STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:2267–2281 (DOI: 10.1002/sim.1818)

The multiple-record systems estimator when registrations refer to different but overlapping populations

Eugene N. Zwane[1,*,†], Karin van der Pal-de Bruin[2] and Peter G. M. van der Heijden[1]

[1] *Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands*
[2] *TNO Prevention and Health, Leiden, The Netherlands*

SUMMARY

In multiple-record systems estimation it is usually assumed that all registration relate to the same population. In this paper, we develop a method which can be used when the registrations relate to different populations, in the sense that they cover, for example, different time periods or regions. We show that under certain conditions ignoring that the registrations relate to different populations results in correct estimates of population size. The EM algorithm is presented as a method that can be used for more general problems. The parametric bootstrap is used to construct a confidence interval. The proposed method is then applied to a data set with five registrations of neural tube defects, that cover different time periods. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: capture–recapture; multiple-records systems; neural tube defects; population size estimation

1. INTRODUCTION

Capture–recapture analysis was developed by ecologist for assessing the size of animal populations in the wild [1]. The population size is estimated from the degree of overlap between two or more samples obtained from the same population. In epidemiology, capture–recapture methods are used to estimate or adjust for the extent of incomplete ascertainment using information from overlapping cases from distinct sources [2, 3]. The common labels for the methods in human populations are, *multiple-system*, *multiple-records systems*, and *multiple-record systems method* [2].

For two overlapping samples (from the same population) the method is used to estimate the part of the population that is not observed (individuals in neither of the two samples). This estimation is accomplished under the assumption of independence of inclusion probabilities [2, 3]. Another assumption is homogeneity of inclusion probabilities over individuals. Although

*Correspondence to: Eugene N. Zwane, Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, Utrecht 3508 TC, Netherlands.
†E-mail: e.zwane@fss.uu.nl

Received April 2002
Accepted January 2004

it has been long thought that the inclusion probabilities for both lists should be homogeneous [2], it has recently been shown that only one of the lists has to have homogeneous inclusion probabilities when the *joint* capture probability is positive [4, 5].

Categorical covariates are frequently used to diminish heterogeneity of inclusion probabilities, and for this log-linear models are widely used as it is possible to incorporate stratification variables, and permit dependence between sources when there are more than two sources [6–8].

In this paper, we deal with a special case of heterogeneous inclusion probabilities, namely the case where the populations from where the lists emanate partially overlap. This results in some individuals being systematically missed by one or more of the lists. Therefore the *joint* inclusion probabilities are zero for some individuals in the combined population. A first example is that lists do not cover the same region. For this example, a stratum is defined as a subregion, and not every list is observed in each subregion. A second example is when lists do not cover the same time periods. Here the strata are defined as the subperiods of time. We approach the absence of observations in certain strata for certain lists as an incomplete data problem.

The EM algorithm [9] is an iterative procedure for obtaining maximum likelihood estimates in incomplete data. In the standard capture–recapture problem the EM algorithm can be used to estimate part of the population missed by all sources [10]. As we have partially overlapping populations there are more entries missing in the contingency table than in the standard capture–recapture problem. When some lists are not operating in some strata, implying that there are several unobservable cells that are a result from non-operating lists, the EM algorithm can *still* be used to estimate these missing entries, and thus the population size.

Section 2 introduces the data set that will be used to illustrate the results. In Section 3, we present two simple capture–recapture models; the first with one list operating in a subperiod of another list, and the other where the two lists operate in different but partially overlapping years (time periods). We show under which conditions stratification by year can be ignored. We then present the EM algorithm in Section 4, and show how it can be used to estimate population size in partially overlapping populations, and further show that for simple models the results will be equivalent with using traditional methods. In Section 5 we analyse the data set on neural tube defects which motivated this article. Finally, Section 6 gives some conclusions.

2. DATA

In this section, we will introduce the data set on neural tube defects (NTD's) in the Netherlands that will be used to illustrate the procedure presented in the paper. 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 [11]. In this analysis, we will use five registrations which we describe briefly.

1. *Dutch Perinatal Database I* ($R1$): 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 1998 are used.
2. *Dutch Perinatal Database II* ($R2$): This list registers anonymous data concerning the birth of a child in secondary care. Data over the period 1988 through 1998 are used.

3. *National Neonate Database* (*R*3): 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 1992–1998.
4. *Dutch Monitoring System of Child Health Care* (*R*4): R4 registers live born infants with a NTD who visit a paediatrician for the first time. All paediatric departments participate. NTD's are registered since 1993.
5. *Dutch Association of Patients with a NTD* (*R*5): A short questionnaire was sent to every member of R5 with a NTD affected child between 1988 and 1998.

Children were linked on date of birth, zip code, mother's date of birth and gender of child [11]. It should be noted that abortions are possible in R1 and R2, 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. Thus multiple-record systems estimation has to be used to estimate the size of babies born with NTD's. The usual approach is to fit log-linear models with a structural zero for the observations missed by all lists [3]. In our situation this usual approach cannot be adopted as some of the registrations are not available for some years: for 1988–1991 only three registrations are available (R1,R2,R5) and in 1992 only four registrations are available (R1,R2,R3,R5). The frequencies for all years are given in Table I.

In 1988–1991, observations with an inclusion profile of 01000 also include observations that could have been 01100, 01010, and 01110 had R3 and R4 been active. In 1992, observations with an inclusion profile of 01100 also include observations that could have been 01110 had R4 been active. Similar statements could be said for some other inclusion profiles.

