Journal of Data Science **6**(2008), 557-572

Capture-recapture Studies with Incomplete Mixed Categorical and Continuous Covariates

Eugene Zwane[1] and Peter van der Heijden[2]
[1]*Imperial College and* [2]*Utrecht University*

Abstract: Registrations in epidemiological studies suffer from incompleteness, thus a general consensus is to use capture-recapture models. Inclusion of covariates which relate to the capture probabilities has been shown to improve the estimate of population size. The covariates used have to be measured by all the registrations. In this article, we show how multiple imputation can be used in the capture-recapture problem when some lists do not measure some of the covariates or alternatively if some covariates are unobserved for some individuals. The approach is then applied to data on neural tube defects from the Netherlands.

Key words: Capture-recapture, log-linear model, multinomial logit model, multiple imputation, population size estimation.

1. Introduction

The estimation of the population size based on multiple incomplete lists has a long history (Chao *et al.*, 2001, Schwarz and Seber, 1999). The advantages of using these methods as a substitute for direct counting in epidemiology has been strongly emphasized (International Working Group for Disease Monitoring and Forecasting, 1995). The basic assumptions are that the population being estimated is closed, i.e., births, deaths and migrations are negligible, the individuals can be matched without error, and for the traditional approach an additional assumption is that all individuals have the same probability of being ascertained by a registration. In recent times this additional assumption is relaxed by allowing the capture probabilities to depend on covariate information or by allowing some of the registrations to be dependent.

A serious problem in capture-recapture models with individual level covariates occurs when the data are missing on one or more covariates which define heterogeneous catchability. Item missing values are usually handled by imputation with a reasonable proxy (Zwane and Van der Heijden, 2004) or by excluding

those observations (Hwang and Huang, 2003; Wang and Yip, 2003). The missing-data problem is more acute when some of the registrations do not contain some of the covariates which define heterogeneous catchability. In epidemiology and public health contexts this is a common problem as the registrations used are usually compiled for different administrative purposes. The standard approach is to simply drop these covariates. On top of being a waste of data, this practice could lead to biases (invalid results) if the dropped covariates are sources of heterogeneity. As a result it is of interest to incorporate missing data techniques into capture-recapture studies (Wang and Yip, 2003). A related problem is when the lists do not measure the same population (Zwane *et al.*, 2004).

This problem confronted us in the estimation of the prevalence of neural tube defects in the Netherlands using the capture-recapture (CRC) methodology. The data utilized three overlapping registrations, where the gender of child and delivery weight are measured in all the registrations. Parity of the child, nationality (proxy for ethnicity) and the age of the mother are measured in only two of the registrations. Under certain assumptions, two valid approaches can be used to estimate the population size, (i) use all covariates and the two lists where all covariates are measured, or (ii) use all lists and the covariates measured in all lists. In the first approach, one assumes that given all covariates the lists are independent, whilst in the second approach dependencies between lists can be entertained. As a result these two approaches can result in different estimates of the population size. If, given all covariates some of the lists are dependent and if all covariates influence inclusion to all lists, both these approaches will be biased.

Due to the arbitrary nature of choosing either estimating approach and that these approaches are only valid under certain assumptions, it is preferable to formulate a model that utilizes all covariates and all lists. In the CRC problem, Zwane and Van der Heijen (2007) considered a log-linear model to describe the multinomial probabilities among discrete (and discretized) covariates and suggested the use of the EM algorithm (Little and Rubin, 1987) for likelihood maximization when some covariates do not appear in some registrations. This approach can be used for our data but then we would ignore part of the information available in the continuous covariates.

