An Overview of Population Size Estimation where Linking Registers Results in Incomplete Covariates, with an Application to Mode of Transport of Serious Road Casualties

Article *in* Journal of Official Statistics · September 2017

DOI: 10.1515/jos-2018-0011

CITATIONS
12

READS
230

4 authors, including:

Peter G.M. van der Heijden
Utrecht University
391 PUBLICATIONS **7,018** CITATIONS
SEE PROFILE

M.J.L.F Cruyff
Utrecht University
76 PUBLICATIONS **1,081** CITATIONS
SEE PROFILE

Bart F M Bakker
Centraal Bureau voor de Statistiek
54 PUBLICATIONS **711** CITATIONS
SEE PROFILE

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 239–263, http://dx.doi.org/10.1515/JOS-2018-0011

An Overview of Population Size Estimation where Linking Registers Results in Incomplete Covariates, with an Application to Mode of Transport of Serious Road Casualties

Peter G.M. van der Heijden[1], *Paul A. Smith*[2], *Maarten Cruyff*[3], *and Bart Bakker*[4]

We consider the linkage of two or more registers in the situation where the registers do not cover the whole target population, and relevant categorical auxiliary variables (unique to one of the registers; although different variables could be present on each register) are available in addition to the usual matching variable(s). The linked registers therefore do not contain full information on either the observations (often individuals) or the variables. By treating this as a missing data problem it is possible to construct a linked data set, adjusted to estimate the part of the population missed by both registers, and containing completed covariate information for all the registers. This is achieved using an Expectation-Maximization (EM)-algorithm. We elucidate the properties of this approach where the model is appropriate and in situations corresponding with real applications in official statistics, and also where the model conditions are violated. The approach is applied to data on road accidents in the Netherlands, where the cause of the accident is denoted by the police and by the hospital. Here the cause of the accident denoted by the police is considered as missing information for the statistical units only registered by the hospital, and the other way around. The method needs to be widely applied to give a better impression of the range of problems where it can be beneficial.

Key words: Dual system estimation; linkage; missing data; register; coverage.

1. Introduction

In recent years there has been continuing pressure on National Statistical Offices (NSOs) and other organisations producing official statistics to produce more, better quality and more detailed statistics, generally with decreasing resources. One of the important ways NSOs have responded has been to increase the use of administrative data sets, which provide relatively large amounts of information, generally at a small marginal cost. Often the desired range of statistical units or variables is not available on a single administrative data set, and therefore linking of administrative data sources (we call them registers) is also becoming more and more popular as a means to provide more comprehensive statistics.

There are several methodological problems that NSOs encounter when they are using registers for the production of official statistics. One is that registers, even when linked,

[1] Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands and University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: P.G.M.vanderHeijden@uu.nl
[2] University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: p.a.smith@soton.ac.uk
[3] Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands. Email: m.cruyff@uu.nl
[4] Statistics Netherlands, P.O.Box 24500, 2490 HA Den Haag, The Netherlands. Email: bfm.bakker@cbs.nl

under-cover the population of interest. A second problem is that missing values will be generated if two or more registers are linked. The values of variables that are only available in a subset of the registers are not known for the records that are not present in this subset. In this article, we present a framework for solving both undercoverage and these missing values in one procedure. We do not consider other methodological problems such as overcoverage and item missingness in single registers, although we return to them in the discussion. Figure 1 is a graphical representation of the linkage of two registers. The representation shows the linked data with observations (often individuals) in the rows and variables in the columns. The data for variables available only in register *A* are on the left and denoted by *a*, the data for variables only in register *B* are on the right and denoted by *b*, and in the middle are the data for variables that register *A* and *B* have in common, denoted by *ab*. Typically the variables in *ab* include the variables used for linking the registers.

As Figure 1 illustrates, each register has some unique variables. In *a* we find data for the covariates in register *A* that are *not* in register *B*. It follows that for the individuals in register *B* that are not in register *A* these covariates are missing. This is represented by the grey bitmap block at the bottom left in the representation in Figure 1. Similarly, individuals that are in register *A* but *not* in register *B* have missing values on the variables that are unique for register *B*, and this is represented by the grey bitmap block top right in Figure 1. In this article we consider the presence of the two grey bitmap blocks as a missing data problem that we solve by estimating the missing data. (It is evident that estimating missing covariate values only makes sense for covariates that pertain to all registers involved. An example where estimation of missing covariate values does not make sense: consider a population register coupled with a hospital register, then the

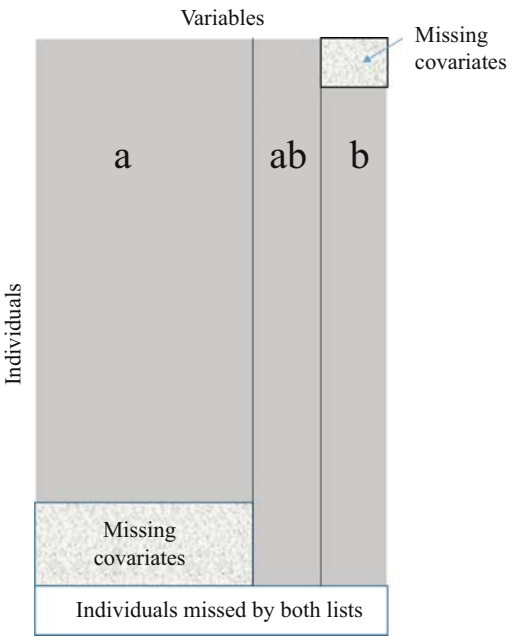

Fig. 1. Graphical representation of two linked registers, see text for details.

hospital register covariate "type of medical problem" should not be estimated for all individuals in the population register who do not appear in the hospital register, as it is likely that they do not have a medical problem at all.)

In addition, the linked registers may not cover the population perfectly, and two-source estimation may be applied to estimate that part of the population missed by both registers. This is depicted by the white area at the bottom of Figure 1. Notice that the aim is here not only to estimate the number of missing individuals but also their covariate data. Also notice that, in common with the basic population size estimation problem using two registers, individuals can be observed in three ways: individuals only in register *A* (on top), individuals in both register *A* and *B* (in the middle), and individuals only in register *B* (at the bottom). We note that when there are two registers, two-source estimation assumes independence conditional on covariates. Heterogeneity of inclusion probabilities can lead to marginal dependence between the registers. When this heterogeneity is caused by observed covariates, using these covariates in the model can compensate for this dependence. However, there may be dependence that is not caused by observed covariates and one way to handle this true dependence is by inclusion of a third register (compare the International Working Group for Disease Monitoring and Forcasting 1995) or by using latent variable models (compare Darroch et al. 1993; Fienberg et al. 1999). We will show that the approach we adopt can also be elaborated for more than two registers.

Thus there is a missing data problem in the covariates and a population size estimation problem, and both problems are handled simultaneously. In earlier work, Zwane and Van der Heijden (2007) and Van der Heijden et al. (2012) studied the situation where the missing variables are categorical and Zwane and Van der Heijden (2008) where they are continuous. In this article we review the case of categorical variables only, where the problem will be solved by applying the Expectation-Maximization (EM)-algorithm to estimate the missing observations in the context of population size estimation. We will in particular investigate the properties of the chosen solution as well as applying it within simulation studies.

Secondly, we will discuss an application where a single concept is measured by one variable in *A* and another in *B*, but where the validity of the variable in *A* is considered to be better than the validity of the variable in *B*. Notice that the concept is not represented by a single variable in part *ab* in Figure 1, but by a variable in part *A* and a different variable in part *B* and in this case the variable is clearly relevant in both (all) registers. Now the focus is on the estimation of the missing data of the grey bitmap part at the bottom left of the representation in Figure 1.

A good overview of two-source and multiple-source estimation where registers are linked is Bishop et al. (1975, Ch. 6). Important work in official statistics includes Wolter (1986), Bell (1993) and Griffin (2014) for the US Census, and by Brown et al. (1999), Brown et al. (2006) and Brown et al. (2011) for the UK. In epidemiology important reviews are by the International Working group for Disease Monitoring and Forecasting (1995) and Chao et al. (2001). For a Bayesian perspective, see Madigan and York (1997).

In this article the covariates in a population size estimation model play an important role. Earlier work in this area is from Bishop et al. (1975, Ch. 6), Alho (1990), Huggins (1989), Baker (1990), Tilling and Sterne (1999), and Zwane and Van der Heijden (2005); for a review see Pollock (2002). Bishop et al. (1975) discuss the use of categorical

covariates, and Alho (1990) and Zwane and Van der Heijden (2005) discuss how inclusion probabilities may be functions of continuous auxiliary information, where Alho (1990) predicts the inclusion probabilities using logistic regressions for two registers and Zwane and Van der Heijden (2005) generalize this to more than two registers. These papers do not discuss the problem of partly missing covariates.

