Centraal Planbureau

Investeren in de arbeidsmarktintegratie van statushouders

Vanaf 1 januari 2022 krijgen gemeenten weer een regierol bij de inburgering en begeleiding naar werk van statushouders. De intensiteit waarmee gemeenten investeren kan afhangen van de verwachting hoe snel een statushouder weer verhuist en of een statushouder elders betere baankansen heeft. Het verhuisgedrag van statushouders is echter lastig te voorspellen. Ook blijken statushouders niet naar gemeenten met hogere baankansen te verhuizen.

De verhuisgeneigdheid van statushouders is berekend met behulp van machine-learning technieken. Van bijna niemand is met meer dan 50% zekerheid te voorspellen of diegene verhuist. Het advies aan gemeenten is dan ook: ga ervan uit dat ze blijven.

CPB Achtergronddocument

Cécile Magnée, Mark Kattenberg, Gijs Roelofs

september 2021

Inleiding

De publicatie 'Investeren in de arbeidsmarktintegratie van statushouders' onderzoekt de financiële risico's voor en door gemeenten onder de Wet Inburgering 2022 (hierna: WI2022). Onder de WI2022 komt een groot deel van de inburgeringstaken weer bij gemeenten te liggen. Dit achtergronddocument beschrijft achtereenvolgens de data en methoden die in de publicatie gebruikt zijn. Daarbij worden ook een aantal beschrijvende statistieken en aanvullende resultaten getoond die niet in de publicatie zijn opgenomen.

Data

In dit onderzoek bekijken we twee cohorten. Het eerste cohort bestaat uit statushouders die tussen 1995 en 1999 in Nederland gevestigd zijn. Het tweede cohort bestaat uit statushouders die tussen 2014 en 2019 naar Nederland gekomen zijn.

Variabelen

Voor beide cohorten hebben we maandelijkse gegevens van de volgende variabelen:

- **Geslacht**
- **Leeftijd**
- **Herkomstland**
- **COA-locaties**
  Voor het 1995-cohort hebben we registratiegegevens van de verschillende COA-locaties die toen actief waren, waardoor we kunnen matchen met de GBA-gegevens van een statushouder. Voor het nieuwe cohort is in de data al aangegeven welke GBA-inschrijvingen COA-locaties betreffen.
- **Woongemeente**
  De GBA-inschrijvingen laten zien in welke gemeente iemand op een bepaalde tijd ingeschreven staat.
- **Uitplaatsing**
  Het moment dat iemand van een COA-locatie naar een niet-COA GBA-inschrijving verandert, zien wij als het zogenoemde 'uitplaatsingsmoment' waarop iemand van een opvanglocatie naar een woning in een Nederlandse gemeente verhuist. We bekijken het verhuisgedrag en bijstandsgebruik vanaf dit moment van eerste vestiging. De achtergrondvariabelen op het moment van eerste plaatsing in een Nederlandse gemeente worden gebruikt in de voorspelmodellen.
- **Huishoudsamenstelling**
  De huishouddata bevat informatie over het hebben van een partner en/of kinderen, het type huishouden, aantal gezinsleden en leeftijd van de kinderen.
- **Bijstand**
  Vanaf 1999 zijn er gegevens over bijstandsgebruik. Dit betekent dat voor het jaren-negentig-cohort geen bijstandsgebruik bekend is voor de beginperiode van de statushouders die voor 1999 een eerste woning in Nederland kregen.
- **Sociaaleconomische positie**
  Aan de hand van de belangrijkste inkomensbron wordt de sociaaleconomische positie van een persoon op een bepaald moment bepaald. We clusteren de informatie in drie groepen: (1) werkend/inkomen, (2) bijstand/uitkering/pensioen en (3) geen inkomen.

Selecties

Om het verhuisgedrag en sociale voorzieningengebruik goed in kaart te brengen, maken we een aantal dataselecties en keuzes. Allereerst nemen we asielmigranten die nooit op een adres anders dan een COA-locatie ingeschreven hebben gestaan niet mee, omdat de kans aanzienlijk is dat zij geen verblijfsvergunning ontvangen hebben en zij nooit gebruik hebben gemaakt van gemeentelijke sociale voorzieningen. Ook de enkele statushouders voor wie belangrijke achtergrondkarakteristieken ontbreken (geslacht, leeftijd en herkomstland) zijn geen onderdeel van de onderzoekspopulatie. Verder kijken we niet naar remigratie. Zodra iemand Nederland verlaten heeft, wordt eventueel verhuisgedrag binnen Nederland en socialevoorzieningengebruik in de toekomst niet meer meegenomen.[1] We bekijken alleen statushouders die ouder dan achttien waren op het moment van uitplaatsing, omdat deze groep recht heeft op bijstand.

Om de uitkomsten van de voorspelmodellen voor verhuisgedrag en bijstandsgebruik te kunnen combineren om tot inzichten te komen over de interactie tussen deze variabelen, maken we verder de keuze om alle analyses te doen voor de personen voor wie zowel het bijstandsgebruik als het verhuisgedrag bekend zijn drie jaar na uitplaatsing. Na deze laatste selectie zijn er voor het jarennegentigcohort 31.831 personen over en voor het recente cohort 60.026.[2]

Tabel 1 Achtergrondkenmerken van de cohorten

		Cohort 1995			Cohort 2014		
		Aantal	Gem.	Std. Dev.	Aantal	Gem.	Std. Dev.
Man		31.831	0,598	0,490	60.026	0,592	0,491
Leeftijd		31.831	33,809	11,356	60.026	32,405	11,199
Herkomstland	Syrië	31.831	0,013	0,112	60.026	0,638	0,480
	Iran	31.831	0,081	0,272	60.026	0,024	0,154
	Irak	31.831	0,247	0,431	60.026	0,028	0,165
	Somalië	31.831	0,057	0,232	60.026	0,014	0,116
	Afghanistan	31.831	0,216	0,412	60.026	0,018	0,133
	Voorm. Joegoslavië	31.831	0,128	0,335	60.026	0,000	0,016
	Voorm. Sovjet Unie	31.831	0,052	0,222	60.026	0,003	0,058
	Overig	31.831	0,206	0,404	60.026	0,274	0,446
Kind/partner		29.857	0,639	0,480	55.488	0,643	0,479
Leeftijd partner		9467	36,391	10,327	17.322	36,951	11,020
Leeftijd jongste kind		13.057	7,370	6,500	25.560	8,080	6,727
Leeftijd oudste kind		14.365	10,850	7,152	26.714	12,867	7,606

Tabel 1 geeft een beschrijving van de belangrijkste achtergrondkenmerken van de twee cohorten. De cohorten verschillen sterk in herkomstland: het nieuwe cohort wordt gedomineerd door Syriërs. Verder zijn de kinderen van het nieuwe cohort gemiddeld gezien iets ouder. De cohorten verschillen niet op basis van geslacht,