The EM algorithm can in principle be used for capture-recapture models with continuous covariates, but the expectation involves complex numerical integration. In semiparametric capture-recapture models in continuous-time Wang (2005) used the Monte Carlo EM (Wei and Tanner, 1990) to estimate the parameters and in turn the population size. In this paper we prefer the use of multiple imputation (MI). The main advantage of MI over maximum likelihood methods is that it is computationally much simpler for most practical situations (Sinharay *et al.*, 2001). MI was developed in the context of non-response in sample surveys,

for handling missing data with arbitrary patterns of missing data on both continuous and categorical covariates, but has been applied successfully in a broad variety of settings (Rubin 1987, 1996). MI has been used extensively in similar problems, *i.e.*, problems utilizing different surveys or registrations (Belin *et al.*. 1993; Gelman *et al.*, 1998; Raghunathan and Sciskovik, 1998). In all these cases MI was found to perform well. For example, Raghunathan and Sciskovik (1998) compared the use of MI in case control studies where the exposure variable is available from several sources to the use of naïve methods (where the exposure variable comes from one source), and they found that using several sources is better in terms of bias, mean square error and confidence coverage.

The rest of the paper is organized into five sections. In Section 2 we briefly discuss the neural tube defects data set. We discuss the multinomial logit model for analyzing CRC data with continuous covariates in Section 3. Section 4 introduces multiple imputation in the capture-recapture problem. In Section 5, we report an analysis of the neural tube defects data set presented in Section 2, and conclude with a discussion in Section 6.

2. Neural Tube Defects Data

The problem that triggered this work is the estimation of the number of children born with a neural tube defect in the Netherlands from 2000. The data are gathered routinely on children born with a neural tube defects (NTD's) in the Netherlands by midwives, obstetricians, or paediatric units (Van der Pal *et al.*, 2002). For this analysis we utilized three incomplete but overlapping databases which we describe briefly;

1. *The Dutch Perinatal Database 1 (LVR_1)*: This is a pregnancy and birth registry of low risk pregnancies and births.

2. *The Dutch Perinatal Database 2 (LVR_2)*: Registers data concerning the birth of a child in secondary care. If a woman is referred from primary care to secondary care she may be registered in LVR_1 and LVR_2.

3. *The National Neonatal Database (LNR)*: Contains information on "all" admissions and re-admissions of newborns to paediatric departments within the first 28 days of life.

In each of these registries, the gender and the birth (or delivery) weight of the child are recorded (fully overlapping covariates). In LVR_1 and LVR_2, there is also information on parity of child, ethnicity/nationality and the age of the mother which are not measured in LNR. A summary of the data is shown in table 1.

Table 1: Neural tube defects data : 2000

	Ascertainment history†							
Covariates	100	010	001‡	110	101	011	111	Total
Gender								
Male	19	14	7	11	5	10	3	69
Female	24	23	9	13	2	7	1	79
Total	*43*	*37*	*16*	*24*	*7*	*17*	*4*	*148*
Birth weight								
Mean	3.21	2.34	2.74	2.15	3.36	2.73	3.05	2.72
Ethnicity/Nationality								
Dutch	34	27	?	21	4	16	2	104
Non-Dutch	9	10	?	3	3	1	2	28
Missing	0	0	16	0	0	0	0	16
Parity								
Mean	1.67	2.14	NA	1.54	2.29	1.59	2.00	1.81
Age of the mother								
Mean	30.02	30.16	NA	29.25	30.29	28.24	29.25	29.68

†The first element of the ascertainment profile refers to LVR_1, the second to LVR_2, and the third to LNR (1 is present, 0 is absent).
‡Observations listed in 'LNR only' have a value for birth weight only.

The variables in table 1 are known (or believed) to be risk factors for neural tube defects (Olney and Mulonnare, 1998; Davidoff *et al.*, 2002; Vieira, 2004) or have been used successfully in the capture-recapture methodology (Madigan and York, 1997). We had no reason to believe that the gender of the child and parity influence the capture probability. Note that we assume that parity doesn't influence the capture probability even though women having second or later births can spend less time in maternity units compared to first-time mothers. This is based on the assumption that conditional on a child having an NTD (where safety of the child is paramount) first-time mothers do not necessarily spend more time in maternity units.

