

Bijlage 1A: Afleiding betrouwbaarheidsinterval

Bustami, R., Heijden, P.G.M. van der., Houwelingen, H van. and Engbersen, G., (february 2001), Point and interval estimation of the population size using the truncated Poisson regression model, *Utrecht*.

Point and Interval Estimation of the Population Size Using the Truncated Poisson Regression Model

Rami Bustami¹, Peter van der Heijden¹, Hans van Houwelingen²
and Godfried Engbersen³

¹ Department of Methodology and Statistics, Utrecht University,
P.O. Box 80.140, 3508 TC Utrecht, The Netherlands

² Department of Medical Statistics, Leiden University Medical Center,
P.O.Box 9604, 2300 RC Leiden, The Netherlands

³ Faculty of Social Sciences, Erasmus University Rotterdam,
Postbus 1738, 3000 DR Rotterdam, The Netherlands

SUMMARY

A method to derive point and interval estimates for the total number of individuals in a heterogenous Poisson population is presented. The method is based on the Horvitz-Thompson approach (Kendall and Stuart, 1991). The zero-truncated Poisson regression model is fitted and results are used to obtain point and interval estimates for the total number of individuals in the population. The method is assessed by performing a simulation experiment computing coverage probabilities of Horvitz-Thompson confidence intervals for cases with different sample sizes and Poisson parameters. We illustrate our method using capture-recapture data from the police registration system providing information on illegal immigrants in four large cities in the Netherlands.

Key words: Capture-recapture; Horvitz-Thompson Confidence Interval; Parametric Bootstrap; Population Size Estimation; Truncated Poisson Regression Model.

1 Introduction

Registration files can be used to generate a list of individuals. Such a list may then show only part of the population. The size of the population is the total number of individuals that,

in principle, could have been registered in the list, but not every member of the population appears in the registration file. The aim of this paper is to estimate the size of the population of those individuals, and its characteristics in terms of a number of covariates.

As an example we discuss the estimation of the number of illegal immigrants in the Netherlands. Other possible applications include the estimation of the number of opiate users from a registration of individuals visiting a center that offers them help, or the estimation of the number of drunk drivers from a police registration of individuals caught by the police.

For the estimation of the number of illegal immigrants in the Netherlands (Van der Heijden *et al.* 1997), police registration data are available for 1995, for four cities in the Netherlands: Amsterdam, Rotterdam, The Hague and Utrecht. The registration data are used to derive count data on how often each illegal immigrant is caught by the police. Due to the nature of police registration data, a zero count can not be observed, so the data are truncated.

These count data can be considered as a special form of capture-recapture data, and traditional capture-recapture methods can be employed to estimate the frequency of the zero count. Once an estimate of this frequency is obtained, we are then able to have an estimate for the size of the population of illegal immigrants which can be obtained by adding the zero-count estimate to the observed number of illegal immigrants apprehended. In these traditional methods it is assumed that a member of the population has a constant probability to be apprehended by the police. The assumption of a constant probability of apprehension can be explained as follows: If illegal immigrants are expelled effectively, they often have a low probability to return and be apprehended again. However, in the Netherlands illegal immigrants apprehended by the police cannot always be effectively expelled because either they refuse to mention their nationality, or their home country does not cooperate in receiving them back. In these cases the police requests them to leave the country, but

it is unlikely that they will abide by this request. In the 1995 police registration for the above-mentioned four large cities, 4392 illegal immigrants were filed, 1880 of whom could not be effectively expelled, 2036 are effectively expelled, and for 476 illegal immigrants this information was not available. These data are given in Table 1, with observed frequencies f_k being the number of individuals caught by the police k times, $k = 1, \dots, 6$. For our analysis, we will consider illegal immigrants that were not effectively expelled (further abbreviated as IINEE), as for those the assumption of a constant probability of apprehension is not a priori unrealistic.

Table 1: Illegal immigrants data, observed frequencies for the three groups

Group	f_1	f_2	f_3	f_4	f_5	f_6	Total
Not effectively expelled	1645	183	37	13	1	1	1880
Effectively expelled	1999	33	2	1	1		2036
Other and missing	430	41	5				476
Total	4074	257	44	14	2	1	4392

Let y_i be the number of times individual i ($i = 1, \dots, N_{obs}$) is apprehended ($y_i = 0, 1, \dots$). Due to the assumption of a constant probability, for each individual, the number of times he/she is apprehended follows a Poisson distribution:

$$P(y_i|\lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}, \quad (1.1)$$

which is determined by the Poisson parameter λ ($\lambda > 0$). In model (1.1), the starting model, this parameter is assumed to be the same for each individual (homogeneity assumption). Since we are using registration data, we do not know f_0 , but we can estimate it from f_k ($k > 0$) by assuming that f_k is generated by a truncated Poisson distribution. The term "truncated" refers to the fact that only data about individuals that are apprehended at least once are in the police registration system.

A more general approach, that we adopt in this paper, is to use the truncated Poisson regression model (see e.g. Cameron and Trivedi, 1998, Winkelmann, 1997, Gurmu, 1991, Long, 1997), where the logarithm of the Poisson parameter λ is a linear function of a number of characteristics (variables) known for an observed individual. All the observed individuals with identical characteristics have the same linear function, and thus the same λ . Thus, for each observed individual there is a truncated Poisson distribution, and the Poisson parameter of that individual is determined by the values of the observed variables. This λ determines the probability to be apprehended once, twice, thrice, and so forth. Here, for each observed individual i , \hat{f}_{0i} is estimated and added up to obtain $\hat{f}_0 = \sum_i \hat{f}_{0i}$, $i = 1, \dots, N_{obs}$. In the statistical literature, this phenomenon is referred to as observed heterogeneity. Heterogeneity implies that the Poisson parameters do not have to be equal for all individuals; the term 'observed' refers to the fact that, although it can be assumed that different individuals can have different Poisson parameters, the Poisson parameter is determined by variables that are actually observed, and is not influenced by unobserved variables.

For our data, the zero-truncated Poisson regression model provides estimators for f_0 , the number of IINEE that were not apprehended by the police, and, by adding the IINEE that were actually apprehended, their total number in the population. The relevance of these estimators increases if their confidence interval is known. For example, there is a significant difference when an estimate of 30000 has a confidence interval of 25000 to 40000 compared with 20000 to 60000. For simple truncated Poisson regression models (with categorical covariates), such confidence intervals have already been derived for subpopulations obtained by subdividing the data according to all categorical covariate combinations (see Zelterman, to appear).

In this paper, we extend this work in a number of ways. These include (1) proposing overall confidence intervals for the population size, (2) estimating those intervals by fitting

the truncated Poisson regression model with covariates that can be both categorical as well as continuous, (3) using more parsimonious models as we are not forced to incorporate all categorical covariate combinations, but can restrict our models to include, for example, main effects only, (4) studying characteristics of the whole population as well as of subpopulations (e.g. the probability that members of subpopulations to be apprehended), and (5) assessing model fit by using graphic diagnostics. All the above is not a trivial problem since we do not only have to take into account individual sample fluctuations, but also the probability of an individual to be observed or not. The method that we use to solve this problem is based on the Horvitz-Thompson estimator (Kendall and Stuart, 1991, page 173).

In Section 2, we review traditional capture-recapture methods employing the homogeneous Poisson model to estimate the number of unobserved individuals in the population. In Section 3, the Horvitz-Thompson method is presented and applied to the homogeneous Poisson model. The zero-truncated Poisson regression model is reviewed in Section 4 and Horvitz-Thompson point and interval estimation method for this model is presented in Section 5. Assessment and performance of the method is done using a simulation experiment described in Section 6. Application and data analysis are presented in Section 7. Section 8 is devoted to a brief and general discussion.

2 Traditional Capture-recapture Methods

The zero-truncated Poisson distribution is defined by a probability function conditional on $y > 0$, that is

$$P(y_i | y_i > 0, \lambda) = \frac{P(y_i | \lambda)}{P(y_i > 0 | \lambda)} = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!(1 - \exp(-\lambda))}, \quad y_i = 1, 2, \dots \quad (2.1)$$

with $p(y_i > 0 | \lambda) = 1 - \exp(-\lambda)$, $i = 1, \dots, N$. An estimate $\hat{\lambda}$ for λ can be obtained by fitting the truncated Poisson distribution using, for example, the EM algorithm (see

Seber, 1982). The algorithm gives an estimate for the probability of an individual not to be observed, $\hat{p}_0 = \exp(-\hat{\lambda})$. The number of unobserved individuals (individuals who were not apprehended but had a positive probability to be apprehended), is denoted by \hat{f}_0 and can be calculated as

$$\hat{f}_0 = \frac{\hat{p}_0}{1 - \hat{p}_0} N_{obs},$$

where N_{obs} is the number of observed individuals in the sample.

3 Horvitz-Thompson Point and Interval Estimation of the Total Number of Individuals: Homogeneous Poisson Case

Consider the zero-truncated homogeneous Poisson model defined by (2.1). A point estimate for the total number of individuals in the population may be defined as (Kendall and Stuart, 1991, page 173)

$$\hat{N} = \sum_{i=1}^N \frac{I_i}{p(\lambda)}, \quad (3.1)$$

where $I_i=1$ if individual i is present and $I_i=0$, otherwise, and $p(\lambda) = 1 - \exp(-\lambda)$ is the probability of an individual to be present in the sample. This probability can be estimated by replacing the parameter λ with its estimated value $\hat{\lambda}$ obtained from fitting the zero-truncated homogeneous Poisson model (2.1). The estimate $\hat{\lambda}$ is conditional (given the I_i 's) and unbiased with variance $\sigma^2(\hat{\lambda})$. The variance of \hat{N} is given by (Kendall and Stuart, 1991, page 173)

$$\text{var}(\hat{N}) = E[\text{var}(\hat{N} | I_i)] + \text{var}(E[\hat{N} | I_i]). \quad (3.2)$$

The first term in (3.2) refers to individual sampling fluctuation in the truncated Poisson distribution and is estimated by $\text{var}(\hat{N} | I_i)$. It can be computed by the δ -method, that is

$$\text{var}(\hat{N} | I_i) = \left(\sum_{i=1}^N I_i \frac{\partial}{\partial \lambda} \frac{1}{p(\lambda)} \right)^T \sigma^2(\lambda) \left(\sum_{i=1}^N I_i \frac{\partial}{\partial \lambda} \frac{1}{p(\lambda)} \right). \quad (3.3)$$

For our case, we have $\sum_{i=1}^N I_i = N_{obs}$. So (3.3) can be re-written as

$$\text{var}(\hat{N} | I_i) = \left(N_{obs} \frac{\exp(-\lambda)}{(1 - \exp(-\lambda))^2} \right)^2 \sigma^2(\lambda). \quad (3.4)$$

The second term in (3.2) refers to the probability of an individual to be observed in the sample or not and is given by

$$\begin{aligned} \text{var}(E[\hat{N} | I_i]) &= \text{var} \left(\sum_{i=1}^N I_i \frac{1}{p(\lambda)} \right) = \sum_{i=1}^N I_i \frac{1 - p(\lambda)}{p^2(\lambda)}, \\ &= N_{obs} \frac{\exp(-\lambda)}{(1 - \exp(-\lambda))^2}. \end{aligned} \quad (3.5)$$

The variance of λ in (3.4), $\sigma^2(\lambda)$, is estimated from the derivatives of the log-likelihood of the truncated Poisson distribution. Consider a random sample $Y_1, \dots, Y_{N_{obs}}$ from the truncated Poisson distribution with parameter λ . Then the log-likelihood is defined by

$$\ell = \sum_{i=1}^{N_{obs}} y_i \log \lambda - N_{obs} \lambda - N_{obs} \log(1 - \exp(-\lambda)) - \log \prod_{i=1}^{N_{obs}} y_i!. \quad (3.6)$$

The estimated variance of λ is

$$\hat{\sigma}^2(\lambda) = - \left(\frac{\partial^2 \ell}{\partial \lambda^2} \right)^{-1}.$$

The first derivative of the log-likelihood (3.6) w.r.t. λ is

$$\frac{\partial \ell}{\partial \lambda} = \sum_{i=1}^{N_{obs}} y_i \lambda^{-1} - N_{obs} - \frac{N_{obs} \exp(-\lambda)}{1 - \exp(-\lambda)},$$

and the second derivative is (after simplification)

$$\frac{\partial^2 \ell}{\partial \lambda^2} = - \sum_{i=1}^{N_{obs}} y_i \lambda^{-2} + \frac{N_{obs} \exp(-\lambda)}{(1 - \exp(-\lambda))^2}.$$

So the variance of λ is

$$\hat{\sigma}^2(\lambda) = - \left(\frac{\partial^2 \ell}{\partial \lambda^2} \right)^{-1} = \left(\sum_{i=1}^{N_{obs}} y_i \lambda^{-2} - \frac{N_{obs} \exp(-\lambda)}{(1 - \exp(-\lambda))^2} \right)^{-1}. \quad (3.7)$$

So the total variance in (3.2) is now obtained from (3.4) and (3.5), that is

$$\text{var}(\hat{N}) = \left(N_{obs} \frac{\exp(-\lambda)}{(1 - \exp(-\lambda))^2} \right)^2 \left(\sum_{i=1}^{N_{obs}} y_i \lambda^{-2} - \frac{N_{obs} \exp(-\lambda)}{(1 - \exp(-\lambda))^2} \right)^{-1} + N_{obs} \frac{\exp(-\lambda)}{(1 - \exp(-\lambda))^2} \quad (3.8)$$

For large values of N_{obs} , the variance of the ML estimator of λ is estimated by (see Johnson, Kotz, and Kemp, 1993)

$$\hat{\sigma}^2(\lambda) \approx \lambda(1 - \exp(-\lambda))^2 (1 - \exp(-\lambda) - \lambda \exp(-\lambda))^{-1} N_{obs}^{-1}. \quad (3.9)$$

Note that the variances in (3.7) and (3.9) coincide at the ML estimate of λ , that is when

$$\frac{\partial \ell}{\partial \lambda} = 0,$$

which implies that

$$\sum_{i=1}^{N_{obs}} y_i = \frac{N_{obs} \lambda}{1 - \exp(-\lambda)}.$$

Thus

$$\begin{aligned} \hat{\sigma}^2(\lambda) &= \left(\frac{N_{obs} \lambda}{1 - \exp(-\lambda)} \lambda^{-2} - \frac{N_{obs} \exp(-\lambda)}{(1 - \exp(-\lambda))^2} \right)^{-1}, \\ &= \left(\frac{N_{obs} [1 - \exp(-\lambda) - \lambda \exp(-\lambda)]}{\lambda(1 - \exp(-\lambda))^2} \right)^{-1}, \end{aligned}$$

which is equal to (3.9). Expressions (3.1) and (3.8) for \hat{N} and $\text{var}(\hat{N})$, respectively, are computed by replacing the parameter λ in these expressions by its estimate $\hat{\lambda}$ obtained from fitting the zero-truncated homogeneous Poisson distribution (2.1). The total variance in (3.8) can be used to compute a 95% confidence interval for N : $\hat{N} \pm 1.96 \text{sd}(\hat{N})$, with $\text{sd}(\hat{N}) = \sqrt{\text{var}(\hat{N})}$.

4 The Zero-Truncated Poisson Regression Model

Let $Y_1, \dots, Y_{N_{obs}}$ be a random sample from the zero-truncated Poisson distribution with parameter λ_i , $i = 1, \dots, N_{obs}$. Consider the regression model (Cameron and Trivedi, 1998)

$$\log(\lambda_i) = \boldsymbol{\beta}^{*T} \mathbf{x}_i, \quad (4.1)$$

where $\boldsymbol{\beta}^* = (\alpha, \beta_1, \dots, \beta_p)^T$, and \mathbf{x}_i is a vector of covariate values for subject i , that is $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$. The log-likelihood is given by

$$\ell(\boldsymbol{\beta}^*) = \sum_{i=1}^{N_{obs}} [y_i \log(\lambda_i) - \lambda_i - \log(1 - \exp(-\lambda_i)) - \log(y_i!)].$$

Model (4.1) can be fitted using a Newton-Raphson procedure. The score function is

$$\mathbf{U}(\boldsymbol{\beta}^*) = \frac{\partial \ell(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*}.$$

The current value of the parameter vector $\boldsymbol{\beta}^{*(t)}$ is updated by

$$\boldsymbol{\beta}^{*(t+1)} = \boldsymbol{\beta}^{*(t)} + \mathbf{W}(\boldsymbol{\beta}^{*(t)})^{-1} \mathbf{U}(\boldsymbol{\beta}^{*(t)}),$$

with \mathbf{W} the observed information matrix, that is,

$$\mathbf{W}(\boldsymbol{\beta}^*) = -\frac{\partial^2 \ell(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^* \partial \boldsymbol{\beta}^{*T}}. \quad (4.2)$$

Fitting model (4.1) provides an estimator for the unknown parameter λ_i for the sampled individuals and thus for the probability to be present, $p(\lambda_i)$, $i = 1, \dots, N_{obs}$.

5 Horvitz-Thompson Point and Interval Estimation of the Total Number of Individuals: Heterogeneous Poisson Case

The fit of model (4.1) can be used to derive the Horvitz-Thompson estimator for the total number of individuals in a heterogeneous Poisson population which is then defined by

$$\hat{N} = \sum_{i=1}^N \frac{I_i}{p(\mathbf{x}_i, \boldsymbol{\beta}^*)}, \quad (5.1)$$

where $I_i=1$ if present and 0, otherwise. As in the homogeneous case, the variance of \hat{N} is given by

$$\text{var}(\hat{N}) = E[\text{var}(\hat{N} | I_i)] + \text{var}(E[\hat{N} | I_i]). \quad (5.2)$$

The first term in (5.2) refers to individual sampling fluctuation in the truncated Poisson regression model and is estimated by the δ -method (Kendall and Stuart, 1991), with $E[\text{var}(\hat{N} | I_i)] = \text{var}(\hat{N} | I_i)$, which can be derived as

$$\text{var}(\hat{N} | I_i) = \left(\sum_{i=1}^{N_{\text{obs}}} \frac{\partial}{\partial \boldsymbol{\beta}^*} \frac{1}{p(\mathbf{x}_i, \boldsymbol{\beta}^*)} \right)^T (\mathbf{W}(\boldsymbol{\beta}^*))^{-1} \left(\sum_{i=1}^{N_{\text{obs}}} \frac{\partial}{\partial \boldsymbol{\beta}^*} \frac{1}{p(\mathbf{x}_i, \boldsymbol{\beta}^*)} \right), \quad (5.3)$$

with $\mathbf{W}(\boldsymbol{\beta}^*)$ the observed information matrix obtained in (4.2), $p(\mathbf{x}_i, \boldsymbol{\beta}^*) = 1 - \exp(-\hat{\lambda}_i) = 1 - \exp(-\exp(\boldsymbol{\beta}^{*T} \mathbf{x}_i))$, the probability of an individual i to be observed in the sample, and

$$\sum_{i=1}^{N_{\text{obs}}} \frac{\partial}{\partial \boldsymbol{\beta}^*} \frac{1}{p(\mathbf{x}_i, \boldsymbol{\beta}^*)} = \sum_{i=1}^{N_{\text{obs}}} \frac{-\mathbf{x}_i \exp(\log(\lambda) - \lambda)}{(1 - \exp(-\lambda))^2}.$$

The second term in (5.2) refers to the probability of an individual to be observed or not and is given by

$$\text{var}(E[\hat{N} | I_i]) = \text{var} \left(\sum_{i=1}^N I_i \frac{1}{p(\mathbf{x}_i, \boldsymbol{\beta}^*)} \right) = \sum_{i=1}^{N_{\text{obs}}} \frac{1 - p(\mathbf{x}_i, \boldsymbol{\beta}^*)}{p^2(\mathbf{x}_i, \boldsymbol{\beta}^*)}. \quad (5.4)$$

Expressions (5.1) for \hat{N} and (5.2) for $\text{var}(\hat{N})$ (obtained by adding expressions (5.3) and (5.4)), are computed by replacing the parameter vector $\boldsymbol{\beta}^*$ in the corresponding expressions

by its ML estimate $\hat{\beta}^*$ obtained from fitting the zero-truncated Poisson regression model (4.1). The total variance in (5.2) can be used to compute a 95% confidence interval for N : $\hat{N} \pm 1.96\text{sd}(\hat{N})$.

We have written a GAUSS-386i (GAUSS, version 3.2.8) procedure that fits the truncated Poisson regression model and computes Horvitz-Thompson point and interval estimates for the total number of individuals in the population.

6 A Simulation Experiment

To assess the performance of the Horvitz-Thompson method, an experiment is carried out to investigate the coverage probability of Horvitz-Thompson confidence interval. At the same time we evaluated the coverage probability of the confidence interval yielded by using parametric bootstrapping (see e.g. Efron and Tibshirani, 1993) . The experiment is performed using a homogenous Poisson model (with intercept only), and is done as follows:

Table 2: Coverage probabilities of Horvitz-Thompson (HT) confidence intervals and confidence intervals generated from 500 parametric bootstrap samples (Boot).

(λ, N)	100	250	500	1000
0.5	<i>HT</i> : 0.934	<i>HT</i> : 0.956	<i>HT</i> : 0.946	<i>HT</i> : 0.942
	<i>Boot</i> : 0.878	<i>Boot</i> : 0.892	<i>Boot</i> : 0.922	<i>Boot</i> : 0.950
1	<i>HT</i> : 0.958	<i>HT</i> : 0.946	<i>HT</i> : 0.940	<i>HT</i> : 0.946
	<i>Boot</i> : 0.924	<i>Boot</i> : 0.946	<i>Boot</i> : 0.928	<i>Boot</i> : 0.934
1.5	<i>HT</i> : 0.946	<i>HT</i> : 0.944	<i>HT</i> : 0.960	<i>HT</i> : 0.950
	<i>Boot</i> : 0.944	<i>Boot</i> : 0.950	<i>Boot</i> : 0.954	<i>Boot</i> : 0.958
2	<i>HT</i> : 0.960	<i>HT</i> : 0.956	<i>HT</i> : 0.944	<i>HT</i> : 0.950
	<i>Boot</i> : 0.956	<i>Boot</i> : 0.952	<i>Boot</i> : 0.956	<i>Boot</i> : 0.948
2.5	<i>HT</i> : 0.968	<i>HT</i> : 0.940	<i>HT</i> : 0.960	<i>HT</i> : 0.924
	<i>Boot</i> : 0.960	<i>Boot</i> : 0.942	<i>Boot</i> : 0.950	<i>Boot</i> : 0.928

1. A sample of size: $N = 100, 250, 500, 1000$ is drawn from a non-truncated homogenous Poisson distribution with parameters: $\lambda = 0.5, 1, 1.5, 2, 2.5$.

2. After omitting the zero count, for each of the above 20 observed samples of size N_{obs} , an EM-algorithm is applied to fit a truncated homogenous Poisson distribution to obtain an estimate \hat{f}_0 for f_0 , the zero-count, as well as an estimate $\hat{\lambda}$ for the Poisson parameter λ . Thus, $\hat{N} = N_{obs} + \hat{f}_0$.
3. Horvitz-Thompson 95% confidence intervals are computed.
4. For each of the above 20 observed samples, five hundred bootstrap samples are drawn from a non-truncated homogenous Poisson distribution with \hat{N} and $\hat{\lambda}$ obtained in 2. 95% bootstrap confidence intervals are obtained using the percentile method. Note that by drawing samples of size \hat{N} from a non-truncated distribution instead of drawing samples of size N_{obs} from a truncated distribution, we take into account that there are two sources for the variance of \hat{N} (see equation (3.2)).
5. Steps 1–4 are repeated 500 times.
6. Coverage probabilities were calculated as the proportion of confidence intervals containing the original sample size N . These probabilities were obtained for both the Horvitz-Thompson confidence interval and the bootstrap confidence interval.

The results are summarized in Table 2. The results indicate that the Horvitz-Thompson confidence interval has a higher coverage probability than that of the bootstrap confidence interval when both λ and N are small ($\lambda = 0.5$ and $N = 100, 250$). For other values of λ and N , bootstrap confidence intervals and Horvitz-Thompson confidence intervals are comparable.

In general, the simulation results indicate that the Horvitz-Thompson confidence interval performs well for different values of N and λ . Thus, we will apply it on the data analyzed in the next section.

7 Data Analysis

Consider the IINEE data described in Section 1. The response of interest is the number of times an individual is apprehended by the Police. Several variables were downloaded from the police registration. For our analysis, we use the following four variables as covariates in the truncated Poisson regression model: Nationality, Gender, Age and Reason for being apprehended. The results of fitting the zero-truncated Poisson regression model to the data are shown in Table 3.

Table 3: Truncated Poisson regression model fit to the IINEE data

Regression parameters	MLE	SE	P -value*
Intercept	-2.317	0.449	
Gender (male = 1, female = 0)	0.397	0.163	0.015
Age (< 40 yrs = 1, > 40 yrs = 0)	0.975	0.408	0.017
Nationality (Turkey)	-1.675	0.603	0.006
(North Africa)	0.190	0.194	0.328
(Rest of Africa)	-0.911	0.301	0.003
(Surinam)	-2.337	1.014	0.021
(Asia)	-1.092	0.302	<0.001
(America and Australia)	0.000		
Reason (being illegal = 1, other reason = 0)	0.011	0.162	0.946

$\text{Log-likelihood} = -848.448$

* P -value for Wald test

Table 3 contains maximum likelihood estimates of regression parameters together with their corresponding standard errors and P -values. We recoded the variable Nationality which had six categories by creating five dummy variables considering America and Australia as the reference category. The variables Gender, Age and Nationality (Turkey, Rest of Africa, Surinam or Asia) showed significant contributions to the average number of times an individual is apprehended by the police. The results show that male individuals, individuals who are less than 40 years of age are, on the average, more frequently apprehended by the police. Individuals from Turkey, most parts of Africa, Surinam and Asia are less frequently apprehended

than those from America and Australia. The variable Reason of being apprehended showed no impact on the average number of times an individual is apprehended by the police. That is, the average number of times an individual is apprehended does not significantly depend on the reason he/she was apprehended for.

The fitted model of Table 3 can be used to estimate the total number of IINEE in the population, together with a 95% confidence interval using the Horvitz-Thompson method described in Section 4. Expression (5.1) is used to obtain an estimate of the total number in the population: this leads to $\hat{N} = 12691$. A 95% confidence interval is computed, using the variance in (5.2), as: (7185, 18198).

For the purpose of model comparisons, we fitted several truncated Poisson regression models to the data, computed point estimates as well as 95% confidence intervals for the total number of individuals in the population and performed likelihood-ratio tests. The results are shown in Table 4. The null model yielded the lowest estimate of the total number of IINEE (see column 2, $\hat{N} = 7080$), among all the other estimates obtained from fitting different truncated Poisson regression models. The corresponding 95% Horvitz-Thompson confidence interval is (6363,7797). The largest estimate of N , $\hat{N} = 12691$, was obtained by fitting the full model of Table 2, with corresponding 95% Horvitz-Thompson confidence interval (7185,18198).

In general, the results show that the more covariates we include in the model the larger the estimate and the wider the confidence interval for N we obtain. That is, for an individual i , including more covariates in the model (accounting for observed heterogeneity between individuals) yielded a higher estimate of p_{0i} , the probability for the individual i not to be observed, and hence a larger estimate of N is obtained. This is a general phenomenon (Long, 1997, p. 221). Model comparisons using the likelihood-ratio test indicate that adding the variable Gender to the null model improved the fit significantly ($P = 0.028$). Additional

improvement to the fit is obtained by including Age ($P = 0.018$). The model fit clearly improves by adding the variable Nationality ($P < 0.001$). Finally, including the variable Reason for being caught showed no further significant improvement to the fit.

Table 4: Estimate \hat{N} and HT : 95% confidence interval for N obtained from fitting different truncated Poisson regression models to the IINEE data. Model comparisons using the likelihood-ratio test are also given.

Model	Estimate	G^2	df	P -value*
Null	$\hat{N} = 7080$ $HT : (6363, 7797)$			
Int.+Gender	$\hat{N} = 7319$ $HT : (6504, 8134)$	4.81	1	0.028
Int.+Gender+Age	$\hat{N} = 7807$ $HT : (6637, 8978)$	5.62	1	0.018
Int.+Gender+Age +Nationality	$\hat{N} = 12690$ $HT : (7186, 18194)$	43.07	5	<0.001
Int.+Gender+Age +Nationality+Reason	$\hat{N} = 12691$ $HT : (7185, 18198)$	0.012	1	0.964

* P -value for likelihood-ratio test.

Note that the choice of \hat{N} in Table 4 should be based on the best model fit which is achieved by fitting models including most or all covariates (models given in the 4th and the 5th row of Table 4. Note that including the covariate Reason did not lead to a better fit, and thus \hat{N} of both models are not much different). When the model is misspecified (e.g. null model or models in the 2nd and the 3rd row of Table 4), the model as well as its corresponding \hat{N} should not be interpreted.

A way of examining how good the model of Table 3 fits the data is to compare the observed and the estimated frequencies obtained from the model fit. Similar diagnostic plots for non-truncated Poisson regression model fits are given in Long (1997). This comparison can also be seen as a way of checking for possible remaining overdispersion (see discussion for more details) in the data. A plot comparing observed and estimated frequencies obtained from

the model fit of Table 3 is shown in Figure 1. The plot indicates that the model fits the data adequately.

It is also possible to make comparisons between observed and estimated number of individuals for subgroups in the data. Table 5 shows such comparisons based on the model fit of Table 3. Note that for all subgroups the Horvitz-Thompson estimate of the number of individuals is much larger than the number of individuals observed in the data. This indicates that the probability that illegal individuals not being apprehended is high for all subgroups in the population. Moreover, it is clear that male individuals, individuals who are less than 40 years of age, individuals from North Africa have larger probability to be apprehended, a confirmation of what was observed in Table 3.

Table 5: Comparisons between observed and estimated N for subgroups based on Model fit of Table 3

Subgroup	Observed	Estimated	Observed/Estimated
Males	1482	8880.10	0.167
Females	398	3811.40	0.104
Individuals with Age < 40 years	1769	10506.72	0.168
Individuals with Age > 40 years	111	2184.73	0.051
Individuals from Turkey	93	1740.03	0.053
Individuals from North Africa	1023	3055.23	0.335
Individuals from Rest of Africa	243	2058.00	0.118
Individuals from Surinam	64	2387.75	0.027
Individuals from Asia	284	2741.96	0.104
Individuals from America and Australia	173	708.47	0.244
Individuals caught for reason Being illegal	259	1631.68	0.159
Individuals caught for Other reason	1621	11059.77	0.147

8 Discussion

The Horvitz-Thompson method was presented to estimate the total number of individuals in a heterogeneous Poisson population. The truncated Poisson regression model was uti-

Figure 1: Comparison of the observed and estimated counts obtained from fitting the model of Table 3 to the IINEE data

lized to estimate f_0 , the number of individuals who were not apprehended by the police, but have positive probability of apprehension. The method was assessed using a simulation experiment and it was proved to be appropriate. Results from fitting the truncated Poisson regression model, a typical model for count data that accounts for heterogeneity between subjects, were utilized to obtain Horvitz-Thompson point and interval estimates. Model comparisons showed that including more significant covariates in the model yielded a larger point estimate and a wider confidence interval for the total number of individuals in the population. Note that other models that account for unobserved heterogeneity (overdispersion) between individuals, such as the zero-truncated negative binomial regression model, were not used in this work. Such models take into account other sources of heterogeneity between individuals which were not observed in the data in terms of covariates. The zero-truncated negative binomial model incorporates overdispersion (which is accounted for by including an additional parameter α in the model) in the sense that the truncated variance of the negative binomial exceeds the truncated variance of the Poisson, which is a limiting case obtained as $\alpha \rightarrow 0$ (see Grogger and Carson, 1991, Greene, 1997 for more details). An implementation of the Horvitz-Thompson method to results from fitting such models will be the subject of a future publication.

Acknowledgments

The authors wish to thank the Dutch Ministry of Justice for their financial support and for making the police registration data available for statistical research.

