AStA Adv Stat Anal (2009) 93: 5–21
DOI 10.1007/s10182-008-0098-6

ORIGINAL PAPER

Structurally missing data problems in multiple list capture–recapture data

Peter G.M. van der Heijden · Eugene Zwane · David Hessen

Received: 19 March 2008 / Accepted: 30 August 2008 / Published online: 18 December 2008

Abstract Multiple-list capture–recapture data can be used to estimate the size of a population. In this manuscript two problems are studied and solved using a common solution. The first problem is that the lists refer to different but overlapping populations. An example is that lists refer to different but overlapping regions, different but overlapping periods in time, or different but overlapping age groups. The second problem is that each list has a set of covariates and the sets of covariates are not identical. By considering both problems as missing data problems, a solution is obtained through the EM algorithm. This approach is illustrated by two examples.

Keywords Capture–recapture · Covariate · Missing data

1 Introduction

One of the ways to estimate the size of an unknown population is by using capture–recapture methods. In epidemiology these methods use two or more lists (or registrations) with individuals. The lists of individuals are linked, and the overlap is studied through the construction of a contingency table. If there are S lists, a contingency table is formed with 2^S cells. For example, if there are three lists, the count in cell

P.G.M. van der Heijden (✉) · D. Hessen
Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, 3508 TC Utrecht, The Netherlands
e-mail: p.g.m.vanderheijden@uu.nl

D. Hessen
e-mail: d.hessen@uu.nl

E. Zwane
Infectious Disease Epidemiology Unit, School of Public Health and Family Medicine, Faculty of Health Sciences, University of Cape Town, Observatory 7925, South Africa
e-mail: eugene.zwane@uct.ac.za

$(1, 0, 1)$ represents the number of individuals that appear in lists 1 and 3 but not in list 2. By definition, the number of individuals in cell $(0, 0, 0)$ is unknown, and the statistical problem is to estimate the population size. For reviews of this research area, see the International Working Group for Disease Monitoring and Forecasting (IWGDMF) (1995), Chao et al. (2001), and Tsay and Chao (2001).

A popular statistical method that is often used for the estimation of the count in cell $(0, 0, 0)$ is the loglinear model (see, for example, Fienberg 1972; Bishop et al. 1975; Cormack 1989). Typically a loglinear model is estimated for the contingency table where the cell $(0, 0, 0)$ is treated as a structural zero. Once the model is found that describes the counts in the cell adequately, the parameter estimates of this model are projected onto cell $(0, 0, 0)$, yielding an estimate of the number of individuals missed by all lists. By adding up this number to the number of individuals observed at least once a population size estimate is obtained.

The loglinear model may reveal different inclusion probabilities for each of the lists, but also—if the number of lists is at least three—dependencies between the lists. It may well be that the inclusion probabilities and dependencies differ for covariates, and if such covariates are available, a loglinear model should take them into account. It is well known that list dependence may be caused by heterogeneity of inclusion probabilities (see, for example, IWGDMF 1995), and therefore it is generally advisable to include covariates into the loglinear model.

However, usually two problems are encountered in taking covariates into account. First, it may be that there are covariates that are related to the definition of the population. For example, it may be that list 1 is listing individuals in regions 1 and 2, whereas list 2 is listing individuals in region 2 and 3. Thus list 1 refers to a population that is different from the population list 2 refers to. Using standard methods, only the population in region 2 could be estimated, as individuals living in region 1 being in list 1 cannot be linked to any individual in list 2, and similarly for the individuals living in region 3 that are listed in list 2. Many other examples can be given, such as lists being operative in different but overlapping time periods, or lists being operative for different but overlapping age groups, or one list being operative for females whereas the other is operative for males as well as females, and so on. What they have in common is that the lists are referring to different but overlapping populations, and covariates are available that make it possible to separate these different but overlapping populations.

A second problem that may be encountered is that the lists refer to the same population, but that the covariates that are available differ over the lists. For example, a set of lists deal with both males and females, but for one of the lists, gender is not recorded. So there are covariates available in all of the lists, and these covariates are used, for example, in linking the individuals over the different lists. But as different lists may be founded with different purposes, it is only natural that the covariates available in the lists are not identical.

Until recently both problems have not been dealt with in an optimal way. That is, for the first problem, if the lists referred to different but overlapping populations, using standard methods, only the overlapping part of the population was estimated; the second problem was solved using standard methods by simply ignoring the covariates that did not appear in each of the lists.

Recently both problems were solved in separate papers but with a common solution (see Zwane et al. 2004; Sutherland et al. 2007; Zwane and Van der Heijden 2007). The common solution to both problems is that there is a missing data problem. In the first problem, when, for example, list 1 is listing individuals in regions 1 and 2, whereas list 2 is listing individuals in region 2 and 3, the missing data problem is that region 3 is missing for list 1 and region 1 is missing for list 2. For the second problem, when a set of lists deal with both males and females, but for one of the lists, females are not recorded, females are considered missing for this latter list.

This missing data problem is solved by estimating the missing data using the Expectation–Maximization (EM) algorithm (Little and Rubin 1987). It can be shown that the assumptions that are made in applying the EM algorithm are less stringent than those that are made with the usual approach to ignore part of the data.

In this manuscript we summarize the work of Zwane et al. (2004) and Sutherland et al. (2007). We present both problems in one framework and discuss several practical issues that may be encountered in their application. We present two applications to illustrate their work. One application deals with a capture–recapture problem where five lists are defined over different but overlapping time periods. Another application deals with a capture–recapture problem where two lists have different sets of covariates.