3. The Multinomial Logit Model in the CRC Problem

Assume that the true population size is N and the individuals are indexed by i ($i = 1, 2, ..., N$) of which n are ascertained by at least one of S registrations. The inclusion profile for individual i is the vector $\boldsymbol{w}_i = [i_1 i_2 \cdots i_S]$, which is a series of binary variables with 1 denoting ascertained and 0 otherwise. The ascertainment profile $\boldsymbol{w}_i$ can be redefined as a nominal categorical variable Y_i with $K = 2^S - 1$ levels, indexed by k ($k = 1, \cdots, K$) with individual i falling in only one of the categories.

Now assume that for individual i there are covariate vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ of length

p and q respectively, where $\boldsymbol{x}_i$ are the covariates observed in all the registrations and $\boldsymbol{z}_i$ are the covariates not observed in all the registrations. Denoting the multinomial logit for individual i as $\eta_i' = [\eta_1(\boldsymbol{x}_i, \boldsymbol{z}_i), \eta_2(\boldsymbol{x}_i, \boldsymbol{z}_i), ..., \eta_K(\boldsymbol{x}_i, \boldsymbol{z}_i)]$, the category probabilities are then given by,

$$\mathbb{P}(Y_i = k|\boldsymbol{x}_i, \boldsymbol{z}_i) = \exp[\eta_k(\boldsymbol{x}_i, \boldsymbol{z}_i)] / \sum_{r=1}^{K} \exp[\eta_r(\boldsymbol{x}_i, \boldsymbol{z}_i)]. \tag{3.1}$$

This model has to be constrained in some way for it to be used in the capture-recapture problem (Zwane and Van der Heijden, 2003, 2004). Alho (1990) and Huggins (1989) constrained the logits such that the lists are independent at the individual level. After fitting the model the parameters can be used to estimate the probability that an individual is registered or listed at least once. Denoting this probability by ϕ_i (the estimated probability is denoted by $\hat{\phi}_i$), the estimate of the population size is

$$\hat{N} = \sum_{i=1}^{n} \hat{N}_i = \sum_{i=1}^{n} \frac{1}{\hat{\phi}_i},$$

where $\hat{N}_i$ is the contribution of individual i to the estimate of the population size (Huggins, 1989).

Rather than use (3.1), the current standard is to use only the covariates observed in all lists, that is

$$\mathbb{P}(Y_i = k|\boldsymbol{x}_i) = \exp[\eta_k(\boldsymbol{x}_i)] / \sum_{r=1}^{K} \exp[\eta_r(\boldsymbol{x}_i)]. \tag{3.2}$$

Equation (3.2) will result in a biased estimate of the population size if the covariates in $\boldsymbol{z}_i$ are related to the inclusion probabilities. In this article we will complete the data set using the multiple imputation approach described in section 4 such that all covariates and lists are utilized.

4. Multiple Imputation in the CRC Problem

In this section, we will briefly describe the idea of multiple imputation methods. Multiple imputation is now standard in statistical literature and thus we will highlight only the most important points (Rubin, 1996). MI involves three steps: 1) imputing the data under an appropriate model and repeating the imputation to obtain m copies of the filled-in data set; 2) analyzing each data set separately to obtain the desired parameter estimates and standard errors; 3) combining the results from the m parameter estimates by computing the mean of the m parameter estimates and a variance estimate that includes both within-imputation and

an across-imputation components. Below we describe how the multiply imputed data sets are created and how the analysis of such data can be performed to result in one estimate of the population size and its standard error.

4.1 Creating multiple imputed data sets

Multiple imputation aims at imputing the missing values in $\boldsymbol{z}_i$ such that they can also be used in generally available software, like the multinomial logit model. Possible multivariate models for the data that can be used to draw the m plausible values for each missing item in the data set are the multivariate normal model, the general location model, or by using "compound conditional specification". A number of software programs are available implementing these models (Horton and Lipsiz, 2001). Below we highlight the features of each of these approaches and situations where they can be used for creating the multiple imputations.

