MEASUREMENT ERRORS IN MULTIPLE SYSTEMS ESTIMATION

Paul A. Smith[1], Peter G.M. van der Heijden[12] and Maarten Cruyff[2]

[1] Department of Social Statistics & Demography, University of Southampton, UK, (e-mail: p.a.smith@soton.ac.uk)

[2] Dept of Social Sciences, Methodology and Statistics, Utrecht University, Netherlands, (e-mail: P.G.M.vanderHeijden@uu.nl, m.cruyff@uu.nl)

ABSTRACT: Dual and multiple system estimation use the presence ('capture') of people in different data sources as the basis for estimation of the population size. Where further characteristics are also available, these can be used to provide estimates of the population size classified by these characteristics. We consider the situation that there are measurement errors in these classifying variables, but not in the linkage of people between data sources. We consider strategies to produce estimates of the population size and breakdown using a consistent, adjusted definition taking account of all the evidence in the collected data sources.

KEYWORDS: capture-recapture, latent class analysis, ethnicity, population size estimation.

1 Introduction

Dual and multiple system estimation have a long history of use to estimate the size of populations which cannot be completely observed, and in recent years there have been many applications to estimating the size of human populations. In the simplest cases this may result from observing people on two sources, and using an assumption of independence between the sources to obtain an estimated population size. When there are more sources, interactions between the sources can be fitted, and an appropriate model needs to be fitted to the (implied) contingency table formed from the presence or absence of people in each source. In general this procedure assumes no errors of observation, and that no errors are made in linking people on the different sources. If an independent estimate of the linkage errors can be obtained, it can be used in an adjusted estimator (Zult *et al.* 2021). However, in this paper we work with the usual framework that assumes that linkage is made perfectly.

Auxiliary information is often available on the different sources, in addition to the existence of a person (or record), and this information may be used in linkage where it corresponds to a stable characteristic. Other variables are of more substantive interest, and may be expected to vary between sources for a number of reasons: they may be characteristics which vary in time, or they may be measured in different ways in different data sources, leading to variations in the measurement. In this latter case

we may consider that there is an underlying 'true' variable, and that one or more of the sources that we are using observe a version of this variable with some added measurement error. The process of linking datasets means that some variables will not be observed for some records, and that no variables are observed for records in none of the sources, the number of which will nevertheless be estimated during population size estimation.

In this paper we consider strategies for dealing with population size estimation, broken down using variables measured in one or more sources and subject to measurement error in this way. Section 2 deals with solutions based on explicit decisions about which measure is the best, and with simple combinations of variables, and Section 3 with the use of a latent class model to derive an underlying measure, which we consider to estimate the 'true' measure based on the available data.

2 Population size estimation with a preferred covariate source

First we consider that there are two sources, and both sources contain what is conceptually the same covariate, though we know or suspect that they are measured differently, or that their resulting quality is different because of the way they are collected. Van der Heijden *et al*. (2018) present an example where characteristics of accidents, specifically whether a motorised vehicle was involved, are recorded both by the police and by hospitals. It would be possible to treat these as the same variable, but investigation of the data where the 'motorised vehicle' variable is available from both sources shows that about 5% of cases have discrepancies. Instead we treat them as two different variables, and construct a four-way contingency table formed from presence/absence on the two sources and the motorised/non-motorised variables in the two sources. We then use loglinear modelling to choose a suitable model for this contingency table, and use this model together with the EM algorithm to produce a completed table (Table 1), which provides an estimate of the missing part of the population, and also estimates of the population sizes in each cell of the contingency table (where they are not observed). This allows us to add up in any way we want to achieve a set of consistent estimates.

In the accidents example, we have reasonable confidence that the police register is better at recording whether a motorised vehicle is involved, as gathering this

		B = 1		B = 0		Total
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	
A = 1	$X_1 = 1$	5970.0	287.0	1289.0	62.0	7608.0
	$X_1 = 0$	28.0	256.0	6.9	63.1	354.0
A = 0	$X_1 = 1$	2933.2	2177.6	633.3	470.2	6214.3
	$X_1 = 0$	13.8	1942.4	3.4	478.8	2438.4
Total		8945.0	4663.0	1932.6	1074.1	16614.7

Table 1: Completed road accidents table. A is the police register, B the hospital register, X_1 is the police record of motor vehicle involvement, and X_2 the hospital record.

information is part of the police function. So we consider the classification of the total according to the variable X_1 in the police register (whether observed or estimated) to be the correct one. And since the full dataset, cross-classified by both police and hospital versions of the motorised vehicle variable is available, we can make inferences about the measurement error in the hospital version.