2 Theory

To motivate the problem, we will first provide the two typical problems that we sketched in the introduction in more detail.

Problem 1: lists refer to different but overlapping sub-populations The first problem is that lists refer to different but overlapping sub-populations. A first example is that sub-populations are defined regionally: lists refer to different but overlapping regions. For example, if there are two lists, it may be that list 1 is observed in both region 1 and region 2 but list 2 is only observed in region 2. Here a sub-population estimate for region 1 and for region 2 may be interesting. Another, slightly more complicated example, is that list 1 is observed in regions 1 and 2, whereas list 2 is observed in regions 2 and 3. Here estimates for each of the three sub-populations may be of interest. Another example that is often encountered in practice is that sub-populations are defined in time: lists are often built up in different time periods, for example, list 1 is referring to individuals observed in time periods 1 and 2, whereas for list 2, only observations for period 2 are available. Here estimates for the sub-populations in period 1 and in period 2 may be of interest. We refer to Zwane et al. (2004), who discuss an example of incidence of spina bifida where five lists cover different time periods.

We will consider one example in more detail: the example is that there are two lists, where in region 1 lists 1 and 2 are available, and in region 2 only list 1 is available.

Let us assume first the standard situation for this example, namely that both lists are observed in both regions. After linking the two lists, Table 1 can be formed. As the elements a and e are not observed, the statistical problem is here to estimate them. This will then lead to a population size estimate for region 1 and for region 2.

Table 1 Multiple list capture–recapture data in two regions

Region	List 1	List 2	
		Not included	Included
1	Not included	a	b
	Included	c	d
2	Not included	e	f
	Included	g	h

The standard approach to do this is using hierarchical loglinear models where the unobserved elements a and e are treated as structural zeros. If we denote loglinear models by placing the variables defining the highest fitted margins between brackets, the most complicated loglinear model is [1R][2R]. We call model [1R][2R] the *maximal model*, since for this model, the fitted values are equal to the observed frequencies. The term [1R] shows that region is related to the inclusion probability of list 1 (i.e., the inclusion probability for list 1 in region 1 may be different from the inclusion probability for list 1 in region 2), and the term [2R] shows that region is related to the inclusion probability of list 2. This notation also shows that the inclusion probabilities of list 1 and list 2 are independent in each region. In this standard approach more restrictive loglinear models can be fitted, for example, [1][2R], where the inclusion probabilities of list 1 are homogeneous over regions 1 and 2, but the inclusion probabilities of list 2 may be different over regions 1 and 2.

Consider now the situation that in region 1 there are two lists available and in region 2 only list 1 is available. In this situation elements a, e, and f are not observed, and the sum $g+h$ is observed. Now in order to arrive at population size estimates for region 1 and region 2, there are two statistical problems, namely first to disentangle the sum $g+h$ into separate elements g and h, and, second, to estimate the unobserved elements a, e, and f.

Interestingly, if one is not interested in the separate estimates for regions 1 and 2 but rather in the sum of the estimates for regions 1 and 2, then an estimate of this population size may be obtained by ignoring region. Zwane et al. (2004) prove that two assumptions need to be fulfilled for this estimate to be unbiased: first, independence of lists 1 and 2 in each of the regions and, second, that the inclusion probability of the list of observed in both regions, i.e., list 1, is homogeneous over the two regions. Though this result may be interesting and useful, it does not lead to separate estimates for regions 1 and 2, and this may regularly be of interest.

Problem 2: lists have different (sets of) covariates We now consider the second typical problem. Often the lists used in a multiple-list capture–recapture problem have one or more covariates in common, namely those covariates that were used for uniquely linking the records in the list, such as age, gender, city of birth, and address. However, there may also be covariates that are unique for a specific list. For example, as we will see in an example that we will discuss later in this manuscript, when civil information collected by local government offices is linked to criminal information, the local government office may have detailed information about the marital status and age of father and mother of an individual, whereas this information may not

Table 2 Lists 1 and 2 with covariates A and B

List 1	Covariate A	List 2			
		Covariate B			
		Not included		Included	
		B_1	B_2	B_1	B_2
Not included	A_1	a	b	c	d
	A_2	e	f	g	h
	A_3	i	j	k	l
Included	A_1	m	n	o	p
	A_2	q	r	s	t
	A_3	u	v	w	x

be available in the list for criminal information. Similarly, in the list for criminal information we may find information about the criminal history of an individual that will most likely not be available in the list of the local government office.

Usually in multiple-list capture–recapture problems only the covariates that both lists have in common are used in an analysis, but covariates that are unique for a list are ignored. Zwane and Van der Heijden (2007) have shown that this may lead to biased estimates of the population size. Second, such an approach of ignoring unique covariates makes it impossible to relate these covariates to the population size estimates.

We will now illustrate this problem. The example is that there are two lists, list 1 and list 2, and covariate A with levels A_1, A_2, and A_3 and covariate B with levels B_1 and B_2. The data can be collected in a table such as Table 2.

In this situation the elements a, b, e, f, i, and j are not observed and have to be estimated. If both covariates A and B are collected in lists 1 and 2, then the traditional approach of using loglinear models can again be employed for estimation the population size. Here the most complicated loglinear model that can be fit is [AB1][AB2]; again we use the phrase *maximal model* for [AB1][AB2] since for this model, the observed frequencies are equal to the fitted counts and no further parameters can be added to the model as it would make the model unidentified. In the maximal model the assumption that has to be made is independence of lists 1 and list 2 given A and B, so that a is estimated using c, m, and o, and similarly for the other unobserved elements. In the maximal model [AB1][AB2] the inclusion probability for list 1 and the inclusion probability for list 2 are functions of A and B jointly, and more restrictive loglinear models can be tried to investigate whether the inclusion probabilities are, for example, only functions of main effects of A and B.

