Wanneer we gemeente FE in plaats van arrondissement FE gebruiken in het model stijgt de performantie: we zien een daling in de MAE naar 0.19. Opnieuw een zeer significante stijging in de performantie. Statistische sector fixed effects, een nog nauwkeurigere indicatie van de locatie van de woning, heeft het tegenovergestelde effect: de MAE stijgt opnieuw naar 0.20. Deze daling in de performantie is een duidelijk voorbeeld van overfitting, zoals hierboven reeds besproken. Ook de afstand tot het centrum toevoegen, naast de gemeente fixed effects, verbetert de performantie niet. De MAE blijft steken op 0.19.

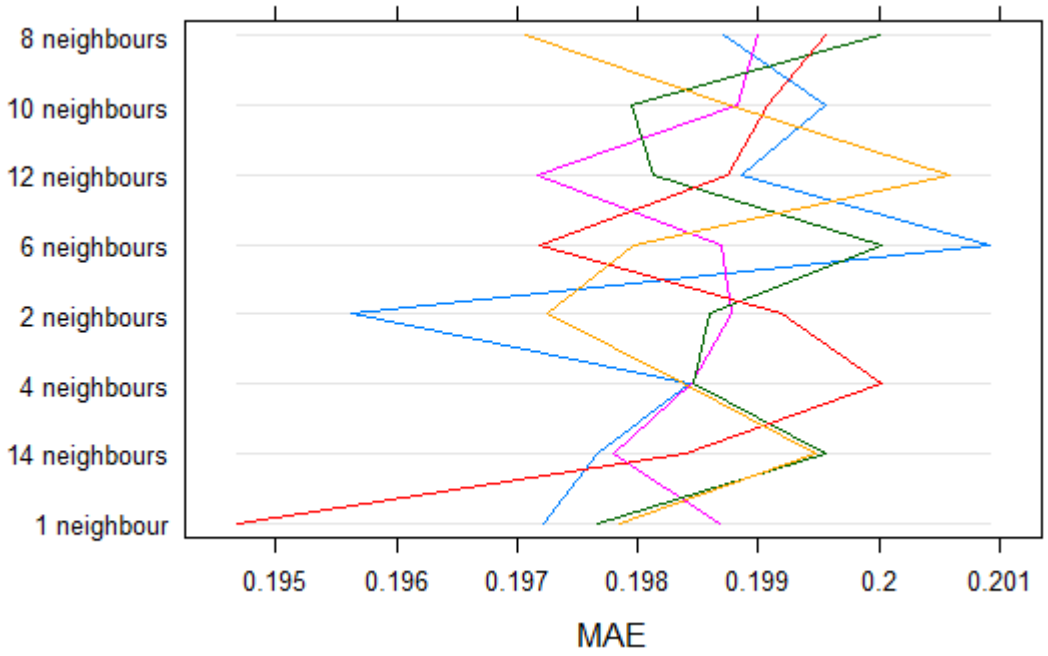
We kunnen dus concluderen dat de gemeente FE en de schattingsfout van de buren elkaar aanvullen wanneer nodig. Daar waar het locatie-effect binnen een gemeente sterk varieert zijn meestal meer verkopen, aangezien we dan meestal spreken over grotere steden. In dat geval is de afstand tot de buren kleiner en onze schattingsfout dus nauwkeuriger voor een bepaald huis. Daar waar minder verkopen zijn en de schattingsfout dus bepaald wordt door huizen die verder van het te schatten huis afliggen, zijn gemeente FE dan vaak weer specifiek genoeg aangezien we over voornamelijk kleinere gemeenten spreken in dat geval waar er minder variatie is binnen hetzelfde dorp.

3.3 Sensitiviteit ten opzichte van constructie locatievariabele

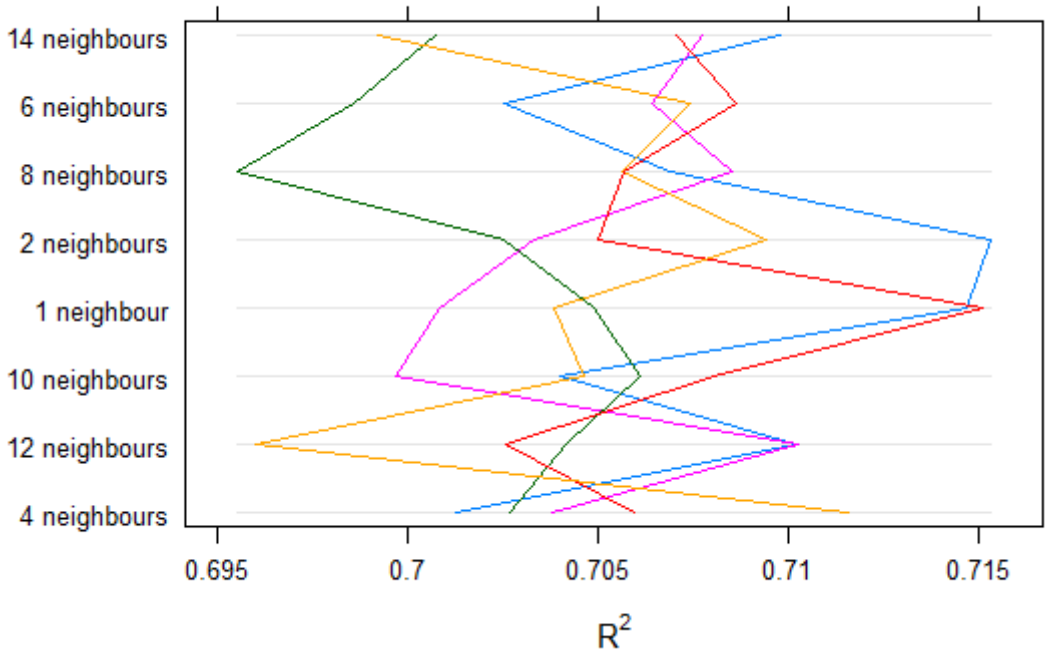
Dan rest er ons nu nog te verifiëren of we de schattingsfout van de buren het best kunnen capteren met de tien dichtstbijzijnde buren dan wel meer of minder buren. Figuren 16 en 17 tonen aan dat de performantie niet noemenswaardig stijgt of daalt indien we meer of minder buren overwegen. In het geval van woonhuizen kunnen we dus besluiten dat de tien dichtstbijzijnde buren, zoals in de schattingen hierboven, optimaal is.

Het is hierbij nogmaals belangrijk om op te merken dat we de schattingsfouten van de buren wegen op basis van de afstand tot het betrokken huis.[13] Door deze weging wordt de performantie automatisch minder beïnvloed door het exacte aantal buren. Meer buren toevoegen zal een beperkt effect hebben op de gemiddelde gewogen schattingsfout aangezien het gewicht van de extra buren per definitie lager is dan de buren die we reeds in rekening namen.

[13] De “kernel”, een functie die de manier van weging aangeeft, wordt bepaald a.d.h.v. de beste performantie. De kernel kan dus verschillen naargelang het onroerend goed dat we behandelen. De uniforme kernel is daarom eigenlijk ook een optie, deze functie zal alle observaties even zwaar wegen en daarom zal het effect niet verminderen naargelang de afstand toeneemt. Deze kernel was echter in geen enkel geval optimaal en werd daarom ook niet gebruikt.



Figuur 16: De MAE van het lineaire regressiemodel voor een variërend aantal buren dat we in rekening nemen voor de gemiddelde gewogen schattingsfout



Figuur 17: De R^2 van het lineaire regressiemodel voor een variërend aantal buren dat we in rekening nemen voor de gemiddelde gewogen schattingsfout

3.4 Toevoeging van EPC data

Voor ongeveer 54% van de transacties beschikken we over de EPC-fiche van het huis en bijgevolg de EPC-score. Om de te zien of de EPC-score de schattingen van de verkoopprijzen nauwkeuriger maakt moeten we eerst nagaan hoe het model zonder EPC-score presteert in deze kleinere dataset. Daarom schatten we exact hetzelfde model zoals hierboven beschreven op deze beperktere dataset. Met 0.18 ligt de MAE in deze dataset iets lager dan de MAE voor de gehele transactie dataset. We gebruiken deze MAE als de baseline voor de beperktere dataset.

Het toevoegen van de EPC-score aan het model verbetert de performantie van het model aanzienlijk: de MAE daalt van 0.18 naar 0.16. Als we het geschatte energieverbruik opnemen in plaats van de EPC score zorgt dit voor een kleinere stijging in performantie. Het energieverbruik toevoegen aan het model met de EPC score verbetert de performantie niet verder. Dit is opnieuw niet geheel onverwacht, de EPC-score is het meest zichtbaar voor potentiële kopers waardoor zij er wellicht meer rekening mee houden dan het totaal geschatte energieverbruik. Daarentegen is het financieel net wel belangrijker voor kopers om na te denken over toekomstige energiekosten. Volgens ons model wordt dit dus niet of slechts beperkt overwogen wanneer kopers en verkopers een prijs onderhandelen. In een vervolgonderzoek zal er trouwens bekeken worden of andere variabelen in de EPC-dataset de prestatie nog verder kunnen verbeteren.

De MAE van 0.16 is de laagste die we voor woonhuizen in dit rapport verkrijgen. Daarom is het daarom nu interessant om deze maatstaf te vertalen naar de gemiddelde absolute schattingsfout uitgedrukt in euro. Dit doen we door de exponentieel te nemen van het natuurlijk logaritme van de geschatte prijs, om daarna deze waarde af te trekken van de geobserveerde prijs en uiteindelijk het gemiddelde te nemen van de absolute waarde van al deze getallen. De MAE van 0.16, uitgedrukt in het natuurlijk logaritme, komt overeen met een gemiddelde absolute schattingsfout van EUR 37,710 voor een gemiddelde verkoopprijs gelijk aan EUR 243,947.

3.5 Out-of-sample predictie met betrekking tot moment van verkoop

In deze sectie bekijken we een andere vorm van out-of-sample predictie, namelijk hoe het model presteert als we de dataset opdelen in de transacties voor 2018 en deze na 2018. Deze manier van werken komt overeen met hoe het vaak in de praktijk gebeurt: men heeft verkopen tot een bepaalde tijdsperiode en men wenst de verkoopprijs te voorspellen van verkopen in het heden of in de toekomst. In deze setting is deze methode echter niet noodzakelijk, men kan ook op het einde van het jaar de prijzen schatten voor het afgelopen jaar. Wel is het interessant om de performantie van het model in deze setting te kennen, het kan namelijk interessant zijn om via deze methode bv. de verwachte belastinginkomsten te voorspellen.

Het is hierbij wel belangrijk om aan te geven dat de verkregen performantie in deze sectie in “speciale” gevallen niet geldig zal zijn. De nauwkeurigheid van de schattingen is namelijk afhankelijk van recente gebeurtenissen in de woningmarkt. Indien in het begin van 2018 een crisis in de woningmarkt plaatsvond verkrijgen we te hoge schattingen, gezien het model nog geen rekening kan houden hiermee. Desalniettemin, gebeurtenissen op nationaal niveau die zorgen da de woningmarkt de trend van de voorbije jaren volledig doorbreek komen zelden voor en daarom is deze vorm van performantie nog altijd interessant om te kennen.