In a situation where the relative merits of the measurements are less clear, we could pragmatically use the average of the population size estimates under the different versions of the auxiliary variable.

3 Latent class models

A further approach is to treat the different measurements separately in the population size estimation, but then to embed them in a latent class model (LCM), which postulates an underlying, unobserved parameter related to all the separate measurements, and which can be interpreted as the true parameter. This approach can be considered when there are at least three measurements. It is conceptually different from using LCMs to deal with heterogeneity in the capture probabilities (as in Stanghellini & van der Heijden 2004). Van der Heijden et al. (2021) apply this approach in analysing four linked data sources in New Zealand – the population census, the health register, the birth registration register, and an education register (covering largely, but not only, tertiary education). Each of these sources includes an ethnicity variable, which we consider in a simplified version recoded to Māori or all other ethnicities. We would like to estimate both the size of the population in New Zealand and the size of the Māori population based on these sources.

Two approaches are possible. In the first, we treat the four sources using multiple system estimation, fitting a loglinear model to the eight-way table formed by the inclusion or not in each source and Māori ethnicity or not. Some of the estimates from a saturated model go to infinity, so a reduced form of the model is needed to obtain parameter estimates with reasonable interpretation and stability. The estimates arising from this model (including the estimates of the size of the unobserved part of the population) are then used as the inputs to a latent class model with two latent classes. This gives a two-part procedure which has the advantage of being close to the original model for the 8-way contingency table. The model produces estimates of the size of the Māori population from one of the latent classes, which can be interpreted as the true Māori variable. It can also be used to give estimates of the errors in the four observed variables in measuring this underlying Māori concept.

The second approach aims to include the latent class model directly in the modelling of the 8-way contingency table. The assumption of the latent class model is that the (unobserved) interactions between the observed variables and the latent variable explain all the interactions within the observed variables in the original data. Therefore we replace [abcd] in the original model with [aX][bX][cX][dX] with the latent variable X (where a, b, c and d label the ethnicity variables in the four data sources). This replaces all interactions of a, b, c, and d, so any terms containing two or more of these parameters are dropped from the model (which serves to make the loglinear model hierarchical with respect to interactions, and therefore more easily

			census	DIA	MOH	DOE
		π_x	$\pi^a_{r=1\|x}$	$\pi^b_{r=1\|x}$	$\pi^c_{r=1\|x}$	$\pi^d_{r=1\|x}$
two-step	class 1	0.827	0.004	0.016	0.003	0.015
	class 2	0.173	0.937	0.937	0.826	0.922
LCMSE	class 1	0.834	0.007	0.014	0.004	0.016
	class 2	0.166	0.957	0.958	0.857	0.959

Table 2: Estimates of probabilities from latent class models with two latent classes. Class $r = 1$ is interpreted as non-Māori, and class 2 as Māori.

interpretable). This leaves a latent class model embedded in the multiple system estimation, and van der Heijden *et al*. (2021) call this the latent class multiple system estimation (LCMSE) model.

In the application to data from the New Zealand Integrated Data Infrastructure (IDI-ERP), the LCMSE has a slightly lower estimate of the number of Māori and a slightly higher overall population estimate than the two-step procedure based on latent class estimation using the multiple system estimation results. The LCMSE therefore takes a more conservative approach to the definition of Māori in this dataset.

The population census has been generally held to be the best measure of Māori ethnicity among the different sources available in New Zealand, and it has low values for measurement error in both Māori and non-Māori in both approaches (Table 2). The Health register has a low error for non-Maori, but the largest measurement error for Māori. The births and education registers are similar to the census in the estimated measurement error in the Māori class, but have more error in estimating the non-Māori class. Therefore overall our results support the conclusion that the census is the best overall measure of ethnicity.

References

STANGHELLINI, E. & VAN DER HEIJDEN, P.G.M. (2004). A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account. *Biometrics*, **60**, 510–516.

VAN DER HEIJDEN, P.G.M., CRUYFF, M., SMITH, P.A., BYCROFT, C., GRAHAM, P. & MATHESON-DUNNING, N. 2021 (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., SMITH, P.A., CRUYFF, M. & BAKKER, B. 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.

ZULT, D., DE WOLF, P.-P., BAKKER, B.F.M. & VAN DER HEIJDEN, P.G.M. 2021 (in press). A general framework for multiple-recapture estimation that incorporates linkage error correction. *Journal of Official Statistics*.