Vlaanderen
is omgeving

Analyse datakwaliteit en geografische verwerking van immodata

DEPARTEMENT
OMGEVING

omgevingvlaanderen.be

Deze opdracht onderzoekt de datakwaliteit en bruikbaarheid van een ruwe immodatabank, afkomstig van Z-immo. Daarnaast wordt ook bekeken of op basis van deze databank beleidsindicatoren ontwikkeld kunnen worden.

COLOFON

Verantwoordelijke uitgever:

Departement Omgeving
Vlaams Planbureau voor Omgeving

Koning Albert II-laan 20 bus 8
1000 Brussel

vpo.omgeving@vlaanderen.be

www.omgevingvlaanderen.be

Bronverwijzing: Antea Group en KULeuven (SADL)(2017), Analyse datakwaliteit en (geografische) verwerking van immodata, uitgevoerd in opdracht van het Vlaams Planbureau voor Omgeving.

D/2017/

PARTNERS

Inhoud

Niet-technische samenvatting

Dit rapport beschrijft het resultaat van het onderzoek dat uitgevoerd werd om de datakwaliteit van de immodatabank van Zimmo te onderzoeken. Het gaat om een eerste exploratieve analyse, waarbij werd onderzocht wat de kenmerken van de databank zijn, in welke mate de gegevens ruimtelijk te lokaliseren zijn, welke variabelen zijn opgenomen, hoe bruikbaar en betrouwbaar deze variabelen zijn, en in welke mate de databank bruikbaar kan zijn in functie van het ontwikkelen van beleidsindicatoren.

De databank omvat bijna 3 miljoen records, met gegevens per pand of grond over wat te koop of te huur staat, met een historische reeks van de afgelopen 10 jaar. Het gaat om data die via een immo-webportaal wordt aangeboden over geheel België en deels erbuiten. De databank omvat een veelheid aan karakteristieken per pand of grond. In totaal werden 104 variabelen ontvangen. Niet alle karakteristieken van elk pand of elke grond zijn gekend en velden zijn vaak niet ingevuld. De belangrijkste doelstelling van deze opdracht is om een eerste inzicht te verkrijgen in de datakwaliteit en bruikbaarheid van de databank in functie van ruimtelijke analyses.

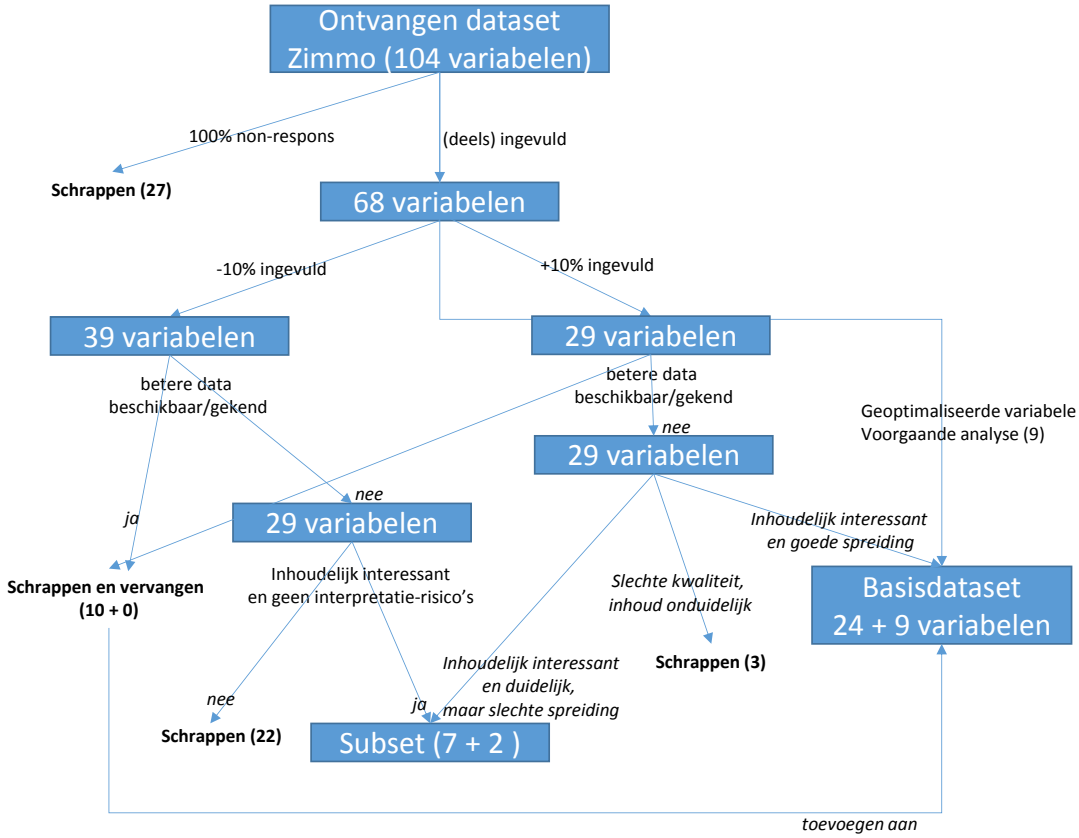
In eerste instantie wordt bekeken tot op welke hoogte de gegevens ruimtelijk gelokaliseerd kunnen worden. Er wordt onderzocht in welke mate de records uit de databank, die in grote mate ook adresgegevens omvatten, gekoppeld kunnen worden aan bestaande adresatlassen voor Vlaanderen, Brussel en Wallonië. Hiervoor worden verschillende stappen doorlopen, waarbij wordt nagegaan of postcode, straatnaam en huisnummer correct zijn ingevuld. Per stap neemt het aantal records drastisch af, waardoor uiteindelijk slechts ongeveer een derde van het oorspronkelijke bestand gegeolocaliseerd kan worden. Dit wordt in onderstaande tabel samengevat. Uit de overgebleven set van ongeveer 1 miljoen records, wordt vervolgens een selectie van unieke adressen gemaakt waardoor uiteindelijk 531.040 records overblijven voor verdere analyse. Aan de records worden bijkomende variabelen gekoppeld die eenvoudig af te leiden zijn vanuit de adresatlassen, waardoor een aantal variabelen uit de immodatabank geoptimaliseerd kunnen worden (zoals bv. oppervlakte van gebouwen en/of percelen).

Ontvangen immo_data bestand	2.900.979
Postcode formaat is 4 cijfers	2.791.856
immo_Precisie_Locatie = -1 : niets correct van het adres	2.640
immo_Precisie_Locatie = 3 : enkel postcode correct	1.490.826
immo_Precisie_Locatie = 2 : enkel postcode & straatnaam correct en voor het huisnr werd het numeriek minst verschillende huisnr genomen	243.594
immo_Precisie_Locatie = 4 : enkel postcode & straatnaam correct en voor het huisnr werd het huisnr uit de CRAB adrespunten genomen waarvan de x, y positie minder dan 10m in afstand verschilt van de immo_X_lamb_input en immo_Y_lamb_input	19.675
immo_Precisie_Locatie = 1 : adres (postcode, straatnaam, huisnr) volledig correct	1.035.121
Selectie unieke adressen op basis van recentste immo_Pub1_start datum en velden a_postcode, a_straat, a_nummer zijn niet leeg	531.040

Vervolgens wordt een exploratieve statistische analyse op de variabelen uitgevoerd om de kwaliteit en bruikbaarheid van de databank verder te onderzoeken. De analyse bestaat in eerste instantie uit een exploratieve of verkennende analyse waarna de feitelijke analysedataset tot stand komt

In eerste instantie wordt de casus Oost-Vlaanderen onderzocht (simultaan aan de operatie van geolocalisatie). Er is in eerste instantie ingegaan op de mate waarin variabelen al dan niet zijn ingevuld (non-respons). Op basis van de non-respons-analyse worden 27 variabelen van bij het begin als niet bruikbaar beschouwd wegens niet ingevuld (100% non-respons). Vervolgens wordt een verder onderscheid gemaakt op basis van een beslissingsboom. Hoe meer een variabele ingevuld is, hoe meer potentie voor verdere opname in de analyse-databank. Bovendien wordt ook nagegaan of voor een aantal variabelen mogelijk betere data beschikbaar zijn die aan de immodatabank gekoppeld kunnen

worden en worden de variabelen gescreend op hun inhoudelijke bruikbaarheid. Zo zijn bijvoorbeeld variabelen met een duidelijk te subjectieve invoer en/of variabelen met duidelijk veel foute ingaven (bv. geen cijfer waar dit verwacht wordt, meer dan 4 cijfers waar een jaartal verwacht wordt, enz…) minder bruikbaar voor verdere analyse dan variabelen waar dit wel eenduidig en objectief gebeurt.



De variabelenset wordt vervolgens aan een aantal bijkomende tests en analyses onderworpen. Zo wordt onderzocht of 'prijs' (de focus van hedonische/multilevel analyse, o.v.v. afhankelijke variabele in de analyse) significante én relevante systematische verschillen vertoont naargelang respons of non-respons. Ook wordt nagegaan hoeveel unieke records uiteindelijk overblijven, indien we tot een respons-only set willen komen voor een combinatie van variabelen. De spreiding van variabelen werd in beeld gebracht, alsook (potentieel) verwantschap tussen records.

Uit de inzichten vanuit de casus Oost-Vlaanderen en de resultaten van de exploratieve statistische analyse, wordt een analyset opgebouwd. De afgeleide subset bevat enkel unieke records, met voor de variabele 'a_status_id' de invulling 'verkocht', 'te koop', 'verhuurd' of 'te huur'; en met voor 'a_type_id' de invulling 'woning' ofwel 'appartement'. Een subset voor 'a_type_id = grond' werd separaat aangemaakt. Hierdoor wordt het aantal records verder gereduceerd tot 340.276 unieke datarijen.

De analyset zelf wordt opnieuw onderworpen aan een exploratieve analyse en de basiskenmerken van de analyseset worden in beeld gebracht. De potentiële outliers van deze analyseset (namelijk de waarden die sterk afwijken op basis van interkwartielafstand) worden bepaald en als variabele toegevoegd in de databank. Indien een afgeleide dataset voor woningen/appartementen zou worden gecreëerd met eliminatie van alle harde uitbijters uitgenomen deze voor f_ki inzake appartementenverkoop (gezien voor f_ki op zich wel reeds werd opgepikt), dan zou de dataset herleid worden van 340.276 tot 298.953 datarijen. Gezien deze outliers echter voornamelijk een knipperlichtfunctie hebben en niet per definitie uit te sluiten zijn, wordt verder gewerkt met de dataset van 340.276 unieke records.

Vervolgens wordt de deze dataset onderworpen aan ruimtelijke analyse. In deze eerste exploratie worden een aantal gegevens aan de databank gekoppeld, onder andere om representativiteit van de Z-immodatabank te onderzoeken door de vraagprijs uit de databank te vergelijken met de gegevens over verkoopprijs uit nationale statistieken. Het hoog aandeel verkochte woningen ten opzichte van het totaal aantal verkopen per gemeente (cfr. kadaster) is hierbij een indicatie van de hoge potentie van de Zimmo-databank. De meerderheid van de gemeenten heeft een afwijking van minder dan 20% in het aantal publicaties ten opzichte van de statistieken van transacties op basis van het kadaster, en de vraagprijs is meestal hoger dan de verkoopprijs, wat ook logisch is.

Er wordt verder nagegaan in welke mate deze gegevens geschikt zijn voor de berekening van beleidsindicatoren over de woningmarkt. De woningmarkt wordt onderverdeeld in (1) de koopmarkt: huizen, appartementen en bouwgronden en (2) de huurmarkt: huizen en appartementen. De berekende beleidsindicatoren zijn: snelheid van verkoop, frictieleegstand, vraagprijs en verkoopprijs, en het verschil ertussen, en schaarste-index. Voor een aantal indicatoren worden hotspot analyse uitgevoerd om ruimtelijke concentraties verder te onderzoeken.

Resultaten per indicator drijven ons voor deze samenvatting te ver (en hiervoor wordt verwezen naar deel 3 van dit rapport), maar er kan wel gesteld worden dat enkele ruimtelijke patronen berekend aan de hand van hotspot analyses aantonen dat de Zimmo-databank kan leiden tot nieuwe inzichten in de woningmarkt. De indicator 'Tijd te koop' (proxy tijd online) is bv. een indicator die weliswaar een ruwe proxy is, maar die in geen andere databank terug te vinden is. In de hotspots analyses op basis van de microdata zijn duidelijke patronen waarneembaar. Deze zijn weliswaar niet allemaal te verklaren, maar bieden potentieel voor bijkomend onderzoek. De potentie van de databank neemt trouwens sinds 2010 jaarlijks toe. Door de trend naar toenemende verkoop en verhuur van woningen via het Internet zal in de toekomst de bruikbaarheid van deze databank voor berekening van beleidsindicatoren nog verder toenemen.

Dit verkennend onderzoek heeft echter ook een reeks beperkingen van de databank blootgelegd, die op basis van deze eerste analyse niet verder uitgeklaard konden worden. De eerste ruimtelijke analyse laat toe om ruimtelijke patronen te detecteren en in beeld brengen, maar brengt tegelijk ook aan het licht dat er nog te veel afwijkingen zijn om in deze fase verder te gaan met een aantal bijkomende analyses (factoranalyse, clusteranalyse, residu analyse). Ook stappen in de richting van een steekproef zijn nog voorbarig; er is nog te veel ruis door ongewenste causale factoren om goede criteria te bepalen voor, bijvoorbeeld, een stratificatie. Uit de ruimtelijke analyse blijken de data dus nog te ruw om voor een aantal beleidsindicatoren (nu al) betrouwbare conclusies te trekken. Het is verder ook duidelijk dat heel wat velden niet altijd correct zijn ingevuld (bv. jaartallen), maar het ontbreekt nog aan een volledig inzicht hierin. Ook zijn er veel variabelen waarvoor de respons laag is. Tot slot wordt ook vastgesteld dat er een aantal gemeenten zijn, zeker in Wallonië, met weinig records in totaliteit. Er worden nu eenmaal in sommige landelijke gemeenten weinig woningen verkocht of verhuurd. Een eenduidige gebied dekkende aggregatie van gegevens tot gemeenteniveau en zeker tot op niveau van statistische sectoren is bijgevolg niet altijd zinvol.

Er zit dus weldegelijk potentie in de onderzochte databank, maar vooraleer verder te kunnen toewerken naar bruikbare beleidsindicatoren, is in principe eerst een verdere uitzuivering van de databank noodzakelijk. Zeker voor wat betreft de prijzen, is nog bijkomend onderzoek nodig naar het effect van outliers. Daarnaast kan de feedback tussen de ruimtelijke analyse en de opmaak van het analysebestand opnieuw leiden tot optimalisatie van de databank en bijkomende ruimtelijke analyses. Ook deelanalyses op datakwaliteit, opgedeeld naar type bebouwing (en eventueel subtype) kan hierin een meerwaarde bieden. Dankzij een iteratief proces (bv. 3 tot 5 iteraties), kan de ruis uit de databank nog sterk gereduceerd worden en kunnen verklarende factoren en causale relaties verder in beeld gebracht worden.

Inleiding

Dit rapport beschrijft het resultaat van het onderzoek dat uitgevoerd werd om de datakwaliteit van de immodatabank van Zimmo te onderzoeken. Het gaat om een eerste exploratieve analyse, waarbij werd onderzocht wat de kenmerken van de databank zijn, in welke mate de gegevens ruimtelijk te lokaliseren zijn, welke variabelen zijn opgenomen, hoe bruikbaar en betrouwbaar deze variabelen zijn, en in welke mate de databank bruikbaar kan zijn in functie van het ontwikkelen van beleidsindicatoren.

De databank omvat bijna 3 miljoen records, met gegevens per pand of grond over wat te koop of te huur staat, met een historische reeks van de afgelopen 10 jaar. Het gaat om data die via een immo-webportaal wordt aangeboden over geheel België en deels erbuiten. De databank omvat een veelheid aan karakteristieken per pand of grond. Niet alle karakteristieken van elk pand of elke grond zijn gekend en velden zijn vaak niet ingevuld. De belangrijkste doelstelling van deze opdracht is om een eerste inzicht te verkrijgen in de datakwaliteit en bruikbaarheid van de databank in functie van ruimtelijke analyses.

In dit onderzoek wordt – na een beschrijving van de databank zelf – in eerste instantie de mogelijkheid voor **geolocalisatie van adressen** onderzocht (DEEL 1). De normalisatie en de geocodering van de immo adressen gebeurt aan de hand van de volgende adresatlassen van de gewesten :

- Voor Vlaanderen : de CRAB adrespunten en de GRB gebouwen.
- Voor Brussel : de Urbis adrespunten en gebouwen.
- Voor Wallonië : de PICC adrespunten

Daarbij worden de volgende stappen doorlopen :

- Voorbereiding adresatlassen van de gewesten.
- Inbreng van het immo bestand.
- Normalisatie van de immo adresinformatie.
- Geocodering van de immo adressen.
- Bepaling van de oppervlaktes van de gebouwen die bij een bepaald immo adres behoren
- Bepaling van de oppervlaktes van de percelen die bij een bepaald immo adres behoren
- Onderzoek naar unieke adressen

Vervolgens wordt ook een eerste statistische analyse op de variabelen uitgevoerd om de **kwaliteit en bruikbaarheid van de databank** verder te onderzoeken (DEEL 2). Hierbij wordt onder meer ingegaan op de mate waarin variabelen al dan niet zijn ingevuld (non-respons). Daartoe vindt vooraf een uitgebreide screening van de data plaats met bijhorende preprocessing om specifieke onvolmaaktheden weg te filteren en de data voor analyse voor te bereiden. De analyse bestaat in eerste instantie uit een exploratieve of verkennende analyse waarna de feitelijke analysedataset tot stand komt, waarmee de beleidsindicatoren bepaald worden. Voor de verkennende analyse worden ook subsets van de totale databank ingezet.

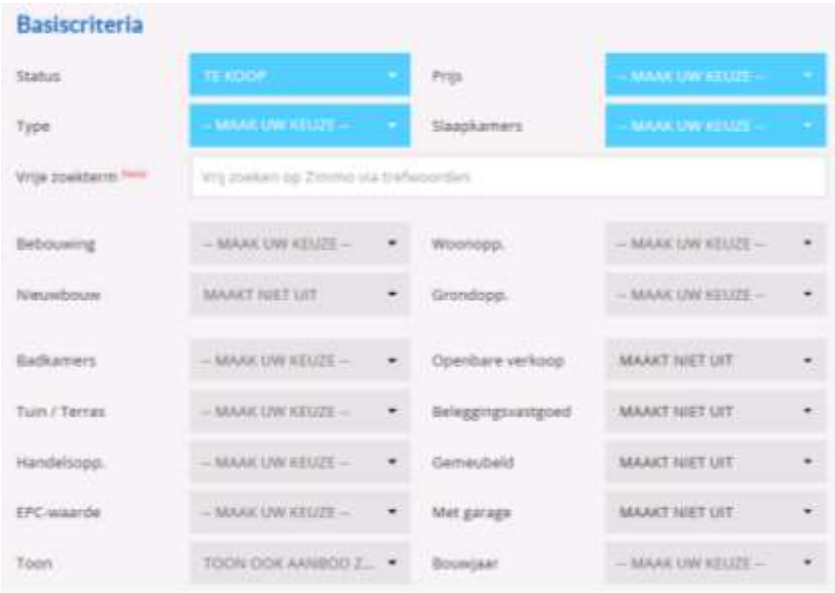
Tot slot wordt ook een overzicht van analyse-testen op de databank toegevoegd, waarbij werd onderzocht of en hoe een aantal **beleidsindicatoren** op basis van de databank berekend kunnen worden (DEEL 3).

Beschrijving van de databank

Voor deze studie werd vertrokken van een bestand met 2.900.979 records. Gezien Zimmo zich in het verleden heeft geconcentreerd op de markt in Vlaanderen, zijn voor Vlaanderen de meeste gegevens beschikbaar (zie bv. verderop Tabel 1 – verdeling van de data over de postcodes).

De databank kan geraadpleegd worden via www.zimmo.be, waarbij een pand kan opgezocht worden op basis van verschillende criteria.

Figuur 1 – printscreen van een zoekopdracht op zimmo.be



Elk record uit de databank gaat over een pand dat te koop of te huur staat, verhuurd is of verkocht. In de databank zijn opgenomen: woning, appartement, garage, bedrijfsvastgoed, grond.

Aan elk record zijn verschillende variabelen verbonden (een totaal van 104[1]):

1. a_beschr_nl: beschrijving van de woning (ligging, aantal slaapkamers, faciliteiten, …)
2. a_beschrkort_nl: beschrijving van de woning (ligging, aantal slaapkamers, faciliteiten, …)
3. a_bus: busnummer
4. a_gemeente: gemeente
5. a_geo_lat: breedteligging
6. a_geo_lon: lengteligging
7. a_geo_precision: precisie van het adres
8. a_land_id: id land (uniek nummer per land)
9. a_nummer: huisnummer
10. a_omgeving_id: beschrijving 'omgeving' (dorp, stad, residentieel, industrieel, …)
11. a_postcode: postcode
12. a_sleutelnummer: nummer van de sleutel (interne informatie Zimmo)
13. a_status_id: te koop, verkocht, verhuurd, …

[1] In het ontvangen bestand zaten 104 variabelen. De volledige databank omvat meer velden (343) die binnen de looptijd van dit project echter niet meer onderzocht konden worden.

14. a_straat: straat
15. a_subtype_id: 111 subtypes:woning, dakappartement, handelspand, villa, garagebox, ….
16. a_titel_nl: titel aankondiging
17. a_type_id: 5 types: woning, appartement, garage, bedrijfsvastgoed, grond
18. a_vrijop_datum: datum waarop het pand beschikbaar komt
19. a_vrijop2_id: bij akte na betaling, onmiddellijk, bij oplevering, …
20. b_bebouwing_id: open, gesloten, halfopen
21. b_bouwjaar: bouwjaar
22. b_bouwopp: bouwoppervlakte
23. b_epc_attest: aanwezigheid van een EPC attest
24. b_epcwaarde: EPC waarde
25. b_grondopp: grondoppervlakte
26. b_kadastraleaard: kadastrale aard
27. b_kadastraleafdeling: kadastrale afdeling
28. b_kadastralesectie: kadastrale sectie
29. b_nieuwbouw: nieuwbouw (0, 1, f)
30. b_perceelbreedte: perceelbreedte
31. b_perceeldiepte: perceeldiepte
32. b_perceelnummer: perceelnummer
33. b_renovatiejaar: renovatiejaar
34. b_staat_id: nieuwe staat, vernieuwd, uitstekende staat, …
35. b_terenoveren: te renoveren (1, 0, f)
36. b_woonopp: woonoppervlakte
37. creatie_datum: datum van aanmaak van het pand in de database
38. f_btwstelsel_grond: toepassing van het BTW stelsel op grondaandeel
39. f_forfait_aansluitingskosten: forfait aansluitingskosten
40. f_forfait_basisakte_notaris: forfait basisakte notaris
41. f_huurwaarborg: waarde van de waarborg die de huurder moet betalen bij afsluiten huurcontract
42. f_jaarhuur: waarde van de huur die de huurder per jaar moet betalen
43. f_ki: kadastraal inkomen
44. f_kiindex: index van het kadastraal inkomen
45. f_kiindex_jaar: het jaar waarop het getal vermeld bij het geïndexeerd kadastraal inkomen (f_kiindex) het laatst werd geïndexeerd
46. *f_locatie_id: deze parameter wordt niet meer gebruikt in zimmo of max-immo*
47. f_maandelijks: maandelijks terugkerende kosten
48. f_nettoopbrengst: geschatte opbrengst na verhuring van een pand
49. f_onroerendevoorheffing: onroerende voorheffing
50. f_onroervh: maandelijkse onroerende voorheffing
51. f_onroervh_jaar: onroerende voorheffing op jaarbasis
52. f_overnameprijs: gevraagde prijs bij een pand met statust ‘over te nemen’ of ‘overgenomen’
53. f_prijs: de gevraagde huurprijs of verkoopprijs
54. f_prijsgrondaandeel: waarde van de grond bij verkoop van nieuwbouwproject
55. f_prijsvm: prijs per vierkante meter
56. f_prijszichtbaar
57. f_provisie: bedrag van de provisie op het pand
58. is_gearchiveerd: wanneer een pand verhuurd/verkocht is of het behoort niet meer tot de portefeuille van de makelaar, dan wordt het gearchiveerd (staat het niet meer op Zimmo)
59. *n_ligging: deze parameter wordt niet meer gebruikt in Zimmo of max-immo*
60. n_project_id: id van het nieuwbouwproject

61. notaris_bouquet
62. notaris_instelprijs
63. notaris_max_looptijd_lijfrente
64. notaris_plaats_openbareverkoop
65. notaris_prijs_gebracht_op
66. notaris_recht_hoger_bod_datum
67. notaris_rente
68. notaris_toegewezen_ovhb_aan
69. notaris_verkooptype_id
70. o_garages: aantal garages aanwezig
71. o_openbaarvervoer: afstand tot openbaar vervoer
72. o_orientatie_id: oriëntatie
73. o_parkings: aantal parkings aanwezig
74. o_school: afstand tot scholen
75. o_strand: afstand tot strand
76. o_tuinaanwezig: tuin aanwezig
77. o_tuintekst_nl: beschrijving van de tuin
78. o_winkels: afstand tot winkels
79. o_zichtopzee: zicht op zee
80. o_zijdelingszichtopzee: zijdelings zicht op zee
81. r_datum: datum waarop het pand naar het archief werd verplaatst (dus waarop is_gearchiveerd = 1 werd gesteld)
82. r_prijs: prijs waaraan het pand uiteindelijk verkocht of verhuurd is
83. showweb: als deze variabele waarde 1 heeft, wordt het pand publiek bekend gemaakt (publicatie op website, naar portalen, …)
84. wijziging_datum: datum van de laatste wijziging van de gegevens van het pand
85. Pub1_start: startdatum van een publicatie van een pand
86. Pub1_stop: einddatum van een publicatie van een pand
87. Pub2_start: startdatum bij nieuwe publicatie
88. Pub2_stop: einddatum bij nieuwe publicatie
89. Pub3_start: enz…
90. Pub3_stop
91. Pub4_start
92. Pub4_stop
93. Pub5_start
94. Pub5_stop
95. Pub6_start
96. Pub6_stop
97. Pub7_start
98. Pub7_stop
99. Pub8_start
100. Pub8_stop
101. Pub9_start
102. Pub9_stop
103. Pub10_start
104. Pub10_stop1

De velden worden ingevuld door vastgoedmakelaars, particuliere eigenaars en crawlers[2]. Er dient rekening te worden gehouden met het feit bij het invullen van velden door de makelaar of particuliere

[2] Een crawler wordt ook wel wanderer, bot or robot genoemd. Dit zijn de softwareprogramma`s / algoritmen die op trefwoorden het internet afzoeken en de bekomen gegevens opslaan in de database

eigenaar, zij bepalen of iets van toepassing is voor hun pand of niet en dit is dus subjectieve invoer. Bij invoer via crawling wordt door Z-immo bepaald welke invulling een variabele krijgt naargelang wat wordt getoond op de website voor een pand. De invulling bestaat dan uit een vast aantal waarden. Normaal gezien is hier geen eigen invoer mogelijk.

Op de invoer vanuit crawling zit een (automatische) kwaliteitscontrole voordat de data in de database wordt verwerkt. De kwaliteitscontrole omvat de juiste toekenning van type, status en adres. Verder wordt nagegaan of de juiste waarden worden ingevuld: een tekst waar tekst wordt verwacht, een getal of cijfercombinatie waar enkel cijfers worden verwacht, etc. Deze kwaliteitscontrole verbetert de databank, maar garandeert niet dat velden niet foutief zijn ingevuld (typfouten, onrealistische waarden,…). *Een voorbeeld is 'renovatiejaar' dat niet consequent ingevuld is met 4 cijfers, maar ook waarden omvat zoals 2, -2, 30, 19, 1994.2, 21012, …. De invoer van dit veld is via de software wel beperkt tot numerieke waarden, maar controleert bv. niet het aantal cijfers. Ook is het zo dat waarden zoals '0, niet aanwezig, geen data' niet altijd eenduidig te interpreteren zijn. Bij de variabele o_tuinaanwezig betekent 0 bijvoorbeeld zowel 'geen tuin' als 'geen informatie bekend'. Enkel wanneer wordt ingevoerd dat een tuin aanwezig is, bevat deze variabele de waarde '1'. Voor andere variabelen kan 0 een standaard waarde zijn, maar het kan ook 'Nee' betekenen, 'niet aanwezig' of 'geen data'. Het is afhankelijk van veld tot veld.*

Met de verschillende variabelen en de wijze waarop de velden zijn ingevuld binnen deze ruwe databank, dient met andere woorden omzichtig omgesprongen te worden. Verderop (zie DEEL 2 - Eerste analyse van variabelen) wordt een eerste exploratieve analyse op de variabelen verder toegelicht.

De databank omvat ook veel dubbele records, namelijk panden die aan eenzelfde adres gelinkt zijn. Dit heeft meerdere oorzaken en dit wordt nader toegelicht onder §7 *Onderzoek naar unieke adressen*.

DEEL 1 - Geolocalisatie

1. Voorbereiding adresatlassen van de gewesten

De volgende 3 adresbestanden van de gewesten worden ingebracht en klaar gemaakt voor gebruik bij de geolocalisatie van de immo adressen :

Vlaanderen :

Inbreng van de CRAB / GRB bestanden die zullen gebruikt worden :

- Gbg : Deze shape bevat de geometrie van de GRB gebouwen aan de grond.
- Adp : Deze shape bevat de geometrie van de administratieve kadastrale percelen.
- TblGbgAdr : Deze tabel legt de relatie tussen de gebouwen in de Gbg shape en de CRAB adressen (zie verder). Enerzijds kan één gebouw meerdere adressen hebben en anderzijds kan aan één adres aan meerdere gebouwen gelinkt zijn.
- CRAB_adresposities : De gegevens (postcode / straatnaam / huisnr) en x,y locatie van de adressen.

Brussel :

Inbreng van de Urbis bestanden die zullen gebruikt worden :

- UrbAdm adres database
- Urb_Adm_BUILDING grafische gebouwen met de contouren en de oppervlaktes van de gebouwen
- URB_P_CAPA omvat de kadastrale perceelsgegevens.

Wallonië :

Inbreng van PICC bestand

- het bestand met x,y adrespunten.
- PICC_CONSTR_BATIEMPRISE omvat de contouren van de gebouwen. De immo adrespunten zijn enkel in het hoofdgebouw gelegen en er is geen link tussen hoofdgebouw en bijgebouwen. Bijgevolg kan de oppervlakte van het hoofdgebouw op basis van deze gegevens niet bepaald worden.
- Cadmap_wallonie_20160101 omvat de kadastrale perceelsgegevens

2. Inbreng van het immo bestand

De immo databank met 2.900.979 records werd aangeleverd door Ruimte Vlaanderen onder de vorm van een .csv bestand : export.csv.

Daar dit databestand te omvangrijk was voor MS Access was het allereerst noodzakelijk om dit op te splitsen. Deze verdeling werd eenvoudig gedaan op postcode-duizendtal (query op het veld a_postcode = "#???" met # een cijfer 1-9 en ? een willekeurig cijfer). Dit heeft bijkomend het voordeel dat op die manier de onmogelijke postcodes geëlimineerd worden (bvb a_postcode leeg, buitenlandse postcodes met meer dan 4 cijfers, …) Hierbij dient opgemerkt te worden dat, daar op het einde van de verwerking alle gegevens terug in 1 bestand worden ondergebracht, deze verdeling geen enkele invloed heeft op het eindresultaat van de geolocalisatie.

Tabel 1 – verdeling van de data over de postcodes

Postcode	Provincie	Aantal records	
1000	VL-Brab (gedeelte) Wal-Brab, Brussel	639,965	22.06%
2000	Antwerpen	491,562	16.94%
3000	Limburg VL-Brab (gedeelte)	316,695	10.92%
4000	Luik	104,595	3.61%
5000	Namen	88,509	3.05%
6000	Luxemburg	112,737	3.89%
7000	Henegouwen	140,040	4.83%
8000	West-VL	569,476	19.63%
9000	Oost-VL	328,277	11.32%
Other		109,123	3.76%
Totaal :		**2,900,979**	

Hierbij wordt met "Other" bedoeld de records waarbij veld a_postcode niet beantwoordt aan het formaat "#???" (met # een cijfer 1-9 en ? een willekeurig cijfer). Voorbeelden daarvan zijn : a_postcode leeg, buitenlandse postcodes, …

Eventuele buitenlandse gemeentes in de x000 bestanden worden gedetecteerd bij de normalisatie van de postcode / gemeente.

Uit bovenstaande tabel blijkt dat de vertegenwoordiging van Vlaanderen en Brussel duidelijk groter is dan van Wallonië. Zimmo heeft zich in het verleden ook vooral op Vlaanderen en Brussel geconcentreerd. Dezelfde conclusie geldt ook na normalisatie (zie tabel 4).

3. Normalisatie van de immo adresinformatie

De normalisatie gebeurt in2 stappen :

- Eerst wordt gecontroleerd of de postcode (het veld a_postcode) bestaat en zo ja, of deze overeenkomt met a_gemeente. Deze controle gebeurt aan de hand van de standaard lijst met Belgische postcodes.
- Vervolgens wordt nagegaan of en, zo ja met welke, adresinformatie (postcode / straatnaam) in de adres atlassen van de gewesten de adresinformatie a_postcode / a_straat van het immo adresbestand overeenkomt. Daarbij wordt getracht om de typische problemen van het verschil in spelling tussen de straatnamen in het immo bestand met deze in de adres atlassen van de gewesten en de aanwezigheid van fouten in de straatnamen in het immo bestand op te lossen.

3.1 Normalisatie van de Postcode

Er wordt gecontroleerd of de a_postcode bestaat en zo ja, of deze overeenkomt met a_gemeente.

Fouten van onlogische postcode/gemeentes combinaties of buitenlandse gemeenten worden zo geëlimineerd. Als de postcode/gemeentes combinatie niet voorkomt in de officiële lijsten dan wordt dit (semi-)handmatig gewijzigd

Hierbij worden de volgende velden ingevuld :

- immo_city = de genormaliseerde gemeentenaam
- immo_postcode = de genormaliseerde postcode
- immo_Remark_PostCode = code die het resultaat van de controle aangeeft :
    - 0 : 100% OK, geen correctie
    - 1 : gemeentenaam aangepast
    - 2 : postcode aangepast
    - 3 : gemeentenaam + postcode aangepast
    - 9 : correctie onmogelijk
- immo_OK_PostCode = Waar als a_postcode / a_gemeente succesvol kon genormaliseerd worden.
- immo_Checked_Postcode = dit veld wordt enkel gebruikt gedurende de normalisatie (heeft dus geen verdere betekenis)

De volgende tabel toont het resultaat :

Tabel 2 – Resultaten van de controle op postcode

Postcode	Aantal records	OK Postcode = Waar	%
1000	639.965	639.699	99,96%
2000	491.562	491.303	99,95%
3000	316.695	316.463	99,93%
4000	104.595	103.972	99,40%
5000	88.509	88.259	99,72%
6000	112.737	112.416	99,72%

Postcode	Aantal records	OK Postcode = Waar	%
7000	140.040	139.609	99,69%
8000	569.476	569.327	99,97%
9000	328.277	328.168	99,97%
Totaal :	**2.791.856**	**2.789.216**	**99,91%**

3.2 Normalisatie van de Straatnaam

Een eerste screening toont aan dat er veel records zijn waarbij het veld a_street leeg is, = “onbekend” of = “inconnu” :

Tabel 3 – Resultaten van de controle op adressen

	Aantal	a_straat leeg	a_straat= Onbekend	a_straat= Inconnu	a_straat leeg of onbekend	a_straat bruikbaar
1000	639.965	397.576	32.535	33.098	463.209	176.756
2000	491.562	140.527	33.781	134	174.442	317.120
3000	316.695	79.203	29.068	539	108.810	207.885
4000	104.595	55.463	5.665	4.155	65.283	39.312
5000	88.509	49.494	6.285	3.106	58.885	29.624
6000	112.737	62.038	9.389	4.246	75.673	37.064
7000	140.040	84.493	7.577	3.466	95.536	44.504
8000	569.476	130.525	40.203	141	170.869	398.607
9000	328.277	91.116	28.219	33	119.368	208.909
other	109.123					
Totaal :	2.900.979					1.459.781

Voor maximaal 1.459.781 records kan dus de straatnaam genormaliseerd worden.

Matching van de velden a_postcode en a_straat met deze van de atlassen van de gewesten

Dit betreft de matching van de adresinformatie a_postcode / a_straat van het immo adresbestand met de adresinformatie (postcode / straatnaam) in de adres atlassen van de gewesten.

De typische problemen bij deze matching zijn het verschil in spelling tussen de straatnamen in het immo bestand met deze in de adres atlassen van de gewesten en de aanwezigheid van fouten in de straatnamen in het immo bestand. Om dit zoveel als mogelijk te corrigeren werden fonetische zoek modules ontwikkeld.

Bij deze normalisatie worden de volgende velden ingevuld :

- immo_Phonetic_Street = fonetisch gedeelte van de straatnaam (enkel gebruikt bij de normalisatie)
- immo _Street = de genormaliseerde straatnaam
- immo _Remark_street = code die het resultaat van de normalisatie van de straatnaam aangeeft :
    - 0 : 100% OK, geen correctie
    - 1 : straatnaam aangepast
    - 2 : postcode aangepast
    - 3 : straatnaam + postcode aangepast
    - 9 : straatnaam niet gevonden

- immo _OK_Street = Waar als a_straat succesvol kon genormaliseerd worden.
- immo_Checked_Street = dit veld wordt enkel gebruikt gedurende de normalisatie (heeft dus geen verdere betekenis)

Deze normalisatie van de straatnamen geeft de volgende resultaten :

Tabel 4 – resultaten van de normalisatie

	Aantal	a_straat bruikbaar	%	a_postcode a_straat OK	%	Verdeling over de postcodes (a_postcode en a_straat OK)
1000	639.965	176.756	27,6%	145.425	22,7%	11%
2000	491.562	317.120	64,5%	291.998	59,4%	22%
3000	316.695	207.885	65,6%	189.528	59,8%	15%
4000	104.595	39.312	37,6%	32.571	31,1%	3%
5000	88.509	29.624	33,5%	24.631	27,8%	2%
6000	112.737	37.064	32,9%	30.722	27,3%	2%
7000	140.040	44.504	31,8%	36.857	26,3%	3%
8000	569.476	398.607	70,0%	356.487	62,6%	27%
9000	328.277	208.909	63,6%	190.176	57,9%	15%
other	109.123					
Totaal :	2.900.979	1.459.781	50,3%	1.298.395	44,8%	100%

Uit de normalisatie kunnen we concluderen dat van een totaal van 2.900.979 records ongeveer 45% een bruikbare straatnaam en postcode oplevert.

4. Geocodering van de immo adressen

Voor ieder immo adres wordt in de betreffende straat het huisnr (veld a_nummer) opgezocht in de adresatlas van de gewesten.
Voor de adressen waarvan het huisnummer a_nummer niet aanwezig is in de adresatlas van de gewesten wordt het numeriek dichtstbije huisnr genomen. Van deze laatste adressen wordt in een latere stap geprobeerd om via de immo velden a_geo_lat: breedteligging en a_geo_lon: lengteligging een geografisch overeenkomend adres te vinden in de adresatlas van de gewesten.

Van deze resulterende huisnrs wordt dan de x,y coördinaat uit de adresatlas van de gewesten gerecupereerd.
Hierbij worden dan de volgende velden ingevuld

- immo_Hnr = het gebruikte huisnr van de adresatlas van de gewesten
- immo_StreetID = het unieke ID van de straat binnen de adresatlas van de gewesten :
    - Voor CRAB adressen (Vlaanderen) : “Cppppiiiiiiiii” met pppp = postcode en iiiiiiiii het Crab StraatnmID eventueel opgevuld met voorafgaande nullen.
    - Voor PICC adressen (Wallonië) : “P” + PICC ID, eventueel opgevuld met voorafgaande nullen.
    - Voor Urbis adressen (Brussel) : “U” + Urbis ID, eventueel opgevuld met voorafgaande nullen.
- immo_X_lamb, immo_Y_lamb = x, y coördinaten van het adres volgens de adresatlas van de gewesten.
- immo_Result_Geocode = code die het resultaat van de geocodering aangeeft :
    - 0 : het adres (met het huisnummer) is aanwezig in de adresatlas van de gewesten.
    - 1 : het huisnummer wordt niet teruggevonden (in dit geval wordt het numeriek minst verschillende huisnr genomen)
    - 2 : het huisnummer wordt niet teruggevonden en er zijn geen huizen gekend in deze straat waardoor het dichtstbij gelegen huisnr niet kon genomen worden.
- immo_OK_Geocode = Waar als het adres succesvol kon gegeocodeerd worden.

Deze geocodering geeft de volgende resultaten:

Tabel 5 – resultaten van de geocodering

	Aantal	a_straat bruikbaar	a_postcode a_straat OK	Geo-codering OK	immo_Result_Geocode		
					0	1	2
1000	639,965	176,756	145.425	145.425	112.823	32.602	0
2000	491,562	317,120	291.998	291.998	238.576	53.422	0
3000	316,695	207,885	189.528	189.527	152.784	36.743	0
4000	104,595	39,312	32.571	32.571	22.370	10.201	0
5000	88,509	29,624	24.631	24.629	14.249	10.380	0
6000	112,737	37,064	30.722	30.720	20.267	10.453	0
7000	140,040	44,504	36.857	36.857	25.643	11.214	0
8000	569,476	398,607	356.487	356.487	295.833	60.654	0
9000	328,277	208,909	190.176	190.176	152.576	37.600	0
other	109,123						
Totaal :	2,900,979	1,459,781	1.298.395	1.298.390	1.035.121	263.269	0

4.1 Controle van coördinaten

In deze stap worden de geografische coördinaten a_geo_lon en a_geo_lat van de immo adressen vergeleken met de x en y coördinaten uit de atlassen van de gewesten. Daarbij worden de a_geo_lon en a_geo_lat eerst omgerekend naar het Lambert 72 coördinatenstelsel.

Het resultaat daarvan is de invulling van de volgende velden in de MS Access tabel met het immo adresbestand :

- immo_X_lamb_input : naar Lambert 72 omgerekende a_geo_lon
- immo_Y_lamb_input : naar Lambert 72 omgerekende a_geo_lat
- immo_Geo_diff_distance : de afstand in meters tussen de geografische locatie in de immo databank en de atlas van de gewesten.

4.2 Diverse aanpassingen

In deze fase worden de volgende velden ingevuld :

- invullen immo_Precisie_Locatie :
  - 1 : adres (postcode, straatnaam, huisnr) volledig correct
  - 2 : enkel postcode, straatnaam correct waarbij het numeriek minst verschillende huisnr werd genomen
  - 3 : enkel postcode correct
  - -1 : niets correct
  - (in een latere fase wordt daarbij ook het cijfer 4 toegevoegd : zie verder bij “Corrigeren adressen met ongekend huisnr adv de x,y coordinaten”)
- invullen immo_Uniek_ID : uniek nr per record
- invullen datumvelden met correct formaat (= MS Access datum veld) :
  - immo_vrijop_datum = a_vrijop_datum
  - immo_creatie_datum = creatie_datum
  - immo_r_datum = r_datum
  - immo_wijziging_datum = wijziging_datum
  - immo_Pub1_start = Pub1_start
  - immo_Pub1_stop = Pub1_stop
- invullen immo_Pub1_dagen = aantal dagen tussen immo_Pub1_start en immo_Pub1_stop
- invullen prijsvelden met correct formaat (= MS Access datum veld)
  - immo_f_ki = f_ki
  - immo_f_kiindex = f_kiindex
  - immo_f_prijs = f_prijs
  - immo_r_prijs = r_prijs
- invullen oppvelden met correct formaat (= MS Access getal veld):
  - immo_bouwopp = b_bouwopp
  - immo_grondopp = b_grondopp
  - immo_perceelbreedte = b_perceelbreedte
  - immo_perceeldiepte = b_perceeldiepte
  - immo_woonopp = b_woonopp

4.3 Corrigeren adressen met ongekend huisnr. adv de x,y coördinaten

Voor de adressen met een niet gekend huisnr (immo_Result_Geocode = 1) werd een bijkomende poging gedaan om het overeenkomstig adres te vinden in de adresatlas van de gewesten . Dit gebeurde op de volgende manier :

- in de immo databank zijn de geografische breedte en lengte van het adres aanwezig : dit zijn de velden a_geo_lat, a_geo_lon
- deze werden omgerekend naar het Lambert 72 coördinatenstelsel : dit zijn de velden immo_X_lamb_input, immo_Y_lamb_input
- in de de atlas van de gewesten met adrespunten zijn voor ieder adres de x,y Lambert 72 coördinaten aanwezig
- indien het punt immo_X_lamb_input, immo_Y_lamb_input minder dan 10m verwijderd is van een adrespunt in de atlas van de gewesten dan wordt aangenomen dat het immo adres overeenkomt met dit adres uit de atlas van de gewesten

De volgende tabel geeft een overzicht van de resultaten van deze correctie :

Tabel 6 – resultaten van de correctie van adressen adv x, y coördinaten

	Aantal	immo_Result_ Geocode = 1	Correctie adv x,y
1000	639,965	32.602	3.394
2000	491,562	53.422	3.338
3000	316,695	36.743	2.068
4000	104,595	10.201	1.327
5000	88,509	10.380	582
6000	112,737	10.453	1.143
7000	140,040	11.214	1.546
8000	569,476	60.654	3.579
9000	328,277	37.600	2.698
other	109,123		
Totaal :	2,900,979	263.269	19.675

In deze stap werden de volgende velden ingevuld :

- immo_Hnr = het gevonden huisnr
- immo_Precisie_Locatie = 4
- immo_X_lamb, immo_X_lamb : dit zijn de x,y coordinaten uit de adresatlas van de gewesten van het gevonden huisnr
- immo_Geo_diff_distance wordt herberekend

5. Oppervlaktes van de gebouwen

Voor Vlaamse adressen kon door middel van de GBG tabel TblGbgAdr de relatie tussen de gebouwen in de Gbg shape en de CRAB adressen bekomen worden. Enerzijds kan één gebouw meerdere adressen hebben en anderzijds kan aan één adres aan meerdere gebouwen gelinkt zijn.

In de Gbg_shape bevinden zich de oppervlaktes van de gebouwen waarbij er een onderscheid gemaakt wordt tussen hoofdgebouwen (GRB_type = 1) en bijgebouwen (GRB_type = 2)

Analoog konden voor de adressen in het Brussels gewest de oppervlaktes van de gebouwen bekomen worden uit de Urbis tabel URB_A_BU.

Voor de adressen in het Waals gewest kan de oppervlakte van het hoofdgebouw bekomen worden uit de laag PICC_CONSTR_BATIEMPRISE (gezien de adrespunten in het hoofdgebouw gelegen zijn en geen link is tussen hoofd- en bijgebouw, kunnen de totale bebouwde oppervlakte en de oppervlakte van het bijgebouw niet bepaald worden).

In deze fase werden de volgende velden ingevuld:

- immo_Opp_HfdGebouw : oppervlakte van het hoofdgebouw
- immo_Opp_BijGebouw : oppervlakte van het bijgebouw (niet voor Wallonië)
- immo_Opp_TotGebouw : totaal bebouwde oppervlakte (niet voor Wallonië)

6. Oppervlaktes van de percelen

Om de oppervlaktes van de kadastrale percelen te bekomen voor de adressen in het Vlaams gewest werd het volgende gedaan :

- Bepaling welke perceel er bij een bepaald immo adres behoort :
    - In het GRB is de tabel TblAdpAdr aanwezig die in principe de relatie tussen de percelen in de Adp shape en de CRAB adressen definieert. Deze tabel kan echter niet gebruikt worden daar een adres enkel wordt gerelateerd aan een administratief perceel indien het om een administratief perceel gaat waarop geen gebouwen aan de grond (Gbg) of kunstwerken (Knw) opgetrokken zijn.
    - Om te bepaling welke perceel er bij een bepaald immo adres behoort wordt daarom als volgt tewerk gegaan. Gebruik makende van de shapefile met de puntobjecten van de immo adressen wordt met de GIS tool bepaald in welk perceel in de Adp shape het adres gelegen is.
- De volgende velden worden hierbij ingevuld :
    - immo_ParcelNr = het unieke identificatienummer van het perceel in het Adp GRB bestand waarin het adres gelegen is (adp.OIDN)
    - immo_ParcelKey = sleutel van het kadastrale perceel (adp.CAPAKEY). Deze sleutel wordt genoteerd als vast formaat (fixed format) en is een concatenatie van :
        - kadastraal afdelingsnummer: cijfer (5 posities);
        - kadastrale sectie: hoofdletter (1 positie);
        - grondnummer: cijfer (4 posities);
        - bisnummer: separator “/”, gevolgd door 2 cijfers (3 posities);
        - letterexponent: “_” (indien geen letterexponent) of hoofdletter (1 positie);
        - cijferexponent: cijfer (3 posities) (= macht).
    - immo_ParcelArea = de berekende oppervlakte van het perceel (dit kan verschillen van de oppervlakte uit de kadastrale legger) (adp.OPPERVL)

Voor Brussel bevinden de kadastrale perceelsgegevens zich in het Urbis bestand URB_P_CAPA.Gebruik makende van de GIS laag met de puntobjecten van de immo adressen kan bepaald worden in welk perceel in de URB_P_CAPA laag het adres gelegen is.

Daarbij kunnen dan volgende velden ingevuld worden.

- immo_ParcelNr = het unieke identificatienummer van het perceel in het Adp van Urbis
- immo_ParcelKey = sleutel van het kadastrale perceel (adp.CAPAKEY) + kadastraal afdelingsnR/kadastrale sectie/grondnummer/bisnummer/…
- immo_ParcelArea = de berekende oppervlakte van het perceel (dit kan verschillen van de oppervlakte uit de kadastrale legger) (adp.OPPERVL)

Voor Wallonië bevinden de Kadastrale perceelsgegevens zich in het PICC bestand cadmap_wallonie_20160101.gdb. Gebruik makende van de GIS laag met de puntobjecten van de immo adressen kan bepaald worden in welk perceel in de cadmap_wallonie_20160101.gdb laag het adres gelegen is.

Daarbij kunnen dan ook volgende velden ingevuld worden.

- immo_ParcelNr = het unieke identificatienummer van het perceel in het Adp van PICC
- immo_ParcelKey = sleutel van het kadastrale perceel (adp.CAPAKEY) + kadastraal afdelingsnR/kadastrale sectie/grondnummer/bisnummer/…
- immo_ParcelArea = de berekende oppervlakte van het perceel (dit kan verschillen van de oppervlakte uit de kadastrale legger) (adp.OPPERVL)

7. Onderzoek naar unieke adressen

Uit een onderzoek naar het aantal unieke adressen in het geleverde immo bestand blijkt dat soms meerdere records toebehoren aan 1 adres. (hetzelfde a_postcode, a_straat, a_nummer)

Hiervoor zijn verschillende oorzaken[3][4]

- dubbele invoer van een pand:
    - Indien een verkoop of verhuur verloopt via verschillende makelaars. Dit kan op hetzelfde moment zijn, maar het kan ook zijn dat een pand eerst via Makelaar X en nadien via Makelaar Y verkocht wordt. Dan komt het pand 2x voor in de databank, maar gekoppeld aan een verschillend kantoor.
    - Bij verkoop/verhuur via dezelfde makelaar zijn er ook meerdere oorzaken voor dubbele invoer:
        - Indien de makelaar een nieuwe website heeft, kan geen link gelegd worden aan de hand van ID of referentie, waardoor hetzelfde pand als nieuw pand wordt aangemaakt.
        - Indien een pand eerder verkocht of verhuurd werd en opnieuw op de markt komt.
            - Indien de makelaar zelf zijn panden beheert, kan het record uit het archief gehaald worden. Dan is er geen dubbel pand (maar dit hangt dus af van een actie van de makelaar).
            - Via crawling wordt het pand opnieuw aangemaakt indien het pand bij de vorige verkoop/verhuur meer dan 6 maand geleden van de markt werd gehaald.
- Er zijn meerdere appartementen, studio's in 1 gebouw met hetzelfde adres (het veld a_bus is niet altijd ingevuld) of meerdere delen worden afzonderlijk verhuurd op 1 adres : bvb afzonderlijke kantoren, handelspanden, opslagplaats, magazijn, garage. Zo lijkt het alsof het om dezelfde panden gaat, maar in principe zijn het verschillende panden op dezelfde locatie.
- huisnr = 0
- Een fout in de crawling, waardoor hetzelfde pand regelmatig opnieuw wordt aangemaakt.
- Soms ontstaan dubbels ook in de software, max-immo. Sommige klanten hebben doorheen de tijd twee accounts in deze software waardoor ook om deze reden dubbels ontstaan. Een nieuwe account kan meerdere oorzaken hebben, bijvoorbeeld door de overgang naar crawler, een nieuwe website, …

Er werd een bestand gemaakt welke enkel de unieke adressen bevat (gecombineerde unieke primary key op de velden a_postcode, a_straat, a_nummer). Bij de selectie van de record voor een uniek adres werd telkens de record genomen met de recentste immo_Pub1_start datum. Daar het hier unieke adressen betreft werden vanzelfsprekend enkel de records beschouwd waarin de velden a_postcode, a_straat, a_nummer niet leeg zijn. Hier dient opgemerkt te worden dat, alhoewel dus de velden a_postcode, a_straat, a_nummer niet leeg zijn, dit niet betekent dat voor al deze adressen een overeenkomstig adres kon gevonden worden in de adres atlas van de gewesten. Met andere woorden: in dit bestand met enkel de unieke adressen kan het veld immo_precisie_locatie toch nog 2, 3, 4 of -1 zijn.

[3] De mogelijke oorzaken voor dubbele records werden doorgegeven door Z-immo

[4] Het is niet zo dat een nieuw record wordt gecreëerd bij elke aanpassing (bv. wijziging van de inhoud van een veld). Aanpassingen van panden worden ofwel zelf ingevoerd door de makelaar ofwel via crawling opgepikt. Deze wijzigingen worden aangebracht in het bestaande record en er wordt dan geen nieuw record aangemaakt.

Dit databestand bevat 531.040 records. Deze reductie van 1.298.390 (= aantal records met bruikbare a_postcode, a_straat, a_nummer) naar 531.040 records toont aan dat er zeer veel adressen zijn die meerdere records bevatten in het immo bestand, omwille van verschillende mogelijke redenen, zoals hierboven aangehaald.

8. Overzicht resultaten

8.1 Overzicht toegevoegde velden

De volgende velden werden toegevoegd tijdens de opmaak van het analyse bestand :

Tabel 7 – overzicht van de toegevoegde velden

immo_Uniek_ID	Uniek nummer van de immo_data record
immo_Postcode	Genormaliseerde postcode
immo_City	Genormaliseerde gemeentenaam
immo_StreetId	ID van de straat in de CRAB / Urbis / PICC adressenatlassen
immo_Street	Genormaliseerde straatnaam
immo_Hnr	Genormaliseerde huisnummer (zie veld immo_Precisie_Locatie)
immo_X_lamb	X Lambert coördinaat volgens CRAB / Urbis / PICC adressenatlassen
immo_Y_lamb	Idem Y
immo_ParcelNr	ID Kadastraal perceel in de GRB databank / Urbis / PICC
immo_ParcelKey	Kadastraal perceelnummer in de GRB databank/ Urbis / PICC
immo_ParcelArea	Oppervlakte van het kadastraal perceel in de GRB databank / Urbis / PICC
immo_X_lamb_input	X Lambert coördinaat berekend uit de velden a_geo_lat, a_geo_lon
immo_Y_lamb_input	Idem Y
immo_Geo_diff_distance	Afstand in meter tussen immo_..._lamb en immo_..._lamb_input
immo_Opp_HfdGebouw	Oppervlakte hoofdgebouw volgens GRB / Urbis /PICC
immo_Opp_BijGebouw	Oppervlakte bijgebouw volgens GRB / Urbis (enkel VL en Brus)
immo_Opp_TotGebouw	Totale oppervlakte van de gebouwen volgens GRB / Urbis (enkel VL en Brus)
immo_Remark_PostCode	immo_Remark_PostCode = code die het resultaat van de normalisatie van de postcode / gemeentenaam aangeeft : • 0 : 100% OK, geen correctie • 1 : gemeentenaam aangepast • 2 : postcode aangepast • 3 : gemeentenaam + postcode aangepast 9 : correctie onmogelijk
immo_OK_PostCode	Waar als de postcode / gemeentenaam kon genormaliseerd worden
immo_Remark_Street	immo _Remark_street = code die het resultaat van de normalisatie van de straatnaam aangeeft : • 0 : 100% OK, geen correctie • 1 : straatnaam aangepast • 2 : postcode aangepast • 3 : straatnaam + postcode aangepast 9 : straatnaam niet gevonden
immo_OK_Street	Waar als de straat succesvol kon genormaliseerd worden.
immo_OK_Geocode	Waar als het adres succesvol kon gegeocodeerd worden
immo_Result_Geocode	code die het resultaat van de geocodering aangeeft : • 0 : het adres (met het huisnummer) is aanwezig in de adresatlas van de gewesten. • 1 : het huisnummer niet teruggevonden (in dit geval wordt het dichtstbij gelegen huisnr genomen) In een volgende stap werd geprobeerd om adv immo_X_lamb_input en immo_Y_lamb_input toch nog het huisnr te vinden in de CRAB adrespunten (zie veld immo_Precisie_Locatie)

	2 : het huisnummer niet teruggevonden en er zijn geen huizen gekend in deze straat waardoor het dichtstbij gelegen huisnr niet kon genomen worden.
immo_Precisie_Locatie	Dit veld geeft de geografische precisie van het adres aan : • 1 : adres (postcode, straatnaam, huisnr) volledig correct • 2 : enkel postcode, straatnaam correct en voor het huisnr werd het numeriek minst verschillende huisnr genomen • 3 : enkel postcode correct • 4 : enkel postcode, straatnaam correct en voor het huisnr werd het huisnr uit de CRAB adrespunten genomen waarvan de x, y positie minder dan 10m in afstand verschilt van de immo_X_lamb_input en immo_Y_lamb_input -1 : niets correct van het adres
immo_a_vrijop_datum	Het originele veld a_vrijeop_datum in het correcte formaat.
immo_creatie_datum	idem
immo_r_datum	idem
immo_wijziging_datum	idem
immo_Pub1_start	idem
immo_Pub1_stop	idem
immo_Pub1_dagen	Het verschil in dagen tussen immo_Pub1_start en immo_Pub1_stop
immo_f_ki	Het originele veld f_ki in het correcte formaat.
immo_f_kiindex	idem
immo_f_prijs	idem
immo_r_prijs	idem
immo_b_bouwopp	idem
immo_b_grondopp	idem
immo_b_perceelbreedte	idem
immo_b_perceeldiepte	idem
immo_b_woonopp	idem

8.2 Overzichtstabel resultaten

De volgende overzichtstabel geeft het aantal adressen volgens de resultaten van de normalisatie en de geocodering en dit in de volgorde van stijgende nauwkeurigheid :

Tabel 8 – overzichtstabel van de resultaten van de normalisatie en de geocodering

Ontvangen immo_data bestand	2.900.979
Postcode formaat is 4 cijfers	2.791.856
immo_Precisie_Locatie = -1 : niets correct van het adres	2.640
immo_Precisie_Locatie = 3 : enkel postcode correct	1.490.826
immo_Precisie_Locatie = 2 : enkel postcode & straatnaam correct en voor het huisnr werd het numeriek minst verschillende huisnr genomen	243.594
immo_Precisie_Locatie = 4 : enkel postcode & straatnaam correct en voor het huisnr werd het huisnr uit de CRAB adrespunten genomen waarvan de x, y positie minder dan 10m in afstand verschilt van de immo_X_lamb_input en immo_Y_lamb_input	19.675
immo_Precisie_Locatie = 1 : adres (postcode, straatnaam, huisnr) volledig correct	1.035.121
Selectie unieke adressen op basis van recentste immo_Pub1_start datum en velden a_postcode, a_straat, a_nummer zijn niet leeg	531.040

DEEL 2 - Eerste analyse van variabelen

1. Exploratie van variabelen

Dit hoofdstuk omvat een eerste exploratie van de immodatabank, enerzijds als geheel, anderzijds toegespitst op de casus Oost-vlaanderen. In eerste instantie wordt de immodatabank in zijn geheel onderzocht op de non-respons graad van de verschillende beschikbare variabelen (§1.2). Dit geeft een eerste ruwe indicatie van de bruikbaarheid van de databank incl. mogelijke valkuilen, gezien meerdere variabelen weinig respons vertonen. Via een verdere analyse voor de casus Oost-Vlaanderen, wordt de (mate van) bruikbaarheid van de verschillende variabelen onderzocht en wordt bepaald welke variabelenset verder wordt opgenomen in de analyse (§1.3 en 1.4). Deze variabelenset wordt aan verdere analyse onderworpen (o.a. het in beeld brengen van outliers en samenhang van variabelen), zie §1.5.

