STAtOR

periodiek van de VvS+OR jaargang 17, nummer 3, december 2016

thema **SMART SOCIETIES**

Ontbrekende waarden bij het schatten van het aantal inwoners van Nederland

Susanna Gerritse, Bart Bakker & Peter van der Heijden

Statistische bureaus zoals het Centraal Bureau voor de Statistiek (CBS) houden periodiek een volkstelling. De meeste landen houden nog een traditionele volkstelling waarbij alle inwoners van een land een vragenlijst invullen (Schulte Nordholt, 2016). In Nederland maakt het Centraal Bureau voor de Statistiek gebruik van een populatieregister gebaseerd op administratieve data van gemeenten om het aantal inwoners te schatten. Vóór 1 januari 2014 was de Gemeentelijke basisadministratie (GBA) het Nederlandse Populatieregister, daarna is deze vervangen door de Basisregistratie Personen (BRP). Aan het Populatieregister worden dan andere registraties en een enquête gekoppeld om de volledige informatie van de volkstelling te kunnen leveren.

Vangst-hervangst-methodologie

Het aantal inwoners van een land wordt in het Engels aangeduid met *usual residents*, ofwel de 'gewoonlijk verblijvende bevolking', welke wij in het vervolg tevens aanduiden met 'inwoners'. Volgens EU Regulation No 1260/2013 worden de inwoners gedefiniëerd als mensen die minstens 12 maanden in Nederland verblijven, of deze intentie hebben. Het Populatieregister omvat een aanzienlijk deel van het aantal inwoners. Echter het Populatieregister alleen is niet voldoende om het aantal inwoners te schatten, aangezien er ook mensen zijn die wel in Nederland verblijven maar die zich niet hebben ingeschreven. Twee zulke grote groepen zijn voormalige asielzoekers die uitgeprocedeerd zijn, en Europese arbeidsmigranten die wegens vrij verkeer van personen legaal in Nederland verblijven, ook als ze zich niet inschrijven in het Populatieregister. Als het Populatieregister gebruikt wordt om het aantal inwoners in Nederland te schatten, is er sprake van onderdekking.

Vangst-hervangst-methodologie kan worden gebruikt om de onderdekking te schatten van het Populatieregister (Bishop, Fienberg en Holland, 1975; International Working Group for Disease Monitoring and Forecasting, 1995). In het simpelste voorbeeld worden twee registers gebruikt, waarbij mensen in ieder van de registers wel of niet zijn opgenomen. Hierdoor kunnen aantallen mensen worden geclassificeerd in een 2 bij 2-tabel. Het aantal mensen dat in beide registers niet voorkomt maar wel tot de bevolking behoort, is per definitie onbekend en dient geschat te worden. Dit aantal kan worden geschat met een loglineair model voor de 2×2-tabel als wordt aangenomen dat de insluitkansen van de registers niet samenhangen.

In dit onderzoek worden twee registers gekoppeld aan het Populatieregister: het HerKenningsdienst Systeem (HKS) van de politie en het werknemersdeel uit de Polisadministratie, dat het WerkNemersBestand (WNB) is genaamd. Het HKS is een register dat bijgehouden word door de politie, waarbij alle verdachten van bij de politie bekende procesverbalen geregistreerd staan. De Polisadministratie is een groot register op het CBS waarin allerlei werknemers- en werkgeversinformatie geregistreerd staat. Voor het WNB hebben wij enkel het deel van de werknemers uit de Polisadministratie gehaald. Er zijn twee niveaus per register: mensen kunnen wel of niet in een register zitten en dus krijgen we een 2×2×2-tabel, met 1 structurele nul. Door middel van de AIC kan er gezocht worden naar het best passende loglineaire model. Vanuit dit model wordt de structurele nul geschat.