The problem of partly missing covariates which we discuss here arises because the registers available describe different parts of populations, for example, the registers cover different but overlapping regions in a country, or cover different but overlapping periods in time. Zwane et al. (2004) and Sutherland et al. (2007) also approach this problem as a missing data problem, where, for the region example, the regional parts of a register that are missed by design are estimated using the EM-algorithm. Here the dependence structure between the registers in those regions that are observed by more than one register is projected onto those regions where one or more registers are missing. Zwane et al. (2004) and Sutherland et al. (2007) illustrate this for an example of six registers on spina bifida that are operative in different but overlapping time periods, where they fit log-linear models in the M-Step, and Pelle et al. (2016) fit multidimensional Rasch models to these data.

In the following we will first present the theory and properties of our approach, including the extension to more than two registers. This is followed by simulation studies showing the circumstances under which our approach is better than ignoring the additional variables. We end with an application to the estimation of the number of serious casualties from traffic accidents in the Netherlands measured by the police and by hospitals.

2. Population Size Estimation in the Presence of Missing Covariates: Theory

The basic idea of the methodology that we review can easily be explained by an example taken from Van der Heijden et al. (2012) and Gerritse et al. (2015b), involving the estimation of the population size of people with Afghan, Iranian, or Iraqi nationality (hereafter "AII") in the Netherlands, see Panel 1 in Table 1. Register A is a population register in the Netherlands and register B is a police register. From the population register A the variable Marital status is used, and denoted by X_1, with $X_1 = 1$ referring to married or living together and $X_1 = 0$ referring to unmarried, divorced, or widowed. From the police register B the variable "Police region where apprehended" is used, and denoted by X_2, with $X_2 = 1$ referring to one of the five biggest cities of the Netherlands and $X_2 = 0$ referring to the rest of the country. Notice that Marital status is not available in register B and "Police region where apprehended" is not available in register A. Clearly Marital status is a relevant variable for people in the police register; "Police region where apprehended" is not so obviously relevant for the population register, since most people will not have been apprehended. However, we can consider it as an approximation to usual residence, and therefore it is a relevant variable (though imperfectly measured in this source). If we compare the variables in Table 1 to Figure 1, we see that X_1 in Table 1 is a variable in region A in Figure 1, and X_2 in Table 1 is a variable in region B in Figure 1. In Table 1 A and B are variables denoting presence in registers A and B respectively, with categories 0 = no and 1 = yes; the variables A and B in Table 1 are dichotomous variables in Figure 1 in the areas A and B.

Table 1. Covariate X_1 (Marital status) is only observed in population register A and X_2 (Police region where apprehended) is only observed in police register B.

Panel 1: Observed counts of AII individuals

		$B=1$		$B=0$
		$X_2=0$	$X_2=1$	X_2 missing
$A=1$	$X_1=0$	259	539	13,898
	$X_1=1$	110	177	12,356
$A=0$	X_1 missing	91	164	–

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

		$B=1$		$B=0$	
		$X_2=0$	$X_2=1$	$X_2=0$	$X_2=1$
$A=1$	$X_1=0$	259.0	539.0	4,510.8	9,387.2
	$X_1=1$	110.0	177.0	4,735.8	7,620.3
$A=0$	$X_1=0$	63.9	123.5	1,112.4	2,150.2
	$X_1=1$	27.1	40.5	1,167.9	1,745.4

The eight counts in Panel 1 of Table 1 correspond to Figure 1 as follows: four counts for AII individuals that are in both register A and register B, which cross–classify individuals using the variables X_1 and X_2; two counts for AII individuals that are only in register A, which categorize them using variable X_1 only, as variable X_2 is missing for the individuals in register A; and two counts for AII individuals that are only in register B, which categorize them using variable X_2 only, as variable X_1 is missing for the individuals in register B.

In Panel 2 of Table 1 the counts 13,898 and 12,356, and the counts 91 and 164, are distributed over the levels of the missing variables. For example, 13,898 is distributed over the levels of X_2 into 4,510.8 and 9,387.2, and the ratio of these two counts is equal to the ratio of the observed counts 259 and 539. Similarly, 91 is split up into 63.9 and 27.1, and the ratio of these two counts is equal to the ratio of the observed counts 259 and 110. As a result, in Panel 2 the odds ratio for the counts 259, 539, 110, and 177 is projected to the four cells on the right and the four cells at the bottom. The theoretical motivation of this projection is given by a Missing At Random (MAR) assumption, and the estimates are found using the EM-algorithm (Zwane and Van der Heijden 2007). The EM-algorithm is an iterative procedure where each iteration has an expectation (E) and a maximization (M) step. In the E-step the expectations of the missing values are found given the observed values and the fitted values under a model, here some log-linear model. The E-step yields completed data. Then, in the M-step, the log-linear model is fitted to the completed data and this updates the fitted values that are used in the next E-step. This proceeds until convergence. The algorithm has linear convergence, which may make the algorithm very slow. Yet the likelihood increases in each step and therefore convergence is guaranteed. We illustrate the EM-algorithm for the maximal model in the next section, but first we elaborate some theoretical properties of this approach.

In the lower right corner of Panel 2 of Table 1 the missing part of the population is estimated (compare the white area in Figure 1). This estimate is a by–product of the

estimation using the EM-algorithm. For example, the missed count for $X_1 = 0$ and $X_2 = 0$ is 1,112.4, and this value is found by assuming independence between A and B given X_1 and X_2, so that $4,510.8 \times 63.9/259 = 1,112.4$. This last step is made under the usual assumptions in population size estimation using two registers taking into account the covariates, that is, (i) perfect linkage, (ii) independence between A and B conditional on X_1 and X_2, (iii) for each of the four subpopulations the population is closed, and (iv) homogeneity of inclusion probabilities for A or B, conditional on X_1 and X_2. The use of the word "or" in assumption (iv) may come as a surprise as in many papers homogeneity is formulated as an assumption that should hold for both A and B. However, if it holds for only one of the registers, this is sufficient, see Chao et al. (2001) and Van der Heijden et al. (2012).

2.1. Maximal Models

The most complicated model that can be fitted is

$$log\,\pi_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2}, \tag{1}$$

with identifying restrictions that the parameters λ, λ_1^A, λ_1^B, $\lambda_1^{X_1}$, $\lambda_1^{X_2}$, $\lambda_{11}^{AX_2}$, $\lambda_{11}^{BX_1}$, $\lambda_{11}^{X_1X_2}$ and $\lambda_{11}^{AX_2}$ are free, and the other parameters are restricted to be zero. The two-factor interactions are closely related to odds ratios; for example, $exp\left(\lambda_{11}^{X_1X_2}\right)$ is the conditional odds ratio between X_1 and X_2. Another way to denote log-linear models is to use the highest fitted interactions to codify the model, the highest interactions implying the inclusion in the model of all lower order effects; for this model this corresponds to the notation $[AX_2][X_1X_2][BX_1]$. Model $[AX_2][X_1X_2][BX_1]$ has eight free parameters, namely an intercept, four main effects, and three interactions. This number of parameters corresponds to the number of counts in Panel 1 of Table 1, that is also eight. Notice that the term AX_1 is not included in the model, because when $A = 0$, X_1 is missing and unknown. Therefore only three counts are available for the term AX_1. On the other hand, for the term AX_2 four counts are available, namely (259 + 110), (539 + 177), 91 and 164, so this term is included in the model (and similarly for BX_1). A maximal model is also a saturated model in the sense that the fitted counts for a maximal model are equal to the observed counts. (Notice that violations of this model, such as dependence between A and B conditional on the covariates, cannot be tested. We come back to this issue in Section 3, where we investigate sensitivity to such model violations.)

The fitted values for this model are obtained with the EM-algorithm. The algorithm starts with the initial estimates $\hat{n}_{10(lk)}^{(0)}$ and $\hat{n}_{01(lk)}^{(0)}$, that are found by evenly distributing the observed frequencies $n_{10(l+)}$ and $n_{01(+k)}$ over the corresponding cells. In the first M-step, the log-linear model $[AX_2][X_1X_2][BX_1]$ is fitted to the completed data, with the cells corresponding to $(i,j) = (0,0)$ specified as structural zeros. This yields the estimates $\hat{\pi}_{ij(lk)}^{(1)}$, which are then used in the first E-step

$$\hat{n}_{10(lk)}^{(1)} = \frac{\hat{\pi}_{10(lk)}^{(1)}}{\hat{\pi}_{10(l+)}^{(1)}} n_{10(l+)}, \quad \hat{n}_{01(lk)}^{(1)} = \frac{\hat{\pi}_{01(lk)}^{(1)}}{\hat{\pi}_{01(+k)}^{(1)}} n_{01(+k)}, \tag{2}$$

to compute the updates $\hat{n}_{10(lk)}^{(1)}$ and $\hat{n}_{01(lk)}^{(1)}$. These estimates are then used in the second M-step to find the updates $\hat{\pi}_{ij(lk)}^{(2)}$, and so on until convergence is reached at iteration t.

Equation (1) also allows for an alternative way to estimate the four cells in the lower right part of Table 1. For example, the upper left element $1,112.4 = \exp(\hat{\lambda})$, as for the cell with indices $(i,j,k,l) = (0,0,0,0)$ the parameter values are zero except for the intercept. Similarly, $2,150.2 = \exp(\hat{\lambda} + \hat{\lambda}_1^{X_1})$. In other words, the parameters of the model are estimated and projected to cells that refer to the part of the population missed by both registers.

