

Benutting HKS en ontwikkeling vangst-hervangstmethoden

**Een onderzoek in opdracht van
het Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC)
van het Ministerie van Justitie**

Drs. M.J.L.F. Cruyff.

Prof. Dr. P.G.M. van der Heijden.

Voorwoord

Binnen de Rijksoverheid is een onderzoeksprogramma gestart om op basis van bestaande politiegegevens de omvang van delinquente populaties te schatten. Het gaat hierbij vooral om de zogenaamde ‘dark number’ populaties, waarbij het ‘dark number’ refereert aan de personen die wel in overtreding zijn (geweest) maar daarvoor niet door de politie zijn aangehouden. Het Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC) van het Ministerie van Justitie richt zich op het verder ontwikkelen van het instrumentarium voor dit type onderzoek.

Dit type onderzoek is gebaseerd op de zogenaamde Vangst-Hervangstmethode. In deze methode wordt een schatting van het onbekende aantal delinquenten verkregen uit politiegegevens over de personen die wel voor het betreffende delict zijn aangehouden, hierbij gaat het vooral om het aantal aanhoudingen per persoon. Het registratiesysteem van de politie – het Dienst Herkenningssysteem (HKS) – omvat gegevens die gemaakt kunnen worden voor de Vangst-Hervangstmethode. Dit rapport beoogt enerzijds aan te geven in welk stadium de ontwikkeling van het onderzoeksinstrumentarium zich bevindt en anderzijds in hoeverre de gegevens uit het HKS bruikbaar zijn om tot verantwoorde schattingen van de omvang van delinquente populaties te komen.

Hiertoe zijn op basis van gegevens uit het HKS voor twee typen delicten – het rijden onder invloed en het voorhanden hebben van wapens – de omvang van de populatie van delictplegers geschat. Daarnaast zijn artikelen in dit rapport opgenomen waarin de Vangst-Hervangstmethodiek is toegepast op andere (delinquente) populaties, en waarin de meest recente ontwikkeling van de theorie is besproken.

Januari 2003.

Drs. M. Cruyff
Prof. Dr. P. van der Heijden
Onderzoekschool IOPS, afdeling Utrecht

Correspondentie naar:
Prof. Dr. P. van der Heijden
Faculteit Sociale Wetenschappen
Universiteit Utrecht
Postbus 80.140
3508 TC Utrecht
tel. 030-2534688, fax 030-2535797
e-mail: p.vanderheijden@fss.uu.nl

Inhoudsopgave

Voorwoord	2
Inleiding	4
1. Het HKS	6
<i>Definitie afhankelijke variabele</i>	6
<i>Definitie covariaten</i>	7
<i>Dubbele registraties</i>	7
<i>Koppeling jaarbestanden</i>	7
<i>Het metabestand</i>	7
2. Ontwikkeling vangst-hervangstmethodiek	9
<i>Het Horvitz-Thompson betrouwbaarheidsinterval</i>	9
<i>Het afgeknotte negatief binomiaalmodel</i>	9
<i>Overige modellen</i>	10
3. Toepassingen en resultaten	11
<i>De regressiegewichten</i>	12
<i>Geobserveerde en geschatte frequenties</i>	13
<i>Populatieschattingen en betrouwbaarheidsintervallen</i>	13
4. Conclusies en discussie	15
<i>Realisatie doelstellingen</i>	15
<i>Poissonproces en populatie</i>	15
<i>Aantal covariaten en populatieschatting</i>	16
<i>Website</i>	17
Bijlage 1A: Afleiding betrouwbaarheidsinterval	18
Bijlage 1B: Toetsing betrouwbaarheidsinterval	39
Bijlage 1C: Toepassing afgeknot negatief binomiaalmodel	44
Bijlage 1D: Simulatiestudie afgeknot negatief binomiaalmodel	51
Bijlage 1E: Voorbeeldanalyses HKS	54
Bijlage 1F: Schatting populatie illegalen	81
Bijlage 2: Handreiking & SQL-programma aanmaak vangst-hervangstdata	95
Bijlage 3: Tabellen ‘rijden onder invloed’.	109
Bijlage 4: Tabellen ‘wapens voorhanden’	121

Inleiding

Begin 1999 is het project “Benutting HKS en ontwikkeling vangst-hervangstmethoden” van start gegaan. Dit project, dat in opdracht van het Wetenschappelijk Onderzoek- en Documentatiecentrum van het Ministerie van Justitie is uitgevoerd, had tot doel: (1) het schatten van ‘dark numbers’ van delinquente populaties op basis van bestaande gegevens uit het HKS en; (2) het verder te ontwikkelen van de vangst-hervangstmethodiek. Het project is, onder supervisie van prof. dr. P.G.M. van der Heijden, uitgevoerd door dr. R. Bustami en, vanaf begin 2002, door drs. M.J.L.F. Cruyff. Inhoudelijk advies is ingewonnen bij prof. dr. H. Van Houwelingen. De begeleidingscommissie van het project werd gevormd door prof. dr. W.R. van Zwet, drs. A.H. van den Berg, mw. drs. A. Slotboom, dr. H. Elffers, dr. M.P. Tummers, dr. F.W. Beijaard, mw. prof. dr. C.C.J.H. Bijleveld en mw. drs. M. Wong. In dit eindrapport zijn de resultaten van dit project gepresenteerd.

