BOA
Nieuwegracht 98
3512 LX Utrecht
Tel. (030) 232 84 49
Fax (030) 230 06 34
BeleidsOnderzoek en Advies

Simulatie Recurrent Events Model

Maarten Cruyff
Ger van Gils
Peter van der Heijden
Utrecht, mei 2013

Simulaties recurrent events model

Maarten Cruyff, Ger van Gils and Peter G.M. van der Heijden

Samenvatting

Het basis vangst-hervangstmodel voor het schatten van de omvang van een verborgen populatie is het Poissonmodel. Dit document evalueert de mogelijkheden van het recurrent events model voor het schatten van de omvang van een populatie. Het verschil tussen de twee modellen is dat het recurrent events model de 'geschiedenis' van de vangsten analyseert, terwijl het Poissonmodel alleen het totale aantal vangsten analyseert. Als gevolg daarvan is het recurrent events model flexibeler dan het Poisson-model, en kan het effecten modelleren zoals tijdelijke afwezigheid uit de bevolking of seizoensgebonden schommelingen in de bevolking. Een nadeel van het model is dat het hogere eisen stelt aan het dataverzamelingsproces, omdat er meer gedetailleerde gegevens nodig zijn. Het doel van dit rapport is om de kosten en baten van het recurrent events model te evalueren.

De eerste drie hoofdstukken beschrijven de theorie van de recurrent events model, en de flexibiliteit van het model met betrekking tot het modelleren van verschillende effecten. De simulatiestudie in hoofdstuk 4 toont dat het model, indien correct gespecificeerd, resulteert in betere schattingen dan het Poissonmodel. Hoofdstuk 5 beschrijft de resultaten van een praktijkvoorbeeld: de schatting van de populatie van illegale immigranten in Nederland in 2009. In tegenstelling tot het Poissonmodel, corrigeert het recurrent events model de schattingen voor de tijd dat de illegale immigranten in detentie hebben doorgebracht. Als gevolg hiervan valt de populatieschatting aanzienlijk lager uit dan die van het Poissonmodel.

De appendix beschrijft het dataverzamelingsproces voor het voorbeeld van de illegale immigranten. Deze beschrijving laat zien dat met name het verzamelen van de detentietijden dusdanig gecompliceerd was, dat bepaalde pragmatische keuzes gemaakt moesten worden. Als gevolg hiervan is de kwaliteit van de detentiegegevens moeilijk te bepalen.

Simulations recurrent events model

Maarten Cruyff, Ger van Gils and Peter G.M. van der Heijden

Summary

The basic capture-recapture model for estimating the size of a hidden population is the Poisson model. This document evaluates the potential of the recurrent events model for population size estimation. The difference between the two models is that the recurrent events model analyzes the 'history' of the captures, while the Poisson model only analyzes the total number of captures. As a consequence, the recurrent events model is more flexible than the Poisson model, and is able to model effects such as a temporary absence from the population or seasonal fluctuations in the population. A down-side of the model is that it requires more detailed data, which may seriously complicate the process of data collection. The aim of this report is to evaluate the costs and benefits of the recurrent events model.

The first three chapters describe the theory behind the recurrent events model, and its flexibility in modeling different effects. The simulation study in Chapter 4 shows that the recurrent events model, if specified correctly, results in better estimates than the Poisson model. Chapter 5 reports the results of a real data example; the estimation of the population of illegal immigrants in the Netherlands in 2009. In contrast to the Poisson model, the recurrent events model corrects the estimates for the time that the illegal immigrants spend in detention. As a consequence, the population estimate is substantially lower than that of the Poisson model.

The appendix describes the process of data collection for the illegal immigrant example. It shows that especially the collection of the detention times has been so complicated, that certain pragmatic choices had to be made. As a consequence, the quality of the required detention data is hard to assess.

Inhoudsopgave

Hoofdstuk 1

Inleiding

In het recente verleden zijn omvangschattingen van populaties gemaakt op basis van de analyse van teldata met behulp van het afgeknotte Poissonmodel. In dit document wordt een plan gepresenteerd om omvangschattingen van populaties te maken op basis 'event histories' met behulp van het recurrent events model. Het recurrent events model is een uitgebreide versie van het Poissonmodel; terwijl het Poissonmodel alleen het totaal aantal gebeurtenissen analyseert, neemt het recurrent events model ook de geschiedenis van de gebeurtenissen (event history) in beschouwing.

Een belangrijk verschil tussen beide modellen is dat het recurrent events model voor elk individu uit de populatie een 'tijd at risk' specificeert waarin gebeurtenissen kunnen optreden. Hierdoor kan het model rekening houden met perioden waarin een individu niet in de populatie aanwezig is geweest, en waarin dus ook geen gebeurtenissen konden optreden. Omdat het Poissonmodel alleen informatie omtrent het totaal aantal gebeurtenissen ter beschikking heeft, kent het deze mogelijkheid in principe niet. Wel is in het verleden een kunstgreep toegepast door personen die voortijdig populatie hebben verlaten in de eerste fase van de analyse (het schatten van de parameters van het Poissonmodel) buiten beschouwing te laten. In de tweede fase van de analyse (het schatten van de populatieomvang op basis van de parameterschattingen) worden deze personen wel weer meegenomen (zie Leerkes et al, 2004). Dit model wordt in dit document aangeduid als het 'tweetraps' Poissonmodel.

Doordat het recurrent events model meer informatie ter beschikking heeft dan het Poissonmodel, levert het in principe meer valide omvangschattingen. Daar staat tegenover dat het recurrent events model hogere eisen stelt aan de dataverzameling, omdat informatie omtrent de tijd at risk be-

schikbaar dient te zijn. Vanwege de extra kosten die dit met zich meebrengt, is het van belang om inzicht te hebben in de situaties waarin het recurrent events model is te prefereren boven het (tweetraps) Poissonmodel. Hiertoe is een aantal simulatiestudies uitgevoerd waarin de populatieschattingen van het recurrents events model en het Poissonmodel onder verschillende situaties worden vergeleken. Hierbij is met name aandacht besteed aan een situatie die in de praktijk veelvuldig voorkomt, namelijk die van een populatie waarvan de leden niet gedurende de gehele observatieperiode at risk hoeven te zijn.

Dit document is als volgt opgebouwd. Hoofdstuk 2 bespreekt het recurrent events model, en het gebruik hiervan voor het schatten van een populatieomvang. Hoofdstuk 3 beschrijft de modelassumpties, en uitbreidingen van het recurrent events model om voor deze schendingen te corrigeren. Hoofdstuk 4 presenteert de resultaten van de simulatiestudies. Hoofdstuk 5 bespreekt de populatieschattingen van de verschillende modellen aan de hand van een voorbeeld uit de praktijk; de illegalen vreemdelingenpopulatie in 2009. De kosten en baten van de verschillende modellen worden besproken in Hoofdstuk 6. De Bijlage beschrijft de preparatie van het databestand van de illegale vreemdelingen over 2009.

Hoofdstuk 2

Het recurrent events model

In de onderstaande paragrafen wordt een beknopte algemene inleiding van het recurrent events model gegeven. Hierbij wordt ingegaan op de overeenkomsten en verschillen met het Poissonmodel bij het modelleren van individuele verschillen en van periodes 'at risk'. De laatste paragraaf bespreekt de afgeknotte versie van het recurrents event model, en de schatting van de populatieomvang die op dit model is gebaseerd.

2.1 Recurrent events

De term 'recurrent events' word gebruikt voor gebeurtenissen die zich met een zekere regelmaat herhalen. De analyse van recurrent events speelt een rol in verschillende disciplines. Vroege voorbeelden hiervan zijn de emissie van radioactieve deeltjes, het voorkomen van aardbevingen of vulkaanuitbarstingen, en de uitbraak van bepaalde ziektes. Recentelijk zijn de analysetechnieken voor recurrent events data uitgebreid met methoden die individuele variabiliteit toelaten middels de opname van covariaten of random effecten (Lawless, 1995). Voorbeelden hiervan zijn te vinden op het gebied van de medische, sociale en technische wetenschappen (Cook and Lawless, 2007).

Het recurrent events model beschrijft de kans op de 'event history' van y gebeurtenissen op de tijdstippen $t_1, \ldots, t_y$ als

$$P(y, t_j|T) = \left(\prod_{j=1}^{y} \lambda(t_j) \right) e^{-\Lambda(T)} \tag{2.1}$$

waarbij λ de intensiteit is waarmee de gebeurtenis optreedt, en Λ(T) de totale intensiteit gemeten over de observatieperiode (t_0, T). Indien we voor het

gemak aannemen dat tijd discreet is, met bijvoorbeeld 1 dag als tijdseenheid en een observatieperiode $(1, 365)$ van één jaar, dan is de totale intensiteit gelijk aan

$$\Lambda(T) = \sum_{j=1}^{365} \lambda(k), \tag{2.2}$$

waarbij $\lambda(k)$ de intensiteit op dag k is. Deze vergelijking laat zien dat in het recurrent events model aannames moeten worden gemaakt t.a.v. intensiteit op dag k. De basisaanname is dat de intensiteit constant is en dus op alle dagen dezelfde waarde heeft, maar er zijn ook andere aannames mogelijk (bijvoorbeeld dat de intensiteit toe- of afneemt in de tijd).

Indien $\lambda(k)$ gelijk is voor alle $k = 1, \ldots, 365$ dagen, dan kan worden aangetoond (zie bijvoorbeeld Cook and Lawless, 2007) dat model (2.1) even informatief is als het Poissonmodel

$$P(y) = \frac{\Lambda^y e^{-\Lambda}}{y!}. \tag{2.3}$$

Het ontbreken van $\lambda(k)$ in dit model impliceert dat het Poissonmodel een constante intensiteit in de tijd veronderstelt. Indien $\lambda(k)$ constant is in de tijd en de populatie gesloten, dan geven het recurrent events model en het Poissonmodel dezelfde schatting voor de totate intensiteit Λ.

2.2 Individuele verschillen

In de voorgaande paragraaf zijn we er steeds vanuit gegaan dat de intensiteit gelijk is voor alle individuen in de populatie. Met de opname van covariaten kunnen individuele verschillen in de intensiteit gemodelleerd worden. We onderscheiden hierbij tussen *tijdsonafhankelijke* en *tijdsafhankelijke* covariaten, en binnen de laatste groep tussen *interne* en *externe* covariaten.

Het kenmerk van tijdsonafhankelijke covariaten is dat de waarde ervan gedurende de observatieperiode niet verandert in de tijd (zoals bijvoorbeeld geslacht). De score van individu i op de m-de tijdsonafhankelijke covariaat wordt aangeduid met x_{im}. Bij tijdsafhankelijke kan de waarde gedurende de observatieperiode wel veranderen. Een tijdsafhankelijke covariaat is extern indien de waarde ervan op tijdstip t_j onafhankelijk is van de event history tot aan tijdstip j. In een studie naar het aantal ziekenhuisopnames n.a.v. ademhalingsproblemen is de mate van luchtverontreiniging bijvoorbeeld een externe covariaat. Rookgedrag is in dit geval een interne covariaat, omdat

dit zeer waarschijnlijk beïnvloed wordt door eerdere ziekenhuisopnames. De score van individu i op de l-de tijdsafhankelijke covariaat wordt aangeduid met z_{ikl}, waarbij $k = 1, \ldots, T$.