Multivariate normal model

A multivariate normal model with arbitrary covariance and correlation structure can be used for the imputation. In the capture-recapture problem this approach can be used when there are no missing values in categorical variables. The variables forming the inclusion profiles are binary, but because they have no missings they can enter the model as continuous covariates (Schafer, 1997). Note that in some cases even in the presence of missing binary or ordinal variables the multivariate normal model can still be used, but as noted by Horton *et al.* (2003) this practice can sometimes lead to a bias.

General location model

This model was introduced by Olkin and Tate (1961) to characterize the joint distribution for data containing a mix of categorical and continuous covariates. This model assumes a multinomial distribution for the categorical variables and a multivariate normal distribution for the continuous variables within each cell of the contingency table. Belin *et al.* (1999) gives a discussion on the performance of the general location model with an ignorable missing data assumption in a mental health services study. They also give several considerations that have to be taken into account before using the general location model.

Compound conditional specification

Also called "incompatible Gibbs sampling"; this approach specifies a different regression model for each variable (Van Buuren and Oudsjoorn, 1999; Raghunathan *et al.*, 2001). For categorical variables, the model could be logistic or

multinomial and for continuous variables, the linear regression model is sufficient. Imputation is done on an equation by equation basis. The problem of dimensionality of multivariate regression is reduced.

4.2 Selection of covariates

The registrations used in capture-recapture problems usually contain a wealth of covariates and these can also be used for imputations. Ideally all variables have to be used in the imputation model to make the missing at random (MAR) assumption more plausible (Rubin, 1996). In some instances, especially in the general location model use of a large number of categorical covariates results in an unestimable model. Belin *et al.* (1999) illustrated an approach which is a trade-off between trying to accommodate more detail in the incomplete data model and the ability to estimate parameters of the model.

4.3 Analysis

Once the model has been chosen and variables selected, the model can be used to generate via a random sampling procedure m imputed values for the missing data points, thus creating m complete data sets. For each completed data set an estimate of the population size ($\hat{N}_d$, where $d = 1, \cdots, m$) and its associated variance which we denote by $\hat{\text{var}}[\hat{N}_d]$ can be computed using capture-recapture models with continuous covariates (Alho, 1990; Zwane and Van der Heijden, 2004). These estimates can then be combined using the approach of Rubin (1987) to arrive at a single estimate of the population size ($\hat{N}$) given by,

$$\hat{N} = \frac{1}{m}\sum_{d=1}^{m}\hat{N}_d \tag{4.1}$$

and an MI variance, $\hat{\text{var}}[\hat{N}]$ given by,

$$\hat{\text{var}}[\hat{N}] = \frac{1}{m}\sum_{d=1}^{m}\hat{\text{var}}[\hat{N}_d] + \left[1+\frac{1}{m}\right]\left[\frac{1}{m-1}\right]\sum_{d=1}^{m}\left(\hat{N}_d - \hat{N}\right)^2. \tag{4.2}$$

This variance includes two parts: the average within-imputation variance, which is the first part of (4.2), and the between-imputation variance.

4.4 Model selection

Most of our presentation thus far assumed that there is a single true model which is usually arrived at by some model search criterion. A cause for concern

in capture-recapture models is that estimated population sizes for models with similar fits can be different, and thus basing inferences on a single model is (Hoeting *et al.*, 1999). To overcome this concern we propose to incorporate model uncertainty into our estimates using the model averaging approach (Stanley and Burnham, 1998). This approach allows for model selection uncertainty to be incorporated into the standard errors and reduces bias in the parameter estimates in cases when there are a number of models with similar Aikake Information Criterion's (AIC's) with (substantially) different estimates of the population size and/or their standard errors. As our model selection is based on the AIC, AIC weights will be used in the model averaging process.

