Some ideas on teaching capture–recapture techniques in epidemiology and the social sciences

Cruyff, Maarten
Utrecht University, Department of Methodology and Statistics
P.O.Box 80.140
3508 TC Utrecht, the Netherlands
E-mail: m.cruyff@uu.nl

van der Heijden, Peter G.M.
Utrecht University, Department of Methodology and Statistics
E-mail: p.g.m.vanderheijden@uu.nl

Böhning, Dankmar
University of Reading, Applied Statistics, School of Biological Sciences
Reading, RG6 6FN, UK
E-mail: d.a.w.bohning@reading.ac.uk

1. Introduction

There are two aspects in teaching capture–recapture techniques in epidemiology and the social sciences that we would like to give more attention than they usually receive. The first aspect refers to the types of data that can be used for capture–recapture techniques, and the second refers to the fact that interest usually does not only go out to the population size estimate, but also to the composition of the observed and unobserved parts of the population.

Two types of data for capture–recapture techniques

The first aspect is that, in teaching, we would like to emphasize that there are two types of data where capture–recapture techniques can be fruitfully applied. Usually the focus is only on the type of data obtained from linking individuals in multiple registrations (compare Bishop, Fienberg and Holland, 1975; the International Working Group for Disease Monitoring and Forecasting, 1995; Chao et al., 2001). Consider three registrations A, B and C, neither of which providing a complete list of the population. Let the variables A, B and C have levels $0 = no$ and $1 = yes$ to denote whether a population member is included in the corresponding registration or not. By counting the number of individuals in each combination of A, B and C, a three-way contingency table $A \times B \times C$ can be constructed with a total of 8 cells. The sum of the counts in the 7 cells corresponding to individuals who are observed in at least one of the registrations equals n, while the count in the cell (0,0,0) is 0 by design, as it refers to the number of individuals that is missed by all registrations. The most important aim of capture–recapture techniques is to estimate the population size N, and this estimate is found as the sum of n and the estimated count for cell (0,0,0).

A second type of data that can be used for estimating a population size and that is also often available in epidemiology and the social sciences, is a single, but incomplete registration in which individuals may appear multiple times. In this respect one can think of a police registration of apprehensions for a specific traffic violation. Some traffic violators will appear multiple times in a specific time span while others will not be apprehended at all. Similarly, for a population of drug users one can think of a registration of contacts in a clinic. Some drug users will seek contact multiple times but others will never seek contact. The idea is that the registration can produce a list of individuals with the count (i.e. the number of times) that they appear in the registration. The frequency distribution of the counts is zero-truncated (drug users with zero contacts do not appear on the list) and the statistical problem is to estimate the frequency of the zero count. For this purpose the (truncated-at-zero) Poisson distribution is used to derive estimators, and these estimators can take covariate information of the individuals into account, (see van der Heijden et al., 2003a, 2003b; Böhning and van der Heijden, 2009; Cruyff and van der Heijden, 2009).

In discussions of population size estimation these single-registration techniques receive much less attention than multiple-registration techniques, but in the social sciences and epidemiology it is often difficult to find multiple lists that can be linked, while single registrations may be ready available. Thus including single registration estimates into the capture–recapture toolbox increases the range of possible applications of capture–recapture.

Composition of the population

Usually interest is not only in the size of a population, but also in its composition, or in other words, in a breakdown of the population size using a set of covariates. For single–registration techniques this can be easily accomplished as the key component of these techniques, a Poisson parameter, can be modeled as a function of covariates. For multiple–registration techniques this is less obvious. First, the contingency table built up of registrations can be stratified by covariates, and a separate subpopulation size estimate can be derived for each of the (cross-classified) levels of the covariate(s). However, in order to be able to do this for a covariate this covariate should be available in each of the registrations and it practice this condition is often not satisfied.

Recently, however, methodology has been proposed that allows to include covariates into the model that are not included in each of the registrations (Zwane and van der Heijden, 2007; Sutherland, Schwarz and Rivest, 2007; van der Heijden, Zwane and Hessen, 2009). In this methodology considers covariates not in one or more of the registrations as missing data, and these missing data are estimated with missing data methodology such as the EM algorithm. We think that this methodology also enlarges the range of possible applications of capture–recapture.