In the next sections, we show that the EM algorithm is a tool which can effectively analyse data of this form, by utilizing information on relations between registrations while stratifying by year. We start by showing for simple cases what can go wrong when one ignores the fact that the registrations do not come from the same population.

3. CAPTURE–RECAPTURE METHODOLOGY

In this section, we discuss the problem of estimating population size in 'dual record systems' when the registrations relate to different but overlapping populations, for example, the registrations may cover different but overlapping time periods. In particular, we study what happens if this fact is ignored, that is, if it is assumed that both registrations refer to the same population. We show one example where the union of the two populations is estimated unbiased, and one where the resulting estimate is biased. This then serves as a motivation for a general solution discussed in Section 4.

3.1. Simple capture–recapture model

The simplest multiple-record systems consists of two lists. Let Π_1 and Π_2 be the probability of capture by list 1 and 2, respectively. The joint probabilities are denoted by π_{ij} $(i = 0, 1; j = 0, 1)$, where π_{10} is the probability to be in list 1 only, π_{01} is the probability to be in list 2 only and, π_{11} is the probability to be in both lists. The corresponding frequencies are shown in

Table I. Numbers ascertained by inclusion profile for all years.

Year	Ascertainment profile*,†																															Total
	00000	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	
1988		9							101	5							4	0							24	2							145
1989		3							114	8							3	1							30	4							163
1990		7							105	5							3	3							43	4							170
1991		3							95	7							4	1							32	8							150
1992		10			15	0			80	3			12	2			9	1			0	0			27	3			7	3			172
1993		3	12	1	4	0	2	0	61	0	18	7	1	0	4	0	5	0	0	1	1	0	0	1	24	1	7	4	0	0	3	0	160
1994		3	15	1	6	0	5	0	34	1	18	1	1	0	5	0	24	0	7	3	1	1	4	1	13	1	9	0	1	1	2	4	162
1995		2	16	2	5	0	4	3	27	2	18	5	2	3	9	1	29	1	15	2	1	1	0	0	15	0	4	1	1	0	5	0	174
1996		5	9	5	10	0	5	1	26	0	11	2	1	0	4	2	26	0	6	6	1	0	0	2	11	0	9	1	1	0	5	4	153
1997		4	12	1	13	2	3	1	26	0	11	1	0	0	6	0	41	2	11	1	2	0	4	2	18	1	7	3	1	0	3	4	180
1998		1	8	0	13	0	6	0	25	0	14	1	2	0	4	1	27	0	7	0	0	0	11	0	20	0	3	2	1	0	7	1	154

*Denotes when a child was included in a list (Included = 1) or not (Not included = 0). 00001 implies the child was seen only in R5.

†Zero frequencies stand for sampling zeros.

Table II. Dual-record system problem.

	List 2	
List 1	Not included	Included
Not included	0*	n_{01}
Included	n_{10}	n_{11}

*Treated as structurally zero cell.

Table II. The probability π_{00} and frequency n_{00} are unknown and have to be estimated in order to compute an estimate of the unknown population size N. Furthermore $n_{ij} = N\pi_{ij}$.

Assume that for each sample, each individual has the same inclusion probability. Then Π_1, Π_2 and N can be estimated by

$$\hat{\Pi}_1 = \frac{n_{11}}{n_{11} + n_{01}}, \quad \hat{\Pi}_2 = \frac{n_{11}}{n_{11} + n_{10}}$$

$$\hat{N} = \frac{n_{11}}{\hat{\Pi}_1 \hat{\Pi}_2} \tag{1}$$

and this result forms the basis of our development.

3.2. Two lists and two strata

Assume now that we have two strata, for example, 2 years (or time periods). Year is a stratifying variable with two categories indexed by k, where $k = 1$ denotes the first year and $k = 2$ the second year. Let $\Pi_{1|1}$ and $\Pi_{2|1}$ be the probabilities to be in the list 1 and 2 in the first year and $\Pi_{1|2}$ and $\Pi_{2|2}$ be the probabilities to be in list 1 and 2 in the second year. Let the joint probabilities for the first year be $\pi_{ij|1}$ and the joint probabilities for the second year be $\pi_{ij|2}$. Let the unknown population size for year 1 and 2 be N_1 and N_2, respectively. By analysing the data from each year (or time period) separately we can estimate $\Pi_{1|1}$, $\Pi_{2|1}$, and N_1 by

$$\hat{\Pi}_{1|1} = \frac{n_{11|1}}{n_{11|1} + n_{01|1}}, \quad \hat{\Pi}_{2|1} = \frac{n_{11|1}}{n_{11|1} + n_{10|1}}$$

$$\hat{N}_1 = \frac{n_{11|1}}{\hat{\Pi}_{1|1} \hat{\Pi}_{2|1}} \tag{2}$$

and $\Pi_{1|2}$, $\Pi_{2|2}$, and N_2 by

$$\hat{\Pi}_{1|2} = \frac{n_{11|2}}{n_{11|2} + n_{01|2}}, \quad \hat{\Pi}_{2|2} = \frac{n_{11|2}}{n_{11|2} + n_{10|2}}$$

$$\hat{N}_2 = \frac{n_{11|2}}{\hat{\Pi}_{1|2} \hat{\Pi}_{2|2}} \tag{3}$$

Now let list 2 be observed only in the first year such that the observed table can be set out as in Table III.

Table III. Dual-record system problem with 2 years.

Year	List 1	List 2		Total
		Not included	Included	
1	Not included	0*	$n_{01\|1}$	
	Included	$n_{10\|1}$	$n_{11\|1}$	
2	Not included	0*	0*	
	Included	?	?	$n_{10\|2} + n^{\dagger}_{11\|2}$

*Structurally zero cells.
†Only the margin is observed.