In dit rapport wordt een overzicht gegeven van de belangrijkste drie componenten van het project:

- i. bewerking van gegevens uit het HKS;
- ii. ontwikkeling van de vangst-hervangstmethodiek;
- iii. toepassingen van de vangst-hervangstmethodiek op het HKS.

Het HKS is oorspronkelijk ontwikkeld voor het opsporen van daders waardoor de gegevens niet direct bruikbaar zijn voor de vangst-hervangstmethodiek. Aan de analyses is een uitgebreid proces van dataselectie en bewerking voorafgegaan. Dit proces en de problemen die bij de uitvoering daarvan zijn ondervonden worden besproken in hoofdstuk 1.

In het kader van het verder ontwikkelen van de vangst-hervangstmethodiek is gewerkt aan de afleiding van een betrouwbaarheidsinterval voor de populatieschatting en aan alternatieve modellen voor afgeknotte Poisson regressiemodel, met name het afgeknotte negatief binomiaal regressiemodel. De stand van zaken op het gebied van de theoretische ontwikkelingen komt aan de orde in hoofdstuk 2.

De toepassing van de vangst-hervangstmethodiek op twee delicten uit het HKS, rijden onder invloed en voorhanden hebben van wapens, wordt besproken in hoofdstuk 3, voor een uitgebreid overzicht van de analyseresultaten wordt verwezen naar de bijlagen 3 en 4. Een artikel dat is geschreven naar aanleiding van deze analyses is opgenomen in bijlage 1E. De vangst-hervangstmethode is eveneens toegepast voor het schatten van het aantal illegalen in Nederland, de resultaten van dit onderzoek worden in bijlage 1F besproken.

In hoofdstuk 4 wordt ingegaan op de vraag in hoeverre de doelstellingen van het project “Benutting HKS en ontwikkeling vangst-hervangstmethoden” zijn gerealiseerd en welke mogelijkheden de resultaten van dit project voor de toekomst bieden. Daarnaast wordt stilgestaan bij twee kwesties die door de begeleidingscommissie naar voren zijn gebracht. De eerste kwestie betreft de validiteit van de aanname dat de aanhoudingen voor de HKS-delicten door een Poissonproces zijn gegenereerd, met name voor delicten waarvoor de kans op aanhouding niet constant is in de tijd. De tweede kwestie betreft het fenomeen dat de populatieschattingen stijgen naarmate er meer covariaten in het model worden opgenomen.

Onze dank gaat uit naar het WODC, de KLPD en het NRI, en met name naar dhr. L. Prins, voor de verleende medewerking bij prepareren van de data. Daarnaast willen we de leden van de begeleidingscommissie bedanken voor hun kritische vragen en opmerkingen, welke hebben bijgedragen tot een dieper inzicht in de problematiek.

1. Het HKS

Het HKS een landelijk politieregistratiesysteem waarin sinds de jaren '80 alle delicten met bekende dader worden geregistreerd. Het HKS maakt het voor de politie mogelijk overzichten van gepleegde delicten en daders op te vragen en heeft als primair doel de opsporing van daders te vergemakkelijken. Het HKS bestaat in feite uit verschillende bestanden waarin onder andere persoonsgegevens en delictgegevens per jaar en per politieregio apart staan geregistreerd.¹ Door deze bestanden aan elkaar te koppelen ontstaat een landelijk overzicht van daders en gepleegde delicten.

Het aanmaken van vangst-hervangstdata uit de gegevensbestanden van het HKS is een zeer complexe taak gebleken. Ondanks dat hierbij de grootste zorgvuldigheid is betracht, is het in de loop van het project toch meermalen voorgekomen dat er fouten in de aangemaakte data werden ontdekt, vaak pas nadat een groot aantal analyses op deze data waren verricht. Hierdoor is erg veel tijd verloren gegaan. Dit hoofdstuk is vooral bedoeld om de valkuilen bloot te leggen die bij de aanmaak van vangst-hervangstdata uit het HKS aanwezig zijn. De valkuilen zijn aanwezig bij het definiëren van de afhankelijke variabelen en van de covariaten en bij de aanmaak van variabelen gemeten over langere perioden dan één jaar. Daarnaast bevatten de HKS-bestanden zelf een aantal fouten, zoals dubbele registraties en ontbrekende gegevens, die van invloed kunnen zijn op de hoogte van de populatieschattingen.

Definitie afhankelijke variabele

Eén oorzaak voor problemen bij de aanmaak van data is dat de registraties in het HKS iets wezenlijk anders meten dan wat de afhankelijke variabele in de vangst-hervangstmethode behoort te meten. De afhankelijke variabele in de vangst-hervangstmethodiek meet het aantal keer dat een individu is aangehouden voor een bepaald delict. Nu worden in het HKS de gepleegde delicten geregistreerd als 'feiten'. Onder een feit wordt elk afzonderlijk gepleegd delict verstaan. De registratie van een feit kan op verschillende manieren tot stand komen, bijvoorbeeld als gevolg van een betrapping op heterdaad of een bekentenis. Feiten maken weer deel uit van een 'antecedent'. Een antecedent refereert aan het proces verbaal waarin alle feiten worden geregistreerd die, volgend op de aanhouding en ondervraging van een dader, aan het licht zijn gekomen. Binnen één antecedent kunnen dus meerdere feiten zijn opgenomen die naar hetzelfde type delict verwijzen, het dragen van een mes en een revolver zijn bijvoorbeeld twee feiten van het illegaal voorhanden hebben van wapens. Bij het aanmaken van de afhankelijke variabelen is als gevolg van deze wijze van registreren de fout gemaakt het aantal feiten te tellen. Dit levert een verkeerde telling op omdat het in de vangst-hervangstmethodiek gaat om het aantal 'vangsten', in