Model 1 can easily be extended when there are additional covariates. For example, consider the situation that in addition to X_1 being observed in A and X_2 in B a variable X_3 is observed in A and B; then the maximal and saturated model is $[AX_2X_3][X_1X_2X_3][BX_1X_3]$. And, as a second example, consider the situation that in addition to X_1 being observed in A and X_2 in B a variable X_4 is only observed in A, then the maximal model is $[AX_2][X_1X_2X_4][BX_1X_4]$.

For each of the three models discussed it is possible to investigate whether more restrictive models also fit the data. For example, for the model $[AX_2]$ $[X_1X_2][BX_1]$ it is useful to investigate whether one of the interactions can be eliminated without the fit deteriorating. For example, if the covariate X_1 is statistically independent from the covariate X_2, then the model becomes $[AX_2][BX_1]$, and under this model A and B are statistically independent, and not independent conditional on X_1 and X_2.

Example. For model 1 the likelihood ratio chi-square is zero with zero degrees of freedom. We may want to investigate whether imposing the additional restriction $\lambda_{kl}^{X_1X_2} = 0$ is allowed, so that the model becomes $[AX_2][BX_1]$. The difference between the likelihood ratio chi-squares for these two models is 3.2 (df is 1), which is not significant at the five percent level. The estimated population size under model 1 is 33,769.9, whereas for model 1 with $\lambda_{kl}^{X_1X_2} = 0$ it is 33,764.2, only marginally different. This corresponds to the odds ratio estimated from the four elements where X_1 and X_2 are both observed, 259, 539, 110, and 177, which yields a value of 0.7732 with a 95 percent confidence interval of (0.5842, 1.0234). The z-statistic to test whether the odds-ratio is significantly different from 1 is 1.798, which is not significant in a two-sided test but is significant in a one-sided test.

2.2. *Collapsibility, Active and Passive Variables*

The maximal models just discussed have interesting properties in terms of collapsibility over variables X_1 and X_2 (Van der Heijden et al. 2012). We use the following terminology. We use the word *marginalize* to refer to the contingency table formed by considering a subset of the original variables. We use the word *collapsibility* to refer to the situation that when a table is marginalized the population size estimate remains invariant. Using these terms the properties of maximal models can be easily explained using interaction graphs of the log-linear models involved. See Figure 2. Log-linear model $[AX_2X_3][X_1X_2X_3][BX_1X_3]$ has graph M_1. This is a maximal model. The log-linear model where X_1 and X_2 are conditionally independent given the variables A, B, and X_3 is $[AX_2X_3][BX_1X_3]$ and this model has graph M_2. What follows are three models where one of X_1, X_2, or X_3 is not available. In model M_3 the variable X_1 is not available, and the log-linear model is $[AX_2X_3][BX_3]$. In model M_4 the variable X_2 is not available, and the log-linear model is $[AX_3][BX_1X_3]$. Finally, in model M_5 the variable X_3 is not available, and the log-linear

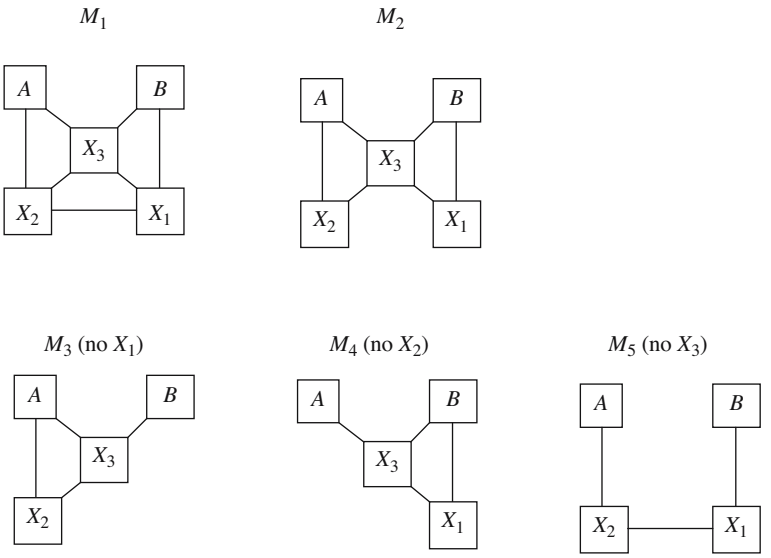

Fig. 2. Graph representations of some log-linear models.

model is $[AX_2][X_1X_2][BX_1]$. This last model is the model that is the focal point in the exposition in this article.

Interaction graphs are useful tools for assessing collapsibility. We make use of the concept of a *short path* (Whittaker 1990): two registers A and B are connected by a path if there is a sequence of adjacent edges connecting the variables A and B in the graph. A short path from A to B is a path that does not contain a sub-path from A to B. The rule is that when a covariate is on a so-called short path from A to B, the contingency table cannot be marginalized over this variable, and vice versa, that is when a covariate is not on a short path, the contingency table can be marginalized over this variable. We now discuss this for models M_1 to M_5. In M_1 the data cannot be marginalized over any of the variables X_1 to X_3. The reason is that there are two short paths, namely $A - X_2 - X_1 - B$ and $A - X_3 - B$, and each of the variables X_1 to X_3 is on one of the two short paths. In M_2 the data are collapsible over X_1 and over X_2, the reason being that the only short path is $A - X_3 - B$. In M_3 the data are collapsible over X_2, the reason being that the only short path is $A - X_3 - B$. In M_4 the data are collapsible over X_1, the reason being that the only short path is $A - X_3 - B$. In M_5 the data cannot be marginalized over X_1 and X_2, because both variables are on the short path $A - X_2 - X_1 - B$.

When a model (or graph) is collapsible over a variable, this means that in both the original model and collapsed model the same estimate of the population size is obtained. For example, models M_2, M_3, and M_4 yield the same population size estimate, and this estimate is identical to the population size estimate of model $[AX_3][BX_3]$. However, it may still be interesting to fit a model M_3, for example, because then this total population size estimate is spread out over the levels of variables X_1 and X_2. In Van der Heijden et al. (2012) the variables X_1 and X_2 in model M_2 are referred to as passive, in the sense that they do not have an impact on the estimate of the total population size. In contrast, variables X_1

and X_2 in model M_1 are referred to as active, because these variables do influence the total population size estimate.

Example. In the former section we saw that in model $[AX_2][X_1X_2][BX_1]$ the additional restriction $\lambda_{kl}^{X_1X_2} = 0$ does not deteriorate the fit, so that a more parsimonious model is $[AX_2][BX_1]$. As there is no short path any more between A and B, this means that we can marginalize over X_1 and X_2, showing that the population size estimate for model $[AX_2] \times [BX_1]$ is identical to the population size estimate for model $[A][B]$. We do not state that the original table should necessarily be marginalized over X_1 and X_2, because the original table can give insight into how the total population size is spread out over the levels of X_1 and X_2. Van der Heijden et al. (2009) and Van der Heijden et al. (2012) consider other examples with a larger number of covariates, namely five. They show that, by estimating the missing covariates and the number of individuals missed completely, the coverage of the population register can be evaluated in terms of the five covariates.

2.3. Precision and Sensitivity

Figure 1 illustrated that there are two estimation problems: estimating the missing covariates (the grey bitmap parts) and estimating the number of individuals (and their covariate values) missed by both A and B (the white parts in Figure 1). For both estimation problems we are interested in the precision when the model assumptions are true, and the sensitivity of the outcomes to deviations from the model assumptions.

We first discuss precision and start with the precision of the estimates for the missing covariates. Here precision is to be understood as an overall term referring to the variance of the estimates. Under the EM approach the model fitted is $[AX_2][X_1X_2][BX_1]$. As can be seen in Table 1, the odds ratio $(259 \times 110)/(539 \times 177) = 0.7732$, is used to calculate the expectations for the part of the table where X_1 is missing and the part where X_2 is missing. Under the model, the more precise this odds ratio, the more precise these expectations. This precision is directly related to the size of the population that is in both A and B: the larger this size, the smaller the standard error of the odds ratio and the standard errors of the estimates and the larger the precision.

The precision of the data for the individuals missed by A and B is the outcome of two sources: first, the precision of the estimates of the missing covariates that we just discussed, and, second, implied coverage. Precision of the estimates of the missing covariates has a direct impact on the precision of the data for the individuals missed. Consider again Table 1. Because, in Panel 2 of Table 1, the estimate $1,112.4 = 4,510.8 \times 63.9/259$, when the estimates 63.9 and 4,510.8 are imprecise, the estimate 1,112.4 will be imprecise as well.

The second source of imprecision is related to implied coverage. We explain this for $(X_1, X_2) = (0, 0)$. For the population register A the coverage of A implied by B is $259/(259 + 63.9) = 0.802$. However, for the police register B the coverage of B implied by A is only $259/(259 + 4,510.8) = 0.057$. The equation $1,112.4 = 4,510.8 \times 63.9/259$ shows that if either or both of these implied coverages is low, the estimated number of missed individuals is large relative to the number of individuals seen, and hence imprecise.