De intensiteit van persoon i op tijdstip k is gelijk aan

$$\lambda_i(k) = \lambda_0 \exp(x_{i1}\beta_1 + \ldots + x_{iM}\beta_M + z_{ik1}\gamma_1 + \ldots + z_{ikL}\gamma_L) \tag{2.4}$$

waarbij λ_0 de basisintensiteit is. De totale intensiteit voor persoon i is dan gelijk aan

$$\Lambda_i = \sum_{k=1}^{365} \lambda_i(k). \tag{2.5}$$

2.3 Tijd 'at risk'

In het voorgaande is steeds aangenomen dat bij de personen in de steekproef op elke dag van het jaar een gebeurtenis kon optreden. Het kan echter zijn dat op bepaalde dagen van het jaar er geen gebeurtenissen kunnen optreden, omdat de persoon tijdelijk niet in de populatie aanwezig is geweest. De tijd waarin er wel gebeurtenissen kunnen plaatsvinden noemen we de tijd 'at risk'. We definiëren de indicator variabele $I_{ik}(risk)$, welke de waarde 1 aanneemt als de persoon i op dag k 'at risk' was en 0 als dat niet het geval was. Bij afwezigheid uit de populatie wordt de dtottale intensiteit van persoon i berekend als

$$\Lambda_i = \sum_{k=1}^{365} I_{ik}(risk)\lambda_i(k). \tag{2.6}$$

waarbij

$$T_i = \sum_{k=1}^{365} I_{ik}(risk) \tag{2.7}$$

de tijd is dat deze persoon in de populatie aanwezig was.

2.4 Het afgeknotte model

Het afgeknotte model geldt wanneer de personen uit de populatie waarbij geen gebeurtenis is opgetreden niet zijn geobserveerd. In het afgeknotte Poissonmodel wordt hiervoor gecorrigeeerd door de kansen op de geobserveerde tellingen van $y_i = 1, 2, \ldots$ gebeurtenissen te delen door $1 - P(y_i =$

$0|\Lambda_i)$, waarbij $P(y_i = 0|\Lambda_i) = e^{-\Lambda_i(T)}$ de kans is op 0 gebeurtenissen voor een persoon met totale intensiteit $\Lambda_i(T)$. Volgens een vergelijkbaar principe vinden we het afgeknotte recurrent events model (Hu and Lawless, 1996) als

$$P(y_i, t_{ij}|T, y_i > 0) = \left(\prod_{j=1}^{y} \lambda(t_j) \right) \frac{e^{-\Lambda_i}}{1 - e^{-\Lambda_i^*}} \tag{2.8}$$

waarbij Λ_i de totale intensiteit van persoon i is zoals in gedefinieerd in (2.6), terwijl

$$\Lambda_i^* = \sum_{k=1}^{365} I_{ik}^*(risk)\lambda_i(k). \tag{2.9}$$

de totale intensiteit is van de niet-geobserveerde personen met een gelijk covariatenpatroon als persoon i.

De waarde van $I_{ik}^*(risk)$ wordt bepaald door de oorzaak van een eventuele tijdelijke afwezigheid uit de populatie van de geobserveerde persoon i. Afwezigheid heeft een *externe* oorzaak indien deze niet is gerelateerd aan de event history van persoon i. In dat geval nemen we aan dat $I_{ik}(risk) = I_{ik}^*(risk) = 1$, zodat $\Lambda_i = \Lambda_i^*$. Een voorbeeld hiervan is vakantie; naar verwachting gaan geobserveerde en niet-geobserveerde personen met gelijke covariaten even lang op vakantie. Afwezigheid heeft een *interne* oorzaak als deze wel is gerelateerd aan de event history. We nemen nu aan dat $I_{ik}(risk) = 0$ terwijl $I_{ik}^*(risk) = 1$, zodat $\Lambda_i < \Lambda_i^*$. Een voorbeeld hiervan is detentie a.g.v. het plegen van een delict; een geobserveerde (lees 'gepakte') delinquent verdwijnt hierdoor een tijd uit de populatie, maar een niet-geobserveerde delinquent blijft gewoon in de populatie aanwezig.

2.5 Omvangschatting

Op basis van de schattingen van de Λ_i kan vervolgens de totale populatieomvang N worden geschat met de Horvitz-Thompson schatter

$$\hat{N} = \sum_{i=1}^{n} \frac{1}{1 - e^{-\hat{\Lambda}_i^*}}. \tag{2.10}$$

waarbij $\hat{\Lambda}_i^*$ de geschatte totale intensiteit is voor de niet-geobserveerde personen, en $e^{-\hat{\Lambda}_i^*} = P(y_i = 0)$. Als die kans op bijvoorbeeld 3/4 wordt geschat, dan zijn volgens de Horvitz-Thompson schatter 3 van de 4 personen met

gelijke covariaten niet geobserveerd, en is de populatieschatting m.b.t. de geobserveerde persoon i gelijk aan $1/(1-3/4) = 4$.

Om inzicht te krijgen in de kwaliteit van de puntschatting is de schatting van een betrouwheidsinterval van belang. Een betrouwbaarheidsinterval van 95% geeft aan dat, bij herhaling van het onderzoek, de ware populatieomvang in ongeveer 95% van de gevallen binnen het het geschatte 95% betrouwbaarheidsinterval zou liggen. Van der Heijden *et al* (2003) beschrijft een methode voor het vinden van het 95% betrouwbaarheidsinterval voor het afgeknotte Poissonmodel. Voor het afgeknotte recurrent events model kan in principe dezelde methodiek worden gehanteerd.

Merk op dat in het geval van een gesloten populatie met een constante $\lambda(k)$ voor het afgeknotte recurrent events model (2.8) geldt dat $\hat{\Lambda}^* = \hat{\Lambda}$. Bij de bespreking van het Poissonmodel in (2.3) kwam al naar voren dat in deze situatie de schattingen van de totale intensiteiten van het Poissonmodel en recurrent events model identiek zijn. Hieruit volgt dus dat – in geval van een gesloten populatie en constante intensiteit $\lambda(k)$ – de populatieschattingen van het afgeknotte Poissonmodel en het afgeknotte recurrent events model identiek zijn.

Hoofdstuk 3

Modelleren van effecten

In dit hoofdstuk gaan we in op het modelleren van speciale effecten m.b.v. het (afgeknotte) recurrent events model. De volgende effecten komen aan bod:

- niet-constante intensiteit
- niet-gesloten populatie
- seizoenseffecten
- ongeobserveerde heterogeniteit
- besmetting

3.1 Niet-constante intensiteit

We hebben gezien dat de intensiteit kan veranderen in de tijd als gevolg van tijdsafhankelijke covariaten. In die zin is een niet-constante intensiteit dus geen schending van het model. In sommige situaties kan er echter een probleem optreden bij het bepalen van de verwachte waarde Λ_i. Uit de vergelijkingen (2.4) en (2.6) valt op te maken dat de waarde van de covariaten $z_{i1k}, \ldots, z_{ikL}$ op alle k dagen wordt geacht bekend te zijn. In de praktijk zal dit echter meestal niet het geval zijn. Het voorbeeld van staandegehouden delinquent kan dit wederom verduidelijken. Voor deze persoon zijn op het tijdstippen t_{ij} (de dagen waarop de aanhoudingen $j = 1, \ldots, y_i$ plaatshadden) een aantal gegevens bekend die als covariaat kunnen dienen, waaronder bijvoorbeeld de verschillende politieregio's waar de aanhoudingen plaatsvonden. Voor de dagen waarop de persoon is aangehouden is de politieregio

bekend, maar voor de overige dagen niet. Deze informatie is echter wel van belang om Λ_i te kunnen berekenen, want daarvoor dienen de intensiteiten $\lambda_i(k)$ voor alle dagen k dat de persoon in de populatie aanwezig was bekend te zijn. De waarden van deze parameters worden immers (mede) bepaald door de politieregio waar de persoon zich op dag k bevond. In die zin is dit dus een 'missing data' probleem; het modelleren van de niet-constante intensiteit wordt bemoeilijkt doordat de noodzakelijke gegevens grotendeels ontbereken.

3.2 Niet-gesloten populatie

De assumptie van een gesloten populatie houdt in dat er gedurende de observatieperiode (t_0, T) geen migratie in of uit de populatie plaatsvindt. Het recurrent events model kan voor schendingen van deze assumptie corrigeren middels de indicator variabelen $I_{ik}(risk)$ en $I^*_{ik}(risk)$, mits voor de geobserveerde personen zowel de afwezigheid als de oorzaak van de afwezigheid bekend zijn. We onderscheiden latere instreding, vervroegde uittreding en tijdelijke afwezigheid, en de mogelijke oorzaken hiervan (zie ook Figuur 3.1):

Latere toetreding: Persoon i treedt op tijdstip $t_{i0} > t_0$ toe tot de populatie (zie event history B in Figuur 3.1). Latere toetreding gaat vooraf aan de event history, en is heeft dus per definitie een externe oorzaak.

Vervroegde uittreding: Persoon i verlaat op tijdstip $t_{i(end)} < T$ de populatie. Als de uittreding gerelateerd is aan het optreden van event y_i (bv. in geval van uitzetting van een illegale vreemdeling), dan is de oorzaak intern en is $T_{y_i(end)} = T - t_{y_i(end)}$ de corresponderende periode van afwezigheid (zie event history D in Figuur 3.1). Als de definitieve uittreding een externe reden heeft en dus niet is gerelateerd aan een event, dan zal over het algemeen $t_{i(end)}$ onbekend zijn (zie event history C in Figuur 3.1).

Tijdelijke afwezigheid: Persoon i is gedurende de periode $(t_{i(out)}, t_{i(back)})$ tijdelijk niet in de populatie aanwezig. Als tijdelijke afwezigheid het gevolg is van event j, voor $j = 1, \ldots, y_i$ (zie event history F in Figuur 3.1), dan is de oorzaak intern en is de periode van afwezigheid $T_{ij(out)}$ gelijk aan $(t_{ij(out)}, t_{ij(back)})$ of aan $(t_{y_i(out)}, T)$ als $t_{y_i(back)} > T$. Als tijdelijke afwezigheid een externe oorzaak heeft, dan zullen over het algemeen $t_{i(out)}$ en $t_{i(back)}$ onbekend zijn (zie event history E in Figuur 3.1).