To compute the AIC weights we first compute the difference in AIC between each model and the model with the lowest AIC as $\Delta_i = AIC_i - AIC_{min}$, where AIC_i and AIC_{min} are the AIC's for model i and the model with the lowest AIC respectively. Using Δ_i the AIC weights are

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_j \exp(-\Delta_j/2)}$$

where '$\exp(-\Delta_i/2)$' is the likelihood of the model given the data (Burnham and Anderson, 2002, Chapter 4.2). Using the weights the model averaged estimate of the population size is $\hat{N}^* = \sum_i w_i \hat{N}_i$ where $\hat{N}_i$ is the estimate of the population size for model i. The variance of the estimate is given by,

$$\hat{\text{var}}[\hat{N}^*] = \left[\sum_i w_i \sqrt{\hat{\text{var}}[\hat{N}_i] + (\hat{N}_i - \hat{N}^*)^2}\right]^2 .$$

In most cases not all the models are included in the estimation as most of them will have insignificant weights. Burnham and Anderson (2002, 70-72) proposed a rule-of-thumb where models with $\Delta_i < 4$ have substantial support and models with $4 < \Delta_i < 7$ have considerably less support. In our analysis the models included in the model averaging are those with at least substantial support, that is models with $\Delta_i < 4$.

In the analysis of multiply imputed data, we opt to create $m + 1$ imputed datasets where the extra dataset is used for model selection (Allison, 2001). In this case we use a somewhat conservative criteria (Allison, 2001) for models to be included in the model averaging process. For this exercise we have considered models with at least considerably less support from the data (Burnham and Anderson, 2002, p.70-72). An alternative is to use the procedure for calculating the complete-data log likelihood ratio (and corresponding p-value) for analysis on multiple imputed data proposed by Meng and Rubin (1992). The *ad hoc* approach is preferred due to its simplicity. However, the *ad hoc* approach might

become misleading in situations where the missingness is severe; this scenario requires investigation.

5. Application

The method presented was applied to the data from the neural tube defects data set described in Section 2. In our imputation model we used all available covariates, but consider only birth weight, ethnicity and the age of the mother as covariates that can possibly have an effect on the inclusion probability. For completeness we first present the results from the traditional approaches.

5.1 Traditional approach

In this section we present the estimates of the population size based on the two valid traditional approaches. We first consider the models using all covariates and the two lists (LVR_1 and LVR_2). The estimates from all possible models range from 228 to 250. The model averaged estimate of the population size is 242 (SE = 41.88, where SE denotes standard error). The log-based confidence interval (Chao, 1987) is $[183, 369]$. All models excluding delivery weight do not have support from the data.

We then consider models using all lists and delivery weight. Estimates from these models range from 183 to 275. The model averaged estimate of the population size is 215 (SE = 43.23), implying the log-based confidence interval is $[168, 373]$.

It is evident that the model averaged estimates from these two approaches are substantially different and it is not clear how one approach can be chosen over the other. Using two lists, one can only assume independence at the individual level but it is possible that this is fulfilled given the available covariates. With three lists, dependencies can be modelled, but it is likely that the dropping of covariates of heterogeneous catchability induces dependence resulting in biased results. In the following section, imputation approaches are used to complete the data set such that all features of the data are utilized.

5.2 Imputation

The difference between the estimate based on "all covariates and two lists" and the estimate based on "all lists and delivery weight" shows that there is a need for an approach which utilizes all lists and all covariates. In this section we use imputation approaches to complete the data set, and then fit a model which uses all covariates and all lists. For ease of exposition we only consider models where the covariates enter linearly. A simple solution is conditional mean

imputation, which is discussed in the next Section.

Conditional mean imputation

As the missingness occurs in only 11% (16/148 = 0.11) of the observations, single imputation can give reasonable results (Harrell, 2001). We used the proportion of women with children listed in LNR which is 0.214 as a proxy for ethnicity to impute the missing values for observations listed in LNR only. The mean age for women with children listed in LNR which is 28.9 was used to impute the age of the mother for the observations listed in LNR only.