This manuscript has two sections, estimates from a single registration and estimates from multiple registrations. In each of these sections we will give special emphasis to models that allow for a description the composition of the population.

2. Estimates from a single registration

Overview

Registers can be used to generate a list of individuals from some population of interest. If each time that an observation of a population member occurs is registered but, for one reason or another, some population members are not observed at all, the list will be incomplete and will show only part of the population. In this section we consider the estimation of population size from one-source capture-recapture data, i.e. a register in which individuals can potentially be found repeatedly and where the question is how many individuals are missed by the register. As a typical example, consider a drug user study where the register consists of drug users who repeatedly contact treatment institutions. Drug users with 1, 2, 3, ... contacts occur, but drug users with zero contacts are not present, requiring the size of this group to be estimated. The register may have a record for every contact a drug user has with the institution. Summation of the individual contacts yields data in which each registered drug user has a single record with the count of the contacts, and some covariates. Statistically, the counts can be considered to come from a zero-truncated count distribution.

We discuss two estimators, namely a homogeneous Poisson estimator for the population size and an estimator for the population size suggested by Zelterman (1988) that is known to be robust under potential unobserved heterogeneity. We first discuss these estimators and then indicate how they can be adjusted so that they take into account covariate information.

Consider a population of size N and a count variable Y taking values in the set of integers

$\{0, 1, 2, 3, ...\}$. For example, in drug user studies Y might represent the number of contacts a drug user has with the treatment institutions. Also denote with $f_0, f_1, f_2, ...$ the frequency with which a $0, 1, 2, ...$ occurs in this population. Consider now a registration where every contact with a treatment institution is registered and assume that a list of drug users is derived from this registration. Since a drug user will only be observed if there has been a positive number of contacts with the treatment institution $y = 0$ will not be observed in the list. Hence the list reflects a count variable truncated at zero that we denote by Y_+. Accordingly, the list has observed frequencies $f_1, f_2, ...$, but the frequency f_0 of zeros in the population is unknown. The size of the list is not N but n, where $N = n + f_0$.

The distribution of the untruncated and truncated counts are connected via $P(Y_+ = j) = P(Y = j)/\{1 - P(Y = 0)\}$ for $j = 1, 2, ...$ For example, if Y follows a Poisson distribution with parameter λ so that

$$(1) \quad P(Y = j) = Po(j \mid \lambda) = e^{-\lambda}\lambda^j/j!,$$

for $j = 0, 1, 2, ...$, then the associated distribution of Y_+ is given as

$$(2) \quad P(Y_+ = j) = Po_+(j \mid \lambda) = \frac{e^{-\lambda}}{1 - e^{-\lambda}}\lambda^j/j!,$$

with $j = 1, 2, 3, ...$

Given that all units of the population have the same probability $P_i(Y > 0) = P(Y > 0) = 1 - P(Y = 0)$ of being included in the list, the population size can be estimated by means of the Horvitz–Thompson estimator

$$(3) \quad \hat{N} = \sum_{i=1}^{n} \frac{1}{P_i(Y > 0)} = \frac{n}{1 - P(Y = 0)} = \frac{n}{1 - g(\lambda)},$$

where $g(\lambda) = e^{-\lambda}$, or more generally, $g(\lambda)$ is the probability of a zero count for a given count distribution. For more details on this type of capture–recapture methodology see van der Heijden *et al.* (2003a,b), Böhning and Schön (2005), and Roberts and Brewer (2006).