Assume now that covariate A is only available in list 1 and covariate B is only available in list 2. Then problems become more complicated. First, as before, the elements a, b, e, f, i, and j are not observed. Second, due to the fact that A is only available in list 1 and not in list 2, the levels of A are unknown in the situation that list 2 is observed but list 1 is not unobserved; as a result, we do not know the elements c, d, g, h, k, and l, but we only know the sums $(c + g + k)$ and $(d + h + l)$. Third, due to the fact that B is only available in list 2 and not in list 1, the levels of B are unknown in the situation that list 1 is observed but list 2 is not unobserved; as a result,

we do not know the elements m, n, q, s, u, and v, but we only know the sums $(m+n)$, $(q+r)$, and $(u+v)$.

Now in order to arrive at population size estimates for list 1 and list 2 stratified for A and B, there are two statistical problems, namely first to disentangle the sums $(c+g+k)$, $(d+h+l)$, $(m+n)$, $(q+r)$, and $(u+v)$ into separate elements $c, g, k, d, h, l, m, n, q, r, u$, and v, and, second, to estimate the unobserved elements a, b, e, f, i, and j.

Solution to the problems What both problems have in common is that both can be considered as missing data problems. A general solution that can be applied to these missing data problems is the EM algorithm (Little and Rubin 1987) proposed in this context by Zwane et al. (2004), Sutherland et al. (2007), and Zwane and Van der Heijden (2007). The EM algorithm can be used if the missing data are missing at random (MAR; Little and Rubin 1987), that is, the probability of missingness depends only on the observed data. This is often a reasonable assumption here, because the data are missing by design. As a result, the missingness provides no information about the underlying process, implying that the missing data mechanism is ignorable.

Due to the similarities between the missing covariate problem and the problem where some lists only operate in a sub-population of the full population, they can both be tackled as one problem, and the EM algorithm can be used to obtain maximum likelihood estimates. The EM algorithm comprises iterations of pairs of steps. In the E-step, the contributions of the missing data to the cell probabilities (sufficient statistics) are estimated, and in the M-step the complete-data analysis is applied, with the contributions estimated in the previous E-step in place of their unknown complete-data values. Since the E-step depends on some of the parameters estimated in the M-step, iterations are necessary. In the capture–recapture problem with categorical covariates the EM-algorithm is practical since both the E- and M-steps are simple. The algorithm is iterated until it converges. After convergence the parameter estimates are used to find point estimates for the structurally zero cells and an estimate of the population size.

Interestingly, for both typical problems that we discussed above, the *maximal models* become more restrictive. Thus far this was only discussed for the models in problem 2, where lists have different (sets of) covariates (see Zwane and Van der Heijden 2007), but this also holds for the models in problem 1 in which lists refer to different but overlapping sub-populations.

Maximal models when lists have different (sets of) covariates We first summarize the results of Zwane and Van der Heijden (2007) for the situation of two lists (for the situation of more than two lists, we refer to their paper). In general, there can be three types of covariates, namely (i) covariates collected in A that only appear in list 1; (ii) covariates collected in B that only appear in list 2; and (iii) covariates collected in C that appear in both lists 1 and 2. The *maximal model* is [1BC][2AC][ABC].

The most complex log-linear model (*maximal* model) that can be fitted to these data is the log-linear model given by [1BC][2AC][ABC]. The maximal model does not include the interactions between 1 and A, and between 2 and B, due to that A exists only when 1 is observed and B exists only when 2 is observed, resulting in

Table 3 Summary of modeling decision for dual list problems

Situation	Model	Decision
A, B, C present	[1BC][2AC][ABC]	Ignore A and B if conditionally independent
No A	[1BC][2C]	Ignore B
No A	[1BC][2C]	Ignore B
No B	[1C][2AC]	Ignore A
No C	[1B][2A][AB]	Ignore A and B if independent
No A and C	[1B][2]	Ignore B
No B and C	[1][2A]	Ignore A
No A, B, C	[1][2]	

them being inestimable. In other words, the MAR assumption is that 1 is not directly related to A, and that 2 is not directly related to B, but indirect relations may exist and will go over the three-factor interaction between A, B, and C.

Simplified situations exist when A, B, or C are not available, and in particular when A and B are available but (conditionally) independent given C. We summarize these results in Table 3 and refer for proofs to Zwane and Van der Heijden (2007).

First, when A, B, and C are available but A and B are independent conditional on C, then the sum of the population size estimates for the full table of 1, 2, A, B, and C equals the sum of the population size estimates for the marginal table of 1, 2, and C. We make a few remarks.

1. If A, B, and C are available but the current practice is followed to ignore A and B, then this will only result in an unbiased population size estimate when in the population A and B are independent given C.
2. Even if A, B, and C are available *and* A and B are independent given C, then it may still be worthwhile to apply the EM-approach because the EM-approach will yield estimates for every combination of levels of A, B, and C.
3. If A, B, and C are available, one approach to find an adequate model is to start with the maximal model [1BC][2AC][ABC], since here the fitted values are equal to the observed counts (likelihood ratio chi-square will be 0 with 0 df). Subsequently interactions may be dropped if this will not significantly deteriorate the fit. The maximal model with an additional conditional independence assumption between A and B given C is equivalent to loglinear model [1BC][2AC]. Therefore, model [1BC][2AC] and all hierarchical loglinear models that are nested in this model by additional parameter restrictions will yield the same sum of the population size estimates found for every combination of levels of A, B, and C.
4. When there are no variables in A, the variables in B can be ignored if interest only goes out into the sum of the population size estimates; in the same way, when there are no variables in B, the variables in A may be ignored.