Estimates of the precision can be obtained using the parametric bootstrap (compare Buckland and Garthwire 1991). The parametric bootstrap provides a simple way to find the

confidence intervals when the contingency table is not fully observed. To compute the bootstrapped confidence intervals for a specific log-linear model, we need to first compute the population size under this model and the probabilities on the completed data under this model, that is, by including the cells that cannot be observed by design. A first multinomial sample is drawn given these parameters, and the sample is then reformatted to be identical to the observed data (for example, the sample in the format of Panel 2 of Table 1 is recoded into the format of Panel 1). The specific log-linear model used is then fitted to the resulting data, resulting in an estimate of the population size. Then this is repeated K times. By ordering the K bootstrap population size estimates, a percentile confidence interval can be constructed. We use this approach later on.

Up to this point we have discussed precision when the model assumptions are correct. We now discuss the sensitivity of the estimates to violations of the assumptions of the model. It is possible to investigate whether maximal models can be reduced by setting some parameters equal to zero. For example, in model M_5 (i.e., Equation 1) it is possible to test whether the parameter $\lambda_{kl}^{X_1X_2}$ is needed to give an adequate description of the data. However, it is not possible to test whether parameters that are *not* included in the maximal model, should be included. In other words, we cannot reject the MAR assumption using the data.

However, as was shown in this context by Gerritse et al. (2015b), it is possible to investigate for a particular data set how sensitive the outcome of the maximal model is to the assumption that certain parameters are zero. Take model M_5. The maximal model assumes that three two-factor interactions are zero, that is, $\lambda_{ik}^{AX_1} = \lambda_{jl}^{BX_2} = \lambda_{ij}^{AB} = 0$, and all three- and four-factor interactions are zero. Consider $\lambda_{ik}^{AX_1} = 0$. The maximal model is extended with a fixed parameter value for $\lambda_{ik}^{AX_1}$. We denote such a fixed parameter with the tilde $\tilde{\lambda}$, and the model to be fitted becomes

$$log\,\pi_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2} + \tilde{\lambda}_{ik}^{AX_1}. \tag{3}$$

Such a model can be fitted for a range of values of $\lambda_{11}^{AX_1}$. Appropriate values can be chosen by making use of the fact that log-linear parameters are closely related to odds ratios. Technically, the model may be fitted as a log-linear Poisson regression with offset $\exp\left(\tilde{\lambda}_{ik}^{AX_1}\right)$ (see Gerritse et al. 2015b, for details).

Gerritse (2016) argues that the sensitivity of outcomes of the analyses to violation of the independence assumption and to violation of perfect linkage is larger when the implied coverage is lower. In the absence of covariates this can be explained as follows. Let m_{ij} be the expected count for cell (i,j) $(i,j = 0,1)$, where m_{00} is the missing count to be estimated. Under independence the odds ratio is 1, that is, $m_{00}m_{11}/m_{01}m_{10} = 1$, so that $m_{00} = m_{01}m_{10}/m_{11}$. Under dependence with odds ratio θ, $m_{00}\theta = m_{01}m_{10}/m_{11}$. Thus, the smaller the overlap in cell (1,1), and hence the smaller the coverage, the larger the estimated value for cell (0,0), and this holds both for independence and dependence. In the same way, when links are missed, this increases the expected values m_{01} and m_{10} and decreases m_{11}, with the result that m_{00} is larger, and this effect is larger the smaller the overlap m_{11}.

Example. We carried out sensitivity analyses for the omission of parameters $\tilde{\lambda}_{ij}^{AB}$, $\tilde{\lambda}_{ik}^{AX_1}$, and $\tilde{\lambda}_{jl}^{BX_2}$. The results are shown in Table 2. Conditional odds ratios of 0.67 and 1.5 are used. For our example the model without the fixed parameters has an estimated missed

Table 2. Sensitivity analyses. The maximal model is $[AX_1][X_1X_2][BX_2]$*, where* $\hat{m}_{00} = 6{,}176$ *and* $\hat{N} = 33{,}770$*. Fixed conditional odds ratios are plugged in for* $\tilde{\lambda}_{ij}^{AB}$, $\tilde{\lambda}_{ik}^{AX_1}$ *and* $\tilde{\lambda}_{jl}^{BX_2}$.

Term	Size(OR)	$\hat{m}_{00}$	N	Size(OR)	$\hat{m}_{00}$	N
$\tilde{\lambda}_{ij}^{AB}$	0.67	4,117	31,711	1.5	9,264	36,858
$\tilde{\lambda}_{ik}^{AX_1}$		6,736	34,330		5,711	33,305
$\tilde{\lambda}_{jl}^{BX_2}$		6,136	33,730		6,220	33,814

population size of $\hat{m}_{00} = 6,176$ and an estimated population size of $\hat{N} = 33,770$. Violation of the model because there is direct dependence between A and B in the form of conditional odds ratios 0.67 or 1.5 has a large effect, because this leads to estimated missed population sizes of $\hat{m}_{00} = 4,117$ and $\hat{m}_{00} = 9,264$ respectively. Plugging in a missed odds ratio $\tilde{\lambda}_{jl}^{BX_2}$ has a minor effect on the estimated missed population size: for conditional odds ratios of 0.67 and 1.5 it leads to estimated values 6,136 and 6,220, both values being close to 6,176. However, plugging in a missed odds ratio $\tilde{\lambda}_{ik}^{AX_1}$ has a larger effect on the estimated missed population size: for conditional odds ratios of 0.67 and 1.5 it leads to estimated values 6,736 and 5,711.

2.4. *Extension to More than Two Registers*

The advantage of being able to use more than two registers is that the restrictive (conditional) independence assumption between variables A and B can be replaced by less restrictive assumptions. For example, in the situation of three registers without covariates, the saturated model is the model with all two-factor interactions. Now it is possible to search for more restrictive models that still describe the data well. One can consult the references provided in the introduction for details, see, for example, Bishop et al. (1975).

For three registers the problem of incomplete covariates has been studied by Zwane and Van der Heijden (2007), who show that for this problem the EM-algorithm can easily be adapted. Van der Heijden et al. (2012) discuss graph representations of the models and collapsibility, but do not touch incomplete covariates.

An interesting official statistics application is found in Gerritse et al. (2015a). The problem is to estimate the number of usual residents for the Dutch census 2011. Here usual residence is defined as, roughly, living in the Netherlands for a continuous period of twelve months before the reference time. Three registers are available, namely the population register, the employment register and a crime suspects register. Given this definition we are interested in a dichotomized version of duration, namely longer than or shorter than a year. From both the population register and the employment register residence duration can be derived (for details see Gerritse et al. 2015a), so when people are only in the population register or only in the job register a measurement for a persons' duration is available. For persons who are both in the population register and the employment register the overlapping durations are reconciled and dichotomized. The crime suspects register has no variable for duration. This is not problematic for persons who are also in the population register or the employment register, because then the residence variable of the latter can be used, but it is problematic for individuals who are only in the crime suspects

Table 3. Polish individuals by the population register, the employment register and the crime suspects register, by usual residence. The counts for the two cells labeled "missing" add up to 1,043. Data from Gerritse et al. (2015a).

Usual residence	Population	Employment	Crime suspects	
			Yes	No
No	Yes	Yes	32	3,523
		No	34	3,225
	No	Yes	149	60,190
		No	missing	0
Yes	Yes	Yes	183	21,309
		No	195	14,052
	No	Yes	81	20,216
		No	missing	0

register. See Table 3, taken from Gerritse et al. (2015a), where counts for people born in Poland and registered in one or more of the three registers are displayed. We find a $2 \times 2 \times 2$ table for residence duration longer than a year and a $2 \times 2 \times 2$ table for residence duration shorter than a year, where two cells are indicated with the label 'missing'. As these two cells refer to persons only in the police register, the sum of the counts for the two cells is known, namely 1,043. The EM-algorithm is used to distribute these 1,043 persons over the two cells under some log-linear model, and the parameters of the final model are projected on the two (0,0,0) cells to find the number of persons missed by all three registers. We refer to the Supplementary materials, Section 1, for further details (Available online at: www.dx.doi.org/10.1515/jos-2018-0011).

A similar example can be found in Héraud-Bousquet et al. (2012), where there are three registers, and in two of the registers place of birth is available, but in a third register it is not. For those individuals only in the third register the missing values are imputed using multiple imputation. Multiple imputation has wider application when covariates are continuous instead of categorical, when the EM-algorithm loses its simplicity. Zwane and Van der Heijden (2008) apply multiple imputation using predictive mean matching in this situation.

3. Simulations

Earlier simulation results can also be found in Zwane and Van der Heijden (2007) for two registers and two partially observed covariates. These results are not completely transparent as the covariates used in the simulation are correlated continuous variables that are dichotomized. Thus the true model structure from which samples are drawn cannot easily be understood from the perspective of a log-linear model. In the simulations that we present here the true model is a log-linear model in which marginal probabilities and conditional odds ratios are specified to describe the dependence between the variables. We refer to the Supplementary materials, Section 2, for details on how true models are generated (available online at: www.dx.doi.org/10.1515/jos-2018-0011).