The model averaged estimate of the population size is 232 (SE = 59.29), implying the log-based confidence interval is $[171, 450]$. This estimate is not that different from the analysis utilizing the two lists (LVR_1 and LVR_2) all covariates save for a higher upper confidence limit, but it is very different from the estimates utilizing all lists and delivery weight. This might lead one to conclude that the dependence between the lists is weak even though all the models with support from the data in the imputed data set incorporate dependencies between the lists. A feature common to both sets of analysis using all covariates is that models including the age of the mother tend to have a higher estimate of the population size.

A problem with mean imputation is that the standard errors are likely to be underestimated. To avoid the underestimation of the standard errors we apply multiple imputation in the next Section.

Multiple imputation

In this section we use multiple imputation techniques to analyze the neural tube defects data set. The covariates with missing values to be used in our analysis model are ethnicity, which is a binary covariate, and the age of the mother which is a continuous covariate. As recommended by Horton *et al.* (2003) it is preferable to use a discrete model even when confronted by a problem with missings in only binary variables. The discrete model we use is the general location model as implemented in the R (Ihaka and Gentleman, 1996) library MIX (Schafer, 1997).

We used the EM and data augmentation (DA) algorithms in MIX to generate the posterior distribution of the parameters of the assumed model. Random draws from the posterior distribution were then taken $m = 10$ times to generate $m = 10$ complete data sets for the final analysis. As recommended the parameter estimates from the EM algorithm are used as starting values for the DA algorithm (Schafer, 1997). To ensure that the successive imputations are statistically independent, the DA algorithm was run 27500 times and at every 2500 iterations one of $m + 1 = 11$ imputations was selected.

All possible models were fitted to the extra data set and the models with least considerably less support from the data, or alternatively, the models with $\Delta_i < 7$ are shown in table 2.

Table 2: Estimates of population size for the all covariates models

			Model	Multiple Imputation		
	Design	Covariate	Selection	Est.		
Model [5]	Matrix [2]	Matrix [3]	AIC	Popu.	*s.e.*	95 % C.I. [4]
1	$[12, 3]$	$1 + B$	507.6	211	23.48	[179, 276]
2	$[13, 2]$	$1 + B$	508.0	199	16.23	[175, 241]
3	$[13, 2]$	$1 + B + E$	507.1	206	20.74	[177, 263]
4	$[13, 2]$	$1 + B + E + A$	506.6	222	31.27	[181, 312]
5	$[1, 23]$	$1 + B$	505.9	236	27.42	[196, 308]
6	$[1, 23]$	$1 + B + E$	509.3	236	27.44	[196, 308]
7	$[12, 13]$	$1 + B$	504.0	183	15.36	[163, 227]
8	$[12, 13]$	$1 + B + E$	503.2	213	58.83	[162, 444]
9	$[12, 23]$	$1 + B + E + A$	504.5	274	142.52	[169, 895]
10	$[12, 23]$	$1 + B$	506.3	275	100.97	[180, 649]
11	$[13, 23]$	$1 + B$	504.0	226	32.46	[184, 319]
12	$[13, 23]$	$1 + B + E$	504.3	238	41.35	[186, 360]
13	$[13, 23]$	$1 + B + E + A$	505.6	259	60.83	[189, 451]
14	$[12, 13, 23]$	$1 + B$	506.9	193	38.74	[158, 342]
15	$[12, 13, 23]$	$1 + B + E$	507.5	306	248.61	[166, 1557]

[†]The first element of the ascertainment profile refers to LVR_1, the second to LVR_2, and the third to LNR (1 is present, 0 is absent).
[‡]Observations listed in 'LNR only' have a value for birth weight only.

Columns 4 in table 2 relates to the AIC of the data set used for model selection, whilst the last three columns relate to the analysis of the rest of the data sets combined using the methods discussed in section 4.3.

The model averaged estimate of the population size, using the AIC's in table 2 is 228 (SE = 64.10), implying the log-based confidence interval is $[167, 476]$. This estimate is marginally different from the estimate using all lists and and birth weight, but also very different from the estimate using the two lists which measure all available covariates (LVR_1 and LVR_2). The estimate also compares favourably with the results using conditional imputation. What is evident from the standard error and confidence interval is that using the multiple imputation approach also adds noise to the analysis.