Indien we de volledige dataset gebruiken, waardoor we de EPC-score dus niet kunnen opnemen, bekomen we een MAE van 0.20. Zoals hierboven reeds gezegd moeten we in dit geval wel gebruik maken van een lineaire tijdstrend in plaats van jaar fixed effects. Gegeven dat er hierboven ook al een stijging in de MAE van 0.01 was indien we deze verandering invoerden kunnen we besluiten dat het model in deze vorm van out-of-sample predictie even goed voorspelt. Wanneer we deze MAE, uitgedrukt in het natuurlijk logaritme van de prijs, omzetten naar de euro-waarde verkrijgen we een

gemiddelde absolute schattingsfout van EUR 48,608. De schattingsfout daalt niet indien we de euro-waarde als uitkomstvariabele gebruiken in plaats van het natuurlijk logaritme. Hieruit merken we op dat de schattingsfout nog altijd groot is. Of het model te grote schattingsfouten produceert om te gebruiken voor de voorspelling van de verkoopprijs en uiteindelijk ook de belastbare basis zal moeten blijken uit een vergelijking met andere opties die voorhanden zijn en de huidige relatie tussen het KI en de huur.

Indien we de beperktere dataset gebruiken waarvoor de EPC-fiche beschikbaar is, bekomen we een MAE van 0.16. Dit is exact dezelfde waarde (zonder afrondingsfouten in rekening te nemen) die we hierboven ook al vonden. Dit komt echter overeen met een grotere gemiddelde absolute schattingsfout van EUR 42,707 indien we de waarde in euro uitdrukken. Een sterke daling in de fout, maar uiteraard is ook dit nog een hoge waarde. Het is waarschijnlijk deze waarde die uiteindelijk vergeleken dient te worden met de andere beschikbare opties. Het zou naar onze mening zonde zijn om deze waardevolle informatie niet te gebruiken in de toekomst, zeker wanneer meer en meer huizen binnenkort over een EPC-fiche zullen beschikken.

3.6 Beschrijving coëfficienten

In deze sectie zullen we de coëfficiënten van de regressie interpreteren alsook enkele opvallende resultaten bespreken. De eerste kolom van Tabel 1 geeft de resultaten weer van de regressie op de volledige sample, terwijl de tweede kolom de resultaten toont van de regressie op de beperktere dataset waarvoor we de EPC-score hebben en deze ook hebben toegevoegd aan het model. Het model in de tweede kolom bevat meer nuttige variabelen en geeft dan ook de meest betrouwbare coëfficiënten, maar in de tekst zullen we beiden bespreken indien dit nuttige inzichten met zich meebrengt of om een interval aan te geven. Belangrijk is op te merken dat de afhankelijke variabele telkens het natuurlijk logaritme van de verkoopprijs is, dit zal de interpretatie van onze coëfficiënten beïnvloeden.

Ook worden in het vervolg van dit hoofdstuk de coëfficiënten van het jaar van verkoop niet besproken, maar we willen hier wel opmerken dat deze zoals verwacht stijgen doorheen de jaren. Meer specifiek stijgt de coëfficiënt van 0.034 in 2007 naar 0.467 in 2020 indien we de volledige dataset bekijken. Tussen 2007 en 2020 is er bovendien elk jaar een stijging merkbaar in de coëfficiënt. Het referentiejaar is hier 2006, deze coëfficiënt stellen we gelijk aan nul. De tendens is gelijkaardig indien we voor de EPC-score controleren, maar de gemiddelde grootte van de coëfficiënten is kleiner. Dit kan verklaard worden door de EPC-score die doorheen de jaren daalt en mee de verkoopprijs positief beïnvloedt.

In de tweede rij van tabel 1 zien we dan dat de prijs met 1.2% tot 1.4% zal stijgen indien de oppervlakte van het perceel met 10% stijgt. De derde rij toont dan weer aan dat een stijging in de bebouwde oppervlakte een negatief effect zal hebben op de prijs, deze zal namelijk met 0.33% tot 0.96% dalen indien de bebouwde oppervlakte met 10% stijgt. Het negatief effect ligt in lijn van de verwachtingen, want als de bebouwde oppervlakte toeneemt en de totale perceel- en woonoppervlakte constant blijft zal de oppervlakte van de tuin kleiner worden. De vierde rij toont aan dat de prijs zal stijgen met 3.5% tot 4.7% indien de nuttige oppervlakte stijgt met 10%. Van de drie verschillende oppervlaktes, heeft de nuttige oppervlakte dus het grootste effect op de verkoopprijs.

In de zesde en zevende rij zien we dan dat een halfopen bebouwing gemiddeld 1.5% tot 4.5% duurder is dan een gesloten bebouwing, terwijl uit de achtste en negende rij blijkt dat een open bebouwing gemiddeld 9% tot 13% duurder is dan een gesloten bebouwing. In vergelijking met een woonhuis met hoge kwaliteit is een woonhuis met lage kwaliteit 9.4% tot 12.1% goedkoper, terwijl een woonhuis met gemiddelde kwaliteit 10.2% tot 12.5% goedkoper is. Een woonhuis met een lage kwaliteit zou volgens het model dus duurder zijn dan een woonhuis met gemiddelde kwaliteit. De kwaliteitsvariabele is echter niet veelzeggend aangezien er maar 1% van de woningen een lage of hoge kwaliteit zou hebben.

We besteden daarom niet al te veel aandacht aan de exacte coëfficiënten van de kwaliteitsfactoren, want deze zijn door de kenmerken van de onderliggende variabele niet betrouwbaar.

Tabel 1: Coëfficienten van het regressiemodel

	(1) België Ln(Prijs)	(2) Vlaanderen met EPC score Ln(Prijs)	(3) Vlaanderen zonder EPC score Ln(Prijs)
Constante	9.66382***	10.32026***	10.11502***
	(0.02831)	(0.01099)	(0.01420)
Ln(Oppervlakte perceel)	0.12486***	0.14310***	0.13099***
	(0.00080)	(0.00076)	(0.00093
Ln(Oppervlakte bebouwd)	-0.09579***	-0.03337***	-0.08361***
	(0.00366)	(0.00182)	(0.00244)
Ln(Nuttige oppervlakte)	0.46687***	0.34697***	0.43511***
	(0.00571)	(0.00234)	(0.00332)
Twee gevels (ref. cat.)	-	-	-
Drie gevels	0.01495***	0.04500***	0.01093***
	(0.00139)	(0.00111)	(0.00122)
Vier gevels	0.09049***	0.13064***	0.08590***
	(0.00216)	(0.00154)	(0.00173)
Lage kwaliteit	- 0.09445***	-0.12139***	-0.13366***
	(0.00931)	(0.00832)	0.01156
Gemiddelde kwaliteit	-0.10195***	-0.12510***	-0.12796***
	(0.00530)	(0.00551)	(0.00667)
Hoge kwaliteit (ref. cat.)	-	-	-
Leeftijd	-0.00254***	-0.00388***	-0.00739***
	(0.00058)	(0.00004)	(0.00005)
Leeftijd²	0.000005	0.00002***	0.00004***
	(0.000003)	(0.0000002)	(0.0000003)
Geen renovatie	-0.12680***	-0.09042***	-0.16598***
	(0.00242)	(0.00215)	(0.00233)
Jaren sinds renovatie	-0.00459***	-0.00362***	-0.00622***
	(0.00009)	(0.00028)	(0.00031)
Jaren sinds renovatie²	0.000002***	0.00001	0.00003***
	(0.0000002)	(0.000008)	(0.000009)
# Verdiepingen	-0.06708***	-0.05260***	-0.03560***
	(0.00688)	(0.00372)	(0.00430)
# Verdiepingen²	0.01605***	0.01523***	0.01072***
	(0.00148)	(0.00096)	(0.00111)
# Kamers	-0.02114***	-0.01249***	-0.01217***
	(0.00207)	(0.00165)	(0.00205)
# Kamers²	0.00162***	0.00147***	0.00129***
	(0.00014)	(0.00013)	(0.00017)
# Garages	-0.01182***	-0.00874***	-0.03116***
	(0.00149)	(0.00151)	(0.00176)
# Garages²	0.00645***	0.00390***	0.01200***
	(0.00067)	(0.00071)	(0.00082)
# Badkamers	0.00574***	-0.00427**	-0.01627***
	(0.00192)	(0.00197)	(0.00255)
# Badkamers²	0.01371***	0.01345***	0.02002***
	(0.00134)	(0.00121)	(0.00161)
Zolder	0.03799***	0.01218***	0.02662***
	(0.00110)	(0.00099)	(0.00112)
Centrale verwarming	0.10642***	0.04485***	0.07440***
	(0.00240)	(0.00102)	(0.00112)
Schattingsfout buren	0.56588***	0.69879***	0.69432***
	(0.01433)	(0.00329)	(0.00420)
EPC-score	-	-0.00048***	-
		(0.000002)	
Obs.	719,368	392,331	392,331
R-kwadraat	0.70	0.73	0.68
FE jaar van verkoop	Ja	Ja	Ja
FE gemeente	Ja	Ja	Ja

Standaardfout tussen haakjes

*** p<0.01, ** p<0.05, * p<0.1

De 11de rij toont aan dat de verkoopprijs daalt met 0.3% tot 0.4% voor elk jaar dat de woning ouder wordt. Merk ook op dat het effect minder uitgesproken is in de tweede kolom ten opzichte van de derde kolom. Dit voorbeeld toont duidelijk aan dat de EPC-score toevoegen zorgt voor minder bias in de andere coëfficiënten en je de EPC-score daarom moet toevoegen indien je hiervoor wilt kunnen controleren. Het model zonder EPC-score heeft een sterker negatief effect van leeftijd op prijs omdat de leeftijd gecorreleerd is met de EPC-score. Met andere woorden, de coëfficiënt van leeftijd neemt in dit model zowel het effect van leeftijd als (deels) dat van EPC op. De geïnteresseerde lezer kan in de tabel hierboven vast nog enkele andere voorbeelden hiervan vinden.

De volgende rij toont ook wel de coëfficiënt van de kwadratische term van leeftijd, maar gezien deze zo klein is kunnen we deze coëfficiënt negeren voor nieuwere woningen. In de eerste kolom is de coëfficiënt zelfs insignificant. Voor oudere woningen zal deze term wel van belang zijn, aangezien het effect kwadratisch stijgt. Hier illustreren we dit effect even in het geval van de tweede kolom. Voor een woning die bijvoorbeeld 100 jaar oud is zal deze term gelijk zijn aan 20% ($100^2 * 0.00002$), terwijl de lineaire term gelijk zal zijn aan -40% (100 * 0.004). Als we beide termen optellen komen we dus uit op -20%. Een woning van 100 jaar oud zal dus gemiddeld 20% goedkoper zijn in vergelijking met nieuwbouw, indien de overige karakteristieken hetzelfde zijn. We kunnen hieruit dus concluderen dat er door de kwadratische term afnemende marginale waardevermindering optreedt, waarbij voor heel oude woningen het effect van leeftijd zelfs omgekeerd wordt. Mogelijks willen mensen meer geld betalen voor het karakter van een oudere woning.

De 13de rij toont aan dat een woning zonder renovatie gemiddeld 9% tot 12.7% goedkoper is dan een woning die minstens één renovatie achter de rug heeft. Het is evenals van belang hoe lang geleden de renovatie gebeurde. Voor elk jaar dat verstrijkt sinds de renovatie zal de verkoopprijs dalen met 0.4% tot 0.5%, zoals weergegeven in de volgende rij. De kwadratische term in de 15de rij is opnieuw heel klein en aangezien een renovatie recenter plaats vindt dan het bouwjaar van een huis is deze term veel minder relevant dan bij de leeftijd van een huis. De berekening gebeurt wel hetzelfde als in het voorbeeld van het 100-jarig huis, dus de geïnteresseerden kunnen dit zelf berekenen. Deze keer is de coëfficiënt wel significant in de eerste kolom, maar niet in de tweede kolom.