Maximal models when lists refer to different but overlapping sub-populations In the typical example that we discussed above there are two lists and two regions; in region 1 both lists 1 and 2 are observed, but in region 2 only list 1 is observed. If both lists were observed in both regions, the maximal model would be [1R][2R]. We

note that there are six frequencies and six parameters. Now assume that in region 2 only list 1 is available. See Table 1. Then elements a, e, and f are not observed, and the elements b, c, d and the sum $g + h$ are observed. Thus there are four observed frequencies, and therefore the maximal model can only have four parameters. These are: the general mean and main effects for list 1, for list 2, and for region.

Finding the maximal model is not always easy, but as a rule of thumb it may be helpful to note that no models can be estimated containing interactions for which there are no corresponding marginal frequencies (sufficient statistics). For example, in Sect. 4 we find an example where certain interaction parameters of a list with time cannot be estimated for those years that the list was not available.

3 Example 1: lists refer to different but overlapping populations

As a first example, we will introduce the data set on neural tube defects (NTDs) in the Netherlands that will be used to illustrate the procedure presented in the paper. It deals with NTD-registrations that are active over different periods. For details about the data, see Van der Pal et al. (2003).

In the Netherlands cases with NTD's are registered in several national databases. Furthermore the Dutch Association of Patients with a NTD also conducts its own surveys. In this analysis we will use five registrations, which we describe briefly.

1. *Dutch Perinatal Database I (LVR1)*: This is an anonymous pregnancy and birth registry of low-risk pregnancies and births, even if care only relates to a part pregnancy or delivery. Data over the period 1988 through 2002 are used.
2. *Dutch Perinatal Database II (LVR2)*: This list registers anonymous data concerning the birth of a child in secondary care. Data over the period 1988 through 2002 are used.
3. *National Neonate Database (LNR)*: This list contains anonymous information about all admissions and re-admissions of newborns to paediatric departments within the first 28 days of life. Data was used for the period 1992–2002.
4. *Dutch Association of Patients with an NTD (BOSK)*: A short questionnaire was sent to every member of BOSK with an NTD affected child between 1988 and 2002.
5. *Dutch Monitoring System of Child Health Care (NSCK)*: NSCK registers live born infants with an NTD who visit a paediatrician for the first time. All paediatric departments participate. Data was used for the period 1993–2001.

Children were linked on date of birth, zip code, mother's date of birth, and gender of child (Van der Pal et al. 2003). It should be noted that abortions are possible in LVR1 and LVR2, whereas they cannot appear in the other registrations. Therefore we consider only children with a pregnancy duration from 24 weeks (the legal limit for pregnancy termination in the Netherlands).

None of these databases include all cases of neural tube defects because of, for instance, non-participation of health care professionals. Therefore capture–recapture methodology has to be used to estimate the size of babies born with NTDs. The standard approach to estimate the number of NTD would be to fit loglinear models with a

structural zero cell for observations that are in none of the registries. In this situation, however, the usual approach could not be adopted, since some of the registrations were not available for all of the years. Before 1992 three registries were available, in 1994 four registries were available, in the period 1993–2001 five registries were available, and in 2002 four registries were available. The frequencies for all years are given in Table 4. Compared to data that have already been published in Zwane et al. (2004), here the period 1999–2002 is added.

Instead of the standard approach, the EM algorithm was used to estimate the number of NTD-affected infants who were not registered in the years one or more registries did not yet exist or no longer existed. A loglinear model was fitted in the M-step of the algorithm.

The results are summarized in Table 5. In Model 1, the loglinear model is the main effect model. To account for unobserved heterogeneity, in Model 2, the procedure proposed by the International Working Group on Disease Monitoring and Forecasting (1995) was followed, by including a first-order heterogeneity term hetterm.2 for the heterogeneity of capture probabilities (this first-order heterogeneity term states that all two-factor interactions are equal). In Model 3, including a second-order heterogeneity term hetterm.3 leads to a lower AIC value, so in the models that follow this term was dropped. Model 4 allows all inclusion probabilities to vary over time. In Model 5, we drop the interaction between nsck by year; the other interactions between list and year cannot be dropped. In Model 6, we add the two-list interactions. By making the model more complicated (for example, by adding three-list interactions or two-list interactions changing over time), it becomes unstable. Model 6 has the lowest AIC, and the deviance of this model of 351 for 251 degrees of freedom is adequate for our purposes. An analysis of residuals did not reveal any clear trends.

To obtain the total number of infants born with NTD for each of the years in the 1988–2002 period, the observed number of infants and the number of missing infants we had calculated via the capture–recapture analysis were summed. The parametric bootstrap was used to calculate confidence intervals for the total number of infants for each year. The advantage of the bootstrap method is its simplicity and the fact that the bootstrap can yield confidence intervals that are nonsymmetric. For analytical formulae, we refer to Sutherland et al. (2007). We note that the parametric bootstrap also allows one to take model uncertainty into account, and we refer to Zwane et al. (2004) and Zwane and Van der Heijden (2007) for examples, but we did not do this here since, in comparison to Model 6, Models 1 to 5 have essentially no support from the data.