1.1 Algemene beschouwingen

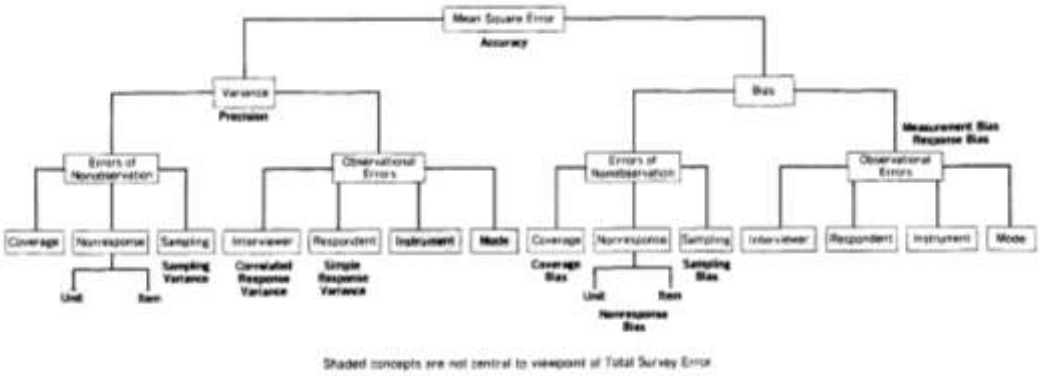
De Z-immo databank omvat een veelheid van informatie, vanuit verschillende bronnen. De databank wordt ingevuld door zowel vastgoedmakelaars, particuliere eigenaars en crawlers. Door deze verschillende wijze van invoer (automatische en ook subjectieve invoer) dient omzichtig met de databank omgesprongen te worden. Men dient zich bewust te zijn van mogelijke afwijkingen of lacunes in de beschikbare immodataset die een vertekend beeld zouden kunnen genereren t.o.v. “de reële” situatie landsbreed.

Enkele vraagstellingen die hiermee verband houden zijn:

-Hoeveel % van de ‘immotransacties’ verloopt via Zimmo? (dit wordt nader in beeld gebracht onder deel 3)
-Welk soort van data zitten niet of nauwelijks in deze immodata: woningen, gronden, appartementen, … (en eventueel subcategorieën, zoals carports, kastelen, hoven, … ?) Welke (categorieën) zijn ondervertegenwoordigd?
-Verhuur %, verkoop % … is dit wat verwacht kan/mag worden wanneer het ganse grondgebied beschouwd wordt?
-En wordt het ganse grondgebied beschouwd (gesampeld) of komen bepaalde gebieden niet/nauwelijks/ondermaats aan bod? (‘coverage’-aspect)
-…

Deze en vergelijkbare vraagstellingen hebben betrekking op de effecten van 1/ coverage, 2/ non-respons en 3/ sampling aspecten. Ze zijn gerelateerd aan nauwkeurigheid via MSE (mean square error) zoals in onderstaand diagram verduidelijkt. Het is belangrijk om een onderscheid te maken tussen de variantie in de data en de bias die er bovenop zit en het beeld dreigt te verstoren en distorties (in resultaten en conclusies) dreigt te veroorzaken.

Figuur 2 – diagram MSE



In kader van deze studie kunnen niet al de facetten zoals hierboven vermeld onderzocht worden. Wel is het doel om via een eerste exploratieve analyse een aantal zaken sterker in beeld te krijgen in functie van de bruikbaarheid van deze immodatabank. Er worden concrete analyses uitgevoerd die hier meer licht op werpen. Met het gebruik van de immodatabank van Zimmo dienen echter hoe dan ook enkele aannames te gebeuren omtrent de representativiteit van de data, vooral inzake coverage en unit-non-respons. Facetten m.b.t. sampling bias of item non-respons komen verderop wel aan bod. Voor sampling bias hangt dit weerom samen met de aanname van een voldoende representativiteit (en betrouwbaarheid) van de beschikbare dataset.

1.2 Globale (non-)respons rate en doelgerichte selectie van relevante variabelen

De verkenning van de data en in het bijzonder van de beschikbare variabelen is erop gericht om belangrijke lacunes en ontbrekende elementen in grote lijnen in kaart te brengen. In eerste instantie gebeurt dit in relatie tot de *respons rate* uitgedrukt als percentage respons (nl. de mate waarin een variabele is ingevuld). Vervolgens wordt ook ingeschat in welke mate de basiskwaliteit van de dataset verdere detailanalyses toelaat en hoe eventuele extra data-acquisitie de analysemogelijkheden significant zou kunnen verhogen.

Per (relevante) variabele wordt bekeken wat de mogelijkheden zijn voor behoud i.k.v. verdere analyse, voor eventuele aanvulling. Tegelijk wordt ook een hiërarchie binnen de variabelen opgesteld; bv. welke variabelen zijn van elementair belang, welke zijn mogelijks te beschouwen als surrogaatvariabele, welke zijn onderling afhankelijk/gerelateerd, en kunnen eventueel op hoger hiërarchisch niveau samen beschouwd worden. Waar zitten er opvallende lacunes? Wat is de algemene betrouwbaarheid van de data/variabelen?

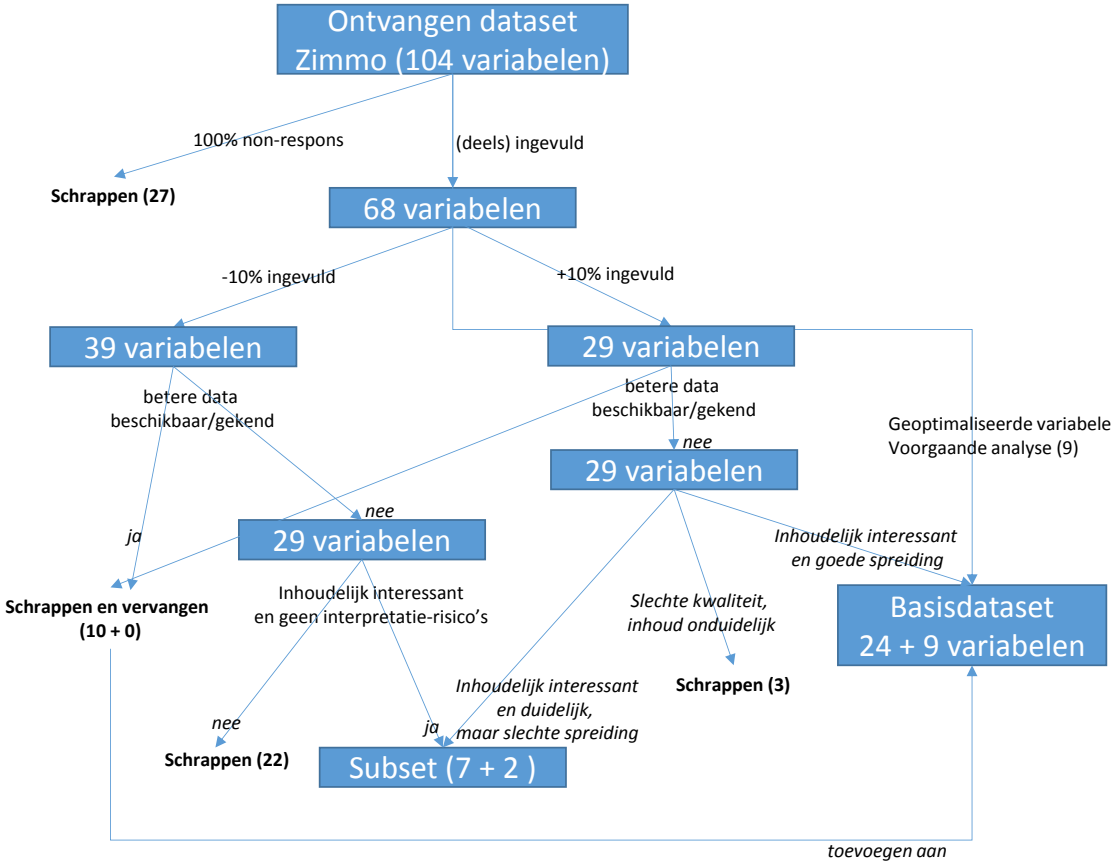
Enkele voorbeelden van lacunes en mogelijke aanvullingen (imputaties) worden hieronder weergegeven. Dit wordt verderop meer in detail besproken.

- ⇨ Voorzieningen zoals ‘openbaar vervoer’ of ‘school’ zijn vaak niet ingevuld. Dit zou automatisch kunnen worden bijgepast op basis van beschikbare data (bv. De Lijn, NMBS, data knooppuntwaarde), zodat deze lacune kan worden opgevuld[5];
- ⇨ Het veld ‘a_vrijop_datum ‘is vaak niet ingevuld en kan mogelijks niet bijkomend worden geimputeerd op basis van koppeling met andere databanken. In dit geval is het weinig zinvol dit veld nog verder mee te nemen in het uiteindelijke analysebestand;
- ⇨ Type gebouw: hier is een verdere aftoetsing mogelijk op basis van de gebouwtypes zoals opgenomen in het GRB, Urbis en Picc.

[5] Er wordt in kader van dit project een koppeling gemaakt met de gegevens omtrent knooppuntwaarde. Dit is een index die berekend is op basis van gegevens over bereikbaarheid van een plek via openbaar vervoer (vito, 2016)

Er werd een lijst van 104 variabelen verkregen. Om een uniforme aanpak te hanteren en de variabelen op consequente wijze te selecteren, werd onderstaand stroomschema opgesteld dat we voor al de variabelen doorlopen. Verderop lichten we elke stap van het schema verder in detail toe en geven we aan welke variabelen werden weerhouden, geschrapt, of geselecteerd[6].

Figuur 3 - Stroomschema voor selectie van variabelen (aantallen indicatief op basis van huidige selectie-voorstel)



1.2.1 Geoptimaliseerde variabelen voorgaande analyse

De variabelen die sowieso in de basisdataset behouden blijven, zijn de geoptimaliseerde variabelen vanuit de voorgaande analyse. Het gaat hierbij over onderstaande 9 variabelen.

Tabel 9 – geoptimaliseerde variabelen voorgaande analyse

Geoptimaliseerde variabele	% ingevuld
a_bus	8,3
a_gemeente	99,7
a_geo_lat	99,0
a_geo_lon	99,0

[6] Het betreft hier de eerste screening van de dataset. Verderop in dit rapport worden verschillende variabelen verder getest op hun bruikbaarheid

Geoptimaliseerde variabele	% ingevuld
a_nummer	46,4
a_postcode	100,0
a_straat	60,9
a_geo_precision	99,7
a_land_id	99,8

We voeren de analyse mbt geschiktheid van variabelen dus verder uit op de overblijvende 95 (=104-9) variabelen.

1.2.2 Globale item non-respons-analyse

De eerste analyse om te bepalen welke variabelen zinvol zijn om te behouden, is de globale item non-respons analyse, namelijk een analyse van de 'missing data' (het in beeld brengen van alle lege cellen per variabele ("IsNull" en "" [leeg]). Dit is ook grafisch samengevat, zie bijlage 1 'nonrespons'.

We schrappen hierbij de variabelen die landsbreed voor 100% niet ingevuld zijn. Uit deze eerste stap worden nog 68 variabelen behouden. De 27 variabelen die geschrapt worden zijn in onderstaande tabel samengevat.

Tabel 10 – Variabelen met 100% non-respons

Variabelen met landsbreed 100% Non-Respons
Netto-opbrengst
Overnameprijs
n_ligging
notaris_bouquet
notaris_instelprijs
notaris_max_looptijd_lijfrente
notaris_plaats_openbareverkoop
notaris_prijs_gebracht_op
notaris_recht_hoger_bod_datum
notaris_rente
notaris_toegewezen_ovhb_aan
Pub3_start
Pub3_stop
Pub4_start
Pub4_stop
Pub5_start
Pub5_stop
Pub6_start
Pub6_stop
Pub7_start
Pub7_stop
Pub8_start
Pub8_stop
Pub9_start
Pub9_stop
Pub10_start
Pub10_stop1

1.2.3 Minder dan 10% ingevuld

Een aantal variabelen zijn ingevuld, maar slechts in beperkte mate. Dit maakt de variabelen weinig bruikbaar voor een gebiedsdekkende data-analyse. Dit betekent echter niet dat deze variabelen niet bruikbaar of zinvol kunnen zijn. In sommige gevallen kunnen ze bv. wel geschikt zijn voor een analyse op lokaal niveau. In andere gevallen zijn alternatieve spatiale datasets voorhanden waardoor het interessanter is de variabele uit de Zimmo-databank te schrappen en te vervangen op basis van een combinatie met zulke datasets (binnen dit project wordt o.a. een koppeling gemaakt met data omtrent knooppuntwaarde, uiteraard zijn meer combinaties met andere datasets mogelijk).

Voor de 39 variabelen die <10% ingevuld zijn, zijn de volgende opties opgenomen:

- De variabele wordt geschrapt en vervangen indien betere spatiale datasets beschikbaar zijn (oranje variabelen in de tabel) – 10 variabelen
- De variabele wordt geschrapt indien de variabele inhoudelijk niet interessant of bruikbaar is (bv. zeer laag percentage, hoog risico voor verkeerde interpretaties, niet zinvol - rode variabelen in de tabel) – 22 variabelen
- De variabele wordt behouden in een subset indien de variabele op het eerste zicht inhoudelijk interessant en bruikbaar lijkt (blauwe variabelen in de tabel, deze kunnen alsnog in de basisdataset worden opgenomen indien na verdere analyse blijkt dat dit zinvol is) – 7 variabelen

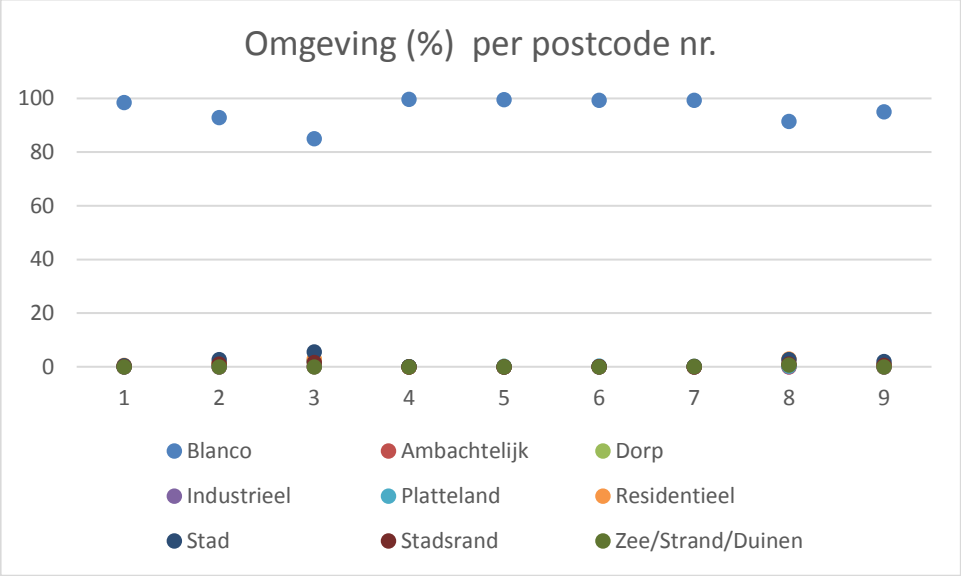
Tabel 11 – variabelen die minder dan 10% zijn ingevuld of een slechte spreiding kennen

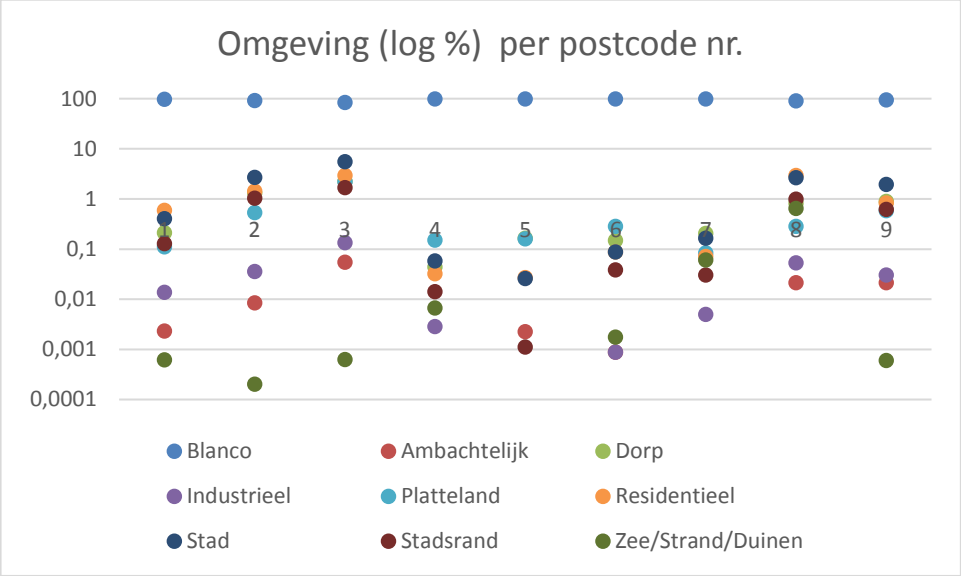
Variabelen die minder dan 10% zijn ingevuld en/of slechte spreiding kennen	Percentage ingevuld	Opmerking
a_omgeving_id	5,7	Te weinig egaal gespreide data – vervangbaar via spatiale data (bv. landgebruikskaart, vectoriële topokaart)
a_sleutelnummer	0,6	Niet zinvol (interne informatie Zimmo)
a_vrijop_datum	3,9	Niet zinvol
a_vrijop2_id	7,1	Niet zinvol
b_bouwopp	4,7	
b_kadastraleaard	0,8	GRB
b_kadastraleafdeling	1,2	GRB
b_kadastralesectie	1,5	GRB
b_perceelbreedte	3,5	
b_perceeldiepte	1,9	
b_perceelnummer	1,5	vervangbaar via koppeling met GRB
b_renovatiejaar	3,5	Risico voor interpretatie – zeer veel foutief ingevuld niet egaal ingevuld, laag %, hooguit beperkt provinciaal
b_staat_id	2,9	niet egaal ingevuld, laag %, hooguit beperkt provinciaal Voor interpretatie vatbaar 'op te frissen' 'in goede staat', 'in uitstekende staat', 'nieuwe staat', 'vernieuwd', …
btwstelsel_grond	7,0	
forfait_aansluitingskosten	0,1	Zeer laag percentage
forfait_basisakte_notaris	0,3	Zeer laag percentage
Huurwaarborg	0,5	Zeer laag percentage
Jaarhuur	0,1	Zeer laag percentage

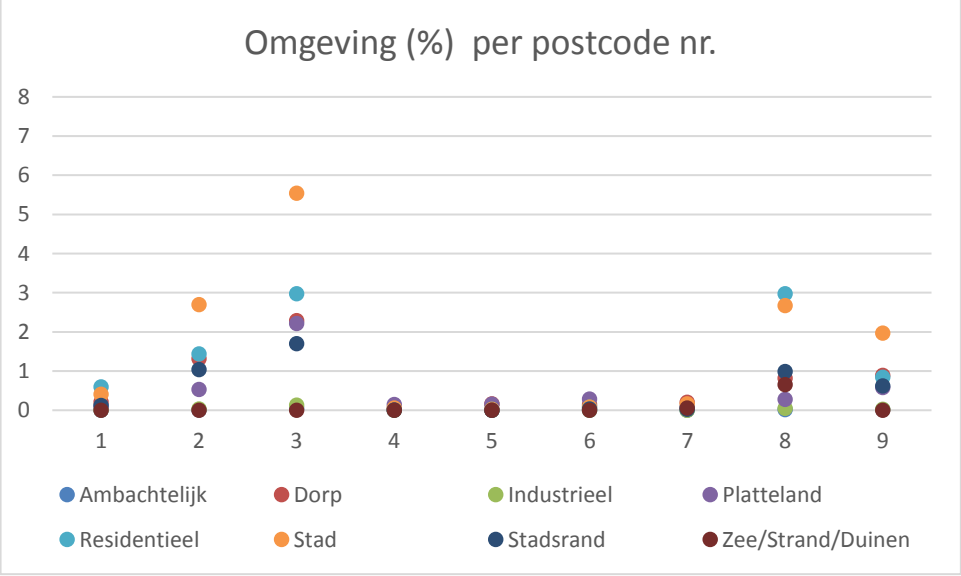
Variabelen die minder dan 10% zijn ingevuld en/of slechte spreiding kennen	Percentage ingevuld	Opmerking
Kiindex	2,2	
kiindex_jaar	0,1	Zeer laag percentage
Maandelijks	0,2	Zeer laag percentage
onroerendevoorheffing	0,1	Zeer laag percentage
Onroervh	0,8	Zeer laag percentage
onroervh_jaar	0,2	Zeer laag percentage
prijsgrondaandeel	1,1	Nauwelijks ingevuld
Prijsvm	0,2	Zeer laag percentage
Provisie	0,2	Zeer laag percentage
notaris_verkooptype_id	0,4	Zeer laag percentage
o_garages	6,1	Risico voor interpretatie Vaak onrealistisch hoge cijfers/aantallen eventueel surrogaat 'faciliteiten'
o_openbaarvervoer	3,6	Knooppuntwaarde eventueel surrogaat 'faciliteiten'
o_orientatie_id	3,2	Mogelijk vervangbaar via geografische analyse
o_parkings	4,0	Risico voor interpretatie Vaak onrealistisch hoge cijfers/aantallen eventueel surrogaat 'faciliteiten'
o_school	3,1	POI - Data; eventueel surrogaat 'faciliteiten'
o_strand	0,2	Landgebruikskaart
o_tuintekst_nl	1,8	Niet zinvol voor analyse – enkel fonetisch, wordt behouden ter info
o_winkels	3,7	POI – Data; eventueel surrogaat 'faciliteiten'
r_prijs	3,4	
Pub2_start	0,3	Zeer laag percentage
Pub2_stop	0,2	Zeer laag percentage

Hieronder wordt de keuze voor een aantal variabelen verder verduidelijkt op basis van een meer gedetailleerde analyse. Telkens wordt de respons weergegeven voor de vernoemde attributen per variabele en opgedeeld naar de postcodes (1000-tallen). Meerdere grafieken worden weergegeven t.b.v. de nodige inzichten/details: procentuele weergave (%, = procentueel aantal data van dat bepaald type), 1x met inclusie van blanco gegevens en 1x zonder de blanco gegevens. Deze nemen immers vaak een groot aandeel in waardoor de andere percentages neergedrukt worden wat de visualisatie erg kan bemoeilijken. Daarnaast wordt ook een grafiek opgenomen die logaritme percentage (log %) weergeeft, dit om voor lage percentages toch nog grafisch detail te bekomen.

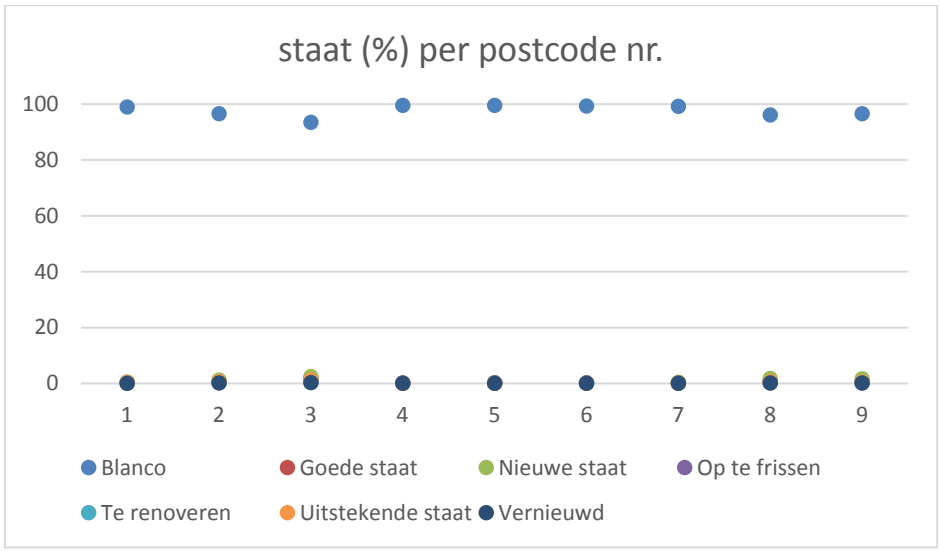
a_omgeving_id: *relatief weinig respons. Slechte spreiding (geen respons Postcodes 1, 4-7 (vnl. Wallonië). Risico voor verkeerde interpretatie van begrippen zoals stadsrand, ambachtelijk, industrieel, … Deze variabele is een subjectieve invoer door de makelaar of particuliere eigenaar. Bij invoer wordt de invulling bepaald naargelang wat op de website wordt getoond voor een pand. Geen duidelijke typering (residentieel vs. stadsrand – dorp vs. platteland - …). Objectieve zijn mogelijk (bv. combinaties met vectoriële topografische kaart, ruimteboekhouding, …)*

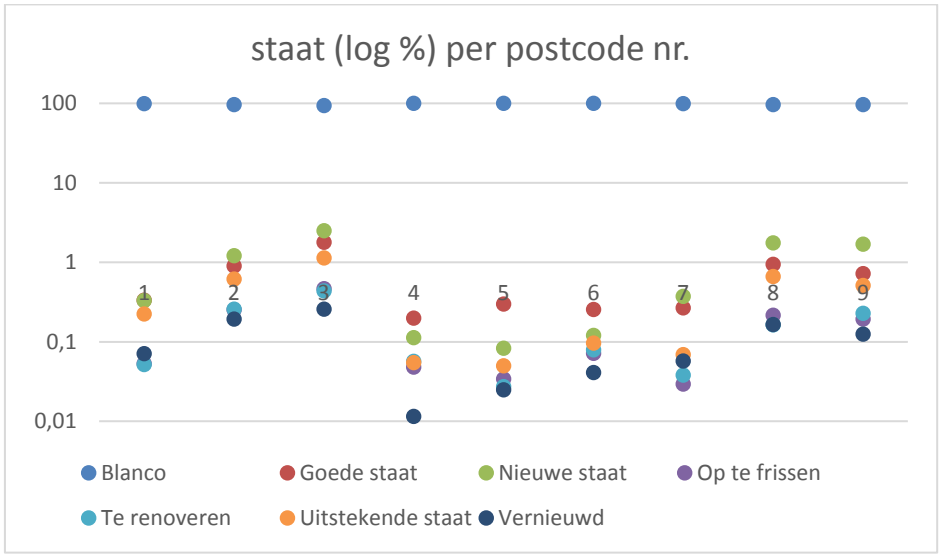

Omgeving (%) per postcode nr.
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
Blanco
Ambachtelijk
Dorp
Industrieel
Platteland
Residentieel
Stad
Stadsrand
Zee/Strand/Duinen

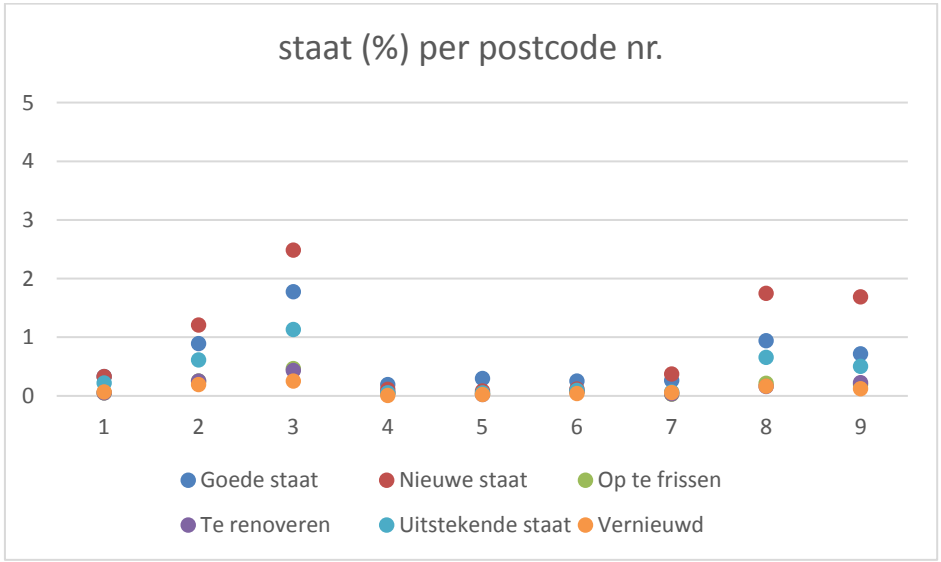

Omgeving (log %) per postcode nr.
100
10
1
0,1
0,01
0,001
0,0001
1
2
3
4
5
6
7
8
9
Blanco
Ambachtelijk
Dorp
Industrieel
Platteland
Residentieel
Stad
Stadsrand
Zee/Strand/Duinen


Omgeving (%) per postcode nr.
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
Ambachtelijk
Dorp
Industrieel
Platteland
Residentieel
Stad
Stadsrand
Zee/Strand/Duinen

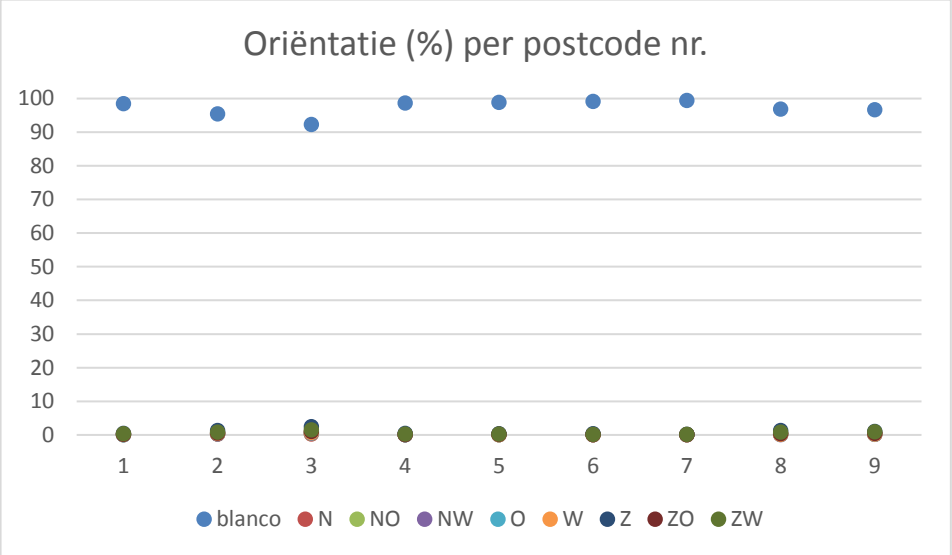
a_staat: *zeer beperkt ingevuld en niet egaal gespreid – betere alternatieven zijn de variabelen 'te renoveren' en 'nieuwbouw' (+10% ingevuld)*

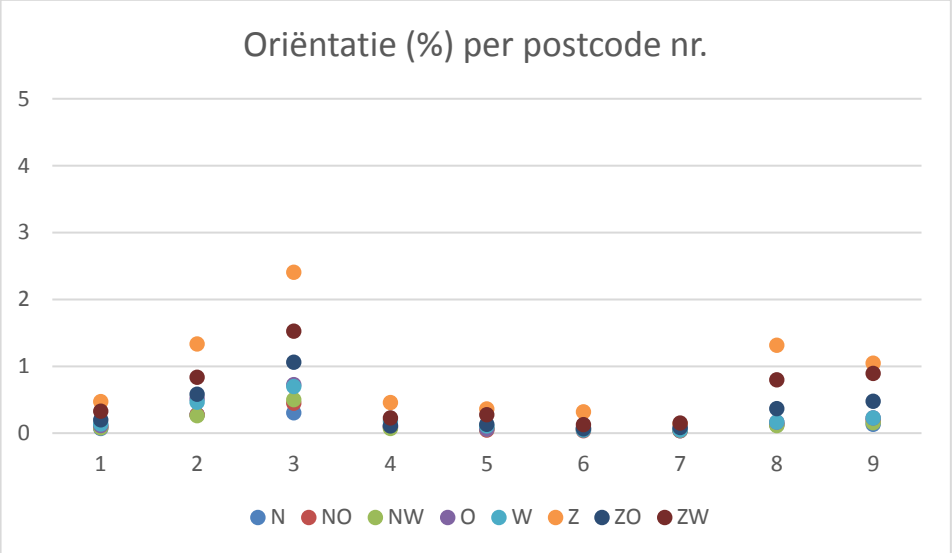


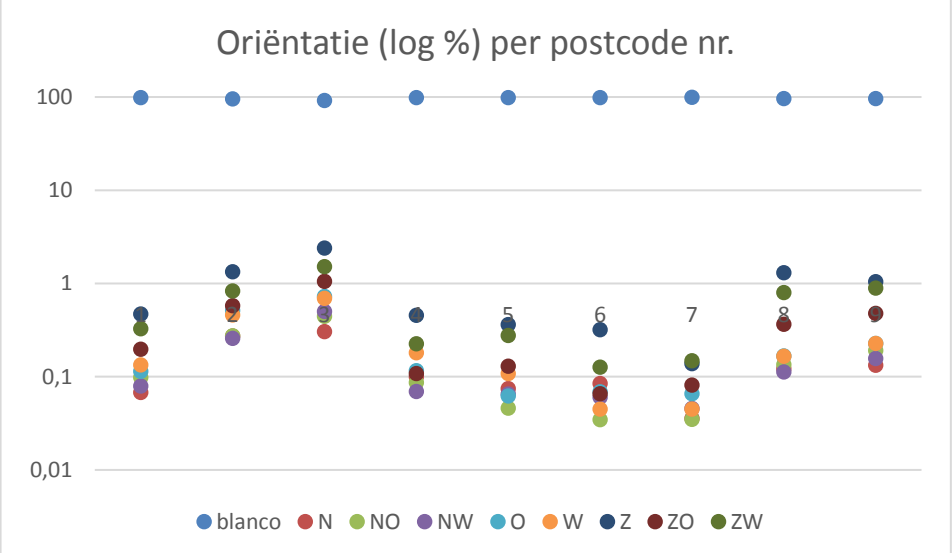




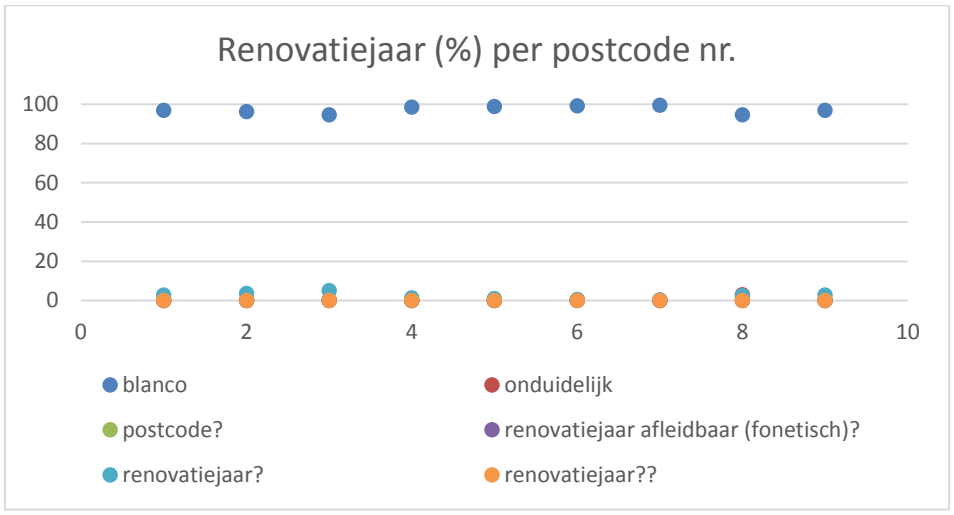
b_oriëntatie: *Zeer veel blanco en slechte spreiding. Aanvulling via ruimtelijke analyse aangewezen.*

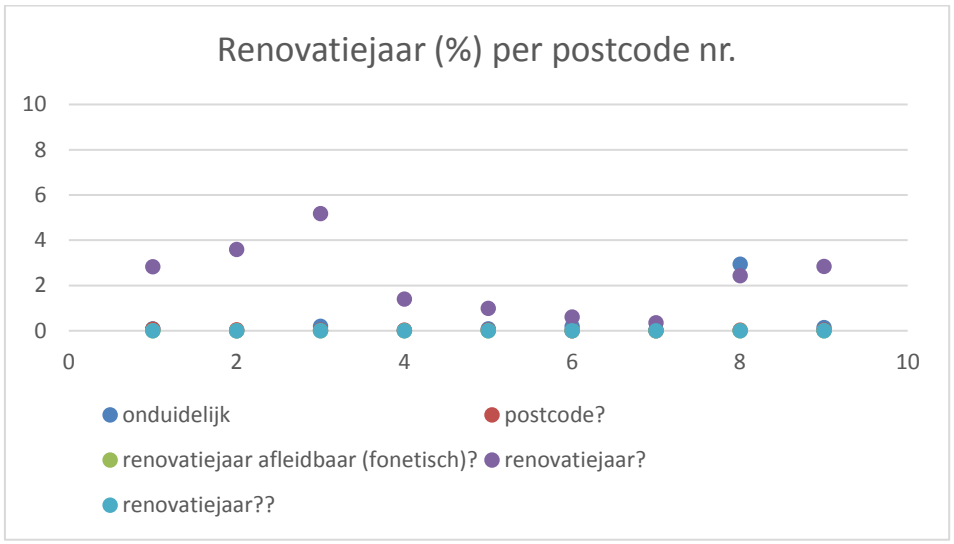


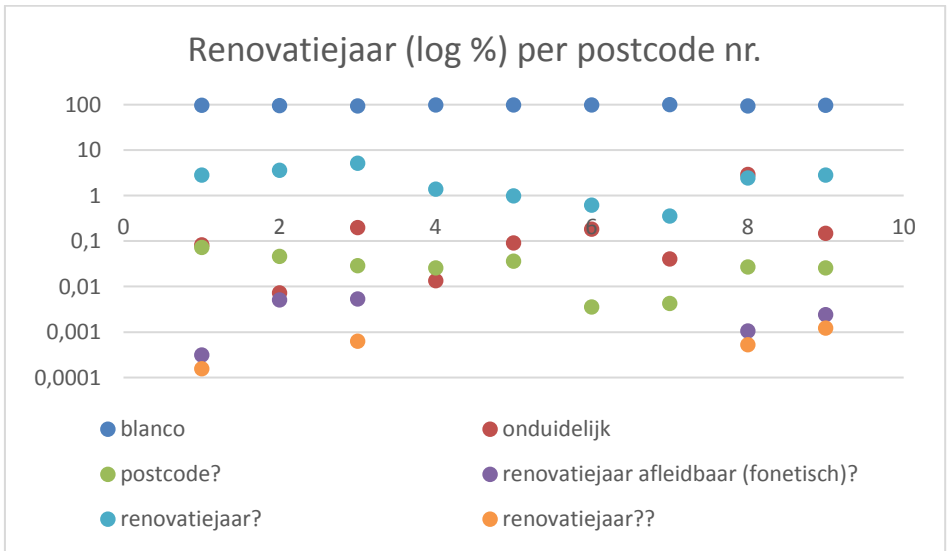




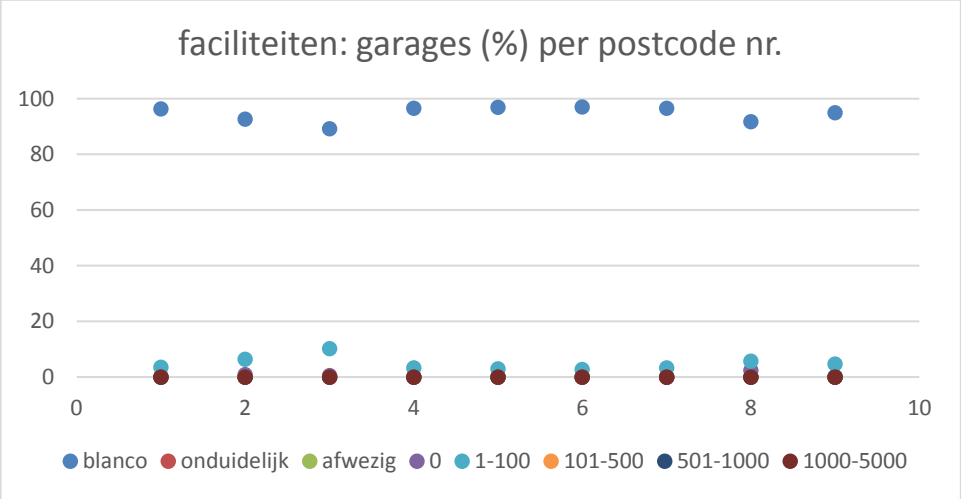
b_renovatiejaar: *Veel blanco. Geen egale spreiding. Vaak onduidelijk en foutief ingevuld. Vermoedelijk risico voor foute interpretatie door persoon die dit invult (aantal jaren geleden vs. jaartal van renovatie?)*. Diverse onduidelijkheden: soms postcode mogelijks ingevuld (postcode?), of renovatiejaartal mogelijks niet correct = twijfelachtig (?) of zeer twijfelachtig (??)…

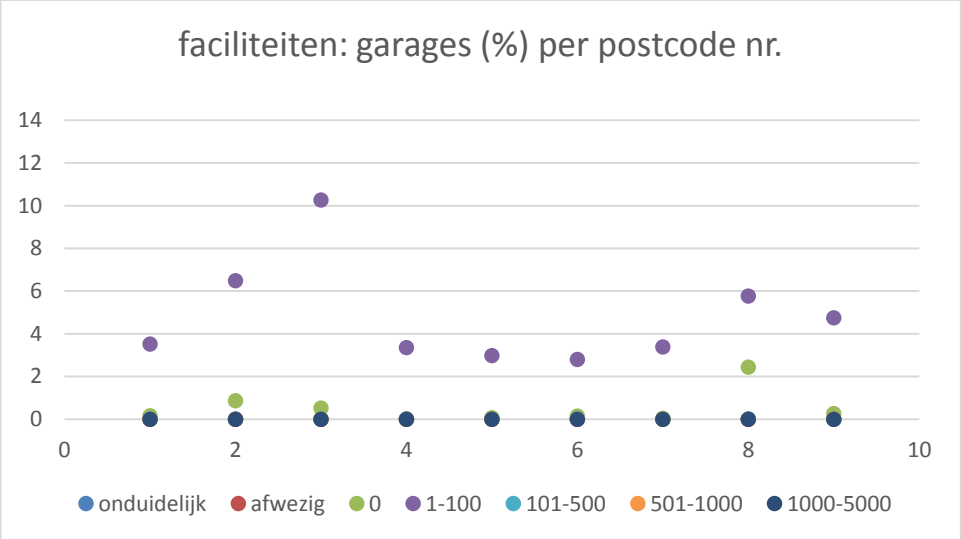


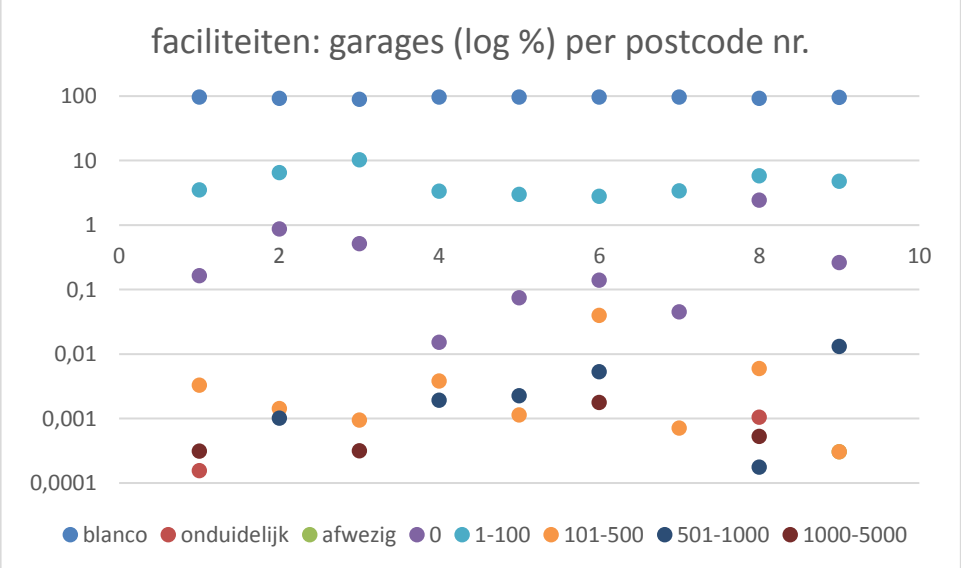


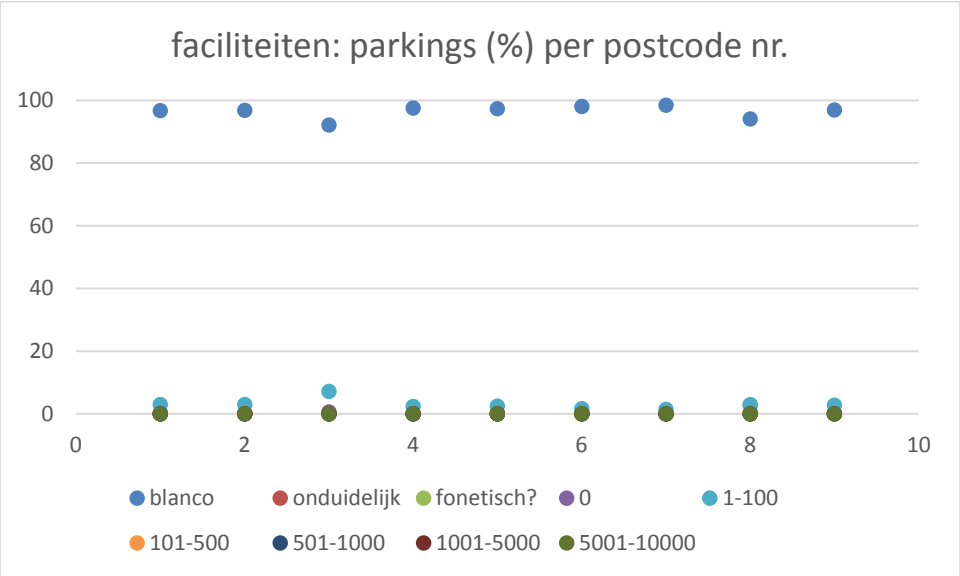


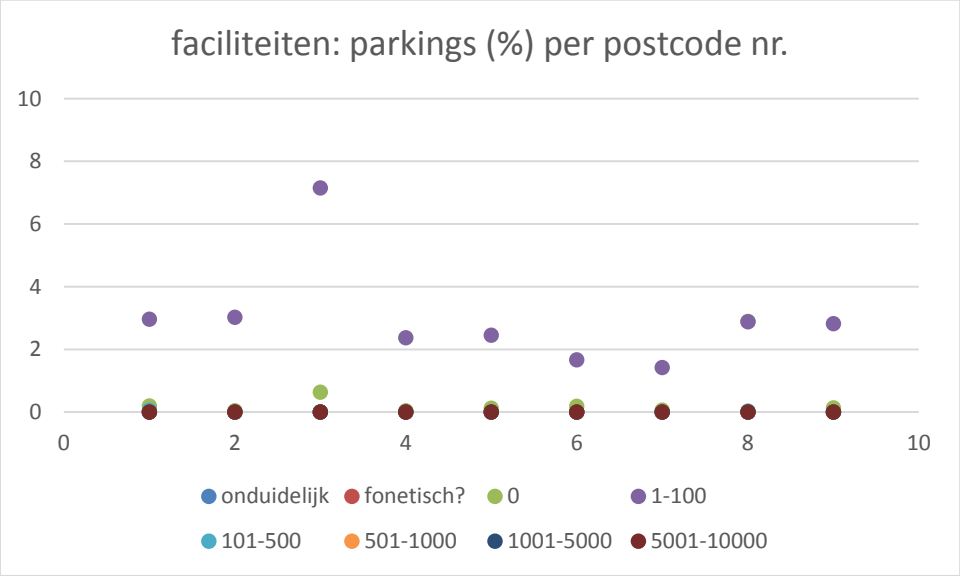
o_garages en o_parkings: onrealistisch hoge aantallen en beperkt ingevuld (de meerderheid is blanco). Het kan hier echter gaan over zowel garages en parkings bij een villa, woning of bij een appartementsgebouw. In dat geval zijn aantallen van 20, 30, 40 realistisch.

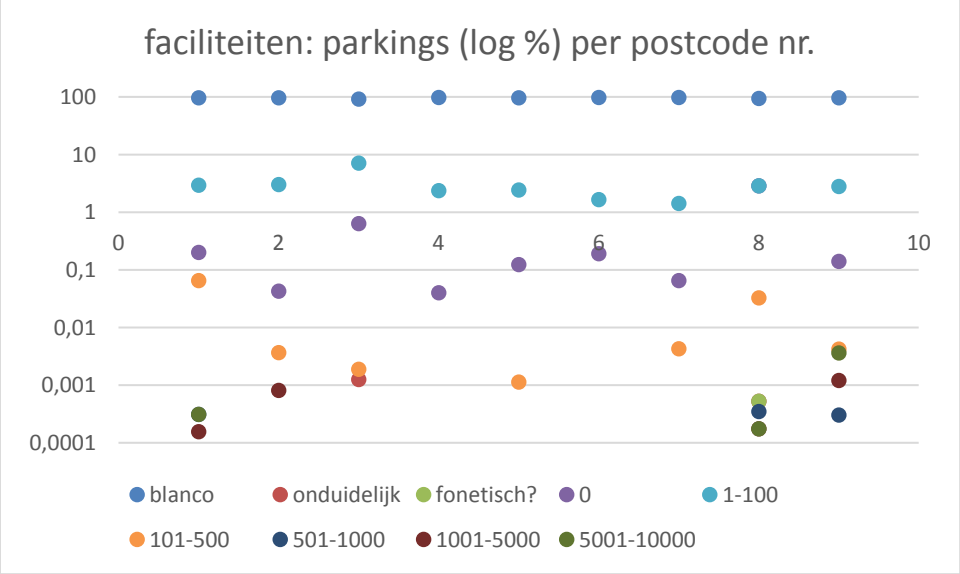







faciliteiten: parkings (%) per postcode nr.
100
80
60
40
20
0
0
2
4
6
8
10
blanco
onduidelijk
fonetisch?
0
1-100
101-500
501-1000
1001-5000
5001-10000


faciliteiten: parkings (%) per postcode nr.
10
8
6
4
2
0
0
2
4
6
8
10
onduidelijk
fonetisch?
0
1-100
101-500
501-1000
1001-5000
5001-10000


faciliteiten: parkings (log %) per postcode nr.
100
10
1
0,1
0,01
0,001
0,0001
0
2
4
6
8
10
blanco
onduidelijk
fonetisch?
0
1-100
101-500
501-1000
1001-5000
5001-10000

1.2.4 Meer dan 10% ingevuld

Voor de 29 variabelen die voor meer dan 10% zijn ingevuld wordt verder onderzocht of ze geschikt zijn om op te nemen in de basisdataset.

Voor de variabelen >10% ingevuld, zijn hiervoor de volgende opties opgenomen:

- De variabele wordt geschrapt en vervangen indien betere spatiale datasets beschikbaar zijn – 0 variabelen
- De variabele wordt geschrapt indien de variabele inhoudelijk niet interessant of bruikbaar is (bv. risico voor interpretatie, niet zinvol - rode variabelen in de tabel) – 3 variabelen
- De variabele wordt opgenomen in de subset, indien ze inhoudelijk wel interessant is, maar risico voor gebiedsdekkende data-analyse kent door een slechte geografische spreiding (blauwe variabelen in de tabel) – 2 variabelen
- De variabele wordt opgenomen in de basisdataset indien de variabele inhoudelijk op het eerste zicht interessant en bruikbaar lijkt en een goede spreiding vertoont (oranje variabelen in de tabel) – 24 variabelen

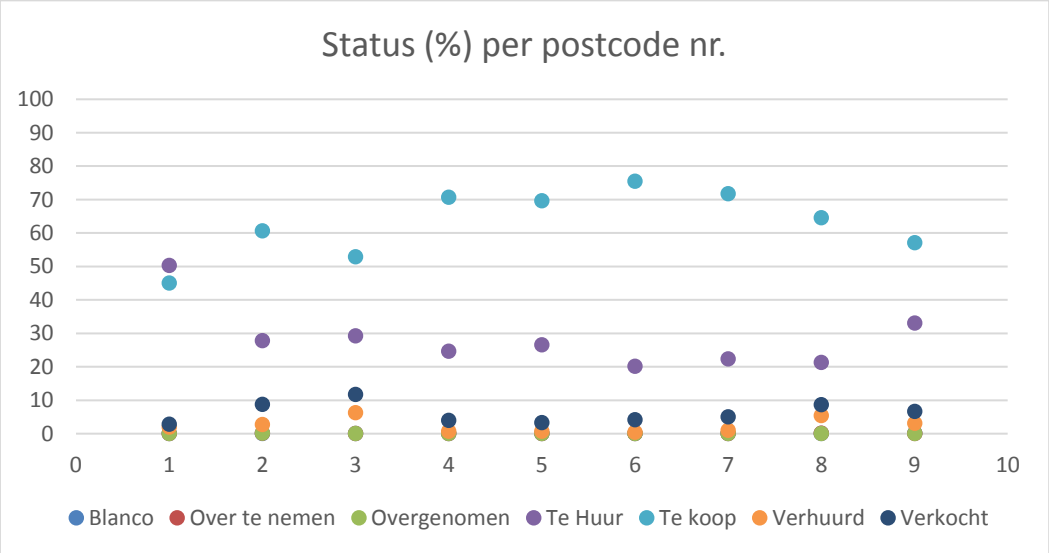
Tabel 12 – Variabelen voor meer dan 10% ingevuld

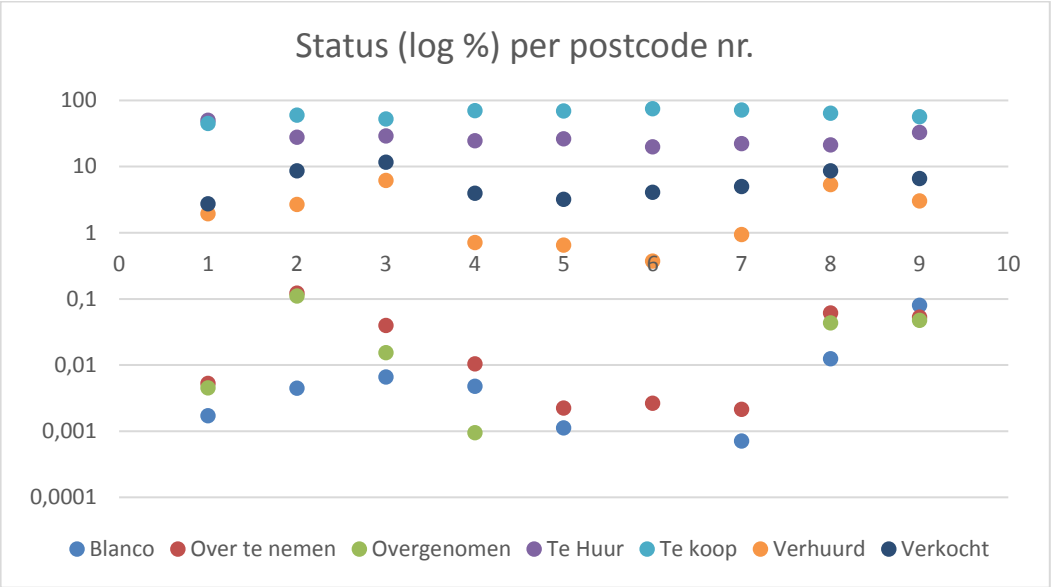
Variabelen die meer dan 10% zijn ingevuld	Percentage ingevuld	Opmerking
a_beschr_nl	74,9	Facultatief
a_beschrkort_nl	10,0	Facultatief
a_status_id	100,0	
a_subtype_id	89,2	Geen hoog analytisch vermogen. Reclassificatie zinvol
a_titel_nl	29,7	Tekstveld, behouden ter info
a_type_id	97,2	Dit is een belangrijke variabele, (woningen vs. andere types (garages, grond, bedrijfsvastgoed, appartement)
b_bebouwing_id	18,5	Mogelijk aan te vullen via andere dataset (ifv onderscheid open, halfopen, gesloten)
b_bouwjaar	26,4	
b_epc_attest	91,7	
b_epcwaarde	20,1	
b_grondopp	28,9	Alternatieve dataset? (Cadmap – GRB)
b_nieuwbouw	98,1	
b_terenoveren	98,0	
b_woonopp	49,9	
creatie_datum	99,6	
Ki	15,0	
locatie_id	100,0	Niet zinvol
f_prijs	97,7	Zie ook r_prijs
f_prijszichtbaar	99,2	
is_gearchiveerd	100,0	

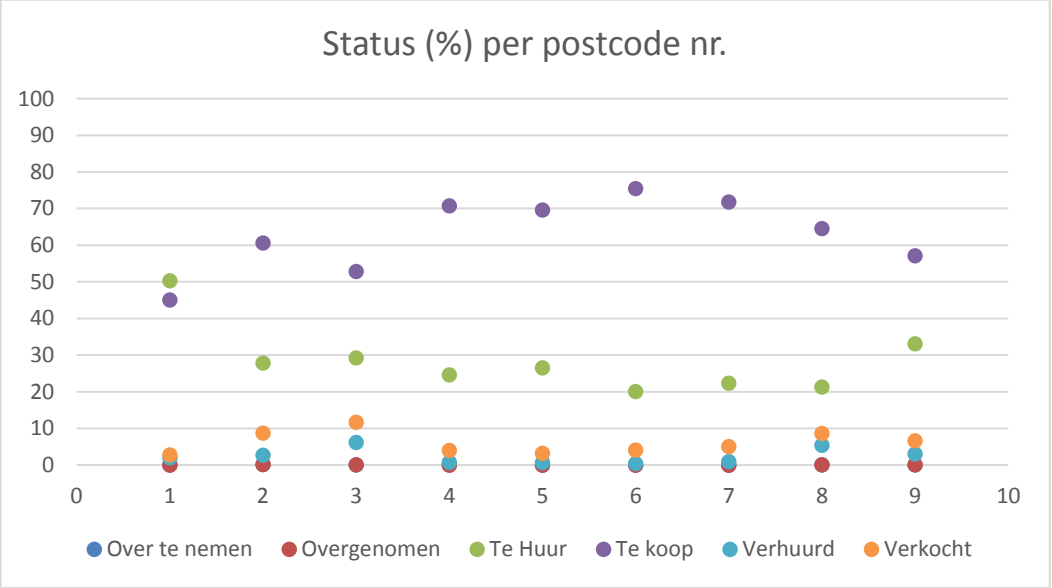
Variabelen die meer dan 10% zijn ingevuld	Percentage ingevuld	Opmerking
n_project_id	100,0	Niet zinvol, intern id Zimmo
o_tuinaanwezig	97,3	Interpretatie-risico's? Grote meerderheidvan de data heeft geen tuin of dit is onduidelijke
o_zichtopzee	97,4	Enkel relevant voor West-vlaanderen
o_zijdelingszichtopzee	98,9	Enkel relevant voor West-vlaanderen
r_datum	39,8	
Showweb	100,0	Niet zinvol
wijziging_datum	99,9	
Pub1_start	82,7	
Pub1_stop	74,8	

Hieronder wordt de keuze voor een aantal variabelen verder verduidelijkt op basis van een meer gedetailleerde analyse. Telkens wordt de respons weergegeven voor de vernoemde attributen per variabelen en opgedeeld naar de postcodes (1000-tallen).