Verblijfsduur bepalen

Om de vraag te beantwoorden hoeveel inwoners Nederland heeft, dient eerst deze vraag te worden beantwoord: hoeveel personen worden gemist door deze drie gekoppelde registers. Verblijfsduur wordt in het loglineaire model als covariaat opgenomen en is van cruciale betekenis om te schatten hoeveel inwoners er zijn, omdat de EU vereist dat de verblijfsduur minimaal 12 maanden is. Echter, de verblijfsduur was enkel te bepalen voor het Populatieregister en het WNB en niet voor het HKS.

Voor het Populatieregister werd de verblijfsduur afgeleid als het verschil tussen de dag van inschrijving en het peilmoment. Door gebruik te maken van het WNB konden we voor het Populatieregister ook gebruik maken van de duur van banen. Als deze duur langer was dan de lengte van inschrijving in het Populatieregister dan werd deze duur gebruikt. Voor personen die alleen in het WNB voorkwamen konden we de duur van banen gebruiken om een verblijfsduur af te leiden, onder de aanname dat iemand die in Nederland een baan heeft ook hier verblijft. Voor de mensen in het HKS die niet gekoppeld konden worden aan het Populatieregister of het WNB was er voor verblijfsduur een ontbrekende waarde. Voor het HKS is besloten deze waarden te imputeren. Verschillende methoden zijn hiervoor beschikbaar en het is onbekend welke methode het beste is. Daarom is besloten een sensitiviteitsanalyse uit te voeren voor de uiteindelijke vangst-hervangst schattingen. Voor de ontbrekende gegevens zijn via drie methoden waarden ingevuld en de resulterende populatieschattingen zijn met elkaar vergeleken. Twee van de drie scenario's gebruikten het Expectation Maximization-algoritme (EM; Schafer, 1997), en het laatste scenario maakte gebruik van Predictive Mean Matching (PMM; Van Buuren, 2012).

We geven hieronder een voorbeeld van personen met een Poolse nationaliteit in 2010. In tabel 1 is de kruistabel gegeven van het al dan niet voorkomen in de drie registers, uitgesplitst naar korter of langer verblijvend dan 12 maanden. In de tabel zien we het aantal Polen in Nederland in 2010 die in minstens 1 van de drie registers staan geregistreerd. Zo zitten er 32 Polen in zowel de PR, HKS als de WNB met een verblijfsduur van korter dan 12 maanden. Maar bijvoorbeeld, 3.523 Polen zitten wel in de WNB en de PR, maar niet in de HKS met een verblijfsduur korter dan 12 maanden. Er zijn twee (structurele nullen): deze nullen verwijzen naar de mensen die in geen van de drie registraties voorkomen. Daarnaast zijn er mensen die alleen in de HKS voorkomen en niet in de andere twee registers. Voor hen is de verblijfsduur niet af te leiden uit het Populatieregister of het WNB. Dat zijn de twee cellen waar 'Ontbreekt' staat. Het totaal van deze twee cellen is een aantal van 1.043 Polen. Echter, hoeveel daarvan korter verblijven dan 12 maanden en hoeveel ervan langer verblijven dan 12 maanden is niet bekend. Er is dus sprake van ontbrekende waarden, en dit probleem moet als eerste opgelost worden. Hiervoor worden twee manieren gebruikt: het EM-algoritme en PMM.

			HKS = Ja	HKS = Nee
Verblijfsduur <12 maand	PR = Ja	WNB = Ja	32	3.523
		WNB = Nee	34	3.225
	PR = Nee	WNB = Ja	149	60.190
		WNB = Nee	Ontbreekt	0
Verblijfsduur >12 maand	PR = Ja	WNB = Ja	183	21.309
		WNB = Nee	195	14.052
	PR = Nee	WNB = Ja	81	21.216
		WNB = Nee	Ontbreekt	0

Tabel 1. De geobserveerde waardes van mensen met een Poolse nationaliteit die korter of langer dan 12 maanden in Nederland verblijven en die in minstens 1 van de 3 registers staan geregistreerd (PR=Populatieregister, HKS=HerKenningsdienst Systeem, WNB=WerkNemersBestand)