In the parametric bootstrap method, random samples are drawn from an estimated probability distribution derived from a fitted model. For the fitted model, Model 6 was used. So, for the first bootstrap sample, a sample of 3 892 observations is drawn from the probability distribution derived from the maximum likelihood estimates of the completed table under Model 6 (3 892 is the estimated population size). The observations can fall into each of the cells of the completed table of dimension $2 \times 2 \times 2 \times 2 \times 2 \times 15$, i.e., including the 15 structurally zero cells and the cells for which only margins were observed. Subsequently, the observations falling into the 15 cells '00000' are omitted, and the observations falling into cells for which only margins are known in the original data are added up. For example, for 1988, the observations falling in cell '00000' are ignored, and the observations of the remaining

Table 4 Neural Tube Defects: Numbers ascertained by inclusion profile for all years. Order of registrations: nsck, bosk, lnr, lvr2, lvr1

Year	Ascertainment profile[a]																															
	00001	00010	00011	00100	00101	00110	00111	01000	01001	01010	01011	01100	01101	01110	01111	10000	10001	10010	10011	10100	10101	10110	10111	11000	11001	11010	11011	11100	11101	11110	11111	Total
1988	10	103	24					9	1	5	2																					154
1989	5	115	30					3	5	8	4																					180
1990	13	105	43					7	8	5	4																					185
1991	14	100	32					4	5	8	9																					172
1992	21	81	27	15	0	12	7	10	2	3	3	0	0	2	3																	186
1993	13	61	24	4	1	2	0	3	1	0	1	0	0	0	0	12	2	18	8	2	1	4	3	1	2	7	4	0	1	0	0	175
1994	27	34	13	6	1	1	1	3	0	1	1	0	1	0	1	15	8	18	9	5	5	5	2	1	4	1	0	0	1	0	4	168
1995	33	27	15	5	1	2	1	2	1	2	0	0	1	3	0	16	16	18	4	4	0	9	5	2	2	5	1	3	0	1	0	179
1996	30	26	11	10	1	1	1	5	0	0	0	0	0	0	0	9	7	11	9	5	0	4	5	5	7	2	1	1	2	2	4	159
1997	43	26	18	13	2	0	1	4	2	0	1	2	0	0	0	12	11	11	7	3	5	6	3	1	1	1	3	1	2	0	4	183
1998	29	25	20	13	0	2	1	1	0	0	0	0	0	0	0	8	7	14	3	6	12	4	7	0	0	1	2	0	1	1	1	158
1999	39	23	16	5	2	6	1	2	0	0	1	0	1	0	0	13	7	8	10	7	5	6	6	3	1	2	3	1	3	1	1	173
2000	32	26	14	9	3	10	1	1	1	1	0	0	0	0	0	12	9	9	5	7	4	5	3	3	1	2	5	0	0	2	0	165
2001	32	20	11	11	6	3	4	1	1	0	1	0	0	0	0	11	6	11	4	6	1	4	7	0	3	0	0	0	1	2	1	147
2002	35	40	7	12	9	6	6	3	1	1	1	1	2	0	1																	125

[a]Denotes when a child was included in a list (Included = 1) or not (Not included = 0). 00001 implies the child was seen only in LVR_1

Table 5 Selected models with deviance and AIC

Model	Design matrix	Num. par.	df	Deviance	AIC	$\hat{N}$
1	lvr1 + lvr2 + bosk + lnr + nsck + year	20	317	996.5	1 036.5	3 178.7
2	lvr1 + lvr2 + bosk + lnr + nsck + year + hetterm.2	21	316	861.5	903.5	4 321.2
3	lvr1 + lvr2 + bosk + lnr + nsck + year + hetterm.2 + hetterm.3	22	315	861.5	905.5	4 332.8
4	lvr1 + lvr2 + bosk + lnr + nsck + year + hetterm.2 + lvr1:year + lvr2:year + bosk:year + lnr:year + nsck:year	81	256	511.9	693.9	4 188.3
5	lvr1 + lvr2 + bosk + lnr + nsck + year + hetterm.2 + lvr1:year + lvr2:year + bosk:year + lnr:year	73	264	518.5	672.5	4 194.5
6	lvr1 + lvr2 + bosk + lnr + nsck + year + hetterm.2 + lvr1:year + lvr2:year + bosk:year + lnr:year + lvr1:lvr2 + lvr1:bosk + lvr1:lnr + lvr1:nsck + lvr2:bosk + lvr2:lnr + lvr2:nsck + bosk:lnr + bosk:nsck	82	255	350.9	522.9	3 891.8

31 cells are added up into the 7 cells that were observed for 1988, see Table 4. Then an analysis is carried out on the resulting data using Model 6. This leads to a set of 15 population size estimates for the first bootstrap sample, one for each year. This procedure is repeated 500 times, yielding 500 sets of population size estimates. Since no condition on years was made, the number of observations for each year may fluctuate across bootstrap samples.

The outcome of the analysis is found in Table 6. The first column shows the estimated number of newborns with NTD older than 24 weeks in the Netherlands in the period 1988–2002. The width of the 95 percent confidence intervals shows that the reliability of the prevalence increases if more registries are available. From a technical point of view it is interesting to see that, in the period 1993–2001, the period when all five registries were available, the 90 percent confidence interval is considerable smaller than in the period 1988–1992. Using the total number of live and stillbirths in the Netherlands for that year (obtained from Statistics Netherlands), the prevalence per 1 000 is calculated.

It is an important question in public health whether the increased intake of folic acid use produces a decrease in the prevalence of NTD. In the years before the recommendation of folic acid use or before an effect of the increased intake of folic acid use could be expected (1988–1997), the average estimated prevalence was 1.37 per 1 000 live and stillbirths. From 1998, a decrease in the estimated prevalence of NTD is noticed (average estimated prevalence over the years 1998 to 2002 equals 1.21).