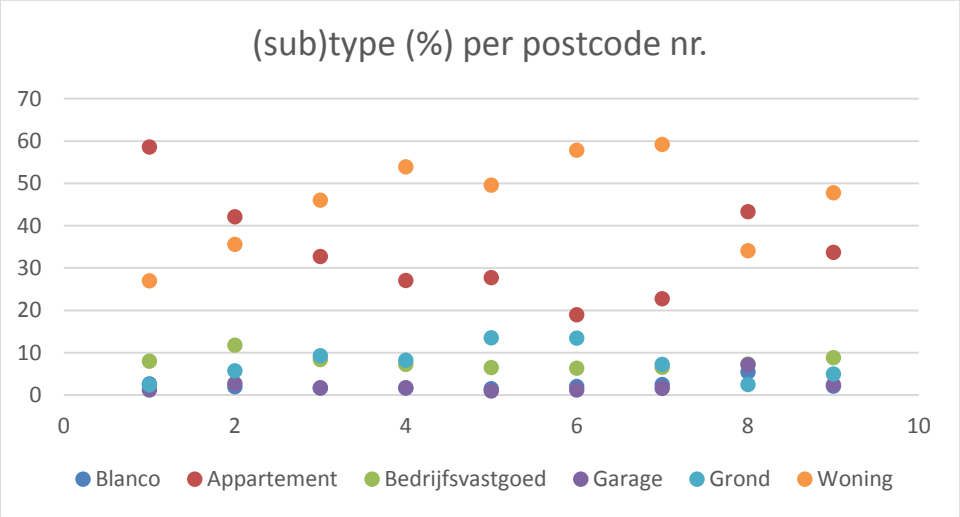
a_status_id: *goede spreiding van de variabele over de postcodes. Te koop, verkocht, verhuurd, te huur duidelijke. 'overgenomen' en 'over te nemen' gelden voor handelszaken waarbij de uitbaatvergunning mee over gaat (deze zijn niet goed verspreid en hun aandeel in de database is beperkt) .*

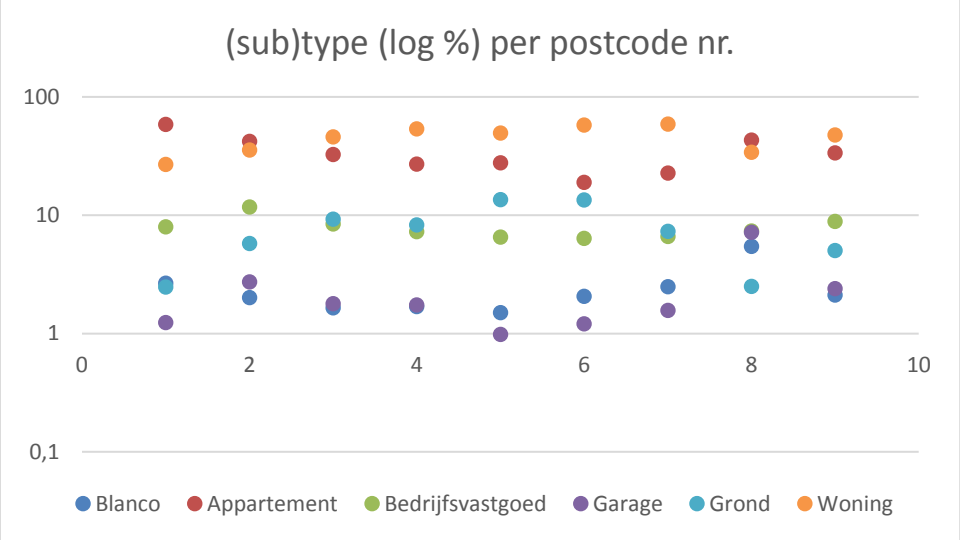


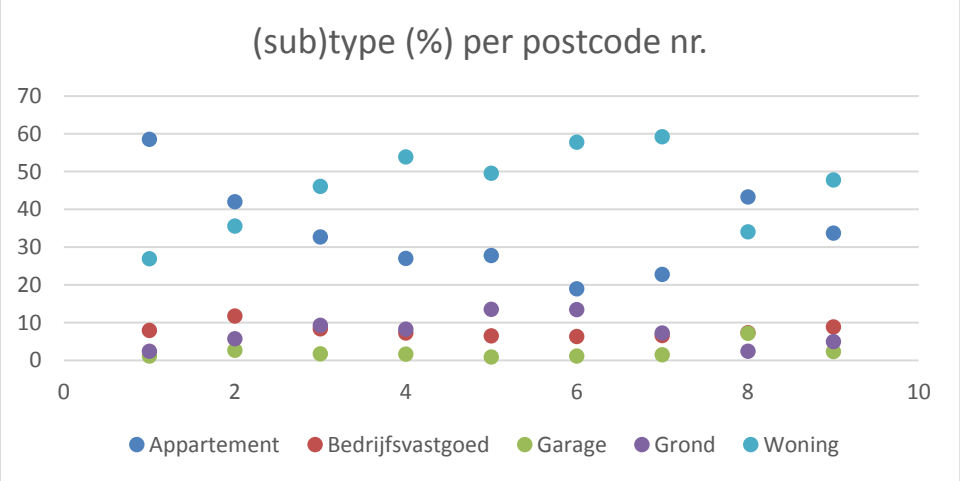

Status (log %) per postcode nr.
100
10
1
0,1
0,01
0,001
0,0001
0
1
2
3
4
5
6
7
8
9
10
Blanco
Over te nemen
Overgenomen
Te Huur
Te koop
Verhuurd
Verkocht


Status (%) per postcode nr.
100
90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
7
8
9
10
Over te nemen
Overgenomen
Te Huur
Te koop
Verhuurd
Verkocht

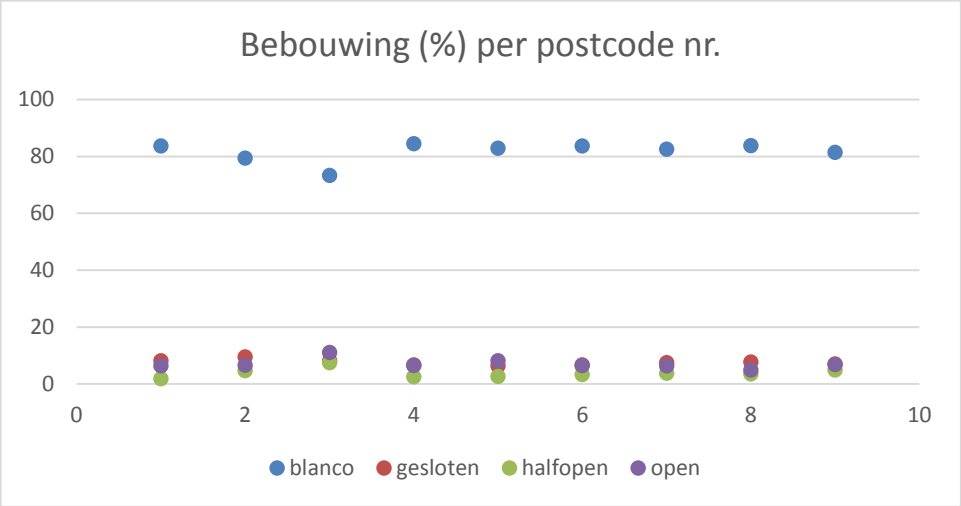
a_type_id: *Deze variabele omvat: woning-bedrijfsvastgoed-appartement-garage-grond-blanco. (Op subtypeniveau zijn er 112 mogelijke subtypes. Dit beperkt het analytische vermogen). Relatief hoge repons, weinig blanco en goede spreiding over de postcodes.*

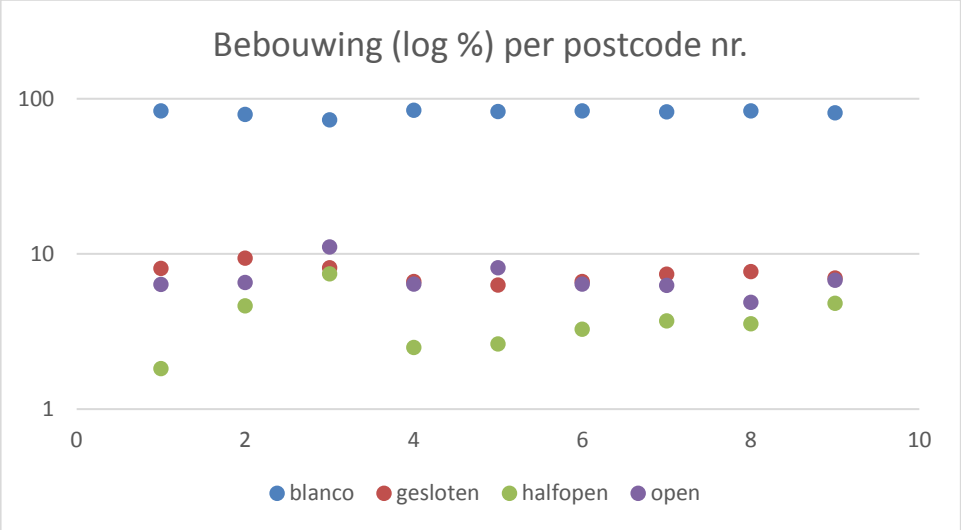


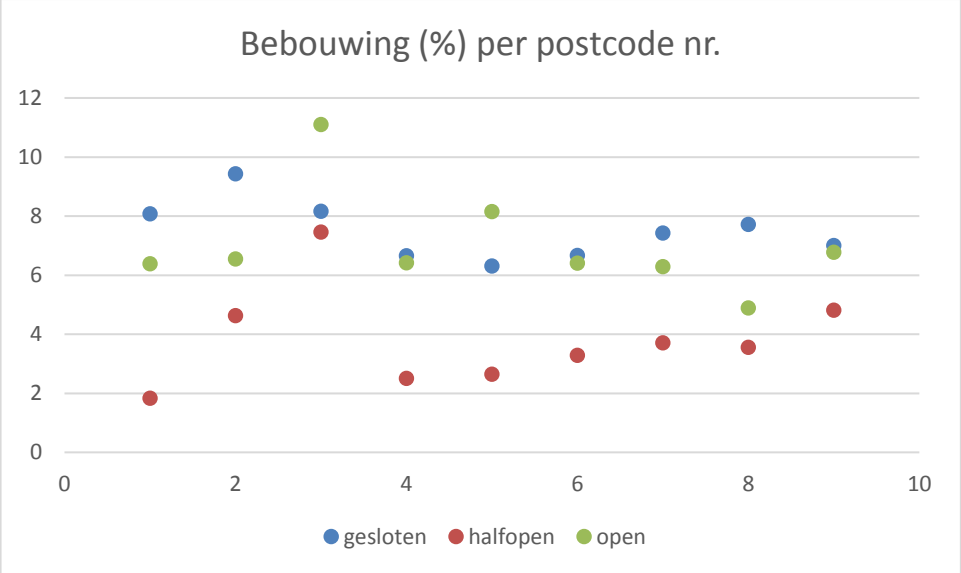




b_bebouwing_id: *Vrij veel blanco waarden; Voor de overige waarden is er wel een relatief goede spreiding.*

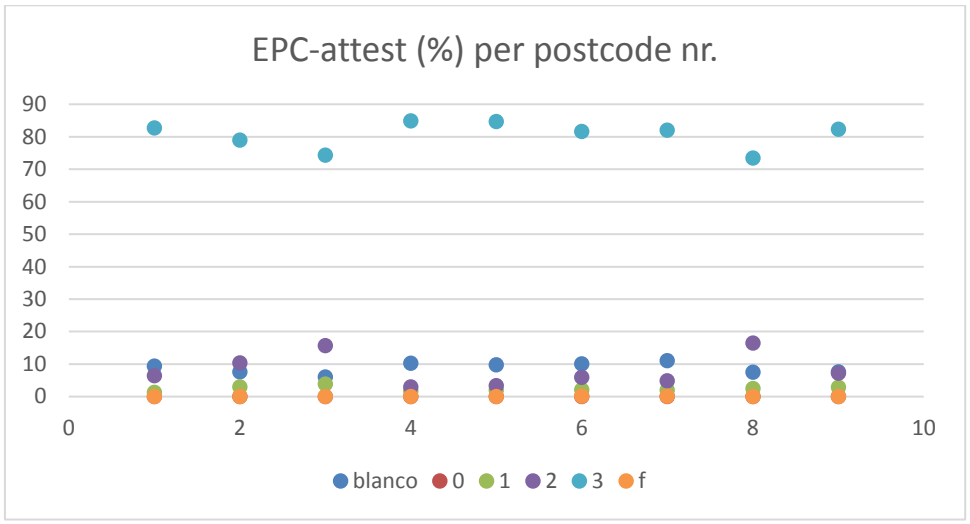


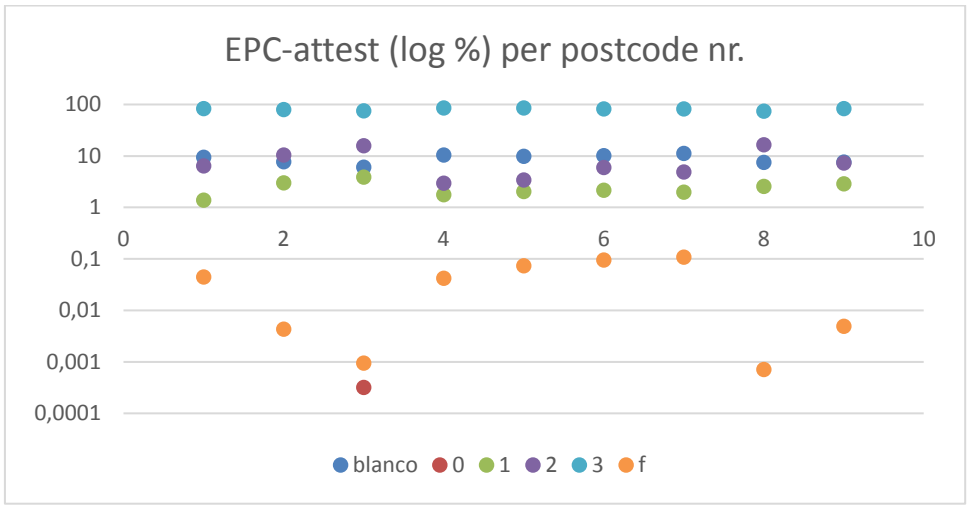


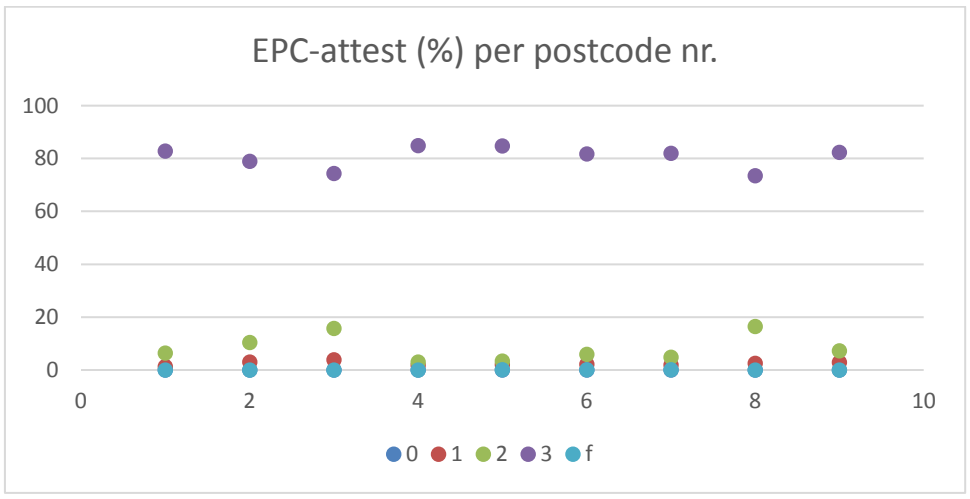


b_EPC-attest: (1 = EPC-attest verkregen, 2 = EPC-attest niet verkregen, 3= *EPC-attest niet ingevuld). Grotendeels ingevuld als 3. De code 0 betekent eveneens niet ingevuld (dit wordt gebruikt bij het wegschrijven van crawling data naar de database. Waarde f betekent 'false' (= foute invoer vanuit de crawling)*

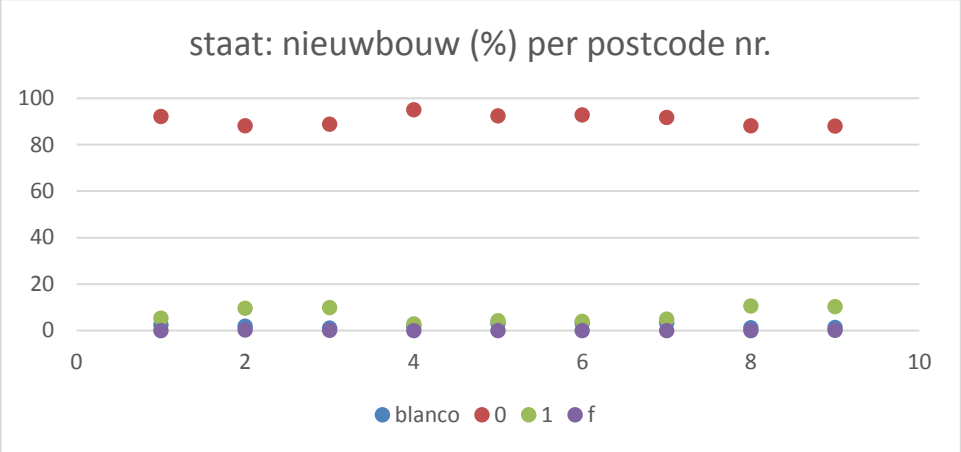
Inhoudelijk interessant, dus behoud in de basisdataset.

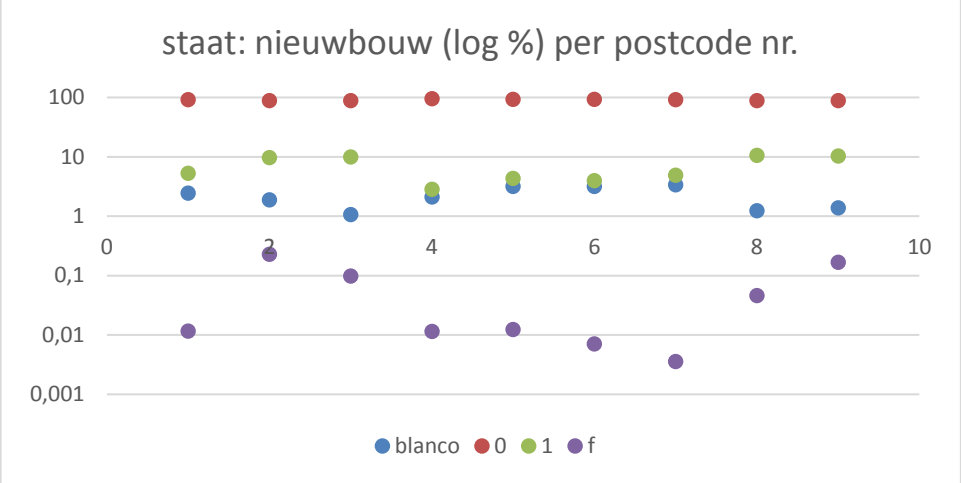


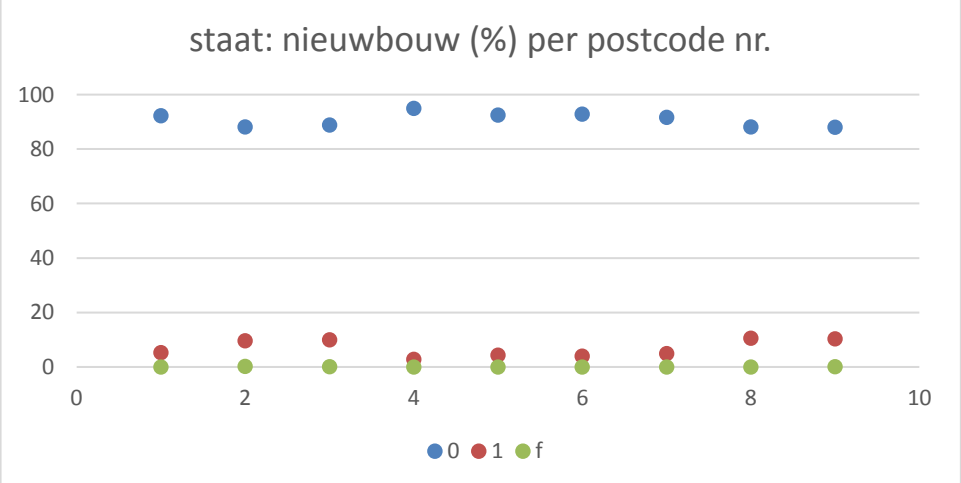


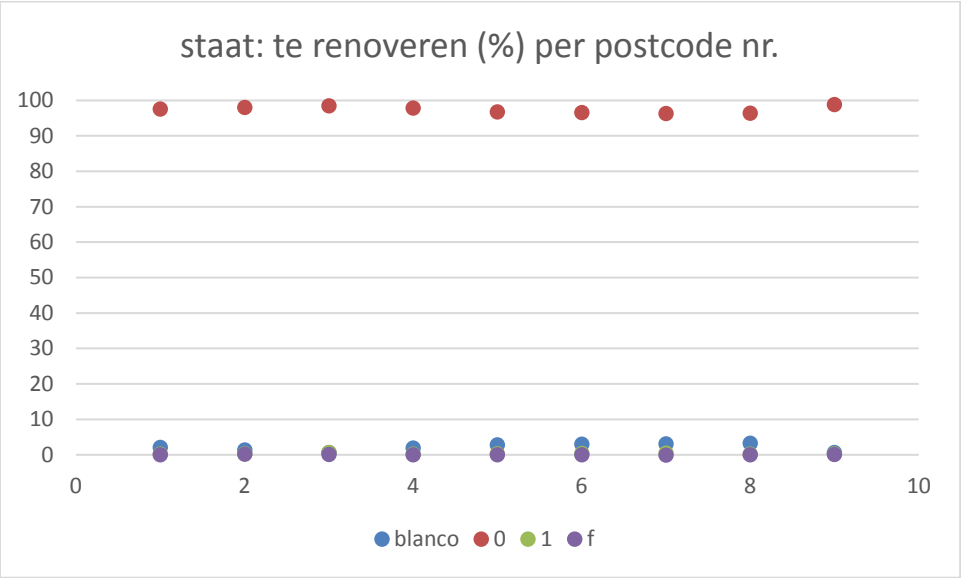


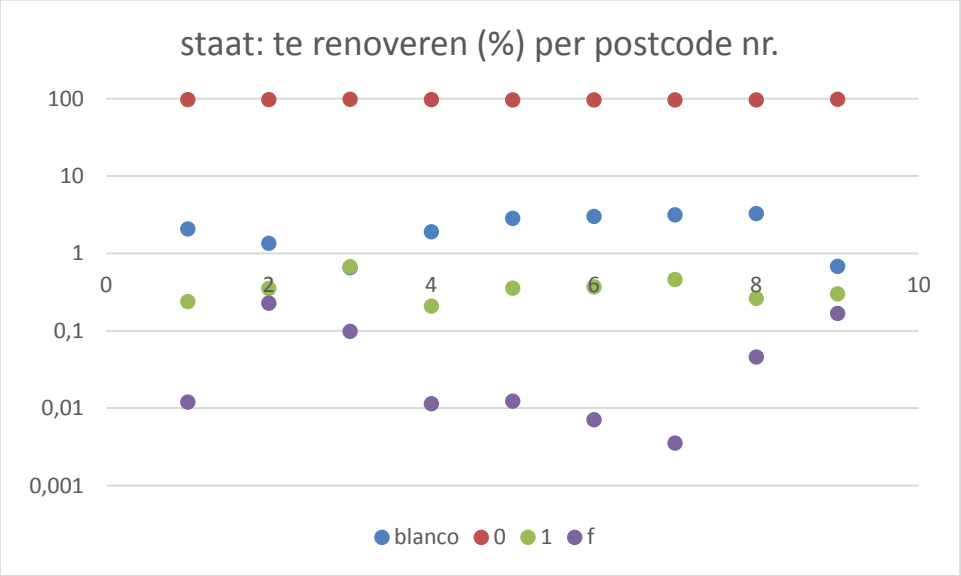
b_nieuwbouw en b_terenoveren: *Vergelijkbare variabelen. Nieuwbouw en te renoveren zijn velden die bestonden voor 'staat' werd toegevoegd. Grotendeels ingevuld als waarde 0 voor beide variabelen (geen nieuwbouw, niet te renoveren). Mogelijk kan hier meer informatie verkregen worden door integratie van de variabelen 'b_nieuwbouw, b_terenoveren (b_staat_id). Voorlopig behoud in de dataset voor verdere analyse..*

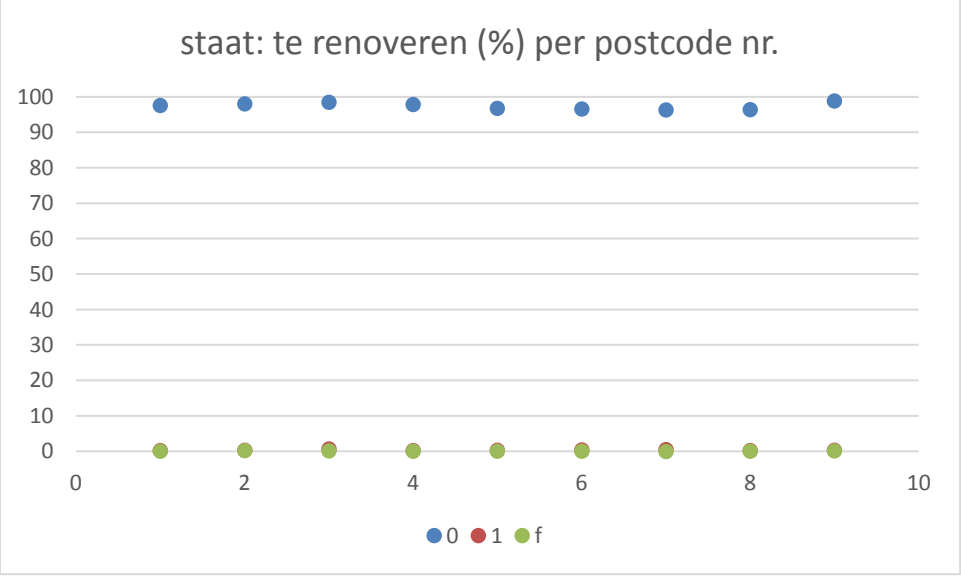




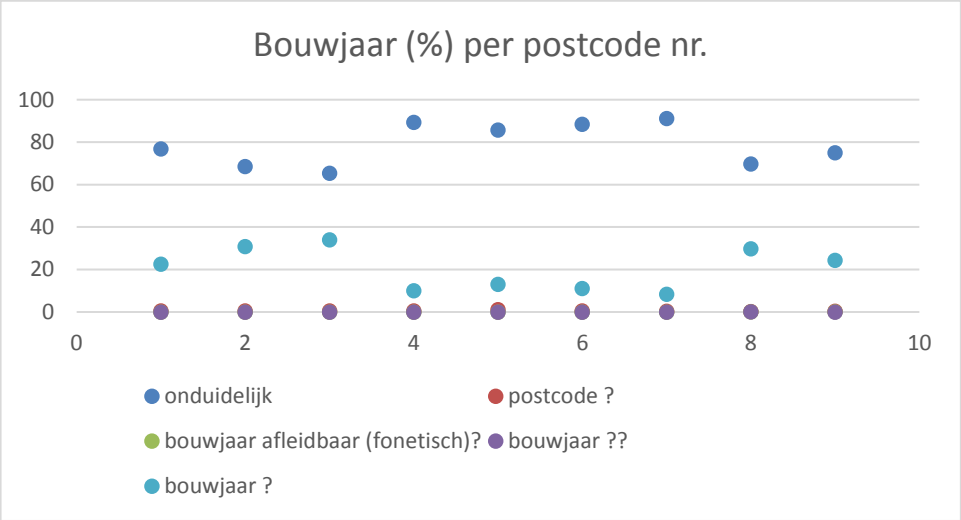


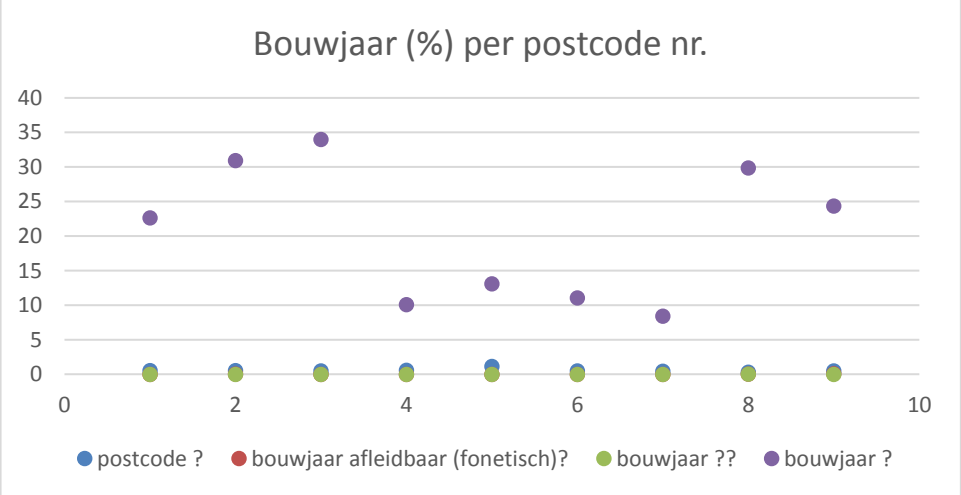

staat: te renoveren (%) per postcode nr.
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
blanco
0
1
f

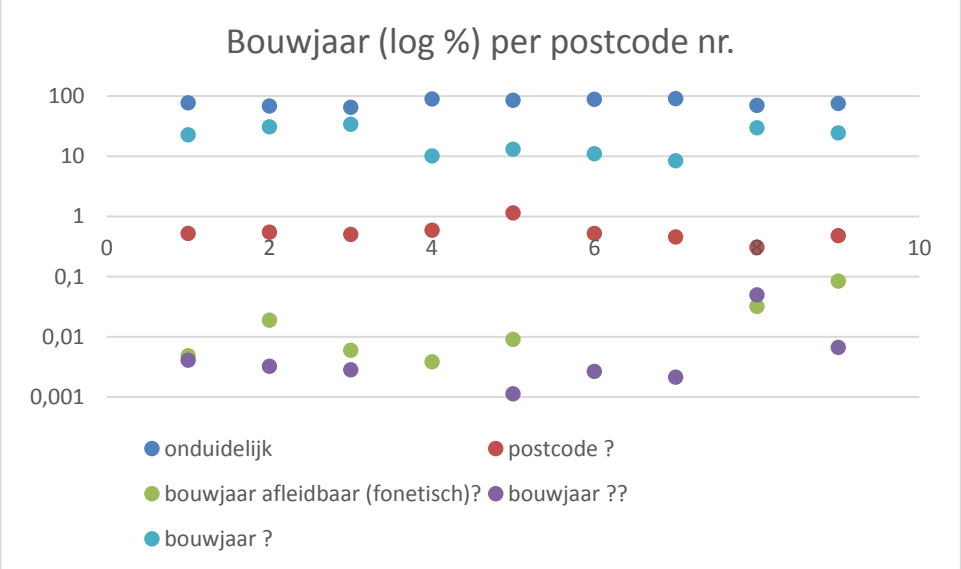

staat: te renoveren (%) per postcode nr.
100
10
1
0,1
0,01
0,001
0
2
4
6
8
10
blanco
0
1
f


staat: te renoveren (%) per postcode nr.
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
0
1
f

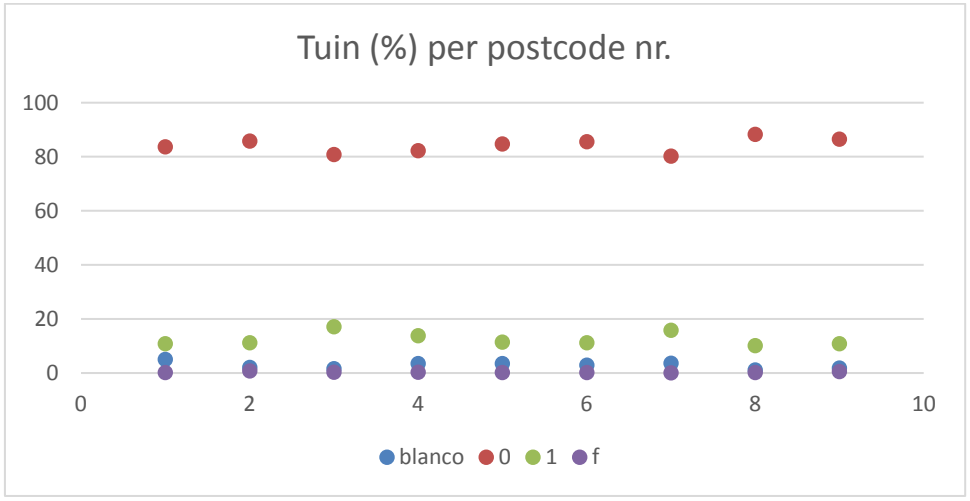
b_bouwjaar: *cfr. renovatiejaar, vaak foutief ingevuld. De invoer van dit veld is beperkt tot numerieke waarde, maar er zit geen controle op de invoer van het aantal cijfers. Fouten kunnen dus afkomstig zijn van een foute invoer of door het wegschrijven van deze info vanuit de crawlers waarop minder controle is. Behoud omwille van inhoudelijk interessant.*

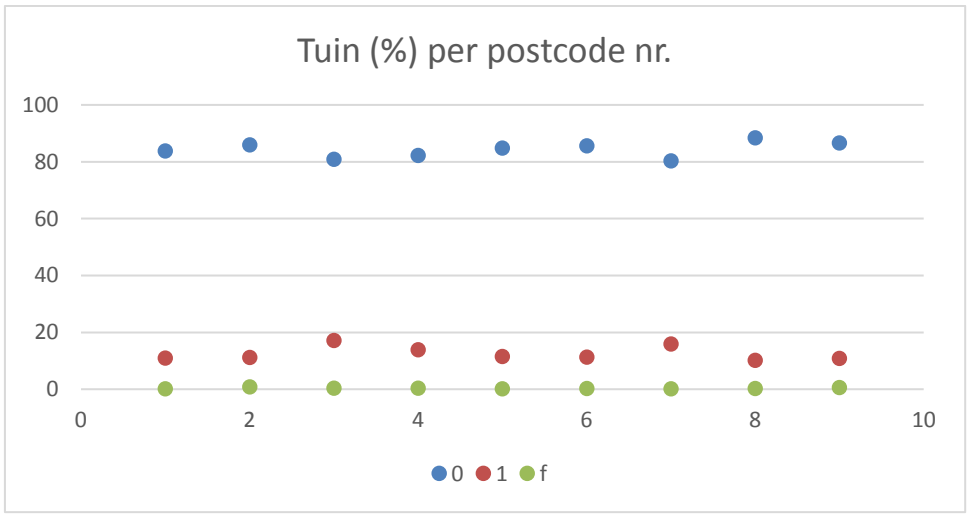


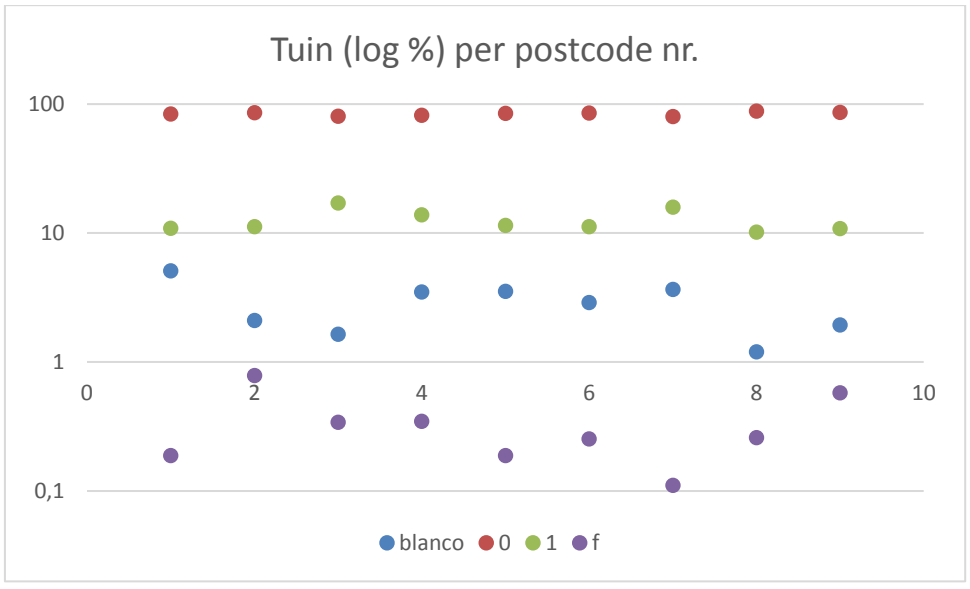




o_tuinaanwezig. *Grotendeels ingevuld als 0. Deze waarde (0) betekent 'geen tuin' of 'geen informatie bekend dat een tuin beschikbaar is'. Vermoedelijk is dit hierdoor een beperkt bruikbare variabele, en is een alternatieve dataset aangewezen.*







1.2.5 Synthese

Bovenstaande item non-respons-analyse heeft er toe geleid dat op basis van deze eerste screening volgende variabelen behouden zijn voor de basisdataset:

Tabel 13 – behouden variabelen

a_bus
a_gemeente
a_geo_lat
a_geo_lon
a_nummer
a_postcode
a_straat
a_geo_precision
a_land_id
a_beschr_nl
a_beschrkort_nl
a_status_id
a_subtype_id
a_titel_nl
a_type_id
b_bebouwing_id
b_bouwjaar
b_epc_attest
b_epcwaarde
b_grondopp
b_nieuwbouw
b_terenoveren
b_woonopp
creatie_datum
f_ki
f_prijs
f_prijszichtbaar
is_gearchiveerd
o_tuinaanwezig
r_datum
wijziging_datum
Pub1_start
Pub1_stop

Opmerking: a_beschr_nl en a_beschrkort_nl zijn niet bruikbaar voor analyse (tekst), maar worden wel behouden ter info.

Volgende variabelen worden opgenomen als subset:

Tabel 14 – behouden variabelen (subset)

btwstelsel_grond
Kiindex
r_prijs
b_perceelbreedte
b_perceeldiepte
b_bouwopp
o_zichtopzee
o_zijdelingszichtopzee
o_tuintekst_nl

Opmerking: o_tuintekst_nl is niet bruikbaar voor analyse (tekst), maar wordt wel behouden ter info.

Voor onderstaande variabelen zullen surrogaatvariabelen mogelijk zijn via combinaties met andere ruimtelijke datasets. Zeker voor wat betreft de variabelen die te koppelen zijn aan GRB. Variabelen zoals afstand tot bepaalde voorzieningen maken in principe deel uit van specifiek onderzoek hierrond (niet voorzien in deze algemene analyse).

Tabel 15 – variabelen die bij voorkeur worden vervangen door surrogaatvariabelen

a_omgeving_id
b_kadastraleaard
b_kadastraleafdeling
b_kadastralesectie
o_openbaarvervoer
o_orientatie_id
o_school
o_strand
o_winkels
b_perceelnummer

Opmerking: de "kadastervariabelen" worden alsnog in de databank opgenomen ter controle.

Als surrogaatvariabele zijn de kadastervariabelen opgenomen (zie ook tabel 7) alsook de variabele "knooppuntwaarde" (als surrogaat voor "openbaar vervoer"). Overige variabelen werden in kader van deze studie niet vervangen.

De overige variabelen werden geschrapt, wegens te weinig respons, te hoog risico voor foutieve interpretatie, of inhoudelijk niet zinvol.

Tabel 16 – niet-behouden variabelen

Netto-opbrengst
Overnameprijs
n_ligging
notaris_bouquet
notaris_instelprijs

notaris_max_looptijd_lijfrente
notaris_plaats_openbareverkoop
notaris_prijs_gebracht_op
notaris_recht_hoger_bod_datum
notaris_rente
notaris_toegewezen_ovhb_aan
Pub3_start
Pub3_stop
Pub4_start
Pub4_stop
Pub5_start
Pub5_stop
Pub6_start
Pub6_stop
Pub7_start
Pub7_stop
Pub8_start
Pub8_stop
Pub9_start
Pub9_stop
Pub10_start
Pub10_stop1
a_sleutelnummer
a_vrijop_datum
a_vrijop2_id
Pub2_start
Pub2_stop
o_parkings
Prijsvm
Provisie
notaris_verkooptype_id
o_garages
kiindex_jaar
Maandelijks
Onroerendevoorheffing
Onroervh
onroervh_jaar
forfait_aansluitingskosten
forfait_basisakte_notaris
Huurwaarborg
Jaarhuur
b_renovatiejaar
b_staat_id
locatie_id
Showweb
prijsgrondaandeel
n_project_id

1.2.6 Betrouwbaarheid van de respons

De item non-respons analyse met erbij horende afgeleide criteria voor variabelenselectie geeft een eerste inzicht in de dataset, maar zegt op zich niets over de kwaliteit of de betrouwbaarheid van de respons zelf.

De respons kan bijvoorbeeld bestaan uit een automatisch ingevuld veld (bv. vanuit crawling), fout of subjectief ingevoerd zijn (bv. in functie van de verkoop), …. Omdat het nagaan van zulke factoren te ver drijven in kader van deze globale analyse, kan de dataset niet volledig getoetst worden op betrouwbaarheid of nauwkeurigheid. Wel zijn een aantal specifieke kanttekeningen te maken bij de bruikbaarheid van de databank. Enkele voorbeelden:

- ⇨ De waarde '0' kan verschillende betekenissen hebben: 'zero', 'nee' of 'afwezig' of 'niet van toepassing'. Dit is afhankelijk van veld tot veld. Is in relatie hiermee hercodering vereist?
  Dit is inhoudelijk niet zonder meer te beoordelen maar duidelijk is dat er een zeker foutenpercentage in de data zit, dus problemen uit die hoek zijn te verwachten. Deels kan dit opgevangen worden in een vervolgstap van deze opdracht door de distributie van de variabelen nader te bestuderen.

- ⇨ Wat jaartallen (bouwjaar) betreft is het niet steeds zeker of deze effectief het bouwjaar betreffen. In sommige gevallen lijkt zelfs per abuis een postcode ingevoerd. Dit is inhoudelijk niet zonder meer te beoordelen maar er zit een zeker foutenpercentage in de data, dus problemen uit die hoek zijn te verwachten. Deels kan dit opgevangen worden door in een vervolgstap van deze opdracht de distributie van de variabelen nader te bestuderen.

- ⇨ verwante datarijen komen meermaals voor: kunnen datarijen die op hetzelfde pand betrekking hebben als aanvullend op elkaar gezien worden, of eerder als overlappend of als replica…? Dit is bijvoorbeeld cruciaal indien men zicht wil krijgen op frictieleegstand, gezien 'duur te koop' dan gekend moet zijn. Het is voorlopig nog onduidelijk in welke mate de immodataset hiervoor geschikt is. We geven hieronder alvast enkele aandachtspunten in relatie tot verwante datarijen:
  - ⇨ Dezelfde of schijnbaar dezelfde data komen meermaals voor o.v.v. extra datarijen waarin bepaalde informatie redundant is, andere info net additief of corrigerend… Het is niet duidelijk of er gecorrigeerd mag/moet worden, er imputaties mogen verricht worden enz. Dit kan een grote impact hebben. Een voorbeeld hiervan is kasteel Sterckshof dat min. 5x voorkomt en zonder eliminatie van overbodige (redundante) datarijen o.a. gemiddelde prijzen sterk kan beïnvloeden indien elke datarij in de berekening zou worden meegenomen (dit veroorzaakt sample bias in de eerste lijn).

<u>De nood aan additionele preprocessing</u>

Gegeven bovenstaande vaststellingen, kan additionele preprocessing in samenhang met screening de betrouwbaarheid, kwaliteit (bv. juistheid) en bruikbaarheid van de dataset effectief vergroten. Dit wordt daarom verder uitgewerkt, in eerste instantie voor de casus Oost-Vlaanderen. Diverse variabelen werden reeds proactief aan preprocessing onderworpen (cf. supra). Voor bepaalde variabelen werden onvolkomenheden (bv. jaartallen niet voluit of onrealistische/foute waarden zoals ook tekst i.p.v. getal, bv. bouwjaar) zo reeds geëlimineerd. Dit neemt niet weg dat ook voor die variabelen nog anomalieën kunnen voorkomen. Een analyse van de variabelen, onder meer een outlieranalyse biedt hier dan mogelijks verder inzicht in (zie verder). Maar ook hierna zijn niet al de onzekerheden uitgeklaard. Voor tekstvelden was soms herclassificatie vereist/gewenst maar dit kon niet systematisch gebeuren. Een meer doorgedreven screening kan – soms pas na imputatie – meer inzicht bieden in een aantal variabelen, bv. 'subtype (met tal van onderscheiden categorieën die soms amper voorkomen in de immodata)'. Een aantal variabelen blijft betwijfelbaar i.f.v. bruikbaarheid zolang de impact en rol van verwante datarijen niet volledig is uitgeklaard (bv. 'aantal dagen online (publicatie)').

1.3 Additionele preprocessing en non-respons bias (casus Oost-Vlaanderen)

1.3.1 Item non-respons bias (vertekening)

Voor elk van de bemeten variabelen bestaat er een reëel risico dat non-respons items een (deels) verschillende populatie vertegenwoordigen vergeleken met de data (samples) waarvoor wel respons verkregen werd. Dit kan een zeer ernstige distortie en verkeerde conclusies veroorzaken bij ongenuanceerde, simplistische dataverwerking.

Om dit nader te onderzoeken werd de casus 'Oost-Vlaanderen' (postcodecluster 9000) genomen, met enkel behoud van data waarvoor 'f_prijs' beschikbaar is (i.e. 320.513 records van de 328.277). Deze casus[7] wordt overigens hierna verder doorgetrokken i.f.v. te nemen besluiten voor samenstelling en criteria voor de creatie van de finale analyseset (een subset van de totale databank i.f.v. spatiale analyse, multilevel modellering e.a.). Voor deze casus werd uitgegaan van de variabelenset zoals tot hiertoe beschreven en aangepast.

Om te onderzoeken of 'prijs' (de focus van hedonische/multilevel analyse, o.v.v. afhankelijke variabele in de analyse) significante én relevante systematische verschillen (i.e. vertekening of '*bias*') vertoont naargelang respons of non-respons, werd als eerste stap een eenwegs variantie-analyse of analoog Mann-Whitney U-test (met begeleidende boxplots, zie bijlage) uitgevoerd. Een dergelijke vertekening bleek inderdaad voor een aantal (potentieel belangrijke) variabelen het geval te zijn: bouwjaar, grondoppervlakte (en eraan gelieerd ook perceelbreedte en –veel zwakker- ook perceeldiepte), bouwoppervlakte, woonoppervlakte en KI (slechts zwak ook de KI-index). Hetzelfde (maar zwakke significantie) geldt voor de variabele die het aantal dagen online publicatie aangeeft. Dit alles impliceert dat bouwjaar, de bewuste oppervlaktes en KI ofwel niet in beschouwing genomen worden (geen optie in deze context) dan wel dat er bv. imputatie plaatsvindt of een inperking van de dataset tot een respons-only set voor de betreffende variabelen. Dit laatste verdient voor exploratief onderzoek de voorkeur. Het zal uiteindelijk ook een indicatie geven hoeveel data(rijen) overeind blijven en gedegen sampling (zonder sampling bias) toelaten.

In het algemeen zijn er een paar strategieën om met het probleem van *missing data* om te gaan. De voornaamste zijn imputatie en bijstelling door wegingsfactoren. Bij dit laatste worden incomplete items genegeerd en wordt het sample-gewicht van de responsen net versterkt. Bij imputatie worden ontbrekende waarden vervangen door geïmputeerde waarden (Mohadjer et al. 1994[8]): *a disadvantage of imputation is the possibility of ending up with a data file after imputation that is more biased than if no imputation had been performed. Bias reductions depend on the suitability of the assumptions made in the imputation. When imputations are performed separately on different variables, the bias may be reduced for univariate statistics based on the variables containing the imputed data, but multivariate relationships among variables could become distorted. Also, researchers may treat the resulting data set as if it were complete, thus affecting the variances of the estimates. In some instances, it is argued that when a small proportion of observations are imputed, the effect of imputation is relatively minor. However, in such cases, analyses performed on subgroups may contain a high proportion of imputed values. If the compensation is done well, it will usually reduce bias in survey estimates. It is almost always preferable to impute for missing items rather than treating them as randomly missing data at the analysis stage. Although imputation obviously will not eliminate all nonresponse biases, it can be expected to dampen their effects considerably*.

Om te komen tot een respons-only set[9] voor de variabelen bouwjaar, grondoppervlakte, bouwoppervlakte, woonoppervlakte en KI, werden non-respons items (en ganse units) geëlimineerd;

[7] De casus betreft feitelijk een willekeurige staalname met evenwel de bemerking dat over alle postcode(cluster)s heen heel wat variatie bestaat waaronder ook provinciespecifieke kenmerken (of minstens gewestspecifieke). Toch kan de casus een goed voorbeeld vormen voor al de (overige) immodata en vooral voor de wijze waarop hiermee best omgegaan wordt.

[8] http://www.nber.org/nhanes/nhanes-III/docs/nchs/manuals/nr_bias.pdf

[9] Een set van records die voor al de beschouwde variabelen zijn ingevuld

in totaal bleven dan 4804 datarijen over (ca. 1,5% van de dataset case Oost-Vlaanderen). Een klein percentage dus dat echter via imputatietechniek kan worden opgedreven.

In deze benadering konden de variabelen perceelsbreedte/-diepte best worden geëlimineerd gezien anders veel minder data zouden overblijven (de non-respons ligt immers hoog) en gezien (de hoogsignificante) grondoppervlakte wel compleet ingevuld staat (waaruit dan omgekeerd vaak wel breedtes of dieptes berekend zouden kunnen worden). Maar er is (uiteraard) geen uitgesproken verband (correlatie) tussen de oppervlakte en diepte/lengte apart genomen (cf. Pearson correlaties). Imputatie via locatiegegevens vormt voor deze variabelen de meest belangrijkste aanbeveling. Dergelijke data-acquisitie kan gebeuren via kadastrale nummers en kan de analysemogelijkheden zeker doen toenemen, vooral in gevallen waar ook oppervlaktes zouden ontbreken.

Tabel 17 - Pearson correlaties tussen grondoppervlakte en perceelsbreedte/diepte.

	BREEDTE	DIEPTE	GRONDOPP
BREEDTE	1.000		
DIEPTE	0.207	1.000	
GRONDOPP	-0.001	0.012	1.000

Number of observations: 1391

Een belangrijk aandachtspunt in het algemeen is dat extra data-acquisitie voor variabelen die *bias* dreigen te veroorzaken de analysemogelijkheden in belangrijke mate doet toenemen. Dit is bv. het geval voor aanvullingen (imputaties) inzake oppervlaktevariabelen zoals woonoppervlakte, grondoppervlakte…

<u>Suggesties voor surrogaatvariabelen / clustering</u>

-oppervlakte grond, breedte & diepte perceel

Tabel 18 - Resultaten van de Mann-Whitney U-test met diverse klassevariabelen (binair: respons = 1, non-respons = 0) als input en met 'immo_f_prijs' als afhankelijke variabele. Weergegeven zijn aantal data per klasse/variabele alsook mediane en gemiddelde waarden voor de prijsvariabele en dit voor elk van de relevante (input)variabelen (zoals type_id, bouwjaar,…). Significantieniveaus zoals gebruikelijk: 1/ p<0,001 = ***, 2/ p<0,01 = **, 3/ p <0,05 = *

N.B. Significante effecten impliceren *bias* vanuit de (non-respons op de (input)variabele naar de prijsvariabele toe.

		aantal	gemiddelde	mediaan	significantie
street / Hnr					
	non-respons	135650	159359	2568	
	respons	184863	156446,7	133000	
creatie_datum					
	non-respons	766	145524,2	120000	
	respons	319747	157708,4	103000	
immo_Pub1_start					
	non-respons	21068	222464,7	168000	0.0177 *
	respons	299445	153121,2	95000	
Pub1_dagen					
	non-respons	47914	197727,2	169000	0.02048 *
	respons	272599	150640,2	80000	
type_id					
	non-respons	6666	137567,4	795	
	respons	313847	158106,5	108000	
status_id					
	non-respons	265	8722,12	0	
	respons	320248	157802,54	105000	
bebouwing_id					
	non-respons	260175	155952,8	87500	
	respons	60338	165123,9	142000	
nieuwbouw					
	non-respons	4518	183292,5	169000	
	respons	315995	157313,1	99500	
terenoveren					
	non-respons	2228	207905,5	185000	
	respons	318285	157327,7	100000	
bouwjaar					
	non-respons	239282	139295,4	1250	1.33e-05 ***
	respons	81231	211832,6	199000	
epc_attest					
	non-respons	24953	149328,4	110000	
	respons	295560	158384,3	101000	
epcwaarde					
	non-respons	225275	155288,5	77500	
	respons	95238	163334,5	150000	
tuinaanwezig					
	non-respons	6385	88188,15	590	
	respons	314128	159091,76	108000	
perceelbreedte					
	non-respons	308353	154234,4	85000	0.01665 *
	respons	12160	245035,1	225000	
perceeldiepte					
	non-respons	313515	155776,1	95000	0.0787
	respons	6998	242943,5	219000	
grondopp					
	non-respons	202232	126859,6	750	2.659e-08 ***
	respons	118281	210373,4	190000	
bouwopp					
	non-respons	301024	152822,8	80000	0.008426 **
	respons	19489	232691,8	208000	
woonopp					
	non-respons	176931	140680,6	850	0.009199 **
	respons	143582	178626,2	168750	
ki					
	non-respons	270610	138866,8	975	1.476e-09 ***
	respons	49903	259694,1	224500	
kiindex					
	non-respons	317052	156162,2	99000	0.04505 *
	respons	3461	296650,6	239000	

1.3.2 Afhankelijkheden: verwante gegevens, quasi-replica incl. pseudodata

In de immodataset zitten diverse sterk verwante datarijen: deze slaan op hetzelfde pand en komen dubbel voor of deze hebben betrekking op een gelijkaardig pand op hetzelfde adres. Hierbij is het niet steeds duidelijk of dit eenzelfde of meerdere verschillende panden betreft. Tevens is het bij herhaalde publicaties over de maanden/jaren heen niet steeds duidelijk of het vernieuwde of nieuwe publicaties betreft.

Verder zijn er nog diverse andere redenen waarom gegevens verwant zijn, zoals ruimtelijk gecorreleerde items, of geclusterde data. Voorbeelden zijn: zoekertjes geplaatst in belangrijke vakantieperiodes, panden in omgeving van stortplaatsen, woningen met erg hoge KI e.a. Dergelijke variatie kan in een multilevel-analyse meegenomen worden. Het effect van pseudodata en data van schijnbaar dezelfde of gelijkaardige panden kan echter direct best bij aanvang geëlimineerd worden.