Assume that we ignore the fact that the registrations refer to different populations, by ignoring (the variable) year. Let us denote the elements in the table where year is ignored by $n_{ij|+}$. These elements are related to the elements in Table III by $n_{11|+} = n_{11|1}$, $n_{01|+} = n_{01|1}$, and $n_{10|+} = n_{10|1} + n_{10|2} + n_{11|2}$. The question is: can Equation (1) be used to estimate N? And if so, under what assumptions? In other words, when would ignoring the fact that list 2 is observed only in 1 year lead to an unbiased estimate of the population size? The observations to be estimated are $n_{00|+} = n_{00|1} + n_{00|2} + n_{01|2}$. Using this we find that

$$\hat{\Pi}_{1|+} = \frac{n_{11|1}}{n_{11|1} + n_{01|1}}, \quad \hat{\Pi}_{2|+} = \frac{n_{11|1}}{n_{11|1} + n_{10|1} + n_{11|2} + n_{10|2}}$$

$$\hat{N}_{+} = \left(\frac{n_{11|1}}{\hat{\Pi}_{1|1}}\right)\left(\frac{n_{11|1} + n_{10|1} + n_{11|2} + n_{10|2}}{n_{11|1}}\right) \tag{4a}$$

$$= \left(\frac{n_{11|1}}{\hat{\Pi}_{1|1}}\right)\left(\frac{1}{\hat{\Pi}_{2|1}} + \frac{n_{11|2} + n_{10|2}}{n_{11|1}}\right) \tag{4b}$$

$$= \hat{N}_1 + \left(\frac{n_{11|2} + n_{10|2}}{\hat{\Pi}_{1|1}}\right) \tag{4c}$$

$$= \hat{N}_1 + \left(\frac{n_{11|2}}{\hat{\Pi}_{1|1}\hat{\Pi}_{2|2}}\right) \tag{4d}$$

Equation (4) shows that if $\hat{\Pi}_{1|1} = \hat{\Pi}_{1|2}$, then $\hat{N}_{+} = \hat{N}_1 + \hat{N}_2$. Thus for two lists and 2 years (or strata) even if the joint inclusion probability of some individuals in the combined population is zero the dual record-systems estimator can still be used, as long as $\hat{\Pi}_{1|1} = \hat{\Pi}_{1|2}$, that is, the list observed in both years (or strata) has to have homogeneous inclusion probabilities.

3.3. Two lists and three strata

Instead of 2 years we now assume that we have two list and 3 years. We denote the third year by $k = 3$. Assume that list 1 operates in the years 1 and 2, and list 2 operates in years

Table IV. Dual-record system problem with 3 years.

Year	List 1	List 2		Total
		Not included	Included	
1	Not included	0*	0*	
	Included	?	?	$n_{10\|1} + n^{\dagger}_{11\|1}$
2	Not included	0*	$n_{01\|2}$	
	Included	$n_{10\|2}$	$n_{11\|2}$	
3	Not included	0*	?	
	Included	0*	?	
	Total		$n_{01\|3} + n^{\dagger}_{11\|3}$	

*Structurally zero cells.
†Only the margin is observed.

2 and 3, such that the years where list 1 and 2 are operational partly overlap but the year for list 2 are not necessarily a subset of the years where list 1 is active. The cells actually observed are: $n_{1+|1}$ in the first year, $n_{10|2}$, $n_{01|2}$, and $n_{11|2}$ in the second year, and $n_{+1|3}$ in the third year (see Table IV).

Only the observations in the year 2 have non-zero joint *inclusion* probabilities. Ignoring year, the elements of the resulting table are related to those in Table IV by $n_{11|+} = n_{11|2}$, $n_{10|+} = n_{10|1} + n_{11|1} + n_{10|2}$, and $n_{01|+} = n_{01|2} + n_{01|3} + n_{11|3}$. The estimates of $\Pi_{1|+}$, $\Pi_{2|+}$, N_+ from this table are

$$\hat{\Pi}_{1|+} = \frac{n_{11|2}}{n_{01|2} + n_{11|2} + n_{01|3} + n_{11|3}}, \quad \hat{\Pi}_{2|+} = \frac{n_{11|2}}{n_{10|1} + n_{11|1} + n_{10|2} + n_{11|2}}$$

$$\hat{N}_+ = n_{11|2}\left(\frac{n_{01|2} + n_{11|2} + n_{01|3} + n_{11|3}}{n_{11|2}}\right)\left(\frac{n_{10|1} + n_{11|1} + n_{10|2} + n_{11|2}}{n_{11|2}}\right) \tag{5a}$$

$$= n_{11|2}\left(\frac{1}{\hat{\Pi}_{1|2}} + \frac{n_{01|3} + n_{11|3}}{n_{11|2}}\right)\left(\frac{n_{10|1} + n_{11|1}}{n_{11|2}} + \frac{1}{\hat{\Pi}_{2|2}}\right) \tag{5b}$$

$$= \frac{n_{11|1}}{\hat{\Pi}_{1|2}\hat{\Pi}_{2|1}} + \hat{N}_2 + \frac{n_{11|3}}{\hat{\Pi}_{1|3}\hat{\Pi}_{2|2}} + \frac{(n_{10|1} + n_{11|1})(n_{01|3} + n_{11|3})}{n_{11|2}} \tag{5c}$$

This shows that, even if $\hat{\Pi}_{1|1} = \hat{\Pi}_{1|2}$ and $\hat{\Pi}_{2|2} = \hat{\Pi}_{2|3}$, collapsing the table over years results in a positively biased estimate of population size, the bias being $(n_{10|1} + n_{11|1})\ (n_{01|3} + n_{11|3})/n_{11|2}$. However, as this quantity has observable values it can be substracted from $\hat{N}_+$ to get an unbiased estimate for the population size.