Table 6 Neural Tube Defects, results for final model. Table shows estimated population size for NTD, 90 percent confidence interval, number of live and stillbirths in the Netherlands, prevalence per 1 000 live and still births, 90 percent CI

Year	Number NTD	Margin	Number born	Prevalence	Margin
1988	260	194–366	187 712	1.39	1.03–1.95
1989	270	217–340	190 079	1.42	1.14–1.79
1990	267	223–321	199 104	1.34	1.12–1.61
1991	222	185–268	199 732	1.11	0.93–1.34
1992	307	260–402	197 848	1.55	1.31–2.03
1993	247	222–281	196 819	1.25	1.13–1.43
1994	268	234–312	196 666	1.36	1.19–1.59
1995	295	256–336	191 474	1.54	1.34–1.75
1996	248	216–284	190 468	1.30	1.13–1.49
1997	284	250–326	193 428	1.47	1.29–1.69
1998	234	206–270	200 378	1.17	1.03–1.35
1999	256	227–290	201 389	1.27	1.13–1.44
2000	261	224–300	207 619	1.26	1.08–1.44
2001	223	195–260	203 599	1.10	0.96–1.28
2002	251	202–320	203 028	1.24	0.99–1.58

Table 7 Two-list capture–recapture data for Dutch Antilleans in the Official Registration and the Police Registration

Police Registration	Official Registration	
	Not included	Included
Not included	?	64 247
Included	201	5 095

4 Example 2: lists have different sets of covariates

In this second example the question of interest is the size of the population of Dutch Antilleans that stay in the Netherlands without being officially registered. This question was asked to us by the Ministry of Justice as they suspected a sudden rise of young Dutch Antilleans who did not register (the Dutch Antilles are a former colony of the Netherlands that still hold a legal relation to the Netherlands) (for details about the research problem, see Van der Heijden et al. 2006).

In order to obtain an answer we take an unusual approach. Consider Table 7, where we display the data for the year 2000. In this table the column list is the Official Register that is kept by separate Dutch city hall administrations and collected by Statistics Netherlands. As the row list, we may take some other list, in this instance we took the Police Registration. Thus 64 247 and 5 095 are counts of Dutch Antilleans being registered in the Official Registration, and these are the Dutch Antilleans we are not interested in since we are only interested in the Dutch Antilleans who are *not* officially registered. The Dutch Antilleans in the police registration, counts 201 and 5 095, are only partly of interest for answering our research question, namely only

the 201 Dutch Antilleans are of interest since they are not in the Official Registration. Yet if we would know the number of Dutch Antilleans who are neither in the Official Registration nor in the Police Registration, then we could add up this estimate to 201 and we would have an estimate of the Dutch Antilleans who are not officially registered.

This approach to estimate the size of a population is unusual since usually in capture–recapture problems the size of the population of interest is found by adding up counts $64\,247 + 201 + 5\,095$ with the estimate for "?". In this context this would be the estimate of the size of the *complete* Dutch Antillean population, whereas we are only interested in the size of the Dutch Antillean population that is not officially registered. We note that, as a second registration next to the Official Registration, any other registration could have been used instead of the Police Registration to obtain the same goal. However, the Police Registration has the advantage that it is already linked to the Official Registration.

In capture–recapture problems like this each registration has its own covariates, and very often the list of covariates is substantial. This is also the case for this example. However, since we are dealing with a contingency table problem, we can only choose a limited number of covariates in order not to let the data in the contingency able to become too sparse. In choosing covariates we pick covariates for which we suspect that they have an impact on the inclusion probabilities. For this problem, there are three types of covariates. The first type of covariates are covariates that both registrations have in common; here we take gender G and age A (with levels 13–17, 18–24, 25–44, 45+). The second type are covariates that are unique for the Official Registration; here we take length of stay in the Netherlands L (0–5 years, 5–18 years, +18 years) and marital status M (married, unmarried, divorced). And last, the third type of covariates are covariates that are unique for the Police Registration, which is the number of times they have been apprehended by the police T (once, twice, or more) and whether they are known as a hard drug user H (no, yes).

Since the Netherlands knows very strict privacy regulations and we were only allowed by Statistics Netherlands to report observed data were counts would not be below 30, we only report the data in Table 7 and other summaries of the data analyses in tables below.

We show two analyses, one "classical" analysis, where we only make use of the two covariates that the Official Registration and the Police Registration have in common, and one analysis where we also include the covariates that are unique to either the Official Registration or the Police Registration. We denote the Official Registration by O and the Police registration by P.

The "classical" analysis is to build a contingency table of the Official Registration O, the Police Registration P, gender G, and age A. Because the combination "not in O" and "not in P" is impossible, the corresponding cells are structurally zero. Therefore the maximal model that can be fitted to the data is the model where O and P are assumed to be independent given age A and gender G jointly. More restrictive models were tried, but they turned out to have a worse fit. Therefore we now discuss the estimates for model $[GAO][GAP]$ (see Panel A of Table 8).