In equation (3) we used the Horvitz–Thompson approach to arrive at an estimate of the population size. This approach requires that λ is known and if it is not, it needs to be estimated. Clearly, λ can be estimated with maximum likelihood under the assumption of a homogeneous truncated Poisson distribution. Instead of estimating λ under the assumption of a homogeneous Poisson distribution, Zelterman (1988) argued that the Poisson assumption might not be valid over the entire range of possible values for Y but it might be valid for small ranges of Y such as from j to $j + 1$, so that it would be meaningful to use only the frequencies f_j and f_{j+1} in estimating λ. Since for any j both the truncated as well as the untruncated Poisson distribution have the property that $Po(j + 1 \mid \lambda)/Po(j \mid \lambda) = \lambda/(j + 1)$ and $Po_+(j + 1 \mid \lambda)/Po_+(j \mid \lambda) = \lambda/(j + 1)$, respectively (see equations (1) and (2)), λ can be written as

$$(4) \quad \lambda = \frac{(j + 1)Po(j + 1 \mid \lambda)}{Po(j \mid \lambda)} = \frac{(j + 1)Po_+(j + 1 \mid \lambda)}{Po_+(j \mid \lambda)}.$$

An estimator for λ is obtained by replacing $Po_+(j \mid \lambda)$ by the empirical frequency f_j:

$$(5) \quad \hat{\lambda}_j = \frac{(j + 1)f_{j+1}}{f_j}.$$

If $j = 1$ we find $\hat{\lambda}_1 = 2f_2/f_1$, and this estimator is often considered for two reasons: for one, $\hat{\lambda}_1$ is using frequencies in the vicinity of f_0 which is the target of prediction, and two, in many application studies for estimating f_0 the majority of counts fall into f_1 and f_2. Clearly, the estimator is *unaffected* by changes in the data for counts larger than 2 which contributes largely to its robustness. We will call $\hat{\lambda}_1 = 2f_2/f_1$ the *Zelterman estimator for* λ and, when this estimate is used in (3), this leads to the *Zelterman estimator of the population size,* $\hat{N}$. If the context is clear we will simply use the term *Zelterman estimator.*

The Zelterman estimator is simple to understand and to use and this might be one of the reasons why it is quite popular in applications such as drug user studies (Hay and Smit 2003; van Hest et al., 2007). It is also thought of being less sensitive to model violations than the estimator that is derived under the assumption of the homogeneous Poisson distribution, that uses the entire range of frequencies f_j. Indeed, the Zelterman estimator also works rather well with contaminated distributions as given by mixtures or approximated by mixtures (compare Zelterman, 1988).

Composition of the population

In most applications with hidden populations, the assumption of a homogeneous truncated Poisson distribution is not realistic since it implies that every individual has the same Poisson parameter. Often, the register also includes some information about the individuals' characteristics. This information can be used to allow for (observed) heterogeneity in the Poisson parameters. For the homogeneous Poisson estimator we use truncated Poisson regression to adjust for covariate information for each individual separately (see van der Heijden et al., 2003a,b). The covariate information is incorporated in the model by specifying the link function $\ln \lambda_i = \boldsymbol{x_i'\beta}$, where $\boldsymbol{x_i}$ is a vector with covariate values including a constant, and $\boldsymbol{\beta}$ is the corresponding parameter vector. See van der Heijden et al. (2003a,b) for details.

Similarly, Böhning and van der Heijden (2008) showed how the Zelterman estimator can be adjusted so that it takes covariate information into account, thus arriving at a Poisson parameter estimate for every individual separately. Specifically, they demonstrated that the Zelterman estimator can be viewed as a maximum likelihood estimator for a locally truncated Poisson likelihood which is equivalent to a binomial likelihood.

Thus both in the truncated Poisson regression model as in the Zelterman regression model we obtain an estimate λ_i for individual i separately. Using this estimate we can find for every individual separately an estimated probability of being observed, and using the Horvitz-Thompson approach we obtain for every individual separately the number of comparable individuals that is *not* observed. For example, when an individual had estimated probability .25 of being observed, then for this individual there are three comparable individuals that are not observed (where 'comparable' is defined in terms of the values on the covariates). This allows to describe the population in terms of the covariates.

Example

We illustrate this with an example of an analysis with the truncated Poisson regression model taken from van der Heijden et al. (2003b). We discuss the estimation of the number of illegal immigrants in the Netherlands from police records. These records contain information on the number of times each illegal immigrant was apprehended by the police and they are incomplete since the illegal immigrants who were never apprehended do not appear in them. For the estimation of the number of illegal immigrants in the Netherlands, police records are available for 1995, for four cities in the Netherlands: Amsterdam, Rotterdam, The Hague and Utrecht. The records are used to derive count data on how often each illegal immigrant is apprehended by the police, and included the following covariate: age, gender, country and reason for being apprehended. To give some insight in the data, we present the apprehension frequencies for each of the levels of the covariates in Table 1. For more details on these data we refer to van der Heijden et al. (2003b).