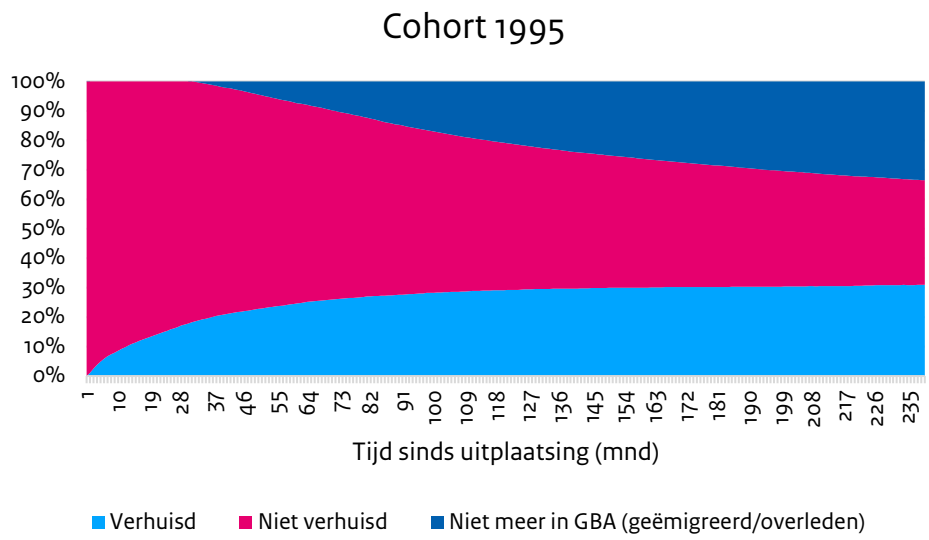
[1] Leerkes et al. (2019) laten zien dat voor het jarennegentigcohort ongeveer 1 op de 10 migranten die eind 2015 in Nederland woonden, een periode uitgeschreven was geweest uit het bevolkingsregister. Het verhuisgedrag en bijstandsgebruik van deze groep na hervestiging in Nederland zijn geen onderdeel van het onderzoek?.

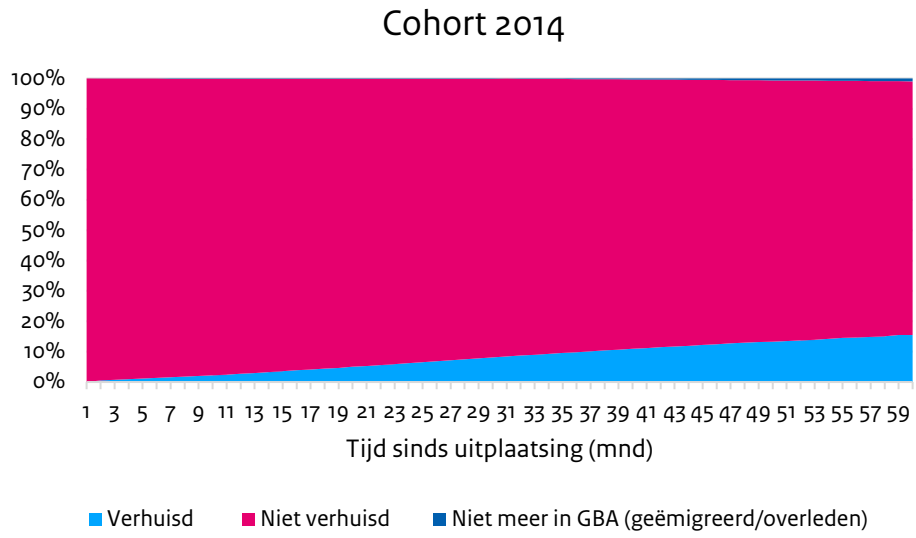
[2] Technisch gezien zijn diegenen die pas in 2018 of 2019 uitgeplaatst zijn nog niet drie jaar in Nederland. Het cohort dat wij 2014-2019 noemen, bestaat dus uit diegenen die vóór 2018 uitgeplaatst zijn.

leeftijd, wonen met een partner en/of kind[3] en de leeftijd van de partner. Verder heeft niet iedereen een partner en/of kinderen.

Als we vervolgens het verhuisgedrag vanaf het moment van uitplaatsing in kaart brengen (figuur 1), zien we dat het jarennegentigcohort veel sneller verhuist vergeleken met het nieuwe cohort. Zo is van het 1995-cohort vijf jaar na uitplaatsing ongeveer een kwart verhuisd, terwijl ongeveer 15% van het nieuwe cohort binnen vijf jaar verhuisd is.

Figuur 1: Verhuisgedrag sinds uitplaatsingsmoment





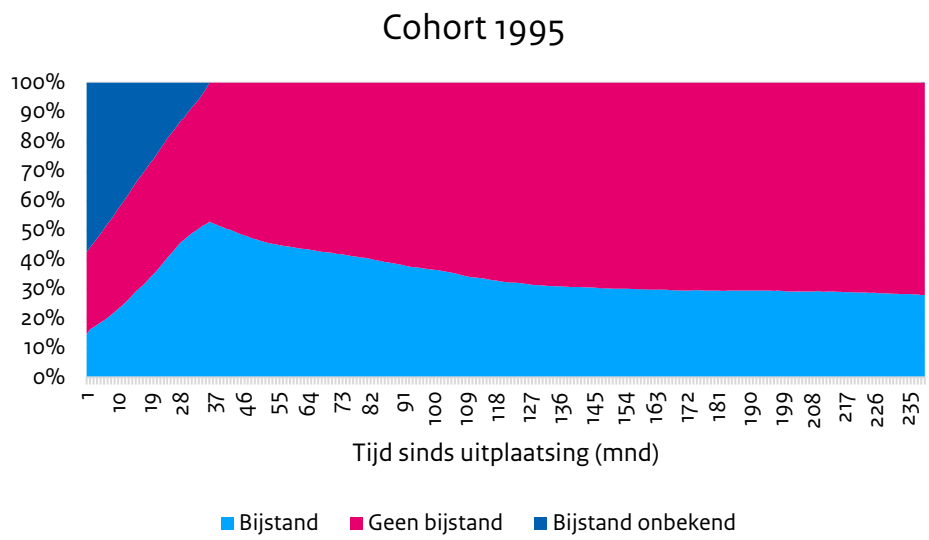
Figuur 2 laat het bijstandsgebruik sinds het moment van uitplaatsing zien. Voor statushouders die vóór 1999 uitgeplaatst zijn, is het bijstandsgebruik vlak na uitplaatsing onbekend, omdat bijstandsgegevens pas vanaf 1999 beschikbaar zijn. Aangezien we alleen de statushouders meenemen van wie we drie jaar na uitplaatsing

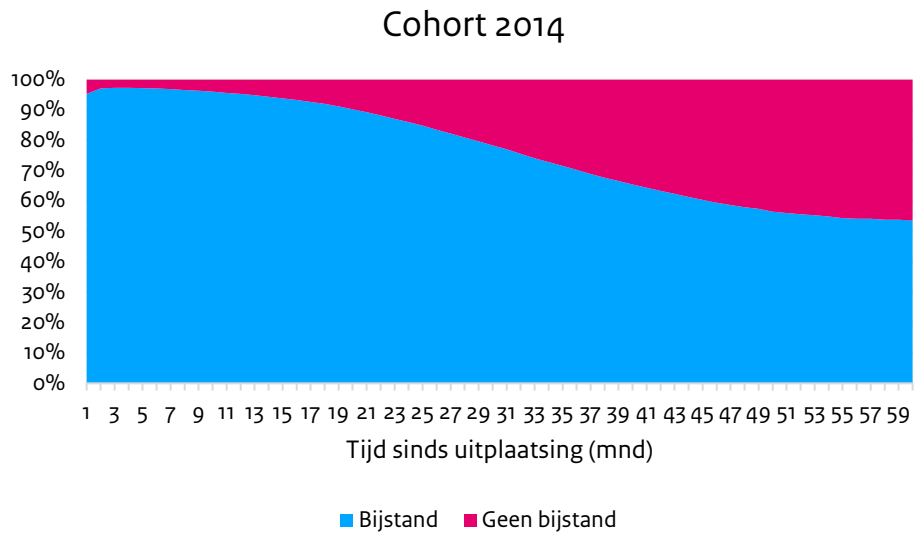
[3] Om een beeld te krijgen of personen met een gezin migreren, direct of via gezinsmigratie, bekijken we wie binnen een jaar na uitplaatsing met een kind en/of partner woont. We kiezen hier voor een jaar, aangezien dit overeenkomt met de termijn waarbinnen gezinsmigratie moet plaatsvinden.

zowel verhuisgedrag als bijstandsgebruik weten, is het verloop van de onbekende categorie een gevolg van deze keuze. Het is opvallend dat voor dit cohort het bijstandsgebruik niet heel hoog is vlak na uitplaatsing. Dit heeft waarschijnlijk meerdere oorzaken. Allereerst werd het bijstandsgebruik nog niet met terugwerkende kracht bijgewerkt door het Centraal Bureau voor de Statistiek (hierna CBS). Verder waren er destijds voorliggende voorzieningen waarvan sommige statushouders mogelijk gebruikmaakten, zoals de toenmalige Regeling Opvang Asielzoekers. Hierdoor stonden zij mogelijk nog niet geregistreerd in de bijstandsdata, maar ontvingen zij wel een andere vorm van financiële steun.