References

- Cameron, A.C. and Trivedi, P.** (1998). *Regression Analysis of Count Data*. Cambridge University Press, USA.
- Efron, B. and Tibshirani, R.** (1997). *An Introduction to the Bootstrap*. Chapman and Hall, USA.
- Greene, W.** (1998). *Econometric Analysis*. Printice-Hall International, Inc., USA.
- Grogger, J.T. and Carson, R.T.** (1991). Models for truncated counts. *Journal of Applied Econometrics*, **6**, No. 2, 225-238.
- Gurmu, S.** (1991). Tests for detecting overdispersion in the positive Poisson regression model. *Journal of Business & Economic Statistics*, **9**, No. 2, 215-222.
- Johnson N.L., Kotz, S., Kemp A.W.** (1993). *Univariate Discrete Distributions*. Second Edition, Wiley, New York.
- Kendall, M., Stuart, A.** (1991). *Advanced Theory of Statistics*. Second Edition, Charles Griffen & Company Limited. London.
- Long, J.** (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage. California, USA.
- Seber, G.** (1982). *The Estimation of Animal Abundance and Related Parameters*. Second edition. Charles Griffen & Company Limited. London.
- Van der Heijden, P.G.M, Zelterman, D., Engbersen, G. and van der Leun, J.** (1997). Estimating the number of illegals in the Netherlands with the truncated Poisson regression model. Unpublished manuscript.

Winkelmann, R. (1997). *Econometric Analysis of Count Data*. Springer-Verlag. Berlin. Heidelberg.

Zelterman, D. (to appear). *GENMOD Applications for Categorical Data*. Cary, N.C.: SAS Institute.

Bijlage 1B: Toetsing betrouwbaarheidsinterval

Bustami, R. (november 2000), Comparison of Confidence Intervals for the Total Number of individuals: homogeneous Poisson Case, without covariates, Utrecht.

Comparison of Confidence Intervals for the Total Number of Individuals: Homogenous Poisson Case, Without Covariates

In this report, an experiment is performed to compare bootstrap confidence intervals with Horvitz-Thompson (Kendall and Stuart, 1991, see the mathematical derivation below) confidence intervals for the total number of individuals in a homogenous Poisson population without covariates. The experiment is done as follows: Samples of size: $N = 100, 250, 500, 1000$ are drawn from the Poisson distribution with parameters: $\lambda = 0.5, 1, 1.5, 2, 2.5$. For the observed sample (after throwing the zeros away) of size $Nobs$, an EM-algorithm (or a Newton-Raphson algorithm) is applied to estimate \hat{m}_0 , the number of individuals who were unobserved, as well as the Poisson parameter $\hat{\lambda}$. The total number of individuals, $\hat{N} = Nobs + \hat{m}_0$. Five hundred bootstraps were performed with \hat{N} and $\hat{\lambda}$ for all different combinations of N and λ . The results are shown in Table 1. Note that Table 1 contains bootstrap confidence intervals obtained using option 5 (B5), these are obtained as follows: to illustrate this option, suppose that the estimated total number of individuals $\hat{N} = 77.4$, then the procedure is done as follows:

- Option 5
 1. Choose $\hat{N}_{random} = 77$ or 78 with probabilities 0.6 and 0.4 , respectively.
 2. A sample of size \hat{N}_{random} is generated from the Poisson distribution with parameter λ and the corresponding Poisson counts are obtained.
 3. A truncated Poisson model is fitted to the above counts to obtain an estimate for the total number of individuals \hat{N} .
 4. Repeat the above steps 500 times to obtain a 500-dimensional vector $\hat{\mathbf{N}} = (\hat{N}_1, \dots, \hat{N}_{500})$.

Note also that in Table 1, for option 5, the bias is calculated and denoted by $BS5$. The bias is equal to $|\hat{N} - \text{median}(\hat{\mathbf{N}})|$.

- Horvitz-Thompson
 1. Estimator

$$\hat{N} = \sum_{i=1}^{Nobs} \frac{\delta_i}{p_i(\hat{\lambda})},$$

where $\delta_i=1$ if present and 0 , otherwise, and $p_i(\lambda)$ is the probability to be present. The unknown parameter λ is estimated from additional information on the sampled individuals. The estimator $\hat{\lambda}$ is conditional (given the δ_i 's) unbiased with estimated covariance matrix $\hat{\Sigma}(\lambda)$.

2. Variance

The variance of \hat{N} is given by

$$\text{var}(\hat{N}) = E[\text{var}(\hat{N} | \delta_i)] + \text{var}(E[\hat{N} | \delta_i]), \quad (1)$$

The first term is estimated by $\text{var}(\hat{N} | \delta_i)$ which can be computed by the δ -method, that is

$$\text{var}(\hat{N} | \delta_i) = \left(\sum_{i=1}^{Nobs} \delta_i \frac{\partial}{\partial \lambda} \frac{1}{p_i(\hat{\lambda})} \right)^T \hat{\sigma}(\lambda) \left(\sum_{i=1}^{Nobs} \delta_i \frac{\partial}{\partial \lambda} \frac{1}{p_i(\hat{\lambda})} \right). \quad (2)$$

For our case, we have $\sum \delta_i = Nobs.p_i(\hat{\lambda}) = \hat{p}_i = \hat{p} = 1 - \exp(-\lambda)$ and $\hat{\Sigma}(\lambda) = \text{var}(\hat{\lambda})$. So (2) can be re-written as

$$\text{var}(\hat{N} | \delta_i) = \left(Nobs \frac{\exp(-\hat{\lambda})}{(1 - \exp(-\hat{\lambda}))^2} \right)^2 \text{var}(\hat{\lambda}). \quad (3)$$

The second term in (1), acting as if $E[1/p_i(\hat{\lambda})] = 1/p_i(\lambda)$, is given by

$$\text{var} \left(\sum_{i=1}^{Nobs} \delta_i \frac{1}{p_i(\lambda)} \right) = \sum_{i=1}^{Nobs} \frac{1 - p_i(\lambda)}{p_i^2(\lambda)},$$

which can be estimated by

$$\sum_{i=1}^{Nobs} \delta_i \frac{1 - p_i(\hat{\lambda})}{p_i^2(\hat{\lambda})} = Nobs \frac{\exp(-\hat{\lambda})}{(1 - \exp(-\hat{\lambda}))^2}. \quad (4)$$

The variance of $\hat{\lambda}$ in (3) is estimated from the derivatives of the log-likelihood of the truncated Poisson distribution. Consider a random sample X_1, \dots, X_N (note that $N = Nobs$) from the Truncated Poisson distribution with parameter λ . Then the likelihood is defined by

$$\ell = \prod_{i=1}^N \frac{\lambda^{x_i} \exp(-\lambda)}{x_i!(1 - \exp(-\lambda))} = \frac{\lambda^{\sum_i x_i} \exp(-N\lambda)}{(1 - \exp(-\lambda))^N \prod_{i=1}^N x_i!}.$$

So,

$$\log \ell = \log \left[\lambda^{\sum_i x_i} \exp(-N\lambda) \right] - \log \left[(1 - \exp(-\lambda))^N \prod_{i=1}^N x_i! \right]. \quad (5)$$

The variance of λ is

$$\text{var}(\lambda) = - \left(\frac{\partial^2 \log \ell}{\partial \lambda^2} \right)^{-1}.$$

(5) is equivalent to

$$\log \ell = \sum_i x_i \log \lambda - N\lambda - N \log(1 - \exp(-\lambda)) - \log \prod_{i=1}^N x_i!.$$

Hence,

$$\frac{\partial \log \ell}{\partial \lambda} = \sum_i x_i \lambda^{-1} - N - \frac{N \exp(-\lambda)}{1 - \exp(-\lambda)}.$$

So,

$$\frac{\partial^2 \log \ell}{\partial \lambda^2} = - \sum_i x_i \lambda^{-2} - \frac{-N(1 - \exp(-\lambda)) \exp(-\lambda) - N \exp(-\lambda) \exp(-\lambda)}{(1 - \exp(-\lambda))^2},$$

which is equivalent to

$$\frac{\partial^2 \log \ell}{\partial \lambda^2} = - \sum_i x_i \lambda^{-2} - \frac{-N \exp(-\lambda) + N \exp(-2\lambda) - N \exp(-2\lambda)}{(1 - \exp(-\lambda))^2}.$$

Thus

$$\frac{\partial^2 \log \ell}{\partial \lambda^2} = - \sum_i x_i \lambda^{-2} + \frac{N \exp(-\lambda)}{(1 - \exp(-\lambda))^2}.$$

So the variance of λ is

$$\text{var}(\hat{\lambda}) = - \left(\frac{\partial^2 \log \ell}{\partial \lambda^2} \right)^{-1} = \left(\sum_i x_i \lambda^{-2} - \frac{N \exp(-\lambda)}{(1 - \exp(-\lambda))^2} \right)^{-1}. \quad (6)$$

For large values of N , the variance of the ML estimator $\hat{\lambda}$ is given by (see Johnson, Kotz, and Kemp, 1993)

$$\text{var}(\hat{\lambda}) \approx \lambda(1 - \exp(-\lambda))^2 (1 - \exp(-\lambda) - \lambda \exp(-\lambda))^{-1} N^{-1}. \quad (7)$$

So the total variance in (1) is now obtained from (3) and (4), that is

$$\text{var}(\hat{N}) = \left(N_{obs} \frac{\exp(-\hat{\lambda})}{(1 - \exp(-\hat{\lambda}))^2} \right)^2 \left(\sum_i x_i \hat{\lambda}^{-2} - \frac{N \exp(-\hat{\lambda})}{(1 - \exp(-\hat{\lambda}))^2} \right)^{-1} + N_{obs} \frac{\exp(-\hat{\lambda})}{(1 - \exp(-\hat{\lambda}))^2}.$$

Note that the variances in (6) and (7) coincide at the ML estimate $\hat{\lambda}$, that is when

$$\frac{\partial \log \ell}{\partial \lambda} = 0,$$

which implies that

$$\sum_i x_i = \frac{N_{obs} \lambda}{1 - \exp(-\lambda)}.$$

Thus

$$\text{var}(\hat{\lambda}) = \left(\frac{N_{obs} \lambda}{1 - \exp(-\lambda)} \lambda^{-2} - \frac{N_{obs} \exp(-\lambda)}{(1 - \exp(-\lambda))^2} \right)^{-1},$$

which is equivalent to

$$\text{var}(\hat{\lambda}) = \left(\frac{N_{obs} [1 - \exp(-\lambda) - \lambda \exp(-\lambda)]}{\lambda(1 - \exp(-\lambda))^2} \right)^{-1},$$

which is equivalent to (7).

- Conclusions: The results in Table 1 show that

1. The bootstrap confidence intervals obtained using option 5 and the Horvitz-Thompson confidence intervals are comparable when λ is large.
2. The bootstrap confidence intervals obtained using option 5 are, in general, wider than the Horvitz-Thompson confidence intervals.
3. The bias is small in percentage of the estimate \hat{N} .
4. A further study needs to be performed to compare Horvitz-Thompson confidence interval with the bootstrap confidence interval obtained using option 5. This study involves double-bootstrapping to help identify which confidence interval is better, the analytical (Horvitz-Thompson) or the bootstrap confidence interval.

Table 1: **Results of 500 bootstraps**: HT: 95% Horvitz-Thompson confidence interval, B5: 95% Bootstrap confidence interval using option 5, BS5: Bias using option 5

(λ, N)	100	250	500	1000
0.5	$\hat{N} = 127.51$ <i>HT</i> : (49.58, 205.44) <i>B5</i> : (87.57, 224.05) <i>BS5</i> = 4.73	$\hat{N} = 222.49$ <i>HT</i> : (147.24, 297.74) <i>B5</i> : (160.24, 327.94) <i>BS5</i> = 6.72	$\hat{N} = 646.94$ <i>HT</i> : (474.77, 819.12) <i>B5</i> : (544.05, 808.80) <i>BS5</i> = 1.30	$\hat{N} = 966.81$ <i>HT</i> : (813.15, 1120.46) <i>B5</i> : (819.72, 1142.96) <i>BS5</i> = 4.64
1	$\hat{N} = 94.34$ <i>HT</i> : (73.65, 115.03) <i>B5</i> : (74.01, 120.64) <i>BS5</i> = 0.99	$\hat{N} = 242.47$ <i>HT</i> : (203.32, 281.62) <i>B5</i> : (210.79, 284.54) <i>BS5</i> = 0.33	$\hat{N} = 530.40$ <i>HT</i> : (467.54, 593.25) <i>B5</i> : (479.06, 582.79) <i>BS5</i> = 0.45	$\hat{N} = 970.09$ <i>HT</i> : (903.62, 1036.57) <i>B5</i> : (895.69, 1043.22) <i>BS5</i> = 0.74
1.5	$\hat{N} = 96.21$ <i>HT</i> : (83.79, 108.63) <i>B5</i> : (84.12, 110.03) <i>BS5</i> = 0.47	$\hat{N} = 238.00$ <i>HT</i> : (216.91, 259.09) <i>B5</i> : (216.56, 261.43) <i>BS5</i> = 0.57	$\hat{N} = 483.99$ <i>HT</i> : (453.36, 514.62) <i>B5</i> : (452.32, 515.97) <i>BS5</i> = 0.85	$\hat{N} = 1005.42$ <i>HT</i> : (960.57, 1050.28) <i>B5</i> : (959.04, 1049.68) <i>BS5</i> = 2.06
2	$\hat{N} = 96.94$ <i>HT</i> : (88.33, 105.55) <i>B5</i> : (87.35, 104.96) <i>BS5</i> = 0.19	$\hat{N} = 256.78$ <i>HT</i> : (241.15, 272.41) <i>B5</i> : (240.70, 271.10) <i>BS5</i> = 0.76	$\hat{N} = 504.96$ <i>HT</i> : (482.68, 527.25) <i>B5</i> : (485.12, 523.00) <i>BS5</i> = 0.63	$\hat{N} = 1006.70$ <i>HT</i> : (976.95, 1036.46) <i>B5</i> : (974.82, 1036.07) <i>BS5</i> = 0.65
2.5	$\hat{N} = 94.88$ <i>HT</i> : (88.89, 100.86) <i>B5</i> : (86.83, 100.71) <i>BS5</i> = 0.42	$\hat{N} = 249.82$ <i>HT</i> : (239.87, 259.77) <i>B5</i> : (239.78, 259.03) <i>BS5</i> = 0.78	$\hat{N} = 498.97$ <i>HT</i> : (483.41, 514.53) <i>B5</i> : (482.98, 512.24) <i>BS5</i> = 0.60	$\hat{N} = 988.97$ <i>HT</i> : (968.09, 1009.85) <i>B5</i> : (966.29, 1007.10) <i>BS5</i> = 1.38

Bijlage 1C: Toepassing afgeknot negatief binomiaalmodel

Bustami, R. (maart 2001), The use of truncated Poisson and Negative Binomial regression models for estimating the number of opiate users in Rotterdam, *Utrecht*, intern manuscript.

March 8, 2001

The Use of Truncated Poisson and Negative Binomial Regression Models for Estimating the Number of Opiate Users in Rotterdam

In this report, we use the truncated Poisson and the truncated negative binomial regression model to compute a confidence interval for the total number of opiate users in the Rotterdam. The data used here were previously used in the analysis that resulted in the paper of Smit, Toet and Van der Heijden (1997): Estimating the number of opiate users in Rotterdam using statistical models for incomplete count data (EMCCDA report).

The response of interest is *cro*: Count of visits (actually, count of episodes during which visits were brought without interruption). The following variables are used in the truncated Poisson and the truncated negative binomial regression model as covariates:

1. Sex: (women = 0, men = 1)
2. Mar: married (no = 0, yes = 1)
3. Dut: Dutch nationality (no = 0, yes = 1)
4. Par: living together with a partner (no = 0, yes = 1)
5. Sur: of Surinam origin (no = 0, yes = 1)
6. Age (in years)
7. Inc: source of income (1= income from work, 0 =else).

Different truncated Poisson and truncated negative binomial regression models were fitted to the above data and 95% confidence intervals for the total number of opiate users were computed. The models fitted are the following:

- Null model: intercept only
- Model 1 (Full model): intercept and the covariate(s): Sex, mar, dut, par, sur, age, inc
- Model 2: intercept and the covariate(s): Sex, Dut, Par, Sur, Age
- Model 3: intercept and the covariate(s): Sex, Mar, Dut, Par, Sur, Age
- Model 4: intercept and the covariate(s): Sex, Mar, Dut, Par, Sur
- Model 5: intercept and the covariate(s): Sex, Mar, Dut, Par
- Model 6: intercept and the covariate(s): Sex, Mar, Dut
- Model 7: intercept and the covariate(s): Sex, Mar
- Model 8: intercept and the covariate(s): Sex
- Model 9: intercept and the covariate(s): Mar
- Model 10: intercept and the covariate(s): Dut

- Model 11: intercept and the covariate(s): Sur
- Model 12: intercept and the covariate(s): Age
- Model 13: intercept and the covariate(s): Inc

The results of fitting the above models are shown in Table 1. Note that Table 1 contains estimate (\hat{N}) as well as *HT*: 95% confidence interval for the total number of opiate users in the population.

Table 1: **Estimate (\hat{N}), and *HT*: 95% confidence interval for the total number of opiate users in the population obtained from fitting the above truncated Poisson and truncated negative binomial regression models to the data**

<i>Model</i>	Truncated Poisson	Truncated negative binomial
<i>Null</i>	$\hat{N} = 2936.76$ <i>HT</i> : (2833.64, 3039.87)	$\hat{N} = 4193.27$ <i>HT</i> : (1469.93, 6916.61)
1	$\hat{N} = 2992.06$ <i>HT</i> : (2879.30, 3104.81)	$\hat{N} = 4122.04$ <i>HT</i> : (1639.92, 6604.16)
2	$\hat{N} = 2990.65$ <i>HT</i> : (2878.14, 3103.17)	$\hat{N} = 4123.68$ <i>HT</i> : (1635.77, 6611.59)
3	$\hat{N} = 2991.14$ <i>HT</i> : (2878.56, 3103.71)	$\hat{N} = 4121.89$ <i>HT</i> : (1637.14, 6606.64)
4	$\hat{N} = 2969.51$ <i>HT</i> : (2860.56, 3078.45)	$\hat{N} = 4157.97$ <i>HT</i> : (1576.11, 6739.82)
5	$\hat{N} = 2961.40$ <i>HT</i> : (2854.05, 3068.75)	$\hat{N} = 4159.11$ <i>HT</i> : (1552.74, 6765.49)
6	$\hat{N} = 2953.56$ <i>HT</i> : (2847.52, 3059.61)	$\hat{N} = 4171.80$ <i>HT</i> : (1534.41, 6809.20)
7	$\hat{N} = 2946.43$ <i>HT</i> : (2841.60, 3051.26)	$\hat{N} = 4179.56$ <i>HT</i> : (1507.43, 6851.69)
8	$\hat{N} = 2944.92$ <i>HT</i> : (2840.36, 3049.48)	$\hat{N} = 4181.07$ <i>HT</i> : (1501.06, 6861.08)
9	$\hat{N} = 2938.54$ <i>HT</i> : (2835.11, 3041.97)	$\hat{N} = 4190.65$ <i>HT</i> : (1477.96, 6903.33)
10	$\hat{N} = 2943.26$ <i>HT</i> : (2839.03, 3047.49)	$\hat{N} = 4182.62$ <i>HT</i> : (1493.88, 6871.36)
11	$\hat{N} = 2937.56$ <i>HT</i> : (2834.31, 3040.80)	$\hat{N} = 4191.76$ <i>HT</i> : (1468.07, 6915.45)
12	$\hat{N} = 2954.72$ <i>HT</i> : (2848.49, 3060.95)	$\hat{N} = 4164.06$ <i>HT</i> : (1533.15, 6794.97)
13	$\hat{N} = 2936.83$ <i>HT</i> : (2833.71, 3039.96)	$\hat{N} = 4193.17$ <i>HT</i> : (1469.79, 6916.55)

Truncated Poisson and truncated negative binomial fits for the null model, and models 1 and 2 are shown in Tables 2-7. Maximum likelihood estimates of regression parameters are given as well as their corresponding standard errors and P -values. We started the model search process by fitting models with intercept only (Tables 2 and 3). The truncated negative binomial dispersion parameter α showed a highly significant P -value ($P < 0.001$).

In a second step, we fitted both models including all covariates. For both the truncated Poisson model and the truncated negative binomial model fits (Tables 4 and 5), the variables Sex, Dut, Par, Sur and Age showed significant contribution to the average number of visits the opiate users receive. The results show that male opiate users receive more visits than female ones. Dutch and Surinamese individuals are on the average more visited than others. Moreover, unmarried individuals receive more visits than married ones, while individuals having partners are on the average more frequently visited than those who do not have partners. Finally, the older the opiate user is the fewer the number of visits he/she receives.

Table 2: *Null model: Truncated Poisson regression with intercept only*

Regression parameters	MLE	SE	P -value*
Intercept	0.160	0.025	
<i>Log-likelihood</i> = -2406.62			
* P -value for Wald test			

Table 3: *Null model: Truncated negative binomial regression with intercept only*

Regression parameters	MLE	SE	P -value*
Intercept	-0.413	0.118	
Alpha (overdispersion)	1.139	0.293	<0.001
<i>Log-likelihood</i> = -2335.58			
* P -value for Wald test			

In a third step, model fits are obtained by fitting both models again after dropping variables with insignificant contributions (Tables 6 and 7). Though it showed no impact on the average number of visits the opiate users receive, we kept the variable Par in the final truncated negative binomial model for the purpose of model comparisons (see Table 7). The model fit of Table 7 showed that the overdispersion parameter is highly significant ($P < 0.001$), indicating that the data are better analyzed using the truncated negative binomial model which accounts for unobserved heterogeneity between individuals that is not explained by the covariates.

Table 8 summarizes results from model comparisons based on the likelihood-ratio test (G^2). For both the truncated Poisson and the truncated negative binomial fits, model 1 (full model including all covariates) is to be preferred over the null model (see lines 1 and 3 of Table 8, $P < 0.001$). Moreover, the truncated negative binomial model with covariates showed to have the best fit for the data (see line 5 of Table 8, $P < 0.001$).

Back to the results shown in Table 1 (second column), the null truncated Poisson model yielded the lowest estimate of the total number of opiate users in the population ($\hat{N} = 2936.76$) among all the other estimates given in the same column which are obtained from fitting different truncated Poisson models. The corresponding 95% Horvitz-Thompson confidence interval is (2833.64, 3039.87). Model 1, which corresponds to the best truncated Poisson fit yielded the largest estimate ($\hat{N} = 2992.06$), with corresponding 95% confidence interval (2879.30, 3104.81).

In general, the results for the truncated negative binomial case (third column in Table 1), showed larger estimates of the total number of opiate users and wider confidence intervals than for those obtained from the truncated Poisson models (column 2). Moreover, in contrast to the truncated Poisson case, the null truncated negative binomial model yielded the largest estimate, $\hat{N} = 4193.27$, among all the other estimates given in the same column which are obtained from fitting different truncated negative binomial models. The corresponding 95% confidence interval is (1469.93, 6916.61). In addition, Model 1 (truncated negative binomial), corresponding to the best fit to the data (see table 8, third row) yielded a low estimate ($\hat{N} = 4122.04$), with corresponding 95% confidence interval (1639.92, 6604.16).

Table 4: *Model 1: Truncated Poisson regression*

Regression parameters	MLE	SE	<i>P</i> -value*
Intercept	0.371	0.146	
Sex (women = 0, men = 1)	0.193	0.060	0.001
Mar (no = 0, yes = 1)	-0.062	0.089	0.488
Dut (no = 0, yes = 1)	0.213	0.066	0.001
Par (no = 0, yes = 1)	0.130	0.054	0.017
Sur (no = 0, yes = 1)	0.264	0.085	0.002
Age (in years)	-0.017	0.004	<0.001
Inc (1= income from work, 0 =else)	-0.059	0.076	0.440

Log-likelihood = -2384.18

* *P*-value for Wald test

Table 5: *Model 1: Truncated negative binomial regression*

Regression parameters	MLE	SE	<i>P</i> -value*
Intercept	-0.128	0.231	
Sex (women = 0, men = 1)	0.226	0.085	0.008
Mar (no = 0, yes = 1)	-0.064	0.123	0.601
Dut (no = 0, yes = 1)	0.251	0.090	0.006
Par (no = 0, yes = 1)	0.159	0.079	0.044
Sur (no = 0, yes = 1)	0.307	0.120	0.011
Age (in years)	-0.019	0.005	<0.001
Inc (1= income from work, 0 =else)	-0.073	0.110	0.506
Alpha (overdispersion)	0.980	0.246	<0.001

Log-likelihood = -2320.34

* *P*-value for Wald test

Table 6: *Model 2: Truncated Poisson regression with significant covarites only*

Regression parameters	MLE	SE	<i>P</i> -value*
Intercept	0.366	0.146	
Sex (women = 0, men = 1)	0.186	0.059	0.002
Dut (no = 0, yes = 1)	0.220	0.065	<0.001
Par (no = 0, yes = 1)	0.114	0.052	0.030
Sur (no = 0, yes = 1)	0.273	0.084	0.001
Age (in years)	-0.017	0.004	<0.001

Log-likelihood = -2384.75

* *P*-value for Wald test

Table 7: *Model 2: Truncated negative binomial regression with significant covarites only*

Regression parameters	MLE	SE	<i>P</i> -value*
Intercept	-0.139	0.232	
Sex (women = 0, men = 1)	0.217	0.084	0.009
Dut (no = 0, yes = 1)	0.258	0.088	0.003
Par (no = 0, yes = 1)	0.141	0.076	0.062
Sur (no = 0, yes = 1)	0.317	0.118	0.007
Age (in years)	-0.019	0.005	<0.001
Alpha (overdispersion)	0.984	0.247	<0.001

Log-likelihood = -2320.72

* *P*-value for Wald test

Table 8: *Model Comparisons using the likelihood-ratio test.*

Model	G^2	df	P -value*
Truncated Poisson (Null model, Full model (model 1))	44.88	7	<0.001
Truncated Poisson (Model 1, Model 2)	1.14	2	0.566
Truncated negative binomial (Null model, Full model (model 1))	30.48	7	<0.001
Truncated negative binomial (Model 1, Model 2)	0.76	2	0.684
Truncated Poisson-Truncated negative binomial (Null model, Null model)	142.08	1	<0.001
Truncated Poisson-Truncated negative binomial (Model 1, Model 1)	127.68	1	<0.001
Truncated Poisson-Truncated negative binomial (Model 2, Model 2)	128.06	1	<0.001

* P -value for likelihood-ratio test.

Bijlage 1D: Simulatiestudie afgeknot negatief binomiaalmodel

Uit:

Bustami, R. (mei 2001b), The use of truncated Poisson and negative binomial regression models for estimating the number of illegal immigrants in the Netherlands, *Utrecht*, intern manuscript.

November 1, 2001

Simulations investigating power to go from the TPRM to the TNBRM

The following is a description of a Monte Carlo study that investigates power to go from the truncated Poisson regression model to the truncated negative binomial regression model:

1. Generate data with overdispersion (unobserved heterogeneity), that is data from the truncated negative binomial distribution with different values of λ , α and N : $\lambda = 0.5, 1, 2$, $\alpha = 0.05, 0.1, 0.5, 1, 2$, and $N = 100, 200, 500, 1000, 2500$.
2. Fit both the truncated Poisson and the truncated negative binomial regression models to the generated data.
3. Test differences between fits above using the likelihood-ratio χ^2 test.
4. Repeat the above steps 500 times.
5. Compute the proportion of times that the truncated Poisson regression model was rejected.

The results are shown in Table 1.

The results indicate that the higher the values of N , α and λ , the higher the number of times the Poisson regression model is rejected. Note that the Poisson model is hardly rejected when the value of the overdispersion parameter α is small ($\alpha = 0.5$). This result is expected since the Poisson model is obtained when there is no overdispersion, that is, when the overdispersion parameter α equals zero.

Note also that for small values of α , λ and N , a maximum was not reached for a number of simulations, this number is given in parentheses (see Table 1). The main problem with the truncated negative binomial is that it allows a value of $\alpha = \infty$. The corresponding estimate of the population sample size is then infinity as well.

Table 1: Monte Carlo study investigating power to go from the TPRM to the TNBRM. Five hundred simulations were performed and models are compared using the likelihood-ratio test. The numbers between parentheses are the number of times a maximum was not reached.

$\lambda = 0.5$					
N	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
100	0.010 ⁽⁸⁾	0.090 ⁽⁵⁾	0.400 ⁽²⁾	0.690	0.900
200	0.014 ⁽⁸⁾	0.096 ⁽³⁾	0.400 ⁽¹⁾	0.720	0.880
500	0.070 ⁽⁵⁾	0.130 ⁽¹⁾	0.600	0.800	0.960
1000	0.096 ⁽²⁾	0.240 ⁽¹⁾	0.660	0.990	1
2500	0.160 ⁽¹⁾	0.300	0.780	0.998	1
$\lambda = 1$					
100	0.042 ⁽⁶⁾	0.170	0.600	0.900	0.960
200	0.040 ⁽⁵⁾	0.180	0.620	0.900	0.962
500	0.060 ⁽²⁾	0.200	0.800	0.950	0.970
1000	0.192 ⁽¹⁾	0.280	0.980	0.998	1
2500	0.280 ⁽¹⁾	0.300	1	1	1
$\lambda = 2$					
100	0.100 ⁽²⁾	0.200	0.800	0.930	0.960
200	0.160 ⁽²⁾	0.350	0.860	0.950	0.940
500	0.240 ⁽¹⁾	0.500	0.982	0.986	0.990
1000	0.300	0.640	0.998	1	1
2500	0.400	0.900	1	1	1

Bijlage 1E: Voorbeeldanalyses HKS

Heijden, P.G.M. van der, Cruyff, M. and Houwelingen H. van der, (2002). Estimating the size of a criminal population from police registrations using the truncated Poisson regression model, Utrecht .

Estimating the size of a criminal population from
police registrations using the truncated Poisson
regression model

Peter G.M. van der Heijden

Faculty of Social Sciences, Utrecht University, PO Box 80140,
3508 TC Utrecht, the Netherlands

Maarten Cruyff

Faculty of Social Sciences, Utrecht University, PO Box 80140,
3508 TC Utrecht, the Netherlands

Hans C. van Houwelingen

Dept. Medical Statistics, Leiden University Medical Center,
PO Box 9600, 2300 RC Leiden, the Netherlands

Correspondence to P.vanderHeijden@fss.uu.nl. Acknowledgements: This research has been funded partly by the Ministry of Justice, the Netherlands. We gratefully acknowledge the comments of guest-editor Catrien Bijleveld and an anonymous referee, and the help of Leen Prins of the National Police Force Service (KLPD) in preparing the data files.

Summary

The truncated Poisson regression model is used to arrive at point and interval estimates of the size of two offender populations, namely drunk drivers and owners of illegal weapons. The dependent capture-recapture variables are constructed from the Dutch police registration and are counts of individual arrests for both violations. The population size estimates are derived under the assumptions that each count is a realization of a Poisson distribution, and that the Poisson parameters are related to covariates through the truncated Poisson regression model. These assumptions are discussed in detail, and the tenability of the second assumption is assessed by evaluating the marginal residuals and by performing tests on overdispersion. For the illegal weapons example the second assumption seems to hold well, but for the drunk drivers example there is some overdispersion. It is concluded that the method presented is useful, provided that it is used with care.

Keywords: Capture-recapture, Truncated Poisson regression, overdispersion, population size estimation.

1 Introduction

For many policy reasons it is important to have knowledge about the size of specific delinquent populations. One such reason is that it gives insight into the threat these populations may impose on society. Another reason is that it provides an estimate of the workload of the police.

However, estimating the size of a delinquent population may be problematic for various reasons. Counting the number of committed crimes from police registrations may lead to a dark number problem. It may be that the crime is registered but the offender is not known or, as often happens with victimless crimes, that a committed crime is not observed at all. Victim surveys lead to an estimate of the number of crimes the subject reports to have been the victim of in a given period of time, like burglary or robbery. However, they do not provide an estimate of the number of offenders, since usually these are unknown to the victim, nor do they give insight in victimless offences.