We carried out simulations to compare the behaviour of the classical model (denoted by LL), where incomplete covariates are ignored, with the model where incomplete

covariates are completed with the EM-algorithm (denoted by EM). For each choice of conditional odds ratios this yields population probabilities from which we sample. In each instance of the simulation study 25,000 samples are taken. For LL, for each sample the classical model $[A][B]$ is estimated on the marginal table formed from A and B, where the sampled count in cell $(A, B) = (0, 0)$ is made missing, and subsequently estimated assuming independence between A and B. Similarly, for EM for the same samples the model $[AX_2][X_1X_2][BX_1]$ is estimated, where the four cells where $(A, B) = (0, 0)$ are made missing.

In the first simulation study the population model is $[AX_2][X_1X_2][BX_1]$, so that the model estimated by EM is identical to the true model. The prespecified marginal probabilities are $P(A = 1) = 0.3$, $P(B = 1) = 0.3$, $P(X_1 = 0) = 0.5$ and $P(X_2 = 0) = 0.5$. Conditional odds ratios different from 1 are specified between A and X_2, between X_1 and X_2 and between B and X_1, so that the true model is $[AX_1][X_1X_2][BX_2]$. We denote the conditional odds ratio between A and X_2 by $OR(A, X_2)$. Note that the theoretical results in earlier sections show that, when one of the three conditional odds ratios $OR(A, X_2)$, $OR(B, X_1)$ or $OR(X_1, X_2)$ is 1, the model is collapsible over the covariates so that identical results are found for LL and EM. Therefore conditional odds ratios equal to 1 are not used. Also note that, for example, $OR(A, X_2) = OR(B, X_1) = 0.5$ leads to the same population probabilities as $OR(A, X_2) = OR(B, X_1) = 2$, as this is equivalent to the recoding of levels 0 and 1 in X_1 and X_2. Therefore we only use odds ratios of 2.

In Table 4 results are reported. In the upper part the true population size is 1,000. We first plug in conditional odds ratios of moderate size. In the first two lines the three odds ratios plugged in are $OR(A, X_2) = OR(B, X_2) = OR(X_1, X_2) = 2$. The average observed n, over 25,000 samples is 511, which is approximately $1,000 \times (1 - 0.7 \times 0.7)$, where 0.7 is the probability of not being selected in A or B. Note that the implied coverage, derived by collapsing over the covariates, is low, namely 0.3, that is, given population A, when linking to population B 70 percent of the observations in B were not seen before (in A). Under LL, the average estimated mean is 1,014.9 (with SE = 76.3 calculated over the 25,000 samples), the average estimated median is 1,009.9 (with SE = 76.4) and the RMSE is 77.7. Under EM, the average mean is 1,004.9 (SE = 75.4), the average median is 1,000.1 (SE = 75.6) and RMSE is 75.6. For $N = 1,000$ two other triples of conditional odds ratios are investigated. As expected, under EM the average mean and (in particular) the average median under the log-linear model are very close to the population value, where under LL there is some bias. Notice that the median has less bias than the mean, due to the non–normality of the distribution of estimates. With the population size of 1,000, the RMSE's of LL and EM are close. In the following four instances the population size is 10,000. The bias of the means and medians become a bit smaller, and as the standard errors become smaller (due to the increased population size) the RMSE's of EM become smaller than those of LL. The same holds for $N = 50,000$. It seems that the bias found for LL is approximately equally large but opposite for conditional odds ratios $OR(X_1, X_2) = 0.5$ and $OR(X_1, X_2) = 2$, and this is in contrast to the results in Zwane and Van der Heijden (2007).

In Table 5 the coverage is higher, with $P(A = 1) = P(B = 1) = 0.6$. When the coverage is higher, the part of the population missed is smaller, and violation of assumptions will have a smaller effect. This is also apparent by comparing Table 5 with Table 4, which

Table 4. Simulations under the model, with lower coverage. $P(A = 1) = 0.3$, $P(B = 1) = 0.3$, $P(X_1 = 0) = 0.5$ *and* $P(X_2 = 0) = 0.5$. *The conditional odds ratios refer to* $OR(A, X_2)$, $OR(B, X_1)$ *and* $OR(X_1, X_2)$.

	N	Odds ratios	Mean (n)	Mean	Median	SE mean	SE med.	RMSE
LL	1,000	2,2,0.5	511	1,014.9	1,009.9	76.3	76.4	77.7
EM				1,004.9	1,000.1	75.4	75.6	75.6
LL	1,000	2,2,2	509	995.2	990.8	73.8	73.9	74.0
EM				1,005.3	1,000.6	75.4	76.0	76.1
LL	1,000	2,2,5	508	984.1	979.1	72.4	72.6	74.1
EM				1,007.1	1,000.6	77.3	77.5	77.6
LL	10,000	2,2,0.5	5109	10,104.0	10,098.2	237.0	237.1	258.8
EM				10,005.1	10,000.3	234.2	234.3	234.3
LL	10,000	2,2,2	5092	9,910.7	9,906.2	228.7	228.8	245.6
EM				10,008.3	10,003.3	234.4	234.5	234.6
LL	10,000	2,2,5	5081	9,792.0	9,789.0	226.0	226.0	307.1
EM				10,007.2	10,003.3	239.5	239.5	239.6
LL	50,000	2,2,0.5	25,546	50,502.5	50,497.4	531.8	531.8	731.6
EM				50,007.4	50,004.5	524.2	524.2	524.3
LL	50,000	2,2,2	25,456	49,522.5	49,524.1	514.5	514.5	702.0
EM				50,008.2	50,010.5	527.2	527.3	527.3
LL	50,000	2,2,5	25,402	48,935.5	48,932.5	499.4	499.4	1,175.8
EM				50,004.7	49,998.3	529.0	529.1	529.1

Table 5. Simulations. $P(A = 1) = 0.6$, $P(B = 1) = 0.6$, $P(X_1 = 0) = 0.5$ *and* $P(X_2 = 0) = 0.5$. *The conditional odds ratios refer to* $OR(A, X_2)$, $OR(B, X_1)$ *and* $OR(X_1, X_2)$.

	N	Odds ratios	Mean (n)	Mean	Median	SE mean	SE med.	RMSE
LL	10,000	2,2,0.5	9906	10,001.7	10,001.9	10.8	10.8	10.9
EM				9,999.9	10,000.0	10.8	10.8	10.8
LL	10,000	2,2,2	9903	9,998.1	9,998.2	11.0	11.0	11.2
EM				9,999.9	10,000.0	11.0	11.0	11.0
LL	10,000	2,2,5	9901	9,995.9	9,996.0	11.0	11.0	11.8
EM				10,000.0	10,000.1	11.1	11.1	11.1

shows that the bias for LL in Table 5 is smaller than the bias in Table 4. The bias in EM is negligible, in particular when N increases.

Simulations suggested by the Census Coverage Survey for England and Wales are reported in the Supplementary materials, section 3 (available online at: www.dx.doi.org/10.1515/jos-2018-0011). We also did simulations where the model $[AX_2][X_1X_2][BX_1]$, assumed in the EM approach, is violated. These results can be found in the Supplementary materials, section 4 (available online at: www.dx.doi.org/10.1515/jos-2018-0011). Overall the simulations show that, when the MAR assumptions are fulfilled, the EM approach does better, though sometimes only slightly better, than the traditional approach. When the MAR assumptions are not fulfilled, the bias can be substantial, in particular when the inclusion probabilities are low.

4. Novel Application: The Same Variable Measured in Both Registers

We present a novel application of the above methodology. It concerns two registers that both measure the same variable, and the measure in one register is generally considered to be more trustworthy, or valid, than the measure of the same variable in the other register. This is closely related to the classical two-phase sampling problem, where there is an inexpensive but low quality measurement which can be obtained from a large sample, and a more expensive and more accurate approach which is used on a subsample. Two-phase sampling concentrates on combining the small sampling variance of the large sample measure with the measurement accuracy of the small sample measure. In our case we will apply the EM-algorithm to complete the missing information on the highest quality measure, and additionally to provide this information for statistical units which are missed in both the registers (a situation which cannot generally be handled by two-phase sampling). The example we deal with is the number of serious road injuries in the Netherlands. The first author was consulted by the Ministry of Transport with the question whether the current methodology applied for estimating this number was sufficiently appropriate. In the Netherlands the number of serious road injuries is important because it is used for assessing the road safety target.

In the Netherlands there are two parties that can deliver information on serious road injuries, namely the police and hospitals. Both parties are usually present after the occurrence of such an accident. The police are supposed to record the accident and its cause in the police crash record database, but this regularly does not happen for some

Table 6. Road accidents in the Netherlands in 2000, from Reurings and Stipdonk (2011). Motorized vehicle involved X_1 is only observed in the Police register (A) and Motorized vehicle involved X_2 is only observed in hospital register (B). Levels of X_1 and X_2 are 1 = yes, 2 = no.