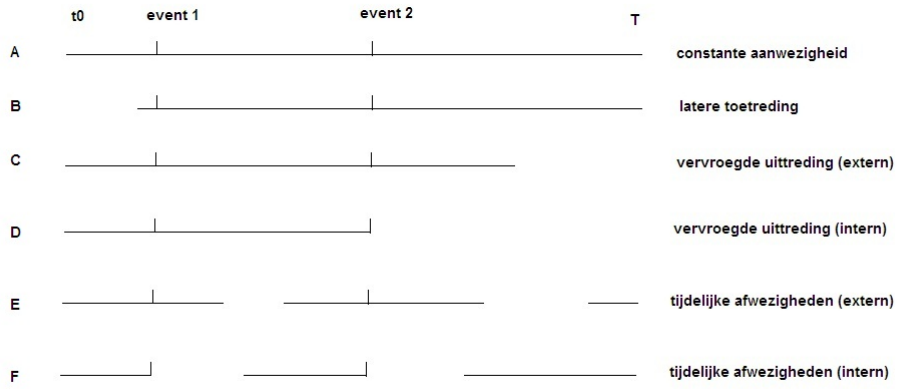
Als we ervan uitgaan dat t_{i0} alleen bekend is voor personen met minimaal 1 gebeurtenis, en dat de overige twee alleen bekend zijn indien gerelateerd aan een gebeurtenis, dan kan voor deze personen de tijd at risk worden berekend als

$$T_i = T_{y_i(end)} - t_{i0} - \sum_{j=1}^{y_i} T_{ij(out)},$$

waarbij $T^*_{ij(out)} = (t_{y_i(out)}, T)$ als $t_{y_i(back)} > T$. Op basis van deze vergelijking kan de verwachte waarde Λ_i voor de geobserveerde personen worden bepaald. Voor de niet-geobserveerde personen geldt dat alleen t_{i0} niet aan de gebeurtenissen is gerelateerd, en dat dus

$$T^*_i = T - t_{i0},$$

dient te worden gehanteerd voor de bepaling van Λ^*_i.



Figuur 3.1: Event histories met afwezigheden (in- en extern) voor een persoon met twee events. De totale lengte van de horizontale lijn(en) geeft de totale tijd T_i weer dat de persoon in de populatie aanwezig is geweest.

3.3 Seizoenseffecten

Een speciaal geval van een niet-gesloten populatie is een seizoenseffect, omdat we dan te maken hebben met een mengsel van twee populaties, één die alleen in het hoogseizoen aanwezig is (we zullen deze voor het gemak aanduiden met de 'hoogseizoeners'), en één die permanent in de populatie aanwezig is (aangeduid met 'permanenten'). Stel dat het hoogseizoen de periode $H = (t_0, t_H)$ omvat en het laagseizoen de periode $L = (t_H + 1, T)$, en dat

$$\pi_i = \frac{\exp(\alpha_0 + x_{i1}\alpha_1 + \ldots + x_{iK}\alpha_K)}{1 + \exp(\alpha_0 + x_{i1}\alpha_1 + \ldots + x_{iK}\alpha_K)}$$

de kans is dat persoon i een hoogseizoener is. Voor hoogseizoeners is de kans op de events $j = 1, \ldots, y_i$ op tijdstippen t_{ij} gelijk aan

$$P(y_i, t_{ij}|H, y_i > 0) = \left(\prod_{j=1}^{y_i} \lambda_i(t_{ij}) \right) \frac{e^{-\Lambda_i(H)}}{1 - e^{-\Lambda_i^*(H)}},$$

waarbij $\Lambda_i(H)$ en $\Lambda_i^*(H)$ zijn gedefinieerd als in (2.6) en (2.9) voor de periode H. Voor permanenten kunnen events op elk tijdstip in T optreden, en is de kans op de events $j = 1, \ldots, y_i$ op tijdstippen t_{ij} gelijk aan

$$P(y_i, t_{ij}|T, y_i > 0) = \left(\prod_{j=1}^{y_i} \lambda(t_{ij}) \right) \frac{e^{-\Lambda_i(T)}}{1 - e^{-\Lambda_i^*(T)}},$$

waarbij $\Lambda_i(T)$ en $\Lambda_i^*(T)$ zijn gedefinieerd als in (2.6) en (2.9) voor de totale observatieperiode T.

De kans voor een hoogseizoener om in de steekproef terecht te komen (ofwel de kans op minimaal één event) is dan gelijk aan

$$\theta_i = \frac{\pi_i(1 - e^{-\Lambda_i^*(H)})}{\pi_i(1 - e^{-\Lambda_i^*(H)}) + (1 - \pi_i)(1 - e^{-\Lambda_i^*(T)})},$$

waarbij de noemer de kans op minimaal één event voor een hoogseizoener aangeeft, en de teller de kans op minimaal één event voor de populatie van hoogseizoeners en permanenten tezamen (vergelijk Böhning en Kuhnert, 2006). Het model voor seizoenseffecten is dan

$$\begin{aligned} P(y_i, t_{ij}|T, y_i > 0) = I(t_{ij} \notin L) \cdot \theta_i P(y_i, t_{ij}|H, y_i > 0) \\ +(1 - \theta_i) P(y_i, t_{ij}|T, y_i > 0), \end{aligned} \tag{3.1}$$

waarbij de indicator $I(t_{ij} \notin L)$ de waarde 1 aanneemt als persoon i géén events heeft in het laagseizoen L, en 0 als deze wel een event heeft in het laagseizoen. Deze indicatior is in het model opgenomen omdat hoogseizoeners geen events kunnen hebben in het laagseizoen. De schatting van de populatieomvang is dan gelijk aan

$$\hat{N} = \sum_{i=1}^{n} \frac{1}{1 - \hat{\pi}_i e^{-\hat{\Lambda}_i^*(H)} - (1 - \hat{\pi}_i) e^{-\hat{\Lambda}_i^*(T)}}.$$

3.4 Ongeobserveerde heterogeniteit

Afwezigheid van heterogeniteit impliceert dat de covariaten in het model de individuele verschillen in intensiteit in voldoende mate verklaren. Indien er echter een belangrijke covariaat ontbreekt, dan is er sprake van ongeobserveerde heterogeniteit. Van der Heijden et al (2003) hebben aangetoond dat ongeobserveerde heterogeniteit leidt tot een onderschatting van de populatieomvang. Ongeobserveerde heterogeniteit kan worden gemodelleerd door een random effect μ_i aan het model toe te voegen zodat

$$\lambda_i^{RE}(k) = \mu_i \lambda(k)$$

Onder de aanname dat μ_i een gammaverdeling heeft, wordt een negatief binomial model verkregen. Dit model is in het verleden toegepast op data met een telvariabele (met het totaal aantal gebeurtenissen), maar het model bleek, behalve voor een populatie druggebruikers (zie Cruyff et al, 2008), niet te schatten. Het is daarom onwaarschijnlijk dat het negatief binomiale recurrent events model wel schatbaar is.

3.5 Afwezigheid van besmetting

Van besmetting is sprake wanneer de (afwezigheid van een) gebeurtenis van invloed is op de intensiteit. In de praktijk betekent dit dat een persoon als gevolg van een gebeurtenis (of van het uitblijven van een gebeurtenis) zijn of haar gedrag zodanig verandert dat de kans op een nieuwe gebeurtenis toeneemt (positieve besmetting) of afneemt (negatieve besmetting). Er bestaan in principe mogelijkheden om dit soort effecten in het model op te nemen, maar deze vallen buiten het kader van dit project. Hiervoor zijn een aantal redenen aan te wijzen. In de eerste plaats bestaat er nog geen uitgewerkte theorie op dit gebied, waardoor de ontwikkeling en toetsing van

een dergelijk model meer tijd zal kosten dan binnen het kader van dit project beschikbaar is. Ten tweede is het praktisch nut van een dergelijk model twijfelachtig, omdat vaak niet duidelijk is of er in de populatie sprake is van positieve of negatieve besmetting. Zo is het onduidelijk of de intensiteit van een niet-uitzetbare illegale vreemdeling na een staandehouding toeneemt omdat de persoon merkt dat er toch geen uitzetting volgt en zich dus vrijer gaat bewegen, of dat deze afneemt omdat de persoon zich meer gaat schuilhouden om zo toekomstige staandehoudingen te vermijden.

Hoofdstuk 4

Simulatiestudies

Om de werking van het Poissonmodel, het tweetraps Poissonmodel, het recurrent events model en het recurrent events seizoenmodel te onderzoeken zijn de volgende simulatiestudies uitgevoerd:

1. Poissonmodel versus tweetraps Poissonmodel
2. Tijdelijke afwezigheid (intern)
3. Latere intreding (extern) en voortijdige uittreding (intern)
4. Combinatie van 1 en 2
5. Seizoenseffecten

Het doel van simulatiestudie 1 is te onderzoeken in hoeverre het tweetraps Poissonmodel een verbetering is van het gewone Poissonmodel in het geval dat er sprake is van vervroegde uitreding uit de populatie. Het doel van simulatiestudies 2 t/m 5 is tweeledig; de studies dienen (1) aan te tonen dat het recurrent events (seizoens)model correct werkt, d.w.z. dat het consistente schattingen geeft indien correct gespecificeerd, en (2) inzicht te geven in de mate waarin de schattingen van het Poissonmodel afwijken van de ware omvang. Hierbij dient men zich te realiseren dat de grootte van de afwijkingen specifiek is voor de waarden van de populatieparameters zoals die in de simulatiestudie zijn gespecificeerd, en zij daarom meer in relatieve dan absolute zin dienen te worden geïnterpreteerd.

Ten behoeve van de simulatiestudies is volgende notatie gehanteerd:

- $(t_0, T) = (0, 365)$: observatieperiode
- $\lambda(k) = 0.002$: kans op een gebeurtenis per dag
- $N \in \{1000, 10000\}$: ware populatieomvang
- $T_{j(out)} \in \{20, 40, 60\}$: perioden tijdelijke afwezigheid a.g.v. event j
- $P(t_{i0} > 0) = .25$: kans op latere instroom: $t_{i0} \sim \text{Uniform}(1, 364)$
- $P(T_{y_i(end)} < T) = .25$: kans definitieve uitstroom a.g.v. event $j = 1, 2, \ldots$.
- $H = (0, 100)$: hoogseizoen, $L = (101, 365)$: laaggseizoen
- y_{iH} events van i in H, y_{iL} events van i in L
- $\pi \in \{.25, .5\}$: kans op aanwezigheid in H maar niet in L

De simulatiestudies worden hieronder afzonderlijk besproken. Bij iedere simulatiestudie wordt kort aangegeven hoe de data zijn gegenereerd en wat het gefitte model is. Voor elke condities wordt het gemiddelde en de root mean squared error (RMSE) van de puntschattingen de coverage (het percentage van de betrouwbaarheidsintervallen dat de ware omvang omvat) gerapporteerd.

4.1 Simulatie 1: Poisson versus tweetraps Poisson

Het Poissonmodel wordt geschonden door individuen die de populatie voortijdig verlaten. Indien bekend is welke van de geobserveerde individuen (individuen met minimaal één gebeurtenis) de populatie voortijdig hebben verlaten, dan kan het tweetraps Poissonmodel kan worden gebruikt. Door de parameters van het Poissonmodel te schatten exclusief deze groep (1ste trap), wordt voorkomen dat deze individuen de parameterschattingen op enige wijze beïnvloeden. Vervolgens wordt een schatting van de populatieomvang verkregen (2de trap van de analyse) met behulp van de Horvitz-Thompson (2.10). In deze stap doen de voortijdige verlaters van de populatie wel weer mee. Het idee achter deze kunstgreep is dat de parameterschattingen door het buiten beschouwing laten van de voortijdige verlaters minder gebiased zullen zijn, waardoor ook de populatieschatting minder gebiased zal zijn.