In conclusion, we note that in certain cases ignoring stratification (strata can be years or time periods) is not a problem in the estimation of the population size but in some cases it is. Furthermore, it is not possible to estimate the population sizes for each of the strata (or

years) separately. This shows that there is a need to develop a general approach which would work for the cases where stratification has to be incorporated in the models.

4. EM ALGORITHM

A widely used method for analysis of partially classified counts is the EM algorithm [9]. This technique was developed for data that are 'missing at random' (MAR). In a survey context, missing values are said to be MAR if the occurrence of the missing value is conditionally independent of the actual response that would have been observed given the observed responses to the other questions; that is, the occurrence of the missing value can depend on observed responses to other questions, but given these, it does not depend on the missing value itself. Loosely speaking, for our case this implies that, observations from years where all registrations are active and observations from years with non-operating registrations that have the same characteristics do not differ systematically by year. If the data are MAR, the missingness is called 'ignorable' because ignoring the missing data mechanism does not affect likelihood-based inferences, such as the maximum likelihood (ML) estimates. ML produces estimates that are asymptotically unbiased [12, p. 264] if the model is true.

Under the ignorability assumption the EM algorithm can be used to obtain the ML estimates [9]. The EM algorithm is an iterative procedure with two steps in each iteration: the E-step and the M-step. The E-step of the EM algorithm computes the expected complete data sufficient statistics given the current parameter estimates and the observed data; in this case this entails distributing partially classified counts using information in other years. The M-step computes new ML estimates of the parameters based on the current values of the expected complete-data sufficient statistics.

Applying the EM algorithm to capture–recapture data with partially overlapping populations is valid, if the non-operating lists are missing by design (such that the missingness is ignorable). For epidemiological capture–recapture data populations might partially overlap, for example:

- by year due to development of registrations which are hoped to be better than active ones or the closing of obsolete existing registrations,
- and by region as some regions might have registrations that are not yet implemented in other regions.

These examples are all design based, implying that the use of the EM algorithm is valid.

4.1. General procedure

In this section, we illustrate the general procedure of the EM algorithm. In our procedure the EM algorithm is used to distribute the observations from years where some registrations are not operational, and as such it is similar to the standard EM algorithm, whilst [10] is a non-standard application of the EM algorithm in that it is used to estimate observations missed by all lists. The notation used here is similar to that in Reference [9, p. 182].

Let n_{ik} denote the frequencies of the hypothetical complete data, where $i=(1,2,\ldots,I)$ is an index denoting a cross-classification of S lists such that $I=2^S-1$, and $k=(1,2,\ldots,K)$ is the index for *Year* (which is fully observed). Note that for each list, $i_s=(0,1)$, where

$s=(1,2,\ldots,S)$ is the index for a list. The hypothetical complete data consist of $I\times K$ cells (with cells denoted by c_{ik}), such that n_{ik} denotes the observed frequency of individuals classified into cell c_{ik}, with corresponding probabilities $\pi_{ik}, \sum_i \sum_k \pi_i k = 1$.

The observed data consist of two sets of years: (1) a set of years, denoted by S_1, where all lists are operating (completely classified observations), and (2) a set of years, denoted by S_2, where not all the lists are operating (partially classified observations). Note that it is a possible that S_1 is empty; if S_2 is empty, then there is no missing data problem and the data can be analysed in a standard way. We partition the partially classified observations into J groups, so that within each group, all units have the same set of possible cells (during partitioning we ignore year). Suppose r_{jk} denotes the count for the partially classified observations in the kth year which fall in the jth group; let S_{jk} denote the set of cells to which the observations might belong. Define indicator functions $\delta(c_{ik}\in S_{jk})$, $i=(1,2,\ldots,I)$ and $j=(1,2,\ldots,J)$ where $\delta(c_{ik}\in S_{jk})=1$ if cell c_{ik} belongs to S_{jk} and 0 otherwise.

Let $\hat{\pi}_{ik}^{(t)}$ be the current estimate of the probability for cell c_{ik} after the tth iteration of the M-step. The $(t+1)$th E-step of the EM algorithm calculates the expectation of the cell frequencies (n_{ik}) for the partially classified frequencies (r_{jk}) using

$$\hat{n}_{ik}^{(t+1)} = \frac{\sum_{p=1}^{K} \hat{\pi}_{ip}^{(t)} \delta(c_{ip}\in S_{jp})}{\sum_{p=1}^{K}\sum_{l=1}^{I} \hat{\pi}_{lp}^{(t)} \delta(c_{lp}\in S_{jp})} \times r_{jk} \tag{6}$$

The above expression distributes the partially classified counts (r_{jk}) using the current estimates of the conditional probabilities of falling in cell c_{ik} given that on observation falls in the set of categories S_{jk}.