Uit de 16de rij kunnen we afleiden dat een extra verdieping een negatief effect heeft op de prijs, maar de 17de rij toont dan weer een positief kwadratisch effect. Voor een woonhuis met 3 verdiepen geldt dan dat de eerste term gelijk is aan -15.9%, terwijl de kwadratische term gelijk is aan 4.5% wat gecombineerd gelijk is aan -11.4% (indien men de coëfficiënten van de tweede kolom gebruikt). Voor het aantal kamers geldt eenzelfde verhaal, de lineaire term is negatief terwijl de kwadratische term positief is.

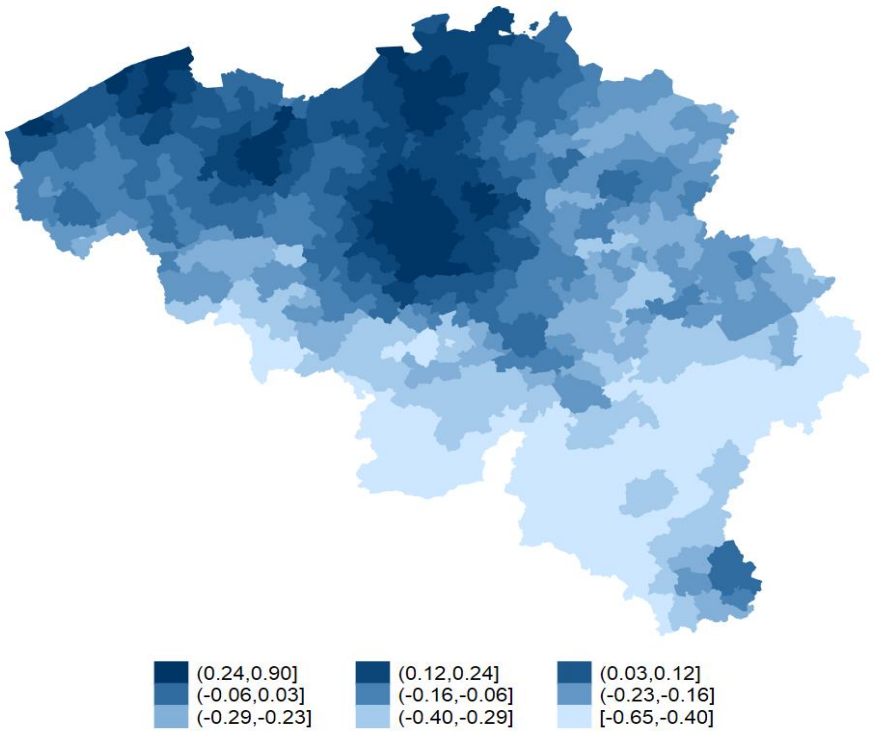
Voor het aantal garages geldt daarentegen het omgekeerde: de lineaire term is negatief, terwijl de kwadratische term positief is. Bij een laag aantal garages zal een extra garage daarom een negatief effect hebben op de verkoopprijs, maar voor een hoog aantal garages zal een extra garage een prijsverhogend effect kennen. Deze bevinding is op het eerste zicht misschien onverwacht. Een mogelijke verklaring is dat een extra garage zorgt voor minder woonoppervlakte, terwijl woonoppervlakte wel belangrijk is voor woonhuizen die niet veel garages hebben. Woonhuizen met veel garages zullen daarentegen doorgaans geen problemen hebben omtrent woonoppervlakte. Dit zijn waarschijnlijk zeer luxueuze huizen, waarbij een extra garage altijd handig kan zijn. Ook zal deze variabele deels de luxe van het woonhuis opvangen aangezien we geen gedetailleerde variabelen hebben die luxe goed kunnen capteren. De kwadratische term zal daarom een deel van het effect van luxe opvangen. Hetzelfde verhaal geldt voor het aantal badkamers, tenminste als we de tweede kolom bekijken.

De 24ste rij toont aan dat een zolder een positief effect heeft op de prijs, meer specifiek verhoogt het de verkoopprijs met 1.2% tot 3.8%. Een centrale verwarming verhoogt de prijs dan weer met 4.5% tot 10.6%. Het is interessant dat de coëfficiënt in de eerste kolom zoveel hoger is voor de centrale

verwarming in vergelijking met de tweede kolom, hoogstwaarschijnlijk een effect van het niet accuraat opnemen van de kwaliteit van de woning in het model. In de tweede kolom doen we dit beter door de toevoeging van de EPC-score, die naast de energiezuinigheid van een woning ook een maatstaf is voor de kwaliteit. Zoals hierboven gezegd zullen variabelen zoals het aantal garages ook deels de kwaliteit van een woning kunnen capteren, maar wij geloven dat de EPC-score een betere maatstaf voor de kwaliteit is. Ook deze is uiteraard niet perfect.

De coëfficiënt van de schattingsfout van de buren, weergegeven in de 26ste rij, is gelijk aan 0.57 tot 0.70. Deze schattingsfout is in het natuurlijk logaritme uitgedrukt, aangezien we ook het logaritme van de prijs gebruikten in het initiële model waarvan we de schattingsfout gebruiken. Deze coëfficiënt kunnen we dus op een gelijkaardige manier als de oppervlakte coëfficiënten interpreteren. Meer specifiek, een stijging van 1% in de schattingsfout van de buren zal resulteren in een positief prijseffect van 0.57% tot 0.70% voor het desbetreffende huis. Hieruit blijkt dus dat een groot deel van de schattingsfout in het initiële model, waarin geen locatie variabelen buiten gemeente fixed effects werden gebruikt, ontstaat uit liggingseffecten die niet goed gecapteerd werden. Dit positief verband ligt daarom geheel in lijn van de verwachtingen. Als laatste heeft de EPC-score een negatief effect op de prijs, een hogere EPC-score zal dus leiden tot een lagere prijs. Merk op dat een hogere EPC-score een slechtere energie-efficiëntie van de woning betekent, waardoor het negatieve verband dus is zoals verwacht.

Figuur 18 hieronder geeft uiteindelijk de gemeente fixed effects weer voor alle verschillende Belgische gemeenten op een kaart. Hoe donkerder de kleur, hoe duurder een bepaalde regio. De kust en Brugge, Brussel, Antwerpen, Gent, Leuven en de omgeving nabij de grens met Luxemburg zijn het duurst. De Ardennen is daarentegen ruim de goedkoopste regio van België. Limburg is over het algemeen de goedkoopste regio van Vlaanderen, maar in vergelijking met grote delen van Wallonië is Limburg iets duurder. In Wallonië is het dan weer duidelijk dat Waals-Brabant er bovenuit steekt met hogere prijzen. Deze bevindingen komen overeen met voorgaande studies omtrent huisprijzen in België. In de bijlagen op het einde van dit rapport tonen we dezelfde kaart indien we de EPC-score opnemen.



Figuur 18: De gemeente fixed effects weergegeven op de kaart van België, waarbij een donkerdere kleur een hogere liggingscoëfficient weergeeft

4. PRIJSSCHATTING VAN APPARTEMENTEN

4.1 Vergelijking van modellen

In dit hoofdstuk zullen we hetzelfde doen als in het vorige hoofdstuk, maar deze keer schatten we de verkoopprijs van appartementen in plaats van woonhuizen. We beginnen weer met het vergelijken van verschillende predictiemodellen, om daarna in de volgende sectie te kijken of het aantal buren dat we in rekening nemen belangrijk is. Uiteindelijk schatten we de prestatie van het best presterende model in een realistische setting zoals we hierboven ook deden. De andere zaken die we in het vorig hoofdstuk gedaan hebben zullen we hier weglaten om ruimte te besparen of uit noodzaak aangezien de data niet beschikbaar is voor appartementen (bv. EPC dataset).

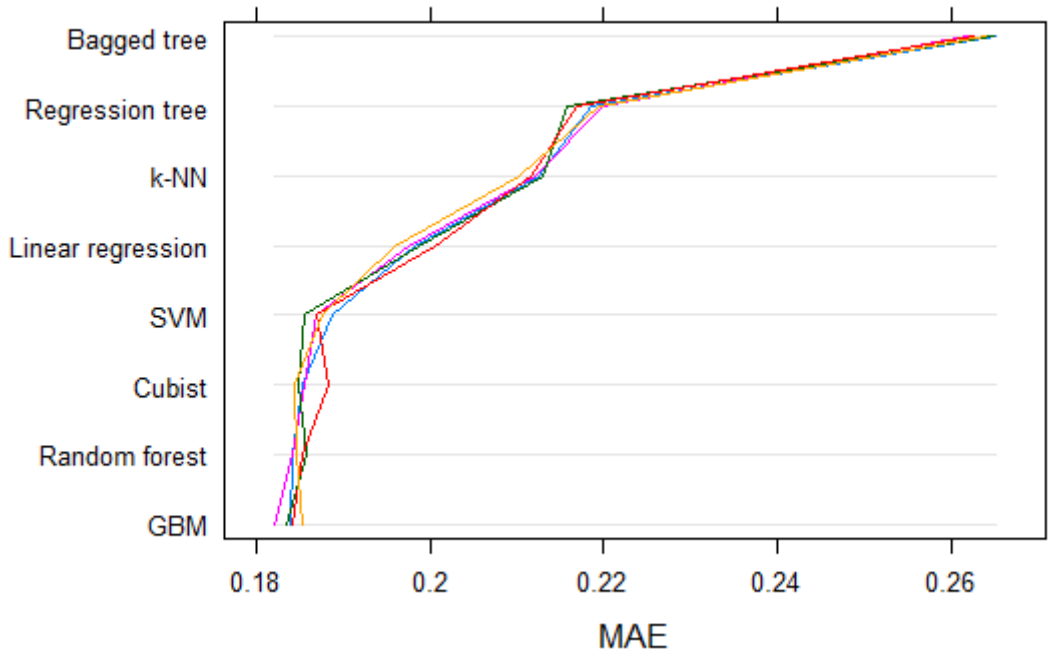
Figuren 19 en 20 tonen opnieuw de MAE en de R^2 van de verschillende modellen die we getest hebben. De gebruikte specificatie is hetzelfde als hierboven met uitzondering van feit dat we volgende variabelen niet gebruiken gezien deze niet beschikbaar zijn voor appartementen: oppervlakte van het perceel, bebouwde oppervlakte, aantal gevels, aantal verdiepingen en zolder. De regressie specificatie die we hier gebruikten voor het OLS model ziet er dus uit als volgt:

$$\begin{aligned}\ln(prijs) = {} & \beta_0 + \beta_1 \ln(nuttige\ oppervlakte) + \beta_2\ lage\ kwaliteit + \beta_3\ gemiddelde\ kwaliteit \\ & + \beta_4\ leeftijd + \beta_5\ leeftijd^2 + \beta_6\ geen\ renovatie \\ & + \beta_7\ jaren\ sinds\ renovatie + \beta_8\ jaren\ sinds\ renovatie^2 + \beta_8\ \#\ kamers \\ & + \beta_{10}\ \#\ kamers^2 + \beta_{11}\ \#\ garages + \beta_{12}\ \#\ garages^2 + \beta_{13}\ \#\ badkamers \\ & + \beta_{14}\ \#\ badkamers^2 + \beta_{15}\ cv + \beta_{16}\ schattingsfout\ buren + A + J + \varepsilon\end{aligned}$$