Panel 1: Observed counts

		$B=1$		$B=0$	
		$X_2=1$	$X_2=2$	X_2 missing	Total
$A=1$	$X_1=1$	5,970	287	1,351	7,608
	$X_1=2$	28	256	70	354
$A=0$	X_1 missing	2,947	4,120	–	7,067
Total		8,945	4,663	1,421	15,029

Panel 2: Fitted values under $[AX_1][X_1Y]$

		$B=1$		$B=0$	
		$X_2=1$	$X_2=2$	X_2 missing	Total
$A=1$	$X_1=1$	5,970.0	287.0	1,351.0	7,608.0
	$X_1=2$	28.0	256.0	70.0	354,0
$A=0$	$X_1=1$	2,509.6	120.6	567.9	3,198.1
	$X_1=2$	437.4	3,999.4	1,093.6	5,530.4
Total		8,945.0	4,663.0	3,082.5	16,690.5

Panel 3: Fitted values under $[AX_2][X_1X_2][BX_1]$

		$B=1$		$B=0$		
		$X_2=1$	$X_2=2$	$X_2=1$	$X_2=2$	Total
$A=1$	$X_1=1$	5,970.0	287.0	1,289.0	62.0	7,608.0
	$X_1=2$	28.0	256.0	6.9	63.1	354,0
$A=0$	$X_1=1$	2,933.2	2,177.6	633.3	470.2	6.214,3
	$X_1=2$	13.8	1,942.4	3.4	478.8	2.438,4
Total		8,945.0	4,663.0	1.932,6	1,074.1	16,614.7

reason, such as that it is not clear which police officer has to file the accident report, or that the injury is not considered very serious. The hospital that treats the seriously injured, can report the cause of the injury in the hospital inpatient registry but this is sometimes forgotten and then such a patient's connection to a traffic accident is lost. Thus there are two register sources that both have coverage problems. Many details of the registration by the police and the hospitals can be found in Reurings and Stipdonk (2011), who report research conducted at the SWOV Institute for Road Safety Research. They state that the police database in particular suffers from serious underreporting, and is inaccurate in indicating injury severity, whereas the hospital database is inaccurate in indicating that a patient was involved in a road crash but in principle contains all serious road injuries.

For the year 2000 Reurings and Stipdonk (2011) present the data in the upper panel of Table 6. (We refer to their paper for a detailed discussion regarding the linking of the two

registers.) The police register has a larger undercoverage than the hospital register. Yet it is reasonable to assume that, where the police registers do record the mode of transport of injured persons, they do this more accurately than the hospital. The reason is that assessing the cause of accidents is a more important function for the police, because liability plays a role, than of the hospital, which is more concerned about the type of serious casualty and who will be focused more on health related issues than on the cause and details of the accident. Notice that in the 2×2 subtable that is fully observed, there are 287 joint classifications not in agreement where the police recorded the involvement of a motorized vehicle but the hospital recorded that no motorized vehicle was involved, and 29 vice versa.

As it turns out, two approaches can be taken for solving the missing data problem and subsequently estimating the number of accidents missed by both registers for the $2 \times 2 \times 2 \times 2$ table. We discuss these options and then generalize to a situation where the number of levels of the variables X_1 and X_2, Cause of the accident, is increased from two to seven.

4.1. The $2 \times 2 \times 2 \times 2$ Table

As a first approach, Reurings and Stipdonk (2011) set up a system of linear equations to estimate the number of seriously injured. They report 10,804 seriously injured in motorized accidents and 5,891 seriously injured in non-motorized accidents. Using a log-linear modelling framework that includes missing data we can obtain their results as follows. We define a new variable Y with three levels, namely $(X_2 = 1, B = 1), (X_2 = 2, B = 1)$ and $(X_2 =$ missing$)$. We then fit model $[AX_1][X_1Y]$ with X_1–values missing for $A = 0$. The estimates using our procedure should in principle be identical to Reurings and Stipdonk's estimates but they are slightly different (probably due to rounding), see Panel 2 of Table 6, in the two last lines, and these lead to estimates of (7,608 + 3,198.1 =) 10,806.1 for motorized and 5,884.4 for non-motorized accidents. In this approach the relative frequencies for 5,970, 287, and 1,351 are identical to those for 2,509.6, 120.6, and 567.9, and similarly for 28.0, 256.0, and 70.0 to 437.4, 3.999.4, and 1,093.6, while at the same time the counts 2,947 and 4,120 are split up over the missing levels of X_1. Notice that we estimate that only (567.9 + 1,093.6 =) 1,661.5 accidents with serious road injuries are missed by both registers, which is approximately ten percent of the total estimated population size. 95 percent confidence intervals of the estimates 10,806.1 and 5,884.4 are obtained using the parametric bootstrap by the percentile method with 10,000 bootstrap samples, and this yields 10,532 – 11,054 and 5,512 – 6,305.

As the second approach, we apply the methodology to this table that we applied before in Table 1. That is, we assume that the hospital Cause of accident is missing for those accidents only registered by the police whereas we assume that the police Cause of accident is missing for those accidents only registered by the hospital, and fit model $[AX_2][X_1X_2][BX_1]$. See Panel 3 in Table 6. This leads to very different estimates for motorized and non-motorized accidents, namely 13,823 (95 percent CI 13,568 – 14,072) and 2,791 (2,551 – 3,037). In this approach the four odds ratios for all combinations of register A and B are assumed to be equal, and the counts 2,947 and 4,120 are now split up in a way different from the first approach.

We make a few remarks. First, when we compare both approaches we have a preference for our own approach using model $[AX_2][X_1X_2][BX_1]$ over the approach by Reurings and

Stipdonk using model $[AX_1][X_1Y]$. This preference is not based on model fit as both models are saturated and have a perfect fit. Instead we make a judgement based on a professional opinion. We find it is reasonable to assume that, for example, the count 2,947 for which X_1 is missing, should be split over motorized and non-motorized in the same way as when X_1 is not missing. Our approach is plausible, simple and transparent, as in the saturated model we present here the estimates can be found by hand. The plausibility of the approach by Reurings and Stipdonk can be argued, but it is less simple and transparent, as it needs an iterative procedure, and in the next section we will see that it can have numerical problems. We obtain additional support from the model-based bootstrap applied to $[AX_2][X_1X_2][BX_1]$ which gives smaller confidence intervals (14,072 – 13,568 = 504 and 486) than the estimates under model $[AX_1][X_1Y]$ (11,054 – 10,532 = 522 and 793). This strategy is in line with Elliott and Little (2000)'s principles for choosing between saturated models where, after a series of principles fail to distinguish models, then principle 5 suggests using the model that gives estimates with reduced variance.

Second, when we compare our log-linear modelling procedure with the approach by Reurings and Stipdonk of solving a system of linear equations, a number of differences are apparent. Our approach is flexible because extra variables can be incorporated easily. This will in principle also be the case in Reurings and Stipdonk's approach. However, when estimates become unstable due to low observed counts, our approach allows for constraints on the log-linear parameters that can stabilize the model. The modelling approach has the advantage that it always produces maximum likelihood estimates, whereas solving a system of linear equations only leads to maximum likelihood estimates when the estimates are non-negative. Also, we think that the flexibility of our approach is important, because Reurings and Stipdonk (2011) report that they applied the method three times separately, namely for the covariates transport mode (reported here), region and injury severity. This has the drawback that three different estimates of the population size will result. In our methodology it is easy to include all three covariates simultaneously, and this will yield a single total population size that is consistent over the three covariates. It also allows investigation of the relationships between the three covariates.

As a third remark, in situations like this a practical approach is often taken (Reurings and Stipdonk 2011 are a noteworthy exception) when a measure of some variable in register A is considered more trustworthy than a different measure of the same variable in register B, so after linking registers A and B a new, composite variable is created that makes the best of the information. In this new measure we fill in the values of the variable from register A when it is available, we fill in the values of the variable from register B for the observations that were missed by register A, and some ad hoc solution is found for the observations that were missed by both registers. In the approaches presented here, however, for those observations that were missed by register A we translate the values in register B into what would have been found in register A using the subtable of $A = 1$ and $B = 1$ to give the structure for those observations only found in register B, 2,947 and 4,120 at the bottom of Panel 1 of Table 1.

Last, notice that the odds ratio in this observed subtable is typically very large (in the upper part of 6 it is almost 200), and in both approaches the odds ratio for the subtable of $A = 1$ and $B = 1$ is used to find the estimates in the subtables of $A = 0$ and $B = 1$.

4.2. The $2 \times 2 \times 7 \times 7$ Table

The reason that the Ministry asked Van der Heijden for a consultation had to do with a generalization of the method applied by the SWOV Institute for Road Safety Research. See Table 7 taken from Reurings and Bos (2012, 25) where we find for 2010 a much more detailed coding of motorized mode of transport: where in Table 6 this only had one coding, it now has six codings, namely "Sitting in car", "Driving motorbike", "Driving moped", "Bicycles in motorized accident", "Pedestrians in motorized accident", and "Other in motorized accident". Of course, this finer coding into seven levels can be useful for assessing the cause of a rise or decline in accidents. Notice the occasional low off-diagonal counts, that are attractive because they make the data plausible (we do not want "non-motorized" to be mixed up a lot with "sitting in car". A second difference between the data for the years 2000 and 2010 is that the police registered many fewer accidents: in 2000 the number in the police register was around 7,000 compared with 8,000 missed by the police but found in the hospital registration, but in 2010 these numbers are approximately 3,500 and 14,000. In the same period, the quality of the hospital register went up: in 2000 1,400 accidents were observed by the police but not by the hospital, but in 2010 this was only approximately 400.