¹ In de handreiking in bijlage 2 is nader ingegaan op de structuur van het HKS. De handreiking is speciaal geschreven voor onderzoekers die zelf vangst-hervangsanalyses op basis van het HKS willen doen maar is ook als aanvulling op dit hoofdstuk te lezen.

dit geval het aantal aanhoudingen voor een specifiek delict. Een telling van het aantal feiten correspondeert daarentegen met het aantal maal dat een specifiek delict is gepleegd, en niet met het aantal aanhoudingen voor dat delict. Dit probleem is opgelost door voor elk antecedent een bepaald type delict maar één keer te tellen, ongeacht het aantal geregistreerde feiten voor dat delict. Wanneer een dader dus is gepakt voor het dragen van een mes en een revolver wordt dat in het kader van de vangst-hervangstmethode geteld als één enkele vangst, terwijl het in het HKS twee feiten zijn.

Definitie covariaten

Een andere valkuil zit verscholen bij de constructie van de covariaten die de criminele geschiedenis van de dader meten. In het HKS worden de minder zware delicten uit het bestand verwijderd indien de persoon 5 jaar lang niet in contact is geweest met de politie. Wanneer nu de criminele geschiedenis wordt gemeten over alle delicten die een persoon heeft gepleegd, ontstaat een niet valide beeld, omdat die geschiedenis deels verwijderd kan zijn. De eerste databestanden zijn op deze wijze aangemaakt en geanalyseerd en moesten later worden geheranalyseerd met covariaten die alleen de delicten tellen die de afgelopen 5 jaar zijn gepleegd.

Dubbele registraties

De mogelijkheid bestaat dat de jaarbestanden van het HKS dubbele registraties bevatten. Deze kunnen bijvoorbeeld ontstaan wanneer een dader in een andere politieregio woont dan waar het delict is gepleegd. De dader wordt dan in eerste instantie geregistreerd in de regio waar het delict is gepleegd. Vervolgens wordt het antecedent doorgegeven aan en geregistreerd in de politieregio waar de dader woont. In geval van verhuizing geldt een vergelijkbare procedure, de politie in de nieuwe woonregio van de dader wordt door de politie in de oude woonregio van de dader op de hoogte gebracht van diens antecedenten.

Koppeling jaarbestanden

Een ander probleem van de jaarbestanden is dat de persoonsidentificatienummers per jaar opnieuw worden toegewezen. Hierdoor heeft dezelfde persoon elk jaar een ander nummer en is het niet mogelijk de delicten van één persoon over de jaren heen te koppelen. In het begin van het project is dit niet duidelijk geweest en zijn personen over de jaren heen gekoppeld aan de hand van het persoonsidentificatienummer, in de veronderstelling dat het bij dezelfde nummers ook om dezelfde persoon zou gaan.

Het metabestand

Het metabestand dat sinds ongeveer een jaar in omloop is voorziet in een oplossing voor beide laatstgenoemde problemen. In het metabestand zijn aan de hand van GBA-nummers de dubbele registraties zoveel mogelijk verwijderd en personen over de jaren heen gekoppeld. Een nadeel van het metabestand is dat in sommige gevallen antecedenten verloren zijn gegaan. In het geval dat een delict in een andere regio dan de woonregio is gepleegd of bij verhuizing van de dader is de informatieoverdracht van antecedenten tussen de betrokken politieregio's niet altijd volledig geweest. Hierdoor is een situatie ontstaan dat personen in de ene regio met andere antecedenten geregistreerd staat dan in de andere regio. Bij het verwijderen van de dubbele registraties is de keuze gemaakt de persoon in die regio te registreren waar deze de meeste antecedenten heeft. Hierdoor kan

het gebeuren dat een persoon bij het ontdebelen antecedenten verliest. Dit probleem kwam aan het licht bij het aanmaken van de vangst-hervangstdata. Nader onderzoek heeft uitgewezen dat bij ongeveer 1% van de personen in het metabestand een deel van de antecedenten verloren is gegaan.

Vanaf het moment dat duidelijk werd dat met de jaarbestanden geen perioden langer dan één jaar konden worden bestudeerd, is met het aanmaken van data gewacht op het beschikbaar komen van het metabestand. Dit heeft enkele maanden geduurd omdat voor het beschikbaar stellen van gegevens uit het HKS aan derden de toestemming benodigd was van alle politieregio's. De aanmaak van de vangst-hervangstdata uit het metabestand is, ondanks het feit dat inmiddels alle valkuilen in het HKS bekend waren, een moeizaam proces geweest dat enkele dagen in beslag heeft genomen. In totaal zijn er over de jaren 1996-2001 per delict datasets aangemaakt over 20 verschillende observatieperioden, zes over een periode van één jaar, vijf over een periode van twee jaar, vier over een periode van drie jaar, drie over een periode van vier jaar en twee over een periode van vijf jaar. In totaal leverde dit voor beide delicten 40 databestanden op. De databewerkingen zijn uitgevoerd met een door SQL-programma (de syntax van dit programma is opgenomen in bijlage 2), geschreven door dhr. L. Prins van het Korps Landelijke Politiediensten (KLPD). Het grote voordeel van dit programma is dat nu met enkele relatief kleine aanpassingen op eenvoudige wijze ook voor andere delicten data kan worden aangemaakt die direct geschikt zijn voor vangst-hervangstanalyse.