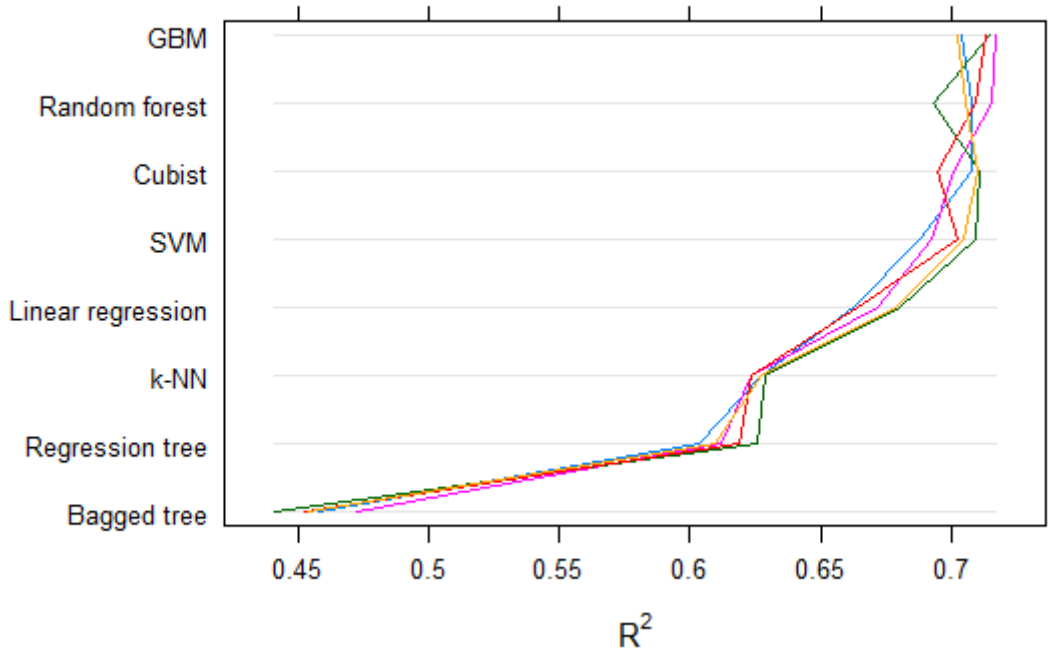
In bovenstaande formule geven A en J de fixed effects voor locatie en tijd aan. Meer specifiek maken we opnieuw gebruik van arrondissement en jaar FE. We houden bovendien rekening met de tien dichtstbijzijnde buren, een aantal dat we in het vervolg van dit hoofdstuk nog zullen onderwerpen aan een test om uit te maken of dit inderdaad het meest optimaal is.

In onderstaande figuren kunnen we zien dat opnieuw GBM, random forest, cubist en SVM het best presteren. Het OLS model presteert deze keer echter significant slechter dan deze modellen, waardoor de afweging tussen de prestatie van het OLS model enerzijds en de aantrekkelijke karakteristieken anderzijds belangrijk wordt. De MAE die we verkrijgen voor het GBM model, dat het best presteert, is gelijk aan 0.18, terwijl de MAE van het OLS model gelijk is aan 0.20.

Deze prestatie verkrijgen we met behulp van arrondissement fixed effects. De modellen met gemeente fixed effects presteren echter beter, net zoals in het vorige hoofdstuk over de prijsschatting van woonhuizen. Opnieuw verbetert de performantie niet indien we de afstand naar het centrum toevoegen. In wat volgt zullen we daarom de specificatie aanpassen naar gemeente fixed effects. Voor de rest verandert er niets aan de specificatie. Met behulp van de nieuwe verkregen specificatie zullen we nu nagaan of het aantal buren dat we in rekening nemen van belang is.



Figuur 19: De MAE van de geteste modellen voor elke fold van de 5-fold cross-validation wanneer deze als test dataset gebruikt werd



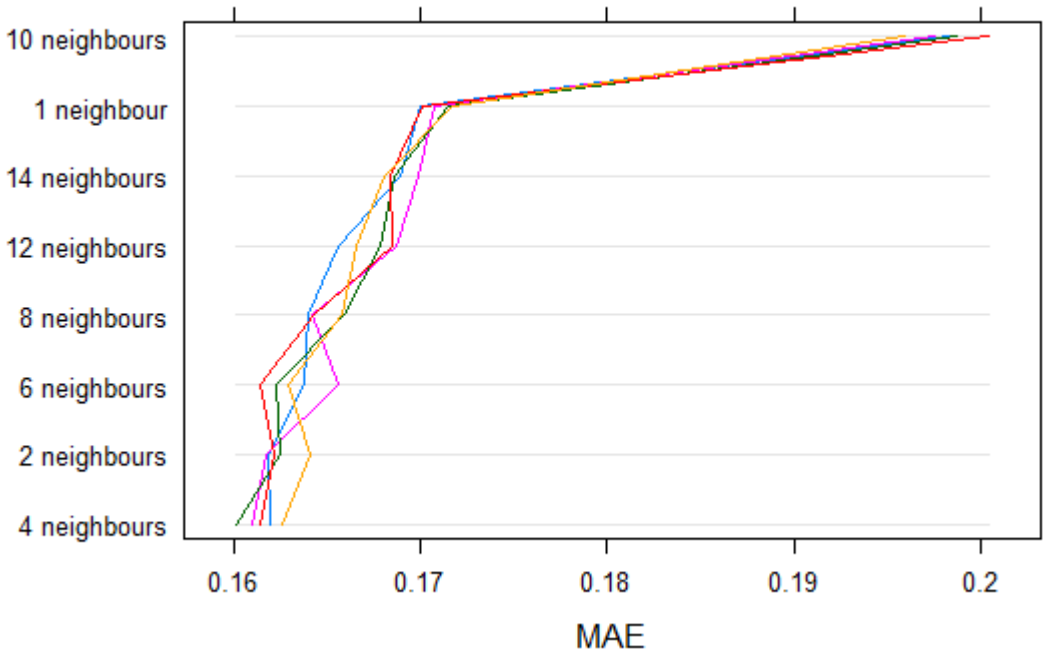
Figuur 20: De R^2 van de geteste modellen voor elke fold van de 5-fold cross-validation wanneer deze als test dataset gebruikt werd

4.2 Sensitiviteit ten opzichte van constructie locatievariabele

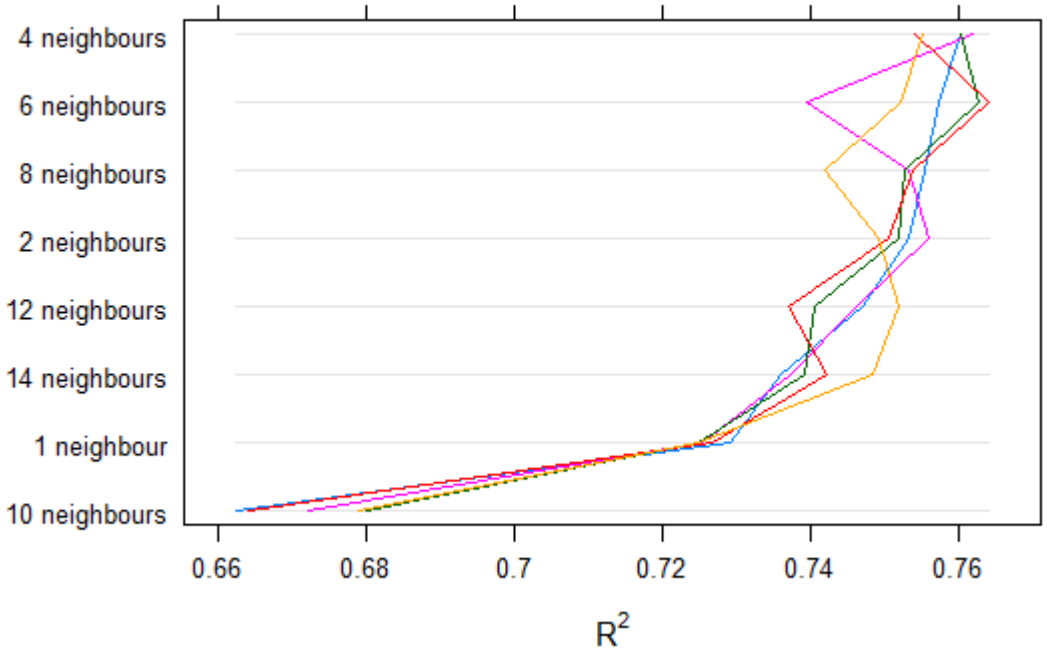
Dan rest er ons nu dus nogmaals te verifiëren of we de schattingsfout van de buren het best kunnen capteren met de tien dichtstbijzijnde buren dan wel meer of minder buren. We testen dit aan de hand van het OLS model om tijd te besparen. Met behulp van de figuren zoals we die ook verkregen in het vorige hoofdstuk zullen we dan het optimale aantal buren selecteren en dit ook toepassen op het GBM model om uiteindelijk de finale performantie te bekomen.

Figuren 21 en 22 tonen aan dat het, in tegenstelling tot in het vorig hoofdstuk, nu wel uitmaakt hoeveel buren we in rekening nemen. Over het algemeen stijgt de performantie naarmate we minder buren in rekening nemen, al daalt de performantie kritisch indien we van twee buren naar één enkele buur gaan. Het optimale aantal buren is vier, zoals blijkt uit onderstaande figuren. Het aantal buren is zelfs heel belangrijk: de baseline MAE waarbij we 10 buren in rekening nemen is gelijk aan 0.20 voor het OLS model, terwijl de MAE daalt tot 0.16 indien we maar vier buren in rekening nemen.

Als we nu dit aantal buren ook gebruiken voor het GBM model verkrijgen we een MAE van 0.15. Deze fout komt overeen met EUR 28,760 voor een gemiddelde prijs in deze dataset van EUR 184,844. De fout is dus nog altijd lager dan de verkregen MAE voor het OLS model met 4 buren, maar het verschil is wel kleiner geworden. Het is dus ook aannemelijk om voor het OLS model te gaan indien de voordelen van dit model hierboven beschreven opwegen ten opzichte van de iets lagere performantie.



Figuur 21: De MAE van het lineaire regressiemodel voor een variërend aantal buren dat we in rekening nemen voor de gemiddelde gewogen schattingsfout



Figuur 22: De R^2 van het lineaire regressiemodel voor een variërend aantal buren dat we in rekening nemen voor de gemiddelde gewogen schattingsfout

4.3 Out-of-sample predictie met betrekking tot moment van verkoop

In deze sectie bekijken we een andere vorm van out-of-sample predictie, namelijk hoe het model presteert als we de dataset opdelen in een periode voor 2018 en na 2018. Dit om een idee te krijgen over hoe het model kan schatten voor een tijdsperiode waarin er nog geen verkopen beschikbaar zijn. Zoals hierboven reeds gezegd moeten we in dit geval wel gebruik maken van een lineaire tijdstrend in plaats van jaar fixed effects.

Met deze procedure te volgen bekomen we een MAE van 0.17 voor het GBM model met vier buren, een stijging van 0.02. Gegeven dat er sowieso een stijging in de MAE te verwachten was door de lineaire tijdstrend bekomen we opnieuw een ongeveer even goede schatting. Wanneer we deze MAE, uitgedrukt in het natuurlijk logaritme van de prijs, omzetten naar de euro-waarde verkrijgen we een gemiddelde absolute schattingsfout van EUR 34,673. De schattingsfout daalt opnieuw niet indien we de euro-waarde als uitkomstvariabele gebruiken in plaats van het natuurlijk logaritme.

Hieruit merken we op dat de schattingsfout nog altijd groot is, ook al is die al kleiner geworden dan de fout die we vonden bij woonhuizen. Gezien de gemiddeld lagere prijs van appartementen dan woonhuizen is het logisch dat de schattingsfout lager is bij een gelijke performantie (uitgedrukt in MAE). De MAE zelf is echter ook lager bij de prijsschatting van appartementen, wat aangeeft dat deze prijzen inderdaad beter te schatten zijn. Of het model te grote schattingsfouten produceert om te gebruiken voor de schatting van de verkoopprijs en uiteindelijk ook de belastbare basis zal opnieuw moeten blijken uit een vergelijking met andere opties die voorhanden zijn en de huidige relatie tussen het KI en de huur.

