Peter G.M. van der Heijden

Multiple record system estimation with overlapping lists, where some lists cover only part of the population[1]

Peter G.M. van der Heijden[1,2]; Paul A. Smith[2]; Maarten Cruyff[1]

[1] Department of Methodology and Statistics, Utrecht University
[2] Department of Social Statistics & Demography, University of Southampton

Abstract:
In dual system estimation and multiple system estimation problems there may be lists available that only cover subsets of the population by design. We discuss under which assumptions such structural undercoverage in the lists can be ignored in these estimation problems.

Keywords:
Capture-recapture; multiple system estimation; population size estimation; list coverage; administrative data

1. Introduction

In population size estimation (Boehning et al., 2017) it regularly happens that lists are available that do not cover the full population. The problem we discuss in this contribution is whether, or under what conditions, such lists can be used in estimating the size of the full population. We encountered this problem in the estimation of ethnicity in New Zealand, where this part coverage is related to the age of individuals.

Zwane, van der Pal-de Bruin and van der Heijden (2004) approached the problem of lists that do not fully cover the population as a missing data problem. They provided solutions assuming the missingness (i.e. the undercoverage) is ignorable, which here means that the relationships in the missing part(s) of the list(s) is the same as the relationship in the overlap between the lists (from which it can be estimated). This is often not a priori unrealistic, as the missingness is due to the design of the lists, for example because lists cover certain age ranges, in contrast to other missing data problems where ignorability is often unrealistic because it is related to unobserved variables).

In the current paper we reframe the part-coverage of a list as a collapsibility problem for a covariate. Using results from van der Heijden et al. (2012), we show under what conditions we may assume collapsibility over such a covariate.

In section 2, we first provide theory and then, in section 3, apply the theory in a descriptive way to the problem of estimating the sizes of ethnic populations in New Zealand. For an extensive discussion, we refer to Van der Heijden et al. (2018 and in press).

[1] This paper contains sections and figures taken from van der Heijden et al. (in press).

2. Methodology

A first paper working on partly overlapping lists in population size estimation was Zwane et al. (2004). We first describe their contribution and then reframe their results using the work of Van der Heijden et al. (2012) on collapsibility over covariates.

Zwane et al. provide two examples, one where partial overlap can be ignored and one where it cannot be ignored. Example 1: There are two regions, south and north. List A covers the full population in the north and the south and list B only the north region. Assume lists A and B are linked, ignoring the region individuals live in. Thus individuals living in the south are not linked. Zwane et al. prove that a standard dual system estimation of the linked data provides an unbiased estimate of the population size under the assumption that the inclusion probability for list A in the south is identical to the inclusion probability for list A in the north region. Example 2: There are three regions, south, middle and north. List A covers south and middle and list B middle and north. Now linking lists A and B, ignoring the region individuals live in, and calculating the dual system estimate will lead to a biased estimate. Zwane et al. propose missing data methodology where, including the variable region, unbiased estimates are provided under specific loglinear models. The general structure is that for the north, list A is missing, and for the south, list B is missing. The missing data methodology uses the information available in the sub table for the middle region, where there is no structural missingness of list A and B, to impute data for when list A is missing (in the sub table for the north region) and for when list B is missing (in the sub table for the south region). Note that in the middle region there are individuals only in list A, only in list B and in both list A and B. For the north region, where list A missing, the missing data methodology also provides an estimate of the individuals that would have been observed only in A, if list A were not missing. Zwane et al. (2004) provide this missing data methodology for any number of lists. The topic of this paper is not the missing data methodology proposed by Zwane et al. (2004), but instead proposing a different understanding of when covariates such as region can be ignored. We use the notion of collapsibility of covariates in population size estimation. See Figure 1, taken from van der Heijden et al. (2012).

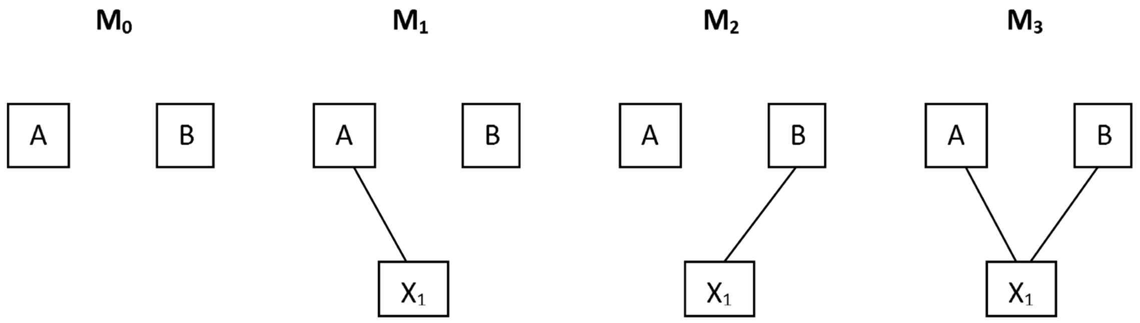

Figure 1. Interaction graphs for loglinear models with two lists and one covariate, taken from van der Heijden et al. (2012),