2. Ontwikkeling vangst-hervangstmethodiek

Bij aanvang van dit project was één model voor het schatten van populaties beschikbaar, het afgeknotte Poisson regressiemodel, en op basis van dit model kon een puntschatting van de populatie worden verkregen. Twee belangrijke doelstellingen van dit project ter uitbreiding van de vangst-hervangstmethodiek waren: (1) het afleiden van een betrouwbaarheidsinterval van de puntschatting en; (2) het uitbreiden van het afgeknotte Poisson regressiemodel met het afgeknotte negatief binomiaal regressiemodel en met modellen voor onderdispersie, negatieve en positieve besmetting en open populaties.

Het Horvitz-Thompson betrouwbaarheidsinterval

De eerste doelstelling is gerealiseerd met de afleiding van een Horvitz-Thompson betrouwbaarheidsinterval voor het afgeknotte Poisson regressiemodel en voor het afgeknotte negatief binomiaal regressiemodel (Bustami *et al.*, 2000, zie bijlage 1A). Deze betrouwbaarheidsintervallen zijn vervolgens getoetst met behulp van een simulatiestudie (Bustami *et al.*, 2001, zie bijlage 1B). Uit deze studie blijkt dat de Horvitz-Thompson betrouwbaarheidsintervallen bevredigende resultaten te zien geven in vergelijking met uit bootstraps verkregen betrouwbaarheidsintervallen. Het Horvitz-Thompson betrouwbaarheidsinterval is inmiddels toegepast in een studie naar het aantal illegalen in Nederland (Engbersen *et al.*, 2002, zie bijlage 1F) en in een studie naar het aantal verkeersdeelnemers onder invloed van alcohol en het illegale wapenbezitters (van der Heijden *et al.*, 2002, zie bijlage 1E).

Het afgeknotte negatief binomiaalmodel

De tweede doelstelling is ten dele gerealiseerd. Een procedure is ontwikkeld voor het schatten van de populatieomvang op basis van het afgeknotte negatief binomiaal regressiemodel. Dit model heeft ten opzichte van het afgeknotte Poisson regressiemodel één extra parameter, de dispersieparameter α . Het afgeknotte negatief binomiaal regressiemodel is van toepassing wanneer de assumptie van homogeniteit van de Poissonparameters is geschonden. Het Poissonmodel veronderstelt dat ieder individu een gelijke pakkans, en dus ook een gelijke Poissonparameter, heeft. In het Poisson regressiemodel wordt middels covariaten heterogeniteit in de Poissonparameters geïntroduceerd, waardoor alleen nog voor personen met identieke scores op de covariaten de assumptie van homogene Poissonparameters geldt. Voor personen met verschillende scores op de covariaten mogen de Poissonparameters heterogeen zijn, men spreekt in dat geval van geobserveerde heterogeniteit. Er is sprake van ongeobserveerde heterogeniteit wanneer niet alle relevante covariaten in het regressiemodel zijn opgenomen en de Poissonparameters ook verschillen voor personen met identieke scores op de covariaten. De Poissonverdeling wordt gekenmerkt door gelijkheid van gemiddelde en variantie, terwijl de aanwezigheid van ongeobserveerde heterogeniteit tot overdispersie leidt, de variantie is dan groter dan het

gemiddelde. Het afgeknotte negatief binomiaal regressiemodel is dankzij de dispersieparameter in staat overdispersie te modelleren.

Het afgeknotte negatief biniaal regressiemodel blijkt soms moeilijk te fitten, de iteratieve schattingsprocedure leidt niet altijd tot convergentie. Uit een simulatiestudie (Bustami et al., 2001, zie bijlage 1D) blijkt de kans op succesvolle convergentie toe te nemen naarmate de steekproef, de Poissonparameters en de dispersieparameter groter is/zijn. In de praktijk is het model toegepast voor het schatten van de populatie druggebruikers (Bustami, zie bijlage 1C). Mede uit deze studies is gebleken dat het afgeknotte negatief binomiaal regressiemodel tot hogere populatieschattingen leidt dan schattingen gebaseerd op het afgeknotte Poisson regressiemodel. Daarnaast zijn een aantal datasets uit het onderhavige project geanalyseerd met het afgeknotte negatief binomiaal regressiemodel. Alhoewel uit tests blijkt dat er wel sprake is van significante overdispersie (zie tabellen 11 van de bijlagen 3 en 4), heeft de schattingsprocedure voor geen van de gevallen tot convergentie geleid. De meest waarschijnlijke reden hiervoor is dat de pakkans voor de in dit project bestudeerde delicten gemiddeld relatief laag is. Een simulatiestudie heeft uitgewezen dat de iteratieve schattingsprocedure van het afgeknotte negatief binomiaal regressiemodel bij kleine waarden van de Poissonparameters vaak niet convergeert (zie bijlage 1D).