Self-report studies can potentially estimate the size of a delinquent population since the subjects are simply asked if they are member of such a population. Problems related to self-report studies are the difficulty of obtaining a representative sample, the risk of socially desirable answers and the need for large samples when offences have low prevalence. For a more elaborate comparison of, and an overview of the literature on police registrations, victim surveys and self-report studies we refer to Wittebrood and Junger (2002).

In this paper we will discuss a way to estimate the number of offenders from police data. The data we use are obtained from the Dutch police registration system HKS (Herkennings Dienst Systeem). In this system offences with a known offender are registered. Each report of an offence has an offender identification, so it is possible to construct an offender-based data set. Assume that we want to estimate the number of offenders of a specific type of crime, say illegal gun ownership. Then we can construct an offender-based data set that will have a number of offenders in possession of a gun apprehended once, a number apprehended twice, a number three times, and so on. Note that offenders never apprehended are not part of this offender-based data set. Yet, if we would be able to estimate the number of offenders never apprehended, we would have an estimate of the total number of illegal gun owners.

The aim is to derive an estimate of the number of offenders never apprehended, using the data about the offenders who are apprehended at least once. We derive such estimates under two assumptions. First, the number of times an offender is apprehended is a realization of a Poisson distribution. Second, the logarithm of the Poisson parameter for an offender is a linear function of his covariates. We will discuss these assumptions in more detail at the end of the introduction.

At this point we want to indicate how we can find an estimate of the size of the population never apprehended using these assumptions. Assume an offender for whom the Poisson parameter specifies that the probability to be apprehended at least once is .25. Yet he is apprehended. This implies that, for this offender, we estimate that there are three additional offenders who are not apprehended. By performing this trick for every offender who is apprehended, and by adding up all individual estimates, we have

an estimate of the total number of offenders that are not apprehended, and this solves our problem.

The methods employed in this paper originate from biology, where they are used to estimate animal abundance. The data we use are a specific form of capture-recapture data. In capture-recapture data (re)captures are made at specific time points and for each animal seen at least once there is a capture history, for example, if there are five capture times, such a history could be 01101 if the animal is seen at captures 2, 3 and 5, and not seen at captures 1 and 4. Here, we use only the total number of times someone is captured as we are collecting data in continuous time. Typically, in the biological application area covariate information is not available or not used, leading to a basic model in which the Poisson parameters are assumed to be homogenous over the animals. For an overview of this area we refer to Seber (1982, chapter 4; 1986), Chao (1988) and Zelterman (1988; 2001, chapter 7). In the statistical literature this problem is also known as the estimation of the size of a truncated sample (Sanathanan, 1977), the estimation of the number of (unseen) species (Efron and Thisted, 1976; Bunge and Fitzpatrick, 1993), or the estimation of the size of a population using samples of size one (Wilson and Collins, 1992).

In criminology we found some early contributions by Greene and Stollmack (1981) who use arrest data to estimate the number of adults committing felonies and misdemeanors in Washington D.C. in 1974/5, Rossmo and Routledge (1990) who estimate migrating (or fleeing) fugitives in 1984, and prostitutes in 1986/7, both in Vancouver, and Collins and Wilson (1990) who use arrest data to estimate automobile theft in the Australian capital territory for adults and juveniles in 1987. These early contributions have in common that they do not systematically pay attention to covariate information on the apprehended individuals, nor to confidence intervals for the point estimates of the number of individuals not apprehended. Our early statistical work in criminology is also in this vein (van der Heijden et al, 1993; Smit et al, 1993).

In more recent work, we have incorporated covariate information by using the truncated Poisson regression model, well known in econometrics (Greene, 1997, chapter 19; Cameron and Trivedi, 1998, chapter 4; Long, 1997, ch. 8). We developed a method to

estimate a frequency for the zero-count and a confidence interval for this point estimate (Bustami et al., 2001). Thus far we have experience with this method in the estimation of the number of illegal immigrants in the four largest Dutch cities in 1995 (see van der Leun et al., 1998; see Hoogteijling, 2002, for a critical evaluation of this method) and in the Netherlands in general from 1997 to 2000 (see Engbersen et al., 2002), and with the number of opiate users in Rotterdam in 1994 (Smit, Toet and van der Heijden, 1996). In this paper we will present two further examples, namely on the number of illegal weapon owners and the number of car drivers under the influence of alcohol.

1.1 Assumptions

As the methodology employed originates from biology, and we employ it in the field of criminology, we discuss the assumptions of the methodology in more detail. It is not straightforward that assumptions realistic for animals will also be realistic for (human) offenders.

The first assumption is that the number of times an individual is apprehended is a realization of a Poisson distribution. Johnson, Kotz and Kemp (1993) discuss the genesis of the Poisson distribution and state that it originally was found by Poisson as the limit of a binomial distribution with success probability p and N realizations, where N tends to infinity and p tends to zero, while Np remains finite and equal to λ . It turns out that even for small N the Poisson distribution approximates the binomial distribution reasonably well if p is sufficiently small. For example, for $N = 3$ and $p = .033$, and for $N = 10$ and $p = .01$, the probabilities of counts 0, 1, 2 and 3 are already very close to those of a Poisson distribution with $\lambda = .1$, and similarly for $N = 3$ and $p = .0033$, and for $N = 10$ and $p = .001$ with $\lambda = .01$.

Referring to Charlier (1905), Johnson et al. (1993) note that the probability of success p does not have to be constant for the Poisson limit to hold. So, loosely speaking, it follows for the type of applications we are discussing that an individual does not need to have a constant probability to be apprehended, but it suffices if he/she can be apprehended a number of times. A property of the Poisson distribution related to the result of Charlier (1905) is that, if X_1 is a realization of a Poisson distribution with

Poisson parameter λ_1 , and X_2 is a realization of a Poisson distribution with Poisson parameter λ_2 , then $X_1 + X_2$ is a realization of a Poisson distribution with Poisson parameter $\lambda_1 + \lambda_2$. So, again we see that the probability to be apprehended does not need to be constant: if we split up the full period of data collection into a larger number of sub-periods where in each of these sub-periods the count is generated by some Poisson distribution, then the sum of these counts will also be generated by a Poisson distribution. For drunk driving this means that an individual does not always have to be drunk when he is driving, but it suffices that this happens at least three times in the period of data collection. Similarly, someone who possesses a gun illegally does not need to have it with him all the time, a small number of times suffices to consider his count to be generated by a Poisson distribution.

We note that the Poisson assumption is only valid if a change in the individual Poisson parameter is unrelated to any prior (non)apprehensions. This follows from the independence of subsequent trials in a binomial distribution. For example, if someone having a gun illegally is apprehended, and subsequently buys a new gun but takes it with him less often, or does not buy a gun at all, this change in capture probability is a violation of the Poisson assumption. Similarly, if someone is apprehended with a gun and the police keeps him under closer inspection after this apprehension, his capture probability is higher as a result of this apprehension and this is a violation of the Poisson assumption. Or, if someone is driving under the influence of alcohol, is not caught and starts doing this with a higher frequency, his probability increases as a function of not being apprehended and this is a violation of the Poisson assumption. In the biostatistical literature these violations are known as positive contagion (when the probability increases) or negative contagion (when the probability decreases).

Closely related to this contagion problem is the problem of an open or closed population. A population is closed if the number of offenders is constant over the period of data collection, and a population is open if offenders may enter or leave the population during this period. Given the above discussion it is clear that the population may be open, as long as entering or leaving the population is not related to a (non)apprehension. An example of leaving the population that is related to an apprehension is detention,

that can be seen as an extreme case of negative contagion.

Thus far we have only discussed the Poisson assumption for an individual count. The second assumption is due to the Poisson regression model we will employ, where the logarithm of the Poisson parameters is a linear function of covariates. In the regression model the Poisson parameters are still assumed to be homogeneous for individuals with identical values on the covariates, but they are allowed to be heterogeneous for individuals with different values. Since the differences in Poisson parameters are assumed to be completely determined by the observed covariates this is referred to as 'observed heterogeneity'. So, even if the count of every individual is Poisson distributed, the assumption of Poisson regression is violated if, additional to observed heterogeneity, there are differences in the Poisson parameters that cannot be explained by the observed covariates. This additional heterogeneity is called 'unobserved heterogeneity'.

In the Poisson regression model, that has only observed heterogeneity, the conditional mean (i.e. the mean conditional on the covariates) is equal to the conditional variance, whereas if there is unobserved heterogeneity, the conditional variance is larger than the conditional mean. This last phenomenon is referred to as overdispersion.

In conclusion, the most important ways in which the Poisson assumption can be violated in criminological applications are contagion and overdispersion.

For some offenses the contagion problem will probably be a larger than for other offenses, and we think that one way to find out whether this plays a role is by studying the behavior of both offenders as well as police officers. When no additional information is available on their behavior, it seems best to interpret the results cautiously. Overdispersion can be assessed in the data as a result of the analysis, and we will return to this and its interpretation in section 2.2.

We start with an introduction to the data we use for the analysis, namely police registrations. Next we discuss zero-truncated Poisson regression. This is followed by the two examples. We end with a critical evaluation of our method.

2 Method

2.1 Police registrations

Since the early eighties the Dutch police registers all violations against more than 70 different criminal laws, provided that the offender is known. From this registration system files over the period 1996 to 2001 for violations against the laws regarding drunk driving and illegal weapon ownership were made available to us. Regarding these files two problems existed. Firstly, within the police it is well-known that the quality of administration in the system varies considerably amongst the 25 Dutch police regions. For that reason we selected the five police regions that were reported to maintain the highest quality of administration for analyses. These regions are Rotterdam Rijnmond, Mid-Gelderland, Mid-Holland, South-Holland-South and Mid-West-Brabant. It is unclear whether these five police regions are representative for the 25 Dutch police regions. Secondly, due to the registration behavior of police officers, who use the registration system to support their job of finding suspects (and not to do statistical analyses), it may occur that the registration files have double entries for the same apprehension. Since double entries result in distorted frequency distributions, population estimates based on these files would be biased. Fortunately, at the time of this study the police had just completed a version of the registration files in which double entries were removed as much as possible.

The registration files contain all violations in the period from 1996 to 2001 for drunk driving or for illegal weapon ownership, respectively. We constructed dependent capture-recapture variables by counting, for every person separately and for fixed periods of time, the number of apprehensions for drunk driving and illegal weapon ownership, respectively. The dependent variables were computed over various fixed periods of time, ranging from one to five years. For illegal gun ownership we present an analysis of the data of the two-year period 1998 and 1999, and for drunk driving of the one-year period 2000.

As covariates we used the background variables age, gender, age of first offence and police region. From the number of apprehensions for the remaining criminal offences we constructed six covariates to measure a person's criminal history. These covariates

were constructed according to a standard categorization employed by Statistics Netherlands (see Huls et al., 2000), distinguishing between violations related to violence, hard-drugs, property, vandalism, traffic and violations of special "economic" laws (for example laws regulating working and environmental conditions). We computed the covariates by adding the number of apprehensions for the remaining offences into the corresponding covariate. The covariates were measured over a five-year period before the last year of measurement of the dependent variable. The five-year period was taken to ensure comparability of criminal histories, since minor offences are deleted from the files after five years without any further registration.

Because the distributions of the six covariates measuring criminal history were very skewed, we transformed them before entering them into the regression model. We used the transformation $\log(1+x)$, since preliminary analysis showed that this logarithmic transformation led to better predictions of the dependent variables.

Only the variable 'age of first offence' had missing values, about a half percent of the total number of observations. These values were imputed by the age of the subject, minus the mean difference between age and age of first offence in the sample.

2.2 Zero-truncated Poisson regression

We start with introducing the zero-truncated Poisson distribution, and then work out zero-truncated Poisson regression.

Consider the data in Table 1. We denote the number of individuals apprehended k times by f_k ($k = 1, \dots, K$). Let y_i be the number of times individual i ($i = 1, \dots, N_{obs}$) is apprehended ($y_i = 0, 1, \dots$). Due to the assumption of a constant probability, for each individual, the number of times he/she is apprehended over a fixed period of time follows a Poisson distribution:

$$P(y_i|\lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}, \quad (1)$$

which is determined by the Poisson parameter λ ($\lambda > 0$). Note the lack of subscript for λ , at this stage the individual Poisson parameters are still assumed to be homogeneous

in the population.

INSERT TABLE 1 ABOUT HERE

Since we are using registration data, we do not have an observed frequency of the individuals not apprehended, f_0 , and this frequency is to be estimated. For this purpose we assume that the observed frequencies f_k ($k = 1, \dots, K$) are generated by a Poisson distribution truncated at zero.

The zero-truncated Poisson distribution is defined by a probability function conditional on $y > 0$, that is

$$P(y_i|y_i > 0, \lambda) = \frac{P(y_i|\lambda)}{P(y_i > 0|\lambda)} = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!(1 - \exp(-\lambda))}, \quad y_i = 1, 2, \dots \quad (2)$$

with $p(y_i > 0|\lambda) = 1 - \exp(-\lambda)$, $i = 1, \dots, N$. Assume that we have an estimate $\hat{\lambda}$ for λ . This estimate can be used to obtain an estimate for the probability of an individual not to be observed, $\hat{p}_0 = \exp(-\hat{\lambda})$. The number of unobserved individuals (individuals who were not apprehended but had a positive probability to be apprehended), is denoted by \hat{f}_0 and can be calculated as

$$\hat{f}_0 = \frac{\hat{p}_0}{1 - \hat{p}_0} N_{obs}, \quad (3)$$

where N_{obs} is the number of observed individuals in the sample. An estimate of the population size, \hat{N} , is then obtained by

$$\hat{N} = \hat{f}_0 + N_{obs}. \quad (4)$$

Bustami et al. (2001) work out a Horvitz-Thompson point and interval estimate for \hat{N} . The point estimate for N is obtained as

$$\hat{N} = \sum_{i=1}^N \frac{I_i}{p(\lambda)}, \quad (5)$$

where $I_i=1$ if individual i is present in the sample and $I_i=0$, otherwise, and $p(\lambda) = 1 - \exp(-\lambda)$ is the probability of an individual to be present in the sample. This probability

can be estimated by replacing the parameter λ with its estimated value $\hat{\lambda}$ obtained from fitting the zero-truncated homogeneous Poisson model (2).

The variance of \hat{N} is given by

$$\text{var}(\hat{N}) = E[\text{var}(\hat{N} | I_i)] + \text{var}(E[\hat{N} | I_i]). \quad (6)$$

The first term in (6) reflects the uncertainty in the estimate λ given the observed individuals. An estimate is obtained by the well-known delta-method. The second term reflects the effect of the variability in the number of observed individuals. It is a well-known term in survey sampling. It depends on the $p(\lambda)$ -values of all individuals and can be estimated from the observed individuals by a Horvitz-Thompson-type estimator. For details on how the variance in (6) is estimated, we refer to Bustami et al. (2001). The first term in (6) will dominate the variance $\text{var}(\hat{N})$ when λ is large, because the probability of being observed is large, and the second term in (6) will dominate the variance $\text{var}(\hat{N})$ when λ is small, because then there is a large probability of not being observed.

We will now introduce zero-truncated Poisson regression, where the Poisson parameter will be a function of one or more covariates. Let Y_1, \dots, Y_N be a random sample from a zero-truncated Poisson distribution with parameter λ_i , $i = 1, \dots, N$. Let y_1, \dots, y_N be the realizations of Y_1, \dots, Y_N . In our applications y_i is the observed number of times individual i is found in the police registration system for different violations of a specific law, for example driving under the influence of alcohol. In the Poisson regression model (see, for example, Cameron and Trivedi, 1998; Greene, 1997; Long, 1997) the Poisson parameter λ_i of individual i is a function of a covariate vector $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ as

$$\log(\lambda_i) = \boldsymbol{\beta}^T \mathbf{x}_i, \quad (7)$$

where $\boldsymbol{\beta} = (\alpha, \beta_1, \dots, \beta_p)^T$. The Poisson parameter is now given with subscript i , since the covariates in the regression model introduce observed heterogeneity in the Poisson parameters. This heterogeneity is called observed since it is determined completely by the covariates, and individuals with identical covariate values will also have an identical

Poisson parameter.

This model (7) provides an estimator for the unknown parameter λ_i for the sampled individuals and thus for the probability to be present, $p(\lambda_i)$, $i = 1, \dots, N_{obs}$. Bustami et al. (2001) derive the Horvitz-Thompson estimator for the total number of individuals in a heterogeneous Poisson population which is defined by

$$\hat{N} = \sum_{i=1}^N \frac{I_i}{p(\mathbf{x}_i, \boldsymbol{\beta}^*)}, \quad (8)$$

where $I_i=1$ if present and 0, otherwise. The variance of \hat{N} can be split in two parts as in (6) and estimated in a similar way as the variance of (7). For details, we refer to Bustami et al. (2001).

For every model fitted we can calculate the deviance, also known as the likelihood ratio, being -2 times log(likelihood of the current model/likelihood of the saturated model). In a contingency table context it is possible to evaluate the fit of a model when the cells in the table are reasonably filled. However, when there are continuous covariates, the deviance cannot be used to assess the absolute fit of a model. Then, when we compare two nested models, differences between deviances can be used to assess the relative fit of two models: this difference is chi-square distributed and allows us to assess the significance of the explanatory variables that are in the more complex model but not in the less complex model.

The Poisson distribution is characterized by equality of mean and variance. An important reason for overdispersion (the variance exceeds the mean) is unobserved heterogeneity. This type of heterogeneity is called unobserved since the differences in the individual Poisson parameters cannot be explained by measured covariates. For the truncated Poisson regression model a Lagrange multiplier test on overdispersion was developed by Gurmu (1991). It compares the model fit of the Poisson model against alternative models with an extra dispersion parameter included, like the negative binomial model. The test statistic is chi-square distributed with one degree of freedom. We have used this test to assess and compare the degree of overdispersion in different models.

Another possibility to assess model fit is to compare the distribution of the count

variable observed in the data with the distribution estimated by the model. For this purpose we will use Pearson residuals, i.e. residuals computed as $(\text{observed} - \text{fitted})/\sqrt{\text{fitted}}$.

We now discuss an interesting property of the truncated Poisson regression model when used for the estimation of population sizes. This property is that a model with fewer covariates has a lower estimate of the population size N . This can be explained as follows. Assume that a truncated regression model with one dichotomous covariate is true, and that the model without this covariate is not true. Thus this latter model is misspecified. We now show that this misspecified model will have a lower estimate of the population size than the true model. As can be deduced from equation (5), the estimate of the population size is directly related to the estimate of the probability of the zero count p_0 , which is $p_0 = \exp(-\lambda)$. Let in the true model the two Poisson parameters be $(x + \delta)$ and $(x - \delta)$, and let in the misspecified model the Poisson parameter be x . Then Figure 1 illustrates that $\exp(E(-\lambda))$ is smaller than $E(\exp(-\lambda))$, which is also known as the Jensen inequality. Generalization to a situation with more covariates is straightforward. In practice this means that significant covariates should always be included in the model, if not, the model is misspecified and the estimated population size will be too low. Also, assume that all observed covariates are included in the final model, but there is still overdispersion, i.e. the observed heterogeneity is taken into account, but there is remaining unobserved heterogeneity. This means that the final model is misspecified, and the final estimate of the population size can only be considered as a lower bound.

INSERT FIGURE 1 ABOUT HERE

3 Examples

We present two examples. The first example deals with the illegal possession of weapons, and the second example with drunk driving.

In the introduction we saw that there are two main threats to the validity of the assumptions we are making: contagion and overdispersion. Overdispersion can be assessed

from the data, and this will be discussed below.

This leaves us with the problem of contagion, namely that the probability of a (non)apprehension may increase as the result of a (non)apprehension of the same offender. This probability can also increase as the result of behavior change of the offender (for example, it may increase when the offender finds out that the punishment following an arrest is relatively minor, or it may decrease when the punishment is evaluated as unpleasant) or as a result change in the behavior of the police. For both offenses studied in this section we have investigated this latter change in behavior, by studying arrest reports and interviewing police officers for a sample of offenders who were caught at least twice. In this research it became evident that contagion resulting from a change in behavior of the police, if present at all, cannot be large. For drunk driving this is because most apprehensions take place at random, because the police apprehends most offenders by stopping and controlling all passing drivers. For fire arms many offenders are apprehended as the result of their own behavior (for example, they show their gun at public places), guns are found when they commit other offences and the like, so here the apprehensions do not take place due to initiative of police officers.

In Table 1 we find the observed recapture frequencies for illegal possession of weapons, being $f_1 = 2,561$, $f_2 = 72$ and $f_3 = 5$. Note that the sample variance in the dependent variable is small, so that the information is limited. In Table 2 we report a model search procedure. The null model has a deviance of 738.1. We then add sex to the model but the deviance does not become significantly better (Δ deviance = .1, for 1 d.f.). Then we add two age variables, namely current age and age at first offence. This improves the fit significantly (Δ deviance is 10.2, for 2 d.f.). Adding six count variables for criminal history (violence, drugs, economic offenses, traffic violations, property crimes and vandalism) also improves the fit (Δ deviance is 29.3, for 6 d.f.) but adding four dummy variables for the five police regions does not improve the fit further, so the distribution of the counts of the regions do not seem to differ when we control for sex, age variables and criminal histories. An indication that the model fits well is given by the match between observed and expected frequencies: the Pearson residuals are relatively small.

INSERT TABLE 2 ABOUT HERE

Another indication of fit is given in Table 2 by the test for overdispersion, i.e. a violation of the Poisson assumption (column indicated with $\chi^2_{(1)}$, with p-value in column 5). This test indicates whether there is evidence for unobserved heterogeneity, given the inclusion of covariates that take observed heterogeneity into account. This test shows that, once we include the six criminal history covariates, there is no evidence in the data for further unobserved heterogeneity, so the Poisson assumption is not violated.

For the most comprehensive model $\hat{f}_0 = 60,084.0$, so for this model the total population size is estimated as $\hat{N} = 62,722$ (C.I. is 43,973-81,471). The population size estimates for the other models are found in Table 2. These estimates illustrate the typical result proven at the end of section 2.2, that the more covariates (i.e. the more observed heterogeneity) in the model, the larger the estimate for \hat{N} becomes. We also see that when covariates are added that are not significant, \hat{N} does not become substantially larger.

Table 3 reports the raw regression coefficients (b) for the most comprehensive model. We also included the exponents of the raw regression coefficients since these are more easily interpretable, they indicate the (multiplicative) factor with which the expected count in the dependent variable changes as a result of a unit change in the covariate, given that all other covariates are held constant. For a general discussion of interpretation of these coefficients, we refer to Long (1997, p.223-226, 241). Long distinguishes an interpretation in terms of expected counts and an interpretation in terms of the distribution of counts. We first give an interpretation in terms of expected counts, and then give an interpretation in terms of the truncated distribution (discussion of Table 4).

INSERT TABLE 3 ABOUT HERE

Given the model selection procedure, it is not surprising that sex is not significant. The sign of the regression coefficient for age of first offence is in the expected direction (the younger the first offence, the larger the count) but not significant. Age is significant (for each year that age increases, the expected count increases with a factor 1.03). The log-transformations of the counts for the violations histories provide some evidence for a positive relation of violence and property crimes with the expected number of times

someone is caught for illegal possession of weapons.

For illustrative purposes we also show the estimated probabilities to be apprehended at least once in the registration system in the left panel of Table 4. We do this for the categorical covariates only. All estimates are around .042, the overall probability to be apprehended at least once. Estimates for the different levels do not differ much, and this is not surprising given that Table 2 showed that both sex and region did not contribute significantly to the model.

INSERT TABLE 4 ABOUT HERE

The second example deals with drunk driving. Here the variance in the counts is larger than for the weapons example (see Table 1). Also the sample size is much larger. The observed and estimated frequencies are not as close as in the previous example, as is also indicated by the higher values of the residuals.

The model search reported in the bottom panel of Table 2 shows that now all sets of variables increase the fit of the model, in particular the criminal histories. Also, note that the estimated population sizes increase substantially with each model. What makes this example different from the weapons example is that in the final model the $\chi^2_{(1)}$ -test shows that there is still evidence for overdispersion (unobserved heterogeneity). This result is in line with the larger residuals in Table 1. Additional covariates are needed to take the unobserved heterogeneity into account, but unfortunately these are not available. However, since every additional covariate leads to an increase in \hat{N} , this means that we should interpret the estimate of 113,771 of the final model as a lower bound estimate of the true population size, so the true population size is estimated to be larger than 113,771. Compare the discussion at the end of section 2.2.

We now turn to the interpretation of the regression coefficients. For drunk driving sex is significant (for a female the expected count is .68 times that of a man), as well as the age of the first offence and the three criminal history covariates 'other traffic violations', 'property crimes' and 'vandalism'. So, the more often someone is caught for other traffic violations, the higher the expected count for drunk driving. The police regions also differ significantly, the number of estimated apprehensions in for example, Rotterdam Rijnmond is about 1.85 times higher than in Gelderland Center (1.85 is the

inverse of the exponent of the regression coefficient of Gelderland Center).

Table 4 now shows interesting differences between the estimated probabilities to be apprehended at least once, for example, males seem to be caught almost twice as much as females. Also, in Rotterdam Rijnmond the police seems to be more effective than in Gelderland Center and Holland Center (the probabilities to be apprehended at least once being 11 percent versus 6 and 6 percent).

4 Discussion

We have shown how police registration data can be used to estimate the size of criminal populations. Such estimates can be used for an evaluation of the effectiveness of the police forces, but they also give insight into differential arrest rates (Collins and Wilson, 1990) for different groups.

A first remark is that, even though the definition of the data is straightforward, the data are contaminated with errors. The reason is that the police does not collect these data to do statistical analyses, but instead uses this registration system to ameliorate the process of apprehending individuals. Therefore the registration is not always as careful as it should be. For example, extensive data cleaning was necessary to make the likelihood of incorrect double counts (i.e. the same apprehension appears twice in the system) as small as possible. Clearly, an incorrect double count incorrectly decreases f_k with 1 but increases f_{k+1} with 1, so that there appear to be more recaptures than there actually are. The result will be that the estimated zero-count, \hat{f}_0 , will be too low. Although the appearance of incorrect double counts has been given careful attention, it is possible that there are still a few in the data.

A second remark is about the meaning of our estimate \hat{N} . What does it stand for? The weapons example may make clear why this meaning is not as straightforward as one may think. Imagine someone having a weapon safely buried somewhere in his house. This individual has a zero probability to be apprehended, and we cannot generalize from the individuals who are actually apprehended to this type of individual. Basically, we can only generalize from the apprehended individuals to similar individuals who are not

apprehended *but who are in principle apprehensible*. Thus the estimate \hat{N} does not stand for the total number of illegal weapon owners, it stands for the *apprehensible* ones only. On the other hand, we may ask ourselves if non-apprehensible individuals really exist and, if so, whether these individuals belong to the population of interest? To have a zero capture probability the individual has to hide his weapon that well that it no longer poses any threat on society. We stress that the population estimate is still useful, as it is exactly the estimate of the number individuals that pose a threat on society and therefore is a good indication of the workload of the police (we may not reasonably expect the police to apprehend non-apprehensible individuals).

We have given an extensive discussion of the assumptions of the model in section 1.1. The main threat to the validity of the outcomes of the model are contagion and overdispersion. Contagion refers to the phenomenon of capture probabilities changing as the result of (non)apprehensions. We gave a few examples of why this might occur in the context of police registration data. In section 3 we discussed a change in behavior of police officers, and concluded that additional research that we conducted (not reported here in detail) leads to the conclusion that it is unlikely that this is a serious problem. However, we do not have any insight in the contagion resulting from a change of behavior of offenders. In general, positive contagion makes the observed counts too large, and this will lead to an estimated population size that is too small. Negative contagion will make the observed counts too small, and this will lead to an estimated population size that is too large. In section 3 we discussed the openness of the population in the context of negative contagion, and concluded that, if leaving the population is a result of a (non)apprehension, this will be a violation of the Poisson assumption.

Overdispersion may result from unobserved heterogeneity. Unobserved heterogeneity becomes evident as a result of the analysis, and it turned out that there is evidence for overdispersion in the drunk driving example, but not in the illegal gun ownership example. The result is that the estimated population size can only be interpreted as an estimate of the lower bound for the true population size.

A last area where violation of the model assumptions may occur is that of (in)dependence of observations (Collins and Wilson, 1990). A violation may occur when certain appre-

hensions involve more than one individual at the same time. This is something that can, in principle, be checked in the police registration system.

Concluding, we have presented a method that provides us with an estimate of the population size, but the assumptions underlying this estimate need careful consideration as they may affect this estimate seriously. We think that it is advisable to compare the estimates found with estimates from other sources, if available.