Er zijn twee simulaties uitgevoerd waarin de prestaties van het tweetrapsmodel t.o.v. het gewone Poissonmodel worden onderzocht. In simulatie A zijn afgeknotte steekproeven gesimuleerd uit een populatie met intensiteit $\lambda_i(k) = \lambda_0 \exp(0.5x_{i1})$, waarbij $\lambda_0 = 0.005$ en $X_1 \sim N(0,1)$. De kans om de populatie voortijdig te verlaten is gerelateerd aan de gebeurtenissen, en is gelijkgesteld aan 25% per gebeurtenis. In simulatie B is $\lambda_i(k) = \lambda_0 \exp(0.5x_{i1} + .25X_{i1})$, waarbij X_2 een dichotome variabele is met $P(X_2 = 0) = P(X_2 = 1) = 0.5$. In deze simulatie is de kans op het voortijdig verlaten van de populatie niet alleen gerelateerd aan de gebeurtenis maar ook aan de score op X_2; personen met score 0 op X_2 blijven met kans 1 in de populatie aanwezig, terwijl personen met score 1 op X_2 per gebeurtenis een kans van 75% hebben om de populatie vortijdig te verlaten. De resultaten van deze simulaties zijn getoond in Tabel 4.1.

In simulatie A geeft het Poissonmodel een overschatting van de ware omvang en een te lage coverage van het 95% betrouwbaarheidsinterval. Het tweetraps Poisson model doet het aanzienlijk beter, en geeft vrijwel perfecte schattingen. In simulatie B overschat het Poissonmodel de ware omvang sterk. Het tweetrapsmodel doet het aanzienlijk beter, maar minder goed dan in simulatie A. Het betrouwbaarheidsinterval is van dit model is te ruim, wat het gevolg is van het feit dat er minder observaties zijn gebruikt bij het schatten van de parameters dan wanneer het gewone Poissonmodel zou zijn gebruikt. Deze resultaten laten zien dat het tweetraps Poissonmodel meer valide schattingen geeft dan het Poissonmodel in geval van het voortijdig verlaten van de populatie. Ze laten ook zien dat de kwaliteit van de schattingen in hoge mate afhangt van de vraag of het voortijdig verlaten

Tabel 4.1: Populatieschattingen, RMSE en coverage percentages.

		$\hat{N}$ (RMSE)		coverage	
	N	Poisson	Poisson*	Poisson	Poisson*
A	1000	1318 (374)	1042 (150)	69%	96%
	10000	12831 (2936)	10094 (528)	0%	95%
B	1000	2911 (2651)	1388 (1017)	91%	98%
	10000	23904 (2936)	10483 (1458)	0%	96%

$*$ tweetraps model

van de populatie wel of niet gerelateerd is aan een predictor in het model; in het laatste geval leidt neemt de mate van overschatting van de omvang sterk toe.

In werkelijkheid zullen individuen de populatie niet allen voortijdig verlaten, maar ook later instromen en/of tussentijds afwezig zijn. Omdat er enkel vervroegde uittreding is gesimuleerd, geven de bovenstaande resultaten dus een optimistisch beeld van de prestaties van beide modellen. In de volgende simulatiestudies simuleren we ook latere toetreding en/of tussentijdse afwezigheid, en vergelijken we de prestaties van het Poissonmodel met die van het recurrent events model.

4.2 Simulatie 2: Tijdelijke afwezigheid

Op elk event volgt steeds dezelfde periode $T_{j(out)}$ van tijdelijke afwezigheid. Iedere persoon i met $y_i > 0$ heeft dus y_i perioden $T_{ij(out)}$ van afwezigheid.

Data: $n \times 2$ matrix met rijen $(y_i, \sum_{j=1}^{y_i} T_{ij(out)})$ als volgt bepaald:

1. stel $y_i = 0$ en $T_{ij(out)} = 0$;
2. voor $i \in \{1, \ldots, N\}$ en $k \in \{1, \ldots, T\}$:
3. als event j voor i op dag k;
   - $y_i = y_i + 1$
   - $T_{ij(out)} = T_{j(out)}$ of $\min\left(T_{j(out)}, T - t_{ij(back)}\right)$
   - $k = k + T_{j(out)}$

Model:

$$P(y_i, t_{ij} | T_{ij(out)}, y_i > 0) = \left(\prod_{j=1}^{y_i} \lambda(t_{ij}) \right) \frac{\exp(-\Lambda_i)}{1 - \exp(-\Lambda_i^*)}$$

waarbij $\Lambda_i = \lambda(T - \sum_{j=1}^{y_i} T_{ij(out)})$ en $\Lambda_i^* = \lambda T$.

Tabel 4.2: Populatieschattingen, RMSE en coverage percentages

		$\hat{N}$ (RMSE)		coverage	
N	$T_{j(out)}$	Poisson	REM	Poisson	REM
1000	20	1098 (117)	1001 (55)	73%	95%
	40	1201 (216)	1002 (59)	25%	96%
	60	1330 (341)	1000 (63)	1%	96%
10000	20	10885 (909)	9977 (174)	2%	94%
	40	11942 (1955)	9979 (176)	0%	95%
	60	13196 (3209)	9992 (203)	0%	93%

Het Poissonmodel overschat de populatieomvang met circa 10% tot 30%, en heeft een veel te lage coverage. Het recurrent events model geeft consistente schattingen en een goede coverage.

4.3 Simulatie 3: Latere toetreding en vervroegde uittreding

Personen treden met kans .25 later tot de populatie toe, en verlaten na elk event met kans .25 voorgoed de populatie.

Data: $n \times 3$ matrix met rijen $(y_i, t_{i0}, T_{y_i(end)})$ als volgt bepaald:

1. stel $y_i = 0$, $t_{ij} = 0$ en $T_{y_i(end)} = 0$;
2. als i late instromer, dan $t_{i0} = \text{Uniform}(1, 364)$;
3. voor $i \in \{1, \ldots, N\}$ en $k \in \{t_{i0}, \ldots, T\}$:
4. als event j voor persoon i op dag k;
   - $y_i = y_i + 1$
   - $T_{y_i(end)} = T - k$ in geval vervroegde uitreding

Model:

$$P(y_i, t_{ij} | t_{i0}, T_{y_i(end)}, y_i > 0) = \left(\prod_{j=1}^{y_i} \lambda(t_{ij}) \right) \frac{\exp(-\Lambda_i)}{1 - \exp(-\Lambda_i^*)}$$

waarbij $\Lambda_i = \lambda(T - t_{i0} - T_{y_i(end)})$ en $\Lambda_i^* = \lambda(T - t_{i0})$.

Tabel 4.3: Populatieschattingen, RMSE en coverage percentages.

	$\hat{N}$ (RMSE)		coverage	
N	Poisson	REM	Poisson	REM
1000	1178 (201)	1014 (77)	52%	95%
10000	11734 (1756)	10083 (272)	0%	95%

Het Poissonmodel overschat de populatieomvang met circa 17%, terwijl het recurrent events model consistente resultaten te zien geeft. Het incorrecte recurrent events model met $\Lambda_i^* = \lambda T$ (niet getoond in tabel) onderschat de populatieomvang met gemiddeld 15%.

4.4 Simulatie 4: Latere toetreding, tijdelijke afwezigheid en vervroegde uittreding

Personen treden met kans .25 later tot de populatie toe, en verlaten na elk event de populatie tijdelijk voor een periode $T_{j(out)}$, of verlaten de populatie voorgoed met kans .25.

Data: $n \times 4$ matrix met rijen $(y_i, t_{i0}, T_{y_i(end)}, \sum_{j=1}^{y_i} T_{ij(out)})$ bepaald als in simulatie 1 en simulatie 2.

Model:

$$P(y_i, t_{ij} | t_{i0}, T_{y_i(end)}, T_{ij(out)}, y_i > 0) = \left(\prod_{j=1}^{y_i} \lambda(t_{ij}) \right) \frac{\exp(-\Lambda_i)}{1 - \exp(-\Lambda_i^*)}$$

waarbij $\Lambda_i = \lambda \left(T - t_{i0} - T_{y_i(end)} - \sum_{j=1}^{y_i} T_{ij(out)} \right)$ en $\Lambda_i^* = \lambda \left(T - t_{i0} \right)$.

Tabel 4.4: Populatieschattingen, RMSE en coverage percentages

		$\hat{N}$ (RMSE)		coverage	
N	$T_{j(out)}$	Poisson	REM	Poisson	REM
1000	20	1292 (305)	1009 (82)	13%	96%
	40	1432 (449)	1010 (85)	2%	93%
	60	1606 (625)	1020 (95)	0%	95%
10000	20	12838 (2855)	10042 (267)	0%	95%
	40	14198 (4215)	10014 (261)	0%	97%
	60	15834 (5852)	10045 (290)	0%	96%

Het Poissonmodel overschat de populatieomvang met circa 30 tot 60% en te lage coverage, terwijl het recurrent events model consistente schattingen geeft.

4.5 Simulatie 5: Seizoenseffecten zonder predictor

De populatie bestaat uit $\pi = .50$ hoogseizoeners, waarbij het hoogseizoen 100 van de 365 dagen in beslag neemt. In de steekproef is θ de kans op een hoogseizoener, en $1 - \theta$ de kans op een permanente verblijver.

Data: $n \times 2$ matrix met rijen $(y_i \in H, y_i \in L)$ als volgt bepaald:

1. stel $y_i \in H = y_i \in L = 0$;
2. voor $i \in \{1, \ldots, N\}$ en $k \in \{1, \ldots, T\}$:
   (a) als event j voor i op dag $k \in H$;
   - $y_i \in H = y_i \in H + 1$

   (b) als event j voor i op dag $k \in L$;
   - $y_i \in L = y_i \in L + 1$
3. $y_i = y_i \in H + y_i \in L$

Model: als gedefinieerd in (3.1).

Tabel 4.5: Populatieschattingen, RMSE en coverage percentages

		$\hat{N}$ (RMSE)		coverage	
N	π	Poisson	REM	Poisson	REM
1000	.25	885 (128)	1002 (73)	42%	95%
	.50	781 (228)	1009 (94)	10%	95%
10000	.25	8835 (1178)	10029 (235)	0%	96%
	.50	7777 (2231)	10037 (276)	0%	95%

Het Poissonmodel onderschat de populatieomvang, en het recurrent events model geeft consistente schattingen.

4.6 Simulatie 6: Seizoenseffecten met predictor

In deze simulatiestudie zijn 1000 random populaties van $N = 1000$ getrokken met de volgende eigenschappen:

- $T = (1, 365)$ is observatieperiode
- $H = (122, 243)$ is hoogseizoen
- $X_i \sim N(0, 1)$, voor $N = 1, \ldots, 1000$
- $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$, met $\beta = (-6, 0.5)$
- $\pi_i = \exp(\alpha_0 + \alpha_1 x_i)/\{1 + \exp(\alpha_0 + \alpha_1 x_i)\}$, met $\alpha_0 = (-1, -0.5)$
- $P(t_{i0} > 1) = 0.25$, waarbij $t_{i0} \sim U(1, 364)$ is de entreetijd
- $t = 30$ is de detentietijd volgend op event j (behalve in geval uitzetting)
- De kans op uitzetting volgend op event j is 0.25

Een random trekking van de covariaat X levert de waarden voor λ_i (de intensiteit) en π_i (de kans op een hoogseizoener), en op basis van die parameters wordt event history van elk persoon bepaald (wel of geen hoogseizoener, en de tijdstippen waarop een event plaatsheeft). Na verwijdering van de personen zonder events is het afgeknotte Poissonmodel, het recurrents eventsmodel en het recurrent events seizoensmodel op de data gefit.