5. PRIJSSCHATTING VAN BOUWGRONDEN

5.1 Vergelijking van modellen

Dit hoofdstuk zal dezelfde structuur volgen als het vorige, maar deze keer voor de prijs van bouwgronden in plaats van appartementen. Opnieuw starten we met de vergelijking van verschillende predictiemodellen, daarna gaan we over naar de zoektocht naar het optimaal aantal buren om vervolgens te eindigen met de predictie van verkoopprijzen in een tijdsperiode waarvoor er nog geen data beschikbaar zijn.

De specificatie van de modellen is echter sterk versimpeld, gezien we veel minder beschikbare variabelen hebben voor bouwgronden. We gebruiken nu de oppervlakte van het perceel, het arrondissement van de bouwgrond, het verkoopjaar van de bouwgrond en de schattingsfout van de buren als predictoren. Daarbovenop hebben we ook nog drie nieuwe variabelen gecreëerd die aangeven welk aandeel van de nabijgelegen gebouwen open, halfopen en gesloten is. Dit om een inschatting te krijgen over de bouwvoorschriften van de bouwgrond in kwestie. De veronderstelling is daarbij dat de kans groter is dat een bouwgrond dient voor gesloten bebouwing indien veel van de buren ook over een gesloten gebouw beschikken. Daarnaast hebben we een vergelijkbare variabele gecreëerd die aangeeft wat het gemiddelde bouwvolume is van de nabijgelegen gebouwen. Dit om opnieuw een inschatting te krijgen van de bouwvoorschriften van de omgeving en de bouwgrond in kwestie.

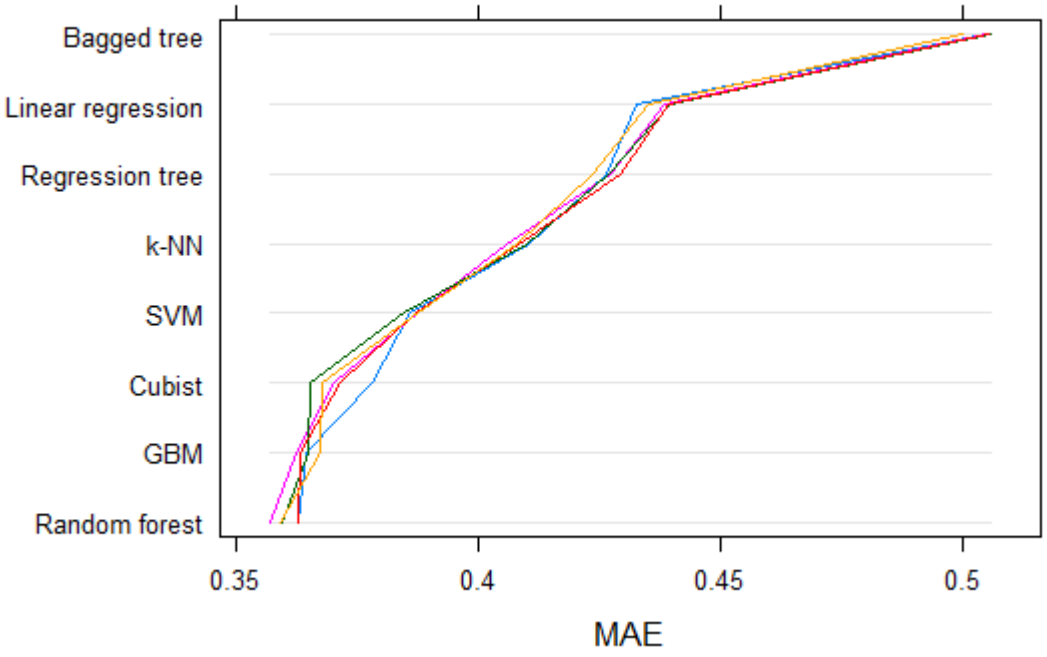
De verkregen regressie specificatie die we hier gebruikten voor het OLS model ziet er dus uit als volgt:

$$\begin{aligned}\ln(prijs) = \ & \beta_0 + \beta_1 \ln(oppervlakte\ perceel) + \beta_2 \ln(volume\ buren) \\ & + \beta_3\ prop.buren\ drie\ gevels\ + \beta_4\ prop.buren\ vier\ gevels \\ & + \beta_5\ schattingsfout\ buren + A + J + \varepsilon\end{aligned}$$

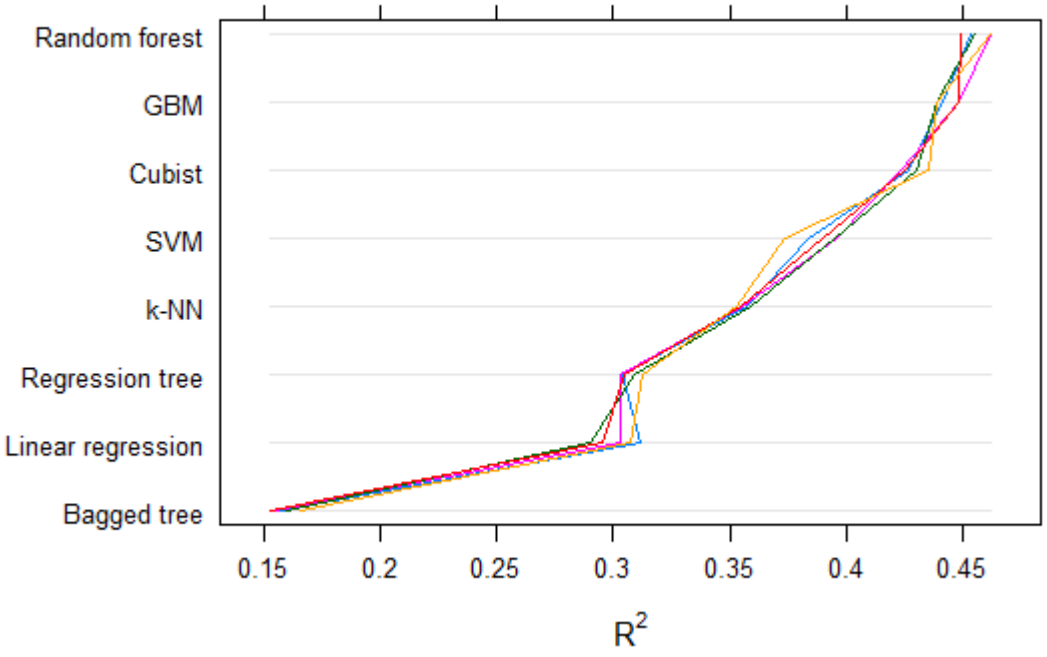
In bovenstaande formule geven A en J de fixed effects voor locatie en tijd aan. Meer specifiek maken we opnieuw gebruik van arrondissement en jaar FE. We houden bovendien rekening met de tien dichtstbijzijnde buren, een aantal dat we in het vervolg van dit hoofdstuk nog zullen onderwerpen aan een test om uit te maken of dit inderdaad het meest optimaal is. We voegen de variabele die aangeeft welk aandeel van de buren een gesloten bebouwing hebben niet toe aan de specificatie, aangezien deze overbodig is door het toevoegen van de andere twee variabelen.

Figuren 23 en 24 tonen beiden aan dat het random forest model het best presteert, op de voet gevolgd door het GBM model. Het OLS model presteert in deze situatie zeer pover en is het op één na slechtste model. In dit geval lijkt het ons daarom geen valabel model om in overweging te nemen voor praktisch gebruik.

Het random forest model presteert nog beter indien we afstand tot het centrum toevoegen, terwijl het in voorgaande hoofdstukken bij woonhuizen en appartementen geen verschil maakte. Het grote verschil is dat in het geval van bouwgronden gemeente fixed effects niet beter zijn dan arrondissement fixed effects. Dus in wat volgt voegen we enkel de afstand tot het centrum toe, de rest blijft hetzelfde. Met behulp van de nieuwe verkregen specificatie zullen we nu nagaan of het aantal buren dat we in rekening nemen van belang is.



Figuur 23: De MAE van de geteste modellen voor elke fold van de 5-fold cross-validation wanneer deze als test dataset gebruikt werd



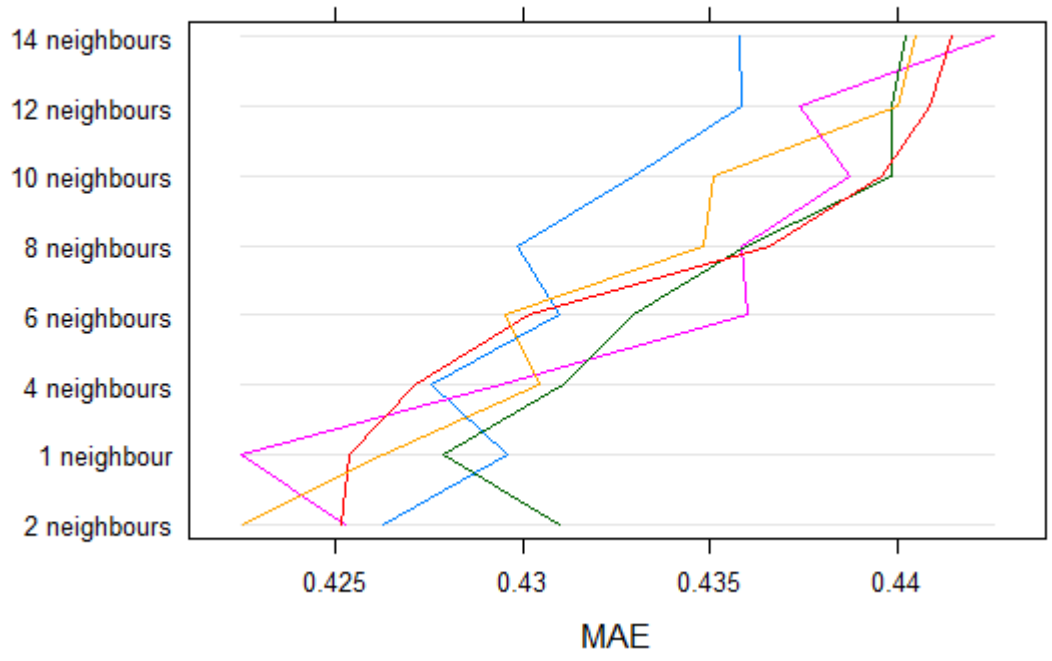
Figuur 24: De R^2 van de geteste modellen voor elke fold van de 5-fold cross-validation wanneer deze als test dataset gebruikt werd

5.2 Sensitiviteit ten opzichte van constructie locatievariabele

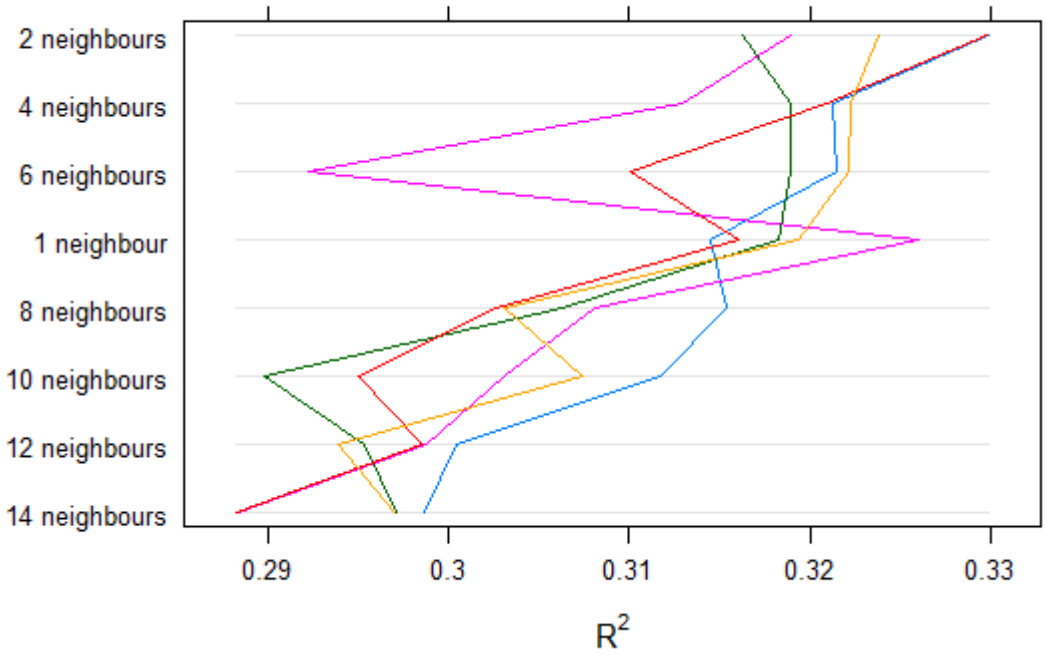
Dan rest er ons nu dus voor een laatste keer te verifiëren of we de schattingsfout van de buren het best kunnen capteren met de tien dichtstbijzijnde buren dan wel meer of minder buren. We testen dit aan de hand van het OLS model om tijd te besparen. Met behulp van de figuren zoals we die ook verkregen in het vorige hoofdstuk zullen we dan het optimale aantal buren selecteren en dit ook toepassen op het random forest model om uiteindelijk de finale performantie te bekomen.