The $\hat{n}_{ik}^{(t+1)}$ in (6) denote the completed data at the $(t+1)$th iteration. In the M-step a log-linear model is fitted to the completed data, with the cells missing by design denoted as structurally zero. Thus we maximize the so-called complete data log-likelihood. Let S^* denote the set of cells corresponding to years where all lists are operating, that is, where all relevant counts are observed. The complete data likelihood at iteration $t+1$ denoted by $\ell^{(t+1)}$ is given by

$$\ell^{(t+1)} = \sum_{c_{ik}\in S^*} n_{ik} \ln \pi_{ik} + \sum_{c_{ik}\in S_{jk}} \hat{n}_{ik}^{(t+1)} \ln \pi_{ik} \tag{7}$$

The fitted probabilities, $\hat{\pi}_{ik}$, from the log-linear model fitted in the M-step are then used in the E-step of the $(t+1)$ iteration, where they are denoted by $\hat{\pi}_{ik}^{(t+1)}$, to derive updates for the completed data. This procedure is repeated until the complete data log-likelihood converges. After convergence the parameter estimates are used to find point estimates for the structurally zero cells, and an estimate of the population size.

4.2. Dual list examples

To illustrate the EM algorithm we use the examples presented in Sections 3.2 and 3.3. For the example given in Section 3.2 there are two lists and two years (strata), with only list 1 operating in the second year. In this example there is only one partially classified frequency and it occurs in the second year, and using the notation presented in the general procedure the partially classified observation is denoted by r_{12} ($r_{12}\equiv n_{1+|2}$, see Table III, and for simplicity we proceed with $n_{1+|2}$). As detailed earlier in the E-step we compute the conditional

expectations for $n_{10|2}$ and $n_{11|2}$ using

$$\hat{n}_{10|2}^{(t+1)} = \frac{\hat{\pi}_{10|1}^{(t)}}{\hat{\pi}_{10|1}^{(t)} + \hat{\pi}_{11|1}^{(t)}} \times n_{1+|2}$$

$$\hat{n}_{11|2}^{(t+1)} = \frac{\hat{\pi}_{11|1}^{(t)}}{\hat{\pi}_{10|1}^{(t)} + \hat{\pi}_{11|1}^{(t)}} \times n_{1+|2}$$

In the M-step the complete data log-likelihood is maximized with respect to the unknown parameters $\pi_{ij|k}$ with $n_{10|2}$ and $n_{11|2}$ replaced by their conditional expectations $\hat{n}_{10|2}^{(t+1)}$ and $\hat{n}_{11|2}^{(t+1)}$. Thus we maximize (compare with Equation (5))

$$\ell^{(t+1)} = \sum_{i,j} n_{ij|1} \ln(\pi_{ij|1}) + \hat{n}_{10|2}^{(t+1)} \ln(\pi_{10|2}) + \hat{n}_{11|2}^{(t+1)} \ln(\pi_{11|2})$$

To maximize this log-likelihood we use a log-linear model with structural zeros for the unknown counts, that is $n_{00|1}$, $n_{00|2}$ and $n_{01|2}$. Thus the complete data likelihood is maximized over 5 cells. The parameters can then be used to estimate the frequencies for the structurally zero cells. It can easily be verified that the total number of observations missed by all lists $n_{00|1} + n_{00|2} + n_{01|2}$ is equal to that obtained by collapsing the table as shown in Section 3.2. Thus the EM is also able to provide the solution to the problem.

Collapsing over years in the example in Section 3.3 resulted in a biased estimate of the population size. Here, we show that the EM algorithm results in an unbiased estimate of the population size. This example has two partially classified counts and they are denoted by r_{11} and r_{23} ($r_{11} \equiv n_{1+|1}$ and $r_{23} \equiv n_{+1|3}$, see Table IV). In the E-step we compute the conditional expectations for $n_{10|1}$, $n_{11|1}$, $n_{01|3}$, and $n_{11|3}$ using

$$\hat{n}_{10|1}^{(t+1)} = \frac{\hat{\pi}_{10|1}^{(t)}}{\hat{\pi}_{10|1}^{(t)} + \hat{\pi}_{11|1}^{(t)}} \times n_{1+|1}; \quad \hat{n}_{11|1}^{(t+1)} = \frac{\hat{\pi}_{11|1}^{(t)}}{\hat{\pi}_{10|1}^{(t)} + \hat{\pi}_{11|1}^{(t)}} \times n_{1+|1}$$

$$\hat{n}_{01|3}^{(t+1)} = \frac{\hat{\pi}_{01|3}^{(t)}}{\hat{\pi}_{01|3}^{(t)} + \hat{\pi}_{11|3}^{(t)}} \times n_{+1|3}; \quad \hat{n}_{11|3}^{(t+1)} = \frac{\hat{\pi}_{11|3}^{(t)}}{\hat{\pi}_{01|3}^{(t)} + \hat{\pi}_{11|3}^{(t)}} \times n_{+1|3}$$

In the M-step the complete data log-likelihood is maximized with respect to the unknown parameters $\pi_{ij|k}$ with $n_{10|1}$, $n_{11|1}$, $n_{01|3}$, and $n_{11|3}$ replaced by their conditional expectations $\hat{n}_{10|1}^{(t+1)}$, $\hat{n}_{11|1}^{(t+1)}$, $\hat{n}_{01|3}^{(t+1)}$, and $\hat{n}_{11|3}^{(t+1)}$, respectively. In this instance we maximize

$$\ell^{(t+1)} = \sum_{i,j} n_{ij|2} \ln(\pi_{ij|2}) + \hat{n}_{10|1}^{(t+1)} \ln(\pi_{10|1}) + \hat{n}_{11|1}^{(t+1)} \ln(\pi_{11|1}) + \hat{n}_{01|3}^{(t+1)} \ln(\pi_{01|3}) + \hat{n}_{11|3}^{(t+1)} \ln(\pi_{11|3})$$

In the M-step a log-linear model with structurally zero cells is fitted to maximize the likelihood. In this instance the complete data has 7 cells and there are 5 structurally zero cells.