The total population of Dutch Antilleans is estimated to be 72 322, where 69 342 are in the Official Registration, and 2 980 are estimated to be not in the Official Registration (201 of these were observed in the Police Registration but not in the Official

Table 8 Dutch Antilleans. Results from standard analysis and EM approach

Panel A: Results for standard loglinear models

Gender	Age	Not in OR	In OR	Total	In OR	In PR	2.5	50	97.5
Male	13–17	41.5	3 127	3 168.5	1.3	12.1	8.1	41.2	81.7
Male	18–24	272.0	6 101	6 373.0	4.3	16.2	193.2	270.4	355.1
Male	25–44	662.1	16 826	17 488.1	3.8	13.9	521.8	659.5	802.6
Male	45+	296.9	7 775	8 071.9	3.7	5.4	158.2	292.7	461.9
Female	13–17	106.7	3 058	3 164.7	3.4	2.8	0.0	100.7	255.0
Female	18–24	317.3	6 210	6 527.3	4.9	4.4	149.7	313.8	514.6
Female	25–44	678.6	16 441	17 119.6	4.0	3.2	401.2	677.8	965.3
Female	45+	605.2	9 804	10 409.2	5.8	0.8	127.2	574.9	1 248.9
Total		2 980.2	69 342	72 322.2	4.1	7.3	2 393.9	2 952.5	3 703.8

Panel B: Results for EM approach

Gender	Age	Not in OR	In OR	Total	In OR	In PR	2.5	50	97.5
Male	13–17	39.4	3 127	3 166.4	1.2	12.1	7.9	37.9	77.8
Male	18–24	258.7	6 101	6 359.7	4.1	16.2	185.7	261.1	338.4
Male	25–44	629.0	16 826	17 455.0	3.6	13.9	508.7	629.0	766.1
Male	45+	280.6	7 775	8 055.6	3.5	5.4	152.5	274.3	421.6
Female	13–17	100.7	3 058	3 158.7	3.2	2.8	0.0	100.4	245.2
Female	18–24	299.7	6 210	6 509.7	4.6	4.4	152.9	298.4	460.0
Female	25–44	640.5	16 441	17 081.5	3.7	3.2	400.2	654.1	934.6
Female	45+	570.3	9 804	10 374.3	5.5	0.8	118.7	556.8	1 175.3
Mar. stat.	L.o.stay	Not in OR	In OR	Total	In OR	In PR	2.5	50	97.5
Unmarried	0–5	926.8	17 230	18 157.2	5.1	10.0	720.8	927.7	1 172.7
Unmarried	5–18	763.4	19 561	20 324.2	3.8	9.0	597.2	765.9	969.1
Unmarried	18+	225.1	6 750	6 975.1	3.2	9.0	175.6	225.7	286.6
Married	0–5	154.2	3 490	3 644.6	4.2	3.7	119.5	154.2	197.6
Married	5–18	207.7	6 237	6 444.6	3.2	2.3	158.1	208.8	267.9
Married	18+	186.8	6 632	6 818.3	2.7	2.2	142.7	187.8	244.0
Divorced	0–5	99.0	1 890	1 989.1	5.0	6.6	76.6	98.9	126.8
Divorced	5–18	147.6	3 988	4 135.5	3.6	5.8	114.8	148.1	189.2
Divorced	18+	108.4	3 564	3 672.1	3.0	5.8	83.9	108.5	137.8
Times seen	Harddrug	Not in OR	In OR	Total	In OR	In PR	2.5	50	97.5
1	No	1 454.9	19 148	20 603.0	7.1	8.4	1 110.6	1 453.1	1 857.9
1	Yes	5.4	236	241.0	2.2	7.7	0.0	0.0	28.1
2+	No	1 279.8	39 147	40 427.0	3.2	7.0	952.5	1 290.2	1 673.9
2+	Yes	78.7	10 811	10 889.8	0.7	6.4	16.2	75.7	166.8
Total		2 818.9	69 342	72 160.9	3.9	7.3	2 202.4	2 827.4	3 580.0

Registration). We calculated a bootstrap confidence interval for this number, and this interval turned out to be skewed, ranging from 2 394 to 3 704. For males, there is a peak in not being registered for the group aged 18–24 (4.3 percent of this gender-age group), and for females, there are two peaks, namely one for 18–24 (4.9 percent) and one for 45 and older (5.8 percent). Notice that from the younger males a substantial part is known in the police registration.

We now turn to the analysis where we make use of covariates that are unique to either the Official Registration or to the Police Registration. The maximal model that can be fit to these data is $[GATHO][GALMP]$ $[GATHLM]$, i.e., (i) the term $GATHO$ shows that the inclusion probabilities for the Official Registration are a function of Gender, Age, number of Times apprehended and known as Hard drug user; (ii) the term $GALMP$ shows that the inclusion probabilities for the police registration are a function of Gender, Age, Length of stay, and Marital status, and (iii) the term $GATHLM$ shows that all covariates are allowed to be related. We note that condition (iii) is important since it is proven by Zwane and Van der Heijden (2007) that, if the unique covariates in the Official Registration are *not* directly related to the unique covariates in the Police Registration, then these unique covariates have no impact on the estimate of the total population size. In other words, model $[GATHO][GALMP]$ or more parsimonious models would lead to the same population size estimate as the model in the "classical analysis" and therefore, when interest only goes out to the total population size—and not to the relation of the covariates with the population size—these covariates could be left out of the analysis.

In our model search we tried several models but we encountered numerical problems due to the fact that, when fitting models with including higher-order margins as $GATHO$ and $GALMP$, many of related observed marginal frequencies were zero. Therefore we had to work with more parsimonious models. First, we fit model $[GAO][THO][GAP][LMP]$, i.e., the inclusion probability for the Official Registration is a function of Gender and Age jointly as well as of number of Times apprehended and known as Hard drug user jointly; the inclusion probability for the Police Registration is a function of Gender and Age jointly as well as of Length of stay and Marital status jointly; we further note that there are no direct relations between the unique covariates in the Official Registration on the one hand and the unique covariates in the Police Registration on the other hand, and this model then leads to a population size estimate that is identical to the population size estimate in model $[GAO][GAP]$ in the "classical" analysis. In this model the deviance is 42 975, 46 parameters are fitted, and the model has an AIC of 43 067. Second, we fit the model $[GAO][THO][GAP][LMP][TL][TM][HL][HM]$, i.e., compared to the first model, we now add direct interactions between the unique covariates in the Official Registration and the Police Registration. In this model the deviance is 36 518, 54 parameters are fit, and the model has an AIC of 36 626. We do not present further models since these models become unstable due to the fact that in more complicated models margins are fitted to the data to contain zero frequencies. Estimates for this model are found in Panel B of Table 8.