EM-algoritme en PMM

Het EM-algoritme is een iteratief algoritme voor het maken van *maximum likelihood* schattingen van een model wanneer er ontbrekende waarden zijn (Little and Rubin, 2002). Iedere iteratie kent een Expectation (E) en een Maximization (M) stap. In de E-stap worden de verwachtingen van de ontbrekende waarden ingevuld op basis van de geobserveerde waarden (hier het geobserveerde aantal 1.043 en de andere waarden in de bovenstaande tabel) en de schattingen afkomstig uit de M-stap. In de M-stap worden parameters geschat op basis van de gegevens verkregen in de E-stap. Na de M-stap volgt weer een E-stap en dit gaat zo door totdat de parameterschattingen geconvergeerd zijn. In de M-stap wordt steeds een loglineair model geschat. Parameter schattingen van het loglineaire model kunnen gebruikt worden voor het schatten van het aantal personen dat door alle drie registers is gemist.

Twee scenario's gebruiken het EM-algoritme voor completering. In het eerste scenario (EM-A) is een verzadigd model gebruikt om de data te completeren. Op deze gecompleteerde data is vervolgens een spaarzaam loglinear model geschat dat de data het beste beschrijft (volgens het Akaike Informatie Criterium) en op basis hiervan is een schatting gemaakt van het aantal inwoners. In scenario twee (EM-B) is een best passend, spaarzaam loglineair model gebruikt voor completering van de data, waarna de parameters van dit laatste model zijn gebruikt om tot een schatting te komen van het aantal inwoners.

Een derde scenario maakt gebruik van multiple imputatie met *Predictive Mean Matching* (PMM). Kandidaat donoren worden gekozen uit de geobserveerde personen die vergelijkbare achtergrondkenmerken hebben als personen met een ontbrekende waarneming. Een willekeurige donor wordt dan gekozen uit alle kandidaten, die dan als donor dient voor de ontbrekende waarneming (hier: van de waarde voor verblijfsduur).

In de context van het huidige onderzoek is er een belangrijk verschil tussen EM en PMM als datacompleteringsmethode. In ons onderzoek is het niet logisch om de gehele populatie te nemen als donoren. De mensen die geregistreerd staan in het HKS en niet ingeschreven zijn in het Populatieregister lijken qua verblijfsduur veel meer op de mensen die een baan hebben in Nederland maar niet zijn ingeschreven zijn in het Populatieregister. Met PMM kunnen we de donoren beperken tot deze subcategorie, terwijl het EM-algoritme de ontbrekende getallen invult aan de hand van de gegevens van de gehele populatie.

Schatting van de onderdekking van het Populatieregister

De schattingen uit de drie scenario's zijn met elkaar vergeleken om tot de beste schatting van de onderdekking van het Populatieregister te komen. Om tot dit besluit te komen is er allereerst literatuuronderzoek gedaan naar eerder onderzoek met vergelijkbare schattingen (Hoogteijling, 2002; Bakker, 2009; Van der Heijden et al., 2012), waaruit een verwachting kon worden afgeleid voor de schatting voor de cijfers uit 2010. Daarnaast gebruikten we ook informatie over de asielaanvragen in 2010 en de jaren ervoor. De verwachte schatting voor 2010 zou mogelijk kunnen liggen tussen 175 en 225 duizend personen.

GEMIST DOOR ALLE DRIE DE REGISTERS		
SCENARIO	schatting × 1000	betrouwbaarheidsinterval
EM-A	139	120 – 176
EM-B	129	111 – 170
PPM	179	121 – 237

Tabel 2. Schattingen voor de 3 scenario's vanuit de vangst-hervangst voor de drie scenario's; tevens zien we de betrouwbaarheidsintervallen

ONDERDEKKING POPULATIEREGISTER		
SCENARIO	schatting × 1000	betrouwbaarheidsinterval
EM-A	172	153 – 209
EM-B	162	143 – 203
PPM	212	154 – 270

Tabel 3. Schattingen voor de onderdekking van het Populatieregister (let op: dit zijn de schattingen van tabel 2, opgehoogd met 33 duizend mensen)