After convergence the parameters can then be used to estimate the frequencies for the structurally zero cells. It can be easily verified that the estimate of population size obtained using the EM algorithm is unbiased, and thus it results in an estimate of population size from the collapsed table corrected for the bias found in equation (5).

4.3. Model selection

As there might be several competing models than can be entertained, it is imperative to find a parsimonious model which best fits the data. The likelihood ratio test can be used to discriminate between two competing (log-linear) models. This test compares the difference in deviance ($-2 \times$ log-likelihood) of the two models with the chi-squared distribution for a given the number of degrees of freedom (difference in number of parameters). For this the observed-data likelihood should be used. The researcher can also use the AIC or BIC statistics, which penalize the maximized likelihood for a model by number of parameters. Thus models with more parameters receive a high penalty. The model with the lowest AIC or BIC is preferred. The AIC is the usually preferred model selection method in capture–recapture studies [13, p. 776].

When there are several adequate models which result in different estimates of the population size, basing inferences on one selected model alone is risky [14]. This is because selecting one model ignores the uncertainty in model selection, leading to overconfident inferences. A striking example in the capture–recapture problem is given in Reference [15]. In this instance it is important to take model uncertainty into account when making inferences [14]. The simplest way to incorporate the uncertainty of the model selection process is to use model averaging. If model selection uses the AIC, one can use AIC weights in the model averaging process [16, 17].

The main advantages of model averaging are that it improves predictive performance and coverage [14]. Model averaging avoids the problem of having to defend model choice and thus simplifies presentation. This is because model averaging allows users to incorporate several competing models in the estimation process. Furthermore, model averaging is more correct in that it takes into account a source of uncertainty that analysis based on model selection ignore. In general, this leads to higher estimates of variance than do estimates that ignore model uncertainty [14, p. 398–399].

4.4. Variance estimation

We propose to use the parametric bootstrap [18, 19] to calculate confidence intervals for the point estimates (of the population size). The advantage of the bootstrap method over asymptotic methods is that it is simple. Also, formulae for asymptotic standard errors are available only for the usual approach of multiple-record systems estimation, but not for the situation where some of the registrations are not operating in some strata.

To illustrate the parametric bootstrap we use the example in Section 3.2, where there are two lists and 2 years. The initial step in the bootstrap is to use the EM algorithm to compute $\hat{p}_{00|1}$, $\hat{p}_{10|1}$, $\hat{p}_{01|1}$, $\hat{p}_{11|1}$, $\hat{p}_{00|2}$, $\hat{p}_{10|2}$, $\hat{p}_{01|2}$, $\hat{p}_{11|2}$, $\hat{N}_1$, and $\hat{N}_2$, where

$$\hat{p}_{00|1} = \frac{\hat{n}_{00|1}}{\hat{N}_1}, \quad \hat{p}_{10|1} = \frac{n_{10|1}}{\hat{N}_1}, \quad \hat{p}_{01|1} = \frac{n_{01|1}}{\hat{N}_1}, \quad \hat{p}_{11|1} = \frac{n_{11|1}}{\hat{N}_1} \tag{8a}$$

$$\hat{p}_{00|2} = \frac{\hat{n}_{00|2}}{\hat{N}_2}, \quad \hat{p}_{10|2} = \frac{\hat{n}_{10|2}}{\hat{N}_2}, \quad \hat{p}_{01|2} = \frac{\hat{n}_{01|2}}{\hat{N}_2}, \quad \hat{p}_{11|2} = \frac{\hat{n}_{11|2}}{\hat{N}_2} \tag{8b}$$

To compute the confidence intervals (or variances) the following steps have to be used.

- Sample from a multinomial distribution with index $\hat{N}_1$ and probability vector ($\hat{p}_{00|1}$, $\hat{p}_{10|1}$, $\hat{p}_{01|1}$, $\hat{p}_{11|1}$). Do the same for the second year. If $\hat{N}_1$ and $\hat{N}_2$ are not integers it is simplest to round to the nearest integer [18].
- Remove cells corresponding to cells not observed in the original data table. That is, delete the observations for $n_{00|1}$, $n_{00|2}$, and $n_{01|2}$. Finally, add $n_{10|2} + n_{11|2}$ such that the form of the resulting table is identical to the form of the observed table.
- Use the EM algorithm to get the estimated population sizes for both years.
- Repeat the above steps B times, to get estimates of $\hat{N}_{1(j)}$ and $\hat{N}_{2(j)}$ ($j = 1, \ldots, B$).

The variance of $\hat{N}_1$ and $\hat{N}_2$ is simply the variance of $\hat{N}_{1(j)}$ and $\hat{N}_{2(j)}$ [18]. Using the parametric bootstrap results in standard errors which are not conditional on the observed sample size.

5. APPLICATION

To apply the EM algorithm to the data presented in Section 2 we note that for 1992 the observed array is $2 \times 2 \times 2 \times 2$, and in the E-step it is spread out into a five-dimensional array of $2 \times 2 \times 2 \times 2 \times 2$ using the five-dimensional arrays for years 1993–1998. In 1993–1998 we have one structural zero cell in a year, namely the cell corresponding to observations missed by all lists. For 1992 we have two structural zeros, one corresponding to the observations missed by all lists and one corresponding to the observations which are only contained in the registration not operating in 1992. The inclusion profiles for these observations are 00000 and 00010 (corresponds to observations that would have been observed in R4 only if the registration was active). In 1988–1991 there are four structural zero cells corresponding to the following inclusion profiles, 00000, 00100, 00110, and 00010. The last 3 inclusion profiles correspond to cells that could have been observed if R3 and R4 were operational.