Overige modellen

De invloed van schendingen van de assumptie van een gesloten populatie is onderzocht door observatieperioden van verschillende duur te analyseren. Een gesloten populatie wordt gekenmerkt door het ontbreken van doorstroming in de populatie, gedurende de observatieperiode komen er geen individuen bij en vallen er geen individuen uit. Het gevolg van doorstroming is dat voor nieuwkomers en uitvallers de kans gepakt te worden gedurende de periode van afwezigheid uit de populatie nul is. Deze personen worden daardoor minder vaak gepakt dan wanneer zij gedurende de volledige observatieperiode in de populatie aanwezig zouden zijn geweest. Hierdoor is in de steekproef het aantal enkelvoudige pakkingen te hoog en het aantal meervoudige pakkingen te laag, hetgeen leidt tot te hoge populatieschattingen. Door de observatieperioden zo kort mogelijk te houden wordt de invloed van doorstroming op de populatieschatting geminimaliseerd.

Aan de ontwikkeling van modellen voor onderdispersie en positieve en negatieve besmetting is in dit project geen aandacht besteed. Eén reden hiervoor is tijdgebrek geweest, door de problemen bij het aanmaken van goede vangst-hervangstdata is veel tijd verloren gegaan. Een tweede reden is dat de data niet genoeg informatie bevatten om gefundeerde uitspraken te doen over de aanwezigheid van eventuele onderdispersie of besmetting.

Voor de toekomst lijken interessante ontwikkelingen op het gebied van de detectie en analyse van uitschieters en invloedrijke datapunten te liggen. Voor de berekening deze diagnostische statistieken is inmiddels software ontwikkeld maar de betekenis ervan is vooralsnog onduidelijk. Deze onduidelijkheid vloeit voort uit het feit dat de statistieken zijn verkregen met een afgeknot, niet-lineair model en met de interpretatie van dergelijke statistieken nog nauwelijks ervaring is opgedaan. Simulatiestudies zouden wat betreft de betekenis van deze statistieken meer duidelijkheid kunnen scheppen.

3. Toepassingen en resultaten

In het kader van dit project zijn over de periode 1996-2001 de landelijke populaties van het aantal plegers van twee slachtofferloze delicten geschat: het aantal verkeersdeelnemers onder invloed van alcohol en het aantal personen met wapens voorhanden. Per delict zijn in totaal verschillende 20 observatieperioden geanalyseerd, de duur van de perioden varieert van één tot vijf jaar (6 perioden van 1 jaar, 5 van 2 jaar, 4 van 3 jaar, 3 van 4 jaar en 2 van 5 jaar). Per delict is uit de landelijke bestanden vervolgens nog een selectie gemaakt van vijf politieregio's die bekend staan om hun zorgvuldige administratie van het HKS, te weten: Gelderland Midden, Hollands Midden, Zuid-Holland-Zuid, Rotterdam Rijnmond, en Midden-West Brabant. In totaal zijn 80 bestanden geanalyseerd, 40 per delict, waarvan 20 landelijke en 20 regionale (de vijf geselecteerde regio's).

Van de analyses rapporteren we de regressiegewichten van de covariaten, de marginale frequenties en residuen, en de populatieschattingen met betrouwbaarheidsintervallen. De resultaten voor rijden onder invloed en het voorhanden hebben van wapens zijn opgenomen in respectievelijk bijlage 3 en bijlage 4. In beide bijlagen geven de tabellen 1 t/m 6 een overzicht van de regressiegewichten van de covariaten. In de tabellen zijn ter indicatie van de significantie van de regressiegewichten eveneens de t-waarden opgenomen. De tabellen 7 t/m 9 geven een overzicht van de geobserveerde en geschatte frequentieverdelingen en de marginale residuen. Tabel 10 geeft een overzicht van de populatieschattingen en betrouwbaarheidsintervallen. In tabel 11 zijn de resultaten van toetsen op overdispersie getoond. In figuur 12 zijn de gemiddelde populatieschattingen voor verschillende observatieperioden opgenomen.

Voor de volledigheid zijn de resultaten van alle 80 geanalyseerde databestanden gerapporteerd. Hierbij dient echter te worden opgemerkt dat de kwaliteit van de data niet voor alle bestanden gelijk is. In de eerste plaats zijn, vanwege het feit dat de administratie van het HKS door de bovengenoemde selectie van vijf politieregio's waarschijnlijk zorgvuldiger is geweest dan die van de overige politieregio's, de analyseresultaten voor de op regio geselecteerde bestanden naar verwachting meer valide dan die van de landelijke bestanden. In de tweede plaats zijn de gevolgen van schendingen van de assumptie van een gesloten populatie groter voor de bestanden die door een langere observatieperioden worden gekenmerkt. Met name de geobserveerde frequenties in de tabellen 7a t/m 9a geven hierdoor voor de langere observatieperioden een vertekend beeld met respectievelijk te hoge frequenties voor de enkelvoudige aanhoudingen en te lage frequenties voor de meervoudige aanhoudingen. Dat patroon leidt tot te hoge populatieschattingen in tabel 10. Tenslotte dient nog te worden opgemerkt dat voor een stabiele schatting van de populatie een minimale spreiding in en vulling van het aantal aanhoudingen vereist is. Deze ontbreken in enkele gevallen in de bestanden van het voorhanden hebben van wapens, met name in de bestanden van de selectie van de vijf politieregio's met een observatieperiode van één

jaar. Een extreem voorbeeld hiervan is het jaar 2000, waarin geen personen zijn geobserveerd die meer dan twee keer zijn aangehouden (zie tabel 7a van bijlage 4). Gevolg van een kleine spreiding in het aantal aanhoudingen en lage frequenties van de meervoudige aanhoudingen is een instabiele populatieschatting en een groot betrouwbaarheidsinterval. In een aantal gevallen is het ook onmogelijk gebleken de regressiegewichten van bepaalde covariaten te schatten (de blanco waarden in tabellen 1 t/m 6 van bijlage 4).