Tabel 4.6: Gemiddelden (stdd) van de parameterschattingen

parameter	ware waarde	Poisson	REM	REMseizoen
β_0	-6.00	-0.78 (.10)	-6.39 (.10)	-6.00 (.10)
β_1	0.50	0.47 (.08)	0.53 (.08)	0.51 (.08)
α_0	-1.00	-	-	-1.07 (.32)
α_1	-0.50	-	-	-0.51 (.31)

Tabel 4.6 geeft een overzicht van de gemiddelden en standaarddeviaties van de parameterschattingen van deze modellen. Het Poissonmodel geeft een schatting voor het intercept β_0 van -0.78, maar dit model schat de totale intensiteit $\Lambda_i = \lambda_i \times 365$. Teruggerekend naar λ_i geeft dit een interceptschatting van ongeveer -6.68, hetgeen een onderschatting van het ware

intercept impliceert. De parameter β_1 wordt eveneens licht onderschat. In het Poissonmodel zijn de parameters α afwezig. Het recurrents eventsmodel geeft een iets geringere onderschatting van β_0, en een lichte verschatting van β_1. Ook in dit model zijn de α parameters afwezig. De schattingen voor de β en α parameters van het recurrent events seizoensmodel wijken nauwelijks van de ware waarden af, hetgeen aantoont dat dit model correct werkt.

Tabel 4.7: Gemiddelden (stdd) van de omvangschattingen

parameter	ware waarde	Poisson	REM	REMseizoen
N	1000	1202 (117)	915 (93)	1006 (112)

Tabel 4.7 presenteert de gemiddelde populatieschattingen (en standaarddeviaties) per model. Hieruit blijkt dat het Poissonmodel de omvang overschat, het recurrents events model de omvang onderschat. Het recurrent events seizoensmodel geeft correcte schattingen.

Hoofdstuk 5

Praktijkvoorbeeld

Als voorbeeld uit de praktijk zijn de data van de illegale vreemdelingen over 2009 genomen (de preparatie van deze data is beschrven in de Bijlage). Voor de geobserveerde personen zijn de absentieperioden bepaald aan de hand van detentiegegevens. Gemiddeld verbleven deze illegale vreemdelingen 242 (SD = 114) dagen in de populatie, hetgeen zo'n 66% van de totale observatietijd is. Tabel 5.1 geeft een overzicht van de schattingen van de populatie illegale vreemdelingen in 2009 (exclusief West-Europeanen) zoals verkregen met het Poissonmodel, het tweetraps Poissonmodel, het REM model en het REM seizoensmodel. Voor elk van deze modellen is zowel het nulmodel (met alleen het/de intercept(s)) als het volledige model (met alle beschikbare predictoren) gefit.

Tabel 5.1: Resultaten van de Poisson- en REM modellen

Nulmodel	# par	logl	Nhat	95% BI
Poisson	1	-849	46423	(40313, 52533)
Poisson*	1	-506	43456	(36204, 50705)
REM	1	-26782	18829	(16643, 21015)
REMseizoen	2	-26781	18907	**
Model met covariaten	# par	logl	Nhat	95% BI
Poisson	14	-829	61531	(44221, 78841)
Poisson*	13	-492	55660	(37567, 73763)
REM	14	-26763	22811	(17782, 27839)
REMseizoen	28	-26755	22839	**

$*$ het tweetrapsmodel

$**$ geen 95% betrouwbaarheidsinterval beschikbaar

Tabel 5.1 laat zien dat de modellen met de covariaten beter fitten en

tot hogere omvangschattingen leiden dan de corresponderende nulmodellen. Het gewone Poissonmodel met covariaten geeft een populatieschatting van 60000 en het Poisson tweetrapsmodel geeft een schatting van 55000. De schattingen van de REM modellen van rond de 23000 vallen aanzienlijk lager uit. Bewijs voor een seizoenseffect ontbreekt, aangezien het REM seizoensmodel niet beter fit dan het gewone REM model.

Tabel 5.2 toont de parameterschattingen van de nulmodellen (bovenste subtabel) en de volledige modellen (onderste subtabel). Het Poissonmodel schat een Poissonparameter voor het gehele jaar, terwijl het REM model een Poissonparameter per dag schat. Om de schattingen van het Poissonmodel te calibreren naar die van het REM model, kan het intercept worden omgezet tot $\hat{\beta}_0^* = \hat{\beta}_0 - \ln(365)$. Dit geeft voor het Poisson nulmodel een intercept $\hat{\beta}_0^* = -8.24$. Het Poissonmodel schat dus een veel kleinere Poissonparameter dan het REM model, hetgeen tot de hogere populatieschatting leidt. Het REM nulmodel met seizoenseffecten geeft een vergelijkbare schatting voor de Poissonparameter, en de $\hat{\alpha}_0 = -3.7$ impliceert een geschatte kans op hoogseizoeners in de populatie van ongeveer 2.4%.

Tabel 5.2: Parameterschattingen van de Poisson- en REM modellen.

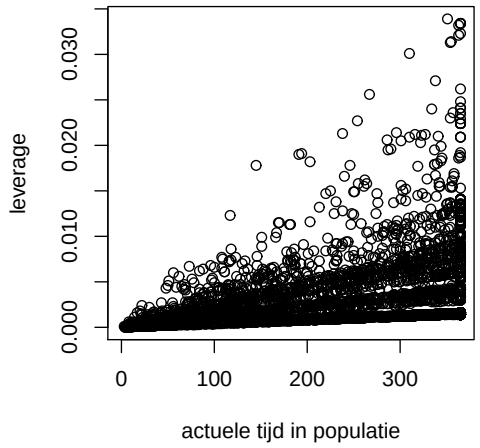
	Poisson	Poisson*	REM	REMseizoen	
predictor	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\alpha}$ (se)
constante	-2.34 (0.07)	-2.27 (0.08)	-7.25 (0.07)	-7.23 (0.66)	-3.70 (0.81)
constante	-2.65 (0.26)	-2.65 (0.31)	-7.65 (0.25)	-7.64 (-)	-5.30 (-)
Geslacht (man)	0.42 (0.25)	0.33 (0.28)	0.41 (0.24)	0.40 (-)	-0.55 (-)
Geslacht (vrouw)	0 (- -)	0 (- -)	0 (- -)	0 (-)	-0 (-)
Leeftijd (>40)	0.29 (0.17)	-0.31 (0.27)	0.37 (0.16)	0.38 (-)	-2.77 (-)
Leeftijf (≤ 50)	0 (- -)	0 (- -)	0 (- -)	0 (-)	-0 (-)
Regio (A'dam)	0.47 (0.21)	0.31 (0.33)	0.51 (0.20)	0.52 (-)	-2.60 (-)
Regio (R'dam)	0.15 (0.29)	0.37 (0.35)	0.25 (0.27)	0.25 (-)	-0.34 (-)
Regio (Haaglanden)	0.51 (0.23)	0.91 (0.28)	0.57 (0.22)	0.57 (-)	2.17 (-)
Regio (Utrecht)	-1.00 (0.58)	-0.96 (0.71)	-0.95 (0.57)	-0.92 (-)	-2.52 (-)
Regio (overige)	0 (- -)	0 (- -)	0 (- -)	0 (-)	-0 (-)
Nat (Turkije)	-1.66 (0.72)	-1.29 (0.73)	-0.91 (0.67)	-0.89 (-)	0.82 (-)
Nat (N-Afrika)	-0.89 (0.35)	-0.87 (0.53)	-0.54 (0.34)	-0.54 (-)	-1.50 (-)
Nat (Overig Afrika)	-0.14 (0.18)	0.16 (0.24)	-0.04 (0.17)	-0.04 (-)	4.04 (-)
Nat (Suriname)*	-1.72 (1.01)	- - (- -)	-1.71 (0.99)	-1.58 (-)	-1.21 (-)
Nat (Oost-EU)	-0.02 (0.25)	-0.09 (0.35)	0.07 (0.24)	0.07 (-)	0.38 (-)
Nat (Azie)	-0.17 (0.19)	0.34 (0.24)	-0.20 (0.18)	-0.20 (-)	-0.56 (-)
Nat (Amerika)	-0.54 (0.51)	0.02 (0.53)	0.02 (0.46)	0.01 (-)	-0.36 (-)
Nat (onbekend)	0 (- -)	0 (- -)	0 (- -)	0 (-)	-0 (-)

* de parameter voor Suriname kon niet worden geschat m.b.v. het tweetraps Poissonmodel

In de volledige modellen komen de schattingen van de β parameters redelijk overeen, waarbij een positieve parameter duidt op een grotere Poissonparameter dan die van de corresponderende referentiegroep. Zo zien we dat mannen en personen ouder dan 40 een grotere Poissonparameter hebben dan respectievelijk vrouwen en personen jonger dan 40. Bij het REM seizoensmodel ontbreken wegens numerieke problemen bij het berekenen van de informatiematrix de standaardfouten (deze problemen zijn mogelijk te verhelpen door de informatiematrix analystisch te berekenen), waardoor het niet mogelijk is om de significantie van de afzonderlijke parameters te bepalen. Voor de β parameters nemen we aan dat deze redelijk zullen overeenkomen met die het gewone REM model, en voor de α parameters nemen we aan dat grotere (absolute) waarden duiden op een groter effect. Een negatieve waarde duidt hier op een geringere kans op hoogseizoeners, hetgeen met name van toepassing is op personen ouder dan 40, personen die zijn staandegehouden in de regio's Amsterdam en Utrecht en personen met Noord-Afrikaanse nationaliteit. Een grotere kans op hoogseizoeners hebben personen die zijn staandegehouden in de regio Haaglanden en personen met de Noord-Afrikaanse nationaliteit.

Uit de analyses blijken de Poissonmodellen meer dan twee keer zo hoge omvangschattingen op te leveren dan de REM modellen. De verklaring hiervoor is dat de Poissonmodellen geen rekening houden met detentietijden. Als gevolg hiervan wordt de Poissonparameter (en met name het intercept) onderschat. Het tweetrapsmodel houdt wel rekening met uitzetting door de uitgezette personen in eerste instantie buiten het model te houden, hetgeen tot een iets lagere omvangschatting leidt. De schattingen van de REM modellen zijn echter aanzienlijk lager dan die van de Poissonmodellen.