This problem can be related to the general procedure described in Section 4 as follows. In 1993–1998 all lists are active, so these years belong to S_1. All observations in other years are in S_2. For this problem, S_2 can be classified into $J = [2^3 - 1] + [2^4 - 1] = 22$ cells as from 1988 to 1991 the observations for each inclusion profile have the same set of possible cells, and in 1992 the observations have their own set of possible cells. Rather than using i, j, and k, for the illustration we will use the corresponding capture profile and year. For example, the partially classified observation $r_{010|1989} = n_{01++0|1989} = 114$ has four possible cells, $S_{010|1989} = \{c_{01000|1989}, c_{01100|1989}, c_{01010|1989}, c_{01110|1989}\}$, that is if all registrations were active, the frequency $n_{01000|1989}$ would have been distributed over these cells.

Table V presents a summary of the models fitted to the data. Year (Y_{cat}) is used as a stratifying variable in the table. The main effects only model has a poor fit. The approach we followed was to first explore heterogeneity followed by dependence [2]. Thus we begin by adding heterogeneity terms (that is H1 and H2). First-order heterogeneity (H1) results in a big improvement of the fit, but second-order heterogeneity (H2) does not significantly fit

Table V. Selected models with deviance and AIC.

Model	Design matrix	Number of parameters	Degree of freedom*	Deviance	AIC	$\hat{N}^{\dagger}$
1	R1+R2+R3+R4+R5+Y_{cat}	16	213	409‡	441	2229
2	1 + H1	17	212	359‡	393	3009
3	2 + H2	18	211	359‡	395	2822
4	2 + (R1+ R2+R3+R4+R5)Y_{cat}	67	162	191	325	2702
5	2 + (R1+ R2+R3+R5)Y_{cat}	57	172	193	307	2708
6	2 + (R1+ R2+R3)Y_{cat}	47	182	203	297	2697
7	2 + (R1+ R2)Y_{cat}	37	192	213	287	2697
8	2 + R1 Y_{cat}	27	202	280‡	334	3212
9	2 + R2 Y_{cat}	27	202	256‡	310	2683
10	7 + R1(R2+R3+R4+R5)+R2 (R3+R4+R5)+ R3(R4+R5)	46	183	156	248§	2777
11	7+R1(R2+R3+R4+R5)+R2 (R3+R4+R5)+ R3 R5	45	184	156	246§	2778
12	11+(R1(R2+R3+R4+R5)+R2 (R3+R4+R5)+ R3 R5)Y_{cat}	135	94	102	372	3034
13	11+(R1(R2+R3+R5))Y_{cat}	85	144	118	288	2988
14	11+R1 R2 Y_{cat}	55	174	140	250	2990

*NB: There are 229 observed cells (see Table I).
†There are 1783 cases observed at least once (see Table I).
‡Significant at the 5 per cent level of significance.
§Model has substantial support from the data [17, p. 70–71,170].

better than model 2. We then allow the inclusion probabilities to vary by year, and it turns out that only the inclusion probabilities for R1 and R2 vary over time (model 7) but the other registrations do not (models 4–6). Model 11 shows that the registrations are pairwise related except for R3 and R4 (as the R4 and R5 interaction is set to zero in order to estimate H1). Models 12–14 allow the interactions to vary over time but none of these models lead to an improvement in fit.

The models with substantial support from the data, that is models with an AIC less than or equal to 2 from the AIC of the model with the lowest AIC [17, p. 70–71,170], are models 10 and 11. The yearly estimates of the population size for these two models is basically the same. If this was not the case, model uncertainty has to be incorporated in the estimates of the population size and their variances [16]. This implies using one of these two models does not lead to overconfident inferences. The model with the lowest AIC is model 11 (see Table V) and this model will be used for the estimation of the yearly estimated population sizes and confidence intervals.

To compute the confidence intervals for the yearly estimates of population size the parametric bootstrap with 500 replications is used (see Table VI). The confidence intervals show that, most often, the distribution of the estimates by year is skewed. Furthermore, years with a higher number of structurally zero cells (1988–1992) have somewhat wider confidence intervals.

The table also shows estimates from the standard capture–recapture methods, that is log-linear models [6–8] and the sampling coverage approach [3, 20]. These models were fitted to

Table VI. Estimates of population size and 95 per cent confidence intervals by year.

		EM algorithm		Log-linear			Sample coverage	
Year	Observed	$\hat{N}$	95 per cent C.I.	Model*	$\hat{N}$	95 per cent C.I.†	$\hat{N}$	95 per cent C.I.
1988	145	238	[161, 290]	[12,5]	311	[200, 648]	231‡	[188, 318]
1989	163	204	[163, 243]	[1,25]	174	[161, 192]	181	[164, 606]
1990	170	231	[185, 269]	[1,25]	177	[168, 189]	190	[171, 581]
1991	150	187	[152, 227]	[12,15]	191	[149, 282]	185	[156, 360]
1992	172	286	[211, 319]	[12,23,5,H1]	782	[326, 2687]	303‡	[249, 395]
1993	160	220	[193, 264]	[12,15,24,34,45]	320	[207, 957]	233	[187, 356]
1994	162	275	[235, 355]	[14,15,24,34,35]	232	[197, 293]	265	[197, 464]
1995	174	307	[263, 396]	[12,13,23,34,35,45]	206	[188, 231]	200	[182, 257]
1996	153	269	[233, 345]	[12,13,24,25,34,45]	317	[220, 583]	255	[181, 527]
1997	180	306	[268, 380]	[12,14,15,24,34,35,45]	351	[259, 595]	340	[213, 952]
1998	154	254	[220, 319]	[14,23,24,25,34,45]	212	[179, 266]	248	[200, 346]