For the zero-truncated Poisson regression analysis on the full model using all available covariates the population size estimate is 12,691, with a 95 percent confidence interval of 7,185-18,198. Table 2

Table 1: Illegal immigrants not effectively expelled. Observed frequencies for the covariate categories)

Covariate category	f_1	f_2	f_3	f_4	f_5	f_6	Total
>40 years	105	6					111
<40 years	1540	177	37	13	1	1	1769
female	366	24	6	1	1		398
male	1279	159	31	12		1	1482
Turkey	90	3					93
North Africa	838	146	28	9	1	1	1023
Rest Africa	229	11	3				243
Surinam	63	1					64
Asia	272	9	1	2			284
America, Australia	153	13	5	2			173
Being illegal	224	29	5	1			259
Other reason	1421	154	32	12	1	1	1621

Table 2: Parameter estimates of the truncated Poisson regression model

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

Log-likelihood$= -848.448$

* P-value for Wald test

shows the maximum likelihood estimates of the regression parameters together with their corresponding standard errors and P-values. The variables Gender, Age and Nationality (Turkey, Rest of Africa, Surinam or Asia) contribute significantly to the average number of times an individual is apprehended by the police. The results show that male individuals and individuals who are less than 40 years of age are, on the average, more frequently apprehended by the police. Individuals from Turkey, rest of Africa, Surinam and Asia are less frequently apprehended than those from America and Australia. The variable Reason for being apprehended appears to have no impact on the average number of times an individual is apprehended by the police.

For the purpose of model selection, we fitted several truncated Poisson regression models. The results are shown in Table 3. The null model yields the lowest estimate of the total number of illegal immigrants ($\hat{N} = 7080$). The corresponding 95% Horvitz-Thompson confidence interval is (6,363 - 7,797). The largest estimate of N, $\hat{N} = 12,691$, is obtained by fitting the full model of Table 3. These estimates illustrate a theoretical result that, in a sequence of nested models, the more covariates that are added to the model the higher the point estimate of N is expected to become.

In order to compare the various models we also computed AIC-values and performed likelihood-

Table 3: Estimates $\hat{N}$ and HT 95% confidence intervals for N obtained from fitting different truncated Poisson regression models. Model comparisons using the likelihood-ratio test and AIC-criterion are also given. $\chi^2_{(1)}$ is the Lagrange multiplier test testing for overdispersion

Model	AIC	G^2	df	P^*	$\chi^2_{(1)}$	$\hat{N}$	C.I.
Null	1805.9				106.0	7080	6363-7797
G	1798.3	9.6	1	.002	99.7	7319	6504-8134
G+A	1789.0	11.2	1	<.001	93.7	7807	6637-8976
G+A+N	1712.9	86.1	5	<.001	55.0	12690	7186-18194
G+A+N+R	1714.9	.004	1	.949	55.0	12691	7185-18198

* P-value for likelihood-ratio test.

ratio tests for the models in (in Table 3. The likelihood-ratio test in Table 3 shows that the variable Reason for being caught can be dropped from the full model ($G^2 = .004$, df = 1, $P = .949$). From the resulting model (G+A+N) the variable Nationality cannot be dropped ($G^2 = 86.1$, df = 5, $P < .001$), nor can the variables Gender and Age (not shown here). Since the AIC-criterion also favors this model and our choice of $\hat{N}$ should be based on the best fitting model, our best estimate seems to be that of the model (G+A+N), $\hat{N} = 12,690$. When models are misspecified (e.g. the null model and the models in the 2nd and the 3rd row of Table 3) their results, including the value of $\hat{N}$, should not be interpreted.