5 References

- Bunge, J. and M. Fitzpatrick (1993), Estimating the number of species: a review. *Journal of the American Statistical Association*, **88**, 364-373.
- Bustami, R., P.G.M. van der Heijden, H. van Houwelingen and G. Engbersen (2001), Point and interval estimation of the truncated Poisson regression model, in: B. Klein and L. Korsholm (eds.) *New trends in statistical modelling. Proceedings of the 16th International Workshop on Statistical Modelling*. Odense: University of Southern Denmark, pp. 87-94.
- Cameron, A.C., and Trivedi, P.K. (1998), *Regression analysis of count data. Econometric Society Monographs No. 30*, Cambridge, Cambridge University Press.
- Chao, A. (1988), Estimating animal abundance with capture frequency data, *Journal of Wildlife Management* **52**, 295-300.
- Charlier, C.V.L. (1905), Die zweite Form des Fehlergesetzes, *Arkiv fur Matematik, Astronomi och Fysik No 15, 2,,* 1-35.
- Collins, M.F., and R.M. Wilson (1990), Automobile theft: estimating the size of the criminal population, *Journal of Quantitative Criminology* **6**, 395-409.
- Efron, B. and R. Thisted (1976), Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* **63**, 435-447.
- Engbersen, G., R. Staring, J. van der Leun, J. de Boom, P. van der Heijden en M. Cruyff (2002), *Illegale vreemdelingen in Nederland. Omvang, overkomst, verblijf*

- en uitzetting*. Erasmus Universiteit: RISBO.
- Greene, M.A., and Stollmack, S. (1981), Estimating the number of criminals. In: Fox, J.A. (ed.) *Models in quantitative criminology*, New York: Academic, p. 1-24.
- Greene, W.H. (1997), *Econometric analysis (3rd ed.)*, New Jersey: Prentice hall.
- Gurmu, S. (1991), Test for detecting overdispersion in the positive Poisson regression model, *Journal of Business and Economic Statistics*, *9*, 215-222.
- Hoogteijling, E. (2002), *Raming van het aantal niet in de GBA geregistreerden.*, Voorburg, Statistics Netherlands.
- Huls, F.W.M., Schreuders, M.M., Ter Horst-van Breukelen, M.H. and Tulder, F.P. van (2000), *Criminaliteit en Rechtshandhaving 2000* **189**, 409-410.
- Johnson, N.L., S. Kotz and A.W. Kemp (1993), *Univariate discrete distributions*, (2nd ed.), New York: Wiley.
- Long, J.S. (1997), *Regression models for categorical and limited dependent variables*, Thousand Oaks, CA: Sage.
- Rossmo, D.K. and Routledge, R. (1990), Estimating the size of criminal populations. *Journal of quantitative criminology* **6**, 293-314.
- Sanathanan, L. (1977), Estimating the size of a truncated sample, *Journal of the American Statistical Association* **72**, 669-672.
- Seber, G.A.F. (1982), *The estimation of animal abundance, 2nd ed.*, London: Griffin and Co.
- Seber, G.A.F. (1986), A review of estimating animal abundance, *Biometrics* **42**, 267-292.
- Smit, F., van der Heijden, P.G.M. en van Gils, G (1993), Enkele weinig gebruikte methoden om de omvang van criminaliteit te schatten, *Tijdschrift voor Criminologie* **36**, 97-119.
- Smit, F., J. Toet and P.G.M. van der Heijden (1996), City Report Rotterdam: Estimating the number of opiate users in Rotterdam using statistical models for

- incomplete count data, in: European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), *Methodological pilot study of local level prevalence estimates*, Lisbon: EMCDDA, pp. 49-69.
- van der Heijden, P.G.M., Smit, F. en van Gils, G. (1993), *Schattingen van het aantal slachtofferloze delicten.*, Politia Nova **3**, Ministerie van Binnenlandse Zaken.
- van der Leun, J., G. Engbersen en Peter van der Heijden (1998), *Illegaliteit en criminaliteit: schattingen, aanhoudingen en witzettingen*. Rotterdam: Erasmusuniversiteit, vakgroep Sociologie.
- Wilson, R.M., and M.F. Collins (1992), Capture-recapture estimation with samples of size one using frequency data, *Biometrika* **79**, 543-553.
- Wittebrood K. and M. Junger (2002), Trends in violent crime: a comparison between police statistics and victimization surveys, *Social Indicators Research* **59**, 153-173
- Zelterman, D. (1988), Robust estimation in truncated discrete distributions with application to capture-recapture experiments, *Journal of Statistical Planning and Inference* **18**, 225-237.
- Zelterman, D. (2001), *Selected applications for categorical data. Advanced log-linear models using GENMOD, in Series: SAS books by users.*

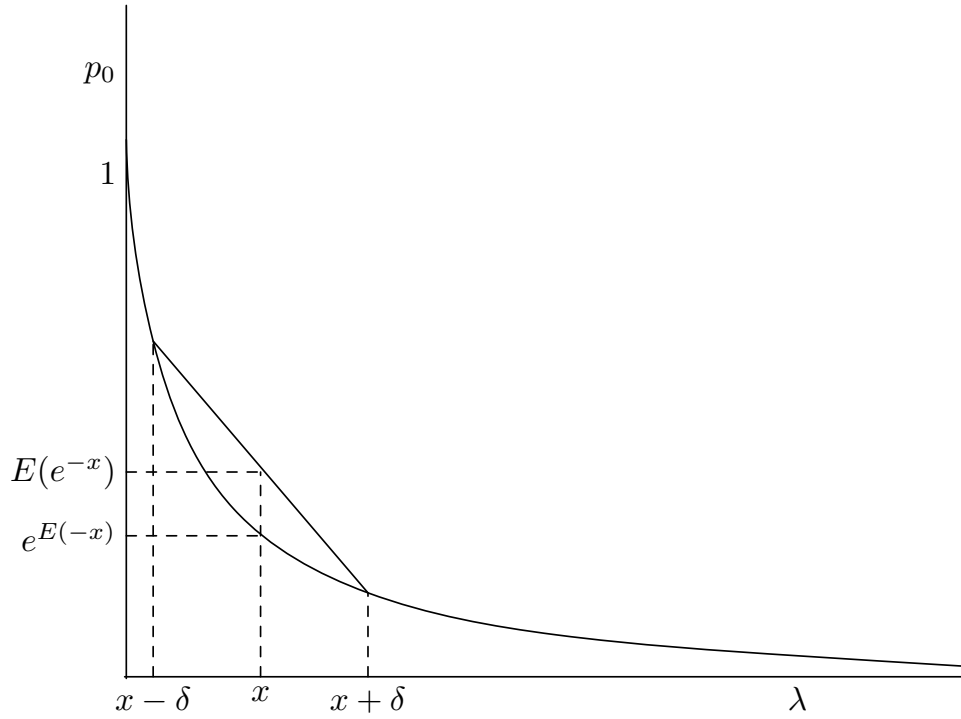


Figure 1: Illustration showing that model with a dichotomous covariate has higher estimates of \hat{N} than model without covariate

Table 1: Observed and estimated counts for illegal possession of weapons (left panel) and for drunk driving (right panel).

k	observed	estimated	residuals	observed	estimated	residuals
0	0	60,084.0	-	0	104,352.0	-
1	2,561	2,558.9	.04	8,877	8,847.2	.32
2	72	76.4	-.50	481	534.4	-2.31
3	5	2.6	1.48	52	34.0	3.08
4	-	-	-	8	2.9	2.98
5	-	-	-	1	.4	1.06

Table 2: Model comparisons. Illegal possession of weapons in top panel, drunk driving in bottom panel.

Model	deviance	Δ dev.	d.f.	$\chi^2_{(1)}$	p	\hat{N}	C.I.
Nul-model	738.1	–	–	6.4	.01	44,201	(34,828-53,574)
+ sex	738.1	.1	1	6.4	.01	44,239	(34,843-53,634)
+ age-var.	727.9	10.2	2	5.4	.02	48,244	(36,863-59,625)
+ violations	698.6	29.3	6	2.0	.16	60,782	(43,321-78,244)
+ regions	696.3	2.3	4	1.8	.18	62,722	(43,973-81,471)
Nul-model	4,652.4	–	–	115.4	.00	78,710	(72,738-84,682)
+ sex	4,634.8	17.5	1	110.3	.00	82,319	(74,828-89,809)
+ age-var.	4,515.2	19.7	2	73.1	.00	99,124	(87,609-110,639)
+ violations	4,389.3	125.8	6	30.7	.00	108,097	(95,581-120,612)
+ regions	4,358.4	30.9	4	24.6	.00	113,771	(99,857-127,685)

Deviance is the likelihood ratio. Δ dev. follows a chi-square distribution with degrees of freedom equal to the number of added covariates and can be used to assess the relative fit of two nested models. $\chi^2_{(1)}$ with its p-value refers to the Gurmur test for overdispersion, the test is chi-square distributed with one degree of freedom. \hat{N} gives the estimated population size of the model, with C.I. the 95 percent confidence interval.

Table 3: Raw regression coefficients, standard errors, t-tests and exponents of regression coefficients for illegal possession of weapons (left panel) and for drunk driving (right panel).

	b	s.e.	t	exp(b)	b	s.e.	t	exp(b)
intercept	-3.82	.40	-9.67	.02	-1.67	.15	-11.20	.19
male	.00	-	-	1.00	.00	-	-	1.00
female	.35	.46	.75	1.42	-.46	.23	-2.01	.63
age first offence	-.02	.02	-1.31	.98	-.03	.01	-5.58	.97
age	.03	.02	2.01	1.03	.01	.01	1.55	1.01
violence	.34	.18	1.86	1.41	.10	.11	.93	1.10
drugs	.21	.24	.90	1.24	-.13	.23	-.54	.88
economic	-.68	.84	-.80	.51	-.31	.57	-.54	.74
traffic	.25	.24	1.04	1.29	.66	.07	9.63	1.94
property	.53	.14	3.81	1.69	.20	.08	2.50	1.22
vandalism	.02	.23	.10	1.02	.24	.12	2.07	1.27
Rotterdam Rijnmond	.00	-	-	1.00	.00	-	-	1.00
Gelderland Center	-.16	.32	-.48	.85	-.61	.16	-3.77	.54
Holland Center	-.44	.45	-1.00	.64	-.56	.15	-3.87	.57
South-Holland-South	.06	.34	.17	1.06	-.35	.13	-2.61	.71
Center and West Brabant	-.34	.31	-1.10	.71	-.13	.10	-1.25	.88

Table 4: Observed N , estimated N and estimated probability of being observed for some subgroups; for illegal possession of weapons (left panel) and for drunk driving (right panel).

	obs. N	est. N	prob.	obs. N	est. N	prob.
male	2,496	60,030.3	.042	8,738	99,070.3	.088
female	142	2,691.9	.053	681	14,700.9	.046
Rotterdam Rijnmond	1,172	25,401.7	.046	3,960	37,481.2	.106
Gelderland Center	423	10,120.7	.042	962	15,699.9	.061
Holland Center	237	7,093.5	.033	1,260	21,151.4	.060
South-Holland-South	262	4,669.1	.056	1,278	18,081.9	.071
Center and West Brabant	544	15,437.3	.035	1,959	21,356.8	.092

Bijlage 1F: Schatting populatie illegalen

Hoofdstuk 3 uit:

Engbersen, G., Staring, R., Leun, J. van der., Boom, J. de., Heijden, P.G.M. van der. en Cruyff, M. (2002), *Illegale vreemdelingen in Nederland. Omvang, overkomst, verblijf en uitzetting*, Rotterdam, Risbo Contactresearch BV / Erasmus universiteit. (conceptversie)

Korte introductie bij hoofdstuk 3:

In hoofdstuk 3 wordt de omvang van de illegalenpopulatie geschat, de definities en operationalisaties van illegalen zijn in de voorgaande hoofdstukken behandeld. In deze introductie wordt kort aangegeven hoe het begrip illegaal geoperationaliseerd is.

In de data die in dit onderzoek is gebruikt is een illegaal gedefinieerd als een vreemdeling die op het moment van aanhouding niet in het bezit is van een geldige verblijfsvergunning. Uit de politieregistraties van deze personen is vervolgens op basis van twee criteria een selectie gemaakt. Het eerste selectie criterium betreft de maatregel die op de aanhouding van de de persoon volgt. De politieregistraties maken hierbij onderscheid tussen: (1) effectief uitzetten, (2) heenzenden en (3) onbekend/overig. Hiervan is alleen de categorie “heenzenden” geselecteerd, dit om zoveel mogelijk te voorkomen dat personen in de data vertegenwoordigd zijn die niet herpakt kunnen worden. Het tweede selectie criterium betreft herkomst van de aangehouden persoon. Personen uit Europa zijn eveneens uit de data verwijderd. Dit is gedaan omdat deze personen over het algemeen slechts voor een korte periode in Nederland zijn, bijvoorbeeld om drugs te kopen (West-Europeanen) of voor seizoensarbeid (Oost-Europeanen), en daardoor evenmin gedurende de gehele observatieperiode herpakt kunnen worden.

3. Een schatting van het aantal illegale migranten in Nederland

Dit hoofdstuk is geschreven door Maarten Cruijff en Peter van der Heijden m.m.v. Joanne van der Leun.

Inleiding

In het voorafgaande zijn de geselecteerde politiedata gebruikt om inzicht te geven in patronen van staandehouding en uitzetting door een viertal jaren heen. De cijfers geven tevens een indicatie van betrokkenheid bij bepaalde activiteiten in de criminele sfeer. Een beperking is dat deze cijfers alleen een beeld geven van degenen die – om welke reden dan ook – met de politie in aanraking zijn geweest. Een betrouwbaar steekproefkader om de staandehoudingspercentage aan te relateren ontbreekt echter omdat de gehele populatie illegalen onbekend is en slechts geschat kan worden.

Het schatten van aantallen illegalen is een omstreden zaak. De economen Jahn en Straubhaar (1998) schrijven hierover: "One of the most challenging and controversial issues in the discussions about illegal immigrants is the estimation of their size." Ondanks het controversiële karakter van dergelijke schattingen zijn met name in de Verenigde Staten diverse methoden ontwikkeld om de omvang van de illegale populatie te schatten (vgl. bijv. Fix and Passel 1995, Espenshade 1995, Delaunay and Tapinos 1998a en 1998b). De keuze voor een schattingsmethode hangt sterk af van de in de nationale context beschikbare data in combinatie met de eisen die een bepaalde schattingsmethode stelt aan die data. Aangezien de toegankelijkheid van registraties van illegale migranten in Nederland is toegenomen sinds 1996 en er zich in dezelfde tijdsspanne geen belangrijke wijzigingen hebben voorgedaan in de beschikbaarheid van alternatieve gegevensbronnen, zullen we ook nu gebruik maken van de capture-recapture of vangst-hervangst methode op basis van continue data, dat wil zeggen data die doorlopend (en niet steekproefsgewijs) worden verzameld (Van der Heijden 1994).

De vangst-hervangst methode kent haar oorsprong in de (populatie)biologie. De techniek wordt wel gebruikt de omvang van een bepaalde diersoort in een nader omschreven gebied te ramen aan de hand van twee metingen. Op basis van het aantal dieren dat alleen op het eerste tijdstip is aangetroffen, het aantal dat alleen op het tweede tijdstip is aangetroffen en het aantal dat beide keren is aangetroffen, valt het aantal dieren te schatten dat beide keren *niet* is aangetroffen. De achterliggende gedachte is dat de laagfrequente overtreders het meest zullen lijken op de groep die nooit is aangehouden. De som van deze getallen geeft een schatting van het totale aantal dieren in het geselecteerde gebied. In het geval van heterogene populaties kan de analyse apart worden uitgevoerd voor subgroepen. Ook zijn er manieren om te corrigeren voor vertekende invloeden. De vangst-hervangsttechniek is een systematische schattingsmethodiek die ook is toe te passen op onderzoek naar menselijke populaties (zie Smit et al. 1994, LaPorte 1994). Hierbij kan ofwel gewerkt worden met registraties van twee of meer onafhankelijke instanties, ofwel met continu verzamelde data, zoals antecedentengegevens in de criminologie. Smit et al. concluderen dat deze statistische technieken adequaat zijn toe te passen op het schatten van de omvang van de plegers van (vooral slachtofferloze) misdrijven (Smit et al. 1994, p. 116). Politieregistraties voldoen aan de eisen die deze methode stelt en voorzover de data niet aan die eisen voldoen, dient een model gemaakt te worden die schendingen van de algemene assumpties zoveel mogelijk ondervangt. Een dergelijk model is uitgewerkt bij de eerder gemaakte schatting door prof dr P. van der Heijden van de vakgroep Methoden en Statistiek van Universiteit Utrecht.

In het navolgende werken we de schatting uit. Eerst wordt echter aan de hand van een eenvoudig voorbeeld uit de biologie - het tellen van het aantal herten in een bos - de methode stap voor stap uitgewerkt (3.2). Vervolgens wordt de stap gemaakt naar het schatten van illegale vreemdelingen. Daarbij worden de aannamen van de methode geconfronteerd met de kwaliteit van de verzamelde data en de kenmerken van de illegale populatie (3.3). Hieruit vloeit een drietal problemen voort. Voor twee daarvan - de heterogeniteit en openheid van de populatie - wordt in paragraaf 3.4 een oplossing geboden. Van het derde bespreken we de consequenties. In paragraaf 3.5 worden de aanpassingen van het model besproken. De resultaten worden in 3.6 gepresenteerd met behulp van een model. Dit model levert verschillende schattingen op, afhankelijk van de gekozen periode. Beargumenteerd wordt dat de schattingen voor de periode van steeds één jaar het meest valide zijn, omdat daarin het meest wordt tegemoet gekomen aan de assumpties van de methode.

Wanneer we deze jaarschattingen over 1997, 1999 en 2000 middelen, komen we op een schatting van ruwweg 78.000 illegalen voor heel Nederland op jaarbasis. In tegenstelling tot in het eerdere onderzoek (van der Leun et al. 1998) zijn we nu ook in staat een betrouwbaarheidsinterval aan te geven. Wanneer we ook hierbij het gemiddelde nemen over de beste datasets (1997, 1999 en 2000), komen we uit op tussen de 65.000 en de 91.000 illegalen exclusief Oost- en West- Europeanen. Voor de laatste groepen kunnen we slechts een zeer ruwe raming maken waarmee we uitkomen op gemiddeld rond de 60.000 tezamen. De betrouwbaarheid van deze berekening is echter aanzienlijk minder dan van die bij de andere groepen. Wanneer we daar toch vanuit gaan zou het totaal op rond de 138.000 uitkomen voor heel Nederland.

Een voorbeeld van de vangst-hervangst methode

Vangst-hervangst methoden worden vaak gebruikt in de biologie voor het bepalen van de omvang van een onbekende populatie. Men voert bijvoorbeeld gedurende een bepaalde periode observaties uit in een bos en telt het aantal herten dat men waarneemt. De gegevens zien er na een jaar als volgt uit:

aantal keer gepakt	1	2	3	4	5	6	...
aantal dieren	f1	f2	f3	f4	f5	f6	

waarbij f1, f2, f3, ... frequenties zijn. Na een tijd observeren heeft men f1 = 100 herten 1 keer geobserveerd, f2 = 20 herten 2 keer geobserveerd, f3 = 7 herten 3 keer geobserveerd, enzovoort. Het aantal herten dat men heeft gezien is, is de som van deze frequenties.

$$f(\text{gezien}) = f1+f2+f3+f4+f5+f6+\dots \quad (1)$$

Het totaal aantal herten in het bos dat men niet heeft gezien is onbekend.

$$f(\text{niet gezien}) = f0 \quad (2)$$

Men weet niet hoe groot f_0 is en het probleem is dan ook hoe f_0 te bepalen, zodat men een schatting kan maken van het totale aantal herten.

$$f(\text{totaal}) = f(\text{gezien}) + f(\text{niet gezien}). \quad (3)$$

We kunnen $f(\text{gezien})$ eenvoudig uitrekenen door frequenties op te tellen, en we dienen een schatting te maken van het aantal herten dat niet is geobserveerd, d.w.z. van $f(\text{niet gezien})$. Voor het maken van een schatting van $f(\text{niet gezien})$ d.w.z. van f_0 , wordt vaak gebruik gemaakt van een Poisson-verdeling. Daarbij gaan we uit van een aantal aannamen.

De Poisson-verdeling

Een Poisson-verdeling is een verdeling voor een telvariabele. Het proces dat bij toeval bepaalt hoe vaak wij dat specifieke hert zien noemt men in de statistiek een Poisson-proces. Wij bekijken nu de kans om één enkel hert te zien. Stel dat in een bepaalde tijdsperiode de kans om een hert te zien constant is, dan is de kans om dat ene specifieke hert in die tijdsperiode nul keer, één keer, twee keer, drie keer, enzovoort te zien, verdeeld als een Poisson-variabele. De kansverdeling voor een Poisson-proces is:

$$p(y; m) = \frac{e^{-m} m^y}{y!} \quad (4)$$

Hierbij noemt men m de zogenaamde Poisson-parameter³, y is het aantal keer dat het hert is geobserveerd in een bepaalde tijdsperiode (bijvoorbeeld een jaar), en e staat voor de exponentiële functie (e is ongeveer 2.81). In het voorbeeld betekent dit, dat als de Poisson-parameter $m=1$, en wij vullen voor y de waarden 0 (keer geobserveerd) tot 6 (keer geobserveerd) in, dat we het onderstaande vinden:

aantal keer gepakt	0	1	2	3	4	5	6	...
kans	.368	.368	.184	.061	.015	.003	.001	

Hieruit blijkt dat als een hert als Poisson-parameter $m=1$ heeft, dat zijn kans 1 maal gepakt te worden .368 is, zijn kans 2 maal gepakt te worden .184, zijn kans 3 maal gepakt te worden .061, enzovoort. Als de Poisson-parameter 1 is, blijkt dat het onwaarschijnlijk is dit specifieke hert 3 of meer keren te zien. In het tweede voorbeeld wordt aangenomen dat $m=1.4$. Dan geldt het onderstaande:

aantal keer gepakt	0	1	2	3	4	5	6	...
kans	.247	.345	.242	.113	.039	.011	.003	

³ De Poisson-parameter wordt geschat met behulp van speciale computerprogramma's. De cijferreeksen in dit hoofdstuk zijn kansen die optellen tot 1.

We zien dus dat het bij een hogere Poisson-parameter waarschijnlijker wordt om een hert vaker waar te nemen. Nu is de kans om zo'n hert 3 of meer keer waar te nemen nog steeds klein, maar een stuk groter dan bij een Poisson-parameter van 1. In een laatste voorbeeld wordt verondersteld dat de Poisson-parameter $m = .75$ is. Dan geldt het volgende:

aantal keer gepakt	0	1	2	3	4	5	6	...
kans	.472	.354	.133	.033	.006	.001	.000	

Uit de voorbeelden blijkt dat bij een kleinere Poisson-parameter de kans om een hert waar te nemen steeds kleiner wordt. Hierbij moet worden opgemerkt dat:

- de bovenstaande kansverdeling in principe steeds voor één hert geldt;
- het toeval bepaalt hoe vaak wij dat specifieke hert zien.

In de praktijk zien wij een bepaald hert bijvoorbeeld twee keer. Als wij alleen deze informatie hebben, kunnen wij niet afleiden wat zijn Poisson-parameter m is. De vraag rijst nu hoe we met de Poisson-verdeling komen tot de schatting van het aantal herten dat wij nul keer hebben gezien (f_0). In het navolgende wordt uitgelegd hoe dit doorgaans wordt gedaan.

Van Poisson-verdeling naar f_0

Bij het schatten van f_0 wordt vaak aangenomen dat alle herten dezelfde Poisson-parameter hebben. Dan zijn de volgende gegevens nodig om te kunnen rekenen:

aantal keer gepakt	1	2	3	4	5	6	...
aantal dieren	f_1	f_2	f_3	f_4	f_5	f_6	

Hierbij is mogelijk de Poisson-parameter m te schatten die zo goed mogelijk past bij de geobserveerde frequenties f_1, f_2, f_3, \dots etc. Als de schatting gemaakt is, wordt deze ingevoerd in formule (4) voor $y=0$. Enkele simpele berekeningen leiden dan tot een schatting van f_0 , zodat het schattingsprobleem is opgelost. We weten dan immers (2) en kunnen (3) berekenen.

Er zijn ook speciale schatters die vooral geschikt zijn voor die gevallen waarin met name de frequenties f_1 en f_2 erg hoog uitvallen. Deze schatters -die vrijwel alleen gebruik maken van f_1 en f_2 - zijn door van Gils en van der Heijden eerder gebruikt in een onderzoek voor het Ministerie van Binnenlandse Zaken, namelijk een studie naar het aantal vuurwapens in Nederland (Van Gils en Van der Heijden 1996).⁴ Inmiddels zijn enkele nieuwe manieren van schatten ontwikkeld die sterke voordelen hebben boven deze schatters.

⁴ De speciale schatters waarvan wordt gesproken zijn momentschatters in plaats van maximum likelihoodschatters.

Centrale aannamen en problemen bij het schatten van illegale vreemdelingen

Om een zinvolle schatting te berekenen moeten we bepaalde aannamen maken. Hiertoe zullen we ook de stap zetten van het voorbeeld van het aantal herten dat zich in een bos bevindt naar de door ons uitgewerkte schatting van het aantal illegale vreemdelingen in de vier grote steden. Daartoe zetten we eerst de aannamen van de door ons gebruikte schattingsmethode op een rij.

Aanname 1: alle geobserveerde wezens (d.w.z. herten of illegale vreemdelingen) hebben dezelfde Poisson-parameter.

Wanneer alle geobserveerde wezens dezelfde Poisson-parameter hebben, spreekt men in de statistische literatuur van een homogene populatie. We weten te weinig van herten om te kunnen beoordelen of deze assumptie realistisch is. In veel toepassingen is deze assumptie waarschijnlijk echter geschonden. Wel kan aangetoond worden dat de schatting van f_0 in dit geval een ondergrens oplevert. Dat wil zeggen, als bijvoorbeeld f_0 geschat wordt op 1.000, dan zijn er *minimaal* 1.000 herten niet geobserveerd. Wanneer er sprake is van een heterogene populatie, kan deze worden opgesplitst in homogene groepen. De schatting wordt voor elke homogene groep apart berekend en de som van de uitkomsten leidt tot de totale schatting.

Het mag duidelijk zijn dat in het geval van illegale vreemdelingen de assumptie van homogeniteit zeker geschonden is. Sommige illegalen zijn nu eenmaal zichtbaarder dan anderen, en lopen daardoor een grotere kans om gepakt te worden. Ook het beleid van de politie is doelbewust selectief en kan bovendien van stad tot stad verschillen (Engbersen et al. 1999). Daarnaast kunnen specifieke steden op specifieke soorten illegalen ook een verschillende aantrekkingskracht uitoefenen. Aanvankelijk was daarom de bedoeling de illegale populatie op te splitsen in homogene groepen. In de volgende paragrafen zullen wij echter een elegantere oplossing voor het probleem met de eerste aanname bespreken.

Aanname 2: de Poisson-parameter verandert niet door de tijd heen. Dat wil zeggen er is sprake van tijdonafhankelijkheid van de observaties.

In hoeverre de Poisson-parameter van herten door de tijd constant blijft weten we niet, en ook voor illegalen is dit nogal duister. Het kan zijn dat illegale vreemdelingen zich 'schuwer' gaan gedragen als ze zijn aangehouden (zodat de pakkans kleiner wordt) maar het kan ook zijn dat ze zich 'roekelozet' gaan opstellen omdat de eerdere staandehouding niet leidde tot een uitzetting (in dit geval wordt de pakkans groter). Uit onderzoek onder illegale vreemdelingen komen beide reacties naar voren.⁵ Als de pakkans kleiner wordt, kan worden aangetoond dat de schatting van f_0 te hoog wordt, omdat er te weinig 'herpakkingen' zijn. Als de pakkans groter wordt, wordt de schatting van f_0 juist te laag, omdat er teveel 'herpakkingen' komen. Eerdere simulatiestudies voor het Ministerie van Binnenlandse Zaken van Van Gils en van der Heijden (1996) hebben dit uitgewezen. Aangezien er geen oplossing voor dit probleem van de tijdfafhankelijkheid bekend is, gaan berekeningen echter toch uit van een constante pakkans.

⁵ Burgers en Engbersen 1999.

Aanname 3: de populatiegrootte is constant: het aantal individuen is constant gedurende één jaar.

Voor herten is de aanname dat de populatie constant is waarschijnlijk niet erg problematisch. Voor illegale vreemdelingen is dit echter een groter probleem. Illegale vreemdelingen die worden aangehouden, worden in principe uit Nederland verwijderd. Sommige illegalen worden effectief uitgezet naar het land van herkomst, met als gevolg dat zij een kleinere kans hebben terug te komen. Als zij niet terugkomen, kunnen zij niet opnieuw aangehouden worden. Voor dit probleem hebben wij een op inhoudelijke gronden gefundeerde oplossing in het model verdisconteerd, die wij hieronder zullen bespreken.

Een tweede probleem met deze aanname is dat illegalen uit vrije wil vertrekken, of lopende het jaar in Nederland arriveren. Ook zijn er categorieën illegale vreemdelingen die slechts korte tijd in Nederland verblijven, zoals drugstoeristen of 'pendelaars'. Dit probleem is enigszins onder controle voor de jaar-schattingen en vormt waarschijnlijk een groter probleem voor de langere periode. Los daarvan zijn er twee categorieën illegale migranten waarvoor men op grond van ander onderzoek mag aannemen dat deze assumptie wordt geschonden. In de eerste plaats gaat het daarbij om de West-Europeanen. Dit zijn veelal toeristen die korte perioden - meestal in de weekeinden - in de Nederlandse steden verblijven, vaak met het oogmerk om drugs te kopen. Het is niet reëel aan te nemen dat zij het hele jaar aan te houden zijn. Bovendien zou het vreemd zijn om een aantal niet-aangehouden West-Europese illegalen 'bij te schatten' aangezien West-Europeanen die niet zijn aangehouden legaal in het land verblijven. Om deze reden worden de West-Europeanen buiten het schattingsmodel gehouden. In de tweede plaats levert de assumptie problemen op in het geval van Oost-Europese illegalen. Veel Oost-Europese arbeidskrachten verrichten seizoenswerk en verblijven daarvoor enkele maanden in Nederland (Odé 1995, p.101). Ook Oost-Europese illegale migranten 'pendelen' waarschijnlijk vaker dan veel andere groepen (Kloosterboer et al. 2002). Het is echter dan ook waarschijnlijk onrealistisch te veronderstellen dat Oost-Europese illegalen het hele jaar zijn aan te houden. Op basis van deze onzekerheden met de aanname van de constante populatie wordt ook de categorie aangehouden illegalen uit Oost Europa niet in de schatting meegewogen.

Oplossingen voor de schendingen van assumpties

Omdat we in de voorgaande paragraaf hebben gezien dat de drie centrale assumpties van de gebruikte methode ten dele worden geschonden, hebben we daarvoor oplossingen gezocht, die hieronder worden uitgelegd.

a) Een oplossing voor de heterogeniteit van Poisson-parameters: afgeknotte Poisson-regressie.

Voor het probleem van de heterogeniteit van de populatie is het schattingsmodel aan te passen door de Poisson-parameter te laten afhangen van een aantal verklarende variabelen. Met andere woorden: we beschouwen het aantal herpakkingen als een functie van een aantal achtergrondkenmerken waar we iets over weten. Dit doen we met behulp van Poisson-regressie. Poisson-regressie is een variant op gewone multiple regressie waarbij er sprake is van een afhankelijke variabele y die voorspeld wordt uit een aantal verklarende variabelen. In Poisson-regressie is elk individu een aantal keer herpakt. Voor individu i noemen we dit aantal $y(i)$. Verder heeft elk individu i een Poisson-parameter $m(i)$. We hebben eerder gezien dat de Poisson-parameter de kans bepaalt dat

iemand veel of weinig gepakt wordt. Bij Poisson-regressie is deze Poisson-parameter $m(i)$ een functie van een aantal verklarende variabelen. Dan ziet het gehele model voor Poisson-regressie er als volgt uit:

$$p(y(i); m(i)) = \frac{e^{-m(i)} m(i)^{y(i)}}{y(i)!} \quad (5)$$

waarbij:

b_0 i.p.v. c in onderstaande formule

$$\log m(i) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki} \quad (6)$$

De i -tjes geven hier aan dat er sprake is van individuele waarden voor elke illegale vreemdeling afzonderlijk. Elke illegale vreemdeling heeft zijn eigen Poisson-parameter $m(i)$, zijn eigen aantal keren dat hij is aangehouden $y(i)$, en de Poisson-parameter van elk individu i is een functie van een aantal verklarende variabelen $x(i)$. Deze verklarende variabelen kunnen voor elk individu invloed uitoefenen op de hoogte van zijn Poisson-parameter $m(i)$, d.w.z. op zijn kans dat hij herpakt wordt, d.w.z. op de hoogte van $y(i)$.

Op deze wijze incorporeren wij de heterogeniteit van pakkansen in het model: de pakkansen mogen over individuen heen verschillen, maar moeten een functie zijn van verklarende variabelen. Met andere woorden: als we van illegale vreemdelingen de reden van staandehouding, de leeftijd, het geslacht en het land van herkomst weten, kunnen we ze op een aantal kenmerken indelen in homogene groepen. Hierbij gebruiken we de volgende verklarende variabelen waar we beschikking over hebben en die hoogstwaarschijnlijk gerelateerd zijn aan de pakkans, namelijk (1) de reden van staandehouding, (2) de leeftijd, (3) het geslacht en (4) het land van herkomst.

b) Het schatten van de niet-geobserveerde illegalen

Het bovenstaande model houdt er geen rekening mee dat de frequentie van f_0 niet geobserveerd kan worden. Dit probleem wordt echter opgelost door gebruik te maken van een computerprogramma voor Afgeknotte Poisson-regressie. In dit computerprogramma wordt (6) geschat terwijl er rekening mee wordt gehouden dat personen met f_0 niet geobserveerd kunnen worden.