Onderzoek naar replica's gebeurde op basis van vergelijkingen via adresgegevens. Items met exact hetzelfde adres betreffen –met abstractie van jaartal van publicatie- een potentieel replica. Voor vage adresgegevens is er duidelijk een risico: er kan immers niet worden uitgesloten dat ze een replica van elkaar vormen. Aan de hand van de concrete casus Oost-Vlaanderen kan dit als volgt worden verduidelijkt. Na classificatie van onderling sterk afhankelijke datarijen in de casus kon de matrix als volgt worden opgedeeld:

-702 units waarvoor enkel postcode bekend – potentieel verwante data

-1 unit waarvoor enkel postcode bekend, maar postcode komt elders in de dataset niet meer voor (dus niet verwant)

-712 units die sterk verwant zijn met 539 andere units die als unieke rijen geselecteerd zijn (– deze data slaan in principe op hetzelfde pand en mogelijks op dezelfde verkoop/verhuur (vandaar enkel het weerhouden van de meest recente publicatie voor betreffend adres waardoor enkel deze 539 recods in aanmerking komen). Evenwel zijn bepaalde variabelen daardoor niet betrouwbaar (om bijvoorbeeld "aantal dagen online publicatie" te bepalen is dan een sommering vereist over de datarijen heen, rekening houdende met mogelijke overlap in de tijd)

-2850 unieke adresgegevens (hier beschouwd als niet sterk verwante data)

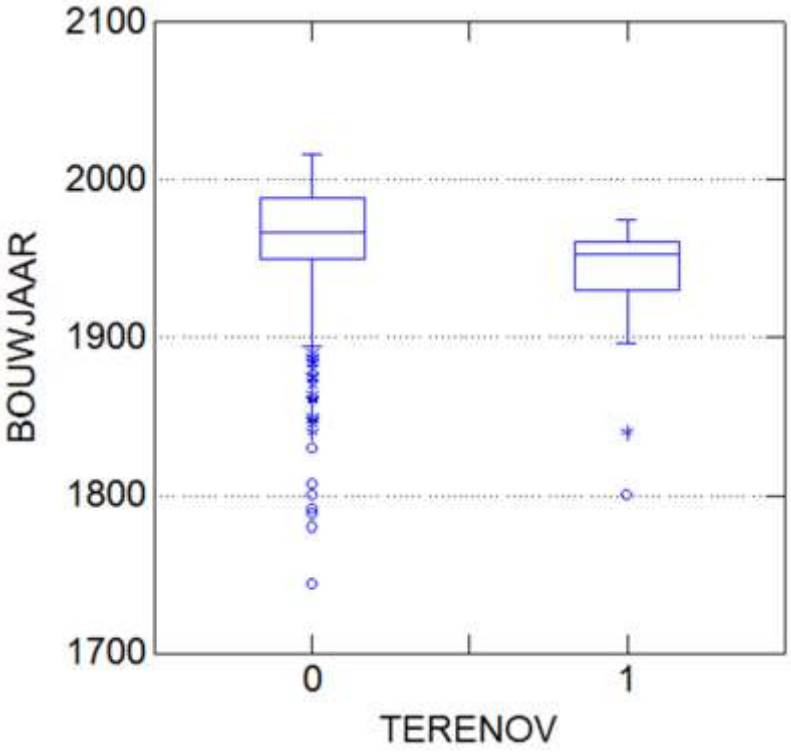
Dit geeft 3390=1+539+2850 "niet sterk verwante data"; of nog 70,57% van de 4804 datarijen is ok, hier weliswaar aangevuld met de 702 postcodedata. De casus Oost-Vlaanderen wijst ook in dit opzicht wellicht op gelijkaardige percentages voor de ganse dataset.

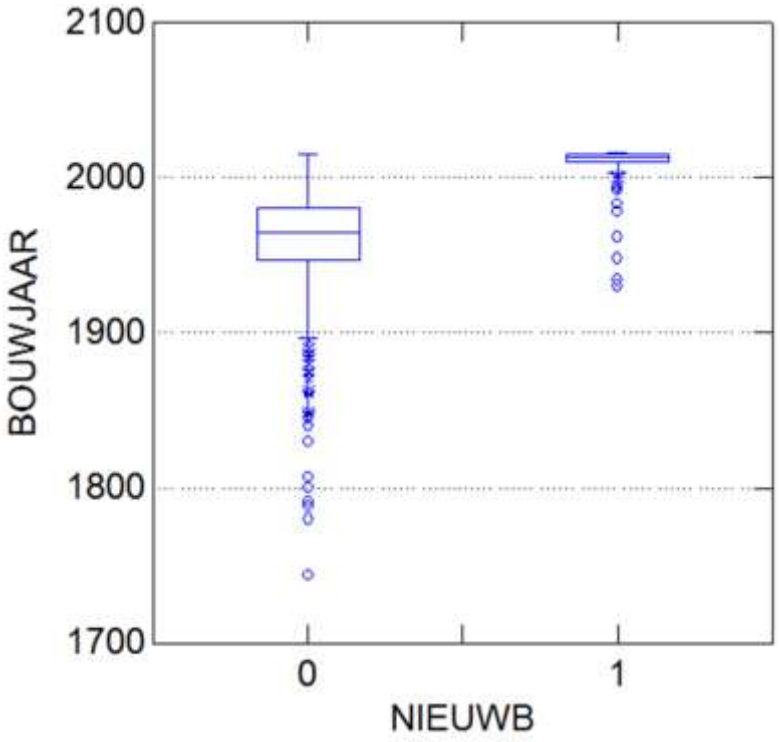
1.3.3 Successieve variabelenfiltering met attentie voor redundantie

De variabelen type_id, status_id, bebouwing_id, epc_attest, tuinaanwezig en creatie_datum werden voor verdere analyse geselecteerd. Deze variabelen veroorzaken geen significante vertekening; maar gezien ze het voorwerp van exploratief onderzoek uitmaken, worden de records zonder respons verder uitgefilterd (waardoor het aantal beschikbare datarijen verder afneemt).

In principe geldt hetzelfde voor de variabelen epcwaarde, nieuwbouw en terenoveren. Maar voor deze laatste twee is de redundantie met bouwjaar te groot, waardoor ze verder niet meer in rekening gebracht werden. Bovendien lijken er in die variabelen ook enkele anomalieën te zitten (1 sterke piek voor bepaald bouwjaar (abnormaliteit), outliers e.a.). Epcwaarde vertoont tot slot een lage respons en kan best enkel ingezet worden o.v.v. subsetanalyse.

Figuur 4 - Boxplots met weergave van spreiding van data over 'bouwjaar' (voor zover data van bouwjaar beschikbaar), dit voor 'nieuwbouw' en 'te renoveren'.





Op basis van deze analyse bleven (voor de casus Oost-Vlaanderen) 2296 records over (dus met respons voor al de relevante velden), waarvan:

-4 met type=grond => geëlimineerd
-3 met status=overname => geëlimineerd
-30 met bouwjaar onduidelijk
-93 met KI<100

Voor 'bouwjaar' bleken 30 datarijen geen betrouwbaar jaartal (o.a. wel tekstveld) te bevatten waardoor eliminatie van deze datarijen gebeurde ingeval bouwjaar geanalyseerd werd.

Ook 93 records met KI <100 (soms 0) bleken verdacht, waardoor deze niet meegenomen werden in analyses waar KI betrokken werd.

Suggesties voor surrogaatvariabelen / clustering

-De variabelen 'terenoveren' en 'nieuwbouw' dienen best in samenhang met bouwjaar bekeken te worden.

1.4 Variabelenselectie

Op basis van de voorgaande analyses volgt hieronder een selectie van de op te nemen variabelen die al dan niet 100% respons dienen te vertonen. Hierdoor wordt meteen ook het aantal resterende datarijen vastgelegd en derhalve de potentie voor verdere analyses afgelijnd. Onderstaande tabel biedt – op basis van de voorgaande stappen – een overzicht van alle variabelen en de variabelenset (in lichtgroen) voor de vervolganalyse.

De variabelenset wordt verder onderzocht om de immodata verder te herleiden tot een werkbare databank. Overige informatie die niet dadelijk wordt toegepast voor analyse wordt uiteraard behouden. Indien nodig kan steeds naar een ruimere subset worden teruggegrepen, uiteraard niet zonder de nodige gevolgtrekkingen en randvoorwaarden m.b.t. specifieke onderzoeksdoelen, analysetechnieken en analytisch vermogen in het algemeen.

Tabel 19 – overzicht van de opgenomen variabelen

Voorselectie	perc_prov	perc_alg	Cluster (cf. surrogaat)	Opmerking	Finale selectie	TypeVariabele	Veldtype
a_beschr_nl	67,7	74,9		tekstveld: niet bruikbaar voor analyse (wel fonetisch), ter info			
a_beschrkort_nl	9,1	10,0		tekstveld: niet bruikbaar voor analyse (wel fonetisch), ter info			
a_bus	7,2	8,3	locatie/postcode	verwerkt ifv adresgegevens			
a_gemeente	99,5	99,7	locatie/postcode	verwerkt ifv adresgegevens	zie immo_City	nominaal	tekst
a_geo_lat	99,2	99,0	locatie/postcode	verwerkt ifv adresgegevens	zie immo_X_lamb	continu	numeriek
a_geo_lon	99,2	99,0	locatie/postcode	verwerkt ifv adresgegevens	zie immo_Y_lamb	continu	numeriek
a_geo_precision	99,8	99,7	locatie/postcode	verwerkt ifv adresgegevens			
a_land_id	99,8	99,8	locatie/postcode	verwerkt ifv adresgegevens			
a_nummer	41,7	46,4	locatie/postcode	verwerkt ifv adresgegevens	zie immo_Hnr	nominaal	numeriek
a_omgeving_id	4,3	5,7		vervalt, data via spatiale analyse en externe databank			
a_postcode	100,0	100,0	locatie/postcode	verwerkt ifv adresgegevens	**zie immo_Postcode**	nominaal	**numeriek**
a_sleutelnummer	0,4	0,6		niet van toepassing			
a_status_id	100,0	100,0	status		a_status_id	nominaal	tekst
a_straat	56,6	60,9	locatie/postcode	verwerkt ifv adresgegevens	zie immo_Street	nominaal	tekst
a_subtype_id	88,7	89,2	(sub)type	teveel subtypes, herclassificatie nodig of beperkte selectie	subtype	nominaal	tekst
a_titel_nl	26,6	29,7		tekstveld: niet bruikbaar (wel fonetisch), ter info			
a_type_id	97,6	97,2	(sub)type		a_type_id	nominaal	tekst
a_vrijop_datum	3,0	3,9		niet zinvol			
a_vrijop2_id	5,6	7,1		niet zinvol			
b_bebouwing_id	18,3	18,5	bebouwing		b_bebouwing_id	nominaal	tekst
b_bouwjaar	21,7	26,4	bouwjaar/renovatie		b_bouwjaar	interval	numeriek
b_bouwopp	4,2	4,7	oppervlakte	enkel in subset(analyse)	zie immo_bouwopp	continu	numeriek
b_epc_attest	91,2	91,7	epc		b_epc_attest	nominaal	numeriek
b_epcwaarde	16,7	20,1	epc		b_epcwaarde	continu	numeriek
b_grondopp	29,9	28,9	oppervlakte		zie immo_grondopp	continu	numeriek
b_kadastraleaard	0,8	0,8	(sub)type	enkel in subset(analyse)	b_kadastraleaard	nominaal	tekst
b_kadastraleafdeling	1,0	1,2	kadaster	enkel in subset(analyse)	b_kadastraleafdeling	nominaal	tekst
b_kadastralesectie	1,3	1,5	kadaster	enkel in subset(analyse)	b_kadastralesectie	nominaal	tekst
b_nieuwbouw	97,8	98,1	bouwjaar/renovatie		b_nieuwbouw	nominaal	binair
b_perceelbreedte	2,8	3,5	oppervlakte	enkel in subset(analyse)	zie immo_perceelbreedte	continu	numeriek
b_perceeldiepte	1,6	1,9	oppervlakte	enkel in subset(analyse)	zie immo_perceeldiepte	continu	numeriek
b_perceelnummer	1,2	1,5	kadaster	vervalt, data via spatiale analyse en externe databank	zie immo_GRB_adpcapakey	nominaal	tekst
b_renovatiejaar	2,7	3,5	bouwjaar/renovatie	vervalt: niet egaal ingevuld, laag %, hooguit beperkt provinciaal			
b_staat_id	2,3	2,9	bouwjaar/renovatie	vervalt: niet egaal ingevuld, laag %, hooguit beperkt provinciaal			
b_terenoveren	97,9	98,0	bouwjaar/renovatie		b_terenoveren	nominaal	binair
b_woonopp	45,8	49,9	oppervlakte		zie immo_woonopp	continu	numeriek
creatie_datum	99,7	99,6	datum	verwerkt ifv datumgegevens	zie immo_creatie_datum	interval	datum
f_btwstelsel_grond	5,4	7,0					
f_forfait_aansluitingskosten	0,1	0,1		zeer laag percentage respons			
f_forfait_basisakte_notaris	0,3	0,3		zeer laag percentage respons			
f_huurwaarborg	0,4	0,5		zeer laag percentage respons			
f_jaarhuur	0,1	0,1		zeer laag percentage respons			
f_ki	17,0	15,0	KI		zie immo_f_ki	continu	numeriek
f_kiindex	1,7	2,2	KI	enkel in subset(analyse)	zie immo_f_kiindex	continu	numeriek
f_kiindex_jaar	0,1	0,1		zeer laag percentage respons			
f_locatie_id	100,0	100,0		vervalt			
f_maandelijks	0,2	0,2		zeer laag percentage respons			
f_nettoopbrengst	0,0	0,0		100% non-respons			
f_onroerendevoorheffing	0,0	0,1		zeer laag percentage respons			
f_onroervh	0,6	0,8		zeer laag percentage respons			
f_onroervh_jaar	0,1	0,2		zeer laag percentage respons			
f_overnameprijs	0,0	0,0		100% non-respons			
f_prijs	97,6	97,7	prijs		zie immo_f_prijs	continu	numeriek
f_prijsgrondaandeel	0,9	1,1		nauwelijks ingevuld			
f_prijsvm	0,1	0,2		zeer laag percentage respons			
f_prijszichtbaar	99,4	99,2	prijs		prijszichtbaar	nominaal	binair
f_provisie	0,2	0,2		zeer laag percentage respons			
is_gearchiveerd	100,0	100,0	datum	verwerkt ifv datumgegevens			
n_ligging	0,0	0,0		100% non-respons			
n_project_id	100,0	100,0		niet zinvol			
notaris_bouquet	0,0	0,0		100% non-respons			
notaris_instelprijs	0,0	0,0		100% non-respons			
notaris_max_looptijd_lijfrente	0,0	0,0		100% non-respons			
notaris_plaats_openbareverkoop	0,0	0,0		100% non-respons			
notaris_prijs_gebracht_op	0,0	0,0		100% non-respons			
notaris_recht_hoger_bod_datum	0,0	0,0		100% non-respons			
notaris_rente	0,0	0,0		100% non-respons			
notaris_toegewezen_ovhb_aan	0,0	0,0		100% non-respons			
notaris_verkooptype_id	0,6	0,4		zeer laag percentage respons			
o_garages	5,3	6,1	faciliteiten	interpretatierisico's	o_garages	ordinaal (deels continu)	numeriek
o_openbaarvervoer	2,8	3,6	faciliteiten	vervalt, data via spatiale analyse en externe databank	o_openbaarvervoer	continu	numeriek
o_orientatie_id	2,7	3,2	oriëntatie	vervalt, data via spatiale analyse en externe databank	o_orientatie_id	nominaal	tekst
o_parkings	3,5	4,0	faciliteiten	interpretatierisico's	o_parkings	ordinaal (deels continu)	numeriek
o_school	2,4	3,1	faciliteiten	vervalt, data via spatiale analyse en externe databank	o_school	continu	numeriek
o_strand	0,1	0,2	strand/zee	enkel West-Vlaanderen + data via spatiale analyse en externe databank	o_strand	continu	numeriek
o_tuinaanwezig	97,2	97,3	tuin		o_tuinaanwezig	nominaal	binair
o_tuintekst_nl	1,5	1,8	tuin	tekstveld: niet bruikbaar voor analyse (wel fonetisch), ter info			
o_winkels	2,8	3,7	faciliteiten	vervalt, data via spatiale analyse en externe databank	o_winkels		
o_zichtopzee	98,2	97,4	strand/zee	enkel West-Vlaanderen	o_zichtopzee		
o_zijdelingszichtopzee	99,3	98,9	strand/zee	enkel West-Vlaanderen	o_zijdelingszichtopzee		
r_datum	39,0	39,8	datum	verwerkt ifv datumgegevens			
r_prijs	2,6	3,4	prijs	enkel in subset(analyse)	r_prijs	continu	numeriek
showweb	100,0	100,0		vervalt			
wijziging_datum	99,9	99,9	datum	verwerkt ifv datumgegevens			
Pub1_start	85,1	82,7	datum	verwerkt ifv datumgegevens	zie immo_Pub1_start	interval	datum
Pub1_stop	77,3	74,8	datum	verwerkt ifv datumgegevens			
Pub2_start	0,2	0,3	datum	verwerkt ifv datumgegevens			
Pub2_stop	0,1	0,2	datum	verwerkt ifv datumgegevens			
Pub3_start => Pub10_start				100% non-respons			
Pub3_stop => Pub10_stop				100% non-respons			
immo_Uniek_ID			ID	NIEUW	immo_Uniek_ID		numeriek
immo_Postcode			locatie/postcode	NIEUW	immo_Postcode		numeriek
immo_City			locatie/postcode	NIEUW	immo_City		tekst
immo_Street			locatie/postcode	NIEUW	immo_Street		tekst
immo_StreetId			locatie/postcode	NIEUW	immo_StreetId		tekst
immo_Hnr			locatie/postcode	NIEUW	immo_Hnr		tekst
immo_Precisie_Locatie			locatie/postcode	NIEUW	immo_Precisie_Locatie		numeriek
immo_X_lamb			locatie/postcode	NIEUW	immo_X_lamb		numeriek
immo_X_lamb_input			locatie/postcode	NIEUW	immo_X_lamb_input		numeriek
immo_Y_lamb			locatie/postcode	NIEUW	immo_Y_lamb		numeriek
immo_Y_lamb_input			locatie/postcode	NIEUW	immo_Y_lamb_input		numeriek
immo_Geo_diff_distance			locatie/postcode	NIEUW	immo_Geo_diff_distance		numeriek
immo_Geo_diff_OK			locatie/postcode	NIEUW	immo_Geo_diff_OK		binair
immo_Pub1_start			datum	NIEUW	immo_Pub1_start		datum
immo_Pub1_stop			datum	NIEUW	immo_Pub1_start		datum
immo_Pub1_dagen			datum	NIEUW	immo_Pub1_dagen		numeriek
immo_f_prijs			prijs	NIEUW	immo_f_prijs		numeriek
immo_f_ki			prijs	NIEUW	immo_f_ki		numeriek
immo_datum			datum	NIEUW	immo_datum		datum
immo_jaartal			datum	NIEUW	immo_jaartal		numeriek
immo_verwant			afhankelijkheid	NIEUW	immo_verwant		numeriek
immo_bouwopp			oppervlakte	NIEUW	immo_bouwopp		numeriek
immo_grondopp			oppervlakte	NIEUW	immo_grondopp		numeriek
immo_woonopp			oppervlakte	NIEUW	immo_woonopp		numeriek
immo_GRB_adpcapakey			kadaster	IMPUTATIE	immo_GRB_adpcapakey		tekst
immo_GRB_adpoppervl			oppervlakte	IMPUTATIE	immo_GRB_adpoppervl		numeriek
immo_GRB_gbgtotopp			oppervlakte	IMPUTATIE	immo_GRB_gbgtotopp		numeriek
immo_creatie_datum			datum	NIEUW (optioneel in databank)	immo_creatie_datum		
immo_perceelbreedte			oppervlakte	NIEUW (optioneel in databank)	immo_perceelbreedte		
immo_perceeldiepte			oppervlakte	NIEUW (optioneel in databank)	immo_perceeldiepte		
immo_BuildingTotArea			locatie/postcode	NIEUW (optioneel in databank)	immo_BuildingTotArea		
immo_Checked_Postcode			locatie/postcode	NIEUW (optioneel in databank)	immo_Checked_Postcode		
immo_Checked_Street			locatie/postcode	NIEUW (optioneel in databank)	immo_Checked_Street		
immo_OK_Geocode			locatie/postcode	NIEUW (optioneel in databank)	immo_OK_Geocode		
immo_OK_PostCode			locatie/postcode	NIEUW (optioneel in databank)	immo_OK_PostCode		
immo_OK_Street			locatie/postcode	NIEUW (optioneel in databank)	immo_OK_Street		
immo_ParcelArea			oppervlakte	NIEUW (optioneel in databank)	immo_ParcelArea		
immo_ParcelKey			kadaster	NIEUW (optioneel in databank)	immo_ParcelKey		
immo_ParcelNr			kadaster	NIEUW (optioneel in databank)	immo_ParcelNr		
immo_Phonetic_Street			locatie/postcode	NIEUW (optioneel in databank)	immo_Phonetic_Street		
immo_Remark_PostCode			locatie/postcode	NIEUW (optioneel in databank)	immo_Remark_PostCode		
immo_Remark_Street			locatie/postcode	NIEUW (optioneel in databank)	immo_Remark_Street		
immo_Result_Geocode			locatie/postcode	NIEUW (optioneel in databank)	immo_Result_Geocode		

1.5 Globaal overzicht van de variabelen, casus Oost-Vlaanderen

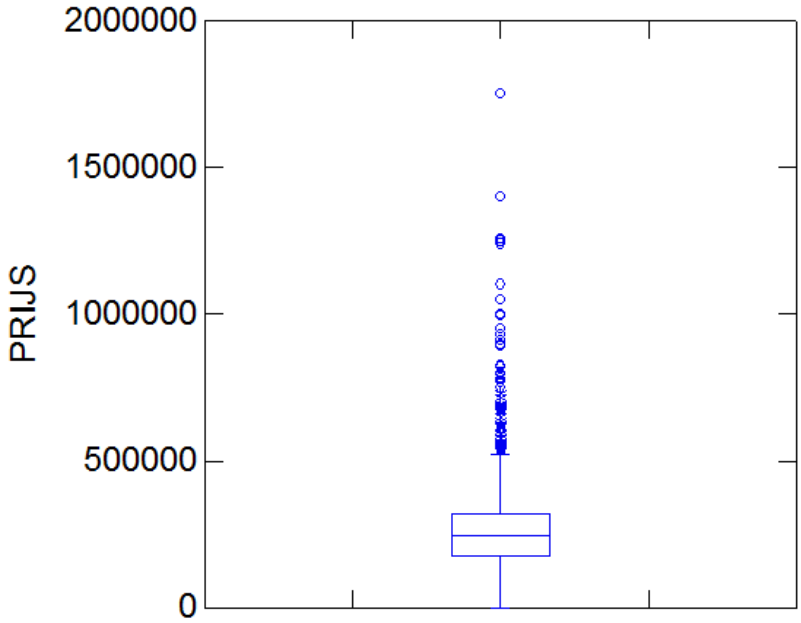
De variabelenset die uit bovenstaande analyse geselecteerd is, wordt verder geanalyseerd, met onderscheid naar de kwantitatieve en nominale data.

1.5.1 Kwantitatieve variabelen

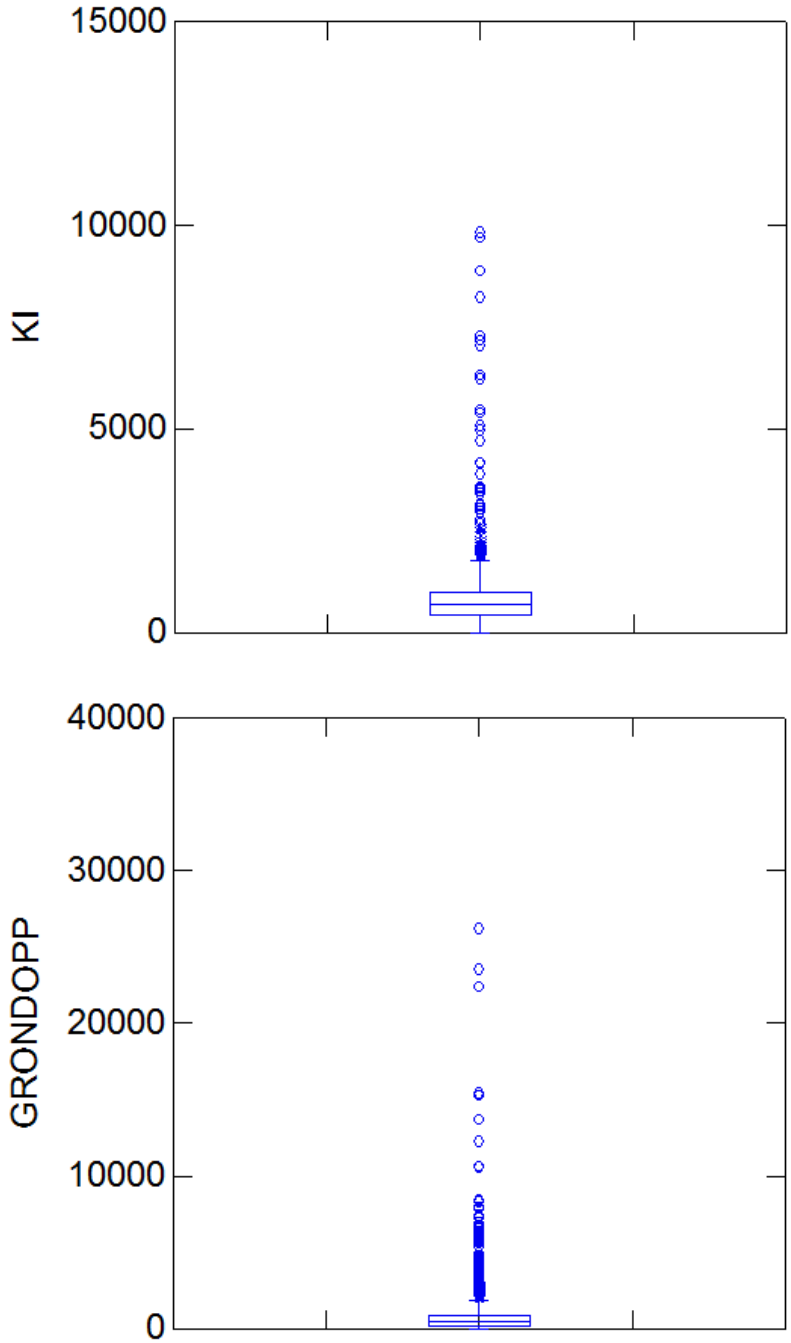
Voor kwantitatieve variabelen kan via boxplots een beeld verkregen worden van de spreiding van de variabelen en het aantal outliers. Dit gebeurt hier voor prijs (met a priori eliminatie van 1 waarde van 3,5 miljoen), KI, grond-/woon-/ en bouwoppervlakte. Voor bouwoppervlakte wordt a priori 1 waarde van 90000 m² geëlimineerd wegens onrealistisch).

Het blijkt dat voor al de genoemde variabelen er tal van outliers (naar boven toe) voorkomen. Voor de hoge waarden kan best controle gebeuren of deze correct kunnen zijn, outliers zijn immers niet per definitie foute data. Vervolgens kunnen de x% hoogste waarden eventueel niet beschouwd worden indien ze effectief het gevolg zijn van foute invoer van data. Hiervoor is echter een realiteitscheck aangewezen, wat geen onderdeel uitmaakte van deze studie. Om indien nodig de verdeling te verbeteren kan eventueel met logaritmen gewerkt worden.

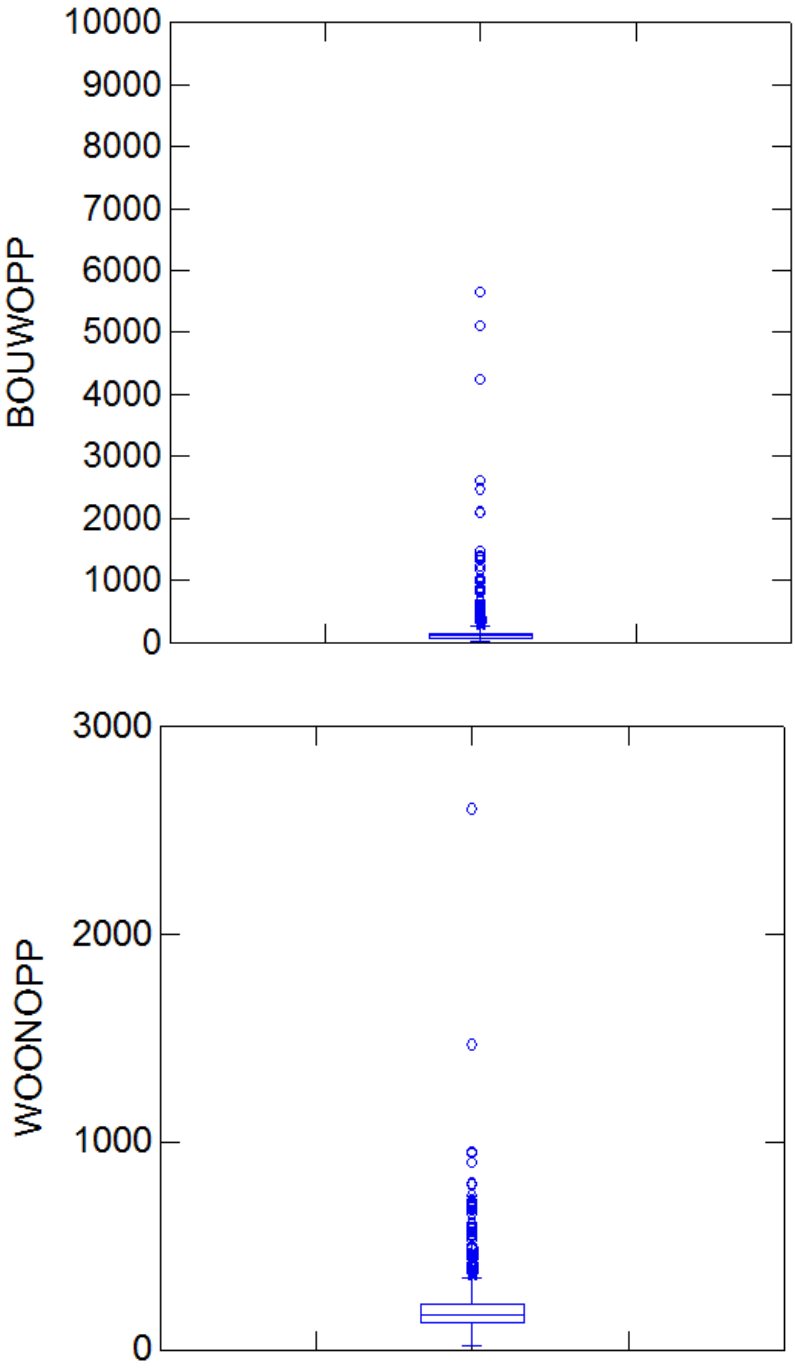
Figuur 5 - Boxplots met weergave van spreiding van data voor "prijs"



Figuur 6 - Boxplots met weergave van spreiding van data voor “KI” en “grondopp”



Figuur 7 - Boxplots met weergave van spreiding van data voor “Bouwopp” en “Woonopp”



1.5.2 Nominale data

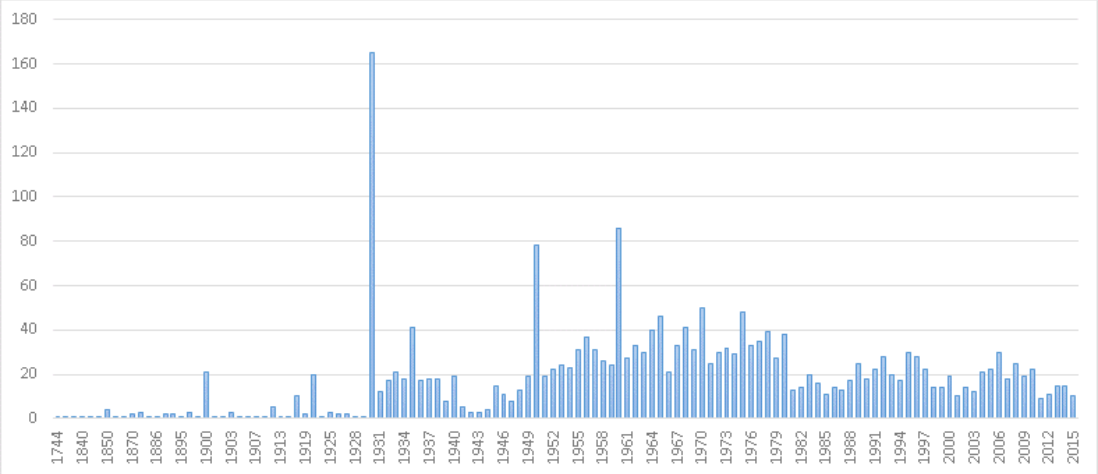
Onderstaande tabellen geven een overzicht van de datakarakteristieken voor de relevantste nominale variabelen in de subset (casus Oost-Vlaanderen). Het valt op dat de huurmarkt beperkt vertegenwoordigd is. Verder is er een redelijke spreiding in bebouwingstype (gesloten tot open) en zijn er relatief veel tuinen aanwezig. Het jaar van publicatie geeft, voor zover ingevuld, 2006 aan als oudste jaartal. Wat bouwjaar aangaat is er na 1930 een vrij redelijke spreiding. Het jaartal 1930 springt er echter opvallend uit wat wellicht een (nep)effect (=artefact) is doordat dit jaartal ook vermeld wordt voor woningen ouder dan 1930.

Tabel 20 - datakarakteristieken voor de relevantste nominale variabelen in de subset (casus Oost-Vlaanderen)

b_bebouwing_id	**#**
gesloten	714
halfopen	534
open	1041
o_tuinaanwezig	**#**
0	583
1	1706
b_epc_attest	**#**
1	331
2	324
3	1633
jaartal	**#**
2006	3
2007	27
2008	27
2009	59
2010	90
2011	86
2012	249
2013	439
2014	610
2015	613
2016	86

type	a_status_id	#
appartement	Te Huur	11
"	Te koop	49
"	Verhuurd	5
"	Verkocht	56
bedrijf	Te Huur	2
"	Te koop	80
"	Verhuurd	2
"	Verkocht	37
woning	Te Huur	16
"	Te koop	1015
"	Verhuurd	18
"	Verkocht	998

Figuur 8 – spreiding van de data over "bouwjaar"



1.5.3 Globale samenhang van variabelen

Een eerste globale multivariate analyse van de dataset van al de numerieke variabelen en de bruikbare categorische variabelen (na introductie van enkele dummies voor o.a. type) levert een aantal opvallende maar ook evidente associaties op (cf. correlatie-biplot):

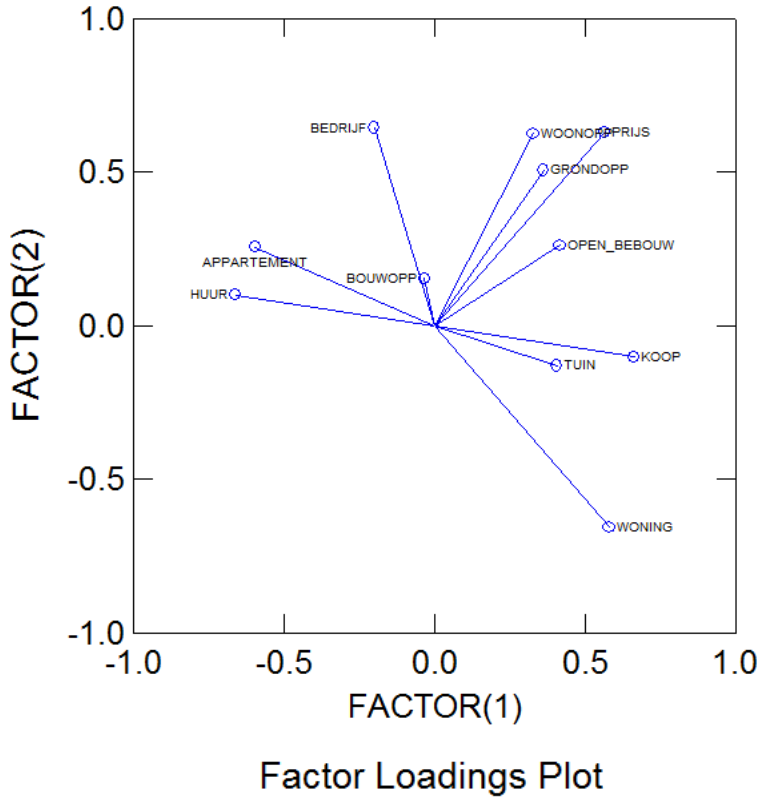
-appartementen in de immodatabank staan vooral te huur, woningen vooral te koop.

-prijs en woonoppervlakte zowel als grondoppervlakte zijn sterk gecorreleerd.

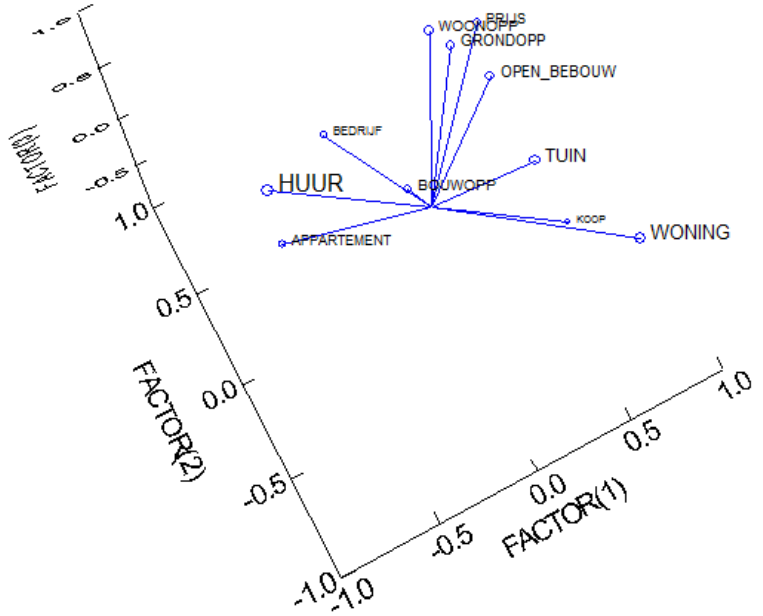
-open bebouwing is in zekere mate positief geassocieerd met tuin en met (hoge) prijs.

-bouwoppervlakte lijkt met overige variabelen niet direct (lineair) samen te hangen.

Figuur 9(a-b) - Globale samenhang tussen numerieke variabelen in de dataset: grondoppervlakte en woningoppervlakte zijn sterk gerelateerd aan een hogere kostprijs. Dit is eveneens het geval bij open bebouwing. Appartementen worden cf. deze dataset overwegend verhuurd, woningen vooral verkocht.



Factor Loadings Plot



Er kan besloten worden dat de eerste drie assen de voornaamste variabelen (incl. dummies) uit de immodata de data voor ruim 58% verklaren (F1: 22,8%, F1-F2: 41,7%) en samenvatten. De belangrijkste variabelen vertonen in de regel dan ook veel samenhang (cf. prijs) waarmee alvast sterk rekening dient gehouden. (voor meer informatie m.b.t. principal components factoring zie Jolliffe 2002).

Een structurele decompositie zou uiteraard meer (specifieke) associaties aan het licht brengen maar vormt in deze fase vooralsnog geen aandachtspunt. Voor de inhoudelijke analyse wordt doorverwezen naar het luik ruimtelijke analyses.

1.6 Sampling bias

Om te vermijden dat er sprake zou zijn van een voor het studiegebied of onderzoeksobject niet representatieve staalname (sample), is een stratified random sample van de voorhanden zijnde dataset wenselijk. De stratificatie volgt na in te stellen criteria (Bv. of evenwicht naargelang de huurmarkt al dan niet moet worden meegenomen in analyses. Bv. conditie inzake spreiding over provincies/gemeenten/regio's. Of nog de verzekering dat alle soorten/types vastgoed er evenwichtig in vervat zitten). Dit kan bv. eisen stellen t.a.v. representativiteit voor gans het grondgebied, waardoor dan bijgevolg rekening moet worden gehouden met aandelen van bv. appartementen in de huurmarkt, of nog met bv. vertegenwoordiging van villa's, boerderijwoningen enz. (alle segmenten voldoende representatief). Cruciaal lijkt vooral het formuleren van minimumeisen (minstens x aantal datarijen voor betreffend type, specifieke condities per sub(markt)segment zoals per postcode minstens x aantal panden met te specifiëren variatie in afstand tot faciliteiten enz…).

Dergelijke criteria kunnen bezwaarlijk op voorhand in detail worden geformuleerd, gezien dit afhankelijk is van specifieke analyses. In volgend hoofdstuk wordt wel dieper ingegaan op de analyseset en haar kenmerken waardoor het mogelijk wordt om een meer concrete inschatting te verkrijgen van stratified sampling voor de diverse variabelen en postcodeclusters.

2. Afleiding van de analyseset en haar kenmerken

In dit hoofdstuk wordt op basis van de gegeolocaliseerde immodatabank met unieke adressen (zie deel 1) in combinatie met de inzichten uit de exploratieve analyse, een analyset opgemaakt in functie van het opstellen van een aantal beleidsindicatoren (deel 3). Deze analyseset wordt opnieuw onderworpen aan een exploratieve statistische analyse.

2.1 Analyseset

De dataset met 'unieke adresgegevens´ met een totaal van 531.040 records (= gecombineerde unieke primary key op de velden a_postcode, a_straat, a_nummer, met selectie van de recentste immo_Pub1_start datum), omvat nog adresgegevens waarvoor geolocaliatie niet correct kon gebeuren. Gezien voor de onvolledige records niet zonder meer kan gesteld worden dat ze onafhankelijk zijn, werden striktere selectiecriteria ingesteld, met eliminatie van data tot gevolg:

1/ data waarvoor 'immo_Precisie_Locatie' een codering '3' of '-1' weergeeft, werden dan verder niet meer beschouwd.

Cf. immo_Precisie_Locatie :

- 1 : adres (postcode, straatnaam, huisnr) volledig correct
- 2 : enkel postcode, straatnaam correct en voor het huisnr werd het numeriek minst verschillende huisnr genomen
- 3 : enkel postcode correct
- 4 : enkel postcode, straatnaam correct en voor het huisnr werd het huisnr uit de CRAB adrespunten genomen waarvan de x, y positie minder dan 10m in afstand verschilt van de immo_X_lamb_input en immo_Y_lamb_input
- -1 : niets correct van het adres

2/ de afgeleide subset bevat enkel nog data met voor de variabele 'a_status_id' de invulling 'verkocht', 'te koop', 'verhuurd' of 'te huur'; en met voor 'a_type_id' de invulling 'woning' ofwel 'appartement'. Een subset voor 'a_type_id = grond' werd separaat aangemaakt en verderop apart beschouwd (slechts 19050 records).

3/ voor woning/appartement subset werden voor de variabele 'immo_f_prijs' datarijen met voor deze cruciale variabele lege cellen of waarden gelijk aan nul geëlimineerd.

Aan de resulterende dataset (met 340.276 records) werden enkele extra 'immo variabelen' toegevoegd:

- immo_jaartal — jaar van publicatie (creatie start)
- immo_postcodecluster — code voor postcodecluster (?000-tallen) waartoe pand behoort
- immo_bouwjaar — bouwjaar (voor zover ingevuld en verondersteld mogelijks correct)
- immo_bouwdecennium — bouwjaar herleid tot eeuw (1500-1800) of decennium (vanaf 1900)

Vervolgens werd een prospectie van uitbijters (outliers) ondernomen (woningen/appartementen, koop/huur). Er werd geopteerd om ook de meer extreme uitbijters voor verdere analyse blijvend in aanmerking te nemen omdat :

1/ deze hier reeds gemerkt worden ('knipperlichteffect')

2/ elke verdere inhoudelijke analyse een specifieke (gerichte) outlieranalyse op maat dient te hanteren

Desalniettemin dient er nadrukkelijk op gewezen dat diverse statistische analyses gevoelig zijn voor alvast meer extreme outliers – wel afhankelijk van type analyse en dus verder op te volgen. Bovendien stellen de meest extreme outliers soms zonder meer regelrechte fouten voor: bv.. een '0' teveel in de variabele 'f_prijs', of negatieve waarden voor oppervlaktes…

Om deze reden werden, op basis van kwartielkwaarden en interkwartielfafstanden (>1,5*IKA=zachte uitbijter, >3,0*IKA=harde uitbijter) indicaties gegenereerd t.b.v. de aanwijzing van records die aandacht vergen gezien hun relatief extreme waarde: zachte uitbijters (1), harde uitbijters (2) en (onterechte) nul-waarden (0) die feitelijk lege cellen (= niet ingevuld) impliceren. Deze oefening werd gedaan voor de relevante numerieke variabelen en de outlierkwalificaties werden telkens o.v.v. een 'stat variabele' in de databank ingebracht, als volgt :

-stat_b_perceelbreedte_outsh zachte/harde uitbijters voor de variabele 'b_perceelbreedte'

-stat_b_perceeldiepte_outsh zachte/harde uitbijters voor de variabele 'b_perceeldiepte'

-stat_ParcelArea_outsh zachte/harde uitbijters voor de variabele 'ParcelArea'

-stat_b_grondopp zachte/harde uitbijters voor de variabele 'b_grondopp'

-stat_b_bouwopp zachte/harde uitbijters voor de variabele 'b_bouwopp'

-stat_b_woonopp zachte/harde uitbijters voor de variabele 'b_woonopp'

-stat_Opp_TotGebouw_outsh zachte/harde uitbijters voor de variabele 'Opp_TotGebouw'

-stat_f_prijs_outsh zachte/harde uitbijters voor de variabele 'f_prijs', na separate analyse van verhuur en verkoop, dit separaat voor woningen en appartementen

-stat_f_ki_generiek_outsh zachte/harde uitbijters voor de variabele 'f_ki'

-stat_f_ki_appart_outsh zachte/harde uitbijters voor de variabele 'f_ki', wanneer louter de verkoop van appartementen beschouwd

Indien een afgeleide dataset voor woningen/appartementen zou worden gecreëerd met eliminatie van alle harde uitbijters uitgenomen deze voor f_ki inzake appartementenverkoop (gezien voor f_ki op zich wel reeds werd opgepikt), dan zou de dataset herleid worden van 340276 unieke gegevens tot 298953 datarijen.

Echter de detectie van outliers rechtvaardigt hun eliminatie niet zonder meer. Dit dient zoals hogerop reeds gesteld bekeken op basis van de analysetechnieken die zullen worden toegepast i.k.v. specifieke analyses (en dus niet in kader van de globale analyse die binnen dit project wordt uitgevoerd).

Verder dient in hypothetisch geval met de gereduceerde set van 298953 records - als gevolg van non-respons voor relevante data - omzichtig omgesprongen gezien de mogelijke bias. Indien vereist zou worden dat voor alle variabelen die potentieel non-responsbias vertonen ook effectief een (positieve) waarde ingevuld staat, dan blijven slechts 4556 datarijen over. Echter een analyse die dergelijke vereiste over al deze variabelen heen zou stellen is eerder onwaarschijnlijk, en de dataset van 298953 records biedt over de regio's/provincies heen toch wel aanzienlijke analytische potenties.

Advies is hoe dan ook met alle datarijen verder te werken, rekening houdende met mogelijke non respons bias, met (zachte) uitbijters alsook (onterechte) nulwaarden, en met de principes van stratified sampling. M.b.t. dit laatste dient informatie inzake de verdeling van belangrijke (klasse)variabelen zich globaal als volgt aan :

- Het relatief beperkt aantal gegevens voor Wallonië kan zekere limieten stellen aan een landsbrede analyse. Bv. gemeenten behorende tot postcodecluster 7: slechts 8 data van verhuurde appartementen in halfopen bebouwing.
- wanneer meerdere categorische variabelen in beschouwing genomen worden, dan kan lang niet voor alle klassen een zinvolle analyse uitgevoerd worden. Bv. verkoop van vrijstaande, te renoveren woning met tuin: slechts 1 enkele casus voor gemeenten behorende tot postcodecluster 5.

- voor gesloten bebouwing zijn er de meeste datarijen voorhanden

Er stelt zich verder de vraag hoe deze analyseset, afgeleid vanuit de immodatabank zich verhoudt tot de volledige populatie. Hiervoor verwijzen we door naar de resultaten van de ruimtelijke analyse (zie deel 3).

2.2 Basiskenmerken van de analyseset

2.2.1 Non-respons

De non-respons voor de analyseset werd opnieuw in beeld gebracht, dit uitsluitend voor de variabelen die geacht worden een (systematische) vertekening (zogenaamde *bias*) te kunnen veroorzaken in relatie tot de prijsvariabele (f_prijs). De analyse van de non-respons gebeurde separaat voor:

- Woningen te koop/verkocht
- Woningen te huur/verhuurd
- Appartementen te koop/verkocht
- Appartementen te huur/verhuurd
- Gronden

In bijlage wordt een systematisch beeld gegeven van het percentage respons per relevante immo-variabele. Omdat voor 'grond' minder variabelen van toepassing zijn resulteert dit in een kleiner aantal figuren.

Volgende zaken vallen globaal op:

- De (non-)responsrate is erg uiteenlopend, niet enkel naargelang de variabele van toepassing maar ook sterk afhankelijk van de postcodecluster: opvallend minder respons voor Wallonië, en lage respons voor immo_f_index, immo_b_perceeldiepte & -breedte (amper ingevuld voor Wallonië behalve voor gronden), immo_b_bouwoppervlakte. De (opname in eender welke) analyse van precies die variabelen zal dus een algeheel probleem kunnen stellen, of bv. kunnen leiden tot een sterke reductie in aantal te betrekken records (met ook gevolgen voor de ruimtelijke representativiteit).
- Bij woningen en appartementen te koop zijn immo_bouwjaar en immo_f_ki relatief goed ingevuld, veel meer dan bij panden te huur.
- De variabelen immo_b_grondoppervlakte en immo_b_bouwoppervlakte zijn enkel redelijk (maar beperkt) ingevuld bij woningen te koop, uitgezonderd bij gronden zelf waar immo_b_grondoppervlakte wel vrij goed scoort in de responsrate.
- Het aantal dagen online (immo_Pub1_dagen) is iets minder goed ingevuld voor panden te koop dan bij panden te huur.

Dit alles stelt zekere limieten naar het uniform en landsbreed analyseren van data die *bias*gevoelig zijn. Zo zullen variabelen als perceelsbreedte en bouwjaar, die toch vrij cruciaal zijn naar prijszetting toe, met expliciete attentie voor al dan niet (non-)respons moeten worden geanalyseed. Dit impliceert dat indien geen aanvullende data ter beschikking zijn, prioritaire aandacht dient besteed aan de effecten van non-respons, en meer bepaald aan de afwijking die dit op de gemiddelde prijs voor regio *x* en conditie *y* meebrengt. De data zouden bv. in geval van non-respons aanleiding kunnen geven tot een overschatting van gemiddelde prijzen indien ook data met respons (bv. ingeval bouwjaar vaker dan gemiddeld recent) zonder onderscheid zouden worden samengenomen. Ook in diverse andere opzichten en zeker ook voor regiospecifieke analyses dienen de nodige afwegingen of gevolgtrekkingen m.b.t. non-respons gemaakt te worden voor minstens de meest relevante (hier geanalyseerde) variabelen uit de Zimmo-databank.

2.2.2 Inhoudelijk

We geven hier een tabelmatig overzicht van de belangrijkste kenmerken van de analyseset, meer bepaald deze van woningen/appartementen (koop/huur), dit voor :

- b_bebouwing_id
- o_tuinaanwezig
- b_terenoveren
- dit alles per postcodecluster

Zie onderstaande tabellen (Tabel 21 -Tabel 26).

Gevolgtrekkingen zijn vanzelfsprekend op basis van de gepresenteerde cijfers in de tabellen. Opvallend is dat te koop staande woningen in open bebouwing vooral uit de provincie Antwerpen komen. Dan is er vaak ook een tuin aanwezig en staat de woning erg zelden als te renoveren aangestipt. De aanwezige 'configuratie' kan daarbij deels indicatief zijn voor de reële situatie, maar dit is zeker niet uitsluitend het geval gezien ook zogenaamde sampling bias meespeelt: sommige gebieden en types bebouwing e.d. zijn nu eenmaal (relatief) minder goed vertegenwoordigd in de Zimmo-databank dan andere. Een indicatieve inhoudelijke detailanalyse gebeurt in ruimtelijk verband, wanneer ook algemene statistieken betrokken worden (zie verder).

Tabel 21 - Samengevatte tabel voor de woning (met telkens aantallen records)

			postcodecluster								
WONING	Verkoop	b_bebouwing_id	1	2	3	4	5	6	7	8	9
		-	7.827	18.903	12.844	6.507	2.603	5.693	5.173	17.802	19.399
		gesloten	1.309	7.427	2.949	1.035	262	597	925	4.563	3.123
		halfopen	1.126	3.802	4.348	835	362	595	594	3.179	2.511
		open	1.460	5.137	7.668	1.130	742	972	651	3.379	3.070
	Verhuur	b_bebouwing_id	1	2	3	4	5	6	7	8	9
		-	1.810	3.727	3.234	464	340	629	951	5.991	6.201
		gesloten	1.872	762	1.074	362	256	274	313	1.484	1.027
		halfopen	543	659	966	64	75	91	29	814	713
		open	593	955	1.689	108	103	88	60	1.071	969

Tabel 22 - Samengevatte tabel voor appartementen (met telkens aantallen records)

			postcodecluster								
APPARTEMENT	Verkoop	b_bebouwing_id	1	2	3	4	5	6	7	8	9
		-	3.506	6.774	2.948	1.023	315	532	609	6.322	2.787
		gesloten	809	2.670	943	134	92	129	77	1.375	554
		halfopen	198	882	499	52	26	36	14	211	152
		open	224	488	265	95	140	54	45	283	150
	Verhuur	b_bebouwing_id	1	2	3	4	5	6	7	8	9
		-	6.161	9.416	5.478	1.083	719	887	1.395	8.013	7.724
		gesloten	5.944	2.560	2.061	704	509	376	477	1.962	1.940
		halfopen	377	799	593	63	64	84	8	349	319
		open	387	690	361	97	129	90	54	356	441

Tabel 23 - detailtabel van de verdeling over de variabelen voor verkoop van woningen (met telkens aantallen records)

					postcodecluster								
		b_bebouwing_id	o_tuinaanwezig	b_terenoverren	1	2	3	4	5	6	7	8	9
WONING	Verkoop				0	0	0	0	0	0	0	11	0
				0	74	269	134	86	37	38	39	111	129
			0		3	534	2	2	37	65	2	453	16
			0	0	6.390	13.936	10.440	4.294	1.860	4.421	2.930	11.749	15.625
			0	1	6	15	11	2	0	0	0	16	10
			1		5	178	0	1	15	23	2	1.751	23
			1	0	1.340	3.851	2.201	2.122	653	1.127	2.195	3.665	3.493
			1	1	3	19	23	0	0	0	3	12	14
			1	f	1	1	1	0	0	1	0	1	1
			f	0	5	99	28	0	1	18	2	32	79
			f	f	0	1	4	0	0	0	0	1	9
		gesloten			0	3	0	0	0	0	0	0	7
		gesloten		0	44	48	18	52	15	6	10	11	21
		gesloten		1	1	0	0	2	0	0	1	0	1
		gesloten	0	0	316	2.230	772	300	104	245	319	1.357	1.243
		gesloten	0	1	9	113	74	4	0	5	2	103	49
		gesloten	1		0	7	1	0	0	0	0	0	2
		gesloten	1	0	905	4.768	1.906	673	142	335	592	2.925	1.714
		gesloten	1	1	34	258	178	4	1	6	1	167	86
		halfopen		0	12	9	13	27	5	6	6	3	11
		halfopen	0	0	174	790	748	179	113	181	152	620	854
		halfopen	0	1	3	23	43	1	0	3	2	35	14
		halfopen	1		0	0	0	0	0	0	0	0	3
		halfopen	1	0	914	2.807	3.317	626	243	393	430	2.399	1.534
		halfopen	1	1	23	164	225	2	1	12	4	116	83
		halfopen	f	0	0	9	2	0	0	0	0	6	12
		open			0	0	2	0	0	0	0	0	1
		open		0	6	5	29	9	6	2	1	2	3
		open		1	1	0	0	0	0	0	0	0	1
		open	0	0	305	1.363	1.209	234	159	299	178	739	1.147
		open	0	1	2	15	37	0	0	5	0	15	18
		open	1		0	0	3	0	0	1	0	0	5
		open	1	0	1.115	3.643	6.140	885	576	657	466	2.537	1.833
		open	1	1	31	110	247	2	1	8	6	84	60
		open	f	0	0	1	1	0	0	0	0	2	2

Tabel 24 - detailtabel van de verdeling over de variabelen voor verhuur van woningen ((met telkens aantallen records)

					postcodecluster								
		b_bebouwing_id	o_tuinaanwezig	b_terenoveren	1	2	3	4	5	6	7	8	9
WONING	Verhuur			0	4	42	18	8	5	1	6	15	46
			0		46	1	82	0	0	0	0	178	2
			0	0	1.267	2.784	2.400	368	290	582	849	4.201	4.720
			1		48	1	46	0	0	0	0	261	6
			1	0	442	885	673	88	45	46	96	1.326	1.398
			1	f	0	1	0	0	0	0	0	1	1
			f	0	3	12	14	0	0	0	0	8	24
			f	f	0	1	1	0	0	0	0	1	4
		gesloten			0	0	0	0	0	0	0	0	2
		gesloten		0	482	73	56	93	35	41	52	22	143
		gesloten	0	0	185	282	384	50	34	45	89	599	394
		gesloten	0	1	0	0	1	0	0	0	0	1	1
		gesloten	1		0	0	0	0	0	0	0	0	1
		gesloten	1	0	1.205	406	633	219	187	188	172	860	486
		gesloten	1	1	0	1	0	0	0	0	0	2	0
		halfopen		0	2	0	8	1	3	3	0	7	2
		halfopen	0	0	96	243	230	16	21	17	14	222	284
		halfopen	0	1	0	0	1	0	0	0	0	1	0
		halfopen	1		0	0	0	0	0	0	0	0	1
		halfopen	1	0	444	415	723	47	51	71	15	583	422
		halfopen	1	1	0	0	2	0	0	0	0	1	0
		halfopen	f	0	1	1	2	0	0	0	0	0	4
		open		0	2	2	4	1	0	1	0	3	2
		open	0		0	0	0	0	0	0	0	0	1
		open	0	0	124	280	320	22	27	15	26	361	451
		open	1	0	467	671	1.365	85	76	72	34	706	514
		open	1	1	0	1	0	0	0	0	0	1	0
		open	f	0	0	0	0	0	0	0	0	0	1
		open	f	f	0	1	0	0	0	0	0	0	0

Tabel 25 - detailtabel van de verdeling over de variabelen voor verkoop van appartementen (met telkens aantallen records)

					postcodecluster								
		b_bebouwing_id	o_tuinaanwezig	b_terenoveren	1	2	3	4	5	6	7	8	9
APPARTEMENT	Verkoop				0	0	0	0	0	0	0	13	0
				0	113	100	63	28	6	10	16	156	38
			0		0	274	0	0	1	1	0	458	2
			0	0	3.204	5.741	2.564	948	294	504	551	5.365	2.572
			0	1	5	2	4	1	0	0	1	6	4
			1		0	18	0	1	0	0	0	66	1
			1	0	165	431	277	44	14	16	39	182	144
			1	1	1	2	0	0	0	0	0	2	0
			f	0	18	203	39	1	0	1	2	73	26
			f	f	0	3	1	0	0	0	0	1	0
		gesloten			0	6	1	0	0	0	0	0	1
		gesloten		0	79	43	25	17	0	20	4	18	14
		gesloten	0	0	604	2.122	752	107	48	99	63	1.221	450
		gesloten	0	1	8	32	10	0	0	0	1	21	5
		gesloten	1		0	0	0	0	0	0	0	0	1
		gesloten	1	0	117	454	155	10	44	10	9	115	82
		gesloten	1	1	1	13	0	0	0	0	0	0	1
		halfopen			0	0	0	0	0	0	0	0	2
		halfopen		0	14	17	15	1	6	0	0	5	2
		halfopen	0	0	133	658	350	39	19	28	14	160	110
		halfopen	0	1	1	11	2	0	0	0	0	2	2
		halfopen	1	0	47	185	126	12	1	8	0	42	31
		halfopen	1	1	1	2	2	0	0	0	0	0	1
		halfopen	f	0	2	9	4	0	0	0	0	2	4
		open			0	0	0	0	0	0	0	1	0
		open		0	33	9	5	7	14	3	2	0	2
		open	0	0	167	413	187	75	113	44	35	254	132
		open	0	1	1	1	1	0	0	0	0	3	0
		open	1	0	23	65	72	13	13	7	8	25	16

Tabel 26 - detailtabel van de verdeling over de variabelen voor verhuur van appartementen (met telkens aantallen records)

					postcodecluster								
		b_bebouwing_id	o_tuinaanwezig	b_terenoveren	1	2	3	4	5	6	7	8	9
APPARTEMENT	Verhuur				0	0	0	0	0	0	0	1	0
				0	133	100	82	47	26	8	28	68	99
			0		32	7	94	1	0	0	0	413	11
			0	0	5.602	8.302	4.939	1.011	655	864	1.328	7.211	7.093
			0	1	0	0	0	0	0	0	0	1	0
			1		11	0	0	0	0	0	0	33	3
			1	0	346	544	267	22	38	14	39	202	300
			f	0	37	433	92	2	0	1	0	82	206
			f	f	0	30	4	0	0	0	0	2	12
		gesloten		0	3.409	427	333	353	177	125	211	112	564
		gesloten	0	0	1.605	1.829	1.367	212	197	148	174	1.602	1.126
		gesloten	0	1	0	1	0	0	0	0	0	1	0
		gesloten	1	0	930	303	361	139	135	103	92	247	250
		halfopen		0	22	13	17	3	6	5	0	6	4
		halfopen	0	0	290	637	443	55	43	59	8	270	242
		halfopen	0	1	0	1	1	0	0	0	0	0	0
		halfopen	1	0	57	119	119	5	15	20	0	65	57
		halfopen	f	0	8	29	13	0	0	0	0	8	16
		open		0	21	18	5	3	9	1	2	5	2
		open	0	0	311	590	285	79	96	63	49	323	378
		open	1	0	52	75	69	15	24	26	3	25	53
		open	1	f	0	2	0	0	0	0	0	0	0
		open	f	0	3	5	2	0	0	0	0	3	8

DEEL 3 - Exploratieve ruimtelijke analyse en opmaak van beleidsindicatoren

Het analysebestand dat in kader van deze studie wordt aangemaakt en onderzocht naar bruikbaarheid, kan naar de toekomst toe een belangrijk instrument zijn in de monitoring van de effectiviteit van bepaalde beleidsmaatregelen op het terrein. Hierbij is het voornamelijk relevant te onderzoeken welke indicatoren kunnen berekend worden op basis van de databank en welke ruimtelijke dataset(s) gekoppeld kan (kunnen) worden aan de Zimmo-databank in functie van zinvolle ruimtelijke analyse.