Four models are shown, M0 to M3. In all models there are two lists, A and B, and no interaction between them, as there is insufficient data to identify it. In model M0 there is no covariate (this is the classical dual system estimator where the interaction is assumed to be zero). In model M1 there is a covariate X1 that is related to list A but not to list B, meaning that inclusion probabilities are heterogeneous across the levels of X1 for A but homogeneous for B. In model M2 it is the other way around. In model M3 inclusion probabilities for both A and B are heterogeneous across the levels of X1. Van der Heijden et al. (2012) show that the total population size estimate is identical in models M0, M1 and M2 and different from the estimate in M3. Therefore, under models M1 and M2 the three-way array of variables A, B and X1 can be added over the values of X1, which collapses it to a two-way array, without affecting the population size estimate. A collapsible model is one

where summing over a covariate does not alter the population size estimate. Model M3 is not collapsible over X1, because summing over X1 changes the population size estimate. Van der Heijden et al. use the concept of a short path in the interaction graph to identify a collapsible model. A short path is a sequence of connected nodes in the graph which does not contain a sub-path – a shorter path between the terminal nodes through at least one of the same intermediate nodes; note that it need not be short in the sense of having few nodes. A model is not collapsible over a variable on a short path, and is collapsible over a variable which is not on a short path (including when there is no path between the nodes). For full details see van der Heijden et al. (2012). In M3 the covariate lies on a short path between A and B and therefore one cannot collapse over X1. In M1 and M2 the covariate X1 does not lie on a short path and in these cases one can collapse over the covariate.
Now we reframe the results from Zwane et al. (2004). Example 1 is represented by model M2 in Figure 1. Region is not on a short path between A and B, so the model can be collapsed over region – in other words, the partial coverage of list B can be ignored (which agrees with Zwane et al.'s result). Example 2 is described by model M3: A has heterogeneous inclusion probabilities over the levels of region, as the inclusion probabilities are positive for the north and the middle but zero for the south, and so does B as the inclusion probabilities are zero in the north and positive in the middle and south. Hence there are two edges, one from X1 to A and another from X1 to B, and therefore X1 is on a short path from A to B and it follows that we cannot collapse over (i.e. ignore) the region covariate (which also agrees with Zwane et al.'s conclusion).
We now extend this to three lists, A, B and C. See Figure 2.

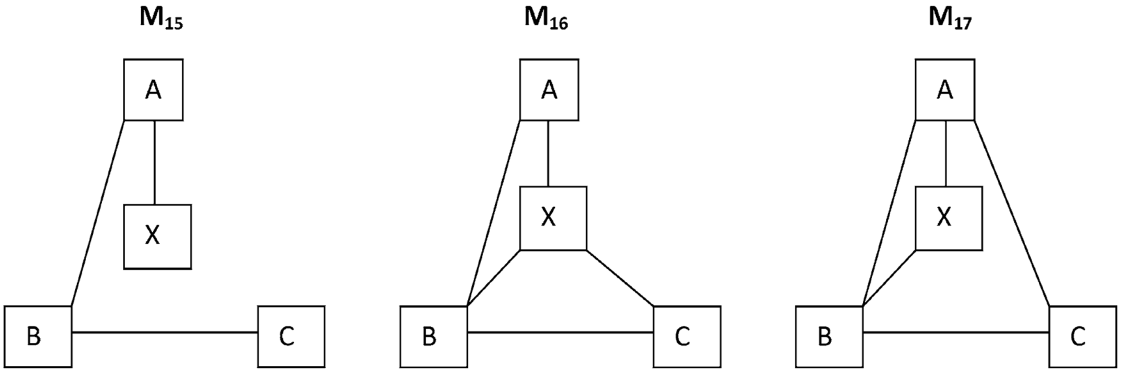

Figure 2. Interaction graphs for loglinear models with three lists and one covariate, taken from van der Heijden et al. (2012).

In M15, list A is conditionally independent from C given B, and A is related to X. Here X is not on any short path and we can collapse over X without affecting the population size estimate. In M16, list A is conditionally independent from C but now X is related to A, B and C. X is on a short path from A to C and collapsing over X will bias the population size estimate. In M17, all lists are pairwise dependent, and X is related to lists A and B. However, X is not on a short path from A to B (as the short path is the edge between A and B), and we can collapse over covariate X without affecting the population size estimate.

3. Application

We applied this methodology in Van der Heijden et al. (in press), where we estimated the size of the Māori and non-Māori population in New Zealand. We refrain from discussing the estimation problem in full, as it also included the problem that individuals did not always provide their ethnicity, as well as that some individuals provided different ethnicities over the different lists. Thus the estimation problem is too difficult to lay out in this paper. The focus is here on the partial coverage of some of the lists that were involved in this estimation problem.