Het 2014-cohort laat een patroon in het bijstandsgebruik zien dat men zou verwachten: bijna iedereen ontvangt bijstand op het moment van uitplaatsing en dit aantal loopt af naarmate statushouders langer uitgeplaatst zijn (bijvoorbeeld doordat de statushouder of zijn/haar partner werk heeft gevonden).

Figuur 2: Bijstandsgebruik sinds uitplaatsingsmoment





Methodes

In het onderzoek? bekijken wij de kansen en risico's voor de arbeidsmarktintegratie van statushouders onder de WI2022. De eerste vraag is of de nieuwe inburgeringswet risico's met zich meebrengt voor onderinvestering of overinvestering door gemeenten bij de inburgering en arbeidsmarktintegratie van nieuwkomers. De tweede vraag is of er financiële risico's zijn voor gemeenten door de bekostigingssystematiek.

Er ontstaat een risico op onderinvestering door gemeenten als gemeenten niet zelf de vruchten plukken – bijvoorbeeld in de vorm van lagere uitgaven aan sociale voorzieningen in de toekomst - van investeren in integratie en inburgering. Als een gemeente weet dat een statushouder geneigd is om te verhuizen, heeft deze gemeente een financiële prikkel om de investering uit te stellen, totdat duidelijk is dat die statushouder waarschijnlijk toch zal blijven. De verhuiskans van een statushouder geeft dus een prikkel aan een gemeente om wel of niet snel te beginnen met integratie. Dit kan een risico met zich meebrengen op onderinvestering in mensen waarvan gemeenten verwachten dat ze niet lang zullen blijven. We onderzoeken dit risico op onderinvestering aan de hand van een voorspelmodel dat we hieronder beschrijven, en vinden voor een beperkte groep een risico op onderinvestering.

De verhuisbewegingen van statushouders leiden mogelijk ook tot financiële risico's voor gemeenten indien er sprake is van een uitsortering, waarbij voorzieningengebruikers in de loop der tijd clusteren in bepaalde gemeenten. Een verhoogd gebruik van bijstand door zo'n uitsortering valt buiten de invloedssfeer van gemeenten, maar wordt niet gedekt via het verdeelmodel voor de bijstandsbudgetten (BUIG). Hieronder beschrijven we hoe we deze uitsortering in kaart gebracht hebben en zien we dat het effect beperkt is en pas op lange termijn statistisch detecteerbaar.

Voorspelmodel

Om te onderzoeken of gemeenten goed kunnen voorspellen wie er verhuist, schatten we de verhuis- en bijstandskansen op basis van observeerbare kenmerken op het moment van uitplaatsing. Als blijkt dat goed te voorspellen is wie er verhuist, bestaat een risico dat gemeenten onderinvesteren in integratietrajecten van de tot verhuizen geneigde statushouders. In dat geval plukt de gemeente immers niet zelf de voordelen (zoals lager bijstandsgebruik in de toekomst) van de investering waardoor de gemeente geneigd kan zijn om de investering uit te stellen of te beperken.

Bij de plaatsing van een statushouders van een opvanglocatie van het Centraal Orgaan opvang Asielzoekers (hierna COA) naar een gemeente, krijgt een gemeente informatie over de statushouder via een overdrachtsdocument. In dit document staan observeerbare kenmerken, zoals geslacht, leeftijd, herkomstland en familiesituatie. Gemeenten kunnen ook aanvullende informatie krijgen zodra zij in gesprek gaan met een statushouder. Uit dit gesprek kan blijken of een statushouder al verhuisplannen heeft, of bepaalde kansen ziet in de huidige gemeente of juist elders.

Wij proberen dit model na te bootsen, met het voorbehoud dat wij alleen de objectieve kenmerken kennen en geen eventuele informatie uit gesprekken tussen gemeenten en statushouders beschikbaar hebben. We schatten dus een model met minder informatie dan een gemeente in potentie heeft, waardoor een gemeente mogelijk iets beter kan voorspellen dan ons model. Echter, een gemeente kent een beperking aan het aantal observaties: ze hebben alleen informatie en ervaring met de statushouders in de eigen gemeente of regio. Ons model heeft wat dat betreft dus meer informatie; wij kunnen op basis van het hele cohort de

verhuisgeneigdheid schatten. Bij dit model moet worden bedacht dat mensen het moeilijk vinden om complexe beslissingen, zoals de beslissing of iemand verhuist, goed in te schatten.[4]

Om deze verschillen aan informatie goed te gebruiken voor het voorspelmodel schatten we de verhuiskans op basis van het hele cohort, en voor de gemeenten waar ten minste honderd statushouders zijn uitgeplaatst, ook apart op gemeenteniveau. We interpreteren de uitkomsten van de schattingen op gemeenteniveau als een robuustheidscheck op het nationale model.

Schattingstechnieken

We gebruiken een ensemble-schatter om de kans op verhuizen te bepalen. De ensemble-schatter combineert de voorspellingen van twee *machine learning*-modellen, een lasso logistische regressie en een *random forest*. De lasso logistische regressie lijkt sterk op een normale logistische regressie, maar kan *out of sample* mogelijk beter voorspellen omdat ze minder gevoelig is voor ruis in de data. Het *random forest* verschilt sterk van een logistische regressie en is veel flexibeler. Verderop gaan we dieper in op de lasso logistische regressie en het *random forest*.

Machine learning modellen moeten afgesteld ('getraind') worden om goed te voorspellen. Dit trainen gebeurt op basis van de data, maar hierdoor hebben *machine learning* modellen de neiging om patronen in de data te vinden die zich enkel voordoen in de data waarop ze getraind zijn. Als deze patronen niet te veralgemeniseren zijn naar 'nieuwe observaties', zal het model slecht voorspellen voor deze nieuwe data. Dit is extra problematisch omdat een voorspelmodel in de praktijk gebruikt wordt om voorspellingen te maken op basis van 'nieuwe observaties'. Om ervoor te zorgen dat de gevonden patronen generaliseerbaar zijn naar nieuwe observaties, verdelen we de data willekeurig in de twee groepen *dataset A* en *dataset B*. Vervolgens delen we *dataset A* willekeurig op in *train A* en *validatie A* (*train A* bevat 90 procent van de observaties in *dataset A*). We gebruiken de data in *train A* om de *machine learning*-modellen te trainen, en we bepalen welke specificatie het beste is door de verhuiskans te voorspellen voor de observaties in *validatie A* en deze te vergelijken met geobserveerde verhuizingen in *validatie A*. De data om te valideren zijn dus niet gebruikt om het model te trainen. Tot slot gebruiken we het beste model, getraind op observaties in *train A* en geselecteerd op basis van prestaties in *validatie A*, om te voorspellen voor de observaties in *dataset B*. Daarna verdelen we de observaties in *dataset B* op dezelfde wijze over de subsets *train B* en *validatie B* om de beste afstelling van *machine learning*-modellen te vinden en voorspellen we op dezelfde wijze voor de observaties in *dataset A*.