The SWOV Institute for Road Safety Research generalized their approach of using a system of linear equations and found unstable estimates for some cells, including estimated counts that were negative. Using log-linear model $[AX_1][X_1Y]$, where Y has eight categories, the EM-algorithm also produces unstable results in the sense that convergence is not reached with 10^6 iterations, where in that last iteration two lines of estimates where $A = 0$ consisted of 0's only. Therefore we will only present results for the approach using model $[AX_2][X_1X_2][BX_1]$.

Table 7. Road accidents in the Netherlands in 2010. Data from Reurings and Bos (2012, 25). Motorized vehicle involved X_1 is only observed in Police register (A) and Motorized vehicle involved X_2 is only observed in hospital register (B). m.a. = motorized accident.

Observed counts

			$B = 1$				$B = 0$		
			X_2				X_2		
X_1	1	2	3	4	5	6	7	missing	Total
$A = 1$									
1. Sitting in car	856	7	12	26	61	62	18	130	1,172
2. Driving motorbike	3	261	33	0	7	5	2	20	331
3. Driving moped	7	83	504	19	8	60	21	47	749
4. Bicycles in m.a.	55	2	10	523	38	29	139	96	892
5. Pedestrians in m.a.	9	0	2	11	208	33	3	35	301
6. Other in m.a.	20	1	18	4	7	17	2	22	91
7. Non-motorized	2	0	0	9	1	7	82	12	113
$A = 0$									
missing	1,100	860	1,530	844	482	540	8,578	–	13,934
Total	2,052	1,214	2,109	1,436	812	753	8,845	362	17,583

Table 8. Motorized vehicle involved X_1 is only observed in Police register (A) and Motorized vehicle involved X_2 is only observed in hospital register (B). Fitted values (rounded) under model $[AX_2][X_1X_2][BX_1]$.

	$B=1$								$B=0$							
	X_2								X_2							
X_1	1	2	3	4	5	6	7	Total	1	2	3	4	5	6	7	Total
$A=1$																
1. Sitting in car	856.0	7.0	12.0	26.0	61.0	62.0	18.0	1,042.0	106.8	0.9	1.5	3.2	7.6	7.7	2.2	129.9
2. Driving motorbike	3.0	261.0	33.0	0.0	7.0	5.0	2.0	311.0	0.2	16.8	2.1	0.0	0.5	0.3	0.1	20.0
3. Driving moped	7.0	83.0	504.0	19.0	8.0	60.0	21.0	702.0	0.5	5.6	33.7	1.3	0.5	4.0	1.4	47.0
4. Bicycles in m.a.	55.0	2.0	10.0	523.0	38.0	29.0	139.0	796.0	6.6	0.2	1.2	63.1	4.6	3.5	16.8	96.0
5. Pedestrians in m.a.	9.0	0.0	2.0	11.0	208.0	33.0	3.0	266.0	1.2	0.0	0.3	1.4	27.4	4.3	0.4	35.0
6. Other in m.a.	20.0	1.0	18.0	4.0	7.0	17.0	2.0	69.0	6.4	0.3	5.7	1.3	2.2	5.4	0.6	21.9
7. Non-motorized	2.0	0.0	0.0	9.0	1.0	7.0	82.0	101.0	0.2	0.0	0.0	1.1	0.1	0.8	9.7	11.9
$A=0$																
1. Sitting in car	989.1	17.0	31.7	37.1	89.1	157.2	578.3	1,899.5	123.4	2.1	4.0	4.6	11.1	19.6	72.1	236.9
2. Driving motorbike	3.5	634.1	87.2	0.0	10.2	12.7	64.3	812.0	0.2	40.8	5.6	0.0	0.7	0.8	4.1	52.2
3. Driving moped	8.1	201.6	1,331.8	27.1	11.7	152.1	674.7	2,407.1	0.5	13.5	89.2	1.8	0.8	10.2	45.2	161.2
4. Bicycles in m.a.	63.6	4.9	26.4	745.6	55.5	73.5	4,465.7	5,435.2	7.7	0.6	3.2	89.9	6.7	8.9	538.6	655.6
5. Pedestrians in m.a.	10.4	0.0	5.3	15.7	303.8	83.7	96.4	515.3	1.4	0.0	0.7	2.1	40.0	11.0	12.7	67.9
6. Other in m.a.	23.1	2.4	47.6	5.7	10.2	43.1	64.3	196.4	7.4	0.8	15.2	1.8	3.3	13.7	20.5	62.7
7. Non-motorized	2.3	0.0	0.0	12.8	1.5	17.7	2,634.4	2,668.7	0.3	0.0	0.0	1.5	0.2	2.1	313.0	317.1

Table 9. Parametric bootstrap point estimates of causes according to the police, with 95 percent confidence interval (percentile method) and median, under model $[AX_2][X_1X_2][BX_1]$.

	Mean	2.5 percent	Median	97.5 percent
1. Sitting in car	3,307.1	2,997.9	3,300.6	3,644.9
2. Driving motorbike	1,195.0	1,071.2	1,190.8	1,336.8
3. Driving moped	3,317.2	2,996.6	3,312.6	3,657.6
4. Bicycles in m.a.	6,981.8	6,382.4	6,980.5	7,583.1
5. Pedestrians in m.a.	884.7	749.4	881.0	1,046.1
6. Other in m.a.	350.8	238.4	344.3	498.9
7. Non-motorized	3,099.3	2,565.1	3,094.4	3,646.8

The estimates under model $[AX_2][X_1X_2][BX_1]$ can be found in Table 8. In order to investigate the stability of the estimates we used a parametric bootstrap. The results are reported in Table 9. The seven estimated total numbers of severely injured are rather stable.

We conclude that, where classification by mode of transport in 2000 was stable, the refined classification of "motorized" into six categories in 2010 is usable for policy purpose when model $[AX_2][X_1X_2][BX_1]$ is applied.

5. Discussion

In this article we have presented a methodological framework that may be useful for the production of official statistics based on linked registers where additional categorical auxiliary variables are available. The methodology has potential for simultaneously solving the problems of undercoverage and of missing covariate values for those persons who are missed in some or all of the registers. This corresponds to solving the missing data problem for the grey bitmap and white parts in Figure 1.

5.1. Extensions

The EM-algorithm can also be used to solve the problem of missing data in covariates that are incompletely measured. There are many reasons why such data may be missing, including administrative errors or lags in recording data. If there is only a single register this is a simple missing data problem, but in the case of more than one register the extra information can help to complete these variables. The software we employ, the CAT-procedure in R, is able to handle this problem (Meng and Rubin 1991; Schafer 1997a,b).

Multiple imputation provides an alternative method for dealing with missing values in covariates. It was used, next to EM, by Gerritse et al. (2015a) and they argue that in their example, multiple imputation is more flexible. Their point is that in Table 3 the persons in the two cells labelled missing are most similar to persons not in the population register, and imputing from this subpopulation is easily accomplished using multiple imputation. But this approach is separate from the estimation of the unobserved part of the population, and does not benefit from the integrated way of dealing with these two issues.

Multiple imputation is however more natural in the case of continuous covariates, as used in Zwane and Van der Heijden (2008). Further research into the benefits of improving estimation using continuous covariates is also desirable. A more general strategy for

official statistics from linked registers which includes options for using categorical and continuous auxiliary variables in the estimation could then emerge, and an important element of that would be to have more examples of the usefulness of the approaches presented in this article.

If the framework is used to produce register based official statistics in a complex system with many registers, then it is more challenging to devise a procedure which ensures consistency between different outputs. Unless all the registers are linked, using the EM approach for different groups of registers using the same covariates, as would be likely in the case for age and gender, would lead to inconsistent outcomes. It is an open research question of how to build in this consistency.

The approaches presented here deal only with the problem of undercoverage. However, many registers also contain overcoverage, and this can have an effect on the undercoverage estimation by increasing the number of records to be linked. This will generally inflate population size estimates by inflating the number of records appearing in only one register, though it could have the opposite effect if the overcovered records appear in both registers. Zhang (2015) provides a framework for models to deal with overcoverage error, but it is important to have at least one source that does not suffer from overcoverage in order to make a suitable adjustment. More work is needed on how the estimation of undercoverage and overcoverage can be integrated into a set of procedures which can be applied in a wide range of situations including the production of official statistics.

5.2. Conclusion

The simulation studies show that, in comparison with the classical method where those partially observed covariates are ignored, the EM approach performs slightly better when the underlying MAR assumption and the conditional independence assumption for inclusion in the registers is met. When these assumptions are violated, both models can be severely biased.

In the last example in this article we showed how this missing data approach can be applied to the situation where a covariate of interest is measured in both registers.

Theoretically, the methodology can also be used when the number of covariates is large, where stability can be improved by making some of the covariates passive (compare Van der Heijden et al. 2012). In this instance there is little practical experience and we hope that this methodology will be used more so that the practical benefits become clearer.