A question remains is the MAR assumption justified in this analysis. First as noted by Zwane and Van der Heijden (2007) the MAR assumption is violated if the true model is more complex than the most complex model that can be fitted to the data. In our case, if the true model includes $LVR_1 : LVR_2 : E : M$ then

the MAR assumption is violated. We have no reason to believe that true model includes this interaction and thus we believe the MAR assumption is justified.

In conclusion, for this problem using conditional mean imputation and MI results in similar estimates of the population size but in the MI analysis there is a added uncertainty in the estimate. Dependence between registrations plays a big role as the estimates using all lists are comparable. All models with support from the data include birth weight of the child and ethnicity and the age of the mother do not seems to have a big influence in the estimate of the population size. That said, as this information is not available 'a priori' it is advisable to use the imputation approaches.

6. Concluding Remarks

In capture-recapture models it is desirable to include individual level covariates to account for any differences in ascertainment by the registrations. When these covariates are not measured by all registrations (or they contain missing data), the commonly used approaches of dropping (or ignoring) these covariates may give biased estimates of the population size. Multiple imputation is proposed to handle the missing covariate problem in the capture-recapture models.

Our results show that mean imputation also performs well with respect to the estimate of the population size but seemingly underestimates the standard error, resulting in narrow confidence intervals. The estimate of the population size from mean imputation is similar to the estimate derived from multiple imputation because the proportion of observations with missing data is very low.

Multiple imputation is applicable to missing covariate problems with arbitrary missing data patterns and arbitrary number of covariates (the categorical covariates do not necessarily have to be binary). Though our application is in epidemiology with only three lists this approach is applicable to wide ranging capture-recapture problems. Based on the results presented in the previous sections we can make a strong recommendation for the use of imputation in capture-recapture models with missing covariates.

A concern in using capture-recapture methods for our data is that some babies born with NTDs might not survive the initial 28 days of life. If children who die during the first 28 days of life are less/more likely to appear in *LNR* the estimate of the population size will be biased. In our analysis this bias is minimized by controlling for birth weight and focusing only on live births. In essence we assume that after controlling for birth weight, the probability of being missed by *LNR* does not depend on whether a baby died within 28 days of life.

Although we have not evaluated the performance of MI in the capture-recapture problem with missing covariates results show that the performance of MI is similar to likelihood methods that make similar arguments (see Schafer and Graham,

2002, p. 170). If the same model is used for imputation and analysis then MI produces answers similar to likelihood analysis under the same model. In our case the imputation model is more complex than the analysis model, and if parity and gender of child are useful in predicting missing ethnicity and the age of the mother then MI analysis improves in power (see Schafer and Graham, 2002, p. 170). Using these arguments we conclude that as confirmed by a simulation that likelihood methods are unbiased if the data are MAR then MI will be unbiased if the data are MAR.

In conclusion, we stress that for problems where standard maximum likelihood methods under the model of interest can be done easily for a data set with missing values, see for example Zwane and Van der Heijden (2007), then maximum likelihood methods will be preferable to MI because they are more efficient (Sinharay *et al.*, 2001). However maximum likelihood estimation is difficult for problems with missing continuous covariates. Another advantage of MI is that standard errors are available as part of model estimation, whilst if one uses the EM algorithm the confidence intervals are computed using the (parametric) bootstrap or other techniques (Zwane and Van der Heijden, 2007).

References

Alho, J. (1990). Logistic regression in capture-recapture models. *Biometrics* **46**, 623-635.

Allison, P. (2001). *Missing data*. Sage University Papers Series on Quantitative Applications in Social Sciences.

Belin, T., Diffendal, G., Mack, S., Rubin, D., Schafer, J. and Zaslavsky, A. (1993). Hierarchical logistic regression model for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Association* **88**, 1149-1159.