In het licht van deze kwaliteitsverschillen tussen de databestanden verdient het aanbeveling voor rijden onder invloed alleen van de analyseresultaten van de observatieperioden van één jaar uit te gaan. Voor het voorhanden hebben van wapens verdienen de resultaten van de observatieperioden van twee jaar de voorkeur. Tenslotte worden voor beide delicten de resultaten voor de bestanden van de selectie van vijf regio's meer valide geacht dan voor de landelijke bestanden. Voor een uitgebreide bespreking van twee voorbeelden hiervan, het bestand van rijden onder invloed over het 2000 en het voorhanden hebben van wapens over de jaren 1998-1999, wordt verwezen naar het artikel in bijlage 1F (van der Heijden *et al.*, 2002). De resultaten van de overige bestanden zijn mede in de rapportage opgenomen om inzicht te geven in het gedrag van het afgeknotte Poisson regressiemodel bij onvoldoende data en schendingen van de assumptie van een gesloten populatie.

De afhankelijke variabelen in de datasets zijn tellingen van het aantal aanhoudingen voor respectievelijk rijden onder invloed en het voorhanden hebben van wapens binnen de periode van observatie. De covariaten zijn geslacht, leeftijd, leeftijd van eerste delict en het aantal aanhoudingen voor een zestal delictgroepen. De delictgroepen hebben betrekking op het aantal aanhoudingen voor vermogens-, gewelds-, opium-, verkeers- en economische delicten en voor vernielingen en zijn gemeten over een periode van vijf jaar voorafgaand aan het laatste jaar waarin de afhankelijke variabele is gemeten. Voor zowel de afhankelijke als onafhankelijke variabelen is maximaal één aanhouding per antecedent geteld (zie ook hoofdstuk 1).

De regressiegewichten

Uit de regressiegewichten en de overwegend significante t-waarden in de tabellen 1 t/m 6 van bijlage 3 is af te leiden dat er een positief verband bestaat tussen het aantal aanhoudingen voor rijden onder invloed enerzijds en leeftijd en het aantal aanhoudingen voor de delictgroepen geweld, verkeer, vermogen en vernieling anderzijds. Dit betekent dat personen die vaker gepakt worden voor dronken rijden over het algemeen ouder zijn ook vaker zijn gepakt voor één van de bovengenoemde delictgroepen. De relatie tussen rijden onder invloed en leeftijd van het eerste delict is negatief, over het algemeen hebben personen die vaker voor rijden onder invloed zijn gepakt op jongere leeftijd hun eerste delict gepleegd. Het teken van het regressiegewicht voor geslacht is negatief en significant. Voor deze covariaat zijn de mannen de referentiecategorie, zodat het negatieve regressiegewicht aangeeft dat vrouwen minder vaak gepakt worden voor rijden onder invloed dan mannen. De regressiegewichten voor regio's geven aan hoe de pakkans in een politieregio zich verhoudt tot de referentieregio Rotterdam Rijnmond. In de tabellen 3 t/m 6 is aan de overwegend negatieve waarden te zien dat de kans gepakt te worden voor rijden onder invloed in Rotterdam Rijnmond groter dan in de andere regio's.

Het aantal aanhoudingen voor het voorhanden hebben van wapens vertoont, net als dat bij het rijden onder invloed het geval is, een positief verband met leeftijd en de delictgroepen geweld, vermogen en vernieling. Bij

het voorhanden hebben van wapens blijkt er ook sprake te zijn van een positief verband met het aantal aanhoudingen voor opiumdelicten maar een verband met het aantal aanhoudingen voor verkeersdelicten lijkt hier te ontbreken. Vergelijkbaar is ook het negatief verband met leeftijd van het eerste delict, ook voor het voorhanden hebben van wapens geldt dat het aantal aanhoudingen toeneemt naarmate het eerste delict op jeugdiger leeftijd is gepleegd. In afwijking met rijden onder invloed blijkt er, gezien het ontbreken van significante regressiegewichten, geen verschil in pakkans voor het voorhanden hebben van wapens tussen de politieregio Rotterdam Rijnmond en de overige politieregio's.

Geobserveerde en geschatte frequenties

De tabellen 7a, 8a en 9a laten de geobserveerde frequenties zien en de tabellen 7b, 8b en 9b de geschatte frequenties. In de tabellen 7c, 8c en 9c zijn de gestandaardiseerde residuen tussen geobserveerde en geschatte frequenties weergegeven. De gestandaardiseerde residuen zijn berekend door het verschil tussen geobserveerde en geschatte frequenties te delen door de wortel uit de geschatte frequenties. Een vergelijking tussen de geobserveerde en geschatte frequenties laat voor beide delicten vrijwel steeds hetzelfde patroon zien, de frequentie van de tweevoudige aanhoudingen wordt overschat en de frequenties van de enkelvoudige en de meervoudige aanhoudingen worden onderschat. In de tabellen met de gestandaardiseerde residuen geven daardoor steeds een plus-min-plus patroon te zien.