Dit computerprogramma levert de schattingen voor de b -waarden en de constante uit formule (6). Met deze schattingen is voor elke illegale vreemdeling i zijn Poisson-parameter $m(i)$ te berekenen. Als dan voor elk individu i zijn Poisson-parameter $m(i)$ bekend is, dan is met behulp van (5) de kans te berekenen dat een dergelijk iemand niet wordt geobserveerd (d.w.z. de kans dat $y = 0$ voor persoon i , die we noteren als $P(y=0|i)$) en de kans dat persoon i wel wordt geobserveerd (d.w.z. de kans dat $y > 0$, $P(y>0|i)$). Met deze kansen is voor elke afzonderlijke illegale vreemdeling i het aantal 'vergelijkbare illegalen' te schatten dat we niet hebben geobserveerd. Dit doen we met de volgende formule:

$$f(0_i) = \frac{p(y = 0_i)}{p(y > 0_i)} \quad (7)$$

Formule (7) is met twee voorbeelden intuïtief aannemelijk te maken. Stel dat voor een illegaal i geldt dat wij op basis van zijn Poisson-parameter $m(i)$ vaststellen dat $P(y>0|i) = .5$. Dat wil zeggen dat de kans om deze illegaal daadwerkelijk te zien slechts 50% was. Toch hebben we hem gezien. Wij verwachten daarom dat er voor deze ene illegale vreemdeling nog een vergelijkbare persoon rondloopt die wij niet hebben geobserveerd. Dit volgt uit (7), want als $P(y>0|i) = .5$, dan geldt dat $P(y=0|i) = .5$, zodat $f(0|i) = 1$. Met “vergelijkbare illegaal” bedoelen we dan een illegaal die dezelfde verklarende variabelen $x(i)$ heeft als de geobserveerde illegaal. Een tweede voorbeeld: stel dat de kans dat wij een bepaalde illegale vreemdeling aantreffen slechts .25 was. Toch hebben wij deze persoon aangetroffen. De verwachting is dan dat er in totaal vier van dergelijke illegale vreemdelingen zijn (25% van 4 geeft de ene illegale migrant die is geobserveerd), en $f(0|i) = .75/.25 = 3$. Op deze manier schatten we de niet-aangehouden illegale vreemdelingen (f_0).

c) Een oplossing voor de open populatie: parameters slechts schatten op basis van niet-effectief uitgezette

De derde assumptie was dat het aantal illegalen constant bleef gedurende het steekproefjaar. Deze assumptie wordt onrealistisch geacht om verschillende redenen. Een belangrijke reden waarover wij informatie hebben, is dat illegalen in een deel van de gevallen effectief worden uitgezet naar het land van herkomst (zie hoofdstuk 2). Het is aannemelijk dat velen van hen niet direct terugkeren, hetgeen ertoe leidt dat zij gedurende een bepaalde periode niet opnieuw aangehouden (herpakt) kunnen worden. Voor de berekening heeft dit tot gevolg dat de waarden $y(i)$ voor deze illegalen te laag kunnen zijn. Met andere woorden, voor effectief uitgezette illegalen vertrouwen wij de $y(i)$ -waarden niet.

Wat wij wel vertrouwen zijn hun $x(i)$ -waarden. Daarom doen we het volgende: we schatten de b -waarden op basis van die illegalen waarvoor we zowel de $x(i)$ -waarden als de $y(i)$ -waarde vertrouwen (d.w.z. op basis van de niet-effectief uitgezette illegalen), en stellen een vergelijking (6) op. Vervolgens berekenen we voor de illegalen waarvan we de $y(i)$ -waarde niet vertrouwen (d.w.z. de effectief uitgezette illegalen) met (6) hun Poisson-parameter $m(i)$. Wij hebben dan voor elke illegale vreemdeling i een Poisson-parameter $m(i)$, en deze gebruiken we om kansen uit te rekenen die nodig zijn om berekeningen met (7) uit te voeren.

Met deze aanpassingen van het rekenmodel komen wij voor zover mogelijk tegemoet aan de bezwaren van de geschonden assumptie. Het moge duidelijk zijn dat niettemin een dergelijke schatting altijd met enige voorzichtigheid moet worden beschouwd.

Aanpassingen van het schattingsmodel

Op basis van de bovenstaande overwegingen is gekozen voor een model waarbij we rekenen met gegevens over de illegale vreemdelingen die in de jaren 1997, 1998, 1999 en 2000 zijn aangehouden. De schattingen die we presenteren behelzen de aangehouden illegale vreemdelingen (exclusief personen uit West- en Oost- Europa) en de bijgeschatte 'vergelijkbare' illegalen. Omdat we nu beschikking hebben over gegevens over meerdere jaren, hebben we schatting uitgetoetst voor steeds één jaar, voor twee jaar tezamen, voor drie jaar tezamen en voor

de gehele periode 1997-2000. Op die manier zijn tien datasets samengesteld die zijn geselecteerd op personen die niet-effectief zijn uitgezet (zonder West- en Oost-Europeanen)⁶.

Een tweede selectieprobleem betrof de uitschieters op het aantal pakkingen die in een aantal datasets aanwezig zijn. Er komen in de VAS-bestanden illegale migranten voor die vele malen in een jaar worden aangehouden. Het effect van deze uitschieters op de populatieschattingen en de betrouwbaarheidsintervallen is onderzocht door in alle datasets pakkingen van en 8 hoger te hercoderen tot 7. Een vergelijking van de resultaten van deze analyses met die van de niet-hergecodeerde data gaf slechts zeer geringe verschillen te zien. Dit geeft aan dat de populatie- en betrouwbaarheidsintervalschattingen robuust zijn tegen uitschieters met extreem veel pakkingen. Dit is te verklaren uit het feit dat personen met veel pakkingen ook een grote poissonparameter hebben en daardoor een kleine nulschatting. De populatieomvang van de overige groepen – de effectief uitgezette illegalen en de groep waarvoor het vervolg van de staandehouding onbekend is – is geschat op basis van de met formule (6) geschatte b-waarden van de niet-effectief uitgezette groep.

Voor de groep West- en Oost-Europeanen is deze benadering niet mogelijk omdat de voor deze groep een deel van de x-en (het land van herkomst) uit formule (6) niet overeenkomt met die van de niet-effectief uitgezette personen. Een (grove) schatting van de populatieomvang kan worden verkregen door op deze groep de gemiddelde verhouding tussen het aantal geobserveerde en het totaal aantal geschatte niet-effectief uitgezette personen toe te passen.

In de tabellen 1a en 1b zijn de verdelingen van respectievelijk de (afgeronde) geschatte en geobserveerde frequenties weergegeven. In het oog springend zijn de verschillen in de lengte van de staarten van de verdelingen; enkele personen zijn in werkelijkheid veel vaker gepakt dan op basis van de schattingen waarschijnlijk geacht wordt. In het algemeen geldt verder dat in verhouding tot de observaties de schattingen van:

1. de frequenties van de 1-pakkingen te laag zijn;
2. de frequenties van de 2-pakkingen te hoog zijn;
3. de frequenties van de hogere pakkingen (over het algemeen) te laag zijn.

Dit patroon lijkt aan te geven dat er sprake is van negatieve besmetting (bijvoorbeeld doordat men vermijdingsgedrag gaat vertonen na een eerste contact met de politie), schending van de assumptie van een gesloten populatie (doordat men geen heel jaar aanwezig is in Nederland) of door ongeobserveerde heterogeniteit (wanneer de illegale migranten verschillen op niet-gemeten kenmerken). Negatieve besmetting leidt tot een daling van de poissonparameter (pakkans) als gevolg van het feit dat de persoon is gepakt en daardoor tot een ondervetegenwoordiging van de meervoudige pakkingen. De schending van de assumptie van de gesloten populatie leidt ertoe dat er personen in de populatie zijn waarvoor de pakkans kleiner is dan de poissonparameter suggereert, omdat deze personen niet gedurende de volledige periode van observatie in de populatie aanwezig zijn geweest. Dit leidt eveneens tot een ondervetegenwoordiging van meervoudige pakkingen. De voornoemde schendingen van het Poissonmodel zouden een verklaring kunnen zijn voor de voornoemde discrepanties tussen geobserveerde en geschatte frequenties van de 1- en 2-pakkingen. Zij geven echter geen verklaring voor de aanwezigheid van de (extreem) hoge pakkingen in de geobserveerde data; als

⁶ Ten aanzien van de groep niet-effectief uitgezette met onbekende nationaliteit bestond aanvankelijk twijfel omtrent de representativiteit van deze groep voor de doelgroep, gezien de mogelijkheid dat deze groep (veel) West- en Oost-Europeanen zou kunnen

gevolg van negatieve besmetting en open populaties neemt de kans op meervoudige pakkingen immers af. Ongeobserveerde heterogeniteit refereert aan mogelijke niet gemeten verklarende variabelen (de x-en uit formule 6) die de discrepanties tussen geobserveerde en geschatte frequenties – waaronder de aanwezigheid van hoge pakkingen - zouden kunnen verklaren. Hierbij moet gedacht worden aan bepaalde – onbekende - persoons- of groepskenmerken die een sterk verband houden met de pakkans.

Tabel 1A: Geschatte frequenties

	periode									
	97	98	99	00	97-98	98-99	99-00	97-99	98-00	97-00
	est	est	est	est	est	est	est	est	est	est
1	2440	2318	2210	2313	4388	4241	4210	6201	6102	7986
2	193	170	142	160	498	438	430	797	775	1148
3	15	17	8	11	61	48	42	104	99	164
4	1	2	0	1	8	5	4	14	13	24
5	1	1	.	2	2	3

Tabel 1B: Geobserveerde frequenties

	periode									
	97	98	99	00	97-98	98-99	99-00	97-99	98-00	97-00
	obs	obs	obs	obs	obs	obs	obs	obs	obs	obs
1	2467	2360	2226	2339	4505	4346	4298	6400	6301	8289
2	158	120	115	121	359	305	297	554	518	777
3	13	15	17	18	55	57	67	99	112	153
4	6	5	2	2	20	10	11	29	31	53
5	3	3	0	1	6	7	7	19	13	24
6	1	0	1	2	5	1	4	7	5	13
7	0	0	.	1	1	1	0	0	2	4
8	1	0	.	.	1	1	1	0	1	1
9	.	1	.	.	0	0	1	1	1	3
10	.	0	.	.	1	0	.	1	1	1
11	.	1	.	.	2	3	.	1	2	1
12	.	0	.	.	0	0	.	1	0	0
13	.	0	.	.	0	0	.	1	1	1
14	.	0	.	.	0	1	.	2	2	2
15	.	1	.	.	2	1	.	2	1	2
16	0
17	0
18	0
19	0
20	1

Tabel 2: Totaal aantal gepakte personen (niet-effectief uitgezet) per periode

	97	98	99	00	97-98	98-99	99-00	97-99	98-00	97-00
aantal gepakte personen	2649	2506	2361	2484	4956	4733	4686	7117	6991	9325

bevatten. Aangezien de groep wel voldoet aan het criterium van niet effectieve uitzetting en zeer waarschijnlijk weinig Europeanen bevat, is zij wel meegenomen in de analyses.

Geschatte populatiegrootte van niet-effectief uitgezette illegalen

De geschatte populatiegrootte van de groep niet effectief- uitgezette illegalen is in tabel 3 weergegeven. Wat in de eerste plaats opvalt is periode-effect; naarmate de geobserveerde periode langer is neemt de grootte van de populatie toe. Een mogelijke verklaring hiervoor is de eerder genoemde schending van de assumptie van een gesloten populatie. Een zekere mate van doorstroming van illegalen leidt tot een geringere pakkans en dus tot te lage frequenties van de meervoudige pakkingen. Relatief lage frequenties op de meervoudige pakkingen impliceren een grote populatie en een schending van de assumptie van een gesloten populatie leidt daarom tot een overschatting van de omvang van de populatie. Dit effect is geringer naarmate de geobserveerde periode kleiner is.

De conclusie van deze exercitie is dat schattingen voor de éénjaarsperioden zijn - uitgaande van een niet gesloten populatie – het meest valide zijn.

Tabel 3: Geschatte populatie en betrouwbaarheidsinterval voor niet-effectief uitgezette illegalen.

	populatie	min 95% CI	max 95% CI
97	26026	19985	32067
98	41156	22715	59597
99	27308	15990	38626
00	28523	21053	35993
97-98	37330	31586	43074
98-99	35972	30962	40982
99-00	37378	29022	45734
97-99	43503	39027	47979
98-00	42119	38082	46156
97-00	49561	45711	53411

Een tweede opmerkelijkheid is de afname van het bereik van het betrouwbaarheidsinterval naarmate de periode langer is. Dit effect is voornamelijk het gevolg van het grotere aantal observaties in de langere perioden en het feit dat, naarmate een groter deel van de populatie is gepakt, de omvang van de populatie nauwkeuriger te schatten is.

Ten derde valt op dat schatting van de populatie in het jaar 1998 aanzienlijk hoger uitvalt dan die van de andere enkele jaren. De meest waarschijnlijke reden hiervoor is de frequentieverdeling van de vrouwen over het aantal pakkingen in dat jaar. In 1998 zijn in totaal 380 vrouwen gepakt, hiervan zijn 375 vrouwen éénmaal en slechts 5 vrouwen tweemaal gepakt. Deze verdeling leidt tot zeer hoge schattingen van het aantal niet geobserveerde vrouwen per gepakte vrouw, er komen zelfs schattingen van 170 voor. Ter illustratie: van de in totaal 464 vrouwen in 1997 zijn er 13 tweemaal gepakt en 1 driemaal, het percentage meervoudige pakkingen ligt hier dus hoger. Daardoor bedraagt de maximale schatting van het aantal niet-geobserveerde vrouwen per gepakte vrouw hier slechts 40 en is de schatting van het gemiddelde aantal niet geobserveerde vrouwen per geobserveerde vrouw ongeveer 4 maal lager dan in 98.

Het effect van deze verdeling op de populatieschatting kan ten dele worden geneutraliseerd door de verklarende variabele geslacht uit het model te verwijderen. Een analyse zonder deze variabele levert een populatieschatting op van 31.500 (t.o.v. 41.156) met een standaarddeviatie van 4238 (t.o.v. 9409). Deze resultaten zijn veel meer in

lijn met die van de overige éénjaarsperioden en lijken er op te duiden dat de verdeling pakkings van de vrouwen in 1998 als niet-representatief voor de normale situatie moet worden beschouwd.

De populatieschattingen van de effectief uitgezette illegalen en de groep waarvan de situatie na staandehouding onbekend is, zijn gebaseerd op de met formule (6) geschatte b-waarden in de niet-effectief uitgezette groep. Alvorens deze populatieschattingen te presenteren zal eerst nader worden ingegaan op de analyse van pakkansen voor subgroepen.

Pakkansen van subgroepen

In het gebruikte rekenmodel zijn vier verklarende variabelen opgenomen; geslacht, leeftijd, nationaliteit en reden van staandehouding⁷. De b-waarden in de tabel geven het verschil in pakkans aan met de corresponderende referentiecategorie. Een negatieve b-waarde duidt op een grotere pakkans en een positieve b-waarde op een kleinere pakkans. Een b-waarde van nul betekent dat de pakkans van die categorie van de verklarende variabele niet van die van de referentiecategorie verschilt. Indien de absolute t-waarde (t) groter is dan 2 is dan verschilt de bijbehorende b-waarde statistisch gezien significant afwijkend van nul, hetgeen er op duidt dat het verschil echt is en niet op toeval berust. Naarmate het aantal observaties kleiner is wordt het echter steeds moeilijker vast te stellen of een verschil echt is of niet. Om die reden is de periode 97-00 ook in de tabel opgenomen, vanwege het grotere aantal observaties kunnen in die dataset ware verschillen beter worden opgespoord.

Tabel 4: Geschatte 'b- en t-waarden voor de enkele jaren en de vierjaarsperiode (niet-effectief uitgezette illegalen)

		periode				
		97	98	99	00	97-00
intercept	b	-2.49	-4.28	-1.86	-2.59	-1.60
	t	-5.44	-5.37	-3.77	-4.81	-9.26
geslacht	b	.84	1.36	.33	.31	.63
	t	3.06	2.96	1.21	1.23	6.40
leeftijd	b	.52	.93	.30	.32	.33
	t	1.60	2.45	.88	.98	3.01
Turks	b	-.61	-.16	-.80	-.71	-.81
	t	-1.48	-.23	-1.74	-1.17	-4.61
Noord Afrikaans	b	.21	1.33	.01	.85	.45
	t	.76	2.28	.03	2.01	3.66
Rest Afrika	b	-.47	.06	-.73	.27	-.29
	t	-1.57	.10	-1.98	.63	-2.27
Surinaams	b	-.09	.01	-.78	.13	-.16
	t	-.20	.01	-1.30	.24	-.91
Aziatisch	b	-.56	.16	-.90	-.77	-.62
	t	-1.87	.27	-2.39	-1.66	-4.70
Nationaliteit onbekend	b	-.53	.74	-2.40	.	-.75
	t	-1.04	1.11	-2.29	.	-3.49
Reden aanhouding	b	-.61	-.66	-.35	-.24	-.53
	t	-4.32	-4.44	-2.15	-1.55	-10.23

Referentiecategorieën:

⁷ De covariaten als dummievariabelen in het model ingevoerd, hetgeen wil zeggen dat één categorie van de covariaat (de referentiecategorie) steeds vergeleken wordt met de overige categorieën van die covariaat. De referentiecategorieën zijn onderaan tabel gepresenteerd.

geslacht: vrouw, leeftijd: ouder dan 40 jaar, nationaliteit: N-Amerika/Z-Amerika/Oceanië, Reden staandehouding: anders dan illegaal verblijf,

intercept: indicatie pakkans voor de referentiegroep.

* vanwege convergentieproblemen is in 00 de groep met onbekende nationaliteit samengevoegd met de groep Aziaten

De positieve 'b-waarden voor geslacht en leeftijd duiden erop dat de pakkans van mannen en jongere illegalen groter is dan van respectievelijk vrouwen en ouderen. Op dezelfde wijze is af te leiden dat van alle nationaliteiten de groep afkomstig uit Noord-Afrika de grootste pakkans heeft en dat de pakkans van de Surinamers niet wezenlijk verschilt van die van de referentiegroep. De reden van staandehouding is, gezien grote t-waarden, een belangrijke verklarende variabele. Uit de negatieve bèta is af te leiden dat illegale migranten die in aanraking komen met de politie in verband met illegaal verblijf, een grotere kans hebben gepakt te worden dan illegalen die voor andere zaken worden verdacht.

Om inzichtelijk te maken hoe de 'b-waarden aan populatieschatting zijn gerelateerd zijn voor 1997 de 'b-waarden de geobserveerde-, geschatte- en de ratio van geobserveerde en geschatte frequenties van de verklarende variabelen opgenomen in tabel 5. De referentiecategorieën hebben een b-waarde van nul. De ratio geobserveerd/geschat geeft aan hoeveel personen van een bepaalde categorie niet zijn geobserveerd per gepakt persoon uit die categorie.

In de tabel is te zien dat per gepakte man 7,7 mannen niet gepakt zijn en dat per gepakte vrouw 19,5 vrouwen niet zijn gepakt. De pakkans van de mannen t.o.v. de vrouwen is dus veel groter hetgeen wordt gereflecteerd in de positieve b-waarde van de mannen. Op deze wijze zijn binnen elk van de verklarende variabelen de categorieën met referentiecategorieën te vergelijken. Door voor alle groepen het aantal niet-geobserveerde personen op te tellen wordt een schatting van de totale populatiegrootte verkregen.

Tabel 5: Aantal geobserveerd, geschatte bèta en aantal niet-geobserveerd per pakking (niet-effectief uitgezette illegalen)

	geobserveerd frequentie	geschatte frequentie	beta	ratio geobserveerd/ geschat
mannen	2171	16689	.84	7.7
vrouwen (ref)	478	9337	.00	19.5
< 40 jr	2413	21974	.52	9.1
> 40 jr (ref)	236	4053	.00	17.2
Turkije	220	3177	.21	14.4
Noord Afrika	684	3043	-.47	4.4
rest Africa	622	7197	-.09	11.6
Suriname	71	566	-.56	8.0
azie	802	9710	-.53	12.1
America/Oceanie (ref)	250	2332	.00	9.3
illegaal	89	1033	-.61	11.6
andere reden (ref)	2560	24994	.00	9.8

Resultaat: schattingen van de totale populatie illegale migranten

Tot zover hebben we ons beperkt tot de niet-effectief uitgezette illegalen. De populatieomvang van de overige twee groepen (de effectief uitgezette personen en degene die in de categorie overig vielen) is berekend door - op basis van de voor de corresponderende periode berekende beta's in de niet-effectief uitgezette groep - het totale aantal geobserveerde en niet-geobserveerde personen te berekenen. De schattingen van de drie groepen zijn bij elkaar opgeteld en vervolgens is - op een vergelijkbare wijze als bij de niet-effectief uitgezette groep - een betrouwbaarheidsinterval berekend. De resultaten van deze berekeningen zijn voor alle perioden in de onderstaande tabel samengevat, achtereenvolgens voor de afzonderlijke jaren en voor de langere perioden. Aangezien de afzonderlijke jaren de meest valide schattingen geven, gaan we daar verder vanuit. De tabel dient als volgt gelezen te worden: in 1997 komt de schatting uit op ruim 77.000 personen, het betrouwbaarheidsinterval ligt tussen de 67.000 en 87.000, en zo verder. Wanneer we de schattingen van 1997, 1999 en 2000 middelen komen we uit op rond de 78.000 personen per jaar, exclusief West- en Oost-Europeanen.

Tabel 6: Schatting totale populatie alle illegale groepen tezamen voor tien perioden, inclusief betrouwbaarheidsinterval, exclusief Oost- en West- Europeanen.

	totale populatie illegalen	min 95% CI	max 95% CI
97	77120	67286	86954
98	128367	91994	164740
99	73401	56511	90292
00	84523	71578	97467
97-98	112559	102399	122719
98-99	103608	95077	112139
99-00	102837	90489	115184
97-99	129468	121790	137146
98-00	121004	114160	127848
97-00	146310	139740	152881

Voor de totale illegalenpopulatie (van buiten de Europese landen) is dezelfde trend te zien als bij de niet-effectief uitgezette; de omvang van de populatie neemt toe en het betrouwbaarheidsinterval wordt kleiner bij toename van de geobserveerde periode. Ook hier steekt de schatting over 1998 af tegen de overige enkele jaren. Dit is te verklaren uit het feit dat voor de schatting van deze populatie dezelfde 'b-waarden (berekend met de verklarende variabele geslacht in het model) zijn gebruikt die voor de niet-effectief uitgezette groep ook al tot afwijkende schattingen leidden.

Wanneer we tenslotte een grove inschatting van het aantal West- en Oost-Europese illegalen willen maken, kunnen we de verhoudingen binnen de geschatte populatie ook toepassen op deze groepen. Het percentage van de groepen West en Oost Europeanen bedragen 40 % op het geheel, circa 12% is afkomstig uit West-Europese

landen en de overige 28% uit Oost Europa. In het totaal komt dit op een aantal van ongeveer 5.000 illegale Europeanen per jaar. Het aantal niet-geobserveerde personen per gepakt persoon bij de niet-effectief uitgezette personen – over de enkele jaren gezien – bedraagt circa 12. Als we deze factor toepassen op de Europese groepen levert dit een populatieschatting van circa 60.000 personen. Een grof betrouwbaarheidsinterval bij deze ruwe schatting kan worden berekend door de populatiegrootte van de niet-effectief uitgezette illegalen te vermenigvuldigen met een bepaalde factor gelijk te stellen aan die van de Europese groepen en door vervolgens de populatievarianties van met dezelfde factor te vermenigvuldigen. Die berekening levert de betrouwbaarheidsintervallen zoals in tabel 7 weergegeven.

Tabel 7: Populatieschatting en betrouwbaarheidsinterval Europese illegalen

	geschatte populatiegrootte	min 95% CI	max 95% CI
97	60000	50828	69172
98	60000	37734	82266
99	60000	43224	76776
00	60000	49166	70834

Uit deze tabel kunnen we – met de nodige slagen om de arm – opmaken dat de groep Oost en West Europeanen tezamen op jaarbasis tussen de 47.000 en de 72.000 zou kunnen bedragen (minimum en maximum gemiddeld over 1997, 1999 en 2000).

Conclusies

In dit hoofdstuk hebben we aan de hand van de vangst-hervangst methode die is toegepast op de in hoofdstuk 2 beschreven VAS-data een aantal schattingen van het aantal illegalen in Nederland gepresenteerd. Het gaat daarbij om schattingen die berekend zijn voor steeds een jaar en voor schattingen over langere periodes (resp. 2, 3, en 4 jaar) . De hoogte van schatting blijkt afhankelijk te zijn van de lengte van de bestudeerde periode: hoe langer de periode, des te groter de geschatte populatie. De veronderstelling is dat dit effect is te wijten aan een schending van de assumptie van een gesloten populatie. De meest valide schattingen zijn dan ook de schattingen voor de periode van een jaar. Opmerkelijk was de consistentie van de schatting door de jaren heen. Alleen 1998 vertoonde een afwijking van dit gelijkmatige beeld. Dit verschil is waarschijnlijk aan toevalligheden in de dataset te wijten, een heranalyse waarin voor deze toevalligheden werd gecorrigeerd gaf een vergelijkbaar beeld als dat van de overige jaren te zien. Gebleken is dat de populatieschatting robuust is tegen uitschieters op de pakkingen. Zelfs observaties met extreem veel pakkingen hebben slechts een geringe invloed op de hoogte van de populatieschatting of op de breedte van het betrouwbaarheidsinterval. Hercodering of zelfs verwijdering van de extreme observaties leidde niet tot noemenswaardige verschillen in deze schattingen. De schattingen zijn echter minder robuust tegen toevallig afwijkende verdelingen binnen een verklarende variabele. Groepen waarbinnen nagenoeg alle personen slechts éénmaal zijn gepakt leiden tot convergentieproblemen in het computerprogramma of tot (te) hoge schattingen van het niet geobserveerde aantal personen.

Wat betreft de uitkomsten beperken we ons dus bij voorkeur tot de schattingen per jaar. Wanneer we ons daarbij baseren op 1997, 1999 en 2000 kunnen we concluderen dat de schatting op jaarbasis exclusief uit Europa afkomstige illegale migranten uitkomt rond de 78.000 illegalen. Dit aantal lijkt tamelijk constant over de jaren heen, hetgeen aansluit bij de bevindingen in het voorafgaande hoofdstuk. In tegenstelling tot in het eerdere onderzoek (van der Leun et al. 1998) zijn we nu ook in staat een betrouwbaarheidsinterval aan te geven. Wanneer we ook hierbij het gemiddelde nemen over de beste datasets (1997, 1999 en 2000), komen we uit op tussen de 65.000 en de 91.000 illegalen exclusief Oost- en West-Europeanen. Voor de laatste groepen kunnen we slechts een zeer ruwe raming maken, die uitkomt op 60.000 tezamen. De betrouwbaarheid van deze berekening is echter aanzienlijk minder dan van die bij de andere groepen. Illegale West- en Oost-Europeanen zijn immers minder goed te schatten omdat in hun geval de assumpties sterker worden geschonden. Vermoedelijk zijn zij vaker 'te kort' in Nederland, hetgeen leidt tot een hoger aantal personen dat niet vaker dan één keer wordt gepakt en daarmee tot een te hoge schatting. Wanneer we daar toch vanuit gaan zou het totaal op rond de 138.000 uitkomen. Hierbij moet echter nogmaals met nadruk gemeld worden dat de onzekerheden groot zijn en dat dit laatste getal slechts een grove indicatie betreft.

Aangezien de verhouding staandehoudingen in de G4 en in de rest van Nederland ongeveer 1:1 is, zie hoofdstuk 2, kunnen we tot slot concluderen dat de uitkomsten van deze schattingen de schatting gebaseerd op politiegegevens uit 1995 nauwelijks ontlopen⁸. Destijds kwamen we uit op minimaal 40.000 illegale migranten (exclusief Oost- en West-Europeanen) in de vier grote steden. De meerwaarde ten opzichte van de schatting van enkele jaren geleden is (1) dat we nu betrouwbaarheidsintervallen kunnen berekenen en (2) dat we over data van meerdere jaren beschikken.

⁸ Deze bevindingen zijn ook consistent met andere schattingen zoals die van het CBS, zie Hoogteijling 2002. Een kanttekening daarbij is wel dat de schattingen van het CBS ten dele op onze eigen onderzoeken zijn gebaseerd.

Bijlage 2:
Handreiking & SQL-programma aanmaak vangst-hervangstdata

L. Prins (2002). Syntax SQL

Handreiking voor het aanmaken van vangst-hervangstdata uit het Herkenningsdienstsysteem (HKS)

Deze handreiking is bedoeld voor onderzoekers die de vangst-hervangstmethodiek willen toepassen op gegevens uit het HKS. Uitgebreide ervaring op dit gebied is opgedaan met het project “Benutting HKS en ontwikkeling vangst-hervangstmethodiek”. Uit dit project is gebleken dat de bewerking van ruwe HKS-gegevens tot vangst-hervangstdata een gecompliceerd proces is waarin een aantal valkuilen verborgen zit. Het doel van deze handreiking is toekomstige onderzoekers te wijzen op belangrijke aandachtspunten bij het aanmaken van vangst-hervangstdata uit HKS-gegevens. Deze handreiking kan gezien worden als een aanvulling op hoofdstuk 1 van dit eindrapport. Het HKS is een bestand dat constant onderhevig is aan aanpassingen en veranderingen, de bijgevoegde SQL-syntax moet daarom worden opgevat als voorbeeld van de wijze waarop vangst-hervangstdata kan worden aangemaakt en niet als dwingend voorschrift.

Het HKS bestaat van oorsprong uit verschillende regionale bestanden. Na koppeling van de regionale bestanden en verwijdering van dubbele registraties (hetzelfde delict kan in meerdere regio's geregistreerd staan) ontstaan de landelijke bestanden. Twee van deze bestanden zijn “**land_ant_del.dbf**” en “**land_result.dbf**”, deze beide bestanden tezamen bevatten alle informatie die benodigd voor het aanmaken van vangst-hervangstdata. Van deze bestanden zijn twee versies in omloop, de jaarbestanden (land_ant_del_jaar.dbf” en “land_result_jaar.dbf) en de metabestanden (land_ant_del_meta.dbf” en “land_result_meta.dbf). In het begin van het project “Benutting HKS en ontwikkeling vangst-hervangstmethodiek” waren alleen de jaarbestanden beschikbaar, de data die uiteindelijk is gebruikt is gebaseerd op de metabestanden. Het belangrijkste verschil tussen de jaar- en metabestanden is dat in de jaarbestanden alleen personen zijn opgenomen die in dat bepaalde jaar een delict hebben gepleegd. In het metabestand zijn alle personen die een delict hebben gepleegd, ongeacht het jaar waarin het delict is gepleegd. De inhoud van beide bestanden en de verschillen tussen beide versies zijn hieronder besproken. Het doel is hiervan is duidelijk te maken welke variabelen de bestanden bevatten, hoe deze geïnterpreteerd dienen te worden en wat de (on)mogelijkheden van beide versies zijn. Voor de wijze waarop in het project “Benutting HKS en ontwikkeling vangst-hervangstmethodiek” de data is aangemaakt wordt verwezen naar de SQL-programmatuur.

In het bestand land_ant_del.dbf zijn ‘feiten’ de observatie-eenheden. Onder een ‘feit ‘ wordt elke afzonderlijke overtreding van een wet of artikel uit het Wetboek van Strafrecht verstaan (feiten worden in dit rapport ook wel aangeduid met de term delict). Deze wetten of wetsartikelen zijn in het bestand “land_ant_del.dbf” aangeduid met een 3-cijferige code, en elke code in de variabele hoofdrubriek (HFD_RUB) verwijst naar de overtreding van een bepaald artikel (355 bijvoorbeeld verwijst naar ‘voorhanden hebben van wapens’, een overtreding van de wet wapens en munitie (VWW)). Feiten kunnen zowel door de politie zijn geobserveerd of door de dader zijn bekend. In het bestand zijn ook ‘antecedenten’ opgenomen. De term antecedent verwijst naar de verzameling feiten die de dader ten laste worden gelegd. Een antecedent volgt op de aanhouding en het verhoor van de dader. In het bestand is per persoon het aantal antecedenten aangegeven in de variabele ANT_ID, elk nummer in de variabele “AUTONR” verwijst naar een apart persoon. Het jaar waarin het antecedent plaats heeft gehad is aangegeven in de variabele JAAR. In de onderstaande tabel, een selectie van 12 feiten uit “land_ant_del.dbf”, is

aan het nummer 155 te zien dat persoon 10162 in 1982 in zijn 9^e antecedent heling heeft gepleegd (nummer 155 verwijst naar artikel SR416/SR417).

Selectie variabelen Land_ant_del.dbf: selectie eerste 40 rijen.

	AUTONR	PERS_ID	JAAR	ANT_ID	HFD_RUB
1	10162	000000001	1982	009	155
2	10162	000000001	1901	015	133
3	10162	000000001	1901	015	135
4	10162	000000001	1976	014	135
5	10162	000000001	1980	013	135
22	107150	000000001	1993	008	355
23	107150	000000001	1993	006	351
24	107150	000000001	1995	005	133
25	107150	000000001	1995	004	137
26	107150	000000001	1995	003	133
36	55391	000000003	1988	206	137
37	55391	000000003	1988	207	133

In het bestand land_result_jaar.dbf zijn de personen de observatie-eenheid. In het bestand zijn een aantal persoonsgegevens en delictgegevens opgenomen. De persoonsgegevens omvatten o.a. het autonummer, de leeftijd, het geslacht, de leeftijd van het eerste delict en de regio waar het delict is gepleegd. De delictgegevens omvatten o.a. de variabelen JR80 t/m JR02 en R101 t/m R357. In het onderstaande voorbeeld zijn een aantal van die variabelen voor de eerste 10 personen uit het bestand (autonr 1 t/m 10) van 1996 getoond. In de tabel is bijvoorbeeld te zien dat persoon 3 in totaal één antecedent heeft met 3 feiten, waarvan er 2 het voorhanden hebben van wapens betreft.