Een belangrijke vraag die bij de lage schattingen van de REM modellen gesteld kan worden is hoe deze tot stand zijn gekomen; zijn hiervoor een klein aantal personen verantwoordelijk die uitzonderlijk kort in de populatie aanwezig zijn geweest, of hebben alle personen die niet gedurende het gehele jaar in de populatie aanwezig zijn geweest er evenredig aan bijgedragen. M.a.w., is dit resultaat een gevolg van een paar uitbijters, en zou door het buiten beschouwing van de uitbijters de populatieschatting weer sterk toenemen? Een statistiek die op deze vraag een antwoord geeft is de zogenaamde *leverage*. De leverage is een maat voor de invloed die een persoon op de parameterschattingen van het model heeft gehad, waarbij een hogere waarden een toenemende invloed representeren. Figuur 5.1 laat de



Figuur 5.1: Leverage als functie van de actuele tijd in de populatie.

leverage zien als een functie van de actuele tijd in de populatie. We zien dat in deze figuur een sterk positief verband tussen de actuele tijd in de populatie en de leverage. Dit betekent dat de personen die zeer kort in de populatie aanwezig zijn geweest slechts een geringe invloed hebben gehad op de parameterschattingen van het model. Een aanvullende analyse bevestigt deze concluisie. In deze analyse is een omvangschatting gemaakt met het REM model, waarbij de personen die het korst in de populatie aanwezig zijn geweest buiten beschouwing zijn gelaten. De schatting van dit model viel slechts enkele tientallen peronen hoger uit dan die het REM model met alle personen.

Hoofdstuk 6

Discussie

Dit document geeft een theoretische onderbouwing voor het schatten van de populatieomvang met het recurrent events model, en laat middels simulatiestudies zien dat het model - mits correct gespecificeerd - consistente schattingen geeft. In dat opzicht is het recurrent events model een duidelijke verbetering van het (tweetraps) Poissonmodel, dat weliswaar weer beter schattingen geeft dan het reguliere Poissonmodel, maar dat bij open populaties toch een overschatting van de ware omvang geeft.

De simulatiestudies laten ook zien hoe de populatieschatting in geval van een open populatie dienen te worden geïnterpreteerd. Indien het recurrent events model correct is gespecificeerd, dan geeft het een schatting van het aantal personen dat gedurende de gehele observatieperiode aanwezig is geweest. Het aantal personen dat op enig moment in de populatie aanwezig is geweest ligt dus lager.

Het praktijkvoorbeeld van de illegale vreemdelingen betreft een open populatie, en de analyses met de verschillende modellen laten duidelijke verschillen in de populatieschattingen zien. De schattingen van het recurrent events modellen zijn superieur aan die van de Poissonmodellen in de zin dat zij voor een open populatie corrigeren. Men moet echter voorzichtig zijn met de conclusie dat deze schattingen ook beter zijn, omdat er andere, niet-gemodelleerde effecten een rol kunnen spelen (zoals bijvoorbeeld niet-geobserveerde heterogeniteit of besmetting) die tot vertekende schattingen kunnen leiden, en omdat de 'event history' data mogelijk van mindere kwaliteit zijn. Zo was het bijvoorbeeld uit de detentiegegevens niet altijd even duidelijk wanneer een detentie begon of ophield, en of een gerapporteerde uitzetting ook daadwerkelijk was geëffectueerd.

In dat opzicht is het van belang om de kosten en baten van de verschil-

lende modellen nog eens op een rijtje te zetten. Als we er voor het gemak even van uitgaan dat de kosten voor het verzamelen van gegevens m.b.t. geslacht, leeftijd en nationaliteit voor alle modellen gelijk zijn, dan is het Poissonmodel is het goedkoopste model in termen van dataverzameling; voor dit model is alleen het totaal aantal gebeurtenissen per individu benodigd. Voor het tweetraps Poissonmodel is ook informatie nodig m.b.t. de vraag of het individu de populatie voortijdig heeft verlaten (het tijdstip waarop is daarbij niet van belang). Voor het recurrent events model is informatie over de event history benodigd. In het geval van het illegalenvoorbeeld was dat informatie over het totaal aantal gebeurtenissen, het tijdstip van toetreding tot de populatie, de duur van de tussentijdse afwezigheden, en het tijdstip waarop de populatie definitief wordt verlaten. Voor het recurrent events seizoensmodel dient er nog een uitsplitsing van deze gegevens te worden gemaakt naar laag- en hoogseizoen.

De keuze om het recurrents events model boven het Poissonmodel te prefereren is in hoge mate afhankelijk van de vraag hoeveel extra inspanning het verzamelen van de aanvullende gegevens vergt, en in hoeverre het gebruik van die gegevens tot betere schattingen leidt. Deze variabelen zullen per populatie verschillen, maar voor de populatie van illegale vreemdelingen speelden beide factoren een grote rol. Het verzamelen van de event histories bleek een zeer tijdrovende bezigheid, maar het gebruik van deze gegevens leidde wel tot een aanzienlijk lagere populatieschatting. Dit resultaat zet vraagtekens bij de validiteit van de schattingen van het Poissonmodel, en hier lijkt het verzamelen van de event histories dus de moeite waard.

Hoofdstuk 7

Literatuur

Böhning, D. and Kuhnert, R (2006). Equivalence of truncated count mixture distributions and mixtures of truncated count distributions. *Biometrics*, **62**, 1207-1215.

Böhning, D. and van der Heijden, P.G.M. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *Annals of Applied Statistics*, **3**, 595-610.

Cook, R.J. and Lawless, J.F. (2007) *The Statistical Analysis of Recurrent Events*, Springer.

Cruyff, M.J.L.F. and van der Heijden, P.G.M. (2008). Point and Interval Estimation of the Population Size Using a Zero-Truncated Negative Binomial Regression Model. *Biometrical Journal*, **50**, 1035-1050.

Gurmu, S. (1991), Test for detecting overdispersion in the positive Poisson regression model,*Journal of Business and Economic Statistics*, **9** 215222.

Lawless, J.F. (1995). The analysis of recurrent events for multiple subjects. *Applied Statistics*, **44**, 487-498.

Leerkes, A., van San, M., Engbersen, G., Cruyff, M. en van der Heijden, P. (2004). *Wijken voor illegalen: Over ruimtelijke spreiding, huisvesting en leefbaarheid*. Sdu Uitgevers, Den Haag.

Lewis, P.A.W. (1972). *Recent results in the statistical analysis of univariate point processes.* In Stochastic Point Processes, 1-54. Ed. P.A.W. Lewis. Wiley, New York.

Nelson, W.B. (2003). Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications. *ASA-SIAM Series on Statistics and Applied Probability*, **10**, Philadelphia.

Van der Heijden, P.G.M. Bustami, R., Cruyff, M.J.L.F., Engbersen, G. and H.C. van Houwelingen (2003). Point and Interval Estimation of the Population Size Using the Truncated Poisson Regression Model. *Statistical Modeling*, **3**, 1-18.

Bijlage A

Datapreparatie

Ger van Gils

Vooraf

Deze korte bijlage bevat een beschrijving van de aanmaak van de belangrijkste data ten behoeve van de schattingen van de populatie illegale vreemdelingen in Nederland in 2009 met het Recurrent Events Model. De ambitie van de datapreparatie was een volledig geschoond bestand te maken met valide tellingen van registraties van staandehoudingen van illegale vreemdelingen en een sluitend beeld te geven van het verloop van een jaar voor de vreemdelingen, inclusief perioden van detentie en perioden dat men om andere redenen uit de populatie was verdwenen. Dat is niet gelukt. De data uit de gebruikte databestanden PSHV, VBS, BVV en KMI over het jaar 2009 en eerder bevatten te veel hiaten en roepen vele moeilijk te beantwoorden vragen op om de gestelde ambitie te realiseren. Tegen het einde van het project is er voor gekozen om een werkbaar bestand te maken, dat wil zeggen een bestand waarmee schattingen waren te maken en de verdiensten van het REM waren uit te proberen.

De ambitie om een bestand te maken dat 'de werkelijkheid' van een jaar uit het leven van vreemdelingen zo goed mogelijk benadert, is daarmee onder druk komen te staan. Tijdens het uitvoeren van de analyses zijn bovendien in korte tijd pragmatische keuzes gemaakt om gebleken tegenstrijdigheden in de data op te lossen. Deze tegenstrijdigheden waren een gevolg van aangebrachte correcties in de data. De keuzen om deze op te lossen zijn onder tijdsdruk onvoldoende gedocumenteerd. Om die reden is een exacte reconstructie van aanmaak van de data gebruikt voor de schattingen niet meer mogelijk. Er is op die manier wel een databestand verkregen waarmee simulaties met het REM kon worden uitgevoerd.

In het navolgende wordt beschreven hoe de gegevens uit verschillende bestanden zijn te combineren tot één analysebestand en welke keuze daar bij zijn te maken. Het betreft gegevens over toelating en vooral verwijdering verkregen uit BVV en KMI en detentiegegevens verkregen van DJI (Dienst Justitiële Inrichtingen). Tot slot volgen enige opmerkingen over de validiteit van herhaalde registraties van staandehoudingen.

Gegevens over staandehoudingen

Het gegevensbestand dat is gebruikt voor het uitproberen van REM is hetzelfde bestand over het jaar 2009 dat voor de eerder in 2012 gerapporteerde schatting is gebruikt. De gegevens over aanhoudingen van illegale vreemdelingen zijn verkregen uit twee bronnen: in de eerste plaats uit PSH-V (PolitieSuite Handhaving Vreemdelingen), het landelijke registratiesysteem van de VreemdelingenPolitie (VP) en in de tweede plaats uit het VBS (Vreemdelingen Basis Systeem) van de Koninklijke Marechaussee (KMar).

Over de volgende aantallen geregistreerde vreemdelingen zijn gegevens verkregen. Het betreft vreemdelingen die op de datum van hun aanhouding of staandehouding geen rechtmatige verblijfstitel hadden, bovendien 'verwijderbaar' waren en dus illegaal in Nederland verbleven.

Tabel 1. Illegale vreemdelingen 2009: PSHV en VBS

	n	%
PSHV	2.889	67
VBS	1.421	33
PSHV en VBS	20	-
Totaal	4.330	100

In PSH-V zijn 2.909 staandehoudingen en aanhoudingen van illegale vreemdelingen voor het jaar 2009 geregistreerd en in VBS 1.441; 20 vreemdelingen hebben een registratie in zowel PSH-V als VBS.

Gegevens over toelating of verwijdering

Gegevens met betrekking tot de afhandeling van zaken van vreemdelingen die zijn staandegehouden door de politie of door de KMar, zijn gehaald uit de Basis Voorziening Vreemdelingen (BVV) en de Keten Management Informatie bestand (KMI). De bestanden registreren maatregelen of stappen in het proces van toelating of verwijdering van vreemdelingen. We noemen deze stappen en maatregelen hier 'acties'.