Populatieschattingen en betrouwbaarheidsintervallen

Tabel 10 in bijlage 3 en 4 geeft een overzicht van de populatieschattingen en betrouwbaarheidsintervallen. De populatieschattingen van observatieperioden van één jaar voor de selectie van regio's van de wapendata zijn niet consistent over de jaren, een gevolg van de geringe spreiding en vulling van het aantal aanhoudingen. De betrouwbaarheidsintervallen zijn hier erg ruim en vertonen in enkele gevallen zelfs een negatieve ondergrens. Ook is duidelijk het effect van een langere observatieperiode te zien, zowel rijden onder invloed als het voorhanden hebben van wapens geeft grotere populatieschattingen te zien bij een toename van de observatieperiode. Op basis van de betrouwbaarheidsintervallen van de observatieperioden van één jaar schatten we in dat op landelijk niveau de populatie verkeersdeelnemers onder invloed van alcohol ongeveer tussen de 400.000 en 500.000 ligt. De betrouwbaarheidsintervallen voor het voorhanden hebben van wapens vertonen iets grotere schommelingen dan bij het rijden onder invloed. Uitgaande van de observatieperioden van twee jaar ligt hier de totale populatie ruwweg tussen de 200.000 en 300.000.

In de tabellen 10 zijn tenslotte ook de gemiddelde pakkansen voor individu uit de populatie berekend. Uit de tabellen is af te leiden dat de pakkans voor rijden onder invloed circa 4 maal zo hoog is dan voor het voorhanden hebben van wapens. Uitgaande van bovengenoemde bestanden is de pakkans voor rijden onder invloed ongeveer 8% en voor het voorhanden hebben van wapens ongeveer 3,5%.

Tabel 11 tenslotte geeft de resultaten van een toets op overdispersie. Met uitzondering van de kortere observatieperioden voor het voorhanden hebben van wapens zijn de toetsen significant, hetgeen een indicatie is voor ongeobserveerde heterogeniteit.

De gemiddelde populatieschattingen in figuur 12 zijn berekend door de populatieschattingen voor respectievelijk de 1-jaars, 2-jaars, 3-jaars, 4-jaars en 5-jaars observatietermijnen te middelen. In de figuur is te zien dat voor zowel rijden onder invloed als voor illegaal wapenbezit de schatting van de omvang van de populatie toeneemt met langere observatieperioden. Dit is waarschijnlijk een gevolg van de schending van de assumptie van een gesloten populatie. Deze schending leidt tot een te hoge schattingen en dat effect neemt toe met de lengte van de observatieperiode. Mogelijk kan voor het effect van een open populatie worden gecorrigeerd door de curven in figuur 12 door te trekken naar het jaar nul. Dit zou bijvoorbeeld onderzocht kunnen worden middels een simulatiestudie. Binnen het huidige onderzoek was hiervoor echter geen ruimte meer.

4. Conclusies en discussie

Realisatie doelstellingen

Het project “Benutting HKS en ontwikkeling vangst-hervangstmethoden” heeft laten zien dat HKS-gegevens na enige bewerking bruikbaar zijn voor het schatten van delinquente ‘dark number’ populaties. De populatieschattingen die in dit project zijn gemaakt van het aantal verkeersdeelnemers onder invloed en het aantal personen met wapens voorhanden lijken realistisch, maar aangezien de werkelijke omvang niet bekend is verdient het aanbeveling de schattingen zoveel mogelijk te toetsen aan schattingen die uit bronnen zijn verkregen.

Dit project heeft aangetoond dat de extractie van data uit het HKS ten behoeve van vangst-hervangstanalyse een gecompliceerd proces is waarin makkelijk fouten kunnen sluipen. Om te voorkomen dat bij toekomstige benutting van het HKS in het kader van de vangst-hervangstmethode dezelfde fouten opnieuw gemaakt worden, is in bijlage 2 het SQL-programma met gevolgde procedure en een handreiking voor het aanmaken van vangst-hervangstdata op basis van HKS-gegevens opgenomen.

In het kader van dit project is de vangst-hervangstmethodiek uitgebreid met een betrouwbaarheidsinterval en een schattingsprocedure voor het negatief binomiaal regressiemodel. Het betrouwbaarheidsinterval geeft inzicht in de nauwkeurigheid van de populatieschatting en draagt daarmee bij aan een meer zinvolle interpretatie van de puntschatting van de populatie. Het afgeknotte negatief binomiaal regressiemodel bleek vanwege enkele specifieke kenmerken van de data, met name een te lage pakkans voor de betreffende delicten, niet toepasbaar. Mogelijk kan dit model in de toekomst wel worden toegepast op andere delicten. Hierbij valt bijvoorbeeld te denken aan opium- of vermogensdelicten, aangezien de pakkans voor deze delicten hoger is.