First we notice that, compared to the "classical" analysis, the estimated population size goes down from 2 980 to 2 819. We show estimated population sizes for subpopulations stratified by, first, Gender and Age, second, by Marital Status and Length of

Stay, and third, by number of Times apprehended and known as Hard drug user. There is no clear picture for Martial status, but for Length of stay, we find that Dutch Antilleans with a length of stay shorter than 0–5 years are relatively more often not known in the Official Registration. Hard drug users are relatively more often known in the Official Registration, whereas Dutch Antilleans who are more often apprehended are also more often known in the Official Registration.

In evaluating the findings for this example, we have to conclude that we are not certain about the validity of the assumptions made, and hence we have to handle the findings with care. First, there is the possibility that the link between the Official Registration and the Police Registration is imperfect. This is discussed in detail in Van der Heijden et al. (2006), and they conclude from discussions with the police and Statistics Netherlands that this problem is likely to be negligible. Second, there is the open population problem. The main problem here is that it is possible that, among the Dutch Antilleans appearing in the Police Registration but not in the Official Registration, there are individuals who just spend a holiday in the Netherlands. The estimates are correct if this assumption is not violated, but if this assumption *is* violated, the estimates found are too high (namely, the 201 in Table 7 should be lower, and this will make the estimates for '?' lower). Thirdly, we assume that being in the Official Registration and being in the Police Registration are independent when we control for the covariates. If there is a negative relation between the two (i.e., being known in the Police Registration goes together with *not* being known in the Official Registration), then the estimates are too low. The validity of these assumptions could be further studied using qualitative field methods—by interviewing police officers, civil servants working for the Official Registrations and interviewing relevant Dutch Antilleans, but this is beyond the scope of this study.

If one is willing to assume that, over the years, the violation of assumptions is constant, then it is possible to study potential rises and falls of the population size estimates. Van der Heijden et al. (2006) report an estimate of 2 818 (in confidence interval 2 202–3 580) in 2000, an estimate of 5 806 (4 939–6 824) in 2001, an estimate of 8 330 (7 548–9 210) in 2002, and 11 147 (10 041–12 413) in 2003. This would then lead to the conclusion that the population size of the Dutch Antilleans not in the Official Register is indeed rising.

5 Discussion

The EM algorithm was used in this paper to solve two problems: the first problem is that the lists refer to different but overlapping populations. The second problem is that each list has a set of covariates and the sets of covariates are not identical.

For this second problem, we showed an example with categorical covariates. The EM algorithm can in principle be used for capture–recapture models with continuous covariates, but the expectation involves complex numerical integration. For these situations, we prefer the use of multiple imputation. The main advantage of multiple imputation over maximum likelihood methods is that it is computationally much simpler for most practical situations (Sinharay et al. 2001). For continuous covariates, we have studied this in Zwane and Van der Heijden (2008).

References

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

Chao, A., Tsay, P., Lin, S., Shau, W., Chao, D.: The applications of capture–recapture models to epidemiological data. Stat. Med. **20**, 3123–3157 (2001)

Cormack, J.M.: Log-linear models for capture-recapture. Biometrics **45**, 395–413 (1989)

Fienberg, S.E.: The multiple recapture census for closed populations and incomplete 2^k contingency tables. Biometrika **59**, 591–603 (1972)

International Working Group for Disease Monitoring and Forecasting: Capture–recapture and multiple-record systems estimation 1: history and theoretical development. Am. J. Epidemiol. **142**, 1047–1058 (1995)

Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)

Sinharay, S., Stern, H., Russell, D.: The use of multiple imputation for the analysis of missing data. Psychol. Meth. **6**, 317–329 (2001)

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

Tsay, P.K., Chao, A.: Population size estimation for capture–recapture models with applications to epidemiological data. J. Appl. Stat. **28**, 25–36 (2001)

Van der Heijden, P.G.M., Zwane, E., Hessen, D.: Schatting van aantal in Nederland verblijvende Antillianen die niet ingeschreven zijn in de GBA. Een "capture–recapture"-analyse in opdracht van het Ministerie van Justitie. Utrecht, Utrecht University, Department of Methodology and Statistics (2006)

Van der Pal, K.M., Van der Heijden, P.G.M., Buitendijk, S.E., Den Ouden, A.L.: Periconceptional folic acid use and the prevalence of neural tube defects in the Netherlands. Eur. J. Obset. Gynecol. Reprod. Biol. **108**, 33–39 (2003)

Zwane, E., Van der Heijden, P.G.M.: Analysing capture–recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations. Stat. Med. **26**, 1069–1089 (2007)

Zwane, E., Van der Heijden, P.G.M.: Capture–recapture studies with incomplete mixed categorical and continuous covariates. J. Data Sci. **6**, 557–572 (2008)

Zwane, E., Van der Pal, K., Van der Heijden, P.G.M.: The multiple-record systems estimator when registrations refer to different but overlapping populations. Stat. Med. **23**, 2267–2281 (2004)