*1,2,3,4, and 5 refer to R1, R2, R3, R4 and R5, respectively. H1 refers to first-order heterogeneity [2].
†C.I. computed using parametric bootstrap with 500 replications.
‡One step estimator used. In 1988 the coverage is low, i.e. less than 55 per cent [3, p. 3137], and in 1992 the *s.e.* is very large rendering the estimate useless.

each year separately using the program CARE-1 (which is downloadable from http://chao.stat.nthu.edu.tw/). Most of the yearly estimates from the sampling coverage approach are consistent with the estimates from the EM algorithm, except for a couple of years. This is not true for the log-linear models. As the estimates from the traditional approaches do not use information from the other years they tend to be more variable.

In conclusion, we stress that although in our example it is possible to use traditional approaches within each year, this is not possible where in one or more strata only one list is operating (compare Section 3). If a stratum has only two active lists, traditional approaches assume independence between the lists, whereas the EM algorithm utilizes the dependence between the lists in other strata. A joint model also decreases the possibility of chance capitalization. Fitting a joint model is also more efficient and the resulting estimates are more stable.

6. CONCLUSIONS AND DISCUSSION

We have show how the population size can be estimated using the multiple system estimator when the registrations emanate from partially overlapping populations. In epidemiology, there is a tendency for different institutions to collect data on the same diseased population, and in most cases not all cases are ascertained. Furthermore, some registrations might concentrate on special subgroups of the population, for example children or the elderly, such that the usual multiple system estimator cannot be used. The approach we presented can be useful in such situations.

This method will also be attractive to ecologists in cases where due to the nature of their surveys, certain groups of animals are excluded *a priori* from the surveys. For instance in the two sample case: both large and small animals are captured and attached tags in the first sample, whereas only large animals are permitted to be caught in the second sample, the

Peterson–Lincoln estimator is still valid because sample one is a random sample under the assumption that the samples observed in both strata have homogeneous capture probabilities. Similar statements can be said for cases where large, medium and small animals denote a strata.

In the two list case the traditional multiple system estimator does not account for any possible dependence between the lists. If there is dependence, this approach also does not provide a correct estimate of the population size. This problem though can be minimized by the inclusion of (categorical) covariates, such that independence is assumed at each level of the covariates.

ACKNOWLEDGEMENTS

The authors wish to thank the three anonymous referees for their helpful comments and suggestions, which have greatly improved the presentation in the paper.

REFERENCES

1. Seber GAF. *The Estimation of Animal Abundance and Related Parameters*. Macmillan: New York, 1982.
2. International Working Group for Disease Monitoring and Forecasting (IWGDMF). Capture–recapture and multiple-record systems estimation 1: history and theoretical development. *American Journal of Epidemiology* 1995; **142**:1047–1058.
3. International Working Group for Disease Monitoring and Forecasting (IWGDMF). Capture–recapture and multiple-record systems estimation 2: applications. *American Journal of Epidemiology* 1995; **142**:1059–1068.
4. Chao A, Tsay PK, Lin S, Shau W, Chao D. The applications of capture–recapture models to epidemiological data. *Statistics in Medicine* 2001; **20**:3123–3157.
5. Alho JM. Logistic regression in capture–recapture models. *Biometrics* 1990; **46**:623–635.
6. Fienberg SE. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* 1972; **59**:591–603.
7. Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis*, *Theory and Practice*. McGraw-Hill: New York, 1975.
8. Cormack JM. Log-linear models for capture–recapture. *Biometrics* 1989; **45**:395–413.
9. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1975.
10. Baker SG. A simple EM algorithm for capture–recapture data with categorical covariates. *Biometrics* 1994; **46**:1193–1200.
11. Van der Pal KM, Van der Heijden PGM, Buitendijk SE, Den Ouden AL. Periconceptional folic acid use and the prevalence of neural tube defects in the Netherlands. *European Journal of Obstetrics Gynecology and Reproductive Biology* 2003; **108**:33–39.
12. Rice JA. *Mathematical Statistics and Data Analysis*. Duxbury Press: California, 1995.
13. Hook EB, Regal RR. Accuracy of alternative approaches to capture–recapture estimates of disease frequency: Interval validity analysis of data from five sources. *American Journal of Epidemiology* 2000; **152**:771–778.
14. Regal RR, Hook EB. The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine* 1991; **10**:717–721.
15. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Statistical Science* 1999; **14**:382–417.
16. Stanley TR, Burnham KP. Information–theoretic model selection and model averaging for closed-population capture–recapture studies. *Biometrical Journal* 1998; **40**:475–494.
17. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference*: *A Practical Information–Theoretic Approach*. Springer: New York, 2002.
18. Buckland ST, Garthwire PH. Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics* 1991; **47**:255–268.
19. Norris JL, Pollock KH. Including model uncertainty in estimating variances in multiple capture studies. *Environmental and Ecological Statistics* 1996; **3**:235–244.
20. Tsay PK, Chao A. Population size estimation for capture–recapture models with applications to epidemiological data. *Journal of Applied Statistics* 2001; **28**:25–36.