6. References

Alho, J.M. 1990. "Logistic Regression in Capture-Recapture Models." *Biometrics* 46(3): 623–635. Doi: http://dx.doi.org/10.2307/2532083.

Baker, S.G. 1990. "A Simple EM Algorithm for Capture-Recapture Data with Categorical Covariates (with discussion)." *Biometrics* 46: 1193–1197. Doi: http://dx.doi.org/10.2307/2532461.

Bell, W.R. 1993. "Using Information from Demographic Analysis in Post-Enumeration Survey Estimation." *Journal of the American Statistical Association* 88(423): 1106–1118. Doi: http://dx.doi.org/10.2307/2290805.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete Multivariate Analysis, Theory and Practice*. New York: McGraw-Hill. Doi: http://dx.doi.org/10.1007/978-0-387-72806-3.

Brown, J., O. Abbott, and I. Diamond. 2006. "Dependence in the 2001 One-Number Census Project." *Journal of the Royal Statistical Society: Series A* (Statistics in Society) 169: 883–902. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00431.x.

Brown, J., O. Abbott, and P.A. Smith. 2011. "Design of the 2001 and 2011 Census Coverage Surveys for England and Wales." *Journal of the Royal Statistical Society: Series A* (Statistics in Society) 174(4): 881–906. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.00697.x.

Brown, J.J., I.D. Diamond, R.L. Chambers, L.J. Buckner, and A.D. Teague. 1999. "A Methodological Strategy for a One-Number Census in the UK." *Journal of the Royal Statistical Society: Series A* (Statistics in Society) 162(2): 247–267. Doi: http://dx.doi.org/10.1111/1467-985X.00133.

Buckland, S. and P. Garthwire. 1991. "Quantifying Precision of Mark-Recapture Estimates Using the Bootstrap and Related Methods." *Biometrics* 47: 255–268. Doi: http://dx.doi.org/10.2307/2532510.

Chao, A., P. Tsay, S. Lin, W. Shau, and D. Chao. 2001. "The Applications of Capture-Recapture Models to Epidemiological Data." *Statistics in Medicine* 20: 3123–3157. Doi: http://dx.doi.org/10.1002/sim.996.

Darroch, J., S. Fienberg, G. Glonek, and B. Junker. 1993. "A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability." *Journal of the American Statistical Association* 88: 1137–1148. Doi: http://dx.doi.org/10.2307/2290811.

Elliott, M.R. and R.J.A. Little. 2000. "A Bayesian Approach to Combining Information from a Census, a Coverage Measurement Survey, and Demographic Analysis." *Journal of the American Statistical Association* 95: 351–362. Doi: http://dx.doi.org/10.1080/01621459.2000.10474205.

Fienberg, S., M. Johnson, and B. Junker. 1999. "Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists." *Journal of the Royal Statistical Society: Series A* 162: 383–406. Doi: http://dx.doi.org/10.1111/1467-985X.00143.

Gerritse, S.C. 2016. "An Application of Population Size Estimation to Official Statistics. Sensitivity of Model Assumptions and the Effect of Implied Coverage." Utrecht University (dissertation), Utrecht, 2016. Available at: https://dspace.library.uu.nl/handle/1874/337476 (accessed February 1, 2018).

Gerritse, S.C., B.F.M. Bakker, and P.G.M. van der Heijden. 2015a. "Different Methods to Complete Datasets Used for Capture-Recapure Estimation: Estimating the Number of Usual Residents in the Netherlands." *Statistical Journal of IAOS* 31: 613–627. Doi: http://dx.doi.org/10.3233/SJI-150938.

Gerritse, S.C., P.G.M. van der Heijden, and B.F.M. Bakker. 2015b. "Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Loglinear Models." *Journal of Official Statistics* 31: 357–379. Doi: http://dx.doi.org/10.1515/jos-2015-0022.

Griffin, R.A. 2014. "Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020." *Journal of Official Statistics* 30: 177–189. Doi: http://dx.doi.org/10.2478/jos-2014-0012.

Héraud-Bousquet, V., F. Lot, M. Esvan, F. Cazein, C. Laurent, J. Warszawski, and A. Gallay. 2012. "A Three-Source Capture-Recapture Estimate of the Number of New HIV Diagnoses in Children in France from 2003–2006 with Multiple Imputation of a Variable of Heterogeneous Catchability." *BMC infectious diseases* 12(1) : 1. Doi: http://dx.doi.org/10.1186/1471-2334-12-251.

Huggins, R.M. 1989. "On the Statistical Analysis of Capture Experiments." *Biometrika* 76(1): 133–140. Doi: http://dx.doi.org/10.1093/biomet/76.1.133.

International Working Group for Disease Monitoring and Forecasting. 1995. "Capture-Recapture and Multiple Record Systems Estimation. Part I. History and Theoretical Development." *American Journal of Epidemiology* 142: 1047–1058. Doi: http://dx.doi.org/10.1093/oxfordjournals.aje.a117558.

Madigan, D. and J.C. York. 1997. "Bayesian Methods for Estimation of the Size of a Closed Population." *Biometrika* 84: 19–31. Doi: http://dx.doi.org/10.1093/biomet/84.1.19.

Meng, X.L. and D.B. Rubin. 1991. "IPF for Contingency Tables with Missing Data via the ECM Algorithm." In Proceedings of the Statistical Computing Section of the American Statistical Association, 244–247. Washington D.C.: American Statistical Association.

Pelle, E., D.J. Hessen, and P.G.M. van der Heijden. 2016. "A Log-Linear Multidimensional Rasch Model for Capture-Recapture." *Statistics in Medicine* 35: 622–634. Doi: http://dx.doi.org/10.1002/sim.6741.

Pollock, K.H. 2002. "The Use of Auxiliary Variables in Capture-Recapture Modelling: an Overview." *Journal of Applied Statistics* 29: 85–102. Doi: http://dx.doi.org/10.1080/02664760120108430.

Reurings, M.C.B. and N.M. Bos. 2012. "Ernstig Verkeersgewonden in de Periode 2009 en 2010. Update van de Cijfers." Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV, ref.; R-2012-7, Leidschendam, 2012. Available at: https://www.narcis.nl/publication/RecordID/oai:library.swov.nl:129380.

Reurings, M.C. and H.L. Stipdonk. 2011. "Estimating the Number of Serious Road Injuries in the Netherlands." *Annals of Epidemiology* 21(9): 648–653. Doi: http://dx.doi.org/10.1016/j.annepidem.2011.05.007.

Schafer, J. 1997a. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall/CRC.

Schafer, J. 1997b. *Imputation of Missing Covariates under a General Linear Mixed Model*. PennState University, Department of Statistics.

Sutherland, J.M., C.J. Schwarz, and L.-P. Rivest. 2007. "Multilist Population Estimation with Incomplete and Partial stratification." *Biometrics* 63: 910–916. Doi: http://dx.doi.org/10.1111/j.1541-0420.2007.00767.x.

Tilling, K. and J.A.C. Sterne. 1999. "Capture-Recapture Models Including Covariate Effects." *American Journal of Epidemiology* 149(4): 392–400. Doi: http://dx.doi.org/10.1093/oxfordjournals.aje.a009825.

Van der Heijden, P.G.M., E. Zwane, and D. Hessen. 2009. "Structurally Missing Data Problems in Multiple List Capture-Recapture Data." *Advances in Statistical Analysis* 93: 5–21. Doi: http://dx.doi.org/10.1007/s10182-008-0098-6.

Van der Heijden, P.G.M., J. Whittaker, M. Cruyff, B. Bakker, and R. van der Vliet. 2012. "People Born in the Middle East but Residing in the Netherlands: Invariant Population Size Estimates and the Role of Active and Passive Covariates." *The Annals of Applied Statistics* 6(3): 831–852. Doi: http://dx.doi.org/10.1214/12-AOAS536.

Whittaker, J. 1990. *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Wolter, K.M. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81(394): 337–346. Doi: http://dx.doi.org/10.1080/01621459.1986.10478277.

Zhang, L.-C. 2015. "On Modelling Register Coverage Errors." *Journal of Official Statistics* 31: 381–396. Doi: http://dx.doi.org/10.1515/jos-2015-0023.

Zwane, E. and P.G.M. van der Heijden. 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. Doi: http://dx.doi.org/10.1002/sim.2577.

Zwane, E. and P.G.M. van der Heijden. 2008. "Capture-Recapture Studies with Incomplete Mixed Categorical and Continuous Covariates." *Journal of Data Science* 6: 557–572.

Zwane, E., K. van der Pal-de Bruin, and P.G.M. van der Heijden. 2004. "The Multiple-Record Systems Estimator when Registrations Refer to Different but Overlapping Populations." *Statistics in Medicine* 23: 2267–2281. Doi: http://dx.doi.org/10.1002/sim.1818.

Zwane, E. and P. van der Heijden. 2005. "Population Estimation Using the Multiple System Estimator in the Presence of Continuous Covariates." *Statistical Modelling* 5(1): 39–52. Doi: http://dx.doi.org/10.1191/1471082X05st086oa.

Received April 2016
Revised September 2017
Accepted September 2017