Table 4: Observed and estimated counts for illegal immigrants for model (G+A+N).

k	observed	estimated	residuals
0	0	10,810.4	
1	1,645	1,612.6	0.81
2	183	233.7	-3.32
3	37	30.1	1.25
4	13	3.2	5.42
5	1	0.3	1.31
6	1	0.0	6.57

A way of examining the goodness of fit of a model is to compare the observed and the estimated frequencies by looking at the Pearson residuals, as presented in Table 4. The residuals for $k = 2$, $k = 4$ and $k = 6$ seem rather large, indicating some lack of fit. The Lagrange multiplier test of Gurmu (1991) (see Section 5) suggests that there still remains some unobserved heterogeneity that cannot be ignored ($\chi^2 = 55.0$, df = 1). Therefore we must conclude that the population size estimate $\hat{N}$ = 12,690 should be interpreted as an underestimate of the true population size.

It is also possible to make comparisons between observed and estimated number of individuals for subgroups in the data. This illustrates that the composition of the population can be studied. Table 5 shows such comparisons based on the model fit of model (G+A+N). 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 are not 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 and individuals from North Africa have a larger probability to be apprehended, a confirmation of what was observed in Table 2.

Table 5: Comparisons between observed and estimated N for subgroups based on model (G+A+N)

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

3. Estimates from a multiple registrations that are linked

Overview

A well known technique for estimating the size of a human population is to find two or more registrations of this population, to link the individuals in the registrations and estimate the number of individuals that occur in neither of the registrations (Fienberg, 1972; Bishop, Fienberg and Holland, 1975; Cormack, 1989; IWGfDMaF, 1995). For example, with two registrations A and B, linkage gives a count of individuals in A but not in B, a count of individuals in B but not in A, and a count of individuals both in A and B. The counts form a contingency table denoted by $A \times B$ with the variable labeled A being short for 'inclusion in registration A', taking the levels 'yes' and 'no', and likewise for registration B. In this table the cell 'no,no' has a zero count by definition, and the statistical problem is to estimate its value in the population. A population size estimate is obtained by adding this estimated count of missed individuals to the counts of individuals found in at least one of the registrations.

With two registrations the usual assumptions under which a population size estimate is obtained are: inclusion in registration A is independent of inclusion in registration B; and in at least one of the two registrations the inclusion probabilities are homogeneous (see Chao et al., 2001 and Zwane et al., 2004). Interestingly it is often, but incorrectly, supposed that *both* inclusion probabilities have to be homogeneous. Other assumptions are that the population is closed and that it is possible to link the individuals in registrations A and B perfectly.

It is generally agreed that these assumptions are unlikely to hold in human populations, but there are three approaches that may be adopted to make the impact of possible violations less severe. One approach is to include covariates into the model, in particular covariates whose levels have heterogeneous inclusion probabilities for both registrations (see Bishop, Fienberg and Holland, 1975; Baker, 1990; compare Pollock, 2002), so that loglinear models can be specified for the higher-way contingency table of registrations A and B and the covariates. The restrictive independence assumption is replaced by a less restrictive assumption of independence of A and B conditional on the covariates; and subpopulation size estimates are derived (one for every level of the covariates) that add up to a population size estimate. Another approach is to include a third registration, and to analyze the three-way contingency table with loglinear models that may include one or more two-factor interactions, thus getting rid of the independence assumption. Here the (less stringent) assumption made is that the three-factor interaction is absent. However, including a third registration is not always possible, as it is not available, or because there is no information that makes it possible to link the individuals in the third registration to both the first and to the second registration. A third approach makes use of a latent variable to take heterogeneity of inclusion probabilities into account (see Fienberg, Johnson and Junker, 1999; Bartolucci and Forcina, 2004). Of course, these three approaches are not exclusive and may be used concurrently in one model.

Composition of the population

Inclusion of covariates into the model allows to get insight into the composition of the population in terms of these covariates, and this one of the aspects that we would like to emphasize in this paper. When the approach is adopted to use covariates, the question is which covariates should be chosen. In the traditional approach, only covariates that are available in both registrations can be chosen. Recently, Zwane and van der Heijden (2007) showed that it is also possible to use covariates that are only available in a subset of the registrations. For example, when a covariate is available in registration A but not in B, the values of the covariate in B are estimated under a missing-at-random assumption (Little and Rubin, 1987); and the subpopulation size estimates are then derived as a by-product.