Selectie variabelen uit land_result_1996.dbf: eerste 10 personen

	AUTO NR	ANT-REGIO	LFT_IDEL	JR 89	JR 96	SOM_ANT	SOM_FEIT	R317	R355
1	1	01	22	.	1	1	1	.	.
2	2	01	28	.	6	6	6	.	.
3	3	01	27	.	1	1	3	.	2
4	4	01	22	.	1	1	1	.	.
5	5	01	81	.	1	1	1	.	.
6	6	01	28	.	1	1	2	.	.
7	7	01	21	.	1	2	7	.	.
8	8	01	27	.	1	1	1	1	.
9	9	01	30	.	1	1	1	.	.
10	10	01	24	.	1	5	6	.	.

Door de persoonsgegevens in land_result_jaar.dbf te koppelen aan de delicten in land_ant_del_jaar.dbf ontstaat een database met een overzicht van feiten en antecedenten, jaartallen en persoonsgegevens. Nu is het mogelijk dat hetzelfde delict meerdere malen per antecedent voorkomt. Dit komt voor wanneer meerdere overtredingen van hetzelfde artikel hebben plaatsgevonden of zijn bekend, bijvoorbeeld bij de persoon met autonummer 3 in de bovenstaande tabel die twee overtredingen van code 355 (het voorhanden hebben van wapens) wegens het dragen van een mes en een pistool in 1 antecedent heeft. Nu dient de afhankelijke variabele in de vangst-hervangstmethode niet het aantal overtredingen maar het aantal aanhoudingen voor een bepaald type overtreding

te tellen, het gaat immers om de vangsten. In het bestand dat na koppeling van beide bovenstaande tabellen is verkregen dient daarom eerst nog te worden geaggregeerd over de antecedenten, zodat elk type overtreding maar één keer per wantecedent wordt geteld. Vervolgens kunnen dan het aantal aanhoudingen voor een bepaald delict, bijvoorbeeld ‘het voorhanden hebben van wapens’, worden opgeteld per persoon. In plaats van een bestand waarin drie rijen voor één persoon zijn opgenomen met elk één telling van het delict wordt zo een bestand verkregen met voor die persoon één rij met een telling van drie delicten.

De criminele achtergrond van een persoon kan worden gemeten middels een aantal delictcovariaten. Deze covariaten kunnen op dezelfde wijze als de afhankelijke variabele worden aangemaakt, na aggregatie wordt voor elk antecedent maximaal één type overtreding geteld en vervolgens worden per type overtredingen het totaal per persoon berekend. Aan te raden is deze periode op vijf te stellen, omdat het HKS-bestand voor een groot aantal overtredingen wordt geschoond nadat een persoon vijf lang geen antecedenten heeft gehad. Door de periode van vijf wordt de vergelijkbaarheid van de criminele achtergrond van de personen gewaarborgd.

De jaarbestanden kunnen niet worden gekoppeld over de jaren heen, omdat de personen per jaar een nieuw autonummer krijgen. Stel dat we de afhankelijke variabele over de jaren T-1 t/m T willen meten. Bij koppeling van jaarbestanden van jaar T en jaar T-1 komen de personen die in beide jaarbestanden geregistreerd zijn ook tweemaal voor in het gekoppelde bestand. Het is echter niet na te gaan voor welke gevallen in het gekoppelde bestand dat geldt, omdat de autonummers tussen bestand T en T-1 niet met elkaar corresponderen. Koppeling van de jaarbestanden leidt daarom tot een hoog percentage dubbele registraties.

Dit probleem is opgelost met de ontwikkeling van de meta-bestanden, dit zijn jaarbestanden die over de jaren heen gekoppeld zijn m.b.v. van het Gemeentelijk Basis Administratienummer (GBA –nummer). Met dit nummer is het wel mogelijk personen over de jaren heen te volgen. I.t.t. de jaarbestanden geven de meta-bestanden een volledig overzicht over alle jaren tegelijkertijd van gepleegde delicten, waarbij dubbele registraties zijn verwijderd. Een langere periode kan nu eenvoudig worden bestudeerd door de personen te selecteren die in de te bestuderen periode een bepaalde type delict hebben gepleegd. Voor de wijze waarop dit kan worden gedaan wordt verder verwezen naar de onderstaande SQL-syntax.

SQL-syntax

```
PUBLIC Install_dir
Install_dir=CHR(34)+SET("directory")+chr(34)
** END HEADER -- do not remove this line
//
// Generated on 29-05-2002
//
parameter bModal
local f
f = new Fss_UtrechtForm()
if (bModal)
    f.mdi = false // ensure not MDI
    f.readModal()
else
    f.open()
endif

class Fss_UtrechtForm of FORM
with (this)
    onOpen = class::FORM_ONOPEN
    scaleFontBold = false
    height = 25.2273
    left = 9.8571
    top = 0
    width = 109.2857
    text = ""
    windowState = 2 // Maximized
endwith

this.RECTANGLE1 = new RECTANGLE(this)
with (this.RECTANGLE1)
    left = 1
    top = 0.5
    width = 91
    height = 21
    metric = 0 // Chars
    text = "Rectangle1"
    colorNormal = "BtnText/BtnFace"
    border = false
    borderStyle = 2 // Lowered
endwith

this.RECTANGLE2 = new RECTANGLE(this)
with (this.RECTANGLE2)
    left = 2
    top = 18
    width = 89
    height = 3
    metric = 0 // Chars
    text = "Rectangle2"
    colorNormal = "BtnText/BtnFace"
    border = false
    borderStyle = 2 // Lowered
endwith

this.PUSHBUTTON1 = new PUSHBUTTON(this)
with (this.PUSHBUTTON1)
    onClick = class::PUSHBUTTON1_ONCLICK
    height = 1.0909
    left = 21
    top = 19
    width = 15.2857
    text = "Start actie"
    metric = 0 // Chars
    upBitmap = "RESOURCE #36"
    group = false
    colorNormal = "BtnText/BtnFace"
    value = false
endwith
```

```
this.PUSHBUTTON2 = new PUSHBUTTON(this)
with (this.PUSHBUTTON2)
  onClick = {;form.close()}
  height = 1.0909
  left = 65
  top = 19
  width = 15.2857
  text = "Sluiten"
  metric = 0 // Chars
  upBitmap = "RESOURCE #1005"
  group = false
  colorNormal = "BtnText/BtnFace"
  value = false
endwith
```

```
this.SPINBOX1 = new SPINBOX(this)
with (this.SPINBOX1)
  height = 1
  left = 23
  top = 5
  width = 8
  metric = 0 // Chars
  picture = "9999"
  colorHighLight = ""
  rangeMax = 2010
  rangeMin = 1992
  fontSize = 8
  value = 2000
  borderStyle = 7 // Client
endwith
```

```
this.SPINBOX2 = new SPINBOX(this)
with (this.SPINBOX2)
  onChange = class::SPINBOX2_ONCHANGE
  height = 1
  left = 42
  top = 5
  width = 8
  metric = 0 // Chars
  picture = "9999"
  colorHighLight = ""
  rangeMax = 2010
  rangeMin = 1992
  fontSize = 8
  value = 2000
  borderStyle = 7 // Client
endwith
```

```
this.SPINBOX3 = new SPINBOX(this)
with (this.SPINBOX3)
  height = 1
  left = 23
  top = 8
  width = 8
  metric = 0 // Chars
  picture = "9999"
  colorHighLight = ""
  rangeMax = 2010
  rangeMin = 1992
  fontSize = 8
  value = 2000
  borderStyle = 7 // Client
endwith
```

```
this.SPINBOX4 = new SPINBOX(this)
with (this.SPINBOX4)
  height = 1
  left = 42
  top = 8
  width = 8
  metric = 0 // Chars
  picture = "9999"
```

```
colorHighLight = ""
rangeMax = 2010
rangeMin = 1992
fontSize = 8
value = 2000
borderStyle = 7 // Client
endwith
```

```
this.TEXT1 = new TEXT(this)
with (this.TEXT1)
    height = 0.7273
    left = 5
    top = 5
    width = 17
    metric = 0 // Chars
    colorNormal = "BtnText/BtnFace"
    text = "Selectie jaren van"
endwith
```

```
this.TEXT2 = new TEXT(this)
with (this.TEXT2)
    height = 0.7273
    left = 32
    top = 5
    width = 10
    metric = 0 // Chars
    colorNormal = "BtnText/BtnFace"
    text = "tot en met"
endwith
```

```
this.TEXT3 = new TEXT(this)
with (this.TEXT3)
    height = 1.3636
    left = 5
    top = 1.5
    width = 63
    metric = 0 // Chars
    colorNormal = "navy/silver"
    fontSize = 19
    fontBold = true
    text = "Project Faculty of Social Sciences"
endwith
```

```
this.TEXT4 = new TEXT(this)
with (this.TEXT4)
    height = 1
    left = 5
    top = 4
    width = 29
    metric = 0 // Chars
    colorNormal = "navy/BtnFace"
    text = "Afhankelijke variabele"
endwith
```

```
this.TEXT5 = new TEXT(this)
with (this.TEXT5)
    height = 0.7273
    left = 5
    top = 7
    width = 29
    metric = 0 // Chars
    colorNormal = "navy/BtnFace"
    text = "Periode delictvariabele"
endwith
```

```
this.TEXT6 = new TEXT(this)
with (this.TEXT6)
    height = 0.7273
    left = 5
    top = 8
```

```
width = 17
metric = 0 // Chars
colorNormal = "BtnText/BtnFace"
text = "Selectie jaren van"
endwith
```

```
this.TEXT7 = new TEXT(this)
with (this.TEXT7)
height = 0.7273
left = 32
top = 8
width = 10
metric = 0 // Chars
colorNormal = "BtnText/BtnFace"
text = "tot en met"
endwith
```

```
this.TEXT8 = new TEXT(this)
with (this.TEXT8)
height = 1
left = 5
top = 10.5
width = 18
metric = 0 // Chars
colorNormal = "BtnText/BtnFace"
text = "Te selecteren rubriek"
endwith
```

```
this.COMBOBOX1 = new COMBOBOX(this)
with (this.COMBOBOX1)
height = 1
left = 23
top = 10.5
width = 8
metric = 0 // Chars
fontSize = 8
dataSource = 'array {"317","355"}'
sorted = true
style = 1 // DropDown
endwith
```

```
this.CHECKBOX1 = new CHECKBOX(this)
with (this.CHECKBOX1)
height = 1.0909
left = 5
top = 12
width = 26
text = "Omzetten naar Ascii"
metric = 0 // Chars
colorNormal = "BtnText/BtnFace"
value = false
group = false
textLeft = true
endwith
```

```
this.ENTRYFIELD1 = new ENTRYFIELD(this)
with (this.ENTRYFIELD1)
height = 1
left = 23
top = 13.5
width = 50
metric = 0 // Chars
colorHighLight = ""
value = "Naam resultaat"
borderStyle = 7 // Client
endwith
```

```
this.PUSHBUTTON3 = new PUSHBUTTON(this)
with (this.PUSHBUTTON3)
onClick = class::PUSHBUTTON3_ONCLICK
```



```

height = 1.0909
left = 73
top = 13.5
width = 3.2857
text = ""
metric = 0 // Chars
upBitmap = "RESOURCE #137"
group = false
colorNormal = "BtnText/BtnFace"
value = false
endwith

```

```

this.PUSHBUTTON4 = new PUSHBUTTON(this)
with (this.PUSHBUTTON4)
onClick = class::PUSHBUTTON4_ONCLICK
height = 1.0909
left = 38
top = 19
width = 15.2857
text = "Tabel tonen"
metric = 0 // Chars
upBitmap = "RESOURCE #36"
group = false
colorNormal = "BtnText/BtnFace"
value = false
endwith

```

```

this.TEXT9 = new TEXT(this)
with (this.TEXT9)
height = 0.7273
left = 5
top = 13.5
width = 15
metric = 0 // Chars
colorNormal = "BtnText/BtnFace"
text = "Bestandsnaam"
endwith

```

```

this.PROGRESS1 = new PROGRESS(this)
with (this.PROGRESS1)
height = 1
left = 23
top = 16.5
width = 55
metric = 0 // Chars
value = 0
rangeMin = 0
rangeMax = 100
endwith

```

```

this.TEXT10 = new TEXT(this)
with (this.TEXT10)
height = 0.7273
left = 5
top = 16.5
width = 15
metric = 0 // Chars
colorNormal = "BtnText/BtnFace"
text = "Voortgang"
endwith

```

```

this.ENTRYFIELD2 = new ENTRYFIELD(this)
with (this.ENTRYFIELD2)
height = 1
left = 23
top = 15
width = 50
metric = 0 // Chars
colorNormal = "red/white"
colorHighLight = ""
value = ""

```

```
borderStyle = 7 // Client
endwith
```

```
this.PUSHBUTTON5 = new PUSHBUTTON(this)
with (this.PUSHBUTTON5)
onClick = class::PUSHBUTTON5_ONCLICK
height = 1.0909
left = 4.1429
top = 18.9545
width = 15.2857
text = "Start batch"
metric = 0 // Chars
upBitmap = "RESOURCE #36"
group = false
colorNormal = "BtnText/BtnFace"
value = false
endwith
```

```
function form_onOpen
PUBLIC sq,cmd,reeks ,fl ,bst ,cmd1,telrub ,g
return
```

```
function PUSHBUTTON1_onClick
```

```
CLASS::bereken(form.combobox1.value,form.entryfield1.value,form.spinbox1.value,form.spinbox2.value,form.spinbox3.value,form.spinbox4.value)
return
```

```
function bereken(p1,p2,p3,p4,p5,p6)
form.combobox1.value=p1
form.entryfield1.value=p2
form.spinbox1.value=p3
form.spinbox2.value=p4
form.spinbox3.value=p5
form.spinbox4.value=p6
form.combobox1.SetFocus()
form.entryfield1.SetFocus()
form.spinbox1.SetFocus()
form.spinbox2.SetFocus()
form.spinbox3.SetFocus()
form.spinbox4.SetFocus()
form.pushbutton4.setfocus()
CLOSE DATABASES
SELECT 1
j3=p3
reeks= "jaar IN ("
for i= p3 to p4
reeks=reeks+chr(39)+LTRIM(STR(j3))+chr(39)+", "
j3++
next
reeks=LEFT(reeks,Len(reeks)-2)+")"
form.entryfield2.value="reeks="+reeks
form.entryfield2.setFocus()
form.Pushbutton1.Setfocus()
```

```
sq="SELECT distinct metanr from 'D:\dex 2000\pe_data\land\Land_ant_del_meta.dbf' where hfd_rub='"+form.combobox1.value+"' and "+reeks
&sq
copy to u1
use u1
form.entryfield2.value="Gevonden: in u1 "+Ltrim(str(reccount()))
form.entryfield2.setFocus()
form.Pushbutton1.Setfocus()
use
sq ="SELECT metanr,geslacht,geb_jaar,regionr,lft_1del from 'D:\dex 2000\pe_data\land\Land_result_meta.dbf' where R"+form.combobox1.value+" >0 "
&sq
copy to u2
form.entryfield2.value="Gekoppeld: in u2 "+Ltrim(str(reccount()))
form.entryfield2.setFocus()
form.Pushbutton1.Setfocus()
sq="select u2.metanr, u2.geslacht, u2.geb_jaar, u2.regionr,u2.lft_1del from u2 INNER JOIN 'u1.dbf' u1 ON (u1.metanr=u2.metanr) "
&sq
copy to u3
```

```

use u3
ar=reccount()
form.entryfield2.value="Gekoppeld: in u3 "+Ltrim(str(reccount()))
form.entryfield2.setFocus()
form.Pushbutton1.Setfocus()
use
SELECT 2
sq="select u2.metanr,L1.ant_id, L1.jaar,L1.hfd_rub from 'D:\dex 2000\pe_data\land\Land_ant_del_meta.dbf' L1 ;
INNER JOIN u2 ON (L1.metanr=u2.metanr) ORDER BY metanr,ant_id "
&sq
copy to u4
use u4
form.entryfield2.value = "indexeren : u4"
form.entryfield2.setFocus()
form.Pushbutton1.Setfocus()
index on metanr tag pi
telrub="R"+form.combobox1.value
ERASE u5.dbf
cmd1="CREATE TABLE u5 (metanr numeric(8,0),;
      hfd_rub char(3),"+ ;
      telrub+" numeric(4,0) ,;
      geweld numeric(4,0),;
      opium numeric(4,0),;
      econom numeric(4,0),;
      verkeer numeric(4,0),;
      verniel numeric(4,0),;
      vermogen numeric(4,0),;
      geslacht numeric(2,0),;
      geb_jaar char(2),;
      antregio char(2),;
      lft_1del numeric(2,0)) "

&cmd1
Use u5 in 3
form.entryfield2.value=" bezig met analyse.."
form.entryfield2.setFocus()
form.Pushbutton1.Setfocus()

SELECT 1
use u3
DO while not eof()
p=metanr
s=IIF(geslacht="V",1,0)
g=geb_jaar
r=regionr
l=lft_1del
SELECT 2
IF SEEK(p)
STORE 0 to vt1,vt2,vt3,vt4,vt5,vt6 ,rubtel
DO WHILE metanr = p
a=ant_id

STORE 0 TO v1,v2,v3,v4,v5,v6
STORE " " to r1,r2,r3,r4,r5,r6,r7
DO while ant_id = a
IF VAL(jaar) =>p5 and VAL(jaar) <=p6
rub=hfd_rub

IF VAL(jaar) =>p3 and VAL(jaar) <=p4
IF form.combobox1.value="317" and "317" $ rub
IF NOT "317" $r7
r7="317"
rubtel++
ENDIF
ENDIF

IF form.combobox1.value="355" and "355" $ rub
IF NOT "355" $ r7
r7="355"
rubtel++
ENDIF
ENDIF
ENDIF
DO CASE
CASE rub $ "115|117|119|121|123|125|127|129|131|139|141"
IF NOT rub $ r1
v1++

```

```

ENDIF
r1=r1+"|"+rub

CASE rub $ "351|353|354"
IF NOT rub $ r2
v2++
ENDIF
r2=r2+"|"+rub

CASE rub $ "335|337|339|343|345|347|349"
IF NOT rub $ r3
v3++
ENDIF
r3=r3+"|"+rub

CASE rub $ "317|321|323|325|327|329|331"
IF NOT rub $ r4
v4++
ENDIF
r4=r4+"|"+rub

CASE rub $ "101|103|105|107|113|147|149|151|153"
IF NOT rub $ r5
v5++
ENDIF
r5=r5+"|"+rub

CASE rub $ "109|111|133|135|137|143|145|155"
IF NOT rub $ r6
v6++
ENDIF
r6=r6+"|"+rub
ENDCASE
ENDIF
SKIP
ENDDO
vt1=vt1+v1
vt2=vt2+v2
vt3=vt3+v3
vt4=vt4+v4
vt5=vt5+v5
vt6=vt6+v6
ENDDO
SELECT 3
IF form.combobox1.value="317"
vt4=vt4-rubtel
ENDIF

APPEND BLANK
replace metanr with p,;
    hfd_rub with form.combobox1.value,;
    &telrub with rubtel,;
    geweld with vt1,;
    opium with vt2,;
    econom with vt3,;
    verkeer with vt4,;
    vermiel with vt5,;
    vermogen with vt6,;
    geslacht with s,;
    geb_jaar with g,;
    antregio with r,;
    lft_1del with l
ELSE
? "niet gevonden"
?? p
ENDIF

SELECT 1
SKIP
form.progress1.value=recno()/ar*100
ENDDO
SELECT 2
// index opruimen
SELECT 3
close databases
form.progress1.value=0
IF form.entryfield1.value # "Naam resultaat" and LEN(TRIM(form.entryfield1.value))>0

```

```

bst=form.entryfield1.value
bestandnaam= left(INSTALL_DIR,len(INSTALL_DIR)-1)+"\ "+BST+CHR(34)
form.entryfield2.value=" bezig met schrijven.."
form.entryfield2.setFocus()
form.Pushbutton1.Setfocus()
IF form.checkbox1.value
  use u5 nosave
  fc=fldcount()
  cmd="SET ALTERNATE TO "+bestandnaam
  &cmd
  SET ALTERNATE ON
  DO WHILE NOT EOF()
  ?
  for i= 1 to fc
  f=field(i)
  if TYPE(f)="N"
  ?? LTRIM(str(&f))
  else
  ?? &f
  endif
  if i<fc
  ?? " "
  endif
  next
  SKIP
  form.progress1.value=recno()/ar*100
  ENDDO
  SET ALTERNATE OFF
  SET ALTERNATE TO
ELSE
  use u5 nosave
  cmd="copy to "+ bestandnaam
  &cmd
ENDIF
ENDIF
  use u1 nosave
  use u2 nosave
  use u3 nosave
  use u4 nosave
  form.progress1.value=0
  form.entryfield2.value="klaar.."
  form.entryfield2.setFocus()
  form.Pushbutton1.Setfocus()
return

function PUSHBUTTON3_onClick
  fl=""
  IF form.checkbox1.value
  fl=putfile("Opslaan bestand","*.txt")
  ELSE
  fl=putfile("Opslaan bestand","*.dbf")
  ENDIF

  IF len(trim(fl))>0
  pos=RAT("\,fl)
  form.entryfield1.value=substr(fl,pos+1,Len(fl-pos))
  form.entryfield1.Setfocus()
  form.pushbutton1.Setfocus()
  ENDIF
return

function PUSHBUTTON4_onClick
  cd &install_dir
  bst=""
  bst=chr(34)+getfile("*.dbf","Open tabel")+chr(34)
  IF LEN(bst) >3
  Use &bst
  BROW
  ENDIF
  cd &install_dir
  return

function PUSHBUTTON5_onClick
CLASS::bereken("317","D_96_96",1996,1996,1992,1996)
CLASS::bereken("317","D_97_97",1997,1997,1993,1997)
CLASS::bereken("317","D_98_98",1998,1998,1994,1998)

```

```

CLASS::bereken("317","D_99_99",1999,1999,1995,1999)
CLASS::bereken("317","D_00_00",2000,2000,1996,2000)
CLASS::bereken("317","D_01_01",2001,2001,1997,2001)
CLASS::bereken("317","D_96_97",1996,1997,1993,1997)
CLASS::bereken("317","D_96_98",1996,1998,1994,1998)
CLASS::bereken("317","D_96_99",1996,1999,1995,1999)
CLASS::bereken("317","D_96_00",1996,2000,1996,2000)
CLASS::bereken("317","D_97_98",1997,1998,1994,1998)
CLASS::bereken("317","D_97_99",1997,1999,1995,1999)
CLASS::bereken("317","D_97_00",1997,2000,1996,2000)
CLASS::bereken("317","D_97_01",1997,2001,1997,2001)
CLASS::bereken("317","D_98_99",1998,1999,1995,1999)
CLASS::bereken("317","D_98_00",1998,2000,1996,2000)
CLASS::bereken("317","D_98_01",1998,2001,1997,2001)
CLASS::bereken("317","D_99_00",1999,2000,1996,2000)
CLASS::bereken("317","D_99_01",1999,2001,1997,2001)
CLASS::bereken("317","D_00_01",2000,2001,1997,2001)
CLASS::bereken("355","W_96_96",1996,1996,1992,1996)
CLASS::bereken("355","W_97_97",1997,1997,1993,1997)
CLASS::bereken("355","W_98_98",1998,1998,1994,1998)
CLASS::bereken("355","W_99_99",1999,1999,1995,1999)
CLASS::bereken("355","W_00_00",2000,2000,1996,2000)
CLASS::bereken("355","W_01_01",2001,2001,1997,2001)
CLASS::bereken("355","W_96_97",1996,1997,1993,1997)
CLASS::bereken("355","W_96_98",1996,1998,1994,1998)
CLASS::bereken("355","W_96_99",1996,1999,1995,1999)
CLASS::bereken("355","W_96_00",1996,2000,1996,2000)
CLASS::bereken("355","W_97_98",1997,1998,1994,1998)
CLASS::bereken("355","W_97_99",1997,1999,1995,1999)
CLASS::bereken("355","W_97_00",1997,2000,1996,2000)
CLASS::bereken("355","W_97_01",1997,2001,1997,2001)
CLASS::bereken("355","W_98_99",1998,1999,1995,1999)
CLASS::bereken("355","W_98_00",1998,2000,1996,2000)
CLASS::bereken("355","W_98_01",1998,2001,1997,2001)
CLASS::bereken("355","W_99_00",1999,2000,1996,2000)
CLASS::bereken("355","W_99_01",1999,2001,1997,2001)
CLASS::bereken("355","W_00_01",2000,2001,1997,2001)
Close databases
RUN(true,'rundll32.exe user,ExitWindows')
return

function SPINBOX2_onChange
    form.spinbox4.value=this.value
    form.spinbox3.value=this.value-4
    return
endclass -

```

Bijlage 3: Tabellen 'rijden onder invloed'.

Tabel 1a: Beta's en t-waarden intercept, geslacht en leeftijdsvariabelen (selectie)

	intercept		geslacht		leeftijd 1e delict		leeftijd	
	b	t	b	t	b	t	b	t
96	-1.55	-9.70	-.41	-1.80	-.02	-4.33	.01	.92
97	-1.91	-11.59	-.07	-.38	-.04	-7.05	.02	3.39
98	-2.06	-12.62	-1.02	-3.36	-.02	-3.47	.01	1.56
99	-1.98	-11.98	-.40	-1.70	-.03	-5.39	.01	2.01
00	-1.67	-11.20	-.46	-2.01	-.03	-5.58	.01	1.55
01	-1.70	-12.28	-.55	-2.65	-.02	-4.26	.01	2.26
96-97	-1.20	-13.24	-.39	-3.16	-.03	-12.29	.02	5.39
97-98	-1.45	-16.16	-.56	-4.21	-.03	-12.27	.02	6.45
98-99	-1.59	-17.40	-1.01	-5.99	-.03	-8.26	.01	4.23
99-00	-1.41	-16.21	-.58	-4.30	-.03	-10.78	.01	4.60
00-01	-1.21	-14.95	-.44	-3.72	-.03	-11.32	.01	4.94
96-98	-1.06	-16.67	-.55	-5.84	-.04	-19.88	.02	10.76
97-99	-1.27	-20.00	-.69	-6.90	-.04	-18.52	.02	11.28
98-00	-1.24	-19.92	-.88	-8.01	-.03	-15.23	.02	8.34
99-01	-1.08	-17.91	-.50	-5.59	-.04	-18.02	.02	9.14
96-99	-.98	-19.61	-.65	-8.36	-.04	-27.56	.03	16.80
97-00	-1.02	-21.04	-.66	-8.65	-.04	-25.38	.03	15.50
98-01	-.99	-20.40	-.72	-9.13	-.04	-24.55	.02	14.41
96-00	-.81	-20.02	-.64	-10.18	-.04	-35.52	.03	22.41
97-01	-.81	-20.44	-.61	-10.01	-.04	-35.68	.03	22.53

Rijden onder invloed: selectie regio's

Tabel 1b: Beta's en t-waarden intercept, geslacht en leeftijdsvariabelen (landelijk)

	intercept		geslacht		leeftijd 1e delict		leeftijd	
	b	t	b	t	b	t	b	t
96	-1.68	-16.91	-.49	-3.85	-.02	-7.44	.01	2.56
97	-1.96	-19.92	-.38	-3.33	-.03	-11.43	.02	5.31
98	-2.00	-20.47	-.73	-5.43	-.04	-12.90	.02	6.59
99	-2.03	-20.32	-.38	-3.28	-.04	-12.41	.02	6.58
00	-1.99	-21.34	-.29	-2.69	-.03	-10.76	.02	5.85
01	-1.66	-18.98	-.50	-4.33	-.03	-12.63	.02	7.54
96-97	-1.29	-23.68	-.47	-7.03	-.03	-20.35	.02	9.58
97-98	-1.46	-27.15	-.48	-7.41	-.04	-25.80	.02	13.70
98-99	-1.54	-28.04	-.57	-8.31	-.04	-27.07	.02	15.46
99-00	-1.53	-28.48	-.41	-6.45	-.04	-23.76	.02	13.39
00-01	-1.28	-25.03	-.39	-6.26	-.04	-27.47	.02	15.35
96-98	-1.11	-29.00	-.50	-10.61	-.04	-37.89	.02	21.53
97-99	-1.21	-31.51	-.51	-10.86	-.04	-42.15	.03	25.01
98-00	-1.25	-32.60	-.56	-11.53	-.04	-38.69	.03	23.09
99-01	-1.14	-30.07	-.44	-9.76	-.04	-40.87	.03	24.00
96-99	-.96	-31.91	-.52	-14.04	-.05	-56.47	.03	34.82
97-00	-1.01	-33.84	-.52	-14.03	-.04	-54.99	.03	33.86
98-01	-.98	-32.79	-.56	-14.77	-.05	-57.18	.03	35.30
96-00	-.82	-32.91	-.54	-17.39	-.05	-71.20	.03	45.67
97-01	-.81	-32.59	-.53	-17.26	-.05	-75.10	.03	47.93

Rijden onder invloed: alle regio's

Tabel 2a: Betas's en t-waarden delict variabelen (selectie)

	geweld		opium		economisch		verkeer		vermogens		vernietiging	
	b	t	b	t	b	t	b	t	b	t	b	t
96	.32	2.86	-.07	-.28	.76	2.52	.71	10.92	.09	1.22	.07	.62
97	.00	.03	.42	2.24	.00	.00	.74	10.82	-.03	-.37	.14	1.09
98	.28	2.28	-.25	-.98	-.56	-.89	.44	5.68	.25	2.78	.23	1.76
99	.23	1.99	.37	1.88	.14	.39	.75	10.08	.17	1.96	-.03	-.25
00	.10	.93	-.13	-.54	-.31	-.54	.66	9.63	.20	2.50	.24	2.07
01	.00	.03	.29	1.52	.14	.31	.59	8.40	.21	2.46	.28	2.36
96-97	.19	2.91	.19	1.63	.38	1.85	.73	18.45	-.01	-.18	.13	1.84
97-98	.08	1.25	.18	1.52	-.11	-.43	.62	14.63	.10	1.98	.21	2.88
98-99	.18	2.58	.07	.54	.14	.60	.54	12.16	.21	4.25	.17	2.23
99-00	.20	3.21	-.03	-.23	-.24	-.86	.67	16.22	.18	3.75	.19	2.82
00-01	.13	2.17	.02	.13	.27	1.03	.63	15.06	.18	3.69	.19	2.79
96-98	.12	2.60	.13	1.55	.21	1.33	.64	20.92	.08	2.34	.22	4.42
97-99	.14	3.01	.09	1.04	.30	2.01	.61	19.26	.11	3.14	.19	3.78
98-00	.17	3.78	-.08	-.91	-.13	-.65	.60	19.01	.20	5.60	.26	5.15
99-01	.16	3.55	-.01	-.12	.02	.09	.60	18.94	.19	5.48	.19	3.75
96-99	.15	4.24	.09	1.36	.42	3.77	.66	25.35	.08	3.01	.21	5.30
97-00	.14	4.01	-.05	-.66	.12	.87	.65	25.20	.11	3.85	.24	5.98
98-01	.16	4.47	-.03	-.43	.05	.30	.57	21.08	.15	5.33	.25	6.36
96-00	.16	5.15	-.01	-.16	.20	1.79	.67	28.81	.09	3.72	.24	7.20
97-01	.13	4.26	-.01	-.17	.21	1.76	.59	24.52	.10	3.95	.25	7.59