Figuren 25 en 26 tonen aan dat de performantie van het OLS model verbetert indien we minder buren in rekening nemen. Het optimale nummer is gelijk aan twee buren, kort gevolgd door maar één buur in rekening te nemen. Buiten het verschil tussen één en twee buren stijgt de MAE min of meer lineair met het aantal buren dat we in rekening nemen. Het verschil is wel veel kleiner dan het geval was bij appartementen. De baseline MAE waarbij we 10 buren in rekening nemen is gelijk aan ongeveer 0.44, terwijl de MAE daalt tot ongeveer 0.42 – 0.43 indien we maar twee buren in rekening nemen.

Als we nu dit aantal buren ook gebruiken voor het random forest model verkrijgen we een MAE van 0.35. Deze performantie is gelijk aan de baseline performantie van het random forest model indien we de afstand tot het centrum toevoegen. De fout komt overeen met EUR 44,958, voor een gemiddelde verkoopprijs gelijk aan EUR 135,032. Net zoals in het vorige hoofdstuk is het complexere model nu ook weer minder gevoelig aan het aantal buren in vergelijking met het OLS model. Dit is echter ook wat we kunnen verwachten, een complexer model kan namelijk meer aanvangen met de reeds verkregen informatie in het baseline scenario. Het is wel belangrijk op te merken dat het random forest model nog altijd veel beter presteert dan het OLS model!



Figuur 25: De MAE van het lineaire regressiemodel voor een variërend aantal buren dat we in rekening nemen voor de gemiddelde gewogen schattingsfout



Figuur 26: De R^2 van het lineaire regressiemodel voor een variërend aantal buren dat we in rekening nemen voor de gemiddelde gewogen schattingsfout

5.3 Out-of-sample predictie met betrekking tot moment van verkoop

In deze sectie bekijken we een andere vorm van out-of-sample predictie, namelijk hoe het model presteert als we de dataset opdelen in een periode voor 2018 en na 2018. Op die manier kunnen we nagaan hoe het model presteert in een tijdsperiode die niet gebruikt werd om het model te schatten. Zoals hierboven reeds gezegd moeten we in dit geval wel gebruik maken van een lineaire tijdstrend in plaats van jaar fixed effects.

Met deze procedure te volgen bekomen we een MAE van 0.41 voor het random forest model met twee buren en de afstand tot het centrum, een stijging van 0.06. Deze stijging is dus groter dan in de twee voorgaande hoofdstukken als we naar de realistische setting overgaan. Wanneer we de verkregen MAE, uitgedrukt in het natuurlijk logaritme van de prijs, omzetten naar de euro-waarde verkrijgen we een gemiddelde absolute schattingsfout van EUR 56,092. De schattingsfout daalt opnieuw niet indien we de euro-waarde als uitkomstvariabele gebruiken in plaats van het natuurlijk logaritme.

Hieruit kunnen we concluderen dat de schattingsfout in het geval van bouwgronden zeer groot is. De gemiddelde absolute schattingsfout uitgedrukt in de euro-waarde is groter dan het geval was voor zowel woonhuizen als appartementen. Dit terwijl de gemiddelde prijs van een bouwgrond veel lager ligt dan die voor woonhuizen en appartementen. We zagen op voorhand al dat de MAE in dit hoofdstuk veel hoger lag dan in de voorgaande, maar het is zelfs zo slecht dat zich dit vertaalt in een hogere schattingsfout in euro-waarde.

CONCLUSIE

In dit onderzoeksrapport onderzochten we welke statistische modellen de verkoopprijs van woonhuizen, appartementen en bouwgronden kunnen schatten. We gebruikten hiervoor data met betrekking tot verkopen van vastgoed in België, de EPC-databank en data over de woningstock. Aan de hand van de capakey, een unieke identificatiecode van het perceel, konden we ook de geografische coördinaten koppelen waardoor we de exacte locatie van de woning konden bepalen.

Uit de analyse komt duidelijk naar voren dat schattingen van de verkoopwaarde van woonhuizen en appartementen accurater zijn dan de schattingen van de prijs van bouwgrond. Een mogelijke verklaring hiervoor is het aantal verkopen dat plaatsvindt: bij bouwgronden moeten we het met veel minder observaties stellen waardoor deze observaties ook verder van elkaar liggen en we minder accurate locatie-effecten kunnen schatten. Daarom testen we in bijlage ook nog de specificatie indien we schattingsfouten van woonhuizen toevoegen. We zien hier dat de performantie inderdaad verbetert. Ook is het mogelijk dat de data bij woonhuizen en appartementen van een hogere kwaliteit zijn. Zo is de oppervlakte van de (bouw)grond in de transactiedata niet altijd correct. Dit is bijzonder problematisch, aangezien voor een grond de oppervlakte de belangrijkste determinant is van de verkoopprijs naast zijn locatie.[14]

Verder is het noemenswaardig dat de prijs van appartementen schatten even goed of zelfs beter lukt dan de prijs van woonhuizen ondanks het feit dat het niet mogelijk is om de appartementen aan de EPC-databank te koppelen. Een mogelijke verklaring ligt in het feit dat appartementen in eenzelfde gebouw vaak heel gelijkaardig zijn. Als één appartement reeds verkocht is in hetzelfde gebouw hebben we daarom veel meer informatie over de andere appartementen in dat complex. Met andere woorden, de schattingsfout van de buren draagt veel meer informatie in de subset van appartementen als dit mechanisme speelt. We zien inderdaad ook dat de schattingsfout sterker daalt bij appartementen wanneer we de schattingsfout van de buren toevoegen in vergelijking met woonhuizen.

Met een gemiddelde absolute schattingsfout van EUR 37,710 en EUR 28,760 voor respectievelijk de verkoopprijzen van woonhuizen en appartementen is er echter nog een substantieel deel van de verkoopprijs dat niet geschat kan worden door bovenstaande modellen. Zoals hierboven al uitgelegd is het onmogelijk om geen schattingsfouten te produceren, daar een deel van de prijs bepaald wordt door ruis. Of één van de modellen in aanmerking kan komen om een nieuwe belastbare basis te schatten zal daarom moeten blijken uit een vergelijking met andere opties die voorhanden zijn en de huidige relatie tussen het KI en de huur of verkoopprijs, in tegenstelling tot het kijken naar de absolute schattingsfout. De huidige relatie tussen het KI en de prijs komt overeen met een R^2 van 45% voor woonhuizen, terwijl deze voor appartementen gelijk is aan 28%.[15] Met een R^2 van 73% en 77% doen bovenstaande modellen het dus veel beter.

Het voordeel van het gebruik van *machine learning* is bovendien dat er geen systematische bias meer zit in de schattingen, iets wat met het huidige KI wel duidelijk het geval is zoals reeds aangetoond in het vorig rapport. Een voorbeeld hiervan is dat momenteel bepaalde regio's systematisch onder- of overschat worden. Dit zal door het gebruik van *machine learning* niet het geval zijn, gegeven dat het model op een correcte manier getraind is.

[14] De prijsschatting van bouwgronden met het gebruik van een andere dataset, waarvan we zeker zijn dat de oppervlakte altijd correct is ingevuld, resulteert echter niet in accuratere schattingen. Het is daarom aannemelijk dat de fouten in de dataset met betrekking tot de oppervlakte niet resulteren in problemen voor de schattingen. Dit is uiteraard wel na het verwijderen observaties waarvan we vermoeden dat er fouten inzitten.

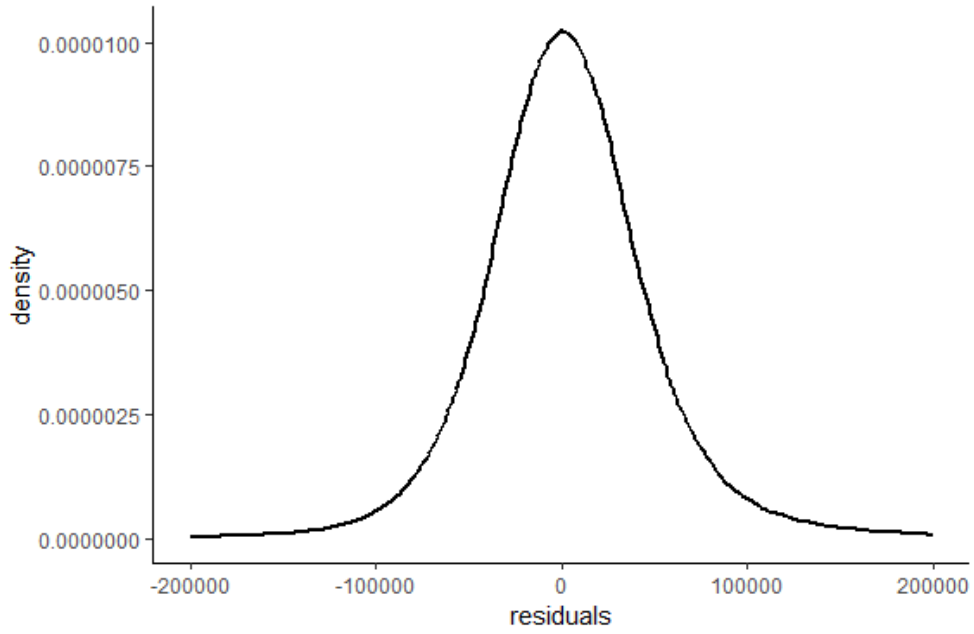
[15] Deze relatie werd eerder aangetoond in een voorgaand verslag getiteld "De staat van het KI" (Boogaerts et al., 2020).

De gemiddelde absolute schattingsfout voor de verkoopprijzen van bouwgronden is gelijk aan EUR 41,465. In absolute getallen ligt deze fout al veel hoger dan de fout die we vonden bij woonhuizen en appartementen, maar het is ook belangrijk om de schattingsfout te vergelijken met de verkoopprijs in plaats van ons blind te staren op de absolute fout. Relatief gezien is deze schattingsfout daarom vele malen groter dan de fout die we vonden bij woonhuizen en appartementen, aangezien de gemiddelde verkoopprijs veel lager ligt bij bouwgronden. De gemiddelde prijs van bouwgrond in onze sample is namelijk EUR 135,029, terwijl deze van huizen en appartementen gelijk is aan respectievelijk EUR 224,449 en EUR 184,839.