Tevens biedt een verkenning van mogelijkheden voor ruimtelijke analyse en koppeling met andere gegevens ook verder inzicht in de bruikbaarheid / betrouwbaarheid van de databank. Sommige beschikbare ruimtelijke datasets die eraan gelinkt kunnen worden, zijn daarnaast mogelijk ook waardevol als extra variabele of ter vervanging van bepaalde onvolledig ingevulde variabelen.

In deze eerste exploratie zijn een reeks indicatoren berekend en werden een aantal gegevens aan de databank gekoppeld[10]. Het hoofddoel van dit onderzoek is nagaan in welke mate deze gegevens geschikt zijn voor de berekening van beleidsindicatoren over de woningmarkt.

De woningmarkt wordt onderverdeeld in volgende deelmarkten (subsets van de databank):

- De koopmarkt: huizen, appartementen en bouwgronden;
- De huurmarkt: huizen en appartementen.

De beleidsindicatoren zijn:

- Snelheid van verkoop,
- Frictieleegstand,
- Vraagprijs en verkoopprijs, en het verschil ertussen,
- Schaarste-index.

In eerste instantie wordt de pre-processing van de data toegelicht, en de methode voor het aanmaken van de subsets. Op basis daarvan wordt de representativiteit van de records per subset geëvalueerd, ter onderbouwing van de selectie van beleidsindicatoren per subset.

Vervolgens wordt elke indicator in een apart hoofdstuk toegelicht, en wordt nagegaan in welke mate verschillende ruimtelijke aggregatieniveaus toelaten om de ruimtelijke verschillen in kaart te brengen.

[10] Uiteraard zijn er nog veel andere datasets die in kader van specifiek onderzoek in de toekomst gekoppeld zouden kunnen worden aan de Zimmo-databank (bv. overstromingsrisico, nabijheid van groen, …), gezien ze tot mogelijk interessante inzichten kunnen leiden bij koppeling met de vastgoeddata of omdat ze interessant kunnen zijn om surrogaatvariabelen te ontwikkelen voor variabelen die onvolledig zijn ingevuld in de immo-dataset. Hiervoor is eerst een verdere uitzuivering van de databank aangeraden (zie ook deel 4 – Conclusies).

1. Opmaak van subsets

1.1 Pre-processing

Het inlezen van het analysebestand in de geodatabase[11] gebeurt in 6 stappen.

Stap 1: Het analyse bestand (versie incl. de outliers) wordt als Access database ingelezen in ArcGIS. Dit levert 2 tabellen op. Eén met woningen (340 276 records) en één met gronden (19 050 records).

Stap 2: Voor gronden waren de coördinaten onvolledig, voornamelijk omdat wanneer er nog geen huis staat meestal nog geen huisnummer toegekend is. De immo_x_lamb en immo_y_lamb werden nl. uit de adresdatabanken van het CRAB/Urbis/PICC gehaald waarbij naar een match met de postcode/straatnaam/huisnr gezocht werd. Bij de gronden zonder coördinaten bleek inderdaad de inhoud van de veld a_nummer veelvuldig het woord "lot" te zijn. De ontbrekende coördinaten van gronden werden daarom berekend op basis van de originele immo velden a_geo_lat en a_geo_lon, tenzij ook deze velden nul of niet correct waren. Deze coördinaten werden aangeleverd in een aparte tabel (grondZonderCoordinaten_immo_XY_lam_input.xls).

Stap 3: De twee tabellen met woningen en gronden worden samengevoegd tot één grote tabel (359 326 records), met drie nieuwe velden:

- *immo_Pub1*_start_*year* — jaar van online plaatsen op basis van het veld *immo_Pub1_start*, omdat afgesproken werd dat een publicatie aan het jaar van online plaatsen toegekend wordt.
- *immo_status* — huur- vs. koopmarkt op basis van het veld *a_status*_id.
- *immo*_grondprijs_*m2* — de grondprijs per m², waar de gegevens van grondoppervlak beschikbaar zijn: in de eerste plaats in het veld immo_b_grondopp en anders in het veld immo_ParcelArea.

Stap 4: Selectie van een subset van bruikbare data. De gekozen voorwaarden zijn:

- Jaartallen 2010 – 2015 (veld *immo_Pub1_start_year*),
- Prijs > 0 (velden *immo_f_prijs* en *immo_grondprijs_m2*),
- Bruikbare coördinaten (*immo_X_lamb* en *immo_Y_lamb*), d.w.z. niet leeg, 0 of negatief,
- Huur- of koopmarkt (*immo_status*),
- Postcode moet aanwezig zijn (*immo_Postcode*),
- Bij de gronden enkel bouwgronden (*a_subtype_id* mag enkel 'bouwgrond', 'projectgrond' of 'woonuitbreidingsgebied' zijn),
- Bij de woningen werden 21 handelspanden weggelaten,
- De punten moeten in België liggen (een tweetal punten ligt buiten het grondgebied, ongeacht of het opgespoord werd via het adres of de coördinaten).

Er blijven na de selectie nog 247 887 records over.

Stap 5: Visualisatie van de data op basis van de velden *immo_X_lamb en immo_Y_lamb* met de tool *Display XY Data*, per deelmarkt per jaar.

De synthese is weergegeven in Tabel 27. Gezien het beperkte aantal gronden te huur werd beslist om de huurmarkt voor bouwgronden niet te weerhouden voor de analyses. Voor de andere subsets werden al de jaren weerhouden voor de ruimtelijke analyses.

[11] Arcmap 10.0 © ESRI

Tabel 27 - Aantal records per deelmarkt per jaar

	KOOPMARKT (164 120)			HUURMARKT (83 767)		
	Huizen	Appartementen	Gronden	Huizen	Appartementen	Gronden
2015	36 493	10 714	1 767	10 086	19 529	9
2014	30 264	6 686	1 597	7 983	13 252	16
2013	22 621	4 526	1 126	5 552	8 133	3
2012	18 744	3 604	675	4 520	6 077	2
2011	14 204	3 108	483	2 973	3 709	4
2010	5 709	1 516	283	802	1 117	1
TOTAAL	128 035	30 154	5 931	31 916	51 817	35

Na het weglaten van de huurmarkt aan bouwgronden, blijven er nog 247 852 records over.

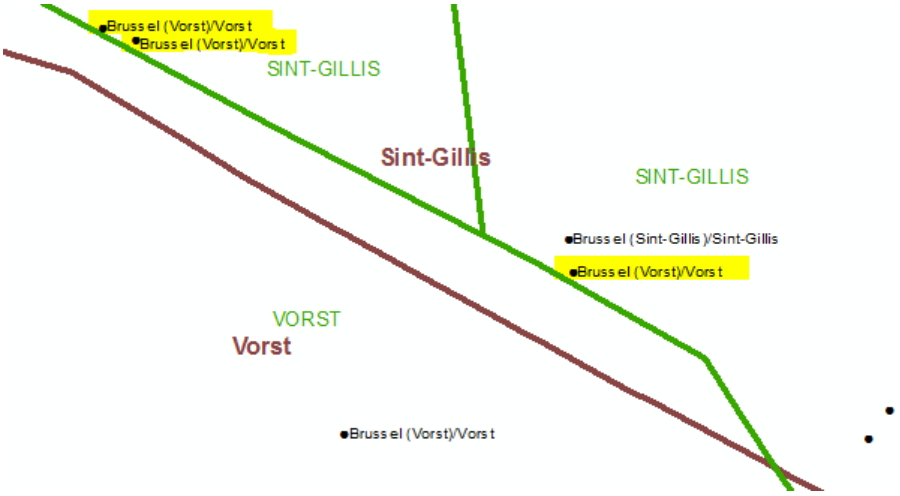
Stap 6: Uitbreiding van de databank met bijkomende gegevens.

De volgende bijkomende gegevens worden als attributen toegevoegd aan de records:

1. De statistische sectoren.
   Bij de spatial joins met statistische sectoren en gemeenten kwamen een aantal onnauwkeurigheden in de referentiebestanden aan het licht. De coördinaten uit de Zimmo-databank komen uit het CRAB (voor Vlaanderen), het PICC (voor Wallonië) en het Urbis (voor Brussel). De relatie met de statistische sectoren bestaat niet op basis van een veld, daarom werd een spatial join toegepast. Ter illustratie wordt naar onderstaand voorbeeld gekeken. De gemarkeerde punten hebben een adres in Vorst, maar zouden volgens de spatial join terecht komen in een statistische sector van Sint-Gillis. Dergelijke onnauwkeurigheden kunnen opgespoord worden bij gemeentegrenzen door vergelijking met het adres, maar niet bij statistische sectoren.

   Verder blijkt dat de grenzen van de statistische sectoren en van de gemeenten verschoven zijn ten opzichte van elkaar. Voor de statistieken op gemeenteniveau wordt daarom geen spatial join maar een aggregatie op basis van de postcodes gebruikt.

Figuur 10 – Inconsistenties bij de toekenning van statistische sector en gemeente aan een record door onnauwkeurigheden in CRAB, PICC en Urbis



Rood: gemeentegrenzen, groen: statistische sectoren, zwart – Urbis punten

Aangezien dit probleem enkel aan gemeentegrenzen gedetecteerd kan worden is het moeilijk exact te zeggen om hoeveel punten het gaat. Om de omvang van het probleem te

kwantificeren werd een test uitgevoerd op de variabele *vraagprijs*. Het gemiddelde van deze variabele werd berekend voor de jaartallen 2011 tot 2015 voor de gemeenten Gent en Antwerpen op twee verschillende manieren:

1. Op basis van de gemeenten in de Zimmo-databank (postcodes van de gemeenten en deelgemeenten: 2000, 2018, 2020, 2030, 2040, 2050, 2060, 2100, 2140, 2170, 2180, 2600, 2610, 2660, 9000, 9030, 9031, 9032, 9040, 9041, 9042, 9050, 9051, 9052).
2. Op basis van de spatial join met de statistische sectoren. In de tabel met statistische sectoren is er namelijk ook een gemeentenaam opgenomen.

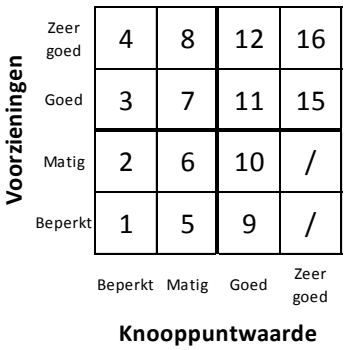
Uit onderstaande tabel blijkt dat het verschil niet zo groot is, waardoor het verder verwaarloosbaar geacht wordt in de analyses. Het wordt hier wel vermeld aangezien het een onnauwkeurigheid betreft in veel gebruikte overheidsdata.

Tabel 28 – Het effect van vastgestelde onnauwkeurigheden aan gemeentegrenzen op de vraagprijs

Zimmo stat. sector		KOOPMARKT			HUURMARKT	
		Huizen	Appartementen	Gronden	Huizen	Appartementen
GemiddeldE vraagprijs	Gent	296 663 296 713	259 658 259 658	281 691 281 691	769 769	653 653
	Antwerpen	303 402 303 395	267 078 267 145	528 415 528 415	1 226 1 226	1 746 1 746
Op basis van hoeveel punten	Gent	3 335 3 334	1 413 1 413	119 119	1216 1216	3 546 3 547
	Antwerpen	4 947 4 948	3 939 3 936	71 71	503 503	3 941 3 941

3. De knooppuntwaarde en voorzieningen, op basis van de kaart *synthesekaart_naturalbreaks_2015_metbus* (Verachtert et.aL, 2016[12]). Ter info: records met een hoge knooppuntwaarde en hoog voorzieningen niveau zijn gelegen in de gridcode 11, 12, 15 of 16 (Figuur 11).

Figuur 11 – Knooppuntwaarde en voorzieningen niveau



4. De verstedelijkingsgraad (op gemeenteniveau), op basis van de urbanisatie-indeling gebruikt door de studiedienst van de Vlaamse Regering. Dit is een indeling in 6 categorieën, die beschikbaar is voor Vlaanderen en Brussel, in totaal 327 gemeenten (Figuur 12).

[12] Verachtert, E., I. Mayeres, L. Poelmans, M. Van der Meulen, M. Vanhulsel, G. Engelen (2016), Ontwikkelingskansen op basis van knooppuntwaarde en nabijheid voor¬zieningen, eindrapport, studie uitgevoerd in opdracht van Ruimte Vlaanderen.

Figuur 12 – Urbanisatie-indeling Vlaanderen en Brussel

urbanisatie

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1) grootsteden	21	6,4	6,4	6,4
	2) centrumsteden	11	3,4	3,4	9,8
	3) stedelijke rand	53	16,2	16,2	26,0
	4) kleinere steden	44	13,5	13,5	39,4
	5) overgangsgebied	97	29,7	29,7	69,1
	6) platteland	101	30,9	30,9	100,0
	Total	327	100,0	100,0	

In de exploratieve ruimtelijke analyse (deel 3) verwijst de term Zimmo-databank naar deze dataset in de geodatabase.

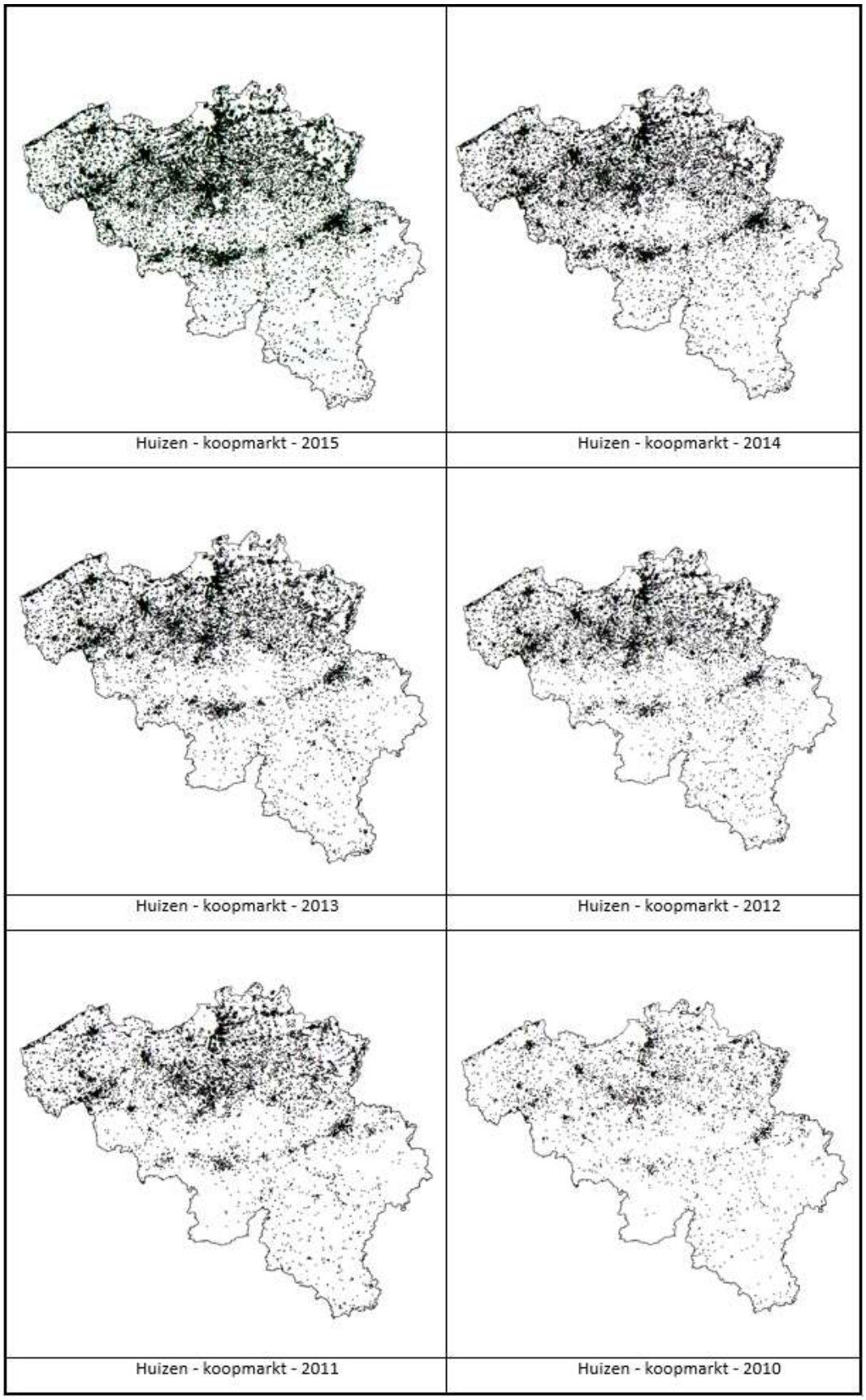
1.2 Koopmarkt

Uit Figuur 13 blijkt dat er in gans België een toename is van het aantal records in de subset 'Koopmarkt van huizen'. Vlaanderen en Brussel zijn al sinds 2010 goed vertegenwoordigd. In 2011 tot 2013 zijn er ook veel huizen te koop gepubliceerd in de Waalse steden, meer bepaald de as Bergen-Charleroi-Namen-Luik. Geleidelijk zijn ook Waals-Brabant en de meer verstedelijkte delen van de provincies Luik, Namen en Henegouwen, goed vertegenwoordigd. Een eerste visuele verkenning van de spreiding geeft voldoende indicaties van mogelijke ruimtelijke patronen om verder te gaan met een ruimtelijke analyse van deze subset per jaar.

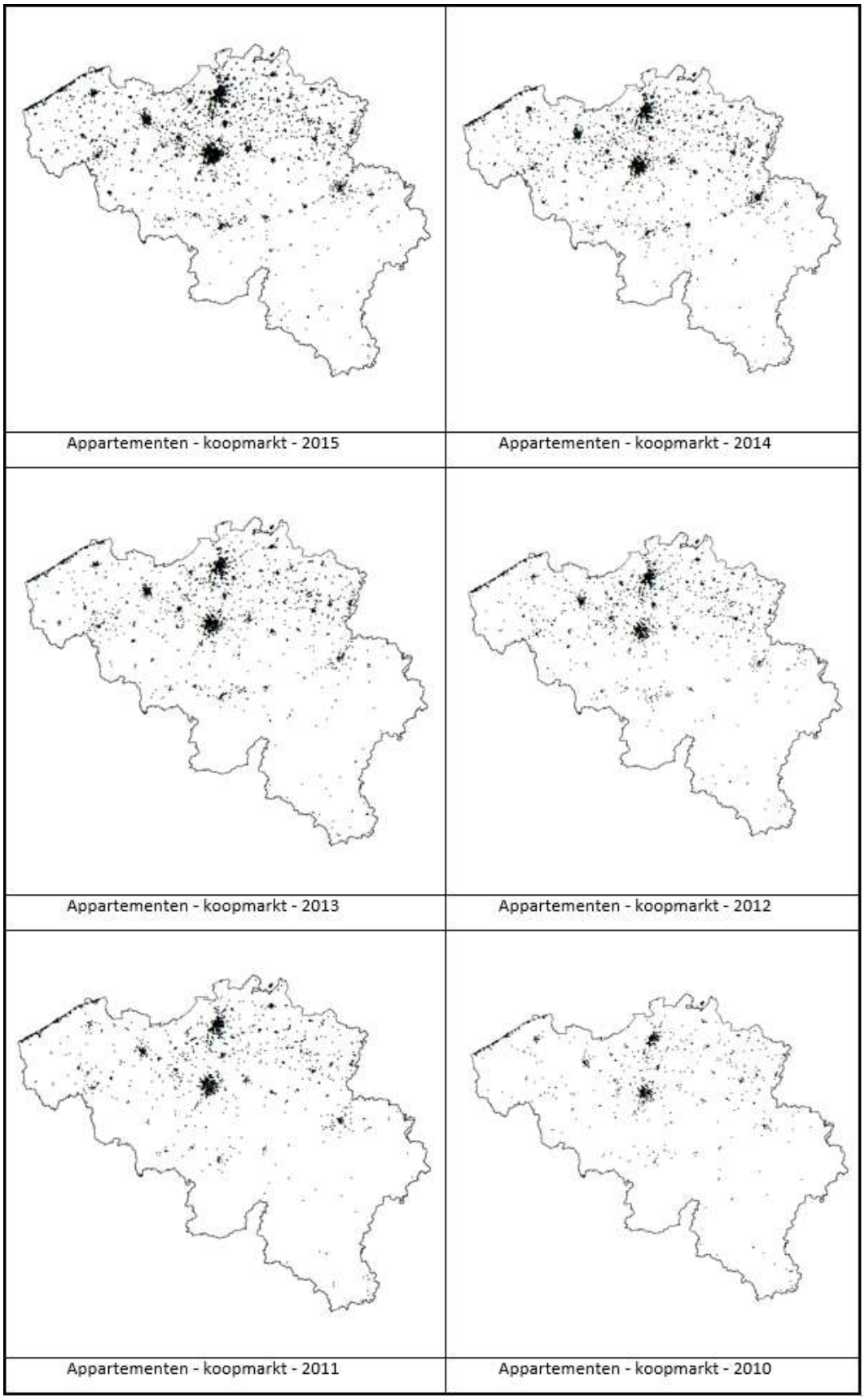
Bij de verkoop van appartementen (Figuur 14) blijft de databank tot 2012 vooral beperkt tot Brussel, Antwerpen en de kust. Het is wachten tot 2013 vooraleer ook kleinere steden in Vlaanderen in beeld komen. In 2015 is de verkoop van appartementen in Vlaanderen niet meer beperkt tot steden en zijn aanzienlijk meer verkopen (zie ook Tabel 27). Deze evolutie in de databank verklaart waarom het niet zinvol was om elke beleidsindicator per jaar te berekenen voor deze subset. Zo werd ervoor gekozen om voor de analyse van de prijzen per statistische sector de jaren 2011-2015 te sommeren. Per beleidsindicator wordt in het rapport de duidelijkste kaart besproken (meestal 2015), tenzij een vergelijking of sommering met voorgaande jaren zinvol is.[13]

[13] In de Geodatabase worden al de aangemaakte kaarten van beleidsindicatoren opgeleverd.

Figuur 13 – Evolutie van de subset koopmarkt - huizen



Figuur 14 – Evolutie van de subset koopmarkt - appartementen



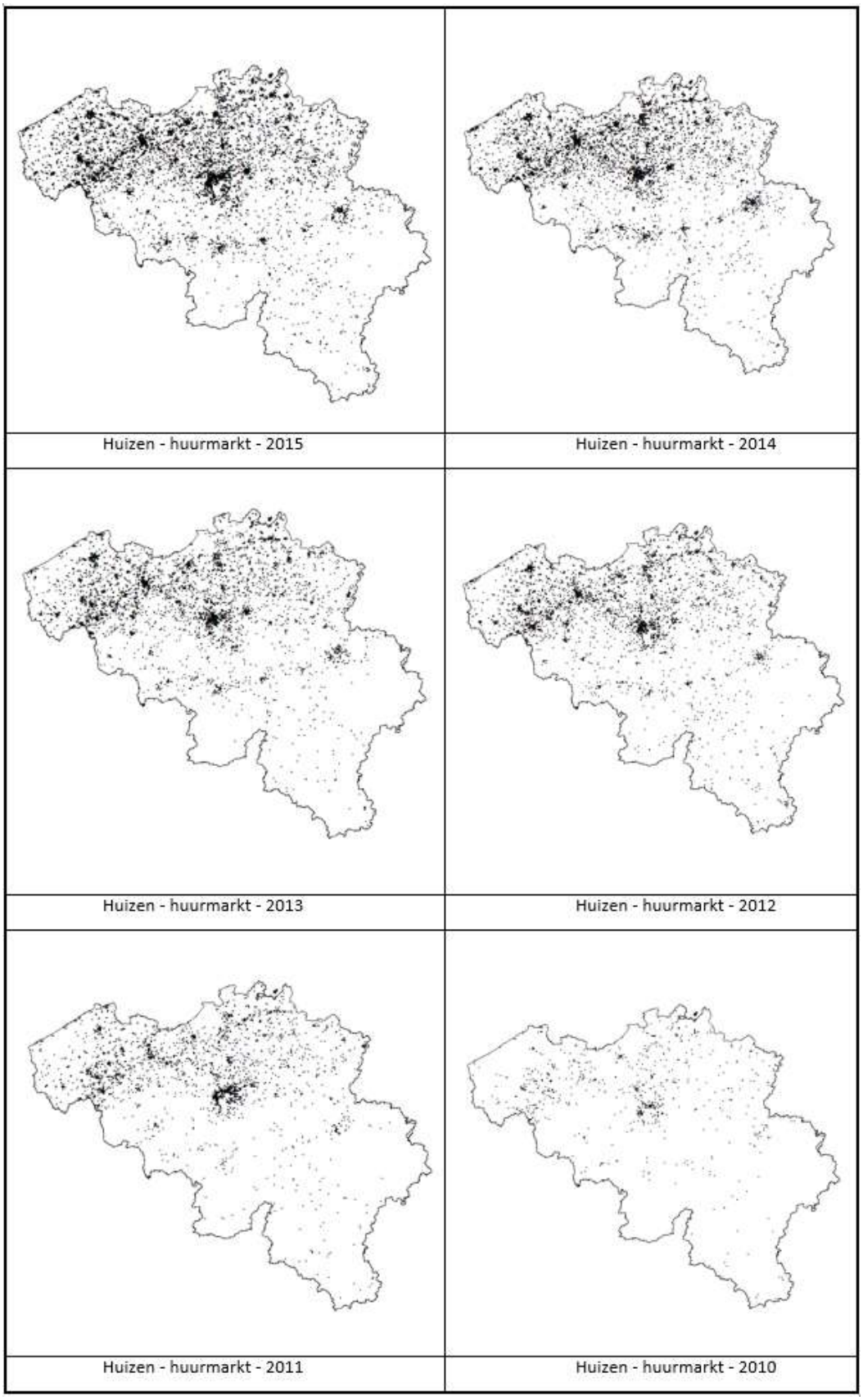
1.3 Huurmarkt

Het aantal records over verhuur van huizen neemt gradueel toe, en bereikt pas in 2015 meer dan 10.000 records (Tabel 27). Er zijn meer records in Vlaanderen en Brussel dan in Wallonië, en vanaf 2012 zijn concentraties merkbaar (

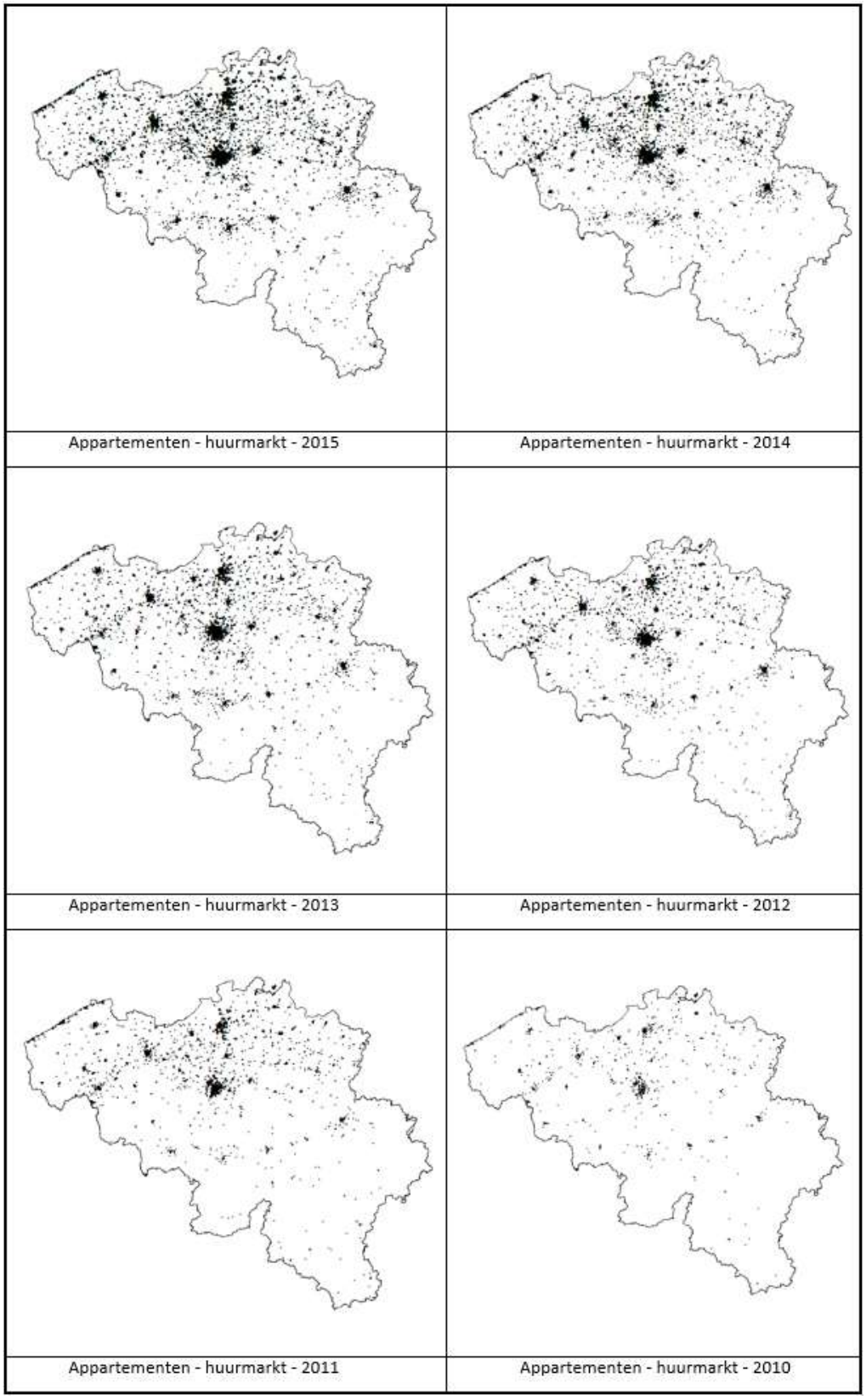
Figuur 15). De hotspots werden voor gans België berekend voor al de beschikbare jaren, maar de bespreking van de ruimtelijke spreiding in het rapport gebeurt, gezien het hoge aantal punten, voor de toestand in 2015.

Voor de verhuur van appartementen zijn meer records aanwezig dan voor de verhuur van huizen, en het aantal neemt sterker toe (Tabel 27). Ook hier bevat de geodatabase hotspots analyses van verschillende jaren, en beperkt het rapport zich tot de interpretatie van 2015.

Figuur 15 – Evolutie van de subset huurmarkt - huizen



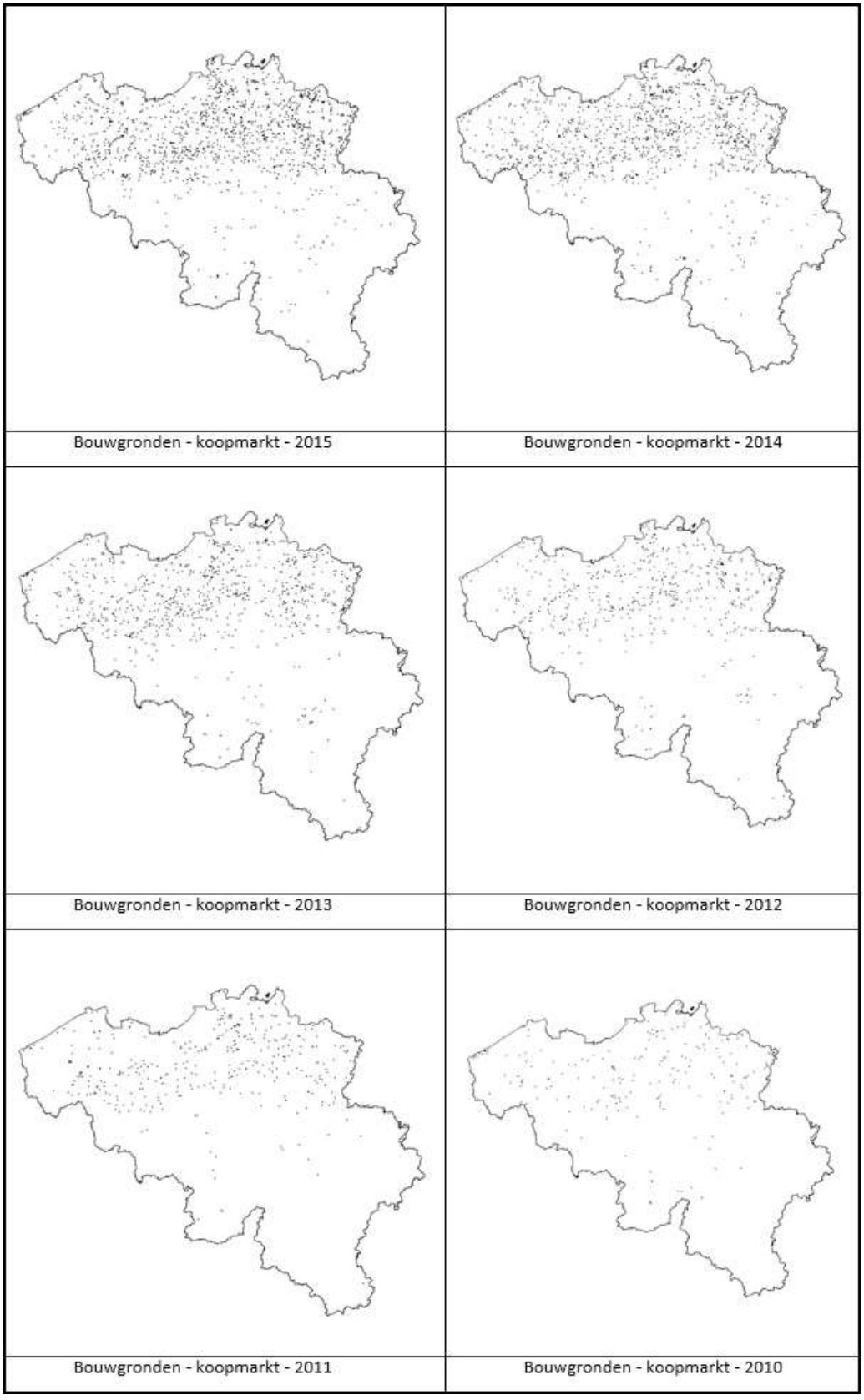
Figuur 16 – Evolutie van de subset huurmarkt - appartementen



1.4 Markt van de bouwgronden

Het aantal te koop aangeboden bouwgronden en hun spreiding in de Zimmo-databank is te beperkt om ruimtelijke patronen in de spreiding van punten te identificeren (Figuur 17). Met een totaal van 5 931 records van bouwgronden te koop (Tabel 27) is een kartering op statistische sector niveau (totaal = 19 781) ook niet zinvol. De ruimtelijke analyses van de bouwgronden zijn daarom geaggregeerd tot het niveau van de gemeenten. Gezien de spreiding van de locaties zijn ook de resultaten van de concentratie berekeningen aan de hand van hotspot analyses voor bouwgronden, niet mee besproken in het rapport.

Figuur 17 – Evolutie van de subset koopmarkt – bouwgronden



1.5 Representativiteit van de subsets

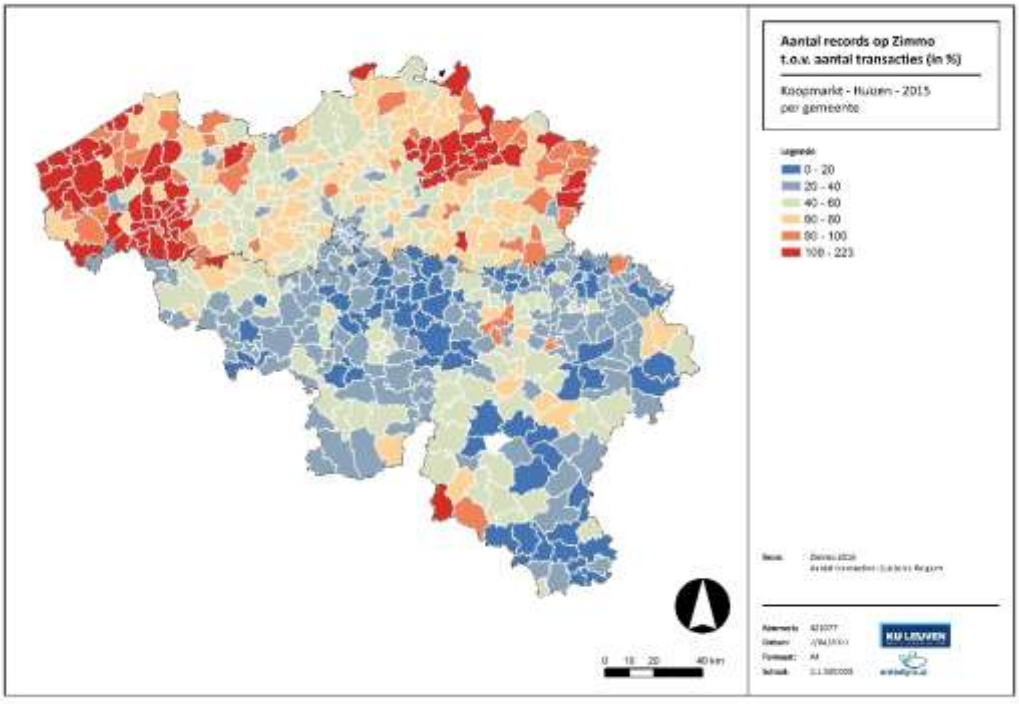
Om de kwaliteit van de Zimmo-databank te testen, wordt het aantal woningen/gronden dat te koop staat op de immodatabank vergeleken met het aantal uitgevoerde transacties volgens de gegevens van Statistics Belgium.

De volgende 3 relevante gegevens werden toegevoegd of berekend:

- *aantal_vraag* het aantal woningen/gronden dat te koop werd aangeboden op Zimmo
- *aantal_verkoop* het aantal woningen/gronden dat verkocht werd volgens Statistics Belgium
- *percent_opZimmo* procentuele vergelijking van beide (aantal_vraag/ aantal_verkoop)

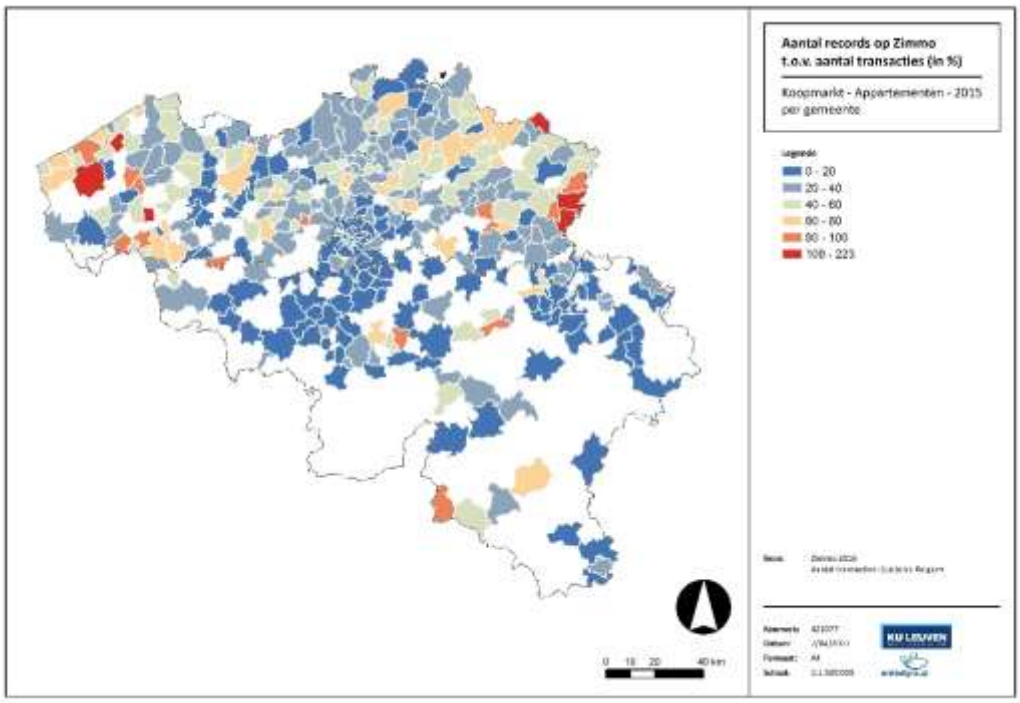
De legende toont de procentuele verschillen in intervallen van 20 punten.

Figuur 18 – Vergelijking van het aantal huizen te koop in 2015 in de Zimmo-databank vergeleken met het aantal transacties per gemeente van Statistics Belgium



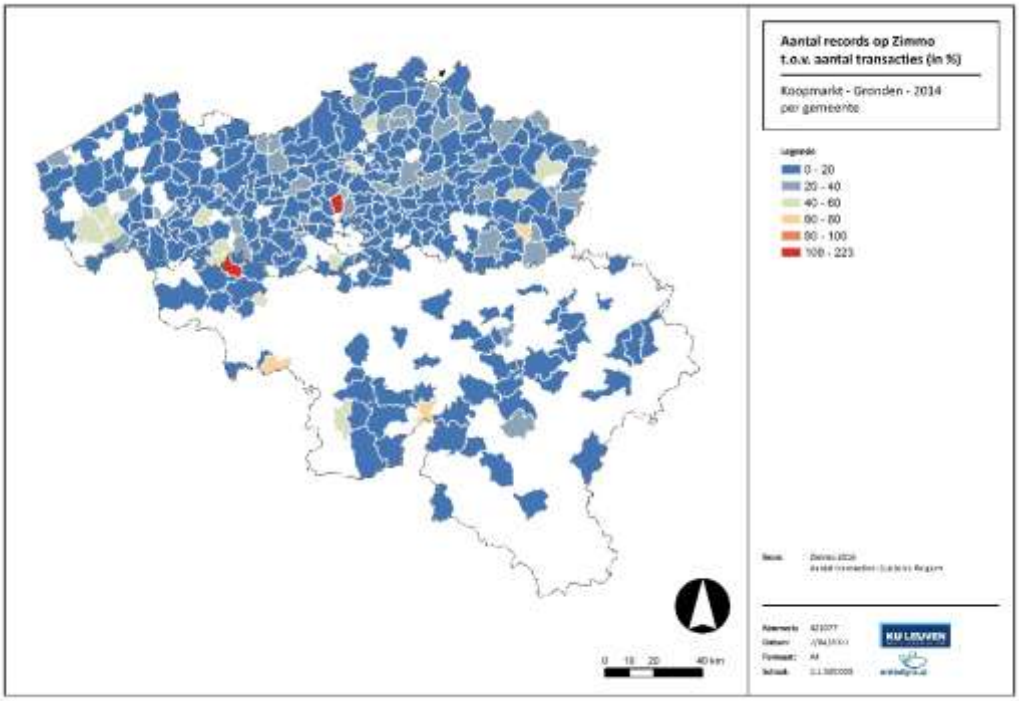
Vooral in Vlaanderen is het aantal records in de Zimmo-databank vergeleken met het aantal verkooptransacties geregistreerd door de FODFIN in 2015, zeer hoog; op enkele gemeenten na meer dan 40%. Merkwaardig is ook het aantal gemeenten waar aanzienlijk meer publicaties van huizen te koop zijn geboden op Zimmo, dan er werkelijk zijn verkocht. In enkele kustgemeenten en industriegemeenten aan het Albertkanaal, is dat meer dan het dubbele. Dat is veelbelovend wat betreft de representativiteit van het aantal records in de Zimmo-databank. Hoewel het logisch is dat er meer huizen te koop worden aangeboden dan er werkelijk verkocht worden, is de reden een interessant onderzoeksonderwerp. In welke mate worden immo sites gebruikt als test om een idee te krijgen van de waarde van een woning, zonder intentie van verkoop, bijvoorbeeld bij verdeling van erfenis, donaties, zoeken naar verzekeringswaarde, *enz.*? Hoe verschillend is dit bij professionele vs. niet professionele publicaties?

Figuur 19 – Vergelijking van het aantal appartementen te koop in 2015 in de Zimmo-databank vergeleken met het aantal transacties per gemeente van Statistics Belgium



Voor de appartementen is de vergelijking moeilijker: in veel gemeenten werden in 2015 geen appartementen verkocht. In landelijke gemeenten met weinig appartementen kunnen enkele publicaties de verhoudingen sterk beïnvloeden. Anderzijds kan dit een indicatie zijn dat de verkoop van individuele appartementen minder online gebeurt. Hier is een verdere analyse van de Zimmo-databank aangewezen: in welke mate worden appartementen eerder gepubliceerd in groep dan individueel, en hoe kunnen individuele appartementen 'te koop' uit de databank worden afgeleid als ze per appartementsgebouw worden gepubliceerd?

Figuur 20 – Vergelijking van het aantal bouwgronden te koop in 2014 in de Zimmo-databank vergeleken met het aantal transacties per gemeente van Statistics Belgium



Slechts een zeer beperkt aandeel van de verkochte bouwgronden werd gepubliceerd op Zimmo. Bij de interpretatie van beleidsindicatoren berekend voor de koopmarkt van bouwgronden dient hiermee rekening te worden gehouden.

2. Beleidsindicator “snelheid van verkoop”

2.1 Berekeningsmethode

Het veld *immo_Pub1_dagen* wordt gebruikt als schatting voor de tijd dat een pand of grond online staat. Dit veld is het verschil tussen *immo_Pub1_start en immo_Pub1_stop*. Dit is een proxy, het geeft een indicatie van de dagen dat een pand of grond online stond en niet of het binnen die tijd al dan niet verkocht is geraakt.

Records waar geen aantal dagen online berekend kunnen worden, of records met een negatief aantal dagen online worden niet weerhouden. Het gaat over volgende aantallen die overblijven (224 466 records in het totaal):

Tabel 29 – Aantal records voor de berekening van de ‘tijd online’, proxy voor snelheid van verkoop

	KOOPMARKT (144 881)			HUURMARKT (79 585)	
	Huizen	Appartementen	Gronden	Huizen	Appartementen
2015	28 746	7 520	1 067	9 320	17 950
2014	27 458	5 807	1 268	7 681	12 763
2013	21 534	4 201	959	5 343	7 815
2012	17 993	3 446	583	4 365	5 900
2011	13 579	2 999	444	2 915	3 647
2010	5 533	1 482	262	782	1 104
TOTAAL	114 843	25 455	4 583	30 406	49 179

Spreidingspatroon

De basisdata (puntgegevens) worden getoond op 30 kaarten (5 deelmarkten en 6 jaartallen):

- Koopmarkt: Huizen, Appartementen, Gronden
- Huurmarkt: Huizen, Appartementen
- Jaartallen 2010 – 2015

Alle kaarten zijn gebaseerd op de Zimmo-databank feature class, maar met een andere Definition Query. De legende is gebaseerd op alle features in de koopmarkt enerzijds en huurmarkt anderzijds, met klassen gebaseerd op natural breaks, om vergelijkingen doorheen de jaren mogelijk te maken.

Aggregaties op niveau van statistische sectoren en gemeenten

Omdat er te weinig statistische sectoren zijn waar voldoende records zijn om een gemiddelde te berekenen, werd beslist om een selectie te maken van statistische sectoren met meer dan 30 records. Het resultaat (70 statistische sectoren) was onvoldoende om een zinvolle statistiek te produceren, zeker voor elk individueel jaar. Daarom werden 2 cases geselecteerd: Gent (met een totaal van 201 statistische sectoren) en Antwerpen (totaal 298 statistische sectoren). Om aan 30 punten per statistische sectoren te geraken werden de 5 jaartallen 2011 tot 2015 geaggregeerd.

In ArcGIS werd de tool *Summary Statistics* gebruikt. In het totaal zouden er 2495 records kunnen zjin (499 statistische sectoren en 5 deelmarkten). Uiteindelijk waren er 1 148 met data en slechts 195 waar de statistiek op basis van minstens 30 punten kon berekend worden. De onderstaande tabel geeft een opdeling van deze 195 statistieken per deelmarkt en per stad:

Tabel 30 – Aantal records voor de berekening van de gemiddelde 'tijd online' per statistische sector, in Antwerpen en Gent

Header	KOOPMARKT			HUURMARKT	
	Huizen	Appartementen	Gronden	Huizen	Appartementen
Gent	37	10	0	5	35
Antwerpen	40	29	0	0	39

Op basis van deze test werd besloten dat statistieken op niveau van statistische sectoren weinig zin hebben.

Samenvatting per gemeente

De samenvatting per gemeente gebeurt wel voor ieder individueel jaartal.

Opnieuw werd de tool *Summary Statistics* gebruikt. In het totaal zijn er 17 670 records mogelijk (589 gemeenten, 6 jaartallen en 5 deelmarkten). Uiteindelijk zijn er 12 112 met data en slechts 1 883 waar de statistiek op basis van minstens 30 punten berekend kon worden. De onderstaande tabel geeft een opdeling van deze 1 883 statistieken per deelmarkt en per jaar:

Tabel 31 – Aantal records voor de berekening van de gemiddelde tijd online' per gemeente

	KOOPMARKT (1 331)			HUURMARKT (552)	
	Huizen	Appartementen	Gronden	Huizen	Appartementen
2015	300	51	0	80	133
2014	277	39	1	61	80
2013	228	29	1	37	54
2012	186	21	1	28	41
2011	134	21	0	15	21
2010	34	7	1	0	2
TOTAAL	1 159	168	4	221	331

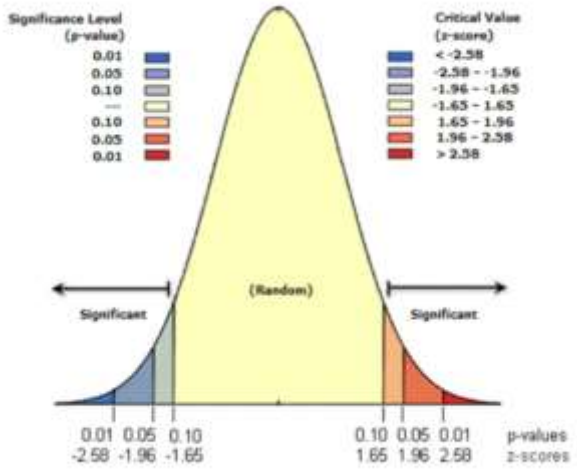
Het is duidelijk dat ook hier de aggregatie weinig zinvol is: het grote aantal zeer korte publicaties en de interne variatie binnen de gemeenten heeft een sterk afvlakkend effect op deze statistiek.

Hotspot analyse

We berekenen een hotspot analyse van dezelfde 30 puntlagen die hierboven besproken staan. In een hotspot analyse wordt de ruimtelijke correlatie gebruikt om concentraties aan te geven. Ruimtelijke correlatie betekent: er is een verband tussen de nabijheid en de gemeten waarde. Punten waar deze relatie significant hoog (nabijgelegen punten lijken meer op elkaar dan wat kan verwacht worden op basis van de gemiddelde waarden) of laag (nabijgelegen punten verschillen meer van elkaar dan wat kan verwacht worden op basis van de gemiddelde waarden) is, behoren tot een hot(cold) spot (

Figuur 21).

Figuur 21 – Berekening van significante concentraties in een hotspot analyse



De **legende** wordt automatisch bepaald door de ArcGIS *Hot Spot Analysis* tool. Echter deze kent automatisch een rode kleur (*hot spot)* toe aan hoge getallen en een blauwe kleur (*cold spot*) aan lage getallen. In dit geval is dit echter niet gewenst. Hoge getallen betekent nu 'lang online' en dus 'moeilijk verkoopbaar'. In de legende werden daarom de kleuren omgekeerd.

2.2 Koopmarkt

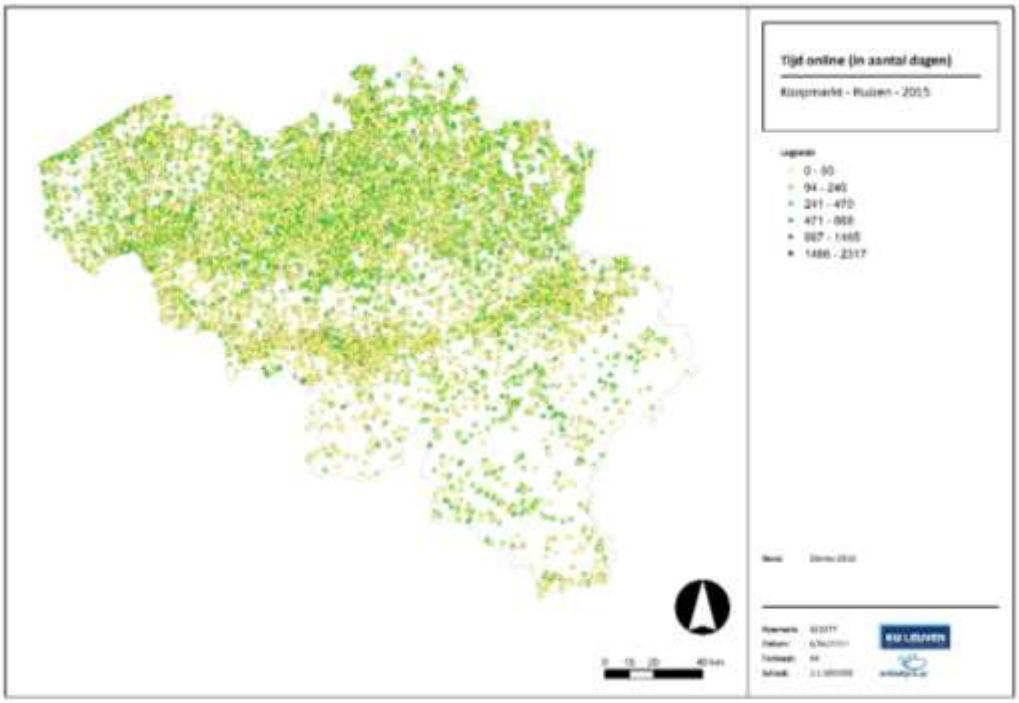
2.2.1 Huizen

Spreidingspatroon

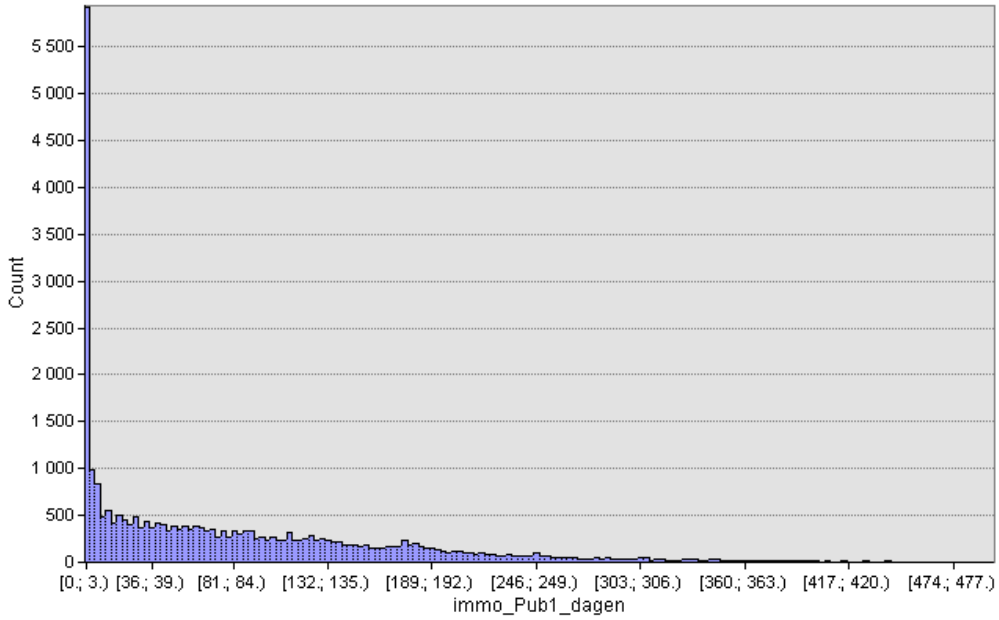
De stippenkaart van de snelheid van verkoop voor woningen (Figuur 22) illustreert dat er geen duidelijke patronen in beeld komen. De verklaring is het zeer grote aantal publicaties die slechts zeer kort (1-2 dagen) online staan (

Figuur 23). Mits aanpassing van de legende voor kortere perioden verschijnen er meer nuances (Figuur 24). Dat motiveert een verdere verkenning van mogelijke patronen aan de hand van hotspot analyse.