Rijden onder invloed: selectie regio's

Tabel 2b: Betas's en t-waarden delict variabelen (landelijk)

	geweld		opium		economisch		verkeer		vermogens		vernietiging	
	b	t	b	t	b	t	b	t	b	t	b	t
96	.17	2.68	.22	1.96	.32	1.52	.70	18.63	.13	3.09	.28	4.46
97	.15	2.58	.14	1.36	-.13	-.53	.79	21.93	.06	1.44	.26	4.37
98	.16	2.77	.06	.57	-.22	-.81	.50	13.04	.18	4.33	.21	3.42
99	.14	2.21	.35	3.34	.13	.60	.56	14.28	.21	4.69	.01	.20
00	.17	3.05	.16	1.56	.09	.34	.65	17.71	.26	6.39	.17	2.77
01	.10	1.72	.04	.37	-.33	-1.18	.55	14.44	.22	4.98	.27	4.32
96-97	.15	4.45	.15	2.52	.10	.86	.75	35.19	.07	2.94	.28	8.20
97-98	.10	2.95	.12	2.07	.01	.10	.65	30.17	.14	5.91	.26	7.57
98-99	.12	3.42	.12	1.97	.06	.51	.49	21.54	.18	7.25	.17	4.67
99-00	.15	4.67	.19	3.20	-.03	-.20	.61	28.08	.20	8.44	.14	4.03
00-01	.11	3.38	.05	.84	.02	.18	.58	26.21	.17	6.75	.23	6.81
96-98	.12	5.13	.12	2.94	.11	1.24	.66	41.01	.09	5.50	.30	12.30
97-99	.09	3.97	.10	2.22	.15	1.81	.59	35.92	.14	8.30	.23	9.07
98-00	.13	5.54	.08	1.89	.03	.27	.56	34.45	.19	10.62	.19	7.56
99-01	.13	5.51	.12	2.69	-.08	-.77	.55	33.30	.17	9.53	.19	7.39
96-99	.11	5.89	.11	3.42	.21	3.29	.61	44.85	.11	7.99	.25	12.80
97-00	.11	6.09	.06	1.84	.09	1.22	.63	47.37	.15	10.52	.22	11.17
98-01	.14	7.72	.08	2.15	.00	-.04	.53	38.38	.15	9.94	.21	10.34
96-00	.13	8.38	.10	3.56	.14	2.33	.65	53.99	.12	9.86	.23	13.71
97-01	.14	8.72	.06	2.20	.	.	.58	46.96	.13	10.50	.22	13.31

Rijden onder invloed: alle regio's

Tabel 3a: Beta's en t-waarden regio's (selectie)

	Gelderland Midden		Hollands Midden		Zuid-Holland-Zuid		Midden & West-Brabant	
	b	t	b	t	b	t	b	t
96	-.65	-4.28	-.64	-4.50	-.31	-2.29	-.41	-3.79
97	-.19	-1.16	-.48	-3.32	.09	.73	-.33	-2.76
98	-.45	-3.01	-.44	-3.01	-.22	-1.59	-.48	-3.77
99	-.67	-3.88	-.42	-2.92	.11	.87	-.51	-4.04
00	-.61	-3.77	-.56	-3.87	-.35	-2.61	-.13	-1.25
01	-1.01	-6.25	-.73	-5.86	-.80	-5.73	-.84	-7.22
96-97	-.46	-5.25	-.59	-7.44	-.20	-2.80	-.54	-8.18
97-98	-.35	-4.19	-.50	-6.14	-.09	-1.21	-.48	-7.01
98-99	-.53	-5.96	-.41	-5.12	-.10	-1.35	-.51	-7.09
99-00	-.60	-6.49	-.45	-5.70	-.03	-.46	-.24	-3.85
00-01	-.68	-7.77	-.58	-7.77	-.47	-6.18	-.40	-6.61
96-98	-.46	-7.67	-.57	-9.98	-.20	-3.95	-.53	-10.96
97-99	-.45	-7.26	-.43	-7.70	-.08	-1.63	-.47	-9.62
98-00	-.53	-8.46	-.43	-7.60	-.09	-1.84	-.35	-7.51
99-01	-.57	-8.95	-.43	-7.98	-.19	-3.72	-.32	-7.21
96-99	-.51	-10.53	-.49	-11.21	-.17	-4.00	-.51	-13.53
97-00	-.44	-9.20	-.43	-9.77	-.07	-1.87	-.40	-10.80
98-01	-.54	-10.96	-.43	-9.87	-.18	-4.24	-.35	-9.68
96-00	-.49	-12.21	-.46	-12.71	-.13	-3.97	-.45	-14.72
97-01	-.47	-11.51	-.42	-11.75	-.15	-4.38	-.38	-12.49

Rijden onder invloed: selectie regio's (referentieregio: Rotterdam Rijnmond)

Tabel 3b: Beta's en t-waarden regio's (landelijk)

	Gelderland Midden		Hollands Midden		Zuid-Holland-Zuid		Midden & West-Brabant	
	b	t	b	t	b	t	b	t
96	-.66	-4.40	-.65	-4.56	-.34	-2.54	-.43	-3.97
97	-.32	-1.98	-.49	-3.40	.06	.48	-.38	-3.22
98	-.44	-2.99	-.42	-2.85	-.21	-1.55	-.46	-3.63
99	-.64	-3.72	-.40	-2.84	.11	.87	-.51	-4.03
00	-.62	-3.80	-.58	-4.04	-.35	-2.62	-.14	-1.34
01	-1.02	-6.34	-.71	-5.76	-.80	-5.76	-.83	-7.15
96-97	-.52	-5.95	-.60	-7.57	-.24	-3.36	-.57	-8.64
97-98	-.40	-4.78	-.49	-6.15	-.10	-1.37	-.49	-7.10
98-99	-.51	-5.79	-.39	-4.85	-.09	-1.24	-.48	-6.78
99-00	-.58	-6.36	-.44	-5.58	-.02	-.31	-.23	-3.72
00-01	-.68	-7.74	-.57	-7.66	-.46	-6.16	-.39	-6.53
96-98	-.49	-8.27	-.57	-10.07	-.22	-4.21	-.54	-11.25
97-99	-.46	-7.52	-.42	-7.55	-.09	-1.69	-.46	-9.43
98-00	-.52	-8.26	-.41	-7.27	-.08	-1.55	-.33	-7.10
99-01	-.55	-8.79	-.42	-7.77	-.19	-3.63	-.31	-7.03
96-99	-.51	-10.67	-.48	-11.07	-.17	-4.04	-.51	-13.38
97-00	-.45	-9.36	-.41	-9.54	-.07	-1.81	-.39	-10.58
98-01	-.53	-10.81	-.41	-9.48	-.17	-4.03	-.34	-9.29
96-00	-.49	-12.34	-.45	-12.56	-.13	-3.89	-.45	-14.57
97-01	-.47	-11.70	-.41	-11.49	-.15	-4.41	-.37	-12.31

Rijden onder invloed: alle regio's (referentieregio: Rotterdam Rijnmond)

Tabel 4: Beta's en t-waarden regio's (vervolg landelijk)

	Groningen		Friesland		Drenthe		IJsselland		Twente		Noord & Oost Gelderland		Gelderland-Zuid	
	b	t	b	t	b	t	b	t	b	t	b	t	b	t
96	-.58	-4.43	-.51	-4.30	-.40	-2.74	-.51	-3.73	-.75	-5.09	-.87	-6.02	-.89	-5.17
97	-.36	-2.76	-.21	-1.88	-.56	-3.24	-.15	-1.13	-.57	-3.93	-.14	-1.19	-.69	-3.98
98	-.05	-.40	-.06	-.48	-.52	-2.43	-.23	-1.66	-.14	-1.11	-.33	-2.63	-.56	-3.11
99	.09	.74	.13	1.19	.01	.07	-.32	-2.05	-.65	-4.14	-.26	-2.06	-.98	-4.54
00	.00	.00	-.18	-1.58	-.12	-.58	-.49	-3.12	-.71	-4.70	-.22	-1.82	-.30	-2.01
01	-.40	-3.46	-.62	-5.26	-.53	-3.43	-.46	-3.54	-.77	-5.80	-.44	-4.10	-1.09	-5.49
96-97	-.40	-5.70	-.37	-5.82	-.42	-4.84	-.45	-5.90	-.60	-7.73	-.54	-7.51	-.78	-8.25
97-98	-.20	-2.91	-.12	-1.98	-.56	-5.13	-.26	-3.39	-.40	-5.38	-.32	-4.62	-.56	-6.03
98-99	.01	.19	.03	.43	-.37	-2.94	-.21	-2.57	-.32	-4.19	-.27	-3.85	-.54	-5.43
99-00	.06	.94	-.04	-.68	-.14	-1.14	-.41	-4.62	-.46	-5.82	-.20	-2.83	-.58	-5.85
00-01	-.10	-1.57	-.29	-4.46	-.38	-3.67	-.38	-4.76	-.57	-7.47	-.36	-5.37	-.47	-5.24
96-98	-.31	-6.12	-.26	-5.61	-.43	-6.18	-.35	-6.45	-.48	-8.98	-.49	-9.64	-.66	-9.93
97-99	-.14	-2.90	-.09	-2.01	-.54	-6.31	-.27	-4.81	-.42	-7.77	-.33	-6.64	-.55	-8.17
98-00	.00	.04	-.07	-1.58	-.36	-3.94	-.32	-5.33	-.36	-6.70	-.25	-5.09	-.41	-6.25
99-01	-.01	-.17	-.08	-1.74	-.34	-4.10	-.37	-6.11	-.44	-7.95	-.27	-5.59	-.53	-7.65
96-99	-.24	-6.17	-.20	-5.32	-.43	-7.23	-.34	-7.63	-.47	-11.02	-.42	-10.70	-.61	-11.60
97-00	-.09	-2.48	-.13	-3.59	-.53	-7.50	-.33	-7.21	-.41	-9.62	-.30	-7.72	-.44	-8.67
98-01	-.02	-.64	-.08	-2.24	-.36	-5.47	-.30	-6.46	-.38	-8.82	-.27	-7.01	-.43	-8.13
96-00	-.19	-5.84	-.20	-6.51	-.44	-8.37	-.39	-10.06	-.45	-12.64	-.39	-11.70	-.52	-12.28
97-01	-.09	-2.86	-.13	-4.12	-.48	-8.68	-.32	-8.36	-.42	-11.71	-.32	-9.87	-.45	-10.53

Rijden onder invloed: alle regio's (referentieregio: Rotterdam Rijnmond)

Tabel 5: Beta's en t-waarden regio's (vervolg landelijk)

	Utrecht		Noord-Holland- Noord		Zaanstreek- Waterland		Kennemerland		Amsterdam- Amstelland		Gooi- en Vechtstreek		Haaglanden	
	b	t	b	t	b	t	b	t	b	t	b	t	b	t
96	-.49	-4.80	-.70	-5.20	-.37	-2.35	-.85	-5.60	-.49	-5.31	-.41	-2.39	-.20	-2.40
97	-.31	-3.18	-.43	-3.34	-.66	-3.39	-.33	-2.64	-.16	-1.67	-.13	-.81	-.18	-1.96
98	-.28	-2.62	-.21	-1.66	-.25	-1.49	-.31	-2.32	-.16	-1.60	-.49	-2.49	-.09	-.99
99	-.47	-4.08	-.27	-2.10	-.37	-1.97	-.22	-1.69	-.26	-2.61	-.44	-2.26	-.14	-1.48
00	-.47	-4.41	-.81	-4.99	-.42	-2.49	-.52	-3.77	-.21	-2.23	-.16	-1.05	-.21	-2.30
01	-1.03	-8.62	-.86	-5.75	-.79	-4.90	-.93	-6.57	-.82	-7.12	-.74	-4.31	-.46	-5.49
96-97	-.37	-6.74	-.53	-7.47	-.44	-4.73	-.55	-7.48	-.34	-6.65	-.29	-3.17	-.16	-3.36
97-98	-.27	-4.86	-.29	-4.23	-.42	-4.41	-.32	-4.48	-.20	-3.86	-.26	-2.77	-.12	-2.37
98-99	-.28	-4.65	-.14	-2.01	-.21	-2.17	-.25	-3.38	-.12	-2.21	-.38	-3.57	-.03	-.53
99-00	-.31	-5.16	-.44	-5.67	-.27	-2.85	-.26	-3.53	-.18	-3.27	-.17	-1.83	-.10	-2.01
00-01	-.56	-9.15	-.72	-8.43	-.51	-5.62	-.50	-6.73	-.47	-8.11	-.39	-4.24	-.19	-3.88
96-98	-.32	-8.08	-.41	-8.22	-.35	-5.45	-.45	-8.72	-.29	-7.88	-.34	-5.02	-.13	-3.84
97-99	-.28	-6.96	-.26	-5.22	-.31	-4.63	-.32	-6.22	-.20	-5.36	-.27	-3.93	-.09	-2.66
98-00	-.26	-6.24	-.29	-5.52	-.20	-2.98	-.27	-5.10	-.15	-3.86	-.24	-3.45	-.08	-2.17
99-01	-.37	-8.52	-.49	-8.52	-.32	-4.82	-.30	-5.75	-.30	-7.38	-.23	-3.51	-.08	-2.22
96-99	-.29	-9.21	-.33	-8.42	-.31	-5.99	-.40	-9.70	-.25	-8.33	-.29	-5.37	-.12	-4.24
97-00	-.27	-8.30	-.30	-7.59	-.27	-5.22	-.31	-7.55	-.20	-6.49	-.20	-3.92	-.11	-3.95
98-01	-.29	-8.77	-.34	-7.97	-.27	-5.29	-.28	-6.70	-.24	-7.48	-.23	-4.23	-.05	-1.91
96-00	-.28	-10.61	-.36	-10.78	-.28	-6.63	-.38	-11.04	-.23	-9.26	-.23	-5.30	-.13	-5.62
97-01	-.28	-10.35	-.35	-10.11	-.31	-7.25	-.31	-9.03	-.26	-9.81	-.20	-4.50	-.09	-3.90

Rijden onder invloed: alle regio's (referentieregio: Rotterdam Rijnmond)

Tabel 6: Beta's en t-waarden regio's (vervolg landelijk)

	Zeeland		Brabant-Noord		Brabant - Zuid-Oost		Limburg-Noord		Limburg-Zuid		Flevoland	
	b	t	b	t	b	t	b	t	b	t	b	t
96	-0.81	-4.82	-0.74	-4.67	-0.84	-5.12	-0.51	-3.55	-0.80	-6.53	-0.70	-2.84
97	-0.05	-0.39	-0.22	-1.60	-0.55	-3.68	-0.61	-3.85	-0.52	-4.14	-0.06	-0.33
98	-0.15	-1.08	-0.35	-2.39	-0.28	-2.00	-0.11	-0.80	-0.33	-2.68	-0.14	-0.66
99	-0.37	-2.08	-0.56	-3.63	-0.24	-1.73	-0.35	-2.41	-0.33	-2.50	-0.17	-0.79
00	-0.48	-2.59	-0.07	-0.58	-0.41	-2.88	-0.74	-4.34	-0.67	-4.59	-0.34	-1.56
01	-0.71	-4.45	-0.47	-4.11	-0.67	-4.57	-1.03	-5.21	-0.77	-6.03	-0.84	-4.28
96-97	-0.36	-4.68	-0.57	-6.87	-0.65	-7.79	-0.64	-7.41	-0.55	-8.43	-0.37	-3.10
97-98	-0.11	-1.50	-0.26	-3.43	-0.51	-6.18	-0.39	-4.70	-0.42	-6.20	-0.22	-1.93
98-99	-0.29	-3.26	-0.40	-4.85	-0.27	-3.40	-0.22	-2.72	-0.31	-4.41	-0.13	-1.12
99-00	-0.47	-4.46	-0.29	-3.94	-0.38	-4.61	-0.41	-4.91	-0.47	-6.10	-0.32	-2.58
00-01	-0.59	-6.04	-0.22	-3.32	-0.57	-6.78	-0.68	-7.07	-0.65	-8.54	-0.44	-3.99
96-98	-0.29	-5.33	-0.47	-8.27	-0.58	-9.89	-0.52	-8.53	-0.46	-9.96	-0.37	-4.34
97-99	-0.20	-3.59	-0.39	-6.87	-0.47	-8.02	-0.31	-5.51	-0.43	-8.63	-0.24	-2.90
98-00	-0.39	-5.85	-0.30	-5.53	-0.38	-6.47	-0.27	-4.72	-0.43	-8.02	-0.27	-3.13
99-01	-0.44	-6.25	-0.26	-5.11	-0.47	-7.74	-0.47	-7.38	-0.53	-9.61	-0.28	-3.49
96-99	-0.29	-6.50	-0.49	-10.78	-0.56	-11.89	-0.45	-9.53	-0.45	-11.74	-0.32	-4.73
97-00	-0.27	-5.79	-0.34	-7.96	-0.49	-10.59	-0.34	-7.42	-0.48	-11.62	-0.30	-4.41
98-01	-0.37	-7.19	-0.26	-6.39	-0.45	-9.47	-0.32	-6.69	-0.46	-10.95	-0.29	-4.38
96-00	-0.32	-8.18	-0.43	-11.83	-0.58	-14.70	-0.42	-10.80	-0.48	-14.55	-0.34	-6.07
97-01	-0.29	-7.27	-0.32	-9.16	-0.52	-13.18	-0.36	-9.19	-0.48	-14.02	-0.30	-5.58

Rijden onder invloed: alle regio's (referentieregio: Rotterdam Rijnmond)

Tabel 7a: Geobserveerde frequenties enkelvoudige jaren

	jaren											
	96		97		98		99		00		01	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	8251	31352	8569	33224	9603	34643	8989	33139	8877	32557	9201	33409
2	521	1586	503	1657	456	1617	430	1502	481	1585	502	1591
3	60	158	48	167	43	152	44	146	52	172	60	173
4	12	24	5	23	2	18	4	19	8	18	10	24
5	1	7	2	6	1	7	1	6	1	4	.	3
6	2	2	.	1	1	2	.	.	.	2	.	0
7	.	0	.	2	.	0	.	.	.	0	.	0
8	.	1	.	.	.	1	.	.	.	0	.	2
9	0	.	.
10	0	.	.
11	0	.	.
12	1	.	.

Tabel 7b: Geschatte frequenties enkelvoudige jaren

	jaren											
	96		97		98		99		00		01	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	8210.9	31255.9	8547.2	33133.1	9575.6	34531.0	8966.5	33038.1	8847.2	32456.8	9164.0	33290.7
2	587.8	1749.7	541.6	1814.5	505.3	1807.1	472.1	1680.8	534.4	1761.5	569.7	1799.1
3	43.1	109.6	34.9	115.9	23.7	94.0	26.8	86.2	34.0	107.2	36.0	102.3
4	4.2	11.3	2.9	12.7	1.3	6.8	2.2	6.1	2.9	10.8	2.9	8.4
5	.7	2.3	.3	2.6	.1	.8	.3	.7	.4	2.0	.	1.1
6	.2	.7	.	.7	.0	.25	.	.2
7	.	.3	.	.2	.	.02	.	.0
8	.	.100	.	.0
90	.	.
100	.	.
110	.	.
120	.	.

Tabel 7c: Residuen enkelvoudige jaren

	jaren											
	96		97		98		99		00		01	
	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk
0
1	.44	.54	.24	.50	.28	.60	.24	.56	.32	.56	.39	.65
2	-2.75	-3.91	-1.66	-3.70	-2.19	-4.47	-1.94	-4.36	-2.31	-4.21	-2.84	-4.91
3	2.58	4.62	2.21	4.75	3.97	5.98	3.33	6.44	3.08	6.26	4.00	6.99
4	3.77	3.77	1.22	2.87	.62	4.26	1.18	5.22	2.98	2.19	4.22	5.38
5	.36	3.12	2.79	2.13	2.80	6.71	1.25	6.40	1.06	1.46	.	1.74
6	4.14	1.45	.	.29	9.33	4.78	.	.	.	2.11	.	-47
7	.	-.55	.	3.53	.	-1.18	.	.	.	-.39	.	-22
8	.	2.49	.	.	.	12.23	.	.	.	-.22	.	19.21
9	-.12	.	.
10	-.07	.	.
11	-.04	.	.
12	52.37	.	.

Tabel 8a: Geobserveerde frequenties 2- en 3-jaarsperioden

	jaren													
	96-97		97-98		98-99		99-00		00-01		96-98		97-99	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	15577	60166	16876	63328	17487	63563	16697	61536	16928	61787	23086	87379	23868	89496
2	1406	4724	1419	4846	1290	4593	1298	4506	1335	4572	2467	8611	2473	8486
3	203	607	175	611	150	589	192	619	207	641	445	1390	376	1351
4	47	135	31	117	26	104	30	111	43	125	100	295	82	271
5	15	40	5	34	5	24	6	22	7	27	20	81	16	79
6	2	13	3	6	2	10	1	5	4	9	9	31	5	19
7	2	7	1	8	1	2	0	4	1	3	6	12	5	16
8	.	1	.	4	.	0	0	1	.	0	.	8	.	5
9	.	1	.	3	.	0	0	0	.	0	.	4	.	3
10	.	0	.	.	.	1	1	1	.	2	.	3	.	1
11	.	1	.	.	.	1	.	0	.	0	.	1	.	0
12	1	.	1	.	2	.	1
13	0
14	1

Tabel 8b: Geschatte frequenties 2- en 3-jaarsperioden

	jaren													
	96-97		97-98		98-99		99-00		00-01		96-98		97-99	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	15443.5	59772.4	16779.8	62914.0	17389.0	63144.6	16587.5	61136.9	16775.3	61333.2	22797.4	86457.0	23643.7	88575.1
2	1615.9	5339.4	1576.6	5493.7	1452.0	5279.9	1487.5	5175.4	1588.3	5315.6	2927.3	10032.0	2829.0	9930.6
3	166.6	495.3	137.3	478.2	109.3	415.8	131.7	433.0	143.6	456.4	349.4	1120.4	306.5	1051.8
4	20.9	65.2	13.9	56.3	9.4	39.9	15.0	48.6	15.2	50.6	47.7	158.0	37.8	136.2
5	3.6	14.0	1.9	10.5	1.1	5.5	2.5	8.6	2.1	8.3	8.4	32.9	6.2	25.3
6	.9	4.6	.3	2.8	.2	1.0	.6	2.3	.4	2.0	1.9	9.9	1.3	6.6
7	.3	2.0	.1	.9	.0	.2	.1	.7	.1	.6	.6	3.8	.4	2.2
8	.	1.0	.	.3	.	.1	.0	.3	.	.2	.	1.7	.	.8
9	.	.5	.	.1	.	.0	.0	.1	.	.1	.	.8	.	.3
10	.	.30	.0	.0	.	.0	.	.3	.	.1
11	.	.20	.	.0	.	.0	.	.2	.	.0
120	.	.0	.	.1	.	.0
130
140

Tabel 8c: Residuen 2- en 3-jaarsperioden

	jaren													
	96-97		97-98		98-99		99-00		00-01		96-98		97-99	
	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk
1	1.07	1.61	.74	1.65	.74	1.66	.85	1.61	1.18	1.83	1.91	3.14	1.46	3.09
2	-5.22	-8.42	-3.97	-8.74	-4.25	-9.45	-4.91	-9.31	-6.36	-10.20	-8.51	-14.19	-6.69	-14.50
3	2.82	5.02	3.21	6.07	3.89	8.50	5.25	8.94	5.29	8.64	5.11	8.05	3.97	9.23
4	5.71	8.65	4.57	8.09	5.44	10.15	3.87	8.94	7.13	10.46	7.57	10.90	7.20	11.55
5	6.02	6.96	2.26	7.25	3.78	7.94	2.22	4.54	3.40	6.47	3.99	8.38	3.97	10.69
6	1.15	3.94	4.55	1.87	4.56	8.85	.60	1.82	6.00	5.04	5.06	6.74	3.16	4.82
7	2.87	3.56	3.41	7.25	5.80	3.66	-3.8	3.76	3.37	3.20	7.31	4.23	7.72	9.38
8	.	.00	.	6.23	.	-.24	-.20	1.39	.	-.44	.	4.92	.	4.65
9	.	.62	.	8.02	.	-.12	-.11	-.32	.	-.26	.	3.70	.	4.73
10	.	-.56	.	.	.	16.88	16.66	4.93	.	12.66	.	4.48	.	2.44
11	.	1.93	.	.	.	34.77	.	-.12	.	-.09	.	2.15	.	-.22
12	14.70	.	18.20	.	7.58	.	7.17
13	-.08
14	20.80

Tabel 9a: Geobserveerde frequenties 3- t/m 5-jaarsperioden

	jaren													
	98-00		99-01		96-99		97-00		98-01		96-00		97-01	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	24349	89186	24076	88212	29584	111524	30094	89496	31170	113753	35334	133329	36427	135824
2	2387	8280	2321	8133	3557	12615	3662	8486	3520	12370	4826	17004	4894	16967
3	379	1351	416	1368	717	2325	696	1351	674	2331	1054	3452	1030	3510
4	73	282	79	254	169	556	159	271	146	496	267	888	247	809
5	17	57	17	76	48	154	38	79	29	131	90	256	62	252
6	5	23	10	23	12	47	9	19	16	49	23	87	25	81
7	3	12	3	9	14	32	5	16	5	20	19	47	9	37
8	0	1	0	5	1	10	1	5	0	7	2	17	2	13
9	0	1	0	2	0	4	1	3	1	3	2	10	2	9
10	1	2	0	1	1	6	1	1	1	1	0	8	0	5
11	.	1	0	0	.	2	.	0	0	2	0	4	1	5
12	.	1	12	2	.	3	.	1	1	3	0	2	1	3
13	.	1	.	1	.	0	.	0	.	1	0	1	.	0
14	0	.	1	.	0	0	2	.	0
15	1	.	.	.	0	1	2	.	1
16	0	.	1	.	0
17	0	.	.	.	2
18	0	.	.	.	0
19	0	.	.	.	0
20	0	.	.	.	0
21	1	.	.	.	0
22	1

Tabel 9b: Geschatte frequenties 3- t/m 5-jaarsperioden

	jaren													
	98-00		99-01		96-99		97-00		98-01		96-00		97-0 ^a	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	24117.6	88269.3	23785.4	87249.2	29095.8	109879.1	29686.4	111305.0	30715.0	112134.8	34608.1	130892.1	35760.9	.
2	2762.9	9756.8	2792.7	9673.1	4311.8	15105.4	4314.8	15053.9	4242.0	14901.6	5916.5	20622.1	5920.3	.
3	290.0	1011.1	300.4	1007.9	582.6	1921.4	561.5	1886.7	520.4	1813.3	900.1	2951.6	854.8	.
4	35.5	128.1	36.9	126.1	88.7	287.3	81.5	281.0	69.9	253.6	148.9	483.5	130.6	.
5	6.0	23.4	6.0	21.9	17.4	58.8	15.4	58.3	12.0	47.0	30.9	105.9	24.3	.
6	1.4	6.1	1.3	5.3	4.4	16.6	3.9	16.8	2.7	11.7	8.3	31.4	5.8	.
7	.4	2.0	.3	1.6	1.4	5.9	1.2	6.2	.7	3.6	2.7	11.7	1.7	.
8	.1	.8	.1	.6	.5	2.5	.5	2.7	.2	1.4	1.0	5.1	.6	.
9	.0	.3	.0	.2	.2	1.1	.3	1.4	.1	.6	.4	2.6	.3	.
10	.0	.1	.0	.1	.1	.5	.2	.8	.0	.3	.2	1.4	.2	.
11	.	.1	.0	.0	.	.2	.	.5	.0	.1	.2	.9	.1	.
12	.	.0	.0	.0	.	.1	.	.3	.0	.1	.1	.6	.1	.
13	.	.0	.	.0	.	.0	.	.2	.	.0	.1	.4	.	.
140	.	.1	.	.0	.1	.3	.	.
1500	.1	.2	.	.
160	.	.1	.	.
170
180
190
200
210
22

a. landelijke schatting ontbreekt wegens geheugentekort computerprogramma

Tabel 9b: Geschatte frequenties 3- t/m 5-jaarsperioden

	jaren													
	98-00		99-01		96-99		97-00		98-01		96-00		97-01	
	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk	regionaal	landelijk
1	1.49	3.09	1.88	3.26	2.86	4.96	2.37	4.63	2.60	4.83	3.90	6.74	3.52	.
2	-7.15	-14.95	-8.93	-15.66	-11.50	-20.26	-9.94	-19.50	-11.09	-20.74	-14.18	-25.19	-13.34	.
3	5.23	10.69	6.67	11.34	5.57	9.21	5.67	10.39	6.73	12.16	5.13	9.21	5.99	.
4	6.28	13.60	6.93	11.38	8.53	15.85	8.59	14.61	9.10	15.23	9.67	18.39	10.18	.
5	4.47	6.95	4.53	11.54	7.35	12.41	5.76	11.61	4.92	12.26	10.64	14.59	7.64	.
6	3.08	6.88	7.81	7.71	3.59	7.47	2.59	6.14	8.17	10.94	5.11	9.91	7.99	.
7	4.29	7.08	4.82	5.82	10.69	10.71	3.37	7.97	5.15	8.57	9.92	10.30	5.61	.
8	-33	.27	-.29	5.89	.69	4.81	.71	3.20	-.44	4.83	.99	5.22	1.82	.
9	-.19	1.23	-.15	3.87	-.46	2.76	1.46	5.65	4.07	3.18	2.39	4.64	3.33	.
10	9.50	5.19	-.08	3.21	2.97	7.68	2.05	3.68	8.17	1.41	-.48	5.52	-.41	.
11	.	4.12	-.04	-.18	.	3.63	.	-.68	-.06	5.24	-.40	3.39	2.35	.
12	.	6.81	52.99	18.62	.	8.86	.	1.37	33.02	11.95	-.36	1.95	2.62	.
13	.	11.32	.	15.71	.	-.22	.	-.41	.	5.82	-.34	1.04	.	.
14	-.14	.	9.09	.	-.11	-.32	3.46	.	.
15	11.14	.	.	.	-.07	3.09	4.33	.	.
16	-.05	.	2.50	.	.
17	-.03
18	-.02
19	-.01
20	-.01
21	307.25
22

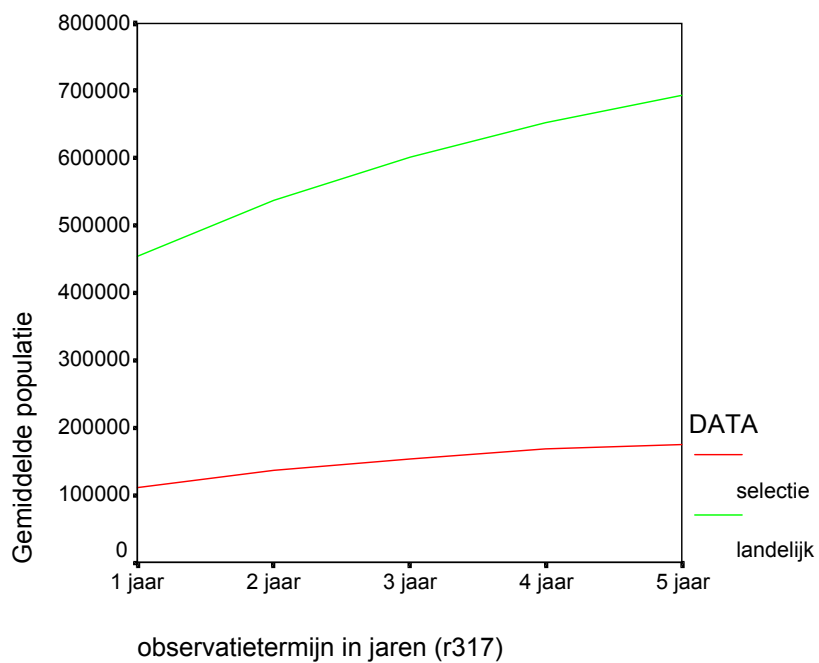
Tabel 10: Populatieschattingen, betrouwbaarheidsintervallen en pakkansen

			geschatte populatie	min. 95% BI	max. 95% BI	aantal gepakte personen	gemiddelde pakkans
selectie	jaar van observatie	96	87610	78102	97118	8847	.10
		97	99661	88802	110520	9127	.09
		98	126677	109598	143756	10106	.08
		99	128764	112771	144757	9468	.07
		00	113771	99857	127685	9419	.08
		01	113598	100673	126523	9773	.09
		96-97	117143	109906	124380	17252	.15
		97-98	133602	125396	141808	18510	.14
		98-99	156159	144022	168296	18961	.12
		99-00	142924	133335	152513	18225	.13
		00-01	136645	128084	145206	18525	.14
		96-98	142078	135868	148288	26133	.18
		97-99	152887	146019	159755	26825	.18
		98-00	165261	156943	173579	27214	.16
		99-01	157073	150081	164065	26923	.17
		96-99	160931	155281	166581	34103	.21
		97-00	164452	158738	170166	34666	.21
		98-01	177456	170999	183913	35563	.20
		96-00	170709	165811	175607	41618	.24
		97-01	178085	173035	183135	42700	.24
landelijk	jaar van observatie	96	421088	394367	447809	33130	.08
		97	461314	432647	489981	35080	.08
		98	476547	445405	507689	36440	.08
		99	468124	438080	498168	34812	.07
		00	444100	416738	471462	34339	.08
		01	457500	428913	486087	35202	.08
		96-97	514758	497444	532072	65695	.13
		97-98	544661	526579	562743	68957	.13
		98-99	557945	538722	577168	68887	.12
		99-00	538277	520189	556365	66806	.12
		00-01	537084	518891	555277	67167	.13
		96-98	581858	568309	595407	97817	.17
		97-99	609874	595471	624277	99729	.16
		98-00	611658	597047	626269	99198	.16
		99-01	604526	590155	618897	98086	.16
		96-99	638419	626714	650124	127279	.20
		97-00	654139	642158	666120	128614	.20
		98-01	667912	655367	680457	129168	.19
		96-00	681584	671278	691890	155110	.23
		97-01	703971	693218	714724	157519	.22