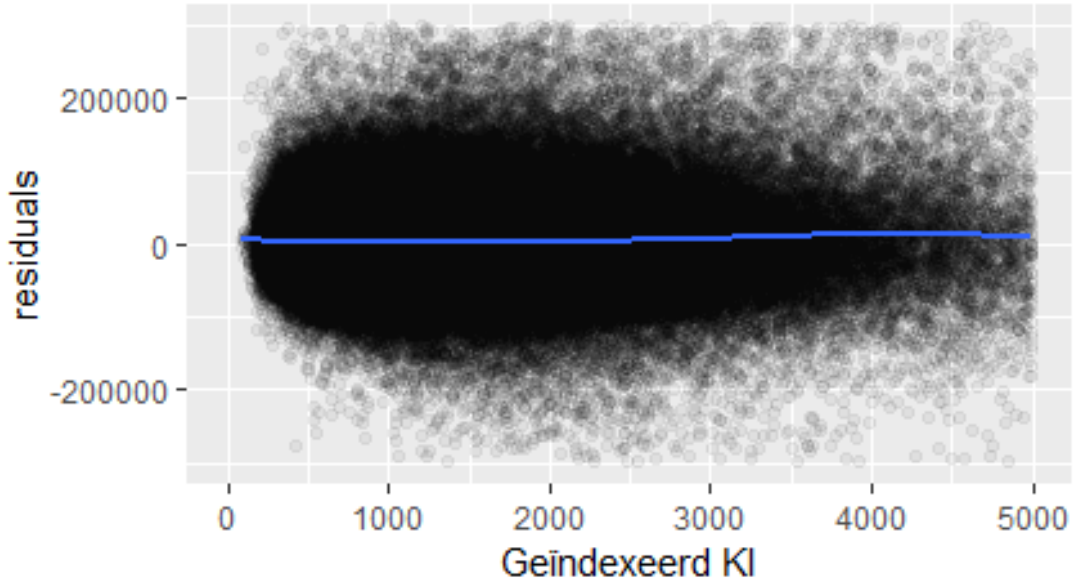
De reden waarom prijzen van bouwgrond schatten veel moeilijker lukt dan die van woonhuizen of appartementen is ons momenteel niet bekend. Hoogstwaarschijnlijk kunnen de schattingen verbeterd worden indien bouwnormen en –regulaties geobserveerd worden, maar voor de rest is het momenteel gissen. De vergelijking met het huidige KI biedt ook geen goed referentiepunt, gezien onbebouwde gronden een zeer laag KI krijgen toegekend en er weinig differentiatie is.

BIJLAGEN

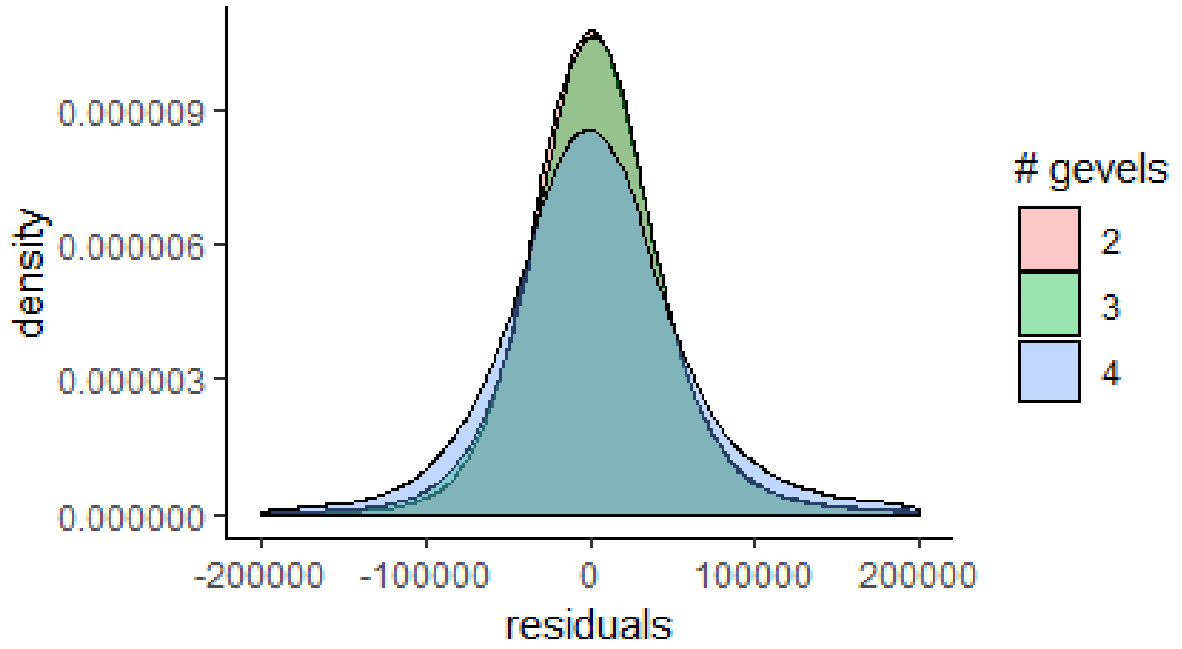
Figuren



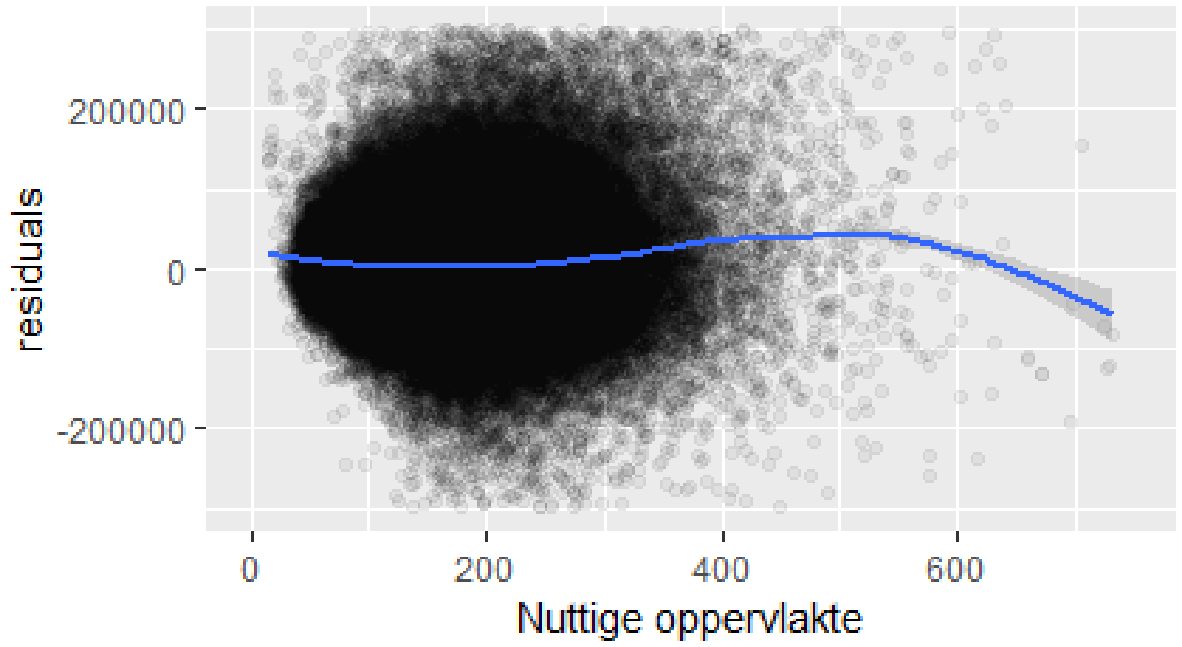
Figuur A1: De verdeling van de predictiefout voor de dataset van woonhuizen waarbij we de EPC-score opnemen in het model



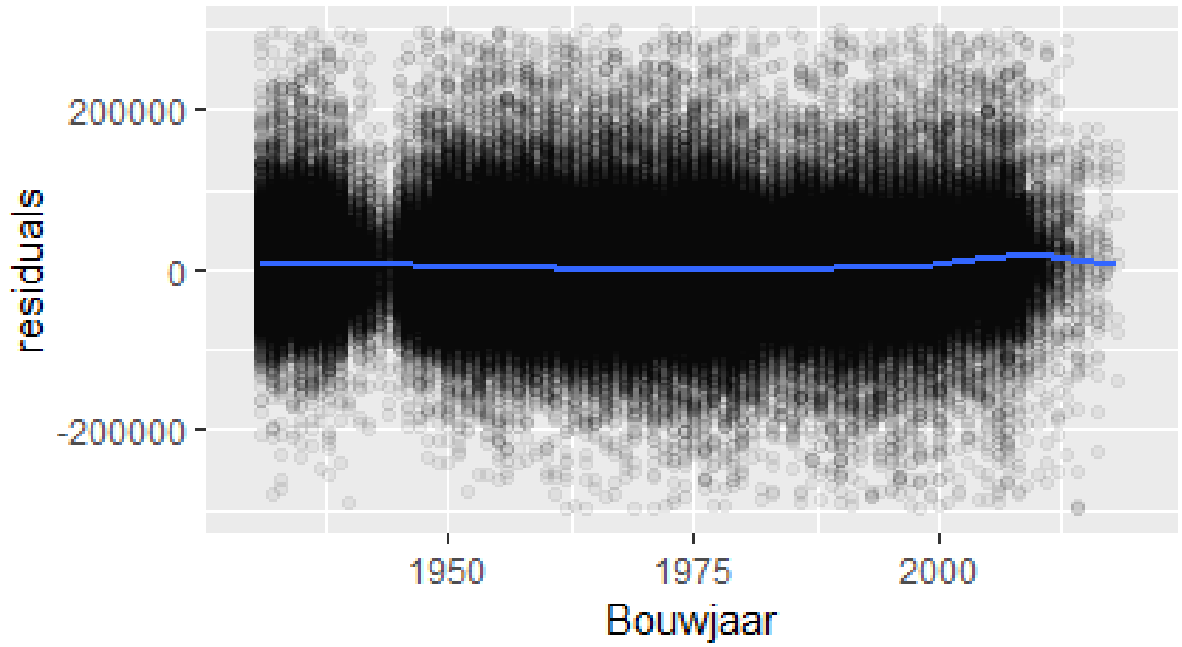
Figuur A2: De individuele predictiefouten en de gemiddelde predictiefout ten opzichte van het geïndexeerd KI voor de dataset van woonhuizen waarbij we de EPC-score opnemen in het model