A simple example illustrates the problem, see Panel 1 of Table 6. It concerns people with Afghan, Iranian or Iraqi nationality being registered in the official registration GBA and in the police registration HKS in 2007. The aim is to estimate the size of the population and the size of the subpopulations in terms of marital status and police region of apprehension. Covariate X_1 (Marital status) is only observed in registration A (GBA) and covariate X_2 (Police region) is only observed in registration B (HKS). As a result X_1 is missing for those observations not in A and X_2 is missing for those observations not in B. Zwane and van der Heijden (2007) show that the missing observations can be estimated using the EM algorithm under a missing-at-random (MAR) assumption (Little and Rubin, 1987, Schafer, 1997) for the missing data process. After EM, in a second step, the population size estimates are obtained for each of the levels of X_1 and X_2. The number of observed cells is lower than in the standard situation. For example, in Panel 1 of Table 6 this number is 8, whereas it would have been 12 if both X_1 and X_2 were observed in both A and B. For this reason only a restricted set of loglinear models can be fit to the observed data. Zwane and van der Heijden (2007) show that the most complicated model is $[AX_2][BX_1][X_1X_2]$, where we use the notation of loglinear models proposed by Bishop et al. (1975). At first sight this model appears counter-intuitive as one might expect an interaction between variables A and X_1, and between B and X_2. However, the parameter for the interaction between A and X_1 (and B and X_2) cannot be identified as the levels of X_1 do not vary over individuals for which $A = 2$.

The loglinear model $[AX_2][BX_1][X_1X_2]$ is the saturated model, since the number of parameters is 8 (the general mean, four parameters for the main effects and three interaction parameters) add up to the 8 observed values. Consequently the 8 observed values equal the corresponding 8 fitted values. The fitted values under this model are presented in Panel 2 of Table 6. Note that the EM algorithm spreads out the missing values over the levels of the corresponding covariate. For example, the observed value $13,898$ is divided over the levels of X_1 into fitted values $4,510.8$ and $9,387.2$; note also that the fitted values ratio $4,510.8/9,387.2$ is identical to the observed values ratio $259/539$.

By comparison, when X_1 and X_2 are observed in both A and B, the saturated model is $[AX_1X_2][BX_1X_2]$. This is a less restrictive model than the model $[AX_2][BX_1][X_1X_2]$, and the difference is due to the MAR assumption.

Concluding remarks

The missing data methodology that we just discussed allows to breakdown the population size in terms of covariates. In Panel 2 of Table 6 the population size is broken down over the variables A (official registration), B (police registration), X_1 (marital status) and X_2 (police region). This four-way array can be marginalized in different ways, depending on the research question. The model has three interaction parameters, related to the margin $A \times X_2$, that may show that there is a relation between

Table 6: *Covariate X_1 is only observed in registrations A and X_2 is only observed in B*

Panel 1: Observed counts

		$A = 1$		$A = 2$
		$X_1 = 1$	$X_1 = 2$	X_1 missing
$B = 1$	$X_2 = 1$	259	539	13,898
	$X_2 = 2$	110	177	12,356
$B = 2$	X_2 missing	91	164	-

Panel 2: Fitted values under $[AX_2][BX_1][X_1X_2]$

		$A = 1$		$A = 2$	
		$X_1 = 1$	$X_1 = 2$	$X_1 = 1$	$X_1 = 2$
$B = 1$	$X_2 = 1$	259.0	539.0	4,510.8	9,387.2
	$X_2 = 2$	110.0	177.0	4,735.8	7,620.3
$B = 2$	$X_2 = 1$	63.9	123.5	1,112.4	2,150.2
	$X_2 = 2$	27.1	40.5	1,167.9	1,745.4

being apprehended in a certain police region and being registered in the official registration GBA; the margin $B \times X_1$, that may show that there is a relation between marital status and being apprehended by the police; and the margin $X_1 \times X_2$ that may show that there is a relation between marital status and police region. Of course, other margins can also be studied and these are not necessarily independent because marginalizing a conditional independence relation may lead to marginal dependence.