Tabel 11: Test op overdispersie 1-jaars observatieperiode (selectie)

		chi-kwadraat	p
jaar	96	58.78	1.77E-14
	97	25.60	4.19E-07
	98	44.36	2.74E-11
	99	14.36	1.51E-04
	00	24.60	7.05E-07
	01	35.16	3.03E-09

Figuur 12: Gemiddelde geschatte populatieomvang per observatieperiode



Bijlage 4: Tabellen 'wapens voorhanden'

Tabel 1a: Beta's en t-waarden intercept, geslacht en leeftijdsvariabelen (selectie)

	intercept		geslacht		leeftijd 1e delict		leeftijd	
	b	t	b	t	b	t	b	t
96	-4.13	-4.69	.	.	-.08	-1.98	.04	1.66
97	-4.18	-5.47	-.17	-.16	-.04	-1.41	.03	1.30
98	-4.33	-5.70	.	.	-.03	-1.14	.04	1.46
99	-3.75	-5.20	1.31	2.35	-.04	-1.20	.04	1.25
00	-2.80	-3.14	.29	.28	-.02	-.48	-.03	-.89
01	-3.48	-3.47	.	.	-.06	-1.44	.02	.81
96-97	-3.99	-8.33	-1.06	-1.05	-.05	-2.65	.04	2.80
97-98	-3.49	-8.15	-1.15	-1.14	-.05	-2.81	.03	2.45
98-99	-3.82	-9.67	.35	.75	-.02	-1.31	.03	2.01
99-00	-3.22	-7.54	.45	.97	-.03	-1.59	.02	1.37
00-01	-2.69	-4.34	-.27	-.26	-.06	-1.96	.00	.10
96-98	-3.50	-10.67	-1.71	-1.70	-.05	-3.61	.03	3.17
97-99	-3.43	-11.64	-.12	-.28	-.03	-2.76	.03	2.92
98-00	-3.26	-11.18	-.04	-.08	-.03	-2.13	.02	2.15
99-01	-2.80	-8.03	-.02	-.04	-.05	-3.64	.03	2.47
96-99	-3.41	-13.71	-.32	-.82	-.04	-3.90	.03	4.05
97-00	-3.12	-13.19	-.37	-.96	-.03	-2.85	.02	2.43
98-01	-2.92	-11.62	-.37	-.89	-.04	-4.14	.03	3.35
96-00	-3.15	-15.11	-.50	-1.41	-.03	-4.08	.03	3.64
97-01	-2.91	-13.84	-.62	-1.61	-.04	-4.95	.03	3.99

wapens: regionaal

Tabel 1b: Beta's en t-waarden intercept, geslacht en leeftijdsvariabelen (landelijk)

	intercept		geslacht		leeftijd 1e delict		leeftijd	
	b	t	b	t	b	t	b	t
96	-3.82	-8.34	-1.21	-1.69	-.02	-1.44	.01	.50
97	-3.35	-6.91	.21	.50	-.06	-3.13	.02	1.38
98	-4.46	-9.17	.	.	-.04	-2.34	.02	1.56
99	-4.11	-9.43	.02	.03	-.03	-2.05	.03	2.46
00	-3.60	-6.80	-1.01	-.99	-.02	-1.05	-.02	-.94
01	-4.65	-9.31	-.31	-.52	-.01	-.81	.01	.90
96-97	-3.05	-11.40	-.43	-1.46	-.03	-3.44	.01	.70
97-98	-3.09	-11.49	-.69	-1.93	-.04	-4.63	.01	1.82
98-99	-3.37	-13.43	-.42	-1.36	-.04	-3.98	.02	3.24
99-00	-3.02	-11.57	-.60	-1.55	-.04	-4.28	.02	3.11
00-01	-3.45	-10.91	-.87	-1.71	-.03	-2.69	.01	1.23
96-98	-3.02	-15.49	-.88	-3.27	-.04	-5.56	.01	1.87
97-99	-3.02	-16.06	-.52	-2.24	-.04	-6.44	.02	3.86
98-00	-2.80	-15.13	-.76	-2.71	-.04	-6.13	.02	3.96
99-01	-3.03	-14.59	-.78	-2.55	-.05	-6.38	.03	4.62
96-99	-2.88	-18.92	-.66	-3.42	-.04	-8.14	.02	4.27
97-00	-2.72	-18.28	-.80	-3.62	-.04	-8.01	.02	4.69
98-01	-2.74	-17.66	-.90	-3.67	-.05	-8.45	.02	5.88
96-00	-2.64	-20.85	-.79	-4.42	-.05	-10.43	.02	5.87
97-01	-2.67	-20.47	-.89	-4.47	-.05	-10.94	.02	7.34

wapens: landelijk

Tabel 2a: Betas's en t-waarden delict variabelen (selectie)

	geweld		opium		economisch		verkeer		vermogens		vernieling	
	b	t	b	t	b	t	b	t	b	t	b	t
96	.92	3.01	.99	2.55	1.64	1.78	.06	.15	.19	.98	-.11	-.29
97	.65	2.17	1.06	3.26	-.60	-.42	-.02	-.04	.26	1.27	-.37	-.92
98	.31	.89	-.74	-1.26	.49	.50	.53	1.20	.35	1.32	.15	.34
99	.66	1.97	-.01	-.02	-.36	-.24	.40	.85	.15	.57	-.01	-.02
00	.83	2.14	1.30	3.32	1.58	1.98	.68	1.44	-.20	-.63	-.34	-.70
01	.20	.46	1.73	4.01	-.69	-.44	.77	1.67	-.03	-.10	.29	.54
96-97	.94	5.33	.93	4.49	.22	.35	-.03	-.10	.17	1.46	-.20	-.88
97-98	.50	2.87	.21	.91	.48	.90	.32	1.45	.25	2.04	.20	.94
98-99	.34	1.86	.21	.90	-.68	-.80	.25	1.04	.53	3.81	.02	.10
99-00	.62	3.09	.56	2.32	.86	1.66	.19	.74	.08	.54	.04	.15
00-01	.57	2.22	1.38	5.13	.07	.09	.09	.28	-.11	-.58	.25	.86
96-98	.70	5.41	.40	2.48	.92	2.57	.21	1.24	.23	2.58	.09	.57
97-99	.32	2.54	.47	2.99	.15	.32	.31	2.00	.39	4.26	.25	1.65
98-00	.51	3.83	.53	3.30	.44	1.07	.39	2.38	.35	3.49	.10	.63
99-01	.40	2.67	.89	5.14	.54	1.23	.26	1.39	.14	1.26	.12	.68
96-99	.49	4.73	.50	3.94	.45	1.32	.24	1.83	.31	4.22	.26	2.13
97-00	.50	4.87	.68	5.58	.48	1.45	.39	3.18	.34	4.65	.18	1.47
98-01	.42	3.75	.66	4.90	.56	1.66	.35	2.58	.29	3.50	.18	1.37
96-00	.58	6.69	.69	6.75	.55	2.01	.37	3.65	.30	4.90	.20	2.05
97-01	.47	5.18	.72	6.71	.67	2.48	.38	3.51	.29	4.54	.18	1.73

wapens: regionaal

Tabel 2b: Betas's en t-waarden delict variabelen (landelijk)

	geweld		opium		economisch		verkeer		vermogens		vernieling	
	b	t	b	t	b	t	b	t	b	t	b	t
96	.58	4.23	.65	4.03	.55	1.20	.26	1.38	.25	2.56	-.12	-.67
97	.38	2.42	.73	3.91	.14	.24	.06	.28	.30	2.77	.21	1.18
98	.68	4.00	.18	.81	.86	1.70	.35	1.69	.22	1.81	-.07	-.36
99	.62	3.66	-.04	-.17	-1.67	-1.19	.13	.54	.32	2.71	.28	1.45
00	.57	2.84	.75	3.15	.86	2.16	.17	.65	.16	1.03	.23	1.03
01	.16	.77	.64	2.80	.53	.73	.63	2.87	.35	2.32	.06	.24
96-97	.61	7.49	.65	6.98	.51	1.93	.22	2.21	.28	5.00	.05	.48
97-98	.62	7.04	.58	5.57	.36	1.20	.16	1.43	.20	3.21	.23	2.37
98-99	.56	6.28	.33	2.98	.05	.14	.18	1.54	.26	4.13	.22	2.16
99-00	.51	5.10	.49	4.02	.62	2.25	.04	.32	.25	3.36	.05	.45
00-01	.58	4.91	.62	4.73	.44	1.57	.34	2.52	.18	2.07	.15	1.16
96-98	.68	11.21	.61	8.80	.41	2.05	.21	2.92	.24	5.62	.11	1.67
97-99	.58	9.14	.54	7.18	.08	.31	.21	2.68	.24	5.29	.30	4.31
98-00	.54	8.10	.56	7.14	.39	1.90	.23	2.74	.25	5.25	.16	2.19
99-01	.51	6.73	.57	6.58	.38	1.78	.15	1.61	.26	4.70	.04	.43
96-99	.61	12.49	.59	10.46	.16	.86	.28	4.78	.24	7.03	.18	3.31
97-00	.56	11.01	.67	11.29	.25	1.45	.25	4.00	.26	6.91	.23	4.11
98-01	.55	9.92	.57	8.91	.36	2.24	.27	4.01	.23	5.73	.15	2.37
96-00	.59	14.10	.70	14.88	.16	1.05	.31	6.51	.25	8.47	.15	3.33
97-01	.58	12.98	.68	13.36	.29	2.04	.28	5.21	.22	6.82	.18	3.56

wapens: landelijk

Tabel 3a: Beta's en t-waarden regio's (selectie)

	Gelderland Midden		Hollands Midden		Zuid-Holland-Zuid		Midden & West-Brabant	
	b	t	b	t	b	t	b	t
96	-1.23	-1.49	-.37	-.48	-.74	-.94	.32	.72
97	-.05	-.08	-.28	-.38	.	.	.47	1.19
98	-.09	-.16	.	.	.52	.90	-.40	-.62
99	-.27	-.48	-.68	-.89	-1.46	-1.41	-.47	-.88
00	-.17	-.23	.	.	.22	.33	-.77	-1.20
01	-1.39	-1.31	-.47	-.60	-1.17	-1.11	-1.05	-1.34
96-97	-.64	-1.54	-.37	-.84	-.65	-1.46	.16	.63
97-98	.04	.12	-.43	-.97	-.30	-.72	.19	.71
98-99	-.16	-.48	-.44	-1.00	.06	.17	-.34	-1.10
99-00	-.16	-.45	-.24	-.58	-.28	-.71	-.64	-1.92
00-01	-.05	-.12	-.68	-1.10	-.01	-.03	-.30	-.68
96-98	-.09	-.37	-.19	-.63	-.07	-.24	.23	1.16
97-99	-.03	-.12	-.26	-.89	-.07	-.27	.00	.02
98-00	-.28	-1.15	-.54	-1.67	-.06	-.25	-.50	-2.20
99-01	-.19	-.67	-.02	-.06	-.19	-.65	-.29	-1.22
96-99	-.15	-.73	-.26	-1.06	.12	.57	.09	.56
97-00	-.15	-.79	-.40	-1.65	.00	.00	-.22	-1.35
98-01	-.18	-.89	-.32	-1.25	-.12	-.54	-.24	-1.33
96-00	-.22	-1.27	-.31	-1.53	.09	.57	-.07	-.51
97-01	-.08	-.49	-.25	-1.21	-.01	-.05	-.08	-.54

wapens: regionaal (referentieregio: Rotterdam Rijnmond)

Tabel 3b: Beta's en t-waarden regio's (landelijk)

	Gelderland Midden		Hollands Midden		Zuid-Holland-Zuid		Midden & West-Brabant	
	b	t	b	t	b	t	b	t
96	-.60	-.78	-.05	-.07	-.34	-.44	.61	1.42
97	-.07	-.12	-.38	-.51	.	.	.47	1.19
98	.38	.66	.	.	.90	1.57	.04	.05
99	-.07	-.12	-.37	-.49	-1.26	-1.22	-.24	-.46
00	.44	.58	.	.	.65	1.01	-.16	-.25
01	.83	1.26	-.33	-.32	-.03	-.04	.65	.98
96-97	-.61	-1.46	-.40	-.90	-.56	-1.26	.14	.58
97-98	.05	.17	-.50	-1.14	-.32	-.78	.15	.59
98-99	-.20	-.62	-.48	-1.10	.04	.11	-.32	-1.03
99-00	-.05	-.15	-.25	-.58	-.22	-.59	-.55	-1.72
00-01	.03	.07	-.51	-.84	-.14	-.31	-.36	-.89
96-98	-.11	-.46	-.23	-.77	-.10	-.35	.18	.96
97-99	-.08	-.33	-.32	-1.10	-.13	-.50	-.04	-.18
98-00	-.28	-1.16	-.52	-1.64	-.07	-.29	-.49	-2.19
99-01	.00	-.01	.14	.47	-.02	-.08	-.10	-.41
96-99	-.17	-.85	-.24	-1.02	.09	.44	.07	.43
97-00	-.18	-.93	-.41	-1.70	-.03	-.14	-.23	-1.41
98-01	-.19	-.94	-.32	-1.29	-.13	-.58	-.23	-1.28
96-00	-.22	-1.32	-.29	-1.47	.08	.50	-.07	-.56
97-01	-.10	-.61	-.27	-1.35	-.03	-.18	-.08	-.59

wapens: landelijk (referentieregio: Rotterdam Rijnmond)

Tabel 4: Beta's en t-waarden regio's (vervolg landelijk)

	Groningen		Friesland		Drenthe		IJsselland		Twente		Noord & Oost Gelderland		Gelderland Zuid	
	b	t	b	t	b	t	b	t	b	t	b	t	b	t
96	-.46	-.60	1.08	2.21	-.59	-.56	.68	1.24	-.41	-.53	.79	1.76	.25	.43
97	.61	1.30	-.34	-.45	-.41	-.39	-.80	-.78	-1.34	-1.29	-.39	-.61	-.45	-.59
98	.68	1.28	1.26	2.1608	.11	1.03	1.94	.76	1.18
99	.00	.00	-1.12	-1.08	.	.	-.85	-.82	-1.25	-1.21	.34	.74	.98	2.39
00	1.03	2.3741	.54	-.53	-.51	.	.	.34	.45
01	.76	1.38	.33	.42	-.01	-.01	.76	1.29	.78	1.18	-.40	-.38	1.53	2.55
96-97	.30	1.00	.41	1.30	-1.08	-1.49	.04	.10	-.72	-1.62	.16	.57	-.50	-1.20
97-98	.15	.52	.12	.32	-1.32	-1.31	-.05	-.11	-1.03	-1.95	-.16	-.46	.05	.13
98-99	.38	1.41	-.15	-.36	-.95	-.95	-.18	-.41	-.45	-1.09	-.08	-.24	.69	2.51
99-00	.09	.31	-.73	-1.24	.	.	-.65	-1.26	-.76	-1.62	-.20	-.58	.31	1.04
00-01	.55	1.83	-.28	-.51	.	.	-.06	-.12	-.14	-.33	-.17	-.36	.24	.59
96-98	.30	1.40	.40	1.58	-1.37	-1.91	.13	.45	-.57	-1.76	.19	.88	-.06	-.20
97-99	.31	1.52	-.22	-.73	-1.24	-1.74	.01	.04	-.63	-2.02	-.33	-1.31	.41	1.87
98-00	.32	1.73	-.36	-1.07	-1.58	-1.57	-.55	-1.58	-.59	-1.97	-.31	-1.24	.19	.83
99-01	.37	1.75	-.12	-.34	.	.	-.12	-.34	-.29	-.95	.09	.32	.36	1.46
96-99	.37	2.22	.16	.73	-1.36	-2.33	.13	.59	-.55	-2.20	.02	.13	.23	1.18
97-00	.34	2.27	-.32	-1.24	-1.60	-2.25	-.12	-.50	-.56	-2.37	-.42	-2.01	.13	.68
98-01	.32	2.04	-.11	.44	-.14	-1.97	-.38	-1.40	-.47	-2.00	-.11	-.54	.22	1.16
96-00	.35	2.64	-.02	-.12	-1.62	-2.79	.04	.22	-.37	-1.96	-.09	-.55	.03	.19
97-01	.32	2.44	-.12	-.60	-1.24	-2.46	-.11	-.54	-.41	-2.17	-.20	-1.15	.19	1.19

wapens: landelijk (referentieregio: Rotterdam Rijnmond)

Tabel 5: Beta's en t-waarden regio's (vervolg landelijk)

	Utrecht		Noord-Holland- Noord		Zaanstreek- Waterland		Kennemerland		Amsterdam- Amstelland		Gooi- en Vechtstreek		Haaglanden	
	b	t	b	t	b	t	b	t	b	t	b	t	b	t
96	.51	1.17	1.14	2.42	-.58	-.55	.	.	.42	1.11	1.64	2.82	.51	1.20
97	.15	.35	-.01	-.02	.64	1.13	.05	.06	-.18	-.44	.97	1.28	.51	1.36
98	.81	1.77	.50	.66	.92	1.20	.82	1.08	-.51	-.89	1.26	1.65	.91	2.23
99	-.68	-1.20	-.35	-.47	.	.	-.02	-.02	-.40	-1.00	.	.	.13	.32
00	.20	.35	.	.	.85	1.12	.13	.12	.18	.39	.	.	.88	2.14
01	.	.	.13	.2378	1.74	.	.	.69	.89
96-97	.34	1.41	.38	1.24	.28	.76	-.57	-1.08	.24	1.13	.87	2.14	.50	2.22
97-98	.09	.33	.46	1.42	.58	1.63	.11	.26	.02	.10	.15	.29	.35	1.56
98-99	-.14	-.53	-.51	-1.07	-.98	-1.36	.20	.48	-.15	-.67	-.16	-.27	.32	1.46
99-00	-.25	-.90	-.51	-.98	-.32	-.62	-.66	-1.12	-.39	-1.73	.	.	.09	.40
00-01	-.54	-1.18	.49	1.08	-.60	-.82	-1.31	-1.28	-.15	-.50	.	.	.30	1.10
96-98	.33	1.80	.56	2.45	.43	1.54	-.17	-.47	.36	2.26	.43	1.22	.50	2.97
97-99	-.07	-.36	.32	1.34	.02	.07	.02	.05	.05	.29	-.12	-.29	.29	1.76
98-00	-.23	-1.20	-.51	-1.46	-.64	-1.52	-.07	-.23	-.23	-1.42	-.56	-1.11	.16	1.02
99-01	-.25	-.97	.40	1.35	-.67	-1.30	-.73	-1.42	-.24	-1.22	.	.	.24	1.34
96-99	.19	1.25	.42	2.22	.10	.39	-.16	-.58	.31	2.44	.23	.77	.41	3.10
97-00	-.13	-.84	.16	.78	-.12	-.46	-.19	-.69	-.05	-.40	-.43	-1.12	.17	1.32
98-01	-.23	-1.32	.15	.67	-.97	-2.32	-.32	-1.07	-.21	-1.53	-.88	-1.73	.14	1.04
96-00	.09	.73	.27	1.62	.08	.36	-.22	-.97	.17	1.65	-.09	-.32	.28	2.56
97-01	-.13	-.89	.38	2.20	-.31	-1.22	-.38	-1.47	-.02	-.14	-.51	-1.42	.15	1.35

wapens: landelijk (referentieregio: Rotterdam Rijnmond)

Tabel 6: Beta's en t-waarden regio's (vervolg landelijk)

	Zeeland		Brabant-Noord		Brabant - Zuid-Oost		Limburg-Noord		Limburg-Zuid		Flevoland	
	b	t	b	t	b	t	b	t	b	t	b	t
96	.65	1.18	.41	.70	-.83	-.79	.36	.66	.26	.54	.78	1.19
97	.95	2.13	-1.19	-1.14	.08	.15	-1.17	-1.12	-1.03	-1.35	-.05	-.06
98	.	.	.59	.91	1.28	2.56	.	.	1.14	2.65	.56	.73
99	.09	.12	-.80	-1.06	-1.13	-1.09	-1.27	-1.23	-.07	-.15	.21	.33
00	1.63	3.25	-.35	-.33	-1.33	-1.28	.	.
01	.	.	1.02	1.74	1.02	2.28
96-97	.57	1.95	-.05	-.14	-.34	-.82	.21	.65	-.09	-.30	.52	1.44
97-98	.40	1.25	.02	.06	.17	.52	-1.43	-1.98	.14	.53	.19	.50
98-99	-.83	-1.39	-.71	-1.63	.18	.57	-2.29	-2.26	.22	.88	.07	.18
99-00	-.26	-.45	-1.00	-1.93	.	.	-.95	-1.83	-.17	-.63	-.39	-.75
00-01	-.27	-.27	-.88	-1.44	.83	2.50	-.63	-1.03	.23	.78	-.70	-.95
96-98	.49	2.15	.23	.92	.01	.03	.25	.98	.17	.83	.63	2.48
97-99	.09	.32	-.37	-1.32	.09	.36	-1.38	-2.98	.07	.37	.21	.80
98-00	-1.02	-1.99	-1.08	-2.92	.12	.53	-1.08	-2.74	-.07	-.39	-.32	-.95
99-01	.03	.06	-.62	-1.72	.35	1.35	-.46	-1.28	.31	1.59	-.16	-.43
96-99	.27	1.36	-.10	-.49	.15	.75	-.02	-.08	.09	.57	.44	2.13
97-00	-.18	-.70	-.51	-2.17	.07	.37	-.79	-2.89	-.12	-.78	-.07	-.30
98-01	-.76	-1.83	-.82	-2.98	.14	.75	-.74	-2.53	.13	.92	-.25	-.92
96-00	.04	.22	-.13	-.75	.07	.45	-.06	-.36	-.07	-.51	.16	.84
97-01	-.16	-.68	-.44	-2.21	.09	.56	-.60	-2.63	.08	.59	-.05	-.23

wapens: landelijk (referentieregio: Rotterdam Rijnmond)

Tabel 7a: Geobserveerde frequenties enkelvoudige jaren

	96		97		98		99		00		01	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	1307	5296	1282	4899	1409	4829	1215	4617	1001	3837	943	3628
2	23	111	25	99	21	85	24	92	21	65	14	64
3	1	13	3	6	1	5	1	3	.	.	1	3
4	1	2

Tabel 7b: Geschatte frequenties enkelvoudige jaren

	96		97		98		99		00		01	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	1305.61	5284.8	1280.45	4895.3	1408.47	4826.9	1214.64	4617.3	1001.84	3836.5	943.11	3627.2
2	24.95	131.9	28.25	105.7	22.08	89.4	24.75	91.7	19.38	65.3	14.02	65.8
3	1.29	5.0	1.18	3.8	.44	2.6	.60	2.9	.	.	.69	1.9
4	.13	.3

Tabel 7c: Residuen enkelvoudige jaren

	96		97		98		99		00		01	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	.04	.15	.04	.05	.01	.03	.01	.00	-.03	.01	.00	.01
2	-.39	-1.82	-.61	-.65	-.23	-.47	-.15	.03	.37	-.03	-.01	-.22
3	-.25	3.59	1.68	1.13	.85	1.51	.52	.06	.	.	.37	.78
4	2.41	2.92

Tabel 8a: Geobserveerde frequenties 2- en 3-jaarsperioden

	96-97		97-98		98-99		99-00		00-01		96-98		97-99	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	2548	9902	2625	9460	2561	9166	2166	8268	1922	7338	3854	14288	3741	13651
2	62	319	76	286	72	297	65	228	46	180	128	581	142	556
3	8	41	6	31	5	17	3	16	1	11	12	69	12	54
4	0	4	1	3	.	.	1	9	.	.
5	1	1	1	1	.	.

Tabel 8b: Geschatte frequenties 2- en 3-jaarsperioden

	96-97		97-98		98-99		99-00		00-01		96-98		97-99	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	2542.31	9875.0	2622.64	9443.7	2558.90	9158.9	2180.56	8256.6	1627.07	7332.2	3846.22	14252.9	3737.61	13633.0
2	72.04	369.4	80.93	318.1	76.35	308.6	69.98	248.4	40.06	190.4	141.39	646.6	149.49	591.3
3	4.10	20.5	3.23	15.8	2.61	14.3	2.35	9.4	1.69	7.7	7.66	43.1	7.26	36.4
4	.46	1.810	.5	.	.	.65	4.6	.	.
5	.08	.207	.7	.	.

Tabel 8c: Geschatte residuen 2- en 3-jaarsperioden

	96-97		97-98		98-99		99-00		00-01		96-98		97-99	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	.11	.27	.05	.17	.04	.07	.05	.13	-.03	.07	.13	.29	.06	.15
2	-1.18	-2.62	-.55	-1.80	-.50	-.66	-.48	-1.30	.31	-.75	-1.13	-2.58	-.61	-1.45
3	1.93	4.51	1.55	3.84	1.48	.73	.42	2.17	-.53	1.21	1.57	3.95	1.76	2.92
4	-.68	1.62	2.92	3.38	.	.	.43	2.06	.	.
5	3.38	1.65	3.43	.38	.	.

Tabel 9a: Geobserveerde frequenties 3- t/m 5-jaarsperioden

	98-00		99-01		96-99		97-00		98-01		96-00		97-01	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	3466	12646	3044	11616	4943	18354	4619	17004	4318	15890	5796	21595	5451	20173
2	128	496	105	401	205	895	205	798	181	707	276	1178	262	1017
3	11	41	7	37	18	97	21	85	15	70	31	133	29	132
4	2	8	2	5	1	12	3	13	3	11	3	24	4	16
5	2	5	1	6	.	.
6	1	1	.	.

Tabel 9b: Geschatte frequenties 3- t/m 5-jaarsperioden

	98-00		99-01		96-99		97-00		98-01		96-00		97-01	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	3459.33	12625.4	3039.00	11595.1	4933.08	18311.3	4609.96	16971.2	4310.27	15860.8	5781.48	21540.0	5439.77	20121.7
2	140.02	531.5	113.43	440.1	221.50	970.8	222.17	858.8	194.71	760.8	301.11	1273.5	283.75	1117.0
3	7.04	31.5	5.17	21.8	12.98	70.8	14.12	62.4	10.93	50.5	22.09	105.3	19.74	87.5
4	.54	3.1	.36	1.7	1.23	8.2	1.48	7.7	.96	5.6	2.72	14.1	2.28	11.6
517	1.447	2.9	.	.
610	.8	.	.

Tabel 9b: Geschatte frequenties 3- t/m 5-jaarsperioden

	98-00		99-01		96-99		97-00		98-01		96-00		97-01	
	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk	selectie	landelijk
1	.11	.18	.09	.19	.14	.32	.13	.25	.12	.23	.19	.37	.15	.36
2	-1.02	-1.54	-.79	-1.87	-1.11	-2.43	-1.15	-2.08	-.98	-1.95	-1.45	-2.68	-1.29	-2.99
3	1.49	1.70	.80	3.25	1.39	3.11	1.83	2.87	1.23	2.74	1.90	2.70	2.08	4.76
4	1.97	2.78	2.75	2.50	-.21	1.31	1.25	1.93	2.09	2.28	.17	2.65	1.14	1.29
5	4.48	3.0176	1.83	.	.
6	2.89	.20	.	.

Tabel 10: Populatieschattingen, betrouwbaarheidsintervallen en pakkansen

		geschatte populatie	min. 95% CI	max. 95% CI	aantal opgepakte personen	pakkans
selectie	96	109688	5137	214238	1332	.01
	97	55202	23258	87146	1310	.02
	98	69686	27257	112116	1431	.02
	99	51058	13749	88367	1240	.02
	00	63070	0	126160	1022	.02
	01	92513	0	191503	958	.01
	96-97	104860	53900	155819	2619	.02
	97-98	75576	40558	110594	2707	.04
	98-99	62722	43973	81471	2638	.04
	99-00	51036	33908	68163	2235	.04
	00-01	74846	20787	128906	1669	.02
	96-98	114343	46400	182286	3996	.03
	97-99	70418	55715	85122	3895	.06
	98-00	68090	52326	83854	3607	.05
	99-01	65545	46976	84113	3158	.05
	96-99	88618	72643	104593	5169	.06
	97-00	78801	64349	93252	4848	.06
	98-01	75783	60841	90725	4517	.06
	96-00	96053	80373	111733	6108	.06
	97-01	87774	72690	102857	5746	.07
	landelijk	96	225897	134435	317359	5422
97		301018	141625	460412	5005	.02
98		262865	159795	365935	4919	.02
99		238279	130607	345951	4712	.02
00		328685	51447	605923	3904	.01
01		180290	107752	252827	3695	.02
96-97		270962	216729	325194	10267	.04
97-98		310771	223141	398400	9779	.03
98-99		279696	158449	400942	9483	.03
99-00		243075	188441	297709	8515	.04
00-01		286248	195525	376970	7531	.03
96-98		330549	274425	386674	14948	.05
97-99		308572	258158	358985	14265	.05
98-00		295547	245511	345583	13192	.04
99-01		279311	230391	328230	12059	.04
96-99		335264	294551	375978	19363	.06
97-00		334341	289949	378733	17902	.05
98-01	326669	279334	374003	16679	.05	
96-00	366907	324545	409270	22937	.06	
97-01	361132	319610	402654	21341	.06	

Tabel 11. Test op overdispersie

	selectie regio's		landelijk	
	chi-kwadraat	p	chi-kwadraat	p
96	2.79	.09	25.84	.00
97	1.13	.29	5.92	.02
98	.26	.61	1.44	.23
99	.32	.57	.00	.97
00	2.43	.12	.00	.96
01	1.37	.24	.30	.58
96-97	6.77	.01	46.15	.00
97-98	2.87	.09	21.71	.00
98-99	1.79	.18	11.31	.00
99-00	3.32	.07	18.42	.00
00-01	2.47	.12	3.30	.07
96-98	18.83	.00	45.27	.00
97-99	2.89	.09	18.53	.00
98-00	9.02	.00	29.75	.00
99-01	6.80	.01	22.16	.00
96-99	22.65	.00	54.79	.00
97-00	10.26	.00	39.60	.00
98-01	9.92	.00	25.14	.00
96-00	26.51	.00	77.21	.00
97-01	12.49	.00	44.00	.00

Figuur 12: Gemiddelde geschatte populatieomvang per observatieperiode