LASSO

LASSO staat voor 'least absolute shrinkage and selection operator' (Tibshirani, 1996). Wij gebruiken de variant van lasso die voorspellingen genereert op basis van een logistische regressie. Deze lasso logistische regressie lijkt op een standaard logistische regressie met het verschil dat de (absolute waarde van de) geschatte coëfficiënten door middel van een L1-norm verkleind worden ('shrinkage'). Voor sommige variabelen zal de coëfficiënt gelijk worden gesteld aan nul, waarmee een variabele dus niet meer meegenomen wordt ('selection'). *Shrinkage* en *selection* zorgen ervoor dat de gevonden patronen tussen uitkomst en voorspellende variabelen bij een lasso logit minder gevoelig zijn voor ruis in de data.

Random Forest

Een random forest (Breiman, 2001) is een populaire voorspelmethode uit de *machine learning*-literatuur. De voorspelling van een random forest is een combinatie van meerdere *regression trees*.[5] Een *regression tree* splitst de trainingsdata aan de hand van een variabele in twee delen: observaties met waarden groter dan de

[4] https://academic.oup.com/qje/article/133/1/237/4095198.

[5] Hoewel we een binaire variabele (wel of niet verhuisd) verklaren, gebruiken we in dit onderzoek *regression trees* omdat we de *verhuiskans* willen bepalen.

grenswaarde voor deze variabele worden bij elkaar gezet en observaties met waarden kleiner dan grenswaarden ook. Voor beide groepen is de voorspelde waarde gelijk aan de gemiddelde waarde van de afhankelijke variabelen. De *regression tree* kiest de variabele en bijbehorende grenswaarde die de voorspelfout in de trainingsdata minimaliseert. Vervolgens deelt de *regression tree* deze groepen opnieuw in tweeën op basis van de beste presterende variabele en grenswaarde volgens de hierboven genoemde beslisregel. Hierdoor zijn er nu vier groepen ('leafs') ontstaan. Dit kan doorgaan tot geen verdere indeling meer mogelijk is. Omdat dit vrijwel zeker leidt tot *overfitting* kan de diepte van de *regression tree* gereguleerd worden aan de hand van een minimale reductie van de gemiddelde voorspelfout die vereist is om een extra indeling te maken, het minimale aantal observaties in een *leaf* of de maximale diepte van de *regression tree*. Desondanks hebben *regression trees* de neiging om te overfitten en zullen ze slecht *out-of-sample* voorspellen.

Een *random forest* overkomt dit door een groot aantal *regression trees* te combineren door middel van *bootstrapped aggregation* ('BAGging'). Iedere *regression tree* is geschat op een willekeurig getrokken sample (met terugleggen) van de trainingsdata (een *bootstrapped sample*) en de uiteindelijke voorspelling voor een observatie is de gemiddelde voorspelling van alle *bootstrapped* regression trees. Daarnaast wordt ruis gecreëerd doordat de *regression tree* bij iedere splitsing mag kiezen uit een aantal willekeurig bepaalde ('random') variabelen. Dit zijn niet noodzakelijk de variabelen die op dat punt in de 'boom' de grootste daling in gemiddelde voorspelfout opleveren. Door het combineren van gecorreleerde, zwakke voorspellers voorspelt een *random forest* nauwkeuriger dan een *regression tree*. Parameters die de voorspelling bepalen zijn het aantal bomen, het aantal variabelen waaruit het algoritme kan kiezen bij iedere splitsing en het minimum aantal observaties per *leaf*.

Modelevaluatie

De meeste voorspelmodellen schatten de kans dat iemand verhuist: die kans is dan bijvoorbeeld 10, 50 of 90 procent. Om te kunnen kijken hoeveel verhuizingen correct voorspeld worden, wordt de geschatte kans geclassificeerd als 'verhuist' als deze groter is dan een bepaalde grenswaarde, bijvoorbeeld 50 procent. Nadat een grenswaarde gekozen is, kan nagegaan worden hoeveel procent van de daadwerkelijke verhuizingen goed voorspeld worden (de *true positive rate*, TPR) en van hoeveel procent van de gevallen waarin een vergunninghouder niet verhuist, het model verwacht dat de statushouder dit wel doet (de *false positive rate*, FPR). Een model dat perfect voorspelt, heeft een TPR van één en een FPR van nul. De grenswaarde bepaalt de hoogte van de TPR en FPR: bij een voorspelmodel dat aangeeft dat niemand verhuist, zijn de TPR en FPR allebei gelijk aan nul, bij een voorspelmodel dat aangeeft dat iedereen verhuist zijn deze waarden allebei gelijk aan één.

In de *machine learning* is het gebruikelijk om de kwaliteit van een model in termen van 'Area Under the Curve' (hierna: AUC) uit te drukken. De AUC zet de FPR op de x-as en zet deze voor alle mogelijke grenswaarden af tegen de TPR op de y-as. Vervolgens geeft de AUC de oppervlakte onder deze lijn. Een model dat alle (niet-) verhuizingen correct voorspelt, zal een AUC hebben die gelijk is aan één, een model dat niet in staat is om verhuizingen van niet-verhuizingen te onderscheiden (zodat de kans op verhuizing willekeurig is) heeft een AUC van een half, terwijl een model dat alle verhuizingen classificeert als niet-verhuizingen en andersom, een AUC heeft van nul.

Oracle

Als een robuustheidscheck creëren we tot slot een *oracle estimator*. De *oracle estimator* is een gewogen gemiddelde van de voorspellingen van ons ensemble en de *te voorspellen* uitkomst. In onze toepassing zijn de gewichten voor het ensemble en de te voorspellen uitkomst 80 en 20 procent. Doordat een *oracle estimator* gebruikmaakt van de te voorspellen uitkomst, 'voorspelt' ze heel goed. Maar natuurlijk kan de oracleschatter niet gebruikt worden in de praktijk. Of een statushouder zal verhuizen, is immers niet met zekerheid bekend bij uitplaatsing. Wij gebruiken de oracle estimator daarom om te bepalen in hoeverre onze uitkomsten van het ensemble lijken op die uit een ideale voorspelling.

Hoe goed voorspelt het model?

Tabel 2 geeft de AUC-waardes weer voor de verschillende schattingstechnieken, waarbij we ook de componenten van het ensemble uitlichten. In beide cohorten liggen de AUC's van de ensemble-indicator tussen de 0,7 en 0,75. Dit is een redelijke voorspelbaarheid, maar niet heel goed. De AUC's voor de bijstandskans liggen iets hoger, maar ook niet boven de 0,8. Deze AUC-waarden geven aan dat het verhuisgedrag en bijstandsgebruik op basis van de gebruikte data niet heel goed te voorspellen zijn op nationaal niveau.