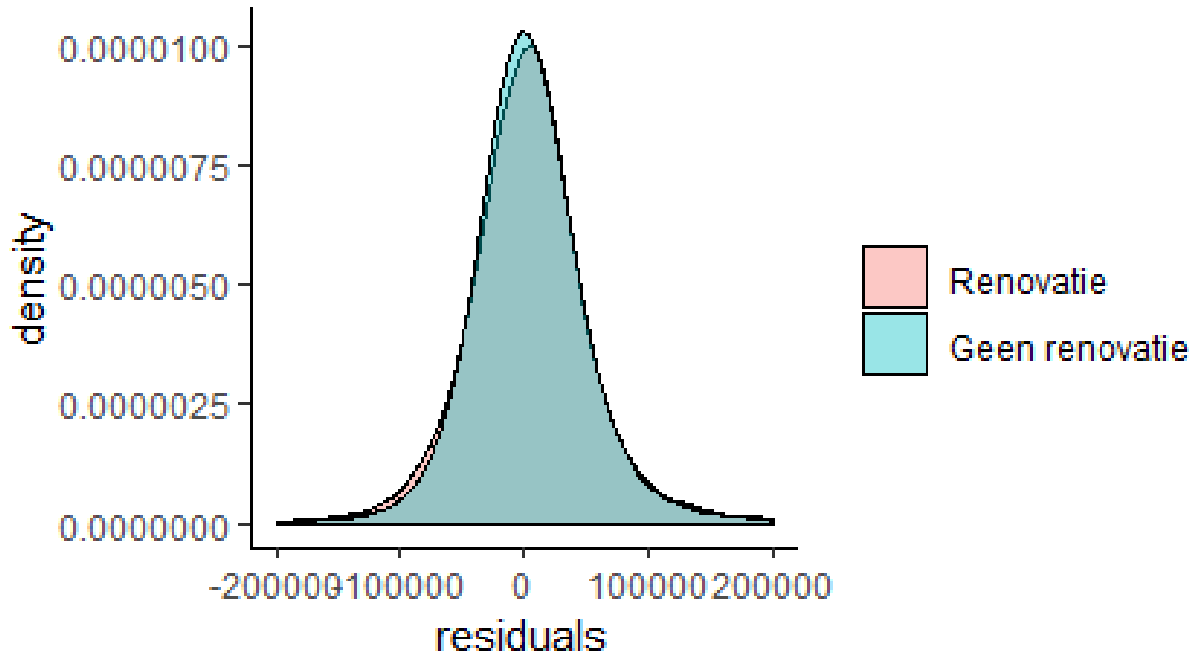
Figuur A3: De verdeling van de predictiefout voor gesloten, halfopen en open bebouwing voor de dataset van woonhuizen waarbij we de EPC-score opnemen in het model



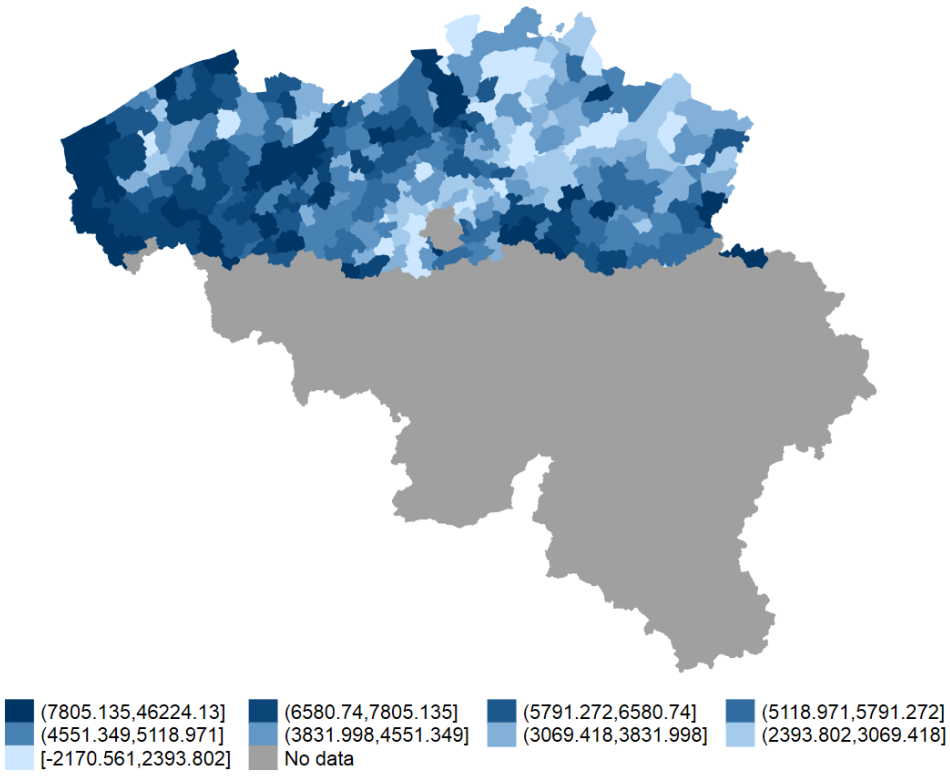
Figuur A4: De individuele predictiefouten en de gemiddelde predictiefout ten opzichte van de nuttige oppervlakte voor de dataset van woonhuizen waarbij we de EPC-score opnemen in het model



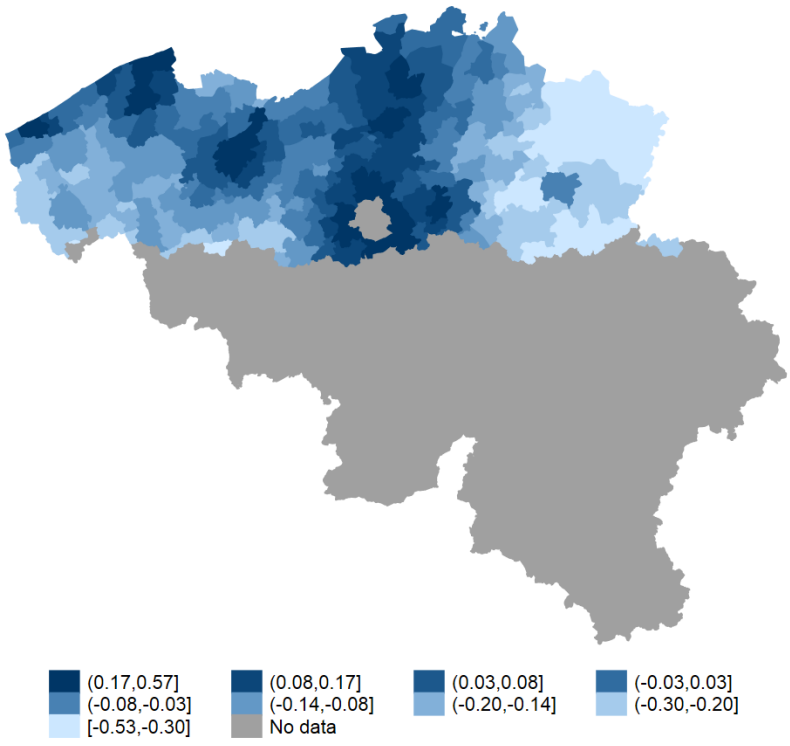
Figuur A5: De individuele predictiefouten en de gemiddelde predictiefout ten opzichte van het bouwjaar voor de dataset van woonhuizen waarbij we de EPC-score opnemen in het model



Figuur A6: De verdeling van de predictiefout voor huizen met en zonder renovatie voor de dataset van woonhuizen waarbij we de EPC-score opnemen in het model



Figuur A7: De gemiddelde predictiefout per gemeente weergegeven op de kaart van België voor de dataset van woonhuizen waarbij we de EPC-score opnemen in het model



Figuur A8: De gemeente fixed effects weergegeven op de kaart van België indien we voor de EPC-score controleren, waarbij een donkerdere kleur een hogere liggingscoëfficient weergeeft

Enkel recent gebouwde huizen

Zoals eerder vermeld zijn de data niet voor elk huis van een even hoge kwaliteit. Veel huizen hebben renovaties ondergaan en achteraf de renovaties en de bijhorende veranderingen in de karakteristieken van het huis niet doorgegeven bijvoorbeeld. Daarnaast zijn er ook variabelen die ronduit verkeerd zijn ingevuld voor bepaalde huizen. Deze hebben we in de mate van het mogelijke telkens verwijderd, maar er zijn ongetwijfeld foutieve observaties in de uiteindelijke dataset gebleven.

Om min of meer de impact van deze fouten na te gaan schatten we hier hetzelfde model als hierboven, maar filteren we de data zodat er enkel huizen met een leeftijd van 20 jaar of minder overblijven. De redenering hierachter is dat observaties van recent gebouwde huizen een grotere kans hebben om correct te zijn. Dit is uiteraard geen exacte methode, maar het zou wel een eerste inschatting kunnen geven.

De predictiefout die we hiermee verkrijgen is gelijk aan EUR 40,747. Een lichte achteruitgang ten opzichte van de schattingsfout van EUR 37,710 die we hierboven verkregen. Proportioneel doet dit model echter wel veel beter dan het model hierboven. De gemiddelde prijs van de huizen in deze dataset is gelijk aan EUR 338,486, terwijl de gemiddelde prijs van de huizen in de volledige dataset waarvoor de EPC fiche beschikbaar is gelijk is aan EUR 243,947. Dit komt daarom overeen met een schattingsfout van respectievelijk 12% en 15.5%. Daarbovenop komt ook het feit dat we in de dataset met nieuwere huizen moeten werken met veel minder observaties, namelijk enkel 30,545 observaties ten opzichte van 392,331 observaties in de volledige dataset. Daarentegen is de dataset bestaande uit recentere huizen wel homogener, iets wat doorgaans de predictie vereenvoudigt.

Met deze informatie kunnen we daarom niet definitief concluderen wat het effect zou zijn van betere data. De vooruitgang die we zien hint wel naar een positief en significant effect en daarom willen we opnieuw aandringen op het belang van goede data. Het resultaat van een predictie staat of valt namelijk met de kwaliteit van de data.

Locatie-effecten van woonhuizen gebruiken voor bouwgrond

Aangezien we over minder observaties beschikken in de transactie dataset met betrekking tot bouwgronden is het ook aannemelijk dat we minder accuraat de locatie fixed effects kunnen schatten met behulp van deze dataset. Om te controleren of dit inderdaad het geval is en we dit kunnen verhelpen maken we gebruik van de gemeente fixed effects die we verkregen met behulp van de transactie dataset met betrekking tot woonhuizen. Deze dataset bevat veel meer observaties en laat ons daarom mogelijks ook toe om specifiekere locatie effecten te schatten.

Als we de gemeente fixed effects verkregen met woonhuizen toevoegen aan het beste model voor het schatten van landprijzen bekomen we een MAE van 0.33, dit correspondeert met een gemiddelde absolute schattingsfout van EUR 41,465. Een verbetering van EUR 3,500 ten opzichte van het vorige model hierboven beschreven. Het toevoegen van gemeente fixed effects voor woonhuizen heeft dus wel degelijk meerwaarde voor het schatten van landprijzen en is dus ook aan te raden in de praktijk.

Als we hierbij ook nog de schattingsfout van de nabijgelegen woonhuizen toevoegen wordt de schatting niet beter. Hiermee bedoelen we dus dat we voor elke bouwgrond nagaan wat de verkregen schattingsfout was van de nabijgelegen woonhuizen en dit dan toevoegen aan het model, naast de schattingsfout van de nabijgelegen bouwgronden. Dit laatstgenoemde zat sowieso al in het model.

REFERENTIES

Boogaerts, T., Damen, S. & Schildermans S. (2020). De staat van het kadastraal inkomen.

Bracke, P. (2015). House Prices and Rents: Micro Evidence from a Matched Data Set in Central London. Real Estate Econ. 43 (2): 403–20.

Damen, S. (2019). Het effect van het EPC en energetische kenmerken op de verkoopprijs van woningen in Vlaanderen. Vlaams Energieagentschap, april 2019.

Mahieu, B., Heyndels, B., Burssens, J., Goeminne, S. & Smolders, C. (2012). Een analyse van de relatie tussen KI en woningprijzen in de Vlaamse centrumsteden. Documentatieblad Federale Overheidsdienst Financiën: 72(2).

Vastmans, F. (2020). De energieprestaties van de Vlaamse woningvoorraad. Cijfers en verklaringen. Leuven: Steunpunt Wonen.

Verbeeck G. & Ceulemans W. (2015). Analyse van de EPC-databank. Resultaten tot en met 2012. Leuven: Steunpunt Wonen

Research Foundation
Flanders
Opening new horizons