Tabel 2 toont de resultaten. In de tweede kolom is de vangst-hervangst uitkomst te zien, en de bijbehorende betrouwbaarheidsintervallen staan in kolom drie. Deze twee kolommen beslaan het aantal mensen die door alle drie de registers worden gemist. Echter in het HKS en WNB bleken er 33.000 mensen die niet in het Populatieregister waren geregistreerd maar wel langer dan 12 maanden verbleven. Deze behoren tot het aantal inwoners en behoren dus tot de onderdekking van het Populatieregister. De schattingen moeten dus met 33 duizend worden opgehoogd. Dit is te zien in tabel 3. Kolom 2 van tabel 3 geeft de met 33.000 personen opgehoogde schatting en kolom 3 van tabel 3 de opgehoogde betrouwbaarheidsinterval.

De verschillende scenario's leiden tot verschillende uitkomsten. Onze voorkeur gaat uit naar de schatting met de PMM, omdat de resulterende schatting inhoudelijk het beste aansluit bij de ontbrekende data in de HKS. In totaal gaat het dan om 212.000 inwoners die gemist worden in de BRP. Dit aantal ligt bovendien binnen de range van het referentiekader en is derhalve plausibel. Let wel, deze schatting had alleen betrekking op mensen met een leeftijd tussen de 15 en 65, vanwege de begrenzing van de registers. De schatting moet nog worden aangevuld met een schatting van mensen jonger dan 15 en ouder dan 65. Al met al is er dan maar een onderdekking van het Populatieregister van 0,5 tot 1,1%.

Literatuur

Bakker, B.F.M. (2009). *Trek alle registers open!* (Open all registers!). Amsterdam: Vrije Universiteit Amsterdam.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis*. MIT press, Cambridge, MA.

Buuren, S. van. (2012). *Flexible imputation of missing data*. Boca Raton: CRC Press.

Gerritse, S.C., Bakker, B.F.M., & Heijden, P.G.M. van der. *(2015)*. Different methods to complete datasets used for capture-recapture estimation: estimating the number of usual residents in the Netherlands. *Statistical Journal of the IAOS, 31,* 613-627.

Hoogteijling, E.J.M. (2002). *Raming van het aantal niet in de GBA geregistreerden* (technical report). Heerlen: Centraal Bureau voor de Statistiek.

International Working Group for Disease Monitoring and Forecasting (1995). Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology, 142*(10), 1047-1058.

Little, R.J., & Rubin, D.B. (2002). *Statistical analysis with missing data*. Hoboken, New Yersey: John Wiley and Sons.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall.

Schulte Nordholt, E. (2016). Volkstellingen in Nederland en elders, *STAtOR, 17*(1), 8-13.

Heijden, P.G.M. van der, Whittaker, J., Cruyff, M.J.L.F., Bakker, B.F.M., & Vliet, H.N. van der. (2012). People born in the Middle East but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics, 6*, 831-852.

Zwane, E.N., & Heijden, P.G.M. van der. (2007). Analysing capture-recapture data when some variables of heterogeneous catchability are not collected or asked in all registries. *Statistics in Medicine, 26*, 1069-1089.

Susanna Gerritse studeerde psychologie aan de Universiteit van Amsterdam en heeft een promotietraject afgerond aan de Universiteit Utrecht. Deze promotie was in samenwerking met het CBS, waar het bovenstaande onderzoek uit voortvloeide. Momenteel is ze statistisch onderzoeker bij het Centraal Bureau voor de Statistiek.
E-mail: sc.gerritse@gmail.com

Bart F.M. Bakker is hoogleraar Methodologie van Registerdata voor sociaalwetenschappelijk onderzoek aan de VU Amsterdam en is Hoofd van de afdeling Methodologie van het Centraal Bureau voor de Statistiek in Den Haag.
E-mail: Bfm.bakker@cbs.nl

Peter G.M. van der Heijden is hoogleraar Statistiek t.b.v. de sociale wetenschappen aan de Universiteit Utrecht en Professor of Social Statistics at the University of Southampton.
E-mail: p.g.m.vanderheijden@uu.nl