Tabel 2 AUC-waarden per schattingstechniek, cohort en verhuis- en bijstandskans

		Cohort 1995				**Cohort 2014**			
		LASSO	RF	Ensemble	Oracle	LASSO	RF	Ensemble	Oracle
Verhuiskans	In-sample	0,748	0,986	0,979	1	0,741	0,983	0,982	1
	Out-of-sample	0,718	0,723	0,744	0,977	0,707	0,695	0,720	0,988
Bijstandskans	In-sample	0,764	0,992	0,974	0,998	0,784	0,992	0,977	0,999
	Out-of-sample	0,749	0,766	0,772	0,941	0,770	0,740	0,776	0,957

Dit blijkt ook uit figuur 3 waar voor het ensemble de verhuiskans voor statushouders ingedeeld is in percentielen voor de cohorten 1995 en 2014. De percentielen zijn gesorteerd: het 1e percentiel bevat de statushouders met de laagste voorspelde verhuiskans, het 100e percentiel statushouders met de hoogste voorspelkans. De doorgetrokken lijn 'voorspelling ensemble' geeft voor ieder percentiel de gemiddelde *voorspelde* verhuiskans. De cirkels geven voor ieder percentiel het percentage dat daadwerkelijk verhuisd is, of de *gerealiseerde* verhuiskans binnen het percentiel. We zien dat voor beide cohorten de voorspelde verhuiskans de gerealiseerde redelijk volgt, de gerealiseerde verhuiskansen zijn geclusterd rondom de voorspelde.

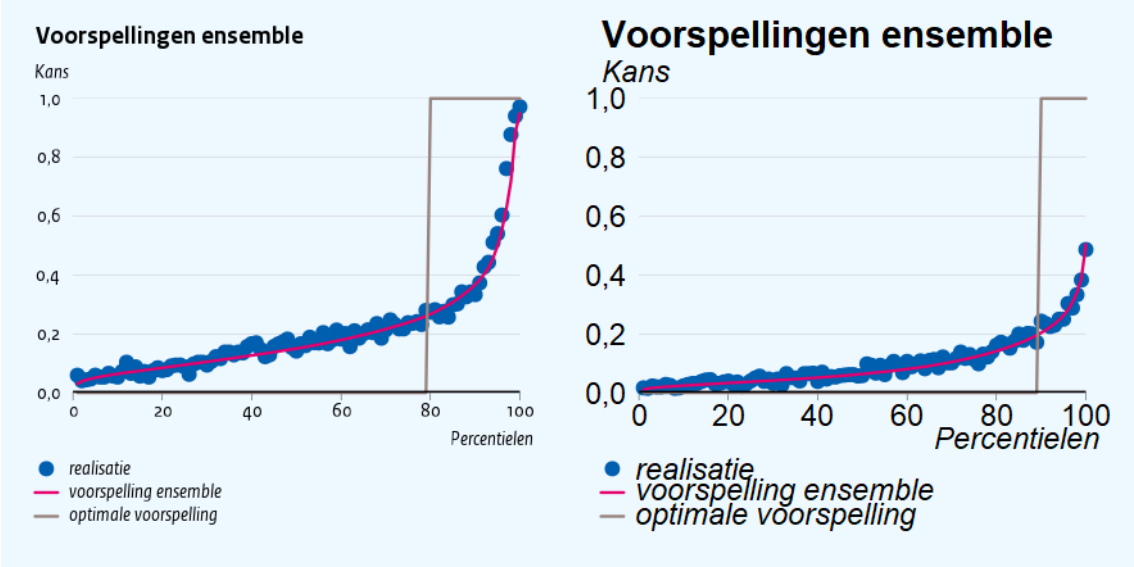
Panelen 3a en 3b tonen dat de verhuiskans slecht te voorspellen is: voor het cohort-1995 is de gerealiseerde verhuiskans alleen voor de drie hoogste percentielen groter dan 80 procent. Voor andere percentielen is dit percentage (veel) lager en is het dus moeilijk om met redelijke waarschijnlijkheid aan te geven of het percentiel verhuist. Voor het cohort-2014 is de gemiddelde gerealiseerde verhuiskans voor geen enkel percentiel hoger dan 60 procent.

Een andere manier om dit te zien is als volgt: de gemiddelde verhuiskans (in test en *train data*) is ongeveer 20 procent in 1995 en 10 procent in 2014. Dit betekent dat een optimaal voorspelmodel aangeeft dat iedereen binnen het 80e percentiel of hoger verhuist en dat iemand binnen de overige percentielen niet verhuist. Dit is aangegeven met de lijn 'optimale voorspelling'. De voorspelling van het ensemble wijkt hier duidelijk van af. Voor het cohort 1995 stijgt de lijn 'voorspelling ensemble' snel en is de gemiddelde voorspelde verhuiskans 20 procent voor het 60e percentiel. Vanaf het 80e percentiel stijgt de gemiddelde voorspelling van het ensemble juist niet snel genoeg en wijkt de gemiddelde voorspelling sterk af van de optimale voorspelling '1'. Alleen voor de onderste en bovenste paar percentielen van het cohort-1995 komen de voorspellingen van het ensemble goed overeen met de optimale voorspelling. Ook voor het cohort-2014 zien we dat de lijn 'voorspelling ensemble' vrij snel afwijkt van de lijn 'optimale voorspelling'. Echter voor de hoogste tien percentielen, waarin het voorspelde aantal verhuizingen dus groot moet zijn, zien we dat de gemiddelde voorspelling juist achterblijft bij de optimale voorspelling.

Figuur 3 De verdeling van voorspelde verhuiskansen door de ensemblemethode

3a :cohort 1995

3b: cohort 2014

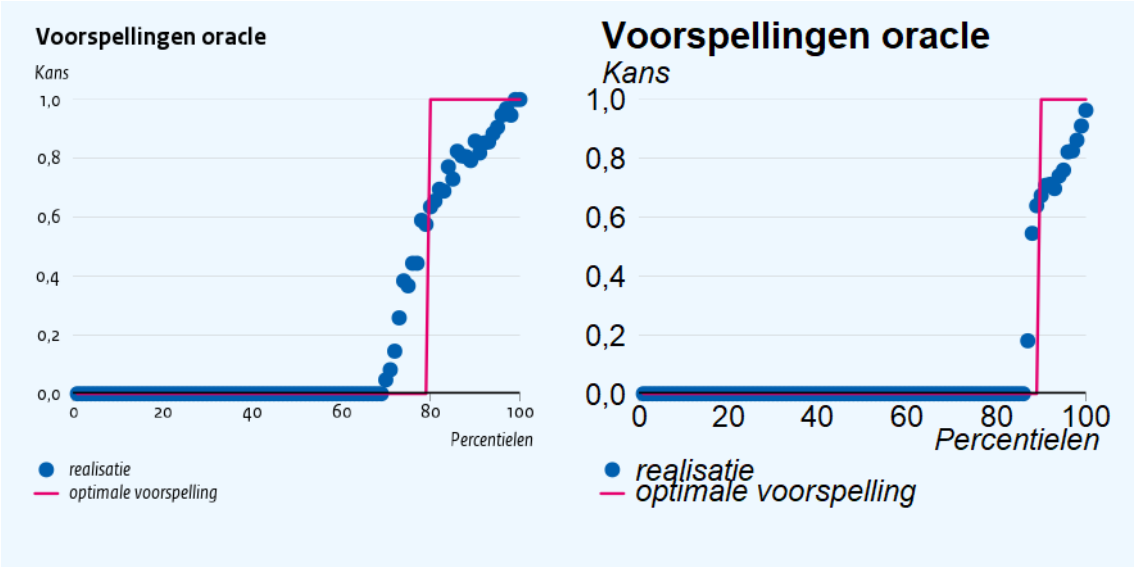


Ter illustratie laten we de gemiddelde voorspelde verhuizing ook zien voor de oracle-schatter in figuur 4. Hier zien we dat de gemiddelde gerealiseerde verhuiskans de optimale voorspelling precies volgt tot aan het 70ste percentiel (cohort 1995) of het 90ste percentiel (cohort 2014). Vervolgens stijgt de lijn heel snel richting 1. Als een voorspelmodel zo goed zou presteren als ons oracle, kunnen gemeenten dus met veel waarschijnlijkheid voor *alle* vergunninghouders goed aangeven wie wel of niet verhuizen.