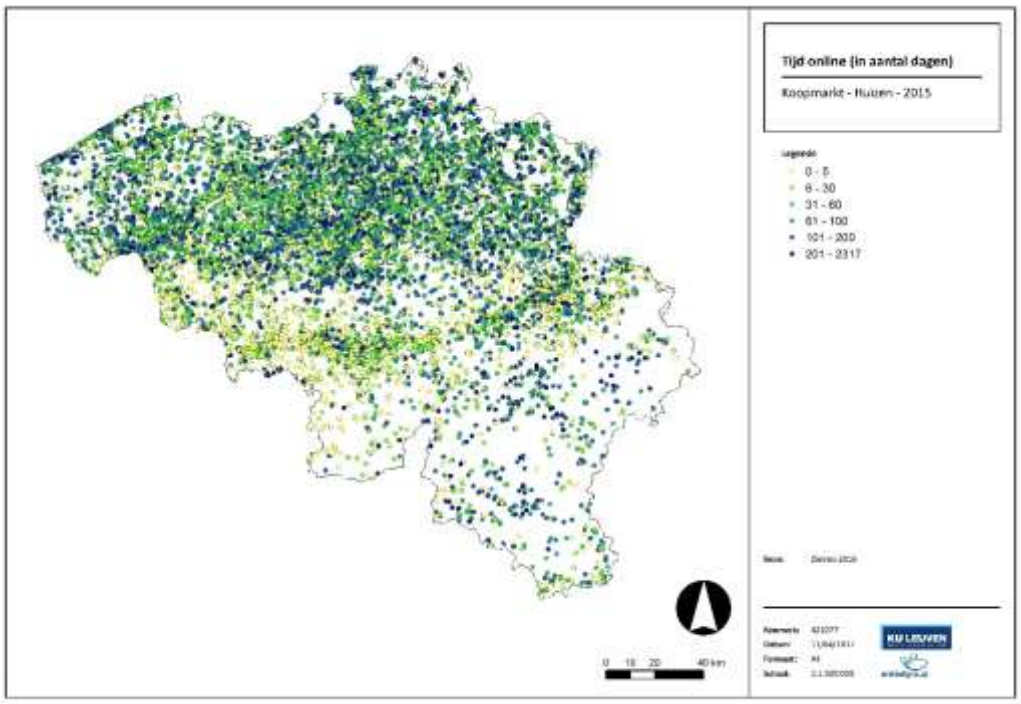
Figuur 22 – Stippenkaart 'snelheid van verkoop', huizen, 2015



Figuur 23 – Histogram 'snelheid van verkoop', huizen, 2015

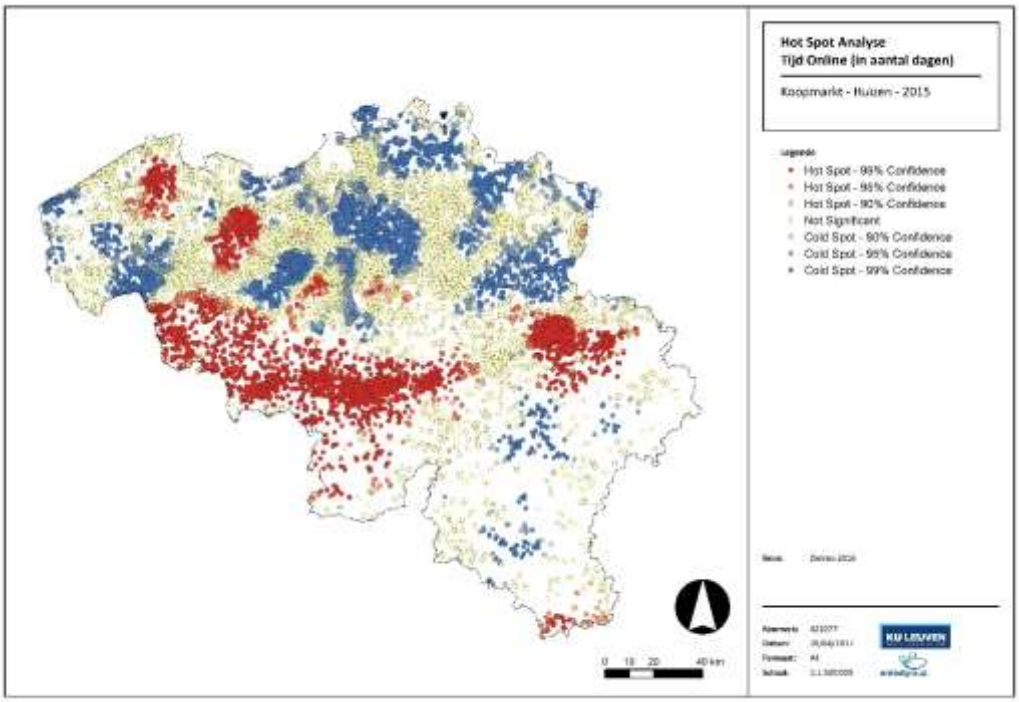


Figuur 24 – Stippenkaart 'snelheid van verkoop' (aanpassing legende voor kortere perioden online), huizen, 2015



Hotspot analyse

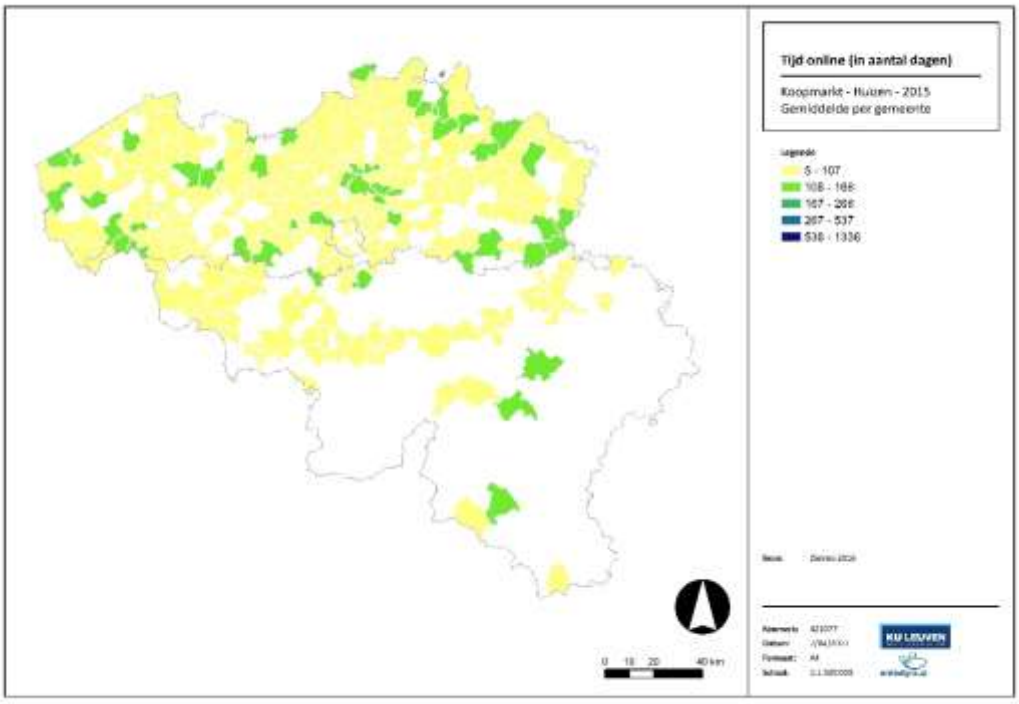
Figuur 25 – Hotspot analyse 'snelheid van verkoop', huizen, 2015



De hotspot analyse geeft duidelijk wel een aantal patronen weer: locaties dicht bij elkaar gelegen worden in bepaalde omgevingen statistisch significant sneller of trager verkocht dan verder afgelegen locaties. In Vlaanderen komen Gent en Brugge sterk in beeld als steden waar huizen snel verkocht zijn. Coldspots zijn er in het gebied Antwerpen-Mechelen-Brussel, tussen Brussel en Gent, de Kempen, delen van Limburg en van de kust en de Westhoek. Opvallend is ook de hotspot die zich uitstrekt van Namen over Henegouwen tot het Zuiden van Oost-Vlaanderen. Daarnaast zijn Luik en de Zuidgrens met Luxemburg hotspots.

Aggregaties op niveau van statistische sectoren en gemeenten

Figuur 26 – Gemiddelde 'snelheid van verkoop' per gemeenten, huizen, 2015



Het grote aantal kortstondige publicaties heeft een nivellerend effect op de gemiddelden per gemeente, waardoor de patronen minder duidelijk zijn dan bij de hotspot analyse.

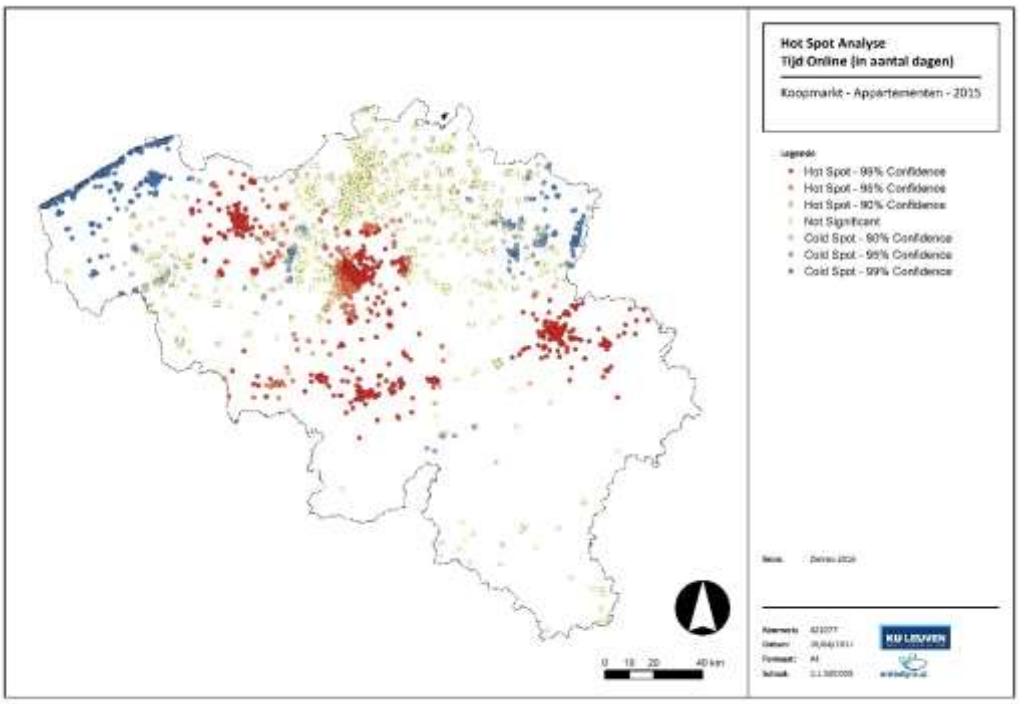
2.2.2 Appartementen

Spreidingspatroon

*Ook bij de appartementen (*Figuur 27*) geeft de stippenkaart met natural breaks klassen in de legende, vooral weer dat er zeer veel appartementen voor korte en zeer korte duur online worden gepubliceerd. Om patronen te herkennen is ook hier gewerkt een hotspot analyse beter geschikt (*

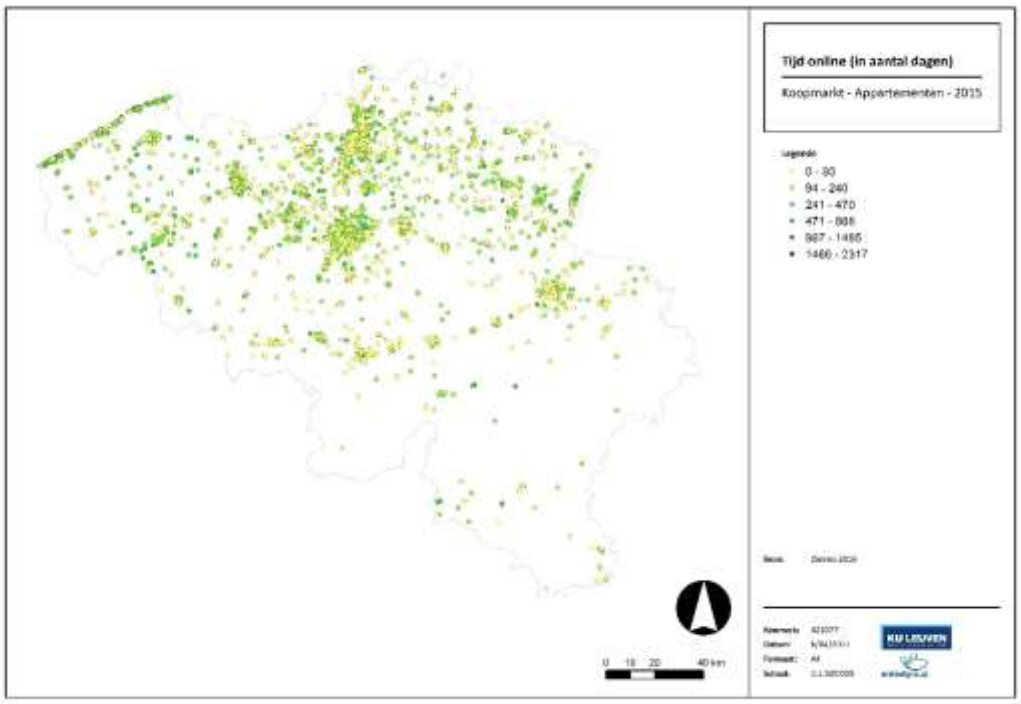
Hotspot *analyse*

Figuur 29 – Hotspot analyse 'snelheid van verkoop' appartementen, 2015

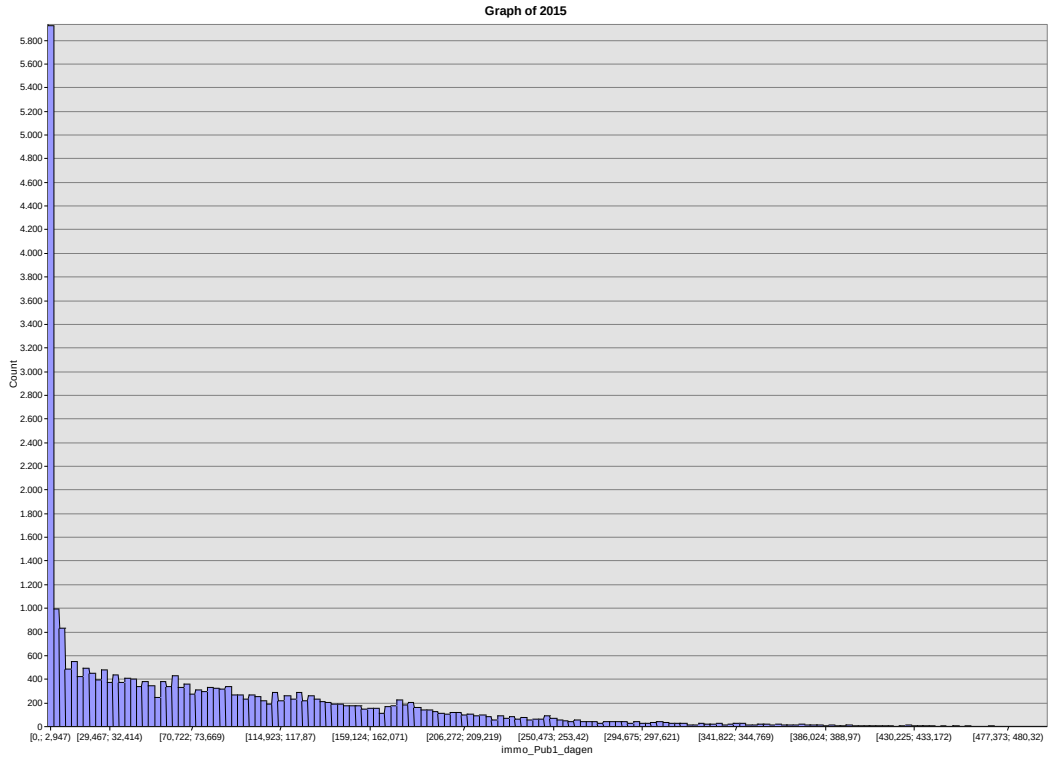


). Daar vallen vooral een aantal steden met hun periferie op Brussel en Gent in Vlaanderen, en Luik in Wallonië. Ook in het ganse gebied tussen Brussel, Charleroi en Namen blijven de appartementen te koop niet lang online. Merkwaardig zijn ook de coldspots, m.a.w. concentraties waar appartementen wel relatief lang online blijven: dat is het geval aan de kust, Brugge, Leiedal en de Westhoek (daar zijn het weliswaar weinig en sterk verspreide appartementen). Ook aan de Maaskant en de omgeving Hasselt-Genk-Bree in Limburg blijven appartementen langer online, en in het gebied Aalst-Ninove.

Figuur 27 – Stippenkaart 'snelheid van verkoop', appartementen, 2015

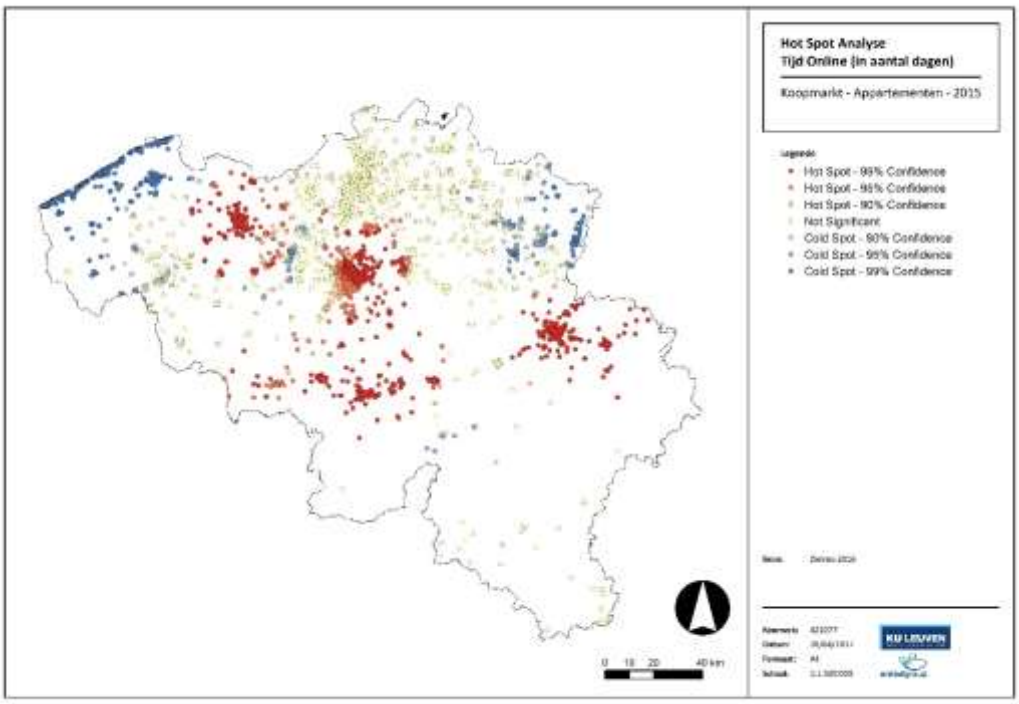


Figuur 28 – Histogram 'snelheid van verkoop', appartementen, 2015



Hotspot analyse

Figuur 29 – Hotspot analyse 'snelheid van verkoop' appartementen, 2015



Aggregaties op niveau van statistische sectoren en gemeenten

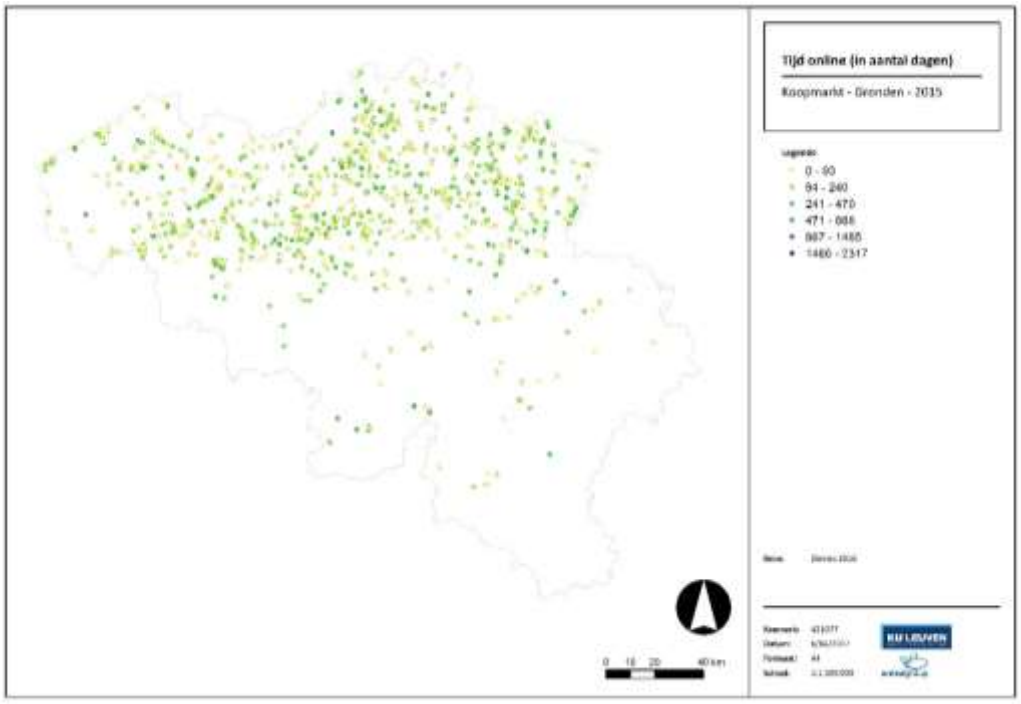
Omdat het aantal verkochte appartementen per gemeente te laag ligt, wordt de kaart van België niet verder besproken.

2.2.3 Bouwgronden

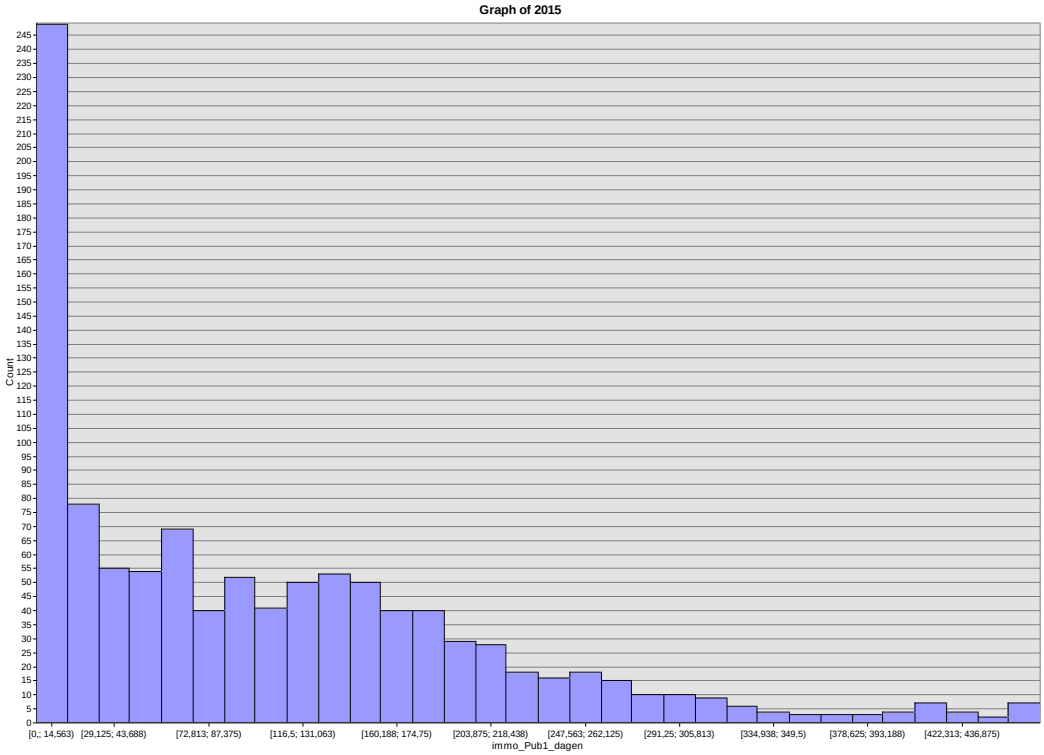
Spreidingspatroon

Ook bij de bouwgronden zijn er veel publicaties van minder dan 3 dagen, maar vergeleken met de huizen en appartementen, zijn er relatief meer bouwgronden die lang online blijven. Opvallend is de hotspot in het Korstrijkse en het Leiedal, en ook bij de bouwgronden de colspot in omgeving Hasselt-Genk-Bree in Limburg. In mindere mate blijven ook in het Zuidoosten van Oost-Vlaanderen bouwgronden langer on-line (weliswaar gaat het om een beperkt aantal records)

Figuur 30 – Stippenkaart 'snelheid van verkoop', bouwgronden, 2015

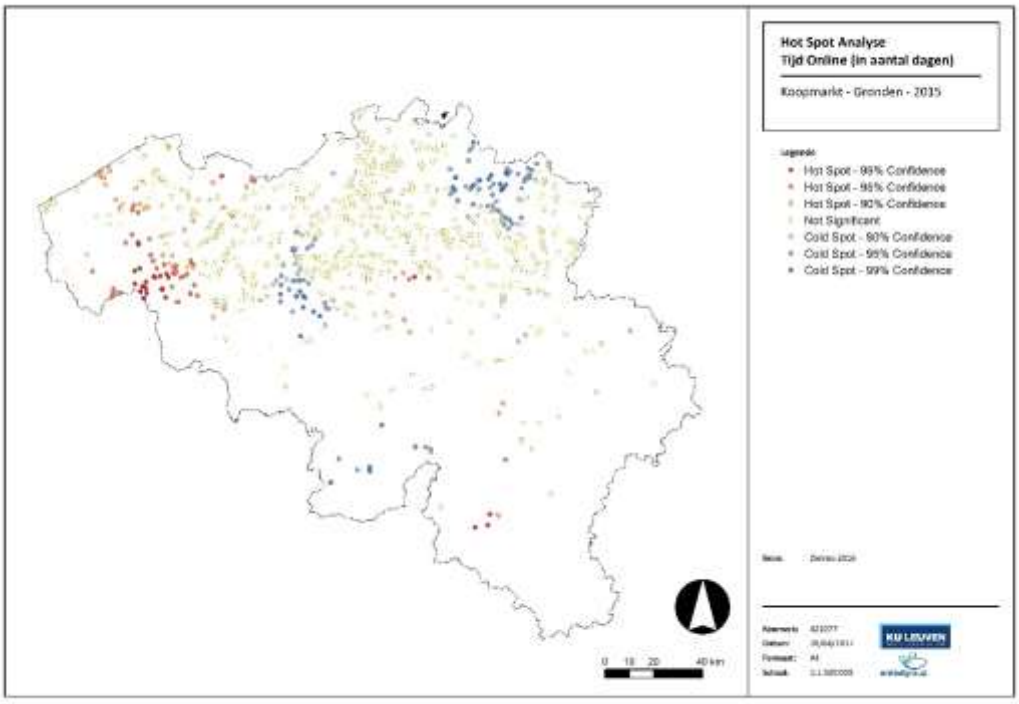


Figuur 31 – Histogram 'snelheid van verkoop', bouwgronden, 2015



Hotspot analyse

Figuur 32 – Hotspot analyse 'snelheid van verkoop' bouwgronden, 2015



Aggregaties op niveau van statistische sectoren en gemeenten

Zoals bij de appartementen, heeft een weergave op niveau van statische sectoren (of gemeente) weinig zin, omdat er te weinig bouwgronden worden verkocht om ruimtelijke variatie na aggregatie in beeld te brengen.

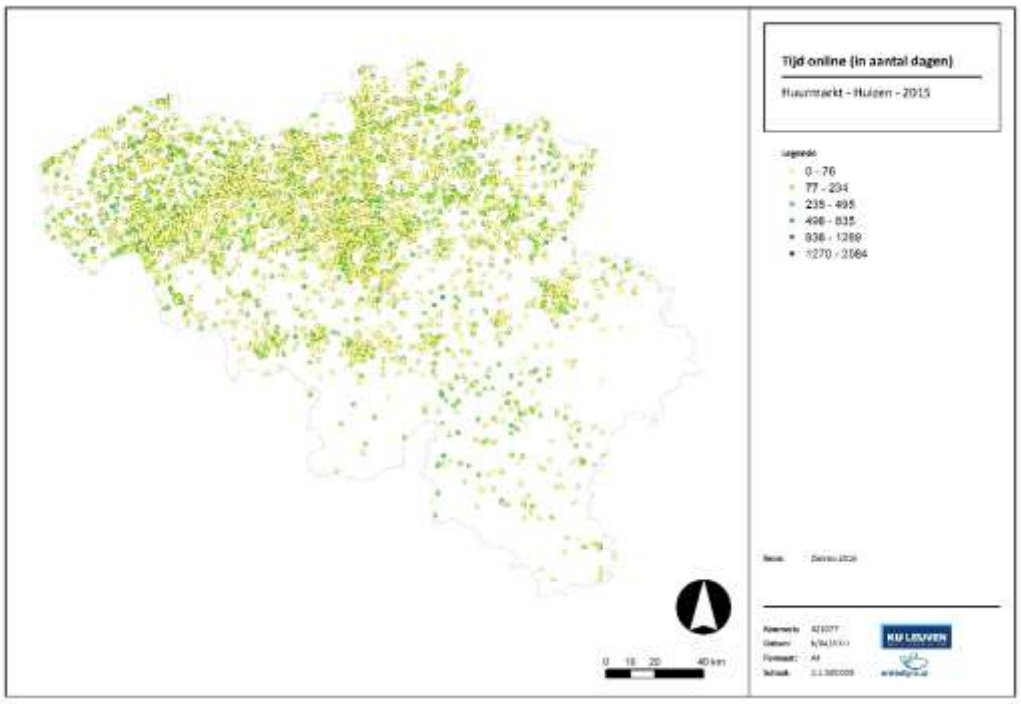
2.3 Huurmarkt

Zoals aangegeven in onderstaande figuren, stellen we ook bij de huurmarkt vast dat de microdata op puntniveau toelaten om ruimtelijke concentraties vast te stellen aan de hand van hostpot analyse. Voor aggregaties op statistische sector niveau zijn er te weinig records per sector. Aggregaties op gemeenteniveau hebben een nivellerend effect waardoor de patronen verdwijnen.

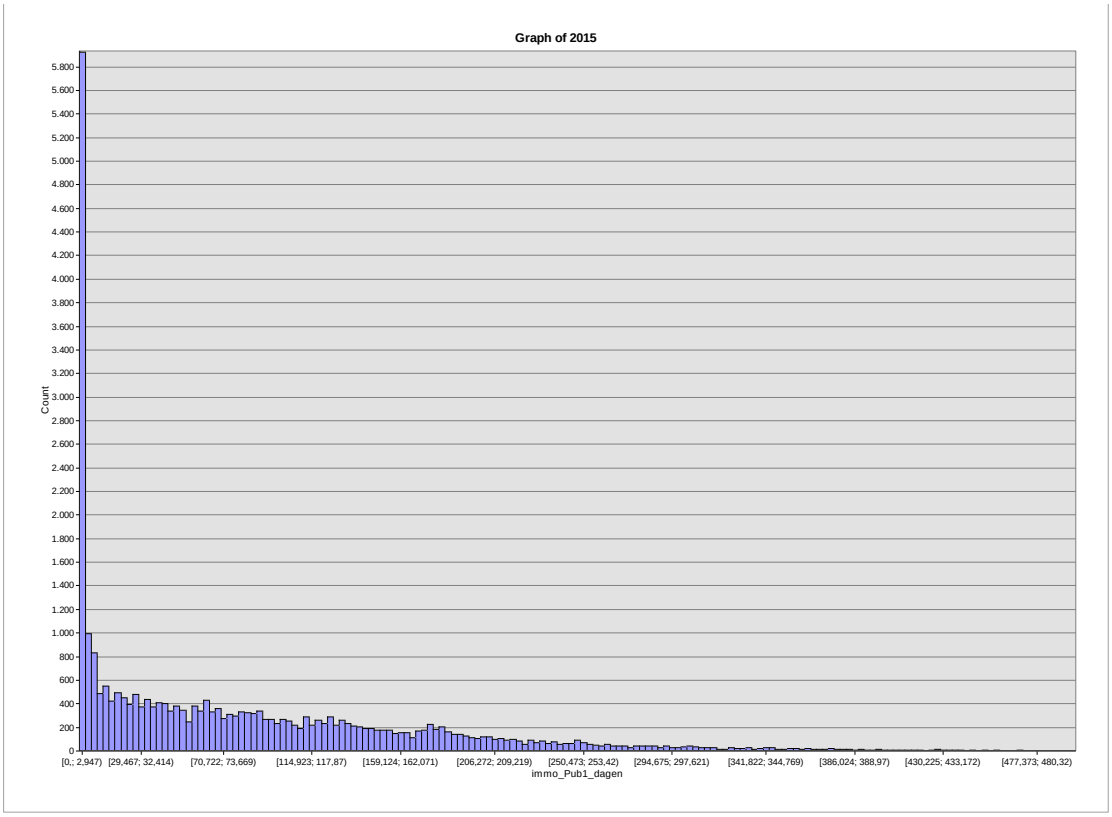
2.3.1 Huizen

Spreidingspatroon

Figuur 33 – Stippenkaart 'snelheid van verkoop', huizen te huur, 2015



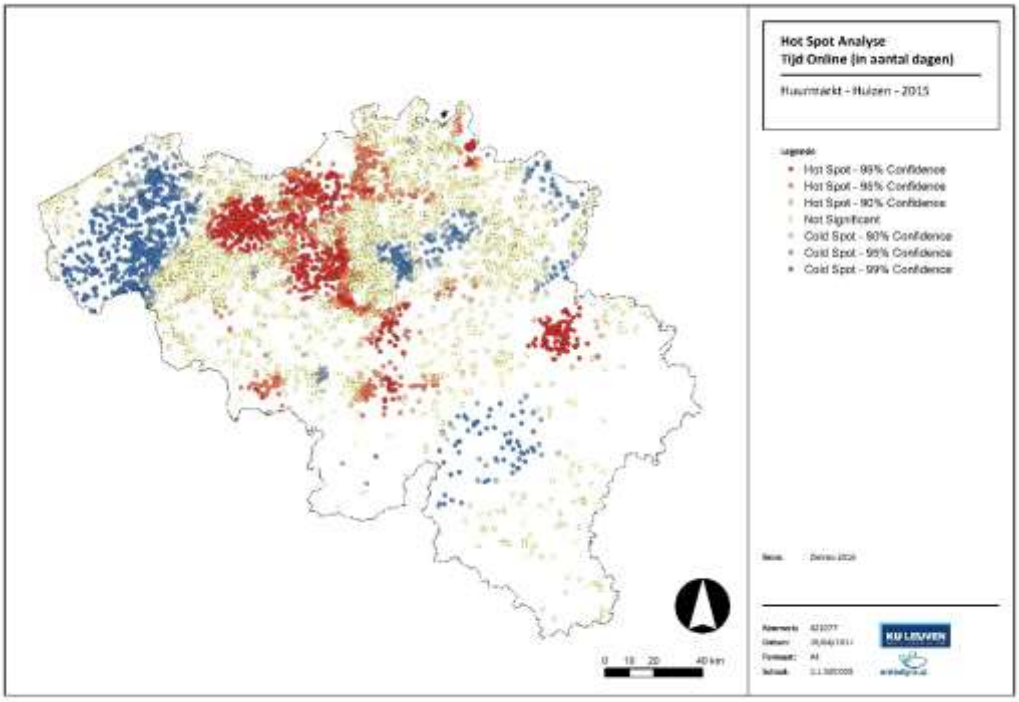
Figuur 34 – Histogram 'snelheid van verkoop', huizen te huur, 2015



Hotspot analyse

In de hostpot analyse van de tijd online in huurmarkt van huizen, is er een duidelijk patroon: in de driehoek Antwerp-Brussel-Gent zijn de huizen sneller verhuurd. Dat sterkt zich ten Noorden en Oosten van Antwerpen uit tot in de Kempen. Ten Zuiden van Brussel loopt deze hotspot door tot Charleroi. Daarnaast is er ook een concentratie sneller verhuurde huizen in Luik. Daarentegen blijven publicaties van huurhuizen in West-Vlaanderen relatief langer online dan

Figuur 35 – Hotspot analyse 'snelheid van verkoop', huizen te huur, 2015



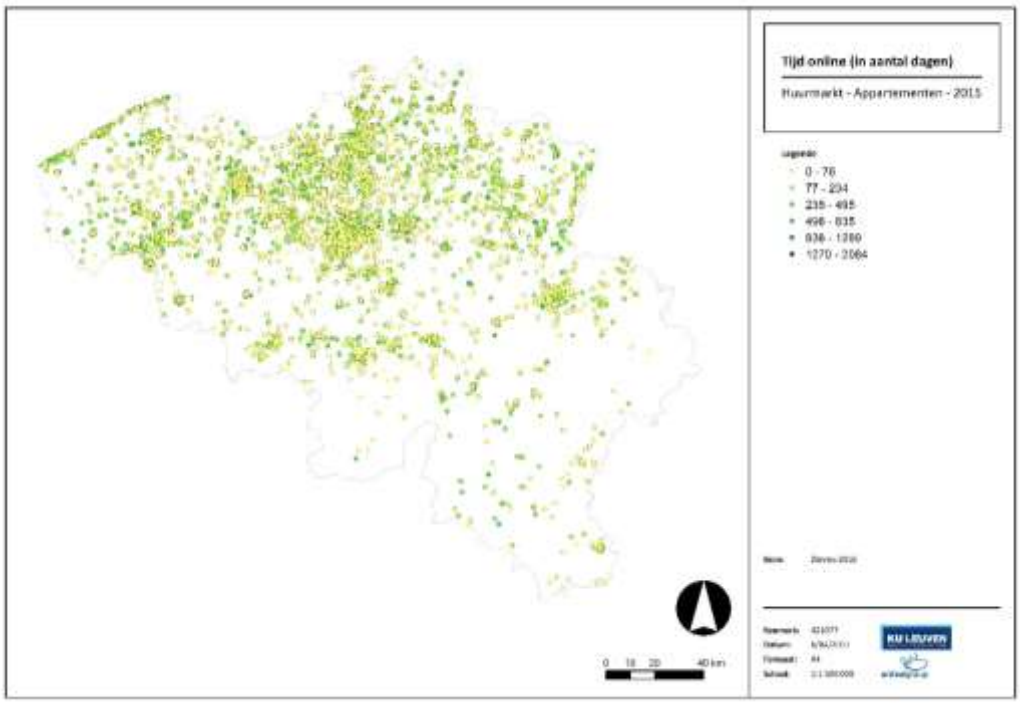
Aggregaties op niveau van statistische sectoren en gemeenten

Ook hier werd een test uitgevoerd voor Antwerpen en Gent, maar bleek het aantal records per statistische sector onvoldoende om op dat aggregatieniveau enige betrouwbare conclusies toe te laten.

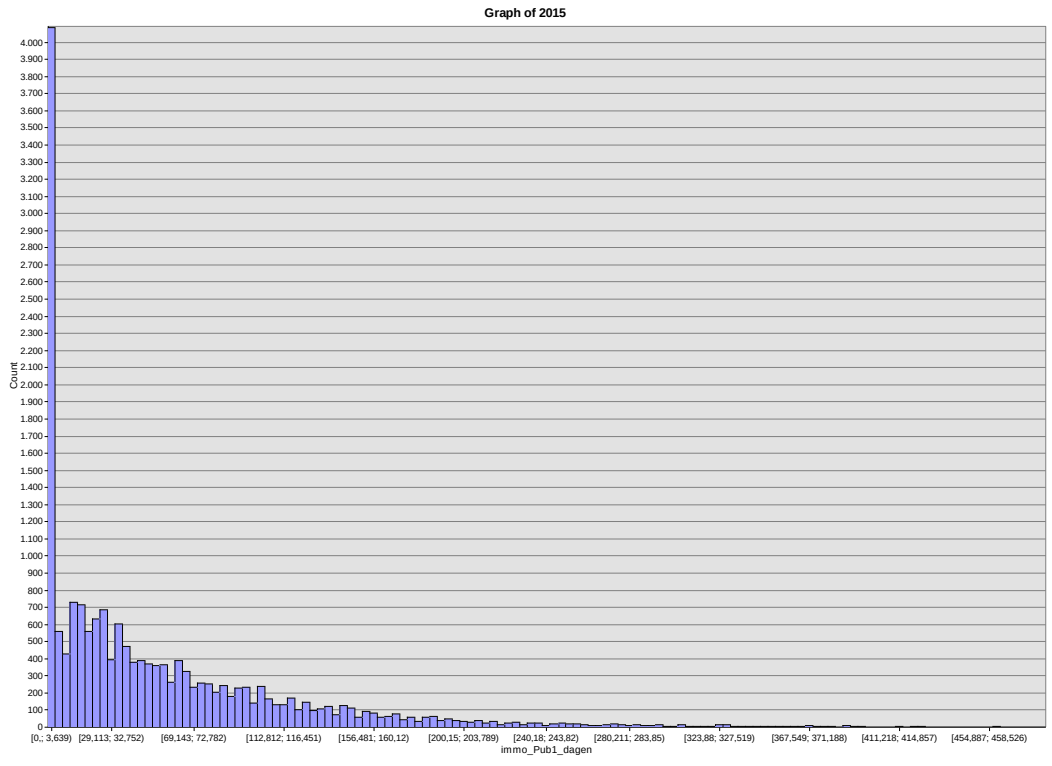
2.3.2 Appartementen

Spreidingspatroon

Figuur 36 – Stippenkaart 'snelheid van verkoop', appartementen te huur, 2015



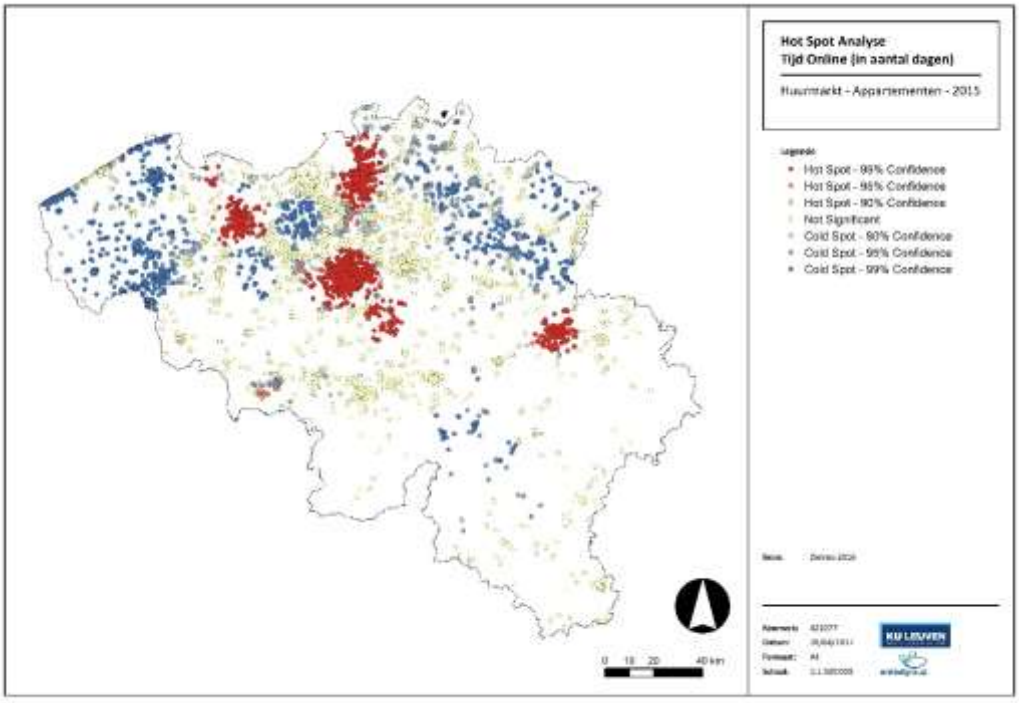
Figuur 37 – Histogram 'snelheid van verkoop', appartementen te huur, 2015



Hotspot analyse

In de hotspot analyse van de tijd online in huurmarkt van appartementen, zijn de steden Antwerpen, Gent, Brussel en Luik duidelijk herkenbaar. Daarnaast zijn er coldspots: West-Vlaanderen, inclusief Kortrijk, Brugge, de kustgemeenten nabij de Franse en de Nederlandse grens, Sint Niklaas en het omliggende het Land van Waas, grote delen van Limburg.

Figuur 38 – Hotspot analyse 'snelheid van verkoop' appartementen te huur, 2015



2.4 Bruikbaarheid van de Zimmo-databank voor de berekening van 'snelheid van verkoop'

Zowel voor huizen als voor appartementen, en voor de koopmarkt en de huurmarkt, biedt de Zimmo-databank de mogelijkheid om via hotspot analyse ruimtelijke concentraties weer te geven waar de tijd online significant langer of korter is dan wat op basis van het gemiddelde verwacht wordt. Het feit dat de gegevens op adres niveau in kaart kunnen worden gebracht, is een grote meerwaarde t.o.v. aggregaties op statistische sector niveau of op gemeenteniveau. De gegevens van 'snelheid van verkoop' kunnen bovendien uit geen enkele andere bron verkregen worden. Daaraan is wel een risico verbonden, het is niet zeker dat deze tijd online zomaar kan geïnterpreteerd worden als een proxy voor de druk op de woningmarkt: bijkomend onderzoek naar hoe en waarom mensen huizen en appartementen online te koop of te huur plaatsen kan helpen om de zeer korte tijd online beter te begrijpen. Anderzijds is het onwaarschijnlijk dat mensen in bepaalde gebieden op een andere manier met de publicatie omgaan, dus ruimtelijke verschillen kunnen wel degelijk helpen om gebieden onderling te vergelijken.

3. Beleidsindicator “frictieleegstand panden”

3.1 Berekeningsmethode

De frictieleegstand wordt berekend als het aandeel woningen te koop en te huur t.o.v. het totaal aantal woningen.

Data voor het totaal aantal woningen zijn online te vinden op de website van van Statistics Belgium (FOD Economie)[14] onder de naam ‘Kadastrale statistieken van het gebouwenpark’. Deze zijn enkel beschikbaar op het niveau van de gemeenten. Specifiek is categorie T8 ‘aantal woongelegenheden’ relevant voor deze indicator. De data met het totaal aantal woningen heeft geen verdere preprocessing nodig. De NIS code, voor de koppeling met de gemeenten, is beschikbaar in deze dataset.

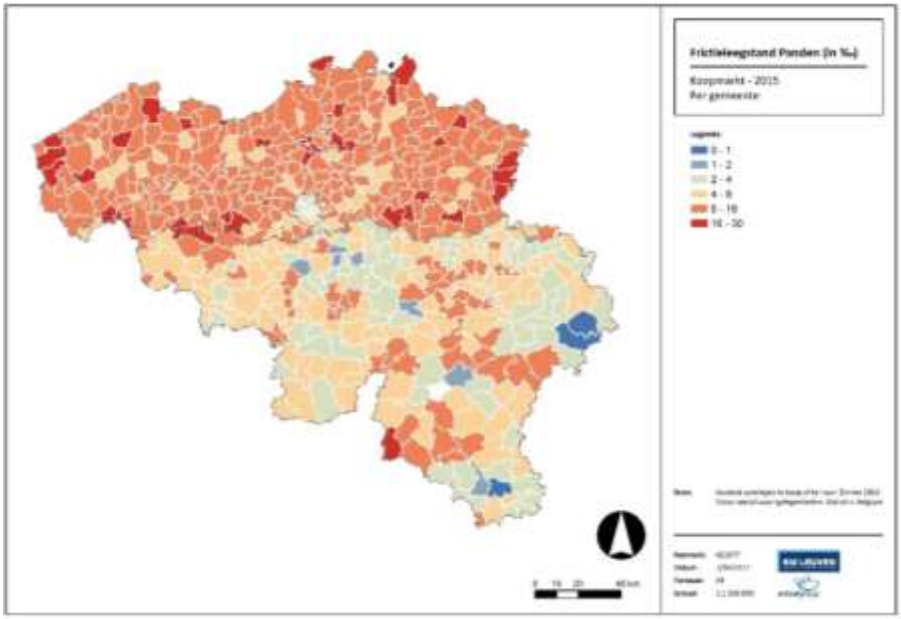
Om het aandeel woningen te koop en te huur te berekenen wordt de tool *Summary Statistics* gebruikt op de ganse Zimmo-databank. De berekening van de fictieleegstand wordt uitgevoerd in MS Access o.w.v. de complexe join functie die hiervoor nodig is (op basis van NIS code en jaartal).

De **legende** toont de verschillen in promille in intervallen van 2^n punten.

3.2 Koopmarkt

In de frictieleegstand ten gevolge van verkoop van woningen is er een duidelijk verschil merkbaar tussen Vlaanderen en Wallonië. De meeste gemeenten hebben een frictieleegstand tussen 8 en 16‰, met een aantal uitschieters tussen 16 en 30‰. In Wallonië is de frictieleegstand aanzienlijk lager.

Figuur 39 – Frictieleegstand van woningen per gemeente, koopmarkt (promille), 2015

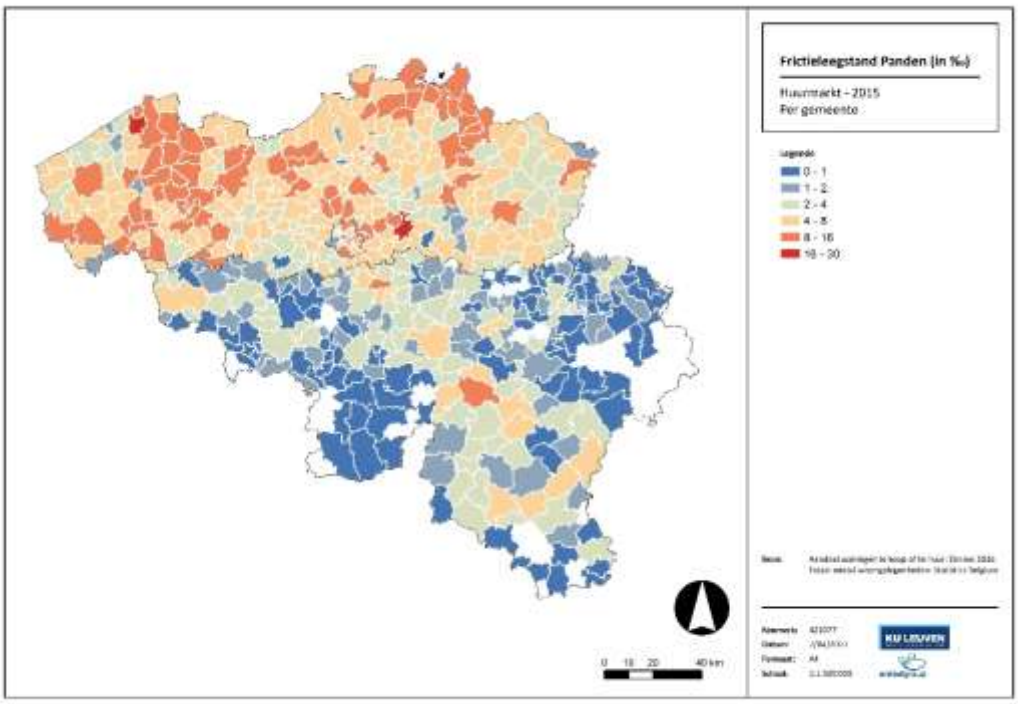


[14] http://statbel.fgov.be/nl/statistieken/cijfers/economie/bouw_industrie/gebouwenpark/

3.3 Huurmarkt

De frictieleegstand op de huurmarkt is aanzienlijk lager, maar ook daar is het onderscheid Vlaanderen (vooral tussen 4 en 8‰, of tussen 8 en 16‰.)-Wallonië (vaak minder dan 1‰) duidelijk.

Figuur 40 – Frictieleegstand van woningen per gemeenten, huurmarkt, 2015



3.4 Bruikbaarheid van de Zimmo data voor de berekening van 'frictieleegstand panden'

De frictieleegstand kan berekend worden aan de hand van de Zimmo-databank. Op dit ogenblik is nog voorzichtigheid geboden bij de interpretatie: deze indicator wordt beïnvloed door de sterkere aanwezigheid van Zimmo in Vlaanderen dan in Wallonië. Bijkomende correcties zijn wenselijk: bijvoorbeeld een weging van het aantal gepubliceerde woningen per jaar t.o.v. het aantal verkopen per jaar per gemeente.

4. Beleidsindicator “vraagprijs vs verkoopprijs”

4.1 Vraagprijs in Zimmo-databank

4.1.1 Berekeningsmethode

De vraagprijs wordt in de velden *immo_f_prijs* (voor de woningen) en *immo_grondprijs_m2* (voor de grondprijs per m²) teruggevonden. Het aantal records per deelmarkt en jaar (ter herhaling):

Tabel 32 – Aantal records voor de berekening van de vraagprijs

	KOOPMARKT (164 120)			HUURMARKT (83 767)	
	Huizen	Appartementen	Gronden	Huizen	Appartementen
2015	36 493	10 714	1 767	10 086	19 529
2014	30 264	6 686	1 597	7 983	13 251
2013	22 621	4 526	1 126	5 552	8 133
2012	18 744	3 604	675	4 520	6 077
2011	14 204	3 108	483	2 973	3 709
2010	5 709	1 516	283	802	1 117
TOTAAL	128 035	30 154	5 931	31 916	51 816

Spreidingspatroon

De basisdata (puntgegevens) worden getoond in 30 kaarten (5 deelmarkten en 6 jaartallen):

- Koopmarkt: Huizen, Appartementen, Gronden
- Huurmarkt: Huizen, Appartementen
- 2010 – 2015

Alle kaarten zijn gebaseerd op de Zimmo-databank, maar met een andere Definition Query.

De **legende** wordt ook op basis van natural breaks gekozen, maar er worden een aantal records uitgesloten omdat er anders geen variatie te zien zou zijn op de kaart. Deze records worden enkel uitgesloten van de berekening van de legenden, niet van de kaart zelf.

- Koopmarkt – huizen: de top 0.02%: 25 records tussen 5 200 000 en 180 160 000 euro
- Koopmarkt – app.: de top 0.02%: 6 records tussen 10 000 000 en 745 003 149 euro
- Koopmarkt – grond: de top 0.5%: 30 records tussen 23 000 en 300 000 euro per m²
- Huurmarkt – huizen: de top 0.9%: 294 records tussen 10 000 en 8 364 750 euro
- Huurmarkt – app.: de top 0.5%: 110 records tussen 10 000 en 6 002 010 euro

Hotspot analyse

We berekenen een hotspot analyse van dezelfde 30 puntlagen die hierboven besproken staan. De **legende** wordt automatisch bepaald door de *Hot Spot Analysis* tool.

Aggregaties op niveau van statistische sectoren en gemeenten

Omdat er te weinig statistische sectoren zijn waar een gemiddelde statistiek kan berekend worden op meer dan 30 punten – zeker niet voor elk individueel jaar, wordt er naar de cases Gent (201 statistische sectoren) en Antwerpen (298 statistische sectoren) gekeken. Om aan 30 punten per statistische sector te geraken wordt er ook over de 5 jaartallen 2011 tot 2015 geaggregeerd.

Het gemiddelde wordt berekend met de tool *Summary Statistics*. In het totaal zijn er 1996 records mogelijk voor woningen (499 statistische sectoren en 4 deelmarkten) en 499 voor bouwgronden. Uiteindelijk zijn er 1 378 en 91 respectievelijk met data en slechts 229 waar de statistiek op basis van

minstens 30 punten kan worden berekend. De onderstaande tabel geeft een opdeling van deze 229 statistieken per deelmarkt en per stad:

Tabel 33 – Aantal records voor de berekening van de gemiddeld vraagprijs per statistische sector, in Antwerpen en Gent

	KOOPMARKT			HUURMARKT	
	Huizen	Appartementen	Gronden	Huizen	Appartementen
Gent	39	13	0	6	35
Antwerpen	58	37	0	0	41

De **legende** is ook hier gebaseerd op natural breaks.

Op basis van deze test werd opnieuw besloten dat statistieken op niveau van statistische sectoren weinig zin hebben.

De samenvatting per gemeente gebeurt wel voor ieder individueel jaartal, met de tool *Summary Statistics*. In het totaal zijn er 14 136 mogelijke records voor woningen (589 gemeenten, 6 jaartallen en 4 deelmarkten) en 3 534 voor bouwgronden. Uiteindelijk zijn er respectievelijk 10 779 en 1 552 records met data en slechts 2 046 waar de statistiek op basis van minstens 30 punten is berekend. De onderstaande tabel geeft een opdeling van deze 2 046 satistieken per deelmarkt en per jaar:

Tabel 34 – Aantal records voor de berekening van de gemiddelde vraagprijs per gemeente

	KOOPKRACHT (1 465)			HUURMARKT (581)	
	Huizen	Appartementen	Gronden	Huizen	Appartementen
2015	346	82	1	92	140
2014	295	48	2	63	83
2013	237	32	1	39	55
2012	188	22	1	28	42
2011	143	22	0	15	22
2010	37	7	1	0	2
TOTAAL	**1 246**	**213**	**6**	**237**	**344**

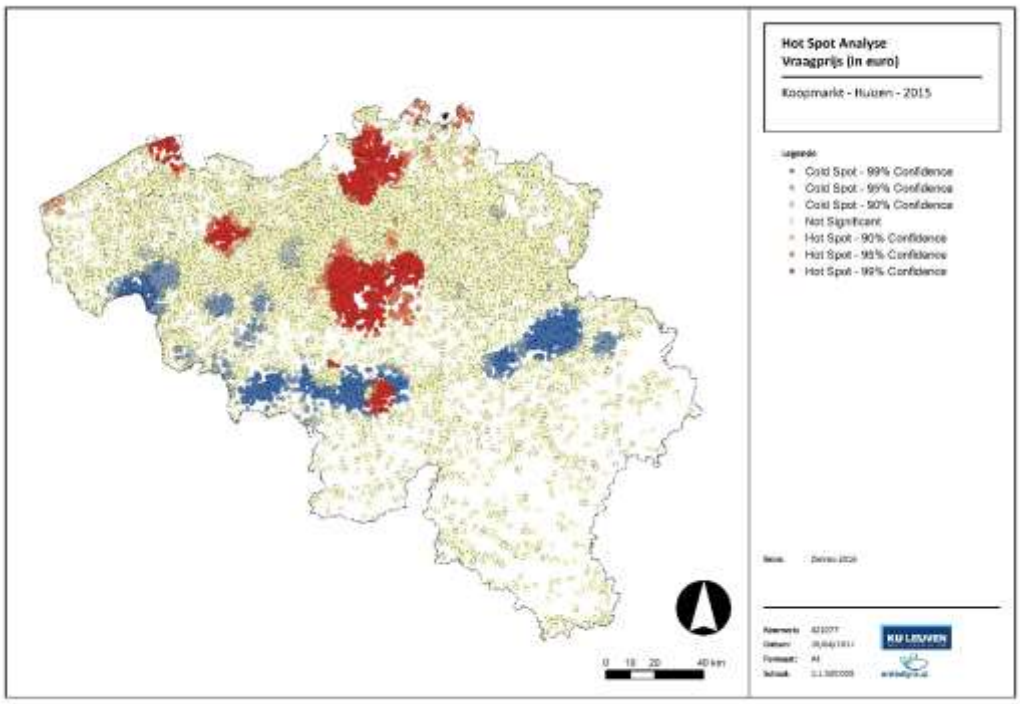
De **legende** is ook hier gebaseerd op natural breaks.