De BVV bevat (o.m.) informatie over (variabele 'verwijzing' in het PSHV-deel van de BVV en 'act_activiteitsoortcode' in het KMar-deel. De variabelennamen liggen echter niet vast en kunnen per extractie verschillen):

10 ontvangst aanvraag VVR (verblijfsvergunning)
11 beslissing op VVR-aanvraag
12 aanmelden bij aanmeldcentrum
13 beslissing op asiel-aanvraag
23 In bewaring stelling ter fine van uitzetting
24 beëindigen bewaring
25 bericht verwijdering

Het bestand bevat ook informatie over zaken als de wijze van verwijdering van een ongewenst vreemdeling uit het land, zoals bijvoorbeeld (variabele 'referentie'in het PSHV-deel van de BVV en 'act_referentiekenmerk' in het KMar-deel):

- Overgave na controle MTV aan landgrenzen
- Uitzetting
- Uitzetting vanuit strafrechttraject (conform VRIS-werkwijze)
- Vertrek onder toezicht MTV
- Vertrek onder toezicht van zelfmelder
- Zelfstandig de woonruimte verlaten in of na de vertrektermijn van de procedure

- Zelfstandig de woonruimte verlaten tijdens de procedure vóór het ingaan van de vertrektermijn
- Zelfstandig vertrek van een bij controle op uitreis illegaal gebleken vreemdeling

Deze gegevens zijn uit de BVV verkregen voor de vreemdelingenzaken uit 2009 aangeleverd door de politie uit PSHV en door de KMar uit VBS.

Het KMI (Keten Magement Informatie) bevat ook informatie over wijze van verwijdering, maar geeft tevens aan hoe deze wijzen zijn te classificeren als:

- zelfstandig vertrek onder toezicht
- aantoonbaar vertrek (gedwongen)
- overgave na controle aan landsgrenzen
- overschrijding vrije termijn

Deze laatste informatie is in de schattingen gebruikt om te bepalen of een vreemdeling daadwerkelijk uit de populatie illegaal in Nederland verblijvende vreemdelingen is verwijderd.

In vrijwel elke zaak worden meerdere stappen gezet. Een veel voorkomend verloop van een zaak is bijvoorbeeld: In bewaring stelling ter fine van uitzetting (code 23), gevolgd door beëindigen bewaring (24), weer gevolgd door bericht verwijdering (25). Het is tevens mogelijk dat in een zaak, of in elk geval betreffende één vreemdeling stappen worden herhaald. Bijvoorbeeld voor sommige vreemdelingen is meerdere malen een uitzetting of vertrek onder toezicht volgend op één staandehouding geregistreerd. Daarom is voor elke vreemdeling de laatste actie in het jaar 2009 geselecteerd. De navolgende paragrafen beschrijven hoe dat in zijn werk is gegaan.

BVV-KMar

De BVV-extractie betreffende KMar-geregistreerden bevatte 4266 records voor het jaar 2009. Daarvan zijn de records met de laatste actiedatum in 2009 geselecteerd. Dit waren 1594 records. Deze groep bevatte nog 158 duplicate cases, d.w.z. meerdere records voor één persoon.
In die gevallen is gekozen voor de records met de hoogste waarde voor de variabele activiteitsoortcode die opeenvolgende stappen in het proces van afhandeling van een vreemdelingenzaak representeert. Hogere waarden staan voor opeenvolgende stappen in het proces.

Op deze manier zijn 135 (duplicate) records verwijderd. Er resteren dan nog 23 duplicates. 20 hiervan hebben gelijke waarden op de variabelen ‘uitvoerdatum van de actie’, ‘act-activiteitsoortcode’, ‘act-referentiekenmerk’ en ‘act-organisatiecode’ en zijn dus volledige doublures. Deze zijn verwijderd. Vervolgens resteren nog 3 duplicates met verschillende waarden op de genoemde variabelen (behalve uitvoerdatum). Van deze 3 paren zijn de records geselecteerd met een waarde (en geen missing) en de records met hogere waarden en dus verdere stappen in het proces van afhandeling van de zaak. Tot slot resteren 1436 unieke records betreffende 1436 vreemdelingen.

BVV-PSHV

De selectie van de laatst in BVV geregistreerde actie in 2009 levert 3241 records op. Na verwijdering van 52 records betreffende een actie ondernomen vóór de geregistreerde staandehouding, resteren 3189 records.

Samenvoeging BVV PSHV KMar

Het PSHV deel BVV bevat 3189 records en het KMar deel 1436, samen 4608 records. 17 records zijn zowel in het PSHV- als in het KMar deel geregistreerd. De relevante variabelen in de verschillende delen hebben verschillende namen en zijn in het samengevoegde bestand gecombineerd. De betreft de variabelen:

- '(laatste) referentie' uit het PSHV-deel die correspondeert met de variabele 'act_referentiekenmerk' uit het KMar deel en
- de variabele 'laatste verwijzing' uit het PSHV-deel die correspondeert met 'act_activiteitsoortcode' uit het KMar deel.

De nieuwe variabele 'Laatste_referentie_PSHV_KMar' bevat 3355 geldige records en 1253 missings (totaal 4605 records) en de nieuwe variabele 'Laatste_verwijzing_PSHV_KMar' bevat 4605 records (geen missings).

KMI voor PSHV en KMar

Het KMI voor KMar 2009 bevat 1177 records en het deel voor PSHV 3520 records, samen 4697 records. Daarvan hebben 3191 records betrekking alleen op acties genomen in 2009 en 1506 op latere acties. Vervolgens zijn de laatste acties in 2009 geselecteerd. Deze vinden inderdaad allemaal plaats na de laatste observatiedatum. Er resteren dan 2746 records betreffende laatste acties in 2009. 58 daarvan zijn nog een duplicate. Deze zijn een gevolg van meerdere observatiedatums per vreemdeling, gevolgd door redundante informatie over laatste acties in 2009. Na verwijdering van de duplicates resteren 2688 records. De variabelen

- 'DMPK_RESULTAAT_pol(itie)' en 'Processtapresultaat_KMar',
- 'resultaat_pol' en 'ResultaatGroepDefinitief_KMar',
- 'datum_300_pol' en 'Datum_start_KMar'

uit respectievelijk het politie deel en KMar deel van KMI worden vervolgens samengevoegd tot de nieuwe variabelen 'Resultaat', 'Resultaatgroep' en 'datum_acti'e in het gecombineerde bestand.

Samenvoeging BVV – KMI (voor PSHV- en KMar-deel)

De beide bestanden, BVV en KMI (beide betreffende geregistreerden in PSHV en KMar) zijn gekoppeld met de variabelen: vreemdelingnummer, observatiedatum en datum van de laatste actie, zodat verschillende acties niet ten onrechte worden gekoppeld (in één record geplaatst). Het nieuwe bestand bevat 4954 unieke records en 2061 duplicates. De paren records met een duplicate bevatten gelijke waarden voor de variabelen 'laatste referentie PSHV KMar' afkomstig uit BVV of 'Resultaat' of 'Resultaatgroep' uit KMI.

De paren verschillen soms doordat maar één ervan een waarde op deze laatste variabelen bevat en de ander een waarde mist. Wanneer de records betreffende een vreemdeling een waarde op de deze

KMI variabelen ‘Resultaat’ of ‘Resultaatgroep’ bevat, is informatie opgenomen als geldig voor de betreffende vreemdeling. Zo is verondersteld dat er meer informatie is over de laatste stappen in de afhandeling van de zaken van de vreemdelingen, dan er in feite is geleverd. Deze aanpassing lijkt gerechtvaardigd omdat de acties geregistreerd in KMI zelden ver verwijderd zijn in de tijd van de laatst geregistreerde actie (voor 90% van de acties is het verschil tussen de uitvoeringsdatum en de laatste uitvoeringsdatum maximaal 10 dagen). We doen de waarheid waarschijnlijk geen grof geweld door acties geregistreerd in KMI als finale actie voor het jaar 2009 te beschouwen. Aldus is alle KMI-informatie uit verschillende (duplicate) records betreffende een vreemdeling in het record van de laatste actie geconcentreerd. Vervolgens is dit record geselecteerd.

Voor 5194 records is de uitvoeringsdatum van de actie tevens de laatste datum van uitvoering van een actie in het jaar 2009. Voor 1821 records is geen informatie over de uitvoeringsdatum bekend. Deze zijn verwijderd. De 5194 records bevatten nog 240 duplicate cases. Deze verschillen niet wat betreft de BVV-variabelen ‘Laatste-referentie_PSHV-KMar’ en ‘Laatste-verwijzing-PSHV-KMar’ en zijn gelijkgesteld voor de KMI-variabelen ‘Resultaat’ en ‘Resultaatgroep’. De duplicates zijn verwijderd en er resteren 4954 unieke records met informatie over de afhandeling van vreemdelingenzaken afkomstig uit BVV en KMI. Vervolgens is voor de records waarvoor een waarde voor de variabele Resultaatgroep ontbreekt, een waarde toegekend op basis van de informatie in de BVV-variabele ‘Laatste referentie PSHV KMar’.

De acties vermeld in deze laatste variabele:
- Aanzegging Nederland te verlaten,
- Opheffing IBS met aanzegging Nederland te verlaten,
- Zelfstandig de woonruimte verlaten in of na de vertrektermijn van de procedure en
- Zelfstandig de woonruimte verlaten tijdens de procedure vóór het ingaan van de vertrektermijn,

zijn gecategoriseerd als ‘Zelfstandig vertrek zonder toezicht’.

De acties:
- Vertrek onder toezicht van zelfmelder en
- Zelfstandig vertrek van een bij controle op uitreis illegaal gebleken vreemdeling

zijn gecategoriseerd als ‘Zelfstandig vertrek onder toezicht’ en Overgave na controle MTV aan landgrenzen.

Uitzetting en Uitzetting vanuit strafrechttraject (conform VRIS-werkwijze) als ‘aantoonbaar vertrek (gedwongen)’.

Het aantal geldige waarden voor deze variabele ‘Resultaatgroep’ neemt hiermee toe van 2669 tot 2881.

Tabel 4. Resultaatgroep

	Aantal	%
1 zelfstandig vertrek zonder toezicht	716	14,5
2 zelfstandig vertrek onder toezicht	132	2,7
3 aantoonbaar vertrek (gedwongen)	1561	31,5
4 rechtmatig verblijf	50	1,0
6 in procedure	413	8,3
8 overgave na controle aan landsgrenzen	3	,1
9 overschrijding vrije termijn	6	,1
Totaal	2881	58,2
Missende waarden	2073	41,8
	4954	100,0

De informatie in deze variabele is gebruikt om te bepalen voor welk deel van het jaar een vreemdeling uit de populatie is vanwege verwijdering uit het land. Voor deze periode is aangehouden de tijd vanaf de dag waarop

- een zelfstandig vertrek onder toezicht
- een zelfstandig vertrek of zonder toezicht of
- een aantoonbaar vertrek

is geregistreerd, tot aan het einde van het jaar.

In het gebruikte analysebestand is van 2462 vreemdelingen de betreffende informatie bekend over vertrek uit het land. Van 1868 vreemdelingen in het bestand is dergelijke informatie niet bekend.

Detentiegegevens - DJI

Het DJI-bestand bevat de volgende variabelen:

- vreemdelingennummer,
- datumbegindetentie,
- insluittitel (strafrechtelijk of vreemdelingendetentie),
- een datum instroom en
-]een datum uitstroom.