Figuur 4 De verdeling van voorspelde verhuiskansen door de oracle-methode

4a: cohort 1995

4b: cohort 2014



Hoe goed voorspelt het model op gemeenteniveau?

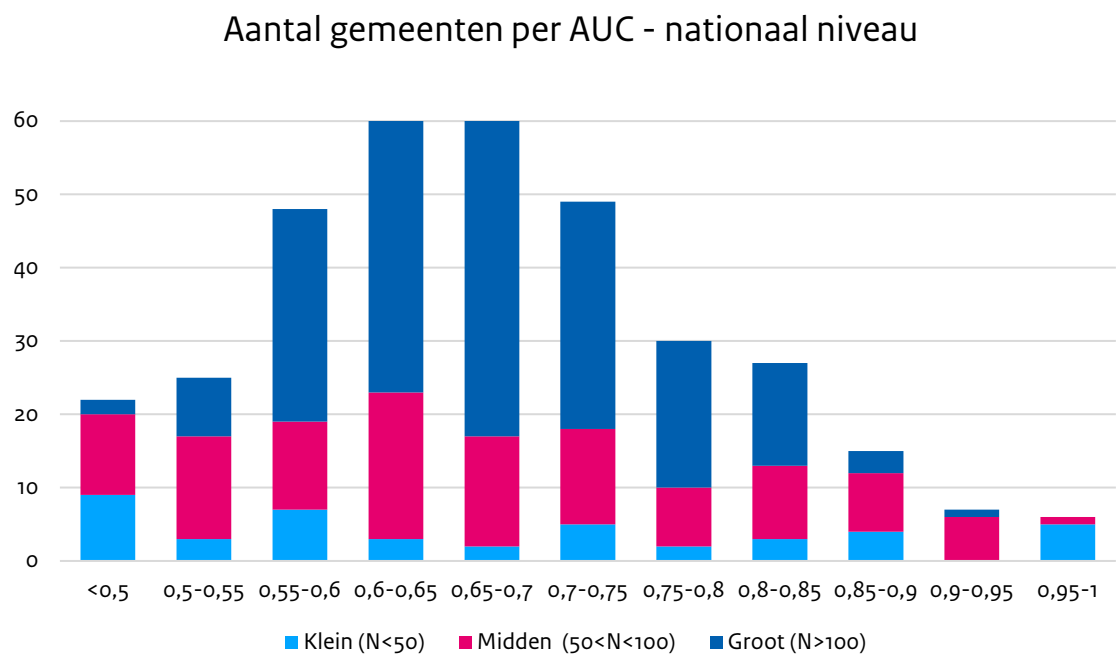
De resultaten tot nu toe zijn gebaseerd op een nationaal voorspelmodel: we hebben de verhuis- en bijstandskansen voorspeld op basis van de hele populatie van statushouders. In werkelijkheid zal een gemeente echter geen informatie hebben over de gehele populatie, enkel over de lokale populatie. Om een

beeld te krijgen van de voorspelbaarheid met informatie op gemeenteniveau maken we voorspelmodellen voor gemeenten waar ten minste honderd statushouders uitgeplaatst zijn. We doen dit voor het 2014- cohort. In werkelijkheid kan een gemeente, buiten de observeerbare kenmerken, ook nog informatie hebben over verhuisgeneigdheid via een gesprek met de statushouder. Onze schattingen op gemeenteniveau bevatten deze extra informatie niet.

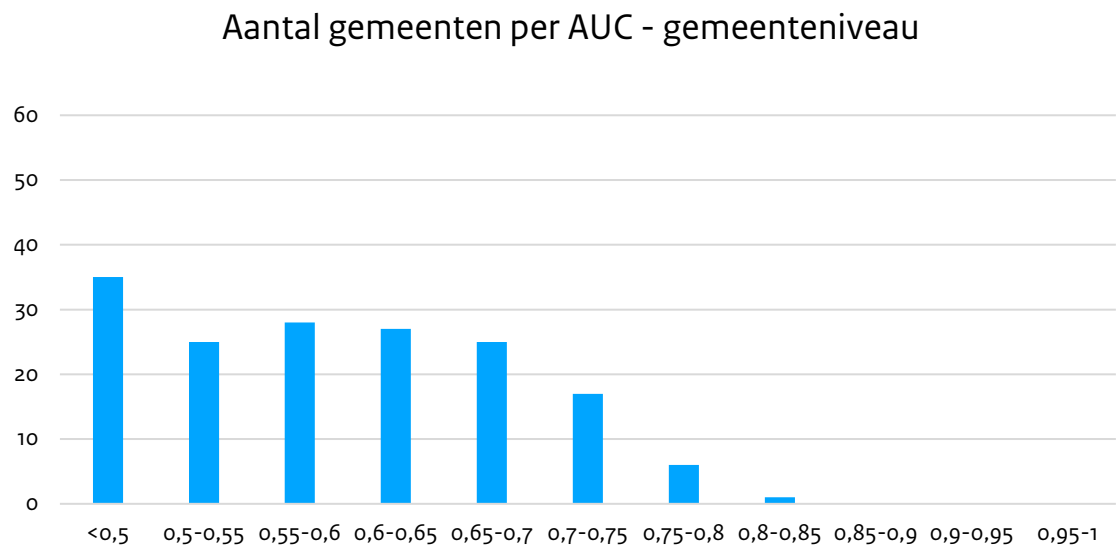
Figuur 5 geeft het aantal gemeenten per AUC-categorie weer, opgedeeld op basis van het aantal uitgeplaatsten. Hierbij is het model getraind op basis van de gehele populatie. Per gemeente wordt dus de kans geschat voor de uitgeplaatste statushouders met informatie van alle uitgeplaatsten. Er zijn maar enkele gemeenten met een AUC hoger dan 0,8.

Figuur 6 geeft het aantal gemeenten per AUC-categorie weer, maar in dit figuur zijn de AUC's gebaseerd op modellen op gemeenteniveau. Per gemeente heeft het model dus alleen de informatie van de statushouders die in die gemeente uitgeplaatst waren in plaats van de gehele populatie. Het aantal gemeenten met een AUC boven de 0,8 is nog lager dan de AUC's op gemeenteniveau met nationale data. De analyses op gemeenteniveau laten zien dat het beeld over verhuisgeneigdheid door het nationale model - verhuisbewegingen vallen lastig te voorspellen– in lijn zijn met bevindingen op gemeenteniveau. Er lijken weinig gemeenten te zijn die ontzettend goed kunnen voorspellen wie er verhuist.

Figuur 5 Verdeling van AUC's per gemeente, uitgesplitst naar gemeentegrootte – geschat op nationaal niveau



Figuur 6 Verdeling van AUC's per gemeente – geschat op gemeenteniveau



Om een beeld te krijgen of de kleine groep sterk verhuisgeneigden verschilt van de minst verhuisgeneigden (top 10 percentielen vs. laagste 50 decielen) vinden we dat statushouders met een heel hoge verhuiskans gemiddeld gezien vaker man en jonger zijn en zonder kind/partner wonen. Voor het nieuwe cohort zien we verder dat diegenen die een partner en/of kind hebben, meer verhuisgeneigd zijn als ze samen met een jongere partner of oudere kinderen wonen (tabel 3).