Het is ook hier opnieuw duidelijk dat de aggregatie weinig zinvol is.

4.1.2 Koopmarkt huizen

De weergave van het spreidingspatroon aan de hand van een stippenkaart is een moeilijk leesbare kaart. Aan de hand van een hotspot analyse komen concentraties in beeld.

Figuur 41 – Hotspot analyse 'vraagprijs', huizen te koop, 2015

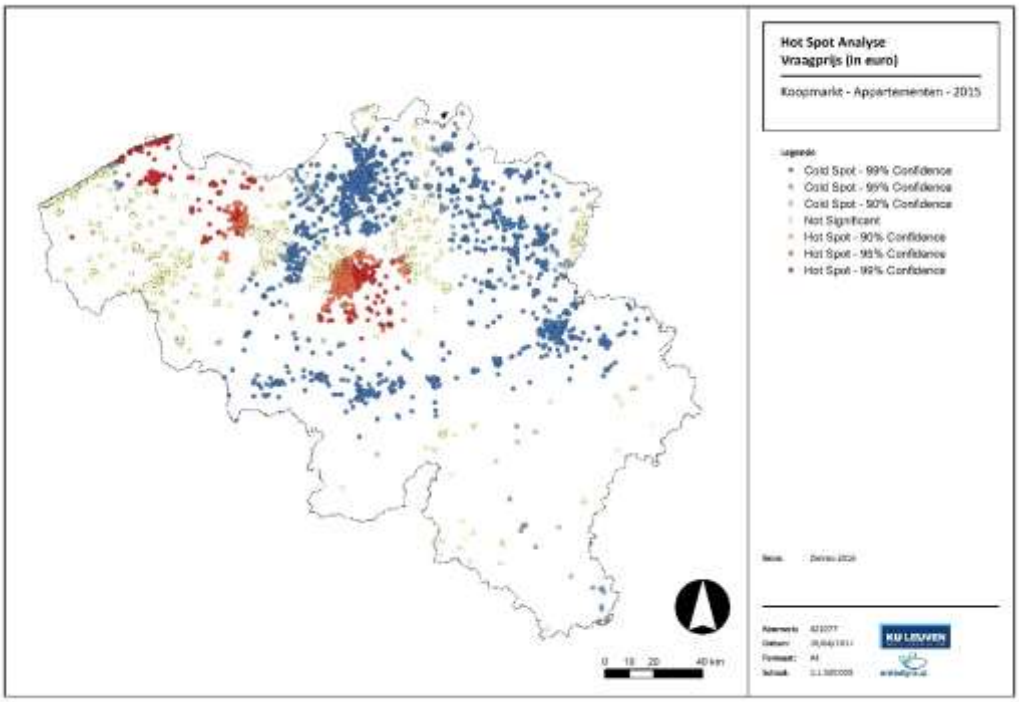


Deze kaart geeft een 'logisch' beeld: gemeenten met hoog aanbod van kleine woningen lijken bv. logischerwijs een lage vraagprijs te hebben. 'Dure gemeenten' (Brusselse rand, Leuven, Knokke, Brasschaat, Gent, …) behoren inderdaad tot de hotspots met hogere vraagprijzen. Luik, Bergen, Kortrijkse regio, … hebben lagere vraagprijzen. Het is vreemd dat de Ardennen hier niet verschijnen als gebied met lagere vraagprijzen. Dit kan o.m. het gevolg zijn dat er geen ruimtelijke concentraties zijn van verkopen van huizen.

4.1.3 Koopmarkt appartementen

Ook voor de koopmarkt van de appartementen zijn hotspots en coldspots van de vraagprijs zichtbaar. Bij de interpretatie van deze kaart is het best om vooral rekening te houden met de steden en gemeenten zoals de kust, waar concentraties van appartementen zijn. De hogere vraagprijzen voor appartementen zien we in en rond Brussel, in en rond Gent, Brugge, en de kust van Oostende tot Knokke.

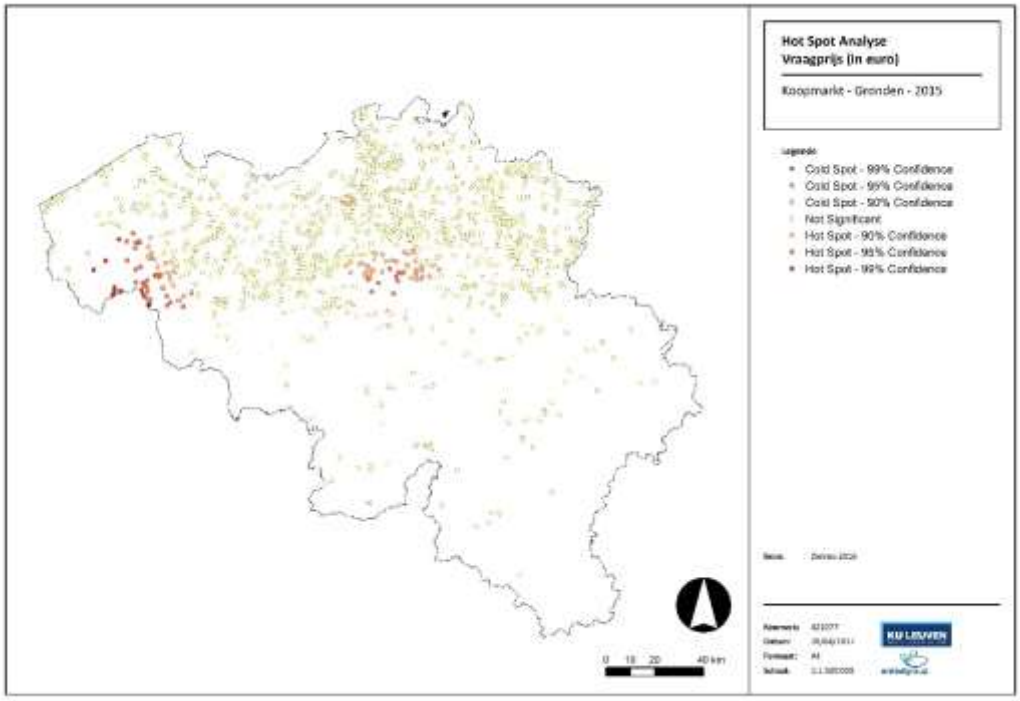
Figuur 42 – Hotspot analyse 'vraagprijs', appartementen te koop, 2015



4.1.4 Koopmarkt bouwgronden

Voor de bouwgronden werd de vraagprijs herberekend per per m^2, zoals gebruikelijk is bij vergelijkingen van prijzen, omdat de oppervlakten sterk kunnen variëren en daardoor bepalend zijn voor de prijs. Ondanks deze berekening m^2, was het niet mogelijke om ruimtelijke concentraties te identificeren, gezien het eerder vermelde spreidingspatroon van de punten.

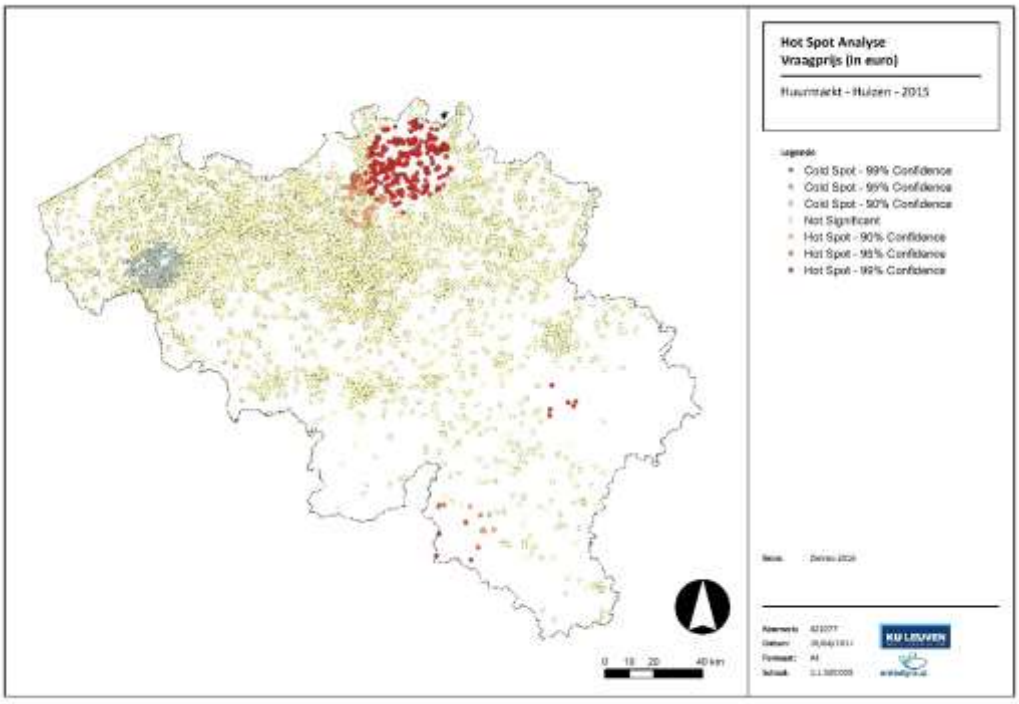
Figuur 43 – Hotspot analyse 'vraagprijs', bouwgronden te koop, 2015



4.1.5 Huurmarkt huizen

In de huurmarkt van huizen komt één hotspot (hogere vraagprijzen) in beeld, nl. de Kempen. Er is ook en één, weliswaar minder uitgesproken, coldspot (lagere vraagprijzen) in het Leiedal nabij Kortrijk.

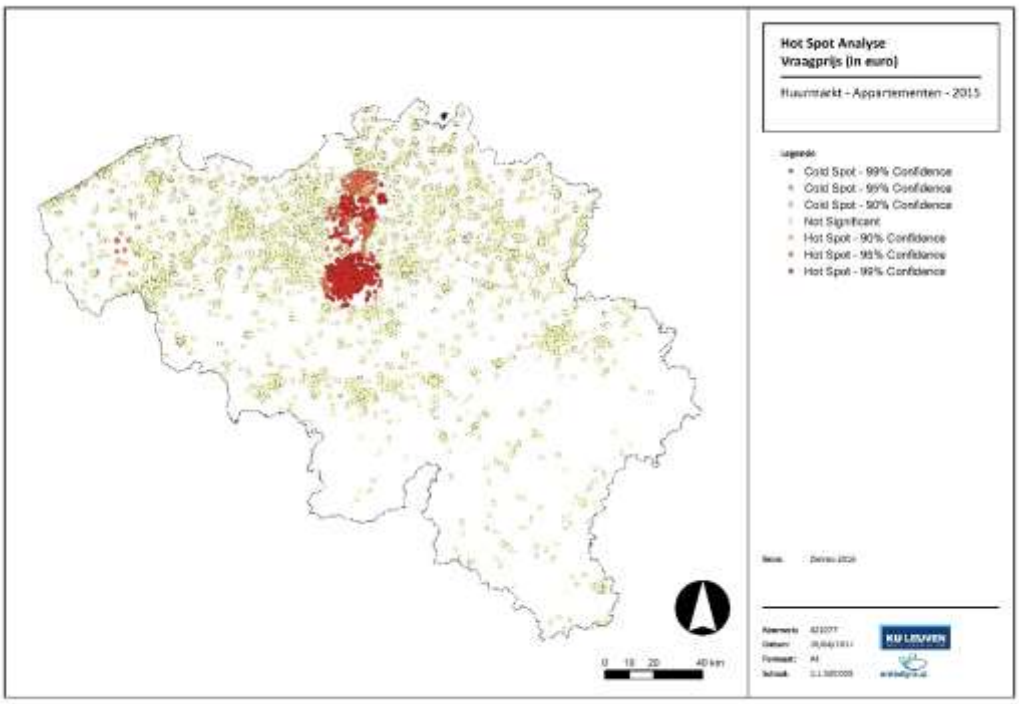
Figuur 44 – Hotspot analyse 'vraagprijs', huizen te huur, 2015



4.1.6 Huurmarkt appartementen

Bij de huurmarkt van appartementen is er één concentratiegebied waar de vraagprijzen hoger zijn, die zich uitstrekt van Antwerpen tot en met Brussel.

Figuur 45 – Hotspot analyse 'vraagprijs', appartementen te huur, 2015



4.1.7 Aggregaties op niveau van statistische sectoren en gemeenten

De aggregaties op niveau van statistische sectoren werden toegevoegd aan de geodatabase, maar geven geen duidelijke patronen weer en worden hier niet verder besproken. Het is wel een indicatie dat de microdata, op adresniveau, een schat aan informatie bieden over ruimtelijke patronen in de woningmarkt prijzen, die niet uit geaggregeerde gegevens kunnen gehaald worden.

4.2 Verkoopprijs FOD Economie

4.2.1 Berekeningsmethode

Data zijn online te vinden op de website van Statistics Belgium (FOD Economie)[15]. Deze data zijn beschikbaar op het niveau van de gemeenten en bevatten info over de gemiddelde prijs (voor bouwgronden is dit de gemiddelde prijs per m²) en over het aantal transacties per jaar.

De gegevens over woningen zijn verdeeld in drie subclasses: 'woningen' 'villa's' en 'appartementen'. Om vergelijking mogelijk te maken wordt er eerst een gewogen gemiddelde gemaakt van de woningen en villa's (weging op basis van het aantal transacties).

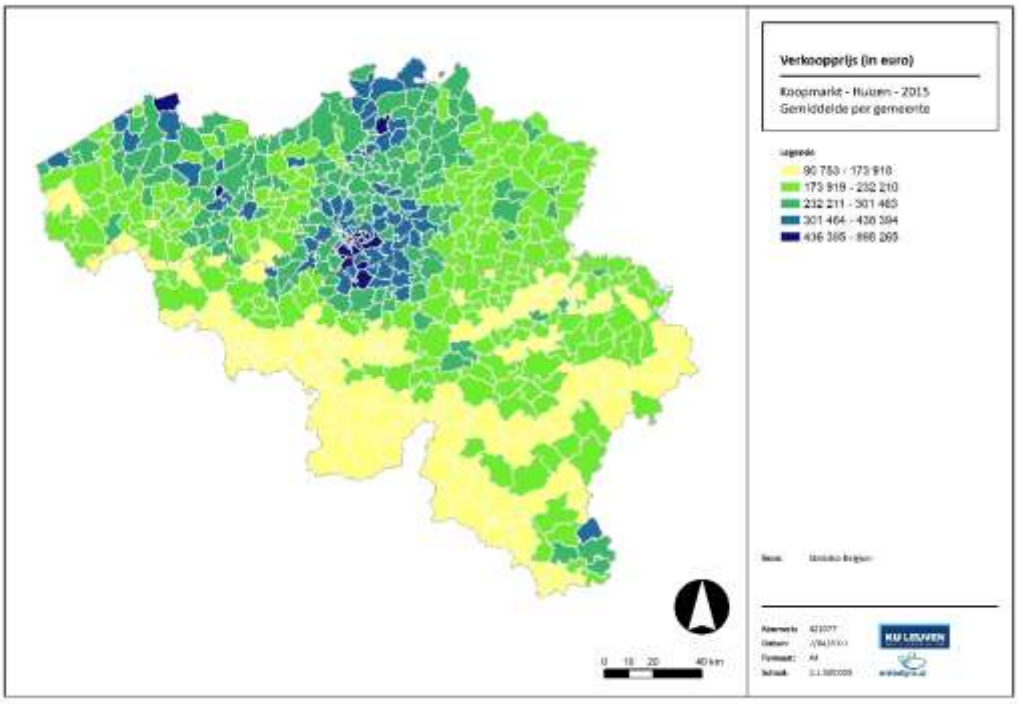
Verder is er nog wat opkuiswerk rond de gemeentenamen. Er is geen NIS code beschikbaar in de dataset, dus moet er een koppeltabel gemaakt worden die de beschikbare gemeentenamen kan koppelen aan deze NIS codes.

[15] http://statbel.fgov.be/nl/statistieken/cijfers/economie/bouw_industrie/vastgoed/

4.2.2 Huizen

Deze kaarten zijn gekend uit de publicaties van de verkoopprijzen. Zij geven het gekende verschil tussen Vlaanderen, Wallonië en Brussel weer. Ook de gekende 'dure' gemeenten zijn duidelijk herkenbaar.

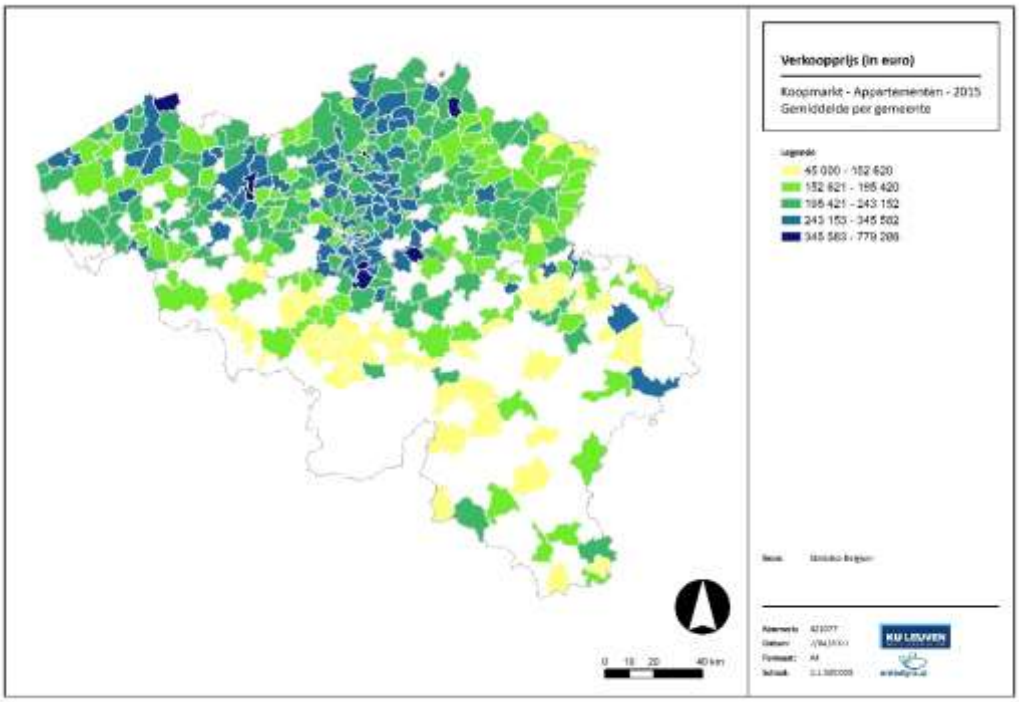
Figuur 46 – Verkoopprijs van huizen, gemiddelde per gemeente (FOD Economie, 2015)



4.2.3 Appartementen

Ook bij de appartement is deze kaart een gekend beeld. We wijzen erop dat er niet in elke gemeente appartementen worden verkocht. Desondanks is het nuttig om deze kaarten gebiedsdekkend te blijven opvolgen: de voorbije jaren werden in veel kleine kernen in Vlaanderen appartementen bijgebouwd, dus de mogelijke analyses van vraagprijs vs. verkoopprijs kunnen in de toekomst nieuwe dynamieken op deze markt in beeld brengen.

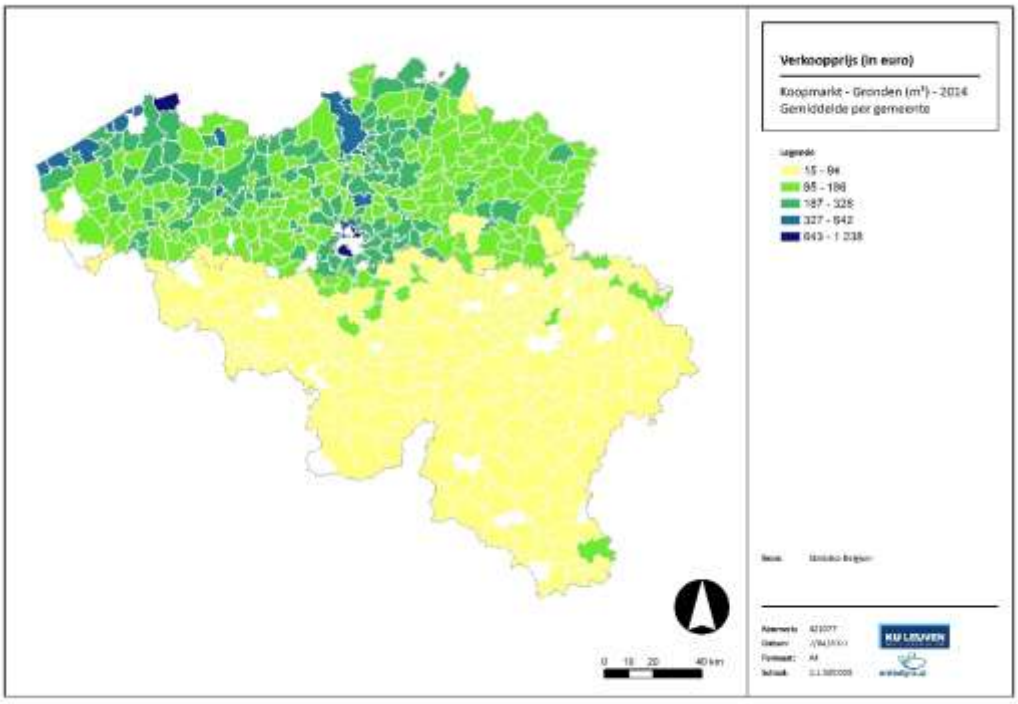
Figuur 47 – Verkoopprijs van appartementen, gemiddelde per gemeente (FOD Economie, 2015)



4.2.4 Bouwgronden

Ook wat betreft de bouwgronden is de kaart van de verkoopprijzen op basis van de gegevens van de FOD Economie, niet nieuw. De hoge vraagprijzen in Vlaanderen zijn gekend, en de evolutie doorheen de jaren kan in de Geodatabase weergegeven worden. Het bespreken van deze kaarten behoort niet tot de doelstelling van deze opdracht. Ze worden hier toegevoegd als referentiekaarten, ter vergelijking met de gegevens van de vraagprijzen Zimmo-databank.

Figuur 48 – Verkoopprijs van bouwgronden, gemiddelde prijs per m² per gemeente (FOD Economie, 2014)



4.3 Verschil tussen vraag en verkoopprijs

4.3.1 Berekeningsmethode

Gezien de verkoopprijzen enkel beschikbaar zijn op gemeenteniveau, kan de vergelijking ook enkel op dit niveau gemaakt worden.

De verwerking van deze gegevens gebeurt buiten ArcMap in MS Access. De bewerkingen zijn:

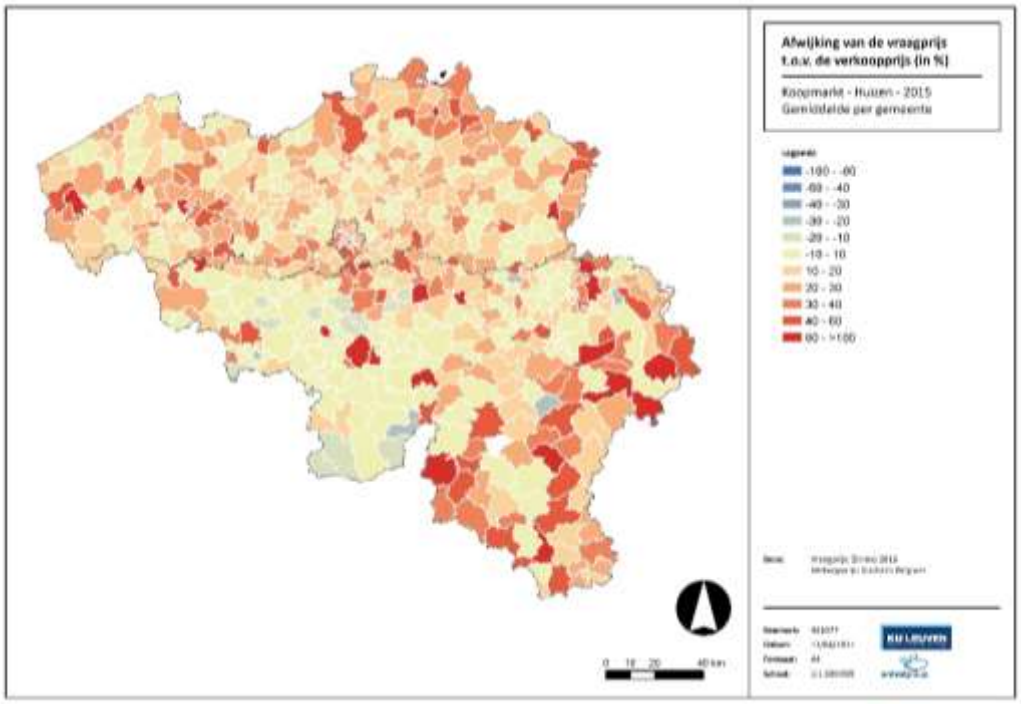
- Enkel records weerhouden waar een gemiddelde verkoopprijs berekend is;
- Koppelen aan de vraagprijs op basis van NIS code, jaar en woning type;
- Statistieken verzamelen in één tabel:
    - *aantal_vraag* het aantal verkopen volgens Zimmo
    - *mean_vraag* de gemiddelde vraagprijs volgens Zimmo
    - *aantal_verkoop* het aantal verkopen volgens Statistics Belgium
    - *mean_verkoop* de gemiddelde verkoopprijs volgens Statistics Belgium
    - *verschil_prijs* het verschil tussen de vraag- en de verkoopprijs
    - *afwijkingPct_prijs* het verschil ten opzichte van de verkoopprijs (%)
    - *percent_opZimmo* het aantal transacties dat terug te vinden is op Zimmo (%)

De **legende** toont het procentuele verschil tussen vraag- en verkoopprijs, ten opzichte van de verkoopprijs, in intervallen van 10 tot 40 punten.

4.3.2 Huizen

Figuur 49 geeft de afwijking aan tussen de gemiddelde vraagprijs in de Zimmo-databank (op het ogenblik dat de publicatie offline wordt gehaald) per gemeente en de gemiddelde verkoopprijs geregistreerd door de FOD Financiën. De gele klassen in de legende geven aan dat de vraagprijs 10% hoger of lager ligt dan de geregistreerde officiële verkoopprijs. De meeste gemeenten kleuren rood, wat wijst op een overschatting (m.a.w. de vraagprijs ligt hoger dan de –officiële- verkoopprijs, meestal tussen 20% en 40%. In enkele landelijke gemeenten in Wallonië (grijsblauwe gemeenten op de kaart) is de vraagprijs lager dan de gemiddelde verkoopprijs in de gemeente

Figuur 49 – Verschil tussen de vraagprijs (Zimmo-databank) en verkoopprijs (FOD Economie, 2015) voor huizen



Deze verschillen zijn berekend op basis van gemiddelden per gemeente, waardoor de extreme waarden het resultaat sterk kunnen beïnvloeden. De correlatie tussen de gemiddelde vraagprijs per gemeente en de gemiddelde verkoopprijs is bijzonder laag: R^2 = 0.15. Als we de mediaan per gemeente vergelijken, dan ligt de waarde zeer dicht bij het gemiddelde van de officiële verkoopprijs (Tabel 35).

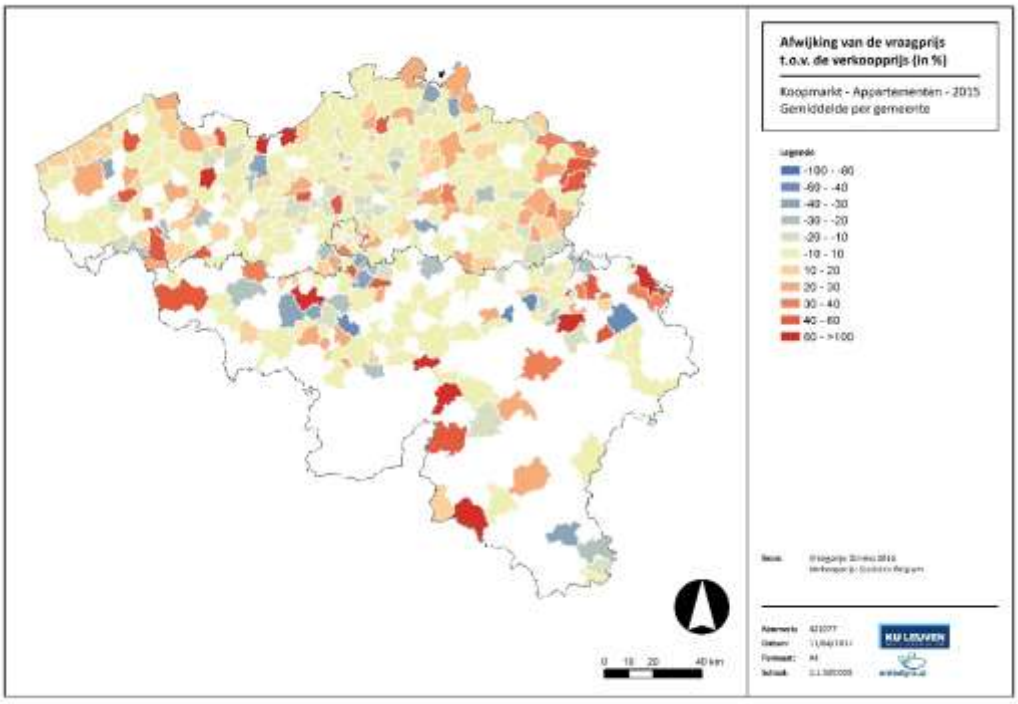
Het probleem van de gemiddelden per gemeente beperkt zich niet tot de extreme waarden. In steden als Brussel, Gent en Antwerpen gaat het om honderden transacties per jaar, met een enorme range in prijzen. Zoals blijkt uit de vergelijking tussen figurenFiguur 41 en Figuur 46, volgen de patronen niet de gemeentegrenzen.

Tabel 35 – Vergelijking tussen de gemiddelde verkoopprijs en de mediane verkoopprijs van huizen

	Vraagprijs (Zimmo-databank) huizen 2015	Gemiddelde verkoopprijs/gemeente (FODFIN) huizen 2015
gemiddelde	284 789	236 285
Mediaan	235 000	

4.3.3 Appartementen

Figuur 50 – Verschil tussen de vraagprijs (Zimmo-databank) en verkoopprijs (FOD Economie, 2015) voor appartementen

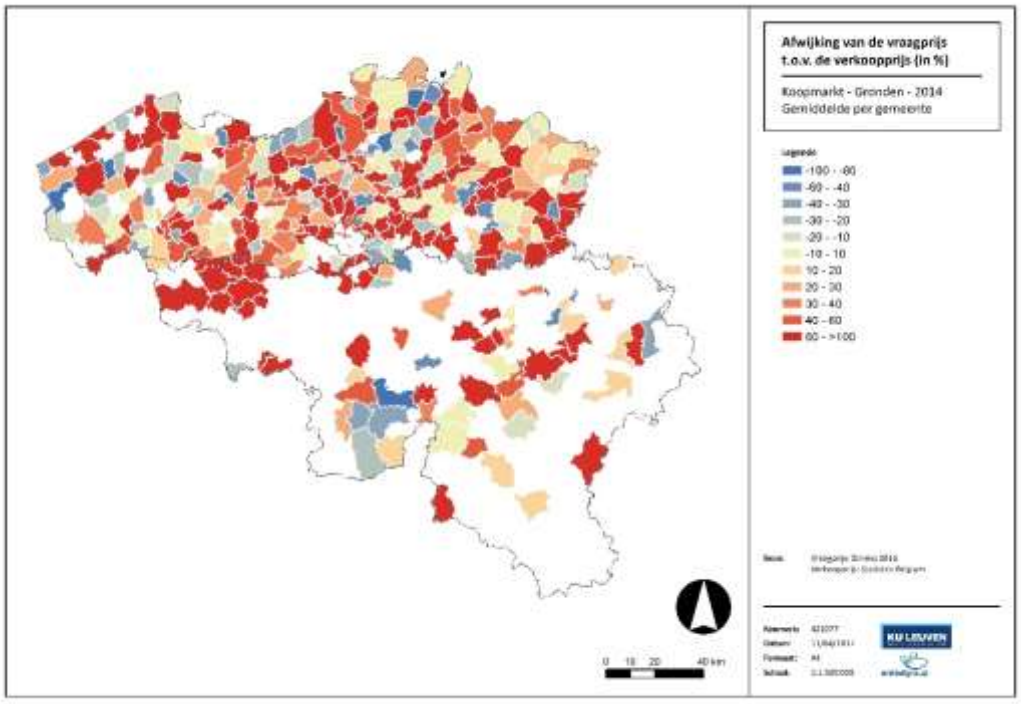


Voor de verkoop van appartementen was het aantal verkopen waar meer dan één prijs werd ingevuld beperkt en kon in 2015 slechts voor 370 gemeenten het verschil tussen vraagprijs en verkoopprijs worden berekend (Figuur 50).

4.3.4 Bouwgronden

Ook voor bouwgronden zijn er weinig transacties, waardoor de schijnbare hoge vraagprijs per m^2 vergeleken met de verkoopprijs mogelijk een niet representatief beeld geeft (Figuur 51).

Figuur 51 – Verschil tussen de vraagprijs (Zimmo-databank) en verkoopprijs (FOD Economie, 2014) voor bouwgronden



4.4 Bruikbaarheid van de Zimmo data voor de analyse van vraagprijzen en verkoopprijzen

Uit de vergelijking tussen vraagprijs en verkoopprijs, blijkt (zoals verwacht) een structurele overschatting. Daarnaast zijn er grote afwijkingen, die kunnen te wijten zijn aan het beperkte aantal records. Ook in deze analyse is de aggregatie van transacties op gemeenteniveau in de publicatie van de FOD Financiën, een handicap. Een vergelijking op micro data niveau (reële transacties koppelen aan records in de Zimmo-databank), kan een schat aan informatie opleveren. De Zimmo-databank bevat nl. een zeer uitgebreide beschrijving van kenmerken van woningen en appartementen, die belangrijk blijken bij het te koop stellen van huizen en appartementen. De FOD Financiën data kunnen deze koppelen aan reële transacties en verkoopprijzen.

Door de grote nog onverklaarde verschillen tussen de prijzen in de Zimmo-data en de verkoopprijzen van de FOD Economie, werd beslist om mogelijke relaties tussen de prijzen in de Zimmo-databank en de knooppuntwaarde en het voorzieningen niveau, en de verstedelijkingsgraad, in deze fase nog niet te onderzoeken. Voorafgaand zijn bijkomende verbeteringen aan de databank wenselijk zodat de prijzen eenduidiger kunnen geïnterpreteerd worden.

4.5 Relatie tussen vraagprijzen en verstedelijkingsgraad, voorzieningen niveau

Er is geen correlatie, noch tussen de vraagprijs en de verstedelijkingsgraad (R^2 = 0.04 bij 6 klassen, -0.05 bij 7 klassen), noch met de gridcode die het voorzieningenniveau en de knooppuntwaarde weergeeft (R^2 = 0.03). Dit is niet verwonderlijk: de vraagprijs van een woning wordt in eerste instantie bepaald door de kenmerken van de woning zelf. 'Locatie is bepalend voor de prijs' betekent m.a.w. 'voor gelijkaardige woningen in verschillende omgevingen'.

5. Beleidsindicator “Schaarste-index bouwgrond”

5.1 Berekeningsmethode

De schaarste-index voor bouwgronden wordt berekend als het aandeel bouwgronden dat te koop staat, t.o.v. het aantal onbebouwde percelen in woongebied.

Het aandeel bouwgronden dat te koop staat wordt gehaald uit het aantal bouwgrond records dat online heeft gestaan in een bepaald jaar (Zimmo-databank).

Het aantal onbebouwde percelen per gemeente wordt gehaald uit het ROP (Register Onbebouwde Percelen). Deze gegevens bestaan enkel voor Vlaanderen en enkel vanaf 2013 (namelijk de situatie op 1 april van elk jaar). Deze dataset bestaat uit lagen met polygonen voor elk perceel. Om het aantal onbebouwde percelen te berekenen wordt er als volgt te werk gegaan:

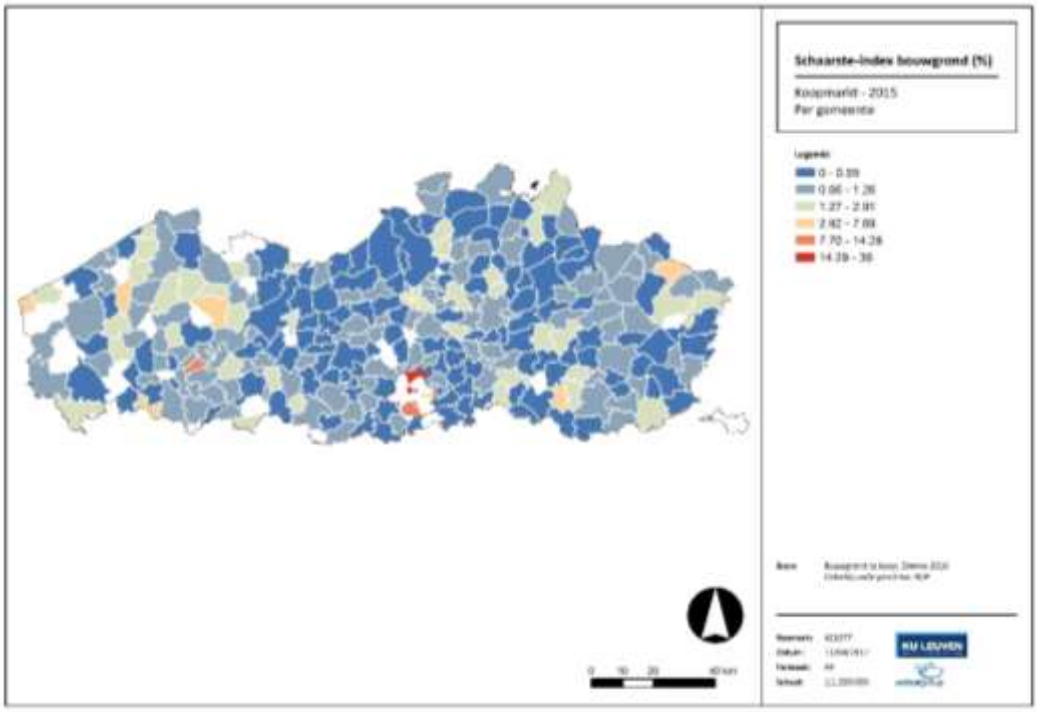
1. Spatial join tussen de percelen en de gemeenten, op basis van het centrum van de polygon, omdat verschillende polygonen over de gemeentegrenzen heen liggen.
2. Optellen van de polygonen (percelen) per gemeente.

Tenslotte worden deze gegevens gekoppeld en wordt de verhouding berekend.

5.2 Schaarste-index kaarten

Hieruit blijkt dat er in de Zimmo-databank een zeer gering aandeel van de bouwgronden te koop wordt aangeboden (Figuur 52).

Figuur 52 – Schaarste-index bouwgronden 2015



5.3 Bruikbaarheid van de Zimmo data voor de berekening van ‘schaarste-index bouwgronden’

Het aantal publicaties van bouwgronden te koop in de Zimmo-databank is te laag om de schaarste-index van bouwgronden te berekenen; de lage index is waarschijnlijk vooral een indicatie dat bouwgronden minder online worden verkocht dan huizen of appartementen.

DEEL 4 – Conclusies en aanbevelingen

Knelpunten en potentie van de Zimmo-databank

Op basis van de analyse van de databank, in combinatie met de berekende beleidsindicatoren, worden de volgende knelpunten en potenties van de databank vastgesteld.

Het groot aantal bruikbare gegevens voor ruimtelijke analyses en het hoog aandeel verkochte woningen ten opzichte van het totaal aantal verkopen per gemeente (cfr. kadaster) is een indicatie van de hoge potentie van de Zimmo-databank, ook in functie van specifieke deelanalyses. De meerderheid van de gemeenten heeft een afwijking van minder dan 20% in het aantal publicaties ten opzichte van de statistieken van transacties op basis van het kadaster, en de vraagprijs is meestal hoger dan de verkoopprijs, wat ook logisch is. Ook dit is veelbelovend. De potentie van de databank neemt trouwens sinds 2010 jaarlijks toe. Door de trend naar toenemende verkoop en verhuur van woningen via het Internet zal in de toekomst de bruikbaarheid van deze databank voor berekening van beleidsindicatoren nog verder toenemen. Enkele ruimtelijke patronen berekend aan de hand van hotspot analyses tonen aan dat de Zimmo-databank leidt tot nieuwe inzichten in de woningmarkt. De indicator 'Tijd te koop' (proxy tijd online) is bv. een indicator die weliswaar een ruwe proxy is, maar die in geen andere databank terug te vinden is. In de hotspots analyses op basis van de microdata zijn duidelijke patronen waarneembaar. Deze zijn weliswaar niet allemaal te verklaren, maar bieden potentieel voor bijkomend onderzoek.

Na deze eerste verkenning van de dataset blijven er nog veel onzuiverheden. Uit de ruimtelijke analyse blijken de data te ruw om voor een aantal beleidsindicatoren (nu al) betrouwbare conclusies te trekken. Niet al de fouten die aanwezig zijn in de dataset, kunnen uitgefilterd worden via een globale analyse. Outliers zijn in grote mate geïdentificeerd, maar komen nog veel voor. Ze kunnen niet zonder meer buiten beschouwing worden gelaten omdat ze nog niet verklaard kunnen worden. Het al dan niet betrekken van outliers is afhankelijk van specifieke analyses. Zo is bijvoorbeeld de vraagprijs een belangrijke variabele met outliers. Enkele mogelijke verklaringen voor deze outliers kunnen zijn: meerdere appartementen in eenzelfde gebouw worden verkocht en de prijs verwijst naar meerdere wooneenheden, de verhuur wordt aan het volledige adres gekoppeld maar betreft slechts een deel van een woning …). Het is verder ook duidelijk dat heel wat velden niet altijd correct zijn ingevuld (bv. jaartallen), maar het ontbreekt nog aan een volledig inzicht hierin. Ook zijn er veel variabelen waarvoor de respons laag is. Dit kan deels opgelost worden via surrogaatvariabelen, maar ook een optimalisatie van invoer van data (zie hieronder) kan hiertoe bijdragen. Tot slot wordt ook vastgesteld dat er een aantal gemeenten zijn, zeker in Wallonië, met weinig records in totaliteit. Er worden nu eenmaal in sommige landelijke gemeenten weinig woningen verkocht of verhuurd. Een eenduidige gebiedsdekkende aggregatie van gegevens tot gemeenteniveau en zeker tot op niveau van statistische sectoren is bijgevolg niet altijd zinvol. Een differentiatie van het aggregatieniveau in functie van de verstedelijkingsgraad is in verdere ruimtelijke analyses aangewezen.

Nood aan verdere uitzuivering van de dataset

Op basis van de opmaak van het analysebestand, gevolgd door een eerste analyse van variabelen en een geografische analyse in functie van het opstellen van beleidsindicatoren, zijn een aantal nieuwe inzichten aan het licht gekomen die kunnen leiden tot bijkomende verbeteringen van de databank. De eerste ruimtelijke analyse laat toe om ruimtelijke patronen te detecteren en in beeld te brengen, maar brengt tegelijk ook aan het licht dat er nog te veel afwijkingen zijn om in deze fase verder te gaan met een aantal bijkomende analyses (factoranalyse, clusteranalyse, residu analyse). Ook stappen in de richting van een steekproef zijn nog voorbarig; er is nog te veel ruis door ongewenste causale factoren om goede criteria te bepalen voor, bijvoorbeeld, een stratificatie.

Hiervoor is in principe eerst een verdere uitzuivering van de databank noodzakelijk. Zeker voor wat betreft de prijzen, is nog bijkomend onderzoek nodig naar het effect van outliers. Daarnaast kan de feedback tussen de ruimtelijke analyse en de opmaak van het analysebestand opnieuw leiden tot

optimalisatie van de databank en bijkomende ruimtelijke analyses. Ook deelanalyses op datakwaliteit, opgedeeld naar type bebouwing (en eventueel subtype) kan hierin een meerwaarde bieden. Dankzij een iteratief proces (bv. 3 tot 5 iteraties), kan de ruis uit de databank nog sterk gereduceerd worden en kunnen verklarende factoren en causale relaties verder in beeld gebracht worden.

Deze vaststellingen kunnen ook vertaald worden naar aanbevelingen voor Zimmo. Het toekennen van een uniek id dat kan ingezet worden bij update van de databank is er één van. Daarnaast is op veel vlakken een optimalisatie van invoer mogelijk (voor zover het geen crawling betreft) en nodig om tot een meer betrouwbare invulling van velden (er is voor veel velden een risico voor foutieve invoer) en tot een hogere respons te komen (door bv. invulling van bepaalde velden te verplichten). Er zijn ook nogal wat fouten bij het toekennen van kenmerken aan een volledig gebouw, vs. aan één wooneenheid. Bijvoorbeeld, wanneer in één herenhuis drie appartementen en één garage te huur gepubliceerd worden.

Extra aanbeveling: naar een dynamische geïntegreerde immo-databank

Behalve de Z-immo databanken, bestaan er nog andere immo-databanken zoals ERA en de gegevens van FOD Economie, met elk hun kwaliteiten en knelpunten. De databanken zijn complementair en kunnen elkaar aanvullen om zo tot een optimale geïntegreerde immo-databank te komen.

Een afstemming en/of onderlinge integratie van databanken impliceert wel tal van aandachtspunten. Dit gaat onder meer over eigendomsrechten op de data(bank) en (toelating tot / rechten op) gegevensgebruik. Het gebruik van (gegevens van) een bestaande databank is wat anders dan het afleiden van een nieuwe databank of, meer nog, het integreren van databanken afkomstig van diverse eigenaars.

Daarnaast spelen ook heel wat juridische elementen die een duurzaam gebruik en beheer zouden kunnen bemoeilijken. We noemen hier in het bijzonder de nieuwe EU-regelgeving inzake de bescherming van persoonsgegevens. In mei 2018 treedt de GDPR (Algemene Verordening inzake Persoonsgegevens) in werking. Een belangrijk aspect is de manier waarop persoonsgegevens in de richtlijn worden gedefinieerd: als gegevens aan de hand waarvan een persoon direct of indirect kan worden geïdentificeerd. Dit betekent dat ook IP-adressen, locatiegegevens en andere elementen aan de hand waarvan iemand kan worden geïdentificeerd onder de nieuwe wet vallen. Wanneer in de toekomst wordt gewerkt aan integratie van databanken, zal in het bijzonder aandacht nodig zijn voor het aggregeren en anonimiseren van de data.

Ook vanuit technisch en statistisch oogpunt is een integratie van databanken verder te onderzoeken. De mogelijkheid voor analyse van de data optimaliseren of maximaal informatiebehoud zou een leidraad kunnen zijn. Het opzet van de integratie dient voorafgaandelijk eenduidig bepaald. Zo zal unilaterale categorisering van variabelen ongewenst informatieverlies met zich kunnen brengen. Anderzijds dient zo weinig mogelijk weinig bruikbare data (bv. deels foutief) geïntegreerd.

Omwille van onder meer de knelpunten die nu nog spelen in de Z-immo databank, zou een integratie met andere immo data wel een belangrijke meerwaarde bieden, zeker indien op termijn kan worden toegewerkt naar een dynamische databank, met regelmatige updates.

Mogelijke bijkomende analyses

Gezien het doel van de exploratieve ruimtelijke analyse was om na te gaan in welke mate de gegevens geschikt zijn voor de berekening van beleidsindicatoren over de woningmarkt, ligt de nadruk in dit rapport op de haalbaarheid van de berekeningsmethode. De interpretatie van de resultaten is een eerste visuele beschrijving van de resulterende kaarten. Bijkomende analyses op deze beleidsindicatoren kunnen ruimtelijke verbanden en evoluties in kaart brengen, voor gebruik van de Zimmo-databank in de monitoring van de effectiviteit van bepaalde beleidsmaatregelen op het terrein.

De resultaten van hotspot analyses zijn voor al de jaren beschikbaar. In het rapport zijn enkel de resultaten van de meest recente analyse besproken. De mogelijkheid bestaat om **trends** te

identificeren, maar vereist bijkomende analyse van de representativiteit van de data voor, onder andere, de verkoopprijs in de opeenvolgende jaren.

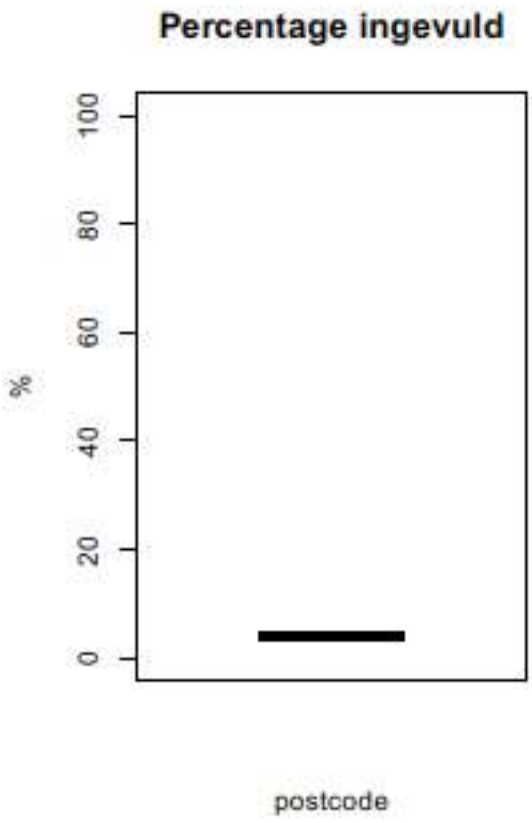
De beleidsindicatoren zijn in deze fase per subset van de deelmarkten besproken. **Dwarsverbanden** tussen de verschillende beleidsindicatoren kunnen tot interessante inzichten leiden. Een voorbeeld: de combinatie van de hotspots/coldspots ontdekt in de vraagprijs en de tijd online. Daar lijkt er een ander verhaal te spelen in Vlaanderen en Wallonië. Het is vreemd dat alle combinaties (goedkoop snel / goedkoop traag / duur snel / duur traag) voorkomen, dit zou verder onderzocht kunnen worden.

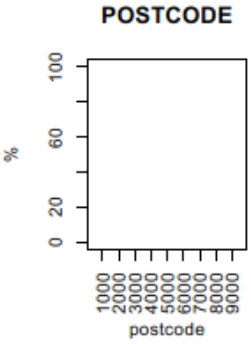
De al dan niet vastgestelde ruimtelijke patronen in de beleidsindicatoren kunnen verder gebruikt worden in woningmarkt analyses, door **bijkomende combinaties met demografische en socio-economische gegevens**. Een interessante denkpiste: combinaties tijd online, vraagprijs, demografische ontwikkelingen, percentage huur/koop en evoluties hierin, van de laatste 10 jaar. Dit kan in beeld brengen waar de combinatie tijd online / vraagprijs gepaard gaat met bepaalde demografische ontwikkeling. Ter illustratie, in de streek van Kortrijk is er weinig demografische groei, dit kan verklaren dat, ondanks de goedkopere woningen deze niet snel verkocht worden. Een ander voorbeeld is de evolutie naar meer kopers t.o.v. huurders in relatie tot interesten. Als interesten laag staan, gaat men mogelijk overstappen van huur naar koop dit kan de vraag stimuleren en de speed of sale en de vraagprijs beïnvloeden. In de huidige ruimtelijke analyses werd, naar analogie met de statistieken van de woningmarkt, per jaar gerekend. De micro data laten toe om, indien relevant, **andere tijdstippen** te kiezen voor, bijvoorbeeld een ex-ante evaluatie. Soms kunnen streken in de lift zitten, maar hebben prijzen zich nog niet aangepast of omgekeerd (men gaat prijs niet snel laten zakken). Deze timelag kan onderzocht worden.

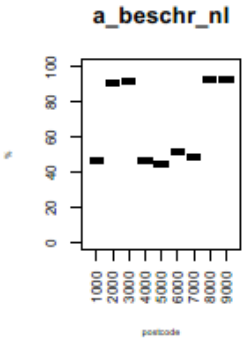
Bijlagen

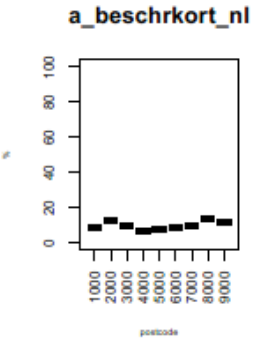
Bijlage 1: Non-respons totale dataset

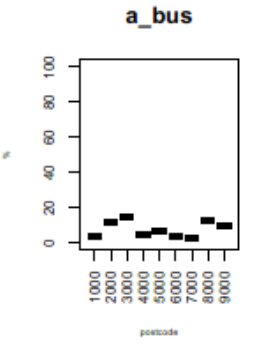
Bijgevoegde grafieken geven weer hoeveel procent van de gegevens zijn ingevuld voor de verschillende variabelen en dit per postcode en geldig op de volledige Z-immo dataset.