Belin, T., Hu, M., Young, A., and Grusky, O. (1999). Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. *Statistics in Medicine* **18**, 3123-3135.

Burnham, K. and Anderson, D. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783-791.

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

Davidoff, M., Petrini, J., Damus, K., Russell, R., and Mattison, D. (2002). Neural tube defects-specific infant mortality in the United States. *Teratology* **66**, S17-S22.

Gelman, A., King, G., and Liu, C. (1998). Not asked and not answered: Multiple Imputation for multiple surveys. *Journal of the American Statistical Society* **93**, 846-857.

Harrell, F. (2001). *Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science* **14**, 382-417.

Horton, N. and Lipsitz, S. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician* **55**, 244-254.

Horton, N., Lipsitz, S., and Parzen, M. (2003). A potential for bias when rounding in Multiple Imputation. *The American Statistician* **57**, 229-232.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**, 133-140.

Hwang, W. and Huang, S. (2003). Estimation in capture-recapture models when covariates are subject to measurement errors. *Biometrics* **59**, 1113-1122.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299-314.

International Working Group for Disease Monitoring and Forecasting (1995). Capture-recapture and multiple record systems estimation 2: applications. *American Journal of Epidemiology* **142**, 1059-1068.

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

Madigan, D. and York, J. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84**, 19-31.

Meng, X. and Rubin, D. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103-111.

Olkin, I. and Tate, R. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annal of Mathematical Statistics* **32**, 448-465, (correction in **36**, pp. 343-344).

Olney, R. and Mulinare, J. (1998). Epidemiology of neural tube defects. *Mental Retardation and Development Disabilities Research Reviews* **4**, 241-246.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27**, 85-97.

Raghunathan, T. and Sciskovik, D. (1998). Combining exposure information from various sources in an analysis of case-control data. *The Statistician* **47**, 333-347.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Rubin, D. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Society* **91**, 473-489.

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

Schafer, J. and Graham, J. (2002). Missing data: our view of the state of the art. *Psychological methods* **7**, 147-177.

Schwarz, C. and Seber, G. (1999). A review of estimating animal abundance III. *Statistical Science* **14**, 427-456.

Sinharay, S., Stern, H., and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods* **6**, 317-329.

Stanley, T. and Burnham, K. (1998). Information-theoretic model selection and model averaging for closed-population capture-recapture studies. *Biometrical Journal* **40**, 475-494.

Van Buuren, S. and Oudshoorn, C. (1999). Flexible multivariate imputation by mice. *Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054*.

Vieira, A. (2004). Birth order and neural tube defects: a reappraisal. *Journal of Neurological Sciences* **217**, 65-72.

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

Wang, Y. (2005). A semiparametric regression model with missing covariates in continuous-time capture-recapture studies. *Australian and New Zealand Journal of Statistics* **47**, 287-297.

Wang, Y. and Yip, P. (2003). A semiparametric model for capture-recapture experiments. *Scandinavian Journal of Statistics* **30**, 667-676.

Wei, G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Society* **85**, 699-704.

Zwane, E. and Van der Heijden, P. (2004). Semiparametric models for capture-recapture studies with covariates. *Computational Statistics and Data Analysis* **47**, 729-743.

Zwane, E. and Van der Heijden, P. (2005). Population estimation using the multiple-system etimator in the presence of continuous covariates. *Statistical Modelling* **5**, 39-52.

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

Zwane, E., Van der Pal, K., and Van der Heijden, P. (2004). The multiple-record systems estimator when registrations refer to different but overlapping populations. *Statistics in Medicine* **23**, 2267-2281.

Received March 7, 2007; accepted May 2, 2007.

Eugene Zwane
Department of Infectious Disease Epidemiology
Imperial College
London W2 1PG, UK
e.zwane@gmail.com

Peter van der Heijden
Department of Methodology and Statistics
Utrecht University
The Netherlands
p.vanderheijden@fss.uu.nl