The missing data methodology is extended to more registrations and covariates, in Zwane and van der Heijden (2007). It makes clear that more covariates can be taken into account than considered in the traditional approach, be it under MAR assumptions. It may provide a useful description of the composition of the population, in addition to the estimate of the population size.

REFERENCES

Baker, S.G. (1990). A simple EM algorithm for capture-recapture data with categorical covariates (with discussion). Biometrics, 46, 1193-1197.

Bartolucci, F. and Forcina, A. (2004). Analysis of Capture-Recapture Data with a Rasch-Type Model Allowing for Conditional Dependence and Multidimensionality. Biometrics, 57, 714-719.

Bishop, Y.M.M. and Fienberg, S.E. and Holland, P.W. (1975). Discrete Multivariate Analysis, Theory and Practice, New York: McGraw-Hill.

Böhning, D, and Schön, D. (2005). Nonparametric maximum likelihood estimation of the population size based upon the counting distribution. J. Roy. Statist. Soc. Ser. C, 54, 721737.

Böhning, D. And P.G.M. van der Heijden (2009). A Covariate Adjustment for Zero-truncated Approaches to Estimating the Size of Hidden and Elusive Populations. Annals of Applied Statistics, 3, 595-610.

Chao, A., Tsay, P., Lin, S., Shau, W. and Chao, D. (2001). The applications of capture-recapture models to epidemiological data. Statistics in Medicine, 20, 3123-3157.

Cormack, R. (1989). Log-linear models for capture-recapture, Biometrics, 45, 395-413.

Cruyff, M.J.L.F. and P.G.M. van der Heijden. (2008). Point and interval estimation of the population size using a zero-truncated negative binomial regression model. Biometrical Journal, 50 (6), 1035-1050.

Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. Biometrika, 59, 409-439.

Fienberg, S., Johnson, M. and Junker, B. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. J. R. Stat. Soc. Ser. A, 162, 383-406.

Hay, G. and Smit, F. (2003). Estimating the number of drug injectors from needle exchange data. Addiction Research and Theory 11 235243.

International Working Group for Disease Monitoring and Forecasting (1995). Capture-recapture and multiple record systems estimation. Part I. History and theoretical development. American Journal of Epidemiology, 142, 1059-1068.

Little, R. and Rubin, D. (1987). Statistical analysis with missing data. New York, J. Wiley & Sons.

Pollock, K. H. (2002). The use of auxiliary variables in capture-recapture modelling: an overview. Journal of Applied Statistics, 29, 85-102.

Roberts, J. M. and Brewer, D. D. (2006). Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capturerecapture method. J. Roy. Statist. Soc. Ser. A 169 745756.

Schafer, J. (1997). Analysis of Incomplete Multivariate Data. New York, Chapman & Hall/CRC.

Sutherland, J.M. and Schwarz, C.J. and Rivest, L.–P. (2007). Multilist population estimation with incomplete and partial stratification. Biometrics, 63, 910-916.

Van der Heijden, P.G.M., M. Cruyff and H. van Houwelingen. (2003a) Estimating the size of a criminal population from police registrations using the truncated Poisson regression model. Statistica Neerlandica, 57, 289-304.

van der Heijden, P.G.M., Bustami, R., M. Cruyff, G. Engbersen and H. van Houwelingen (2003b). Point and interval estimation of the truncated Poisson regression model. Statistical Modelling, 3, 305-322.

Van der Heijden, P.G.M., E. Zwane and D. Hessen (2009). Structurally missing data problems in multiple list capture-recapture data. Advances of Statistical Analysis, 93, 5-21.

Van der Heijden, P.G.M., J. Whittaker, M. Cruyff, B. Bakker and R. van der Vliet (submitted). Invariant population size estimattes and the role of active and passive covariates.

Van Hest, N. H. A., Grant, A. D., Smit, F., Story, A. and Richardus, J. H. (2007). Estimating infectious diseases incidence: Validity of capturerecapture analysis and truncated models for incomplete count data. Epidemiology and Infection 136 1422.

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

Zwane, E.N., K. van der Pal-de Bruin and P.G.M. van der Heijden. (2004) The multiple records system estimator when registrations partly overlap in time and by region. Statistics in Medicine, 23, 2267-2281.