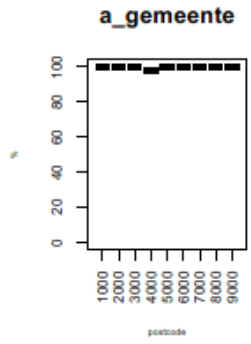


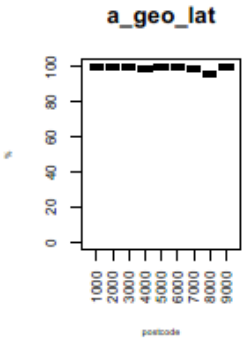

POSTCODE
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

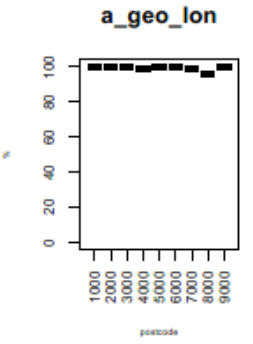

a_beschr_nl
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

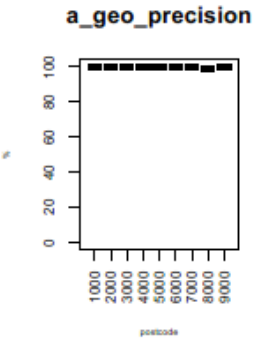

a_beschrkort_nl
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

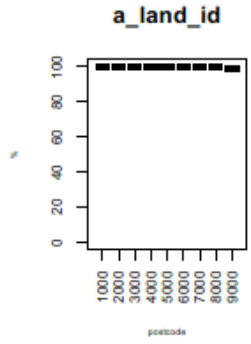

a_bus
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

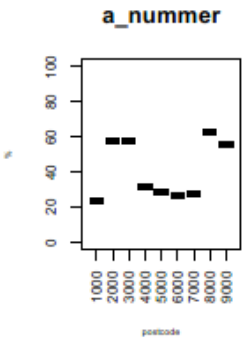

a_gemeente
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

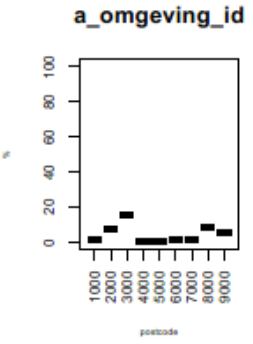

a_geo_lat
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

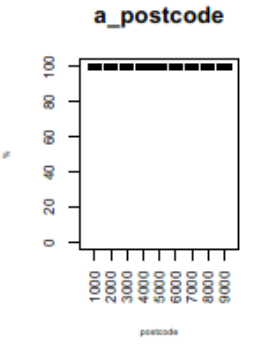

a_geo_lon
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

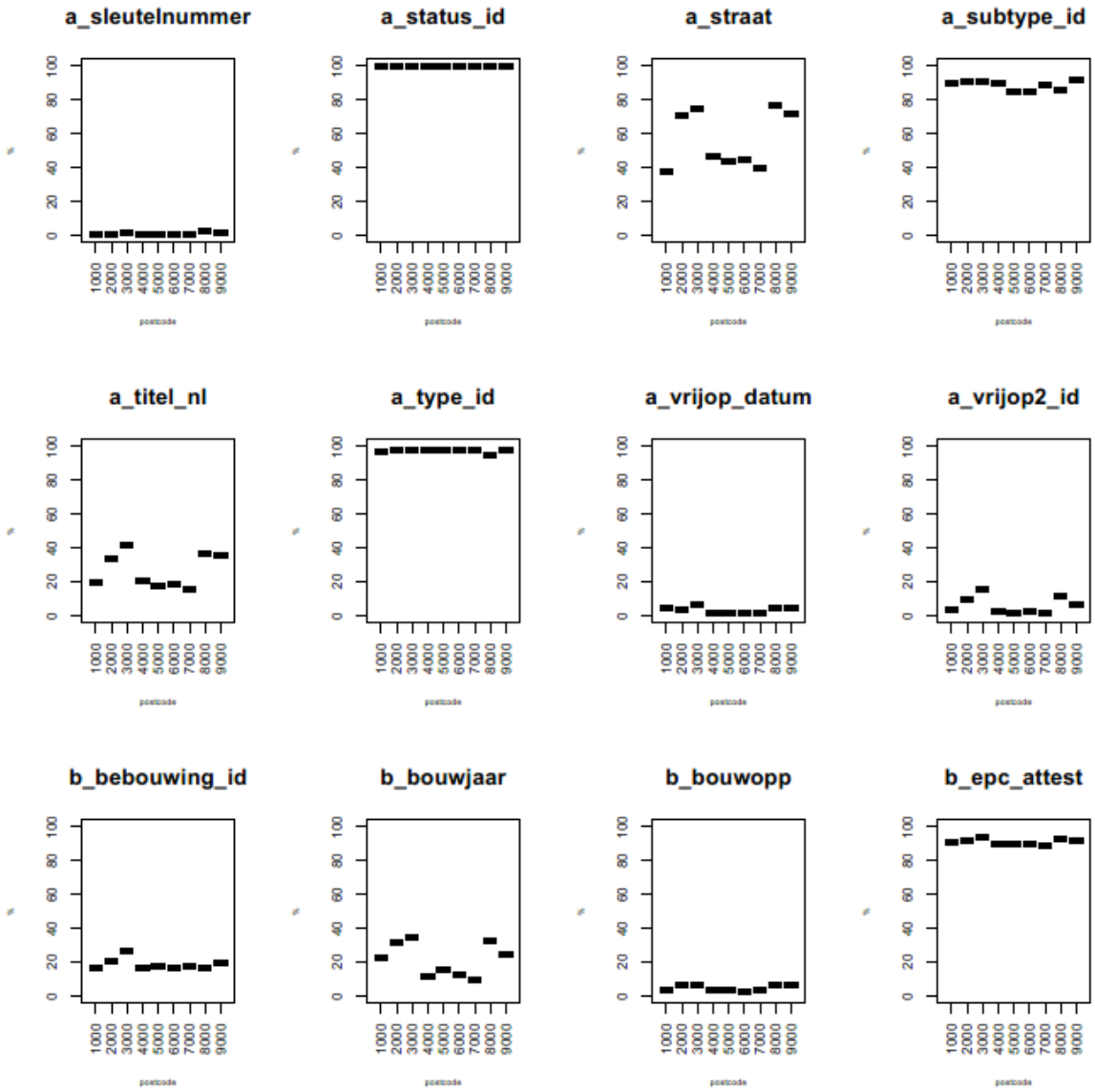

a_geo_precision
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

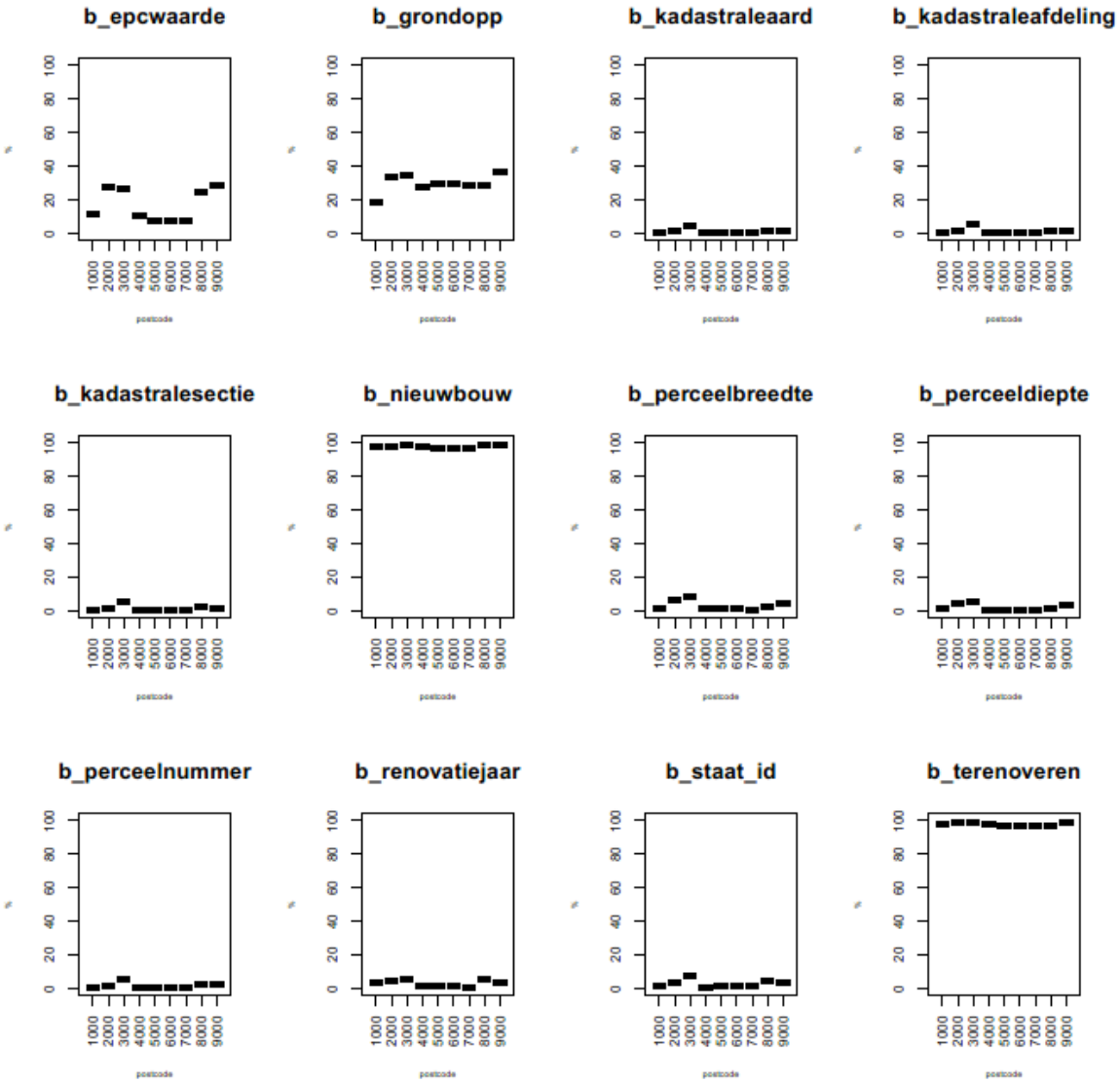

a_land_id
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

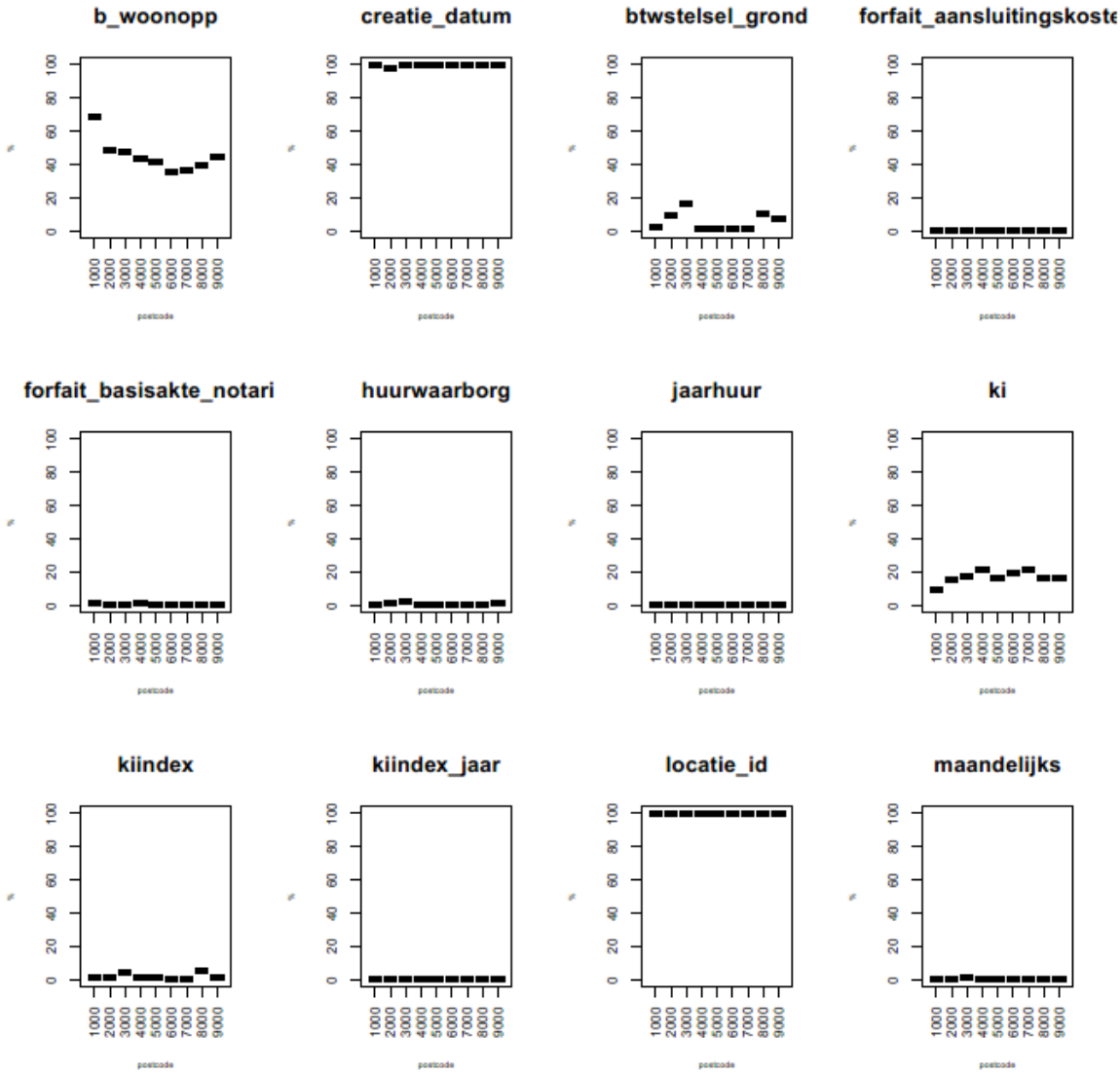

a_nummer
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

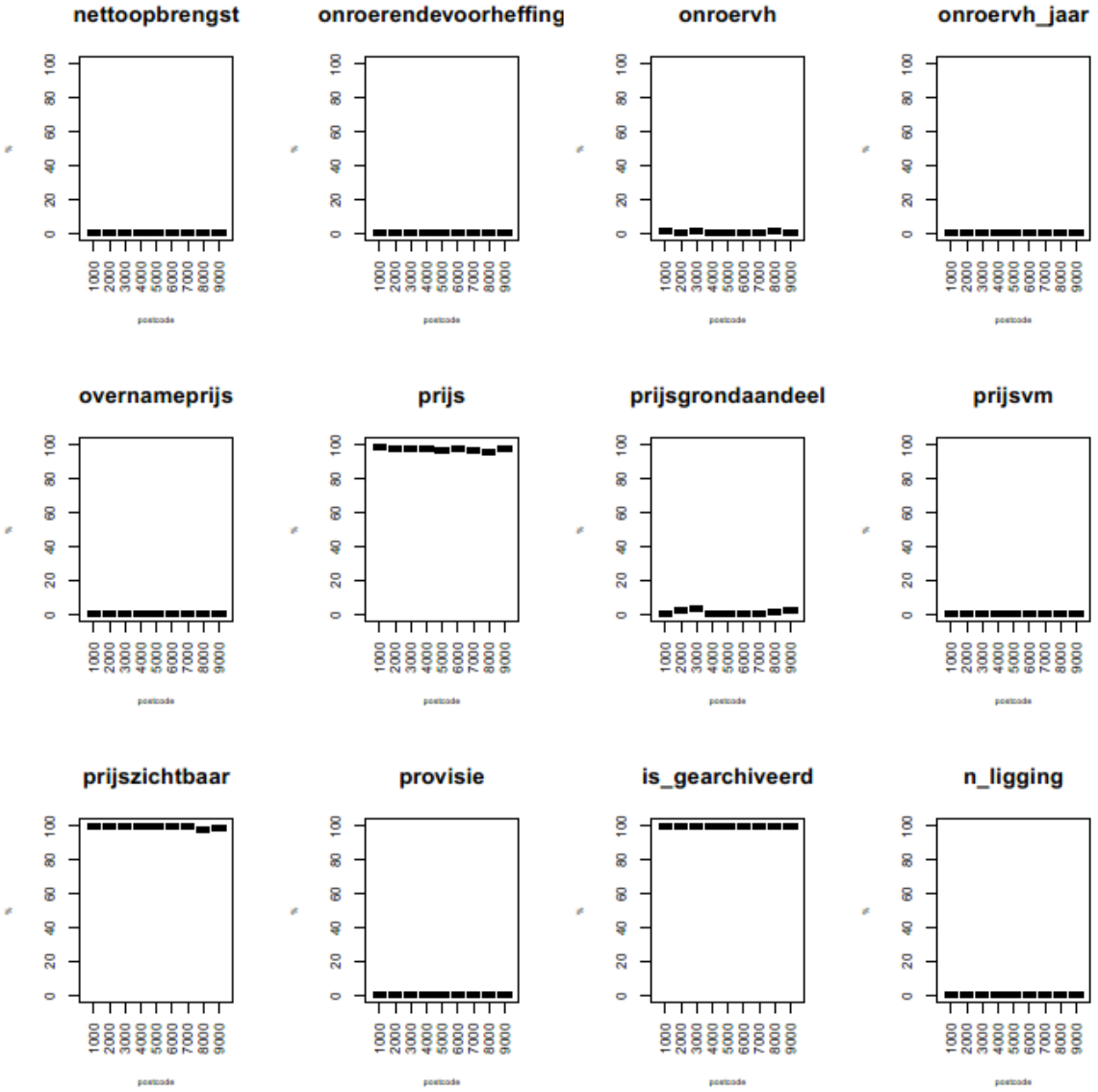

a_omgeving_id
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

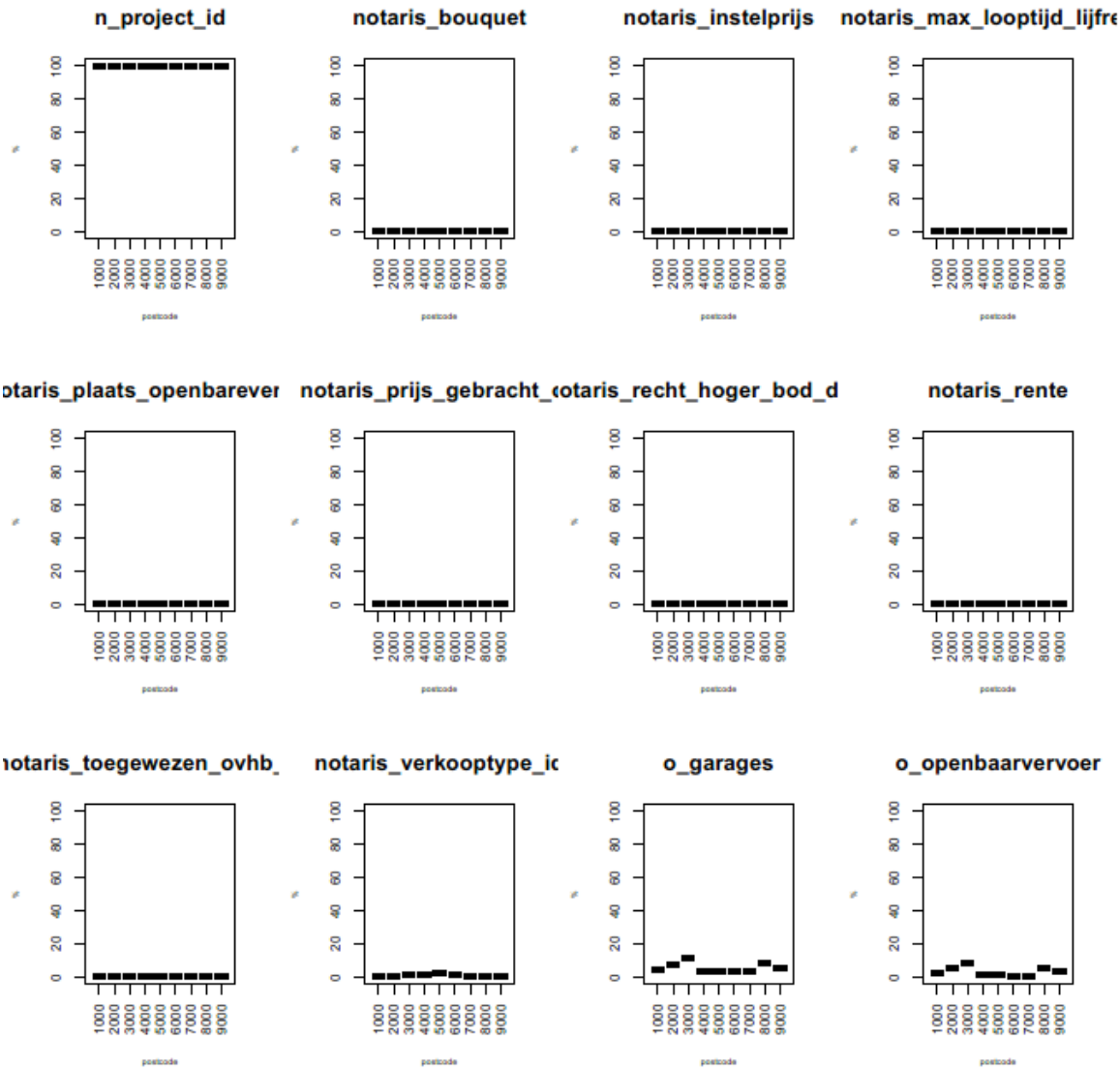

a_postcode
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

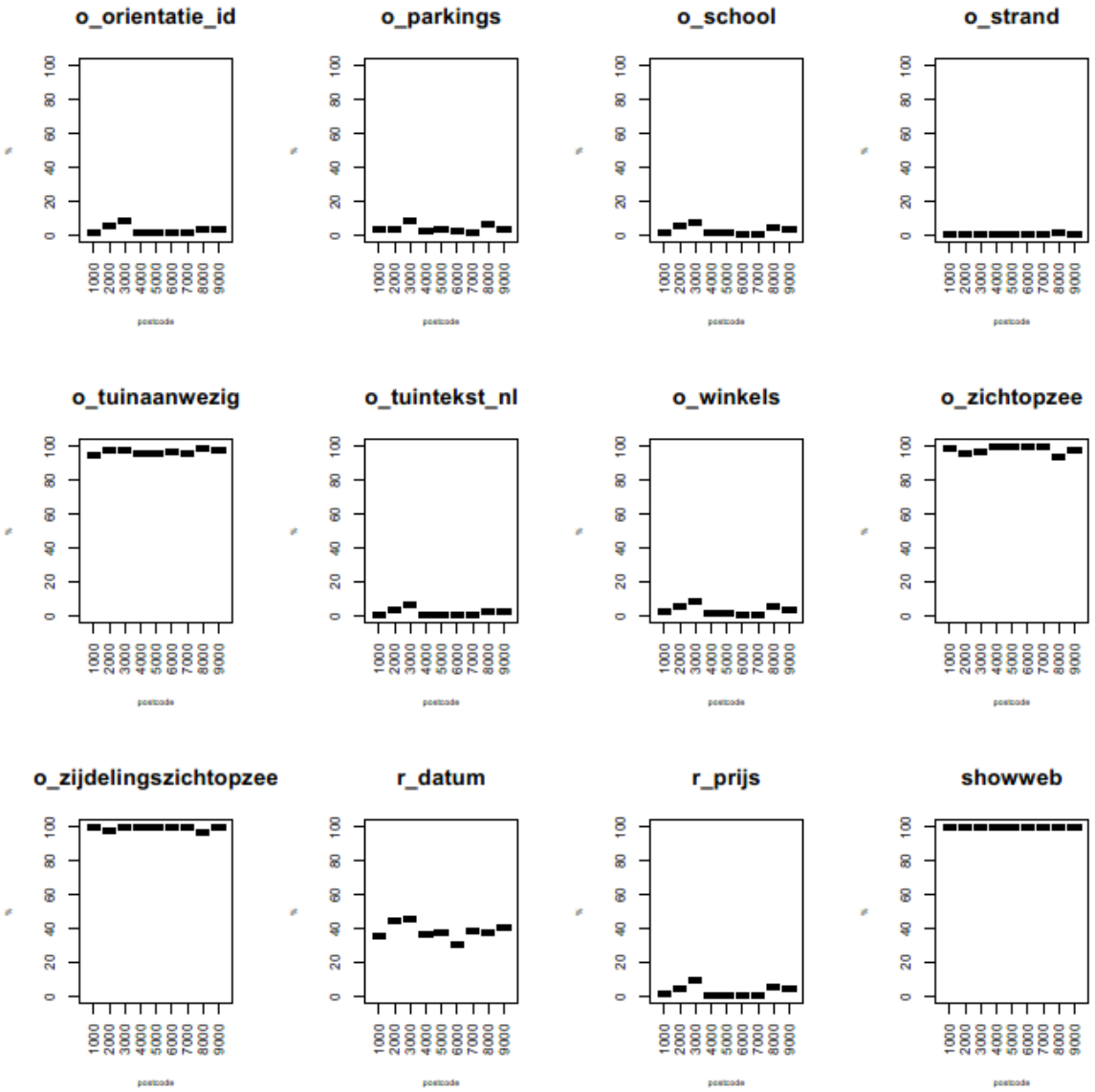

a_sleutelnummer
a_status_id
a_straat
a_subtype_id
a_titel_nl
a_type_id
a_vrijop_datum
a_vrijop2_id
b_bebouwing_id
b_bouwjaar
b_bouwopp
b_epc_attest
%
0 20 40 60 80 100
1000 2000 3000 4000 5000 6000 7000 8000 9000
postcode

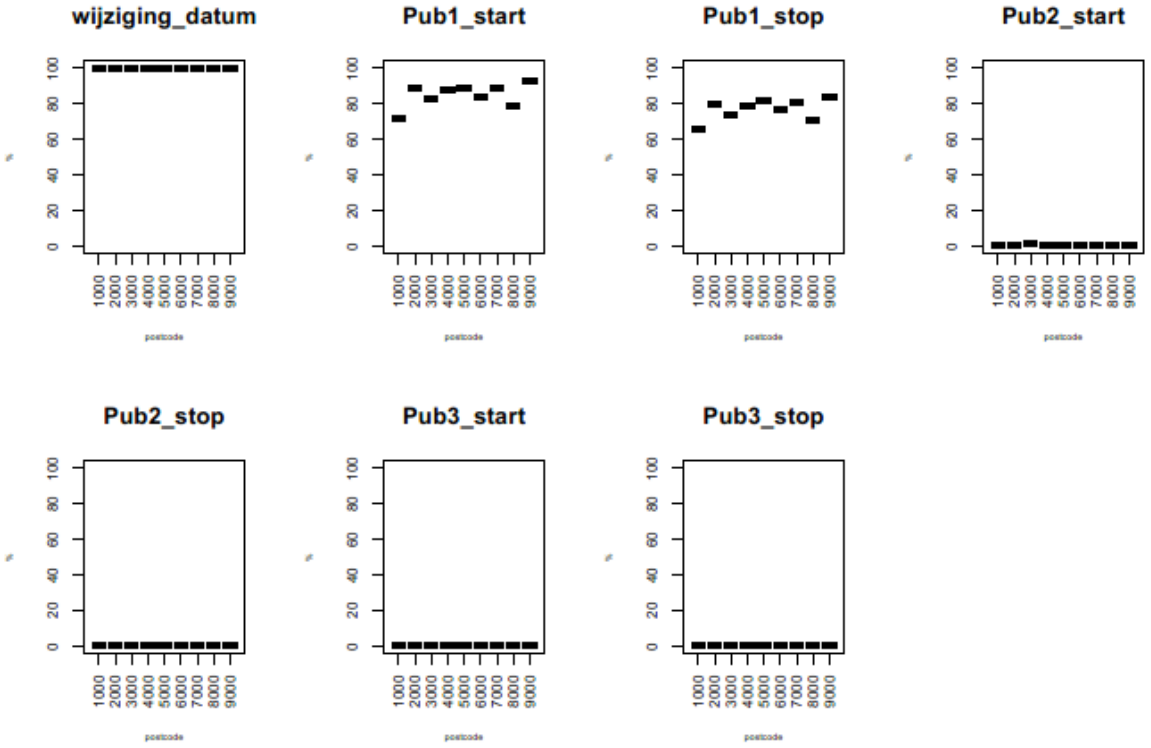

b_epcwaarde
b_grondopp
b_kadastraleaard
b_kadastraleafdeling
b_kadastralesectie
b_nieuwbouw
b_perceelbreedte
b_perceeldiepte
b_perceelnummer
b_renovatiejaar
b_staat_id
b_terenoveren
%
0 20 40 60 80 100
1000 2000 3000 4000 5000 6000 7000 8000 9000
postcode


b_woonopp
creatie_datum
btwstelsel_grond
forfait_aansluitingskoste
forfait_basisakte_notari
huurwaarborg
jaarhuur
ki
kiindex
kiindex_jaar
locatie_id
maandelijks
%
postcode
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000

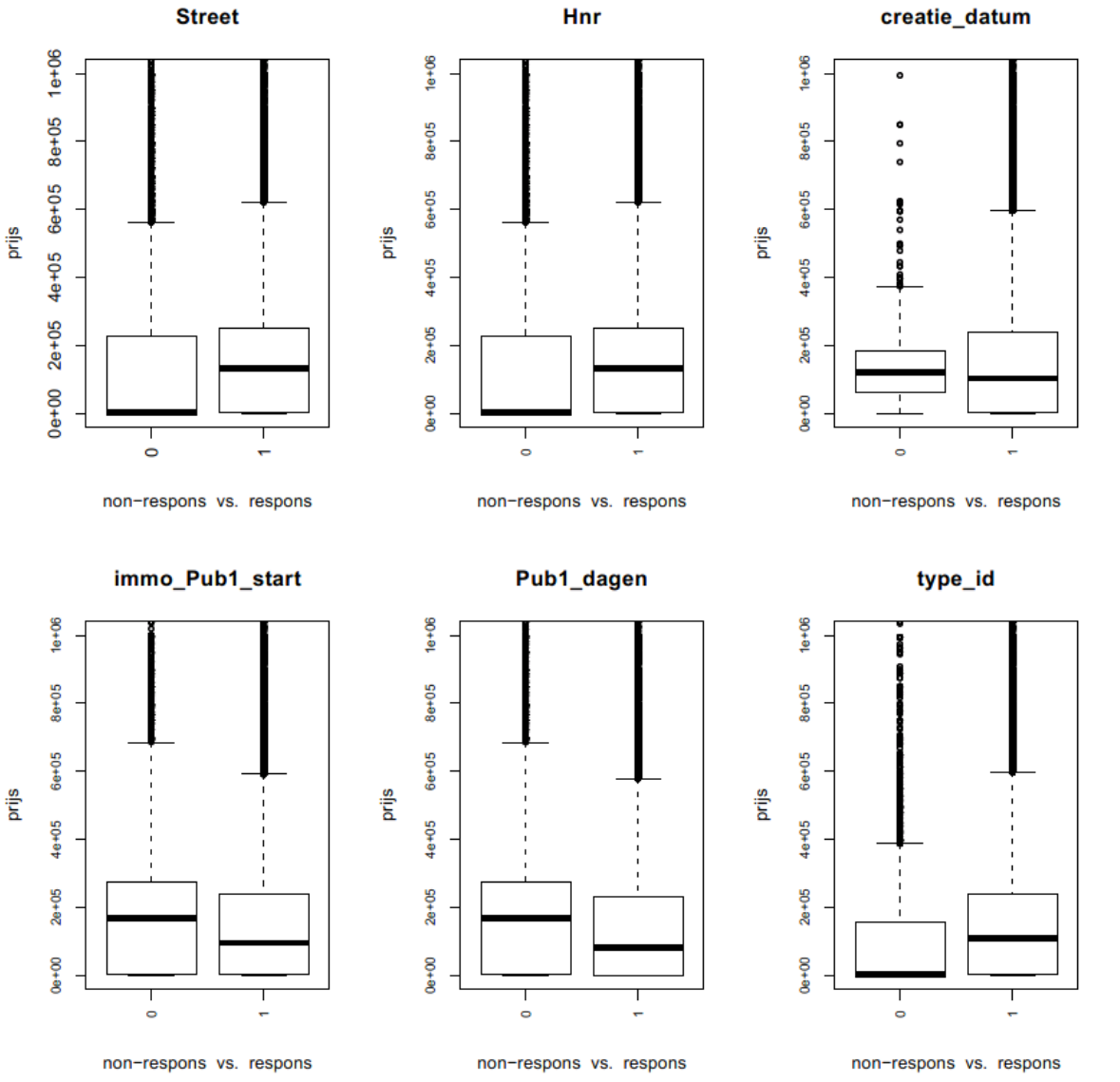

nettoopbrengst
onroerendevoorheffing
onroervh
onroervh_jaar
overnameprijs
prijs
prijsgrondaandeel
prijsvm
prijszichtbaar
provisie
is_gearchiveerd
n_ligging
%
0 20 40 60 80 100
1000 2000 3000 4000 5000 6000 7000 8000 9000
postcode

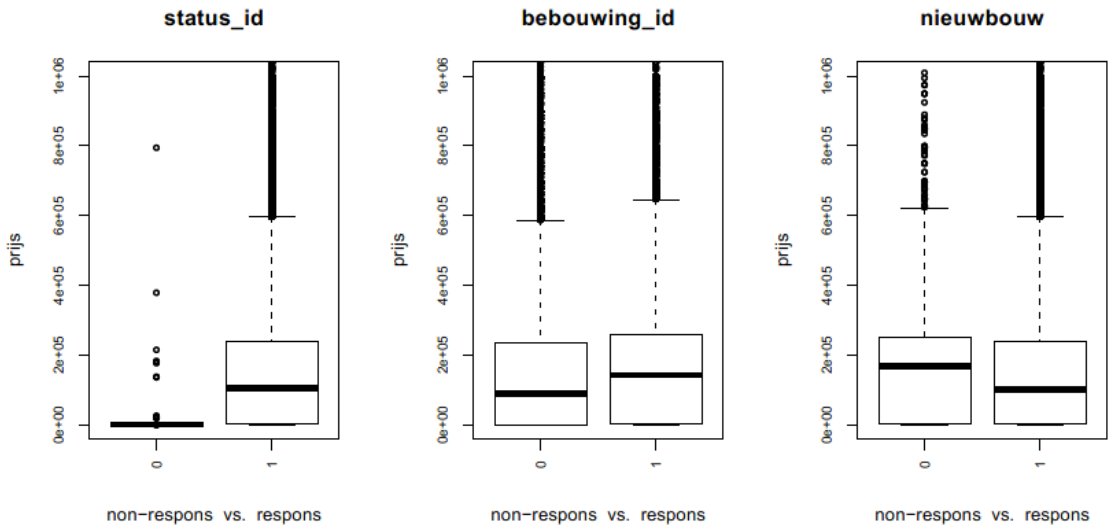

n_project_id
notaris_bouquet
notaris_instelprijs
notaris_max_looptijd_lijfre
otaris_plaats_openbarever
notaris_prijs_gebracht_c
otaris_recht_hoger_bod_d
notaris_rente
notaris_toegewezen_ovhb_
notaris_verkooptype_ic
o_garages
o_openbaarvervoer
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

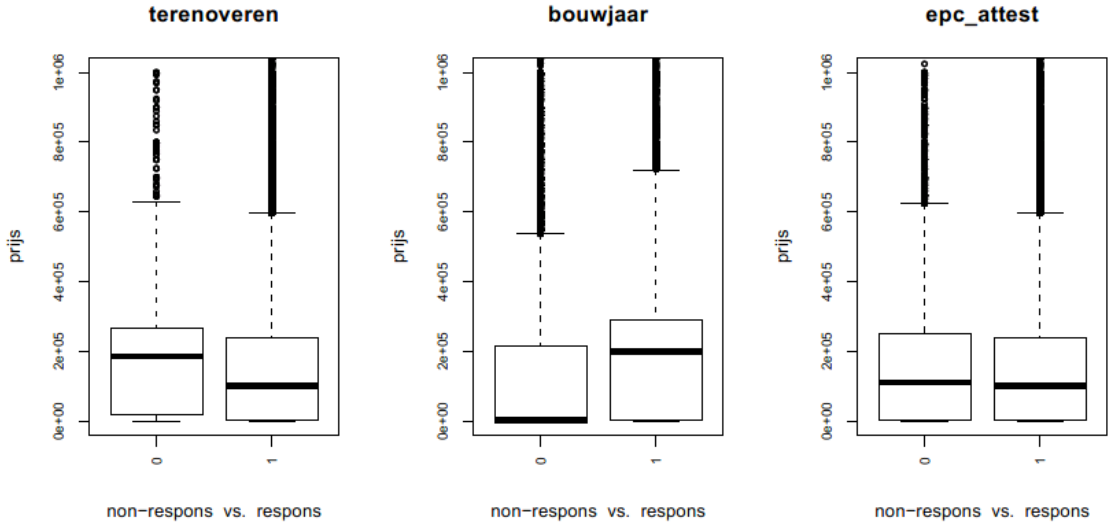

o_orientatie_id
o_parkings
o_school
o_strand
o_tuinaanwezig
o_tuintekst_nl
o_winkels
o_zichtopzee
o_zijdelingszichtopzee
r_datum
r_prijs
showweb
%
0 20 40 60 80 100
1000 2000 3000 4000 5000 6000 7000 8000 9000
postcode

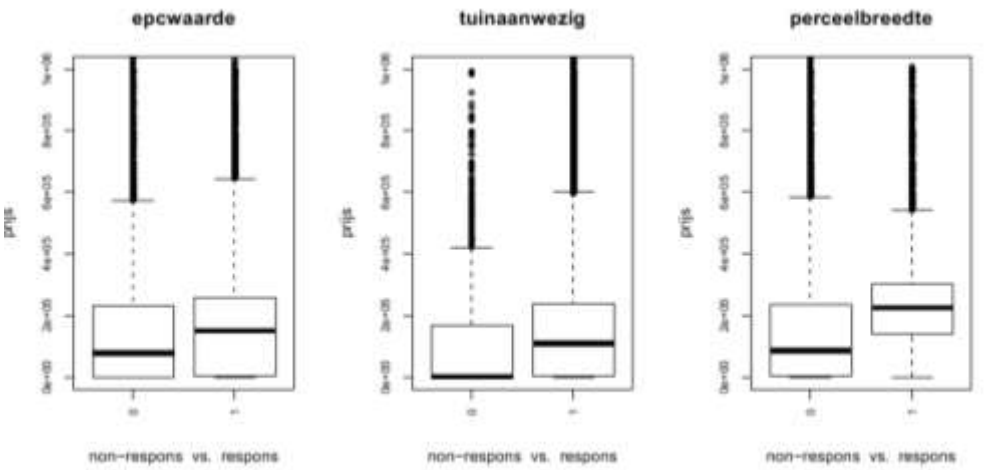

wijziging_datum
Pub1_start
Pub1_stop
Pub2_start
Pub2_stop
Pub3_start
Pub3_stop
%
0
20
40
60
80
100
1000
2000
3000
4000
5000
6000
7000
8000
9000
postcode

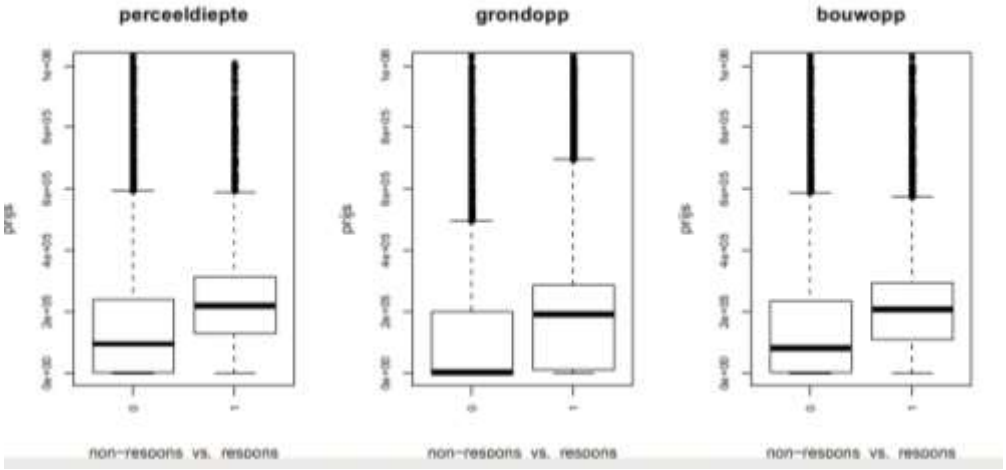
Bijlage 2: Boxplots

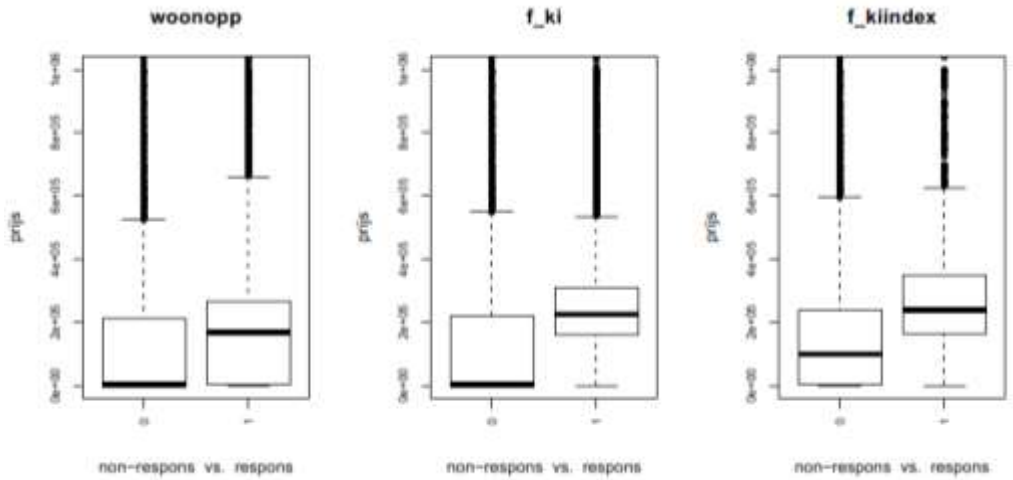



status_id
bebouwing_id
nieuwbouw
prijs
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
0
1
non−respons vs. respons


terenoveren
bouwjaar
epc_attest
prijs
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
0
1
non−respons vs. respons


epcwaarde
tuinaanwezig
perceelbreedte
prijs
non-respons vs. respons

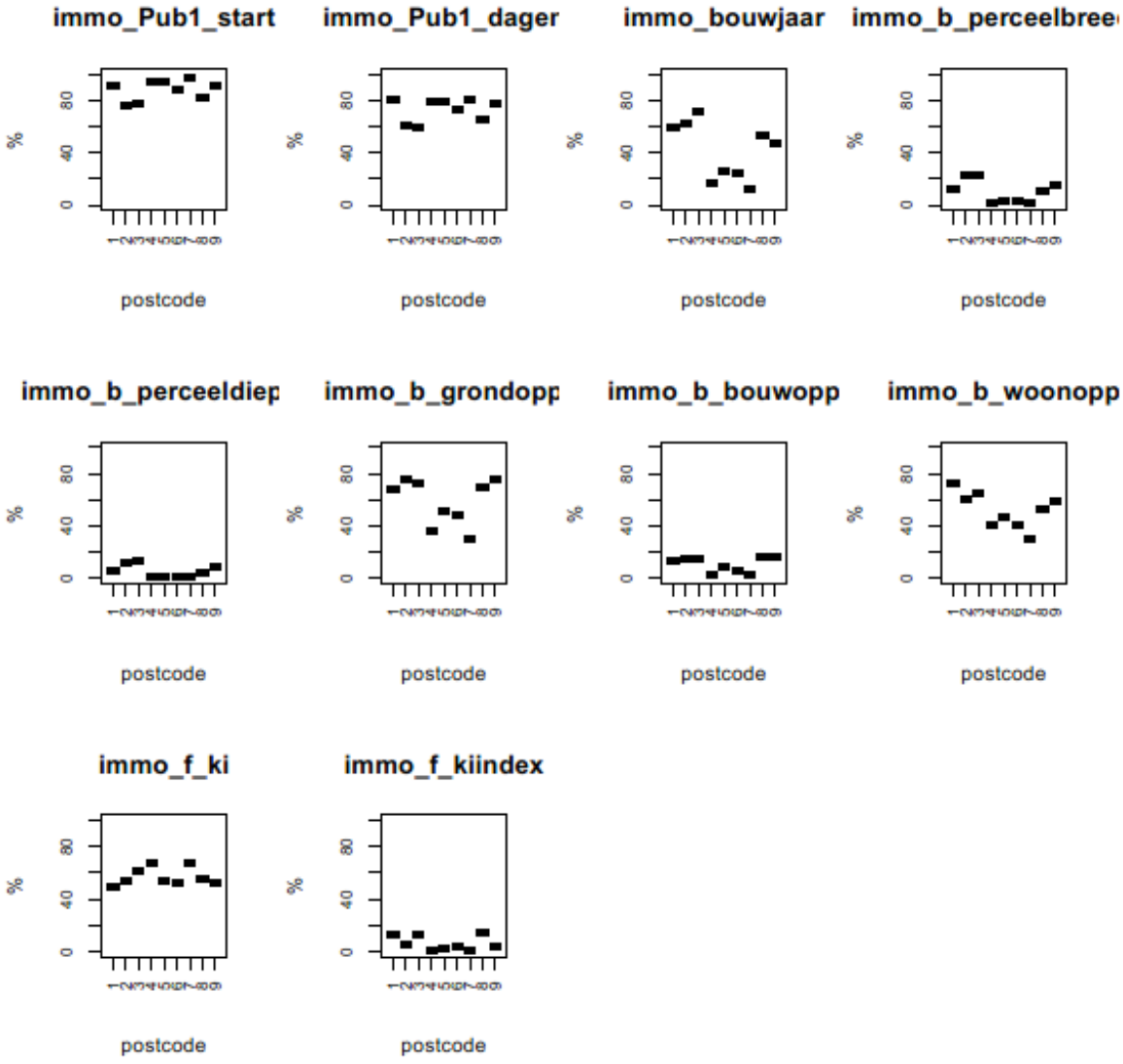

perceeldiepte
grondopp
bouwopp
prijs
non-respons vs. respons


woonopp
f_ki
f_kiindex
prijs
non-respons vs. respons

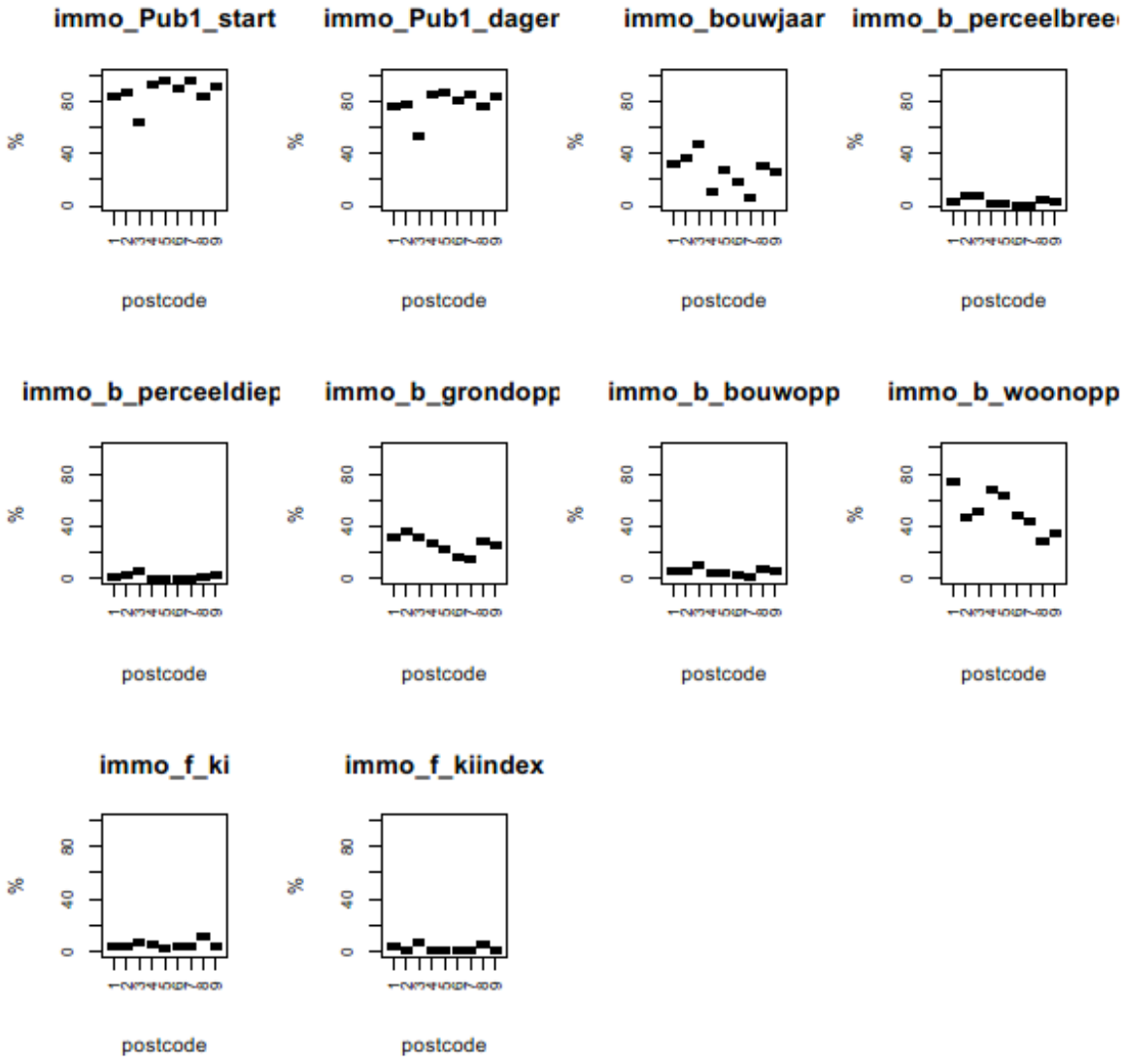
Bijlage 3: Non-respons van de analyse-set

Bijgevoegde grafieken geven weer hoeveel procent van de gegevens zijn ingevuld voor de verschillende variabelen en dit per postcode en geldig op de analyse-set. Er wordt een bijkomend onderscheid gemaakt naar non-response voor woning te koop of te huur, appartement te koop of te huur en grond

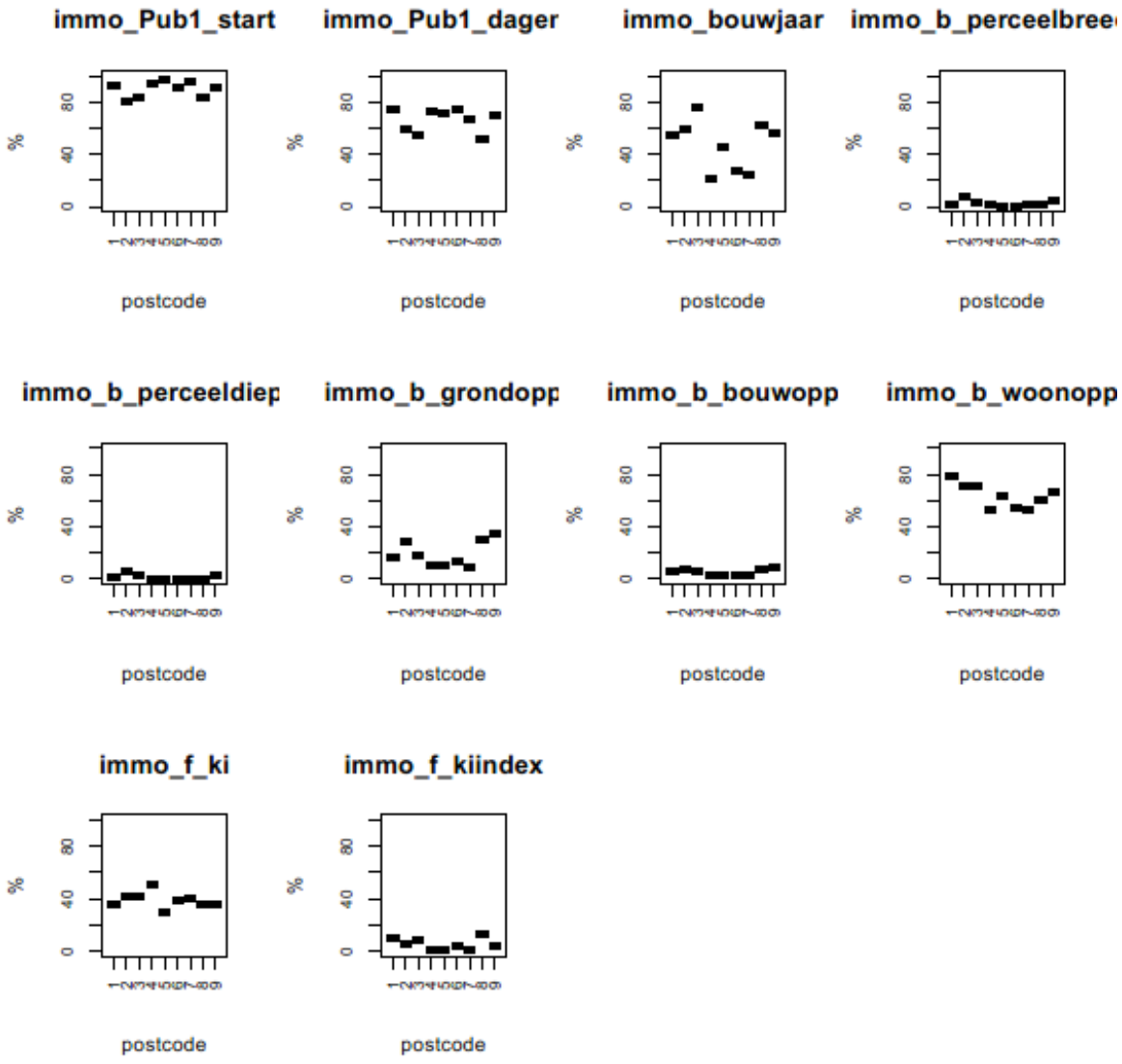
Non-respons Woning te koop



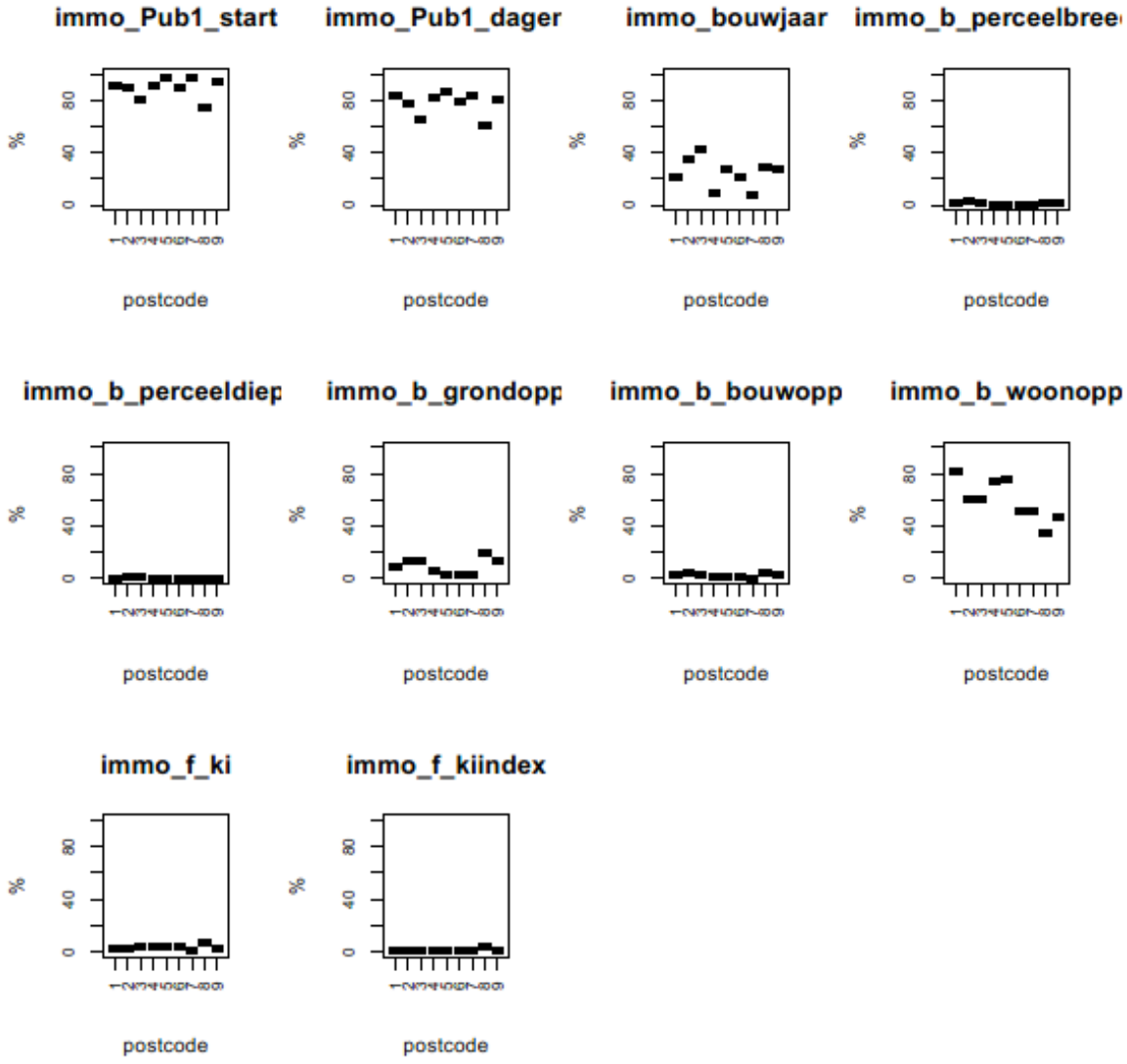
Non-respons Woning te huur



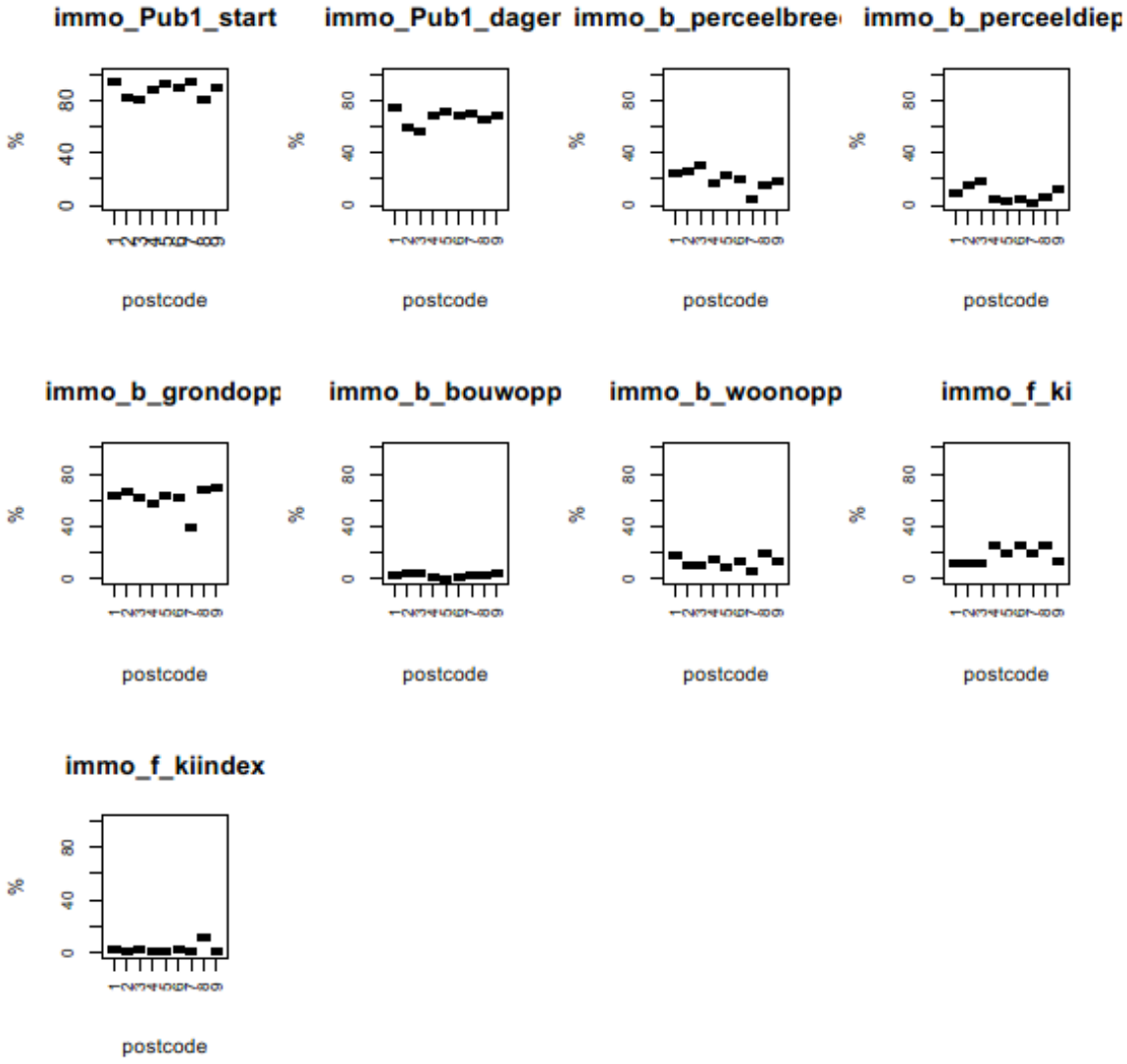
Non-respons appartement te koop



Non-respons appartement te huur



Non-respons grond



Bijlage 4: Lijst van extra velden

• immo_Pub1_start_year	wat?	jaar van online plaatsen
	hoe?	o.b.v. immo_Pub1_start – vb.net functie year()
• immo_status	wat?	huur- of koopmarkt
	hoe?	op basis van het veld a_status_id
• immo_grondprijs_m2	wat?	de grondprijs per m²
	hoe?	immo_f_prijs delen door de grondoppervlakte. In de eerste plaats o.b.v. het veld immo_b_grondopp en anders o.b.v. het veld immo_ParcelArea
• Cs012011	wat?	ID van de statistische sector
	hoe?	spatial join met de statistische sectoren
• Nis_012011	wat?	NIS-code (ID gemeente) van de statistische sector
	hoe?	spatial join met de statistische sectoren
• Sec012011	wat?	ID van de statistische sector binnen de gemeente
	hoe?	spatial join met de statistische sectoren
• Sector_nl	wat?	Nederlandstalige naam van de statistische sector
	hoe?	spatial join met de statistische sectoren
• Gemeente	wat?	Gemeente waarin de statistische sector ligt
	hoe?	spatial join met de statistische sectoren
• Prov_nl	wat?	Provincie waarin de statistische sector ligt
	hoe?	spatial join met de statistische sectoren
• gridcode	wat?	Combinatiewaardie van de knooppuntwaarde en voorzieningen
	hoe?	spatial join met synthesekaart_naturalbreaks_2015_metbus (Verachtert et.aL, 2016[16])
• rsv	wat?	verstedelijkingsgraad voor Vlaanderen en Brussel in 9 categoriën
	hoe?	spatial join met tabel ontvangen van de studiedienst van de Vlaamse Regering
• urbanisatie	wat?	verstedelijkingsgraad voor Vlaanderen en Brussel in 6 categoriën
	hoe?	spatial join met tabel ontvangen van de studiedienst van de Vlaamse Regering

[16] Verachtert, E., I. Mayeres, L. Poelmans, M. Van der Meulen, M. Vanhulsel, G. Engelen (2016), Ontwikkelingskansen op basis van knooppuntwaarde en nabijheid voor¬zieningen, eindrapport, studie uitgevoerd in opdracht van Ruimte Vlaanderen.