De datumbegindetentie is een formele startdatum voor de detentie, maar hoeft niet gelijk te zijn aan het daadwerkelijk begin van de insluiting. Die wordt aangegeven door de datum instroom. De datum uitstroom geeft uiteraard het einde van de detentie weer.

De informatie over één detentie van een vreemdeling is niet weergegeven in één record, maar is in principe weergegeven in twee records, waarbij een hulpvariabele aangeeft of het record betrekking heeft op de instroom(datum) of de uitstroom(datum) van de detentie. Het gevolg is dat de sortering van het bestand van invloed is op de vraag of de juiste records betreffende één detentie bij elkaar worden geplaatst. Voor het bij elkaar plaatsen van records betreffende de detenties is geen gebruik gemaakt van de informatie over de insluittitel. De reden is dat deze variabele als enige missende waarden bevat die sortering incompleet maken. Bovendien is de insluittitel voor het bepalen van de

detentietijd van geen belang. Om een detentie te identificeren is dus alleen gekeken naar de variabelen 'datumbegindetentie', 'datuminstroomuitstroom' en de hulpvariabele die aangeeft of een record betrekking heeft op een instroom of uitstroom in of uit detentie.

We zijn gestart met detentiegegevens over het jaar 2009 van 2375 van de 4330 illegale vreemdelingen in PSHV en VBS geregistreerd als zijnde staandegehouden door politie of KMar in 2009. Voor 1955 van deze vreemdelingen beschikken we niet over detentiegegevens.

De gegevens zijn gebruikt om de periode in aantal dagen in detentie en in delen van het jaar te berekenen voor de geregistreerde vreemdelingen. De gegevens zijn niet in alle opzichten compleet en geschikt om deze perioden te berekenen. De volgende problemen zijn geconstateerd en bijbehorende oplossingen zijn gekozen.

Over een deel van de detenties zijn de gegevens niet compleet (n=237). Deze gegevens zijn op de volgende manieren aangevuld:

- Voor 148 detenties is een instroomdatum de laatst bekende datum en is geen datum uitstroom geregistreerd. We nemen aan dat deze personen tot het einde van het jaar, 31-12-2009, in detentie zijn gehouden.
- Voor 75 detenties is geen datum uitstroom geregistreerd, maar volgt op een later moment in het jaar nog wel een nieuwe detentie. We nemen aan dat deze 75 detenties eindigen op de dag dat de volgende detentie in gaat.
- Voor 14 detenties is alleen een datum uitstroom bekend en ontbreekt een datum instroom. In dit geval nemen we aan dat de datumbegindetentie de start is van de betreffende detentie.

Voorts zijn er bij 11 detenties meerdere datums voor één en dezelfde gebeurtenis geregistreerd. In 7 gevallen betreft het een (dubbele) registratie van een instroom en ontbreekt een uitstroomdatum. In 4 gevallen is een datum voor de uitstroom beschikbaar, maar ontbreekt de instroomdatum. Bij een dubbele registratie van een instroom is de eerste datum die is geregistreerd als de juiste gekozen en bij een dubbele registratie van een uitstroom de laatste. Op die manier wordt verondersteld dat de detentie eerder langer duurde dan korter. Bij 1 detentie is sprake van een doublure: een zelfde gebeurtenis (een instroom in detentie) is twee keer geregistreerd met verschillende datums (1 dag verschil). Hier is de 2^{e} datum uit het bestand verwijderd. Ten behoeve van de schattingen is voor 2377 vreemdelingen een detentieperiode berekend als deel van het jaar.

Geldigheid registratie PSHV KMar

Illegale vreemdelingen kunnen worden staande gehouden door de vreemdelingenpolitie (staandehouding) of worden overgenomen van de basispolitiezorg (overname basispolitiezorg). Illegale vreemdelingen staandegehouden en geregistreerd door de KMar kennen dat onderscheid niet.

Het komt voor dat er weinig dagen tussen 2 opeenvolgende registraties van een zelfde vreemdeling liggen. Dat roept vragen op over de geldigheid van die registraties. Is er daadwerkelijk sprake van een op vrije voeten stellen en vervolgens weer, onafhankelijk van een eerdere keer, aanhouden van eenzelfde illegale vreemdeling, of is er sprake van een administratief artefact? In dat laatste geval wordt er bijvoorbeeld een nieuwe registratie gemaakt van nieuwe aanvullende feiten betreffende een eerdere aanhouding, voor correctie van een fout, of van een nieuwe stap in de afhandeling van de zaak van de eerder staandegehouden vreemdeling.

Bij 8% van de opeenvolgende registraties zijn er 0 tot maximaal 3 dagen verstreken voor dat de nieuwe registratie wordt gemaakt. Het lijkt duidelijk dat herhaalde observaties op eenzelfde dag niet kunnen worden meegeteld als observaties die onafhankelijk van elkaar worden gedaan. Na 3 dagen neemt het aantal dagen tussen opeenvolgende registraties snel toe. Gezien het gewicht van herhaalde staandehoudingen in vangs-hervangstschattingen, is het zaak aandacht te schenken aan de geldigheid van de registraties.

In het navolgende kijken we naar een aantal kenmerken van opeenvolgende registraties die mogelijk aanwijzingen kunnen bevatten voor onjuistheden of onvolledigheden van de eerste registratie die daarom correctie of aanvulling nodig maakten, waarvoor een nieuwe registratie werd gemaakt. In dat geval zou er geen sprake zijn van een nieuwe onafhankelijke observatie van de betreffende illegale vreemdeling. Het gaat om de volgende omstandigheden:

- Registratie in het weekeinde
- Registratie op bepaalde dagen van de week
- Reden staandehouding
- Overdracht van politie aan KMar of vice-versa
- Overdracht tussen regio's
- Verandering van procestype (overdracht van basispolitiezorg, staandehouding door vreemdelingenpolitie)

Weekeinde

Als eerste registraties van opeenvolgende registraties in het weekeinde plaats vinden, verlopen er minder dagen tot de volgende registratie(s), dan wanneer de eerste registratie door de week plaats vindt (zie tabel 1). Dit kan een aanwijzing zijn dat de registraties in het weekeinde minder compleet en juist zijn en worden gecorrigeerd door nieuwe registraties.

Tabel 1: aantal dagen tussen opeenvolgende registraties in verschillende omstandigheden

Variabele	*Gemiddeld aantal dagen tussen opeenvolgende registraties*		*Sig.*
Weekeinde of door de week:	Door de week (n = 1020)	Weekeinde (n = 180)	
	187.04	173.40	
Overdracht KMar – politie of vice-versa	Geen overdracht (n = 1138)	Overdracht (n = 70)	
	180.86	250.91	**
Overdracht tussen politieregio's	Geen verandering (n = 672)	Regio-overdracht (n = 536)	
	167.72	206.48	**
Verandering van proces	Nee (n = 794) 179.46	Ja (n = 243) 196.62	

Alle significanties 2-zijdig: ** = 5%; * = 10%

Overdracht Politie – KMar

Een overdracht van een vreemdeling van de KMar aan de politie gaat eerder gepaard met een groter aantal dagen tussen opeenvolgende observaties dan met minder, zoals te verwachten zou zijn indien (ten onrechte) bij de overdracht een nieuwe registratie zou worden aangemaakt ($t(1206) = -3.411$, $p = .001$).

Overdracht tussen politieregio's

Een vergelijkbare conclusie geldt indien er een overdracht tussen 2 politieregio's plaats vindt. Gemiddeld liggen er in dat geval meer dagen tussen opeenvolgende registraties dan wanneer er geen regio-overdracht plaats vindt tussen opeenvolgende registraties ($t(1206) = - 4.019$, $p < .001$).

Verandering van proces

Een verandering van proces (PSHV: staandehouding of overname van basispolitiezorg) gaat niet gepaard met een significant kleiner of groter aantal dagen tussen 2 opeenvolgende registraties ($t(1035) = -1.375$, $p = .169$).

Dagen van de week

Ongeacht op welke dag van de week de 1e registratie plaats vindt, het aantal dagen tot de volgende registratie is niet (significant) groter of kleiner ($F(6) = 1{,}140$, $p = .337$) (zie tabel 2).

Tabel 2: aantal dagen tussen registratie voor verschillende weekdagen van 1e registratie

Dag 1e registratie	Gemiddeld aantal dagen tussen registraties	N	Std. Deviation
Zondag	161,05	106	166,222
Maandag	175,92	169	168,585
Dinsdag	191,96	234	171,681
Woensdag	199,46	194	174,141
Donderdag	192,79	225	163,988
Vrijdag	186,25	166	161,855
Zaterdag	163,80	114	161,334
Totaal	184,92	1208	167,529

Reden staandehouden

Er zijn geen verschillen in het aantal dagen waarin een nieuwe registratie volgt voor de verschillende redenen voor staandehouding ($F(3) = .442$, $p = .723$) (tabel 3).

Tabel 3. Aantal dagen volgend op verschillende redenen staandehouding

	Gemiddeld aantal dagen tussen registraties	N	Std. Deviation
identiteit kon worden vastgesteld en bleek dat betrokkene geen rechtmatig verblijf had	191,16	178	156,366
identiteit niet onmiddellijk kon worden vastgesteld	175,89	38	143,496
identiteit onmiddellijk kon worden vastgesteld en niet onmiddellijk bleek dat betrokkene rechtmatig verblijf had	299,50	2	265,165
Onbekend	183,91	990	170,286
Total	184,92	1208	167,529

Conclusies

1. Het aantal dagen tussen opeenvolgende registraties is soms zo gering dat het niet aannemelijk is dat er opeenvolgende van elkaar onafhankelijke observaties van de betreffende illegale vreemdeling hebben plaats gevonden;
2. Het is moeilijk om een onderscheid te maken tussen geldige herhaalde observaties en onterechte nieuwe registraties. Opvallend is dat vanaf 4 dagen het aantal dagen tussen opeenvolgende registraties sneller toeneemt; 8% van meervoudige registraties in de jaren 2008 en 2009 vindt binnen 4 dagen na de voorgaande registratie plaats;
3. Er is een aantal kenmerken van de opeenvolgende registraties onderzocht, met als resultaat:
    a. Een 1e registratie in een weekeinde wordt sneller gevolgd door een nieuwe registratie. Dit kan worden opgevat als een aanwijzing dat een registratie van een staandehouding vervolgens in een nieuwe registratie wordt gecorrigeerd of aangevuld;
    b. Bij een overdracht van een illegale vreemdeling van de KMar aan de politie , of tussen verschillende politieregio's, is er gemiddeld genomen juist sprake van meer dagen tussen de registraties;
    c. Er zijn geen significante verschillen in het aantal dagen tussen opeenvolgende registraties bij registratie op verschillende weekdagen, bij verschillende geregistreerde redenen van staandehouding of bij het starten van een nieuw procestype (staandehouding of overname van basispolitiezorg)
4. De kenmerken van opeenvolgende registraties bieden weinig aanknopingspunten voor normen voor beoordeling van de geldigheid van registraties als onafhankelijke observatie.

Mogelijk dat een onderzoek van actuele registraties met een beperkt aantal dagen ertussen een beter zicht geeft op het registratieproces, de keuzes die daarbij worden gemaakt en de gevolgen voor de geldigheid van registraties.