Poissonproces en populatie

Een belangrijke overweging bij de toepassing van de vangst-hervangstmethode is de vraag in hoeverre de aanname gerechtvaardigd is dat de aanhoudingen voor het delict een Poissonverdeling volgen. Deze vraag komt voort uit het feit dat voor bepaalde delicten de pakkans niet constant is in de tijd. De pakkans voor rijden onder invloed bijvoorbeeld is alleen groter dan nul wanneer de persoon daadwerkelijk dronken achter het stuur zit, in alle andere situaties is er geen sprake van een overtreding en is de pakkans gelijk aan nul. Voor wapenbezit en illegaliteit is de situatie wezenlijk anders, illegalen en personen met een wapen zijn continu in overtreding, en hebben daarom ook constant een pakkans die groter is dan nul.

Dat de verdeling voor deze type delicten beide een Poissonverdeling volgen kan worden aangetoond door de Poissonverdeling op te vatten als de limiet van de binomiaalverdeling met oneindig groot aantal trials en een oneindig kleine kans op succes. Voor de illegalen en wapenbezitters kunnen we nu de totale observatieperiode opdelen in kleine tijdsintervallen waarin maximaal één aanhouding mogelijk is. Het is aannemelijk dat dergelijke intervallen zijn te specificeren, aangezien het onwaarschijnlijk is dat een persoon binnen een termijn van bijvoorbeeld een uur meer dan eens wordt aangehouden voor hetzelfde delict. Wanneer we zo'n tijdsinterval opvatten als één trial hebben we een binomiaalverdeling met een zeer groot aantal trials en per trial een zeer kleine pakkans, en deze verdeling kan benaderd worden met de Poissonverdeling. Het is hierbij niet van belang dat de pakkans in sommige intervallen verschilt van die in andere intervallen, wat bijvoorbeeld het geval zou zijn wanneer de pakkans overdag groter zou zijn dan 's nachts. We krijgen twee binomiaalverdelingen met verschillende pakkans die we door twee Poissonverdelingen met verschillende Poissonparameter kunnen benaderen. Deze twee Poissonverdelingen mogen we vervolgens bij elkaar optellen, omdat geldt dat de som van twee Poissonverdeling weer Poisson verdeeld is.

Voor de delicten die niet continu zijn in de tijd, zoals het rijden onder invloed, kunnen we de totale observatieperiode ook opdelen in kleine intervallen waarin het maximaal aantal aanhoudingen één is. We zouden dit bijvoorbeeld kunnen doen door een interval van het begin van een dronken tot het begin van de volgende dronken rit te nemen, er daarbij van uitgaand dat er hooguit één aanhouding per dronken rit mogelijk is. We krijgen dan weer een binomiaalverdeling waarvan het aantal trials gegeven wordt door het aantal dronken ritten. Over het algemeen zal dit aantal veel kleiner zijn het trials dat voor continue delicten kan worden gespecificeerd, en daarom zal de benadering van de binomiaal door de Poissonverdeling voor niet continue delicten minder goed zijn dan voor continue delicten. Op het eerste gezicht lijkt de Poissonverdeling echter zelfs voor een zeer klein aantal trials de binomiaalverdeling redelijk te benaderen.

Dus alhoewel de pakkans gedurende de observatieperiode niet constant hoeft te zijn in de tijd, wordt de aanname van de Poissonverdeling wel geschonden wanneer de Poissonparameters fluctueren als gevolg van besmetting. Besmetting treedt op wanneer de pakkans als gevolg van een eerdere pakking verandert, en vormt als zodanig een schending van de onafhankelijkheid van de observaties.

Aantal covariaten en populatieschatting

Een fenomeen dat in alle uitgevoerde studies met het Poisson regressiemodel is geobserveerd is dat de populatieschattingen stijgen wanneer er meer covariaten aan het model worden toegevoegd. Twee uitgewerkte voorbeelden hiervan zijn gegeven in het artikel in bijlage 1E. Dat dit fenomeen niet op toevalligheden berust kan worden aangetoond met behulp van Jensen's ongelijkheid. Jensen's ongelijkheid stelt dat de verwachte waarde van de kans op nul aanhoudingen voor modellen met homogene Poissonparameters lager is dan die van modellen met heterogene Poissonparameters. Een schets van het bewijs van deze stelling (voor een uitgebreidere uiteenzetting wordt verwezen naar het artikel in bijlage 1E) wordt gegeven door de nul kans $p(0)$ als functie van de Poissonparameter μ te zien, waarbij geldt dat $p(0) = \exp(-\mu)$. In een homogeen model, dus zonder covariaten, heeft ieder individu een gelijke Poissonparameter, en wordt is de verwachting gelijk

aan $\exp[E(-\mu)]$. In een heterogeen model, dus met covariaten, is de verwachting $E[\exp(-\mu)]$. De term voor het heterogene model is, vanwege het feit dat $\exp(-\mu)$ een convexe functie is, altijd groter dan de laatste term. Samenvattend kunnen we nu zeggen, hoe meer covariaten, hoe groter spreiding in Poissonparameters, des te groter de verwachte waarde van de nul kans en des te hoger de populatieschatting.

Website

Tenslotte kan worden vermeld dat de schattingsprocedures die in het kader van dit project zijn ontwikkeld op korte termijn openbaar zullen worden gemaakt via een website op het internet. De bedoeling van deze website is meer onderzoekers van de mogelijkheden van de vangst-hervangstmethodiek op de hoogte te stellen.