Tabel 3 Descriptieve statistieken achtergrondkenmerken sterk en minst sterk verhuisgeneigden

3.A Cohort 1995

		Top 10 percentielen			Laagste 50 percentielen			
		N	Gem.	Std. Dev.	N	Gem.	Std. Dev.	p-waarde t-test
Man		3501	0,623	0,485	15.916	0,581	0,493	0,000
Leeftijd		3501	31,850	10,692	15.916	36,026	11,066	0,000
Herkomstland	Syrië	3501	0,016	0,127	15.916	0,011	0,104	0,009
	Iran	3501	0,070	0,256	15.916	0,088	0,283	0,000
	Irak	3501	0,286	0,452	15.916	0,247	0,432	0,000
	Somalië	3501	0,064	0,245	15.916	0,032	0,177	0,000
	Afghanistan	3501	0,211	0,408	15.916	0,241	0,428	0,000
	Voorm. Joegoslavië	3501	0,102	0,303	15.916	0,158	0,365	0,000
	Voorm. Sovjet Unie	3501	0,046	0,209	15.916	0,061	0,240	0,000
	Overig	3501	0,204	0,403	15.916	0,160	0,367	0,000
Kind/partner		3501	5,550	14,355	15.916	15,986	23,040	0,000
Leeftijd partner		3215	0,586	0,493	15.480	0,718	0,450	0,000
Leeftijd jongste kind		183	36,399	9,750	7565	36,674	10,146	0,717
Leeftijd oudste kind		1263	7,280	6,511	7928	7,254	6,375	0,892

3.B Cohort 2014

		Top 10 percentielen			Laagste 50 percentielen			
		N	Gem.	Std. Dev.	N	Gem.	Std. Dev.	p-waarde t-test
Man		6602	0,622	0,485	30.013	0,567	0,495	0,000
Leeftijd		6602	25,532	8,298	30.013	35,674	11,285	0,000
Herkomstland	Syrië	6602	0,668	0,471	30.013	0,641	0,480	0,000
	Iran	6602	0,039	0,194	30.013	0,011	0,105	0,000
	Irak	6602	0,036	0,186	30.013	0,019	0,138	0,000
	Somalië	6602	0,036	0,185	30.013	0,006	0,078	0,000
	Afghanistan	6602	0,019	0,137	30.013	0,011	0,106	0,000
	Voorm. Joegoslavië	6602	0,001	0,025	30.013	0,000	0,012	0,019
	Voorm. Sovjet Unie	6602	0,004	0,063	30.013	0,002	0,042	0,000
	Overig	6602	0,198	0,399	30.013	0,309	0,462	0,000
Kind/partner		4281	0,427	0,495	29.961	0,719	0,449	0,000
Leeftijd partner		558	30,400	9,862	11.661	38,976	10,746	0,000
Leeftijd jongste kind		2924	12,240	7,289	15.029	7,100	5,729	0,000
Leeftijd oudste kind		2976	19,037	6,918	15.725	11,837	6,609	0,000

Uitsortering van voorzieningengebruikers

Om te bepalen of er selectieve uitsortering heeft plaatsgevonden, willen we weten of het werkelijke bijstandsgebruik in een gemeente significant afwijkt van het bijstandsgebruik dat je zou verwachten op basis van de oorspronkelijk uitgeplaatste groep statushouders in een gemeente. We berekenen dit door te testen of K_{gt} in de 10 procent uitersten van de binomiale verdeling zit (positieve of negatieve afwijking):

$$x = Binomiaal(N_{gt}, \Theta_{gt})$$

Waarbij;
N_{gt} = aantal statushouders dat daadwerkelijk in gemeente g woont op tijd t
K_{gt} = aantal statushouders dat bijstand gebruikt in gemeente g op tijd t
Θ_{gt} = aantal oorspronkelijk uitgeplaatste inwoners van gemeente g die bijstand gebruiken op tijd t /aantal oorspronkelijke bewoners van gemeente g die in Nederland wonen op tijd t
De peilmomenten liggen op drie, vijf, tien, vijftien of twintig jaar (of in de drie maanden daarvoor) na uitplaatsing.

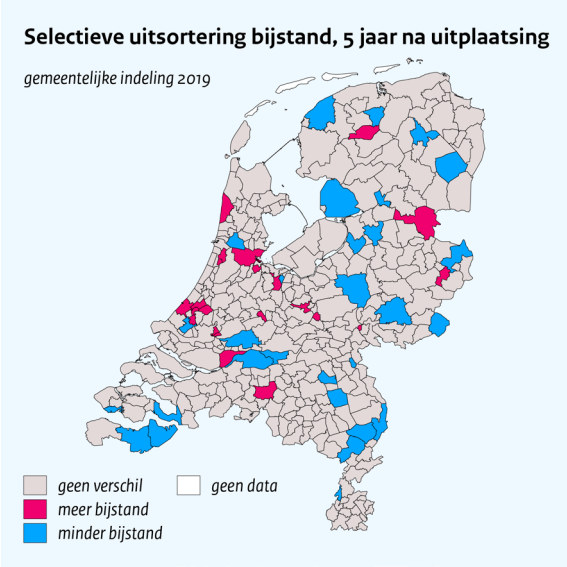
Als het minimum van de binomiale staarten kleiner is dan 0,1 beschouwen we de gemeente als een gemeente waar het bijstandsgebruik significant veranderd is door verhuisgedrag. Of een gemeente een positieve of negatieve significante afwijking van het verwachte bijstandsgebruik heeft, hangt af aan welke kant van de verdeling de gemeente valt. Het teken is negatief als het aantal bijstandsgebruikers 'links' in de staart ligt, en positief als het 'rechts' in de staart van de binomiale verdeling ligt.

De aanname die we doen, is dat de bijstandskans intrinsiek is aan de persoon. We weten bijvoorbeeld dat de bijstandskans sterk afhangt van iemands leeftijd bij binnenkomst. We willen kijken of er clustering van bijstandsgebruikers plaatsvindt doordat mensen met een relatief hoge intrinsieke bijstandskans in dezelfde gemeenten clusteren door verhuisbewegingen. Het is echter ook mogelijk dat de bijstandskans wordt beïnvloed door verhuizing. Figuur 5 in de Policy Brief laat zien dat er geen sterk verband lijkt te zijn tussen verhuizen en de bijstandskans; in het bijzonder hebben de ex-post verhuizers en ex-post blijvers dezelfde bijstandskans, behalve binnen het deciel met de hoogste voorspelde verhuiskans. Dit suggereert een beperkt effect van verhuizing op de bijstandskans. Figuur 5 in de Policy Brief laat desondanks zien dat de bijstandskans van degenen die verhuizen, gemiddeld genomen wel enigszins stijgt na verhuizing; het effect is dus niet nul. Ons analysemodel maakt een eerste-orde-correctie voor een stijging van de bijstandskans na verhuizing; immers de stijgende bijstandskans van de vertrekkers uit een gemeente leidt tot een hogere Θ waartegen de stijgende bijstandskans van de aankomers wordt afgezet. Desondanks moet worden bedacht dat het onmogelijk is een perfecte *counterfactual* te maken van het bijstandsgebruik per gemeente in afwezigheid van verhuisbewegingen; het is dus niet uit te sluiten dat een deel van de waargenomen uitsortering in werkelijkheid het gevolg is van bijstandskansen die worden beïnvloed door verhuizingen.