The data are probabilistically linked in Stats NZ's Integrated Data Infrastructure (IDI). The IDI provides safe access to anonymized linked microdata for research and statistics in the public interest. Data sources in the IDI (including the census) are linked to a central population spine. The population used here is the experimental administrative{based New Zealand resident population known as the `IDI-ERP'. Perfect linkage is an essential assumption for MSE. An incorrect link could mean that the wrong ethnicity is associated with a person. In this application, if records in the lists have not been linked to the IDI spine, they do not enter the analysis, and become part of the unobserved population for the list.

For 2013 we have three administrative sources and the census. The three administrative sources are (i) Department of Internal Affairs (DIA) birth registrations data, (ii) Ministry of Education (MOE) tertiary education enrolment data, and (iii) Ministry of Health (MOH) National Health Index system, a unified national person list. Each of the administrative registers relates to different parts of the population. Birth registrations are for babies born in NZ since 1998, or those up to age 14 in 2013; tertiary education enrolments are available from the late 1990s, and include a range of education enrolments for those aged around 13 and older in 2013; both census and health data include all ages, and each list has an ethnicity reported for around 90 % of the IDI-ERP population. Overall, almost 99 % of the IDI-ERP population have ethnicity information from at least one of these lists, and many people have information from more than one list. Table 1 provides the observed counts for ethnicity, where - stands for item missingness (individuals that do not have their ethnicity registered in this list) and x stands for individuals that are not part of a list. For example, in the Census 3,225,804 are registered as non-Māori, 560,427 as Māori, for 20,619 individuals in the census no ethnicity is reported, and 595,140 individuals are missed by the census but appear in at least one of the other lists.

Table 1: Summary of Census linked to DIA, MOH and MOE, observed numbers. '-' is for being on the list (or in the Census) but no ethnicity is provided. 'x' is for not being in the list (or in the Census).

	Census	DIA	MOH	MOE
non--Māori	3,225,804	574,077	3,527,874	1,763,463
Māori	560,427	236,673	617,205	405,063
-	20,619	6,045	188,781	20,424
x	595,140	3,585,195	68,130	2,213,040
TOTAL	4,401,990	4,401,990	4,401,990	4,401,990

In van der Heijden et al. (in press) there is a model for the Census and MOH, both of which aim to cover the full population, and DIA, the birth registration started in 1998. Conceptually, we consider there to be a covariate Age, for which we do not have data: inclusion probabilities are high for individuals born from 1998 onwards and zero before that time. This gives a graph similar to M15 in Figure 2. Age is not on a short path, so the model can be collapsed over it; that is, we can treat DIA as if it covered the whole population, without affecting the estimated population size. When we add the MOE register, we assume for this register that there is an unmeasured covariate that predicts entry to the forms of education covered by the register and which is also related to age, as enrolments are available starting in the late 1990s. We also assume that there are other factors that lead one to go into tertiary education, and that there are unmeasured covariates which capture this information. Age is therefore related to two lists but if there is also a direct link between the two lists Age is not on a short path, and the model is collapsible over Age (see van der Heijden et al., 2012). The other factors leading one to go to tertiary education are further covariates that are only related to MOE and not to the other lists and the model is therefore collapsible over these covariates too.

Summarizing, in van der Heijden et al. (in press) we take the position that, even though DIA and MOE cover only part of the population, this should not withhold us from using standard modelling approaches in multiple system estimation that include DIA and MOE, as long as the direct edge between DIA and MOE is in the model used in population size estimation.

4. Discussion and Conclusion:

We have shown that, in the presence of overlapping lists that cover only part of the population, standard dual and multiple system estimation can be used as long as certain conditions are met. If these conditions are violated, and covariates describing the presence of individuals in the lists are available (such as age in the example discussed in Section 3), then the missing data methodology provided in Zwane et al. (2004) can be used to obtain population size estimates.

References:

Boehning, D., P.G.M. van der Heijden & J. Bunge (2017). *Capture-recapture methods for the social and medical sciences.* Boca Raton: CRC Press. (429 pages).

Van der Heijden, P.G.M., M. Cruyff, P.A. Smith, C. Bycroft, P. Graham and N. Matheson-Dunning (in press). Multiple system estimation using covariates having missing values and measurement error: estimating the size of the Māori population in New Zealand. *Journal of the Royal Statistical Society, Series A.*

Van der Heijden, P.G.M., P.A. Smith, M. Cruyff and B.F.M. Bakker (2018). An overview of population size estimation where linking registers results in incomplete covariates, with an application to mode of transport of serious road casualties. *Journal of Official Statistics. 34,* 239-263.

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. *Annals of Applied Statistics, 6,* 831-852.

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