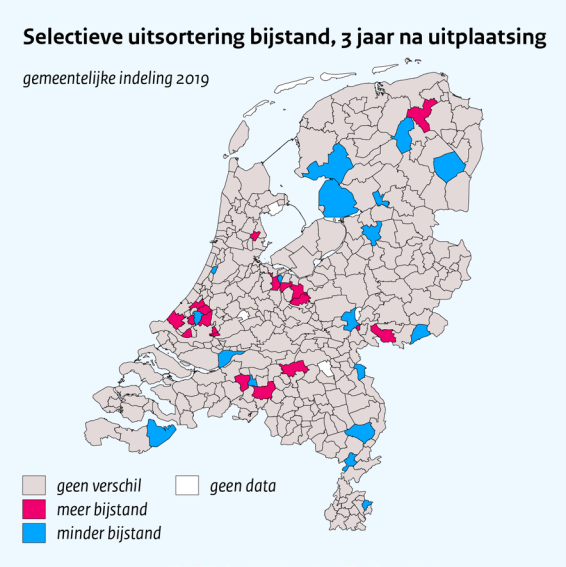
Figuur 7 laat aan de hand van de kaart van Nederland voor verschillende jaren na uitplaatsing zien welke gemeenten een significant hoger of lager bijstandsgebruik hebben door verhuisgedrag. Drie, vijf en tien jaar na uitplaatsing laten geen selectieve uitsortering zien, maar voor vijftien en twintig jaar na uitplaatsing zijn er meerdere gemeenten met een significant hoger bijstandsgebruik. Het gros van deze gemeenten bevindt zich in midden-Nederland. Het aantal significante gemeenten komt echter grotendeels overeen met statistische ruis, als we een p-waarde van 0,1 aanhouden. Twintig jaar na uitplaatsing zijn voor 324 gemeenten data beschikbaar over het bijstandsgebruik. Met een p-waarde van 0,1 en het meenemen van beide staarten zijn er voor circa 65 gemeenten significante waarden door statistische ruis. In de kaart met de selectieve uitsortering twintig jaar na uitplaatsing (figuur 7.E) is te zien dat 77 gemeenten een significante afwijking hebben; hier valt echter ca. 85% van onder statistische ruis. Bovendien is het bijstandsgebruik maar heel beperkt hoger door selectieve uitsortering in de gemeenten met de meest significante afwijkingen (zie Policy Brief sectie 4).

Figuur 7 Selectieve uitsortering - bijstand

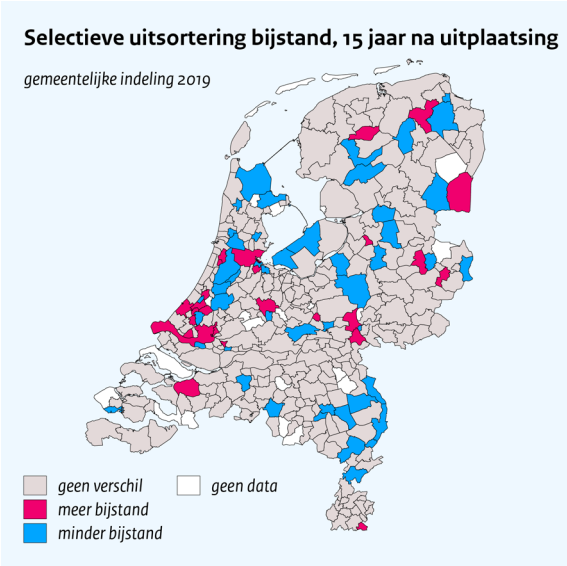
7.A) 3 jaar na uitplaatsing



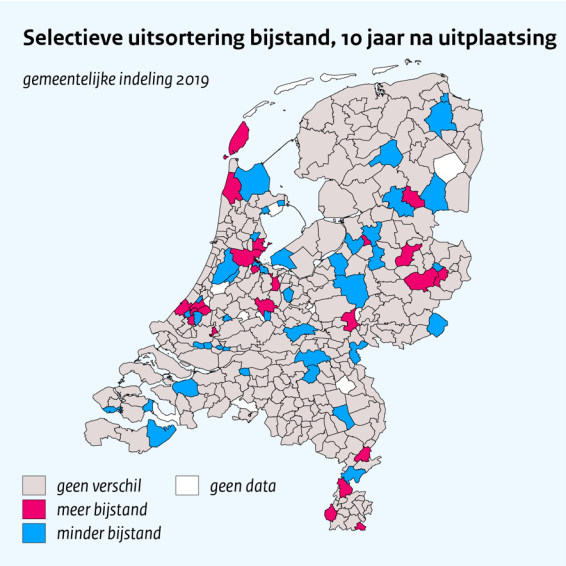
7.B) 5 jaar na uitplaatsing



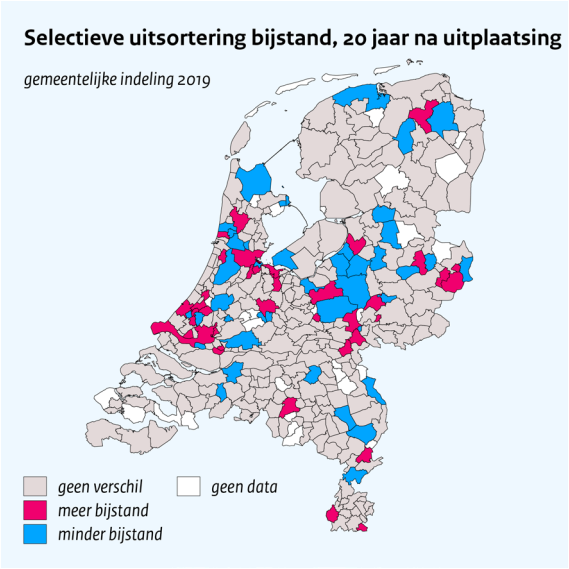
7.C) 10 jaar na uitplaatsing



7.D) 15 jaar na uitplaatsing

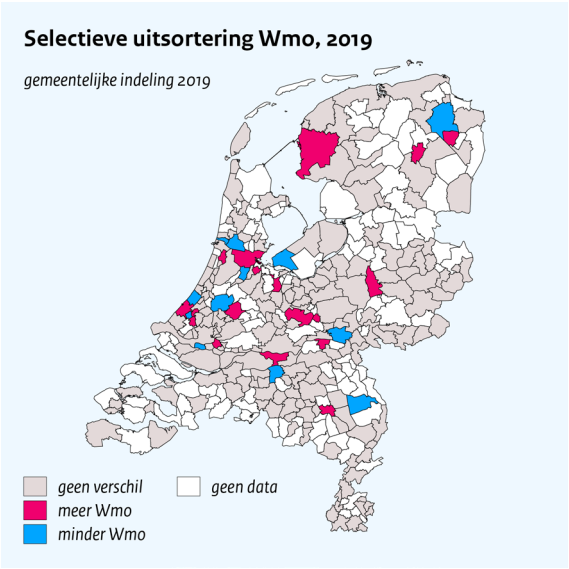


7.E) 20 jaar na uitplaatsing



Wat betreft het Wmo-gebruik hanteren we dezelfde methode, maar hier kijken we niet naar tijd sinds uitplaatsing maar naar het Wmo-gebruik in januari 2019. Er ontstaat dan het beeld zoals in figuur 8 staat; het aantal gemeenten dat significant meer of minder Wmo-gebruik heeft, komt overeen met statistische ruis (221/5=ca. 44, we zien maar voor 28 gemeenten significante afwijkingen). Er is dus geen indicatie dat uitsortering van Wmo-gebruikers zorgt voor significant hogere of lagere uitgaven aan Wmo in bepaalde gemeenten.

Figuur 8 Selectieve uitsortering – Wmo



Referenties

Breiman, L. 2001. Random Forests, *Machine Learning* 45 (1): 5-32.

Tibshirani, R., 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B (Methodological)*, *58*(1): 267-288.