Received: 28 August 2020 | Accepted: 31 May 2021

DOI: 10.1111/rssa.12731

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Multiple system estimation using covariates having missing values and measurement error: Estimating the size of the Māori population in New Zealand

Peter G. M. van der Heijden[1,2] | **Maarten Cruyff**[1] | **Paul A. Smith**[2] | **Christine Bycroft**[3] | **Patrick Graham**[3] | **Nathaniel Matheson-Dunning**[3]

[1]Utrecht University, Utrecht, The Netherlands

[2]University of Southampton, Southampton, UK

[3]Statistics New Zealand, Wellington, New Zealand

Correspondence
Peter G. M. van der Heijden, Utrecht University, Utrecht, The Netherlands.
Email: p.g.m.vanderheijden@uu.nl

Abstract

We investigate the use of two or more linked lists, for both population size estimation and the relationship between variables appearing on all or only some lists. This relationship is usually not fully known because some individuals appear in only some lists, and some are not in any list. These two problems have been solved simultaneously using the EM algorithm. We extend this approach to estimate the size of the indigenous Māori population in New Zealand, leading to several innovations: (1) the approach is extended to four lists (including the population census), where the reporting of Māori status differs between registers; (2) some individuals in one or more lists have missing ethnicity, and we adapt the approach to handle this additional missingness; (3) some lists cover subsets of the population by design. We discuss under which assumptions such structural undercoverage can be ignored and provide a general result; (4) we treat the Māori indicator in each list as a variable measured with error, and embed a latent class model in the multiple system estimation to estimate the population size of a latent variable, interpreted as the true Māori status. Finally, we discuss estimating the Māori population size from administrative data only. Supplementary materials for our article are available online.

KEYWORDS
administrative data, capture–recapture, latent class model, list coverage, population size estimation

1 | INTRODUCTION

The use of dual system estimation (DSE, also known as capture–recapture or the Lincoln—Peterson estimator) to estimate the size of a population which cannot be completely observed has become widespread in official statistics, particularly as a key part of making estimates from a population census (e.g., Brown et al., 1999, 2019; Wolter, 1986), although also in situations involving the use of linked administrative data sources (Bakker et al., 2015, and references therein; Zhang & Dunne, 2018). The need to make efficient use of data already available to government in the construction of official statistics outputs has led to better access to administrative data, but there are important challenges in making use of it (see, e.g. Hand, 2018). Linkage of the records from these sources is being widely used to understand the under- and overcoverage within them and estimate coverage corrections. We will use 'lists' as a generic term for all sources containing lists of identifiable units.

When two lists are linked, in general, there will be some records which remain unlinked, as there is no corresponding record in the other source. This leads to missing data for any variables which appear in only one list (item missingness). The linked data can be used to estimate the size of the population that is not present in either list, and for these unobserved records *all* the variables are missing (unit missingness). There is an extensive line of research that treats this problem from a missing data perspective, starting with Zwane and van der Heijden (2007), and summarised and extended in van der Heijden et al. (2012, 2018). De Waal et al. (2019, 2020) provide an overview of what they call multisource statistics, where they categorise this line of research under the subject 'Overlapping variables and overlapping units' and 'Undercoverage'. Another overview of the field is provided by Zhang and Chambers (2019). The missing data methodology involves the EM algorithm, using a suitable loglinear model specified by the researcher. In the EM algorithm, an expectation step and a maximisation step are alternated until convergence. In the E-step, the expected values of the missing data are derived given the observed data and the current best fitted values found in the preceding M-step. In the M-step, maximum likelihood estimates are found for the chosen model using the completed data from the E-step; see Zwane and van der Heijden (2007) for the full details. Standard errors for the estimates can be calculated using the parametric bootstrap (Buckland & Garthwaite, 1991). Van der Heijden et al. (2018) concluded that further practical experience with these methods is needed to demonstrate their usefulness in a variety of situations and encouraged their wider application.

Here we apply this existing approach to a larger problem, using four lists to estimate the size of the Māori ethnic population in New Zealand. Section 2 describes the context of this research, the lists available and the procedures which have been used to link them. We build up the estimation problem in Section 3, starting with two lists, and then progressing to three and four lists. In Section 3.5, we consider estimates obtained only from the administrative registers to examine how good estimates would be in the absence of a population census. The lists also contain item missingness because people do not always provide their ethnicity. We extend the approach of van der Heijden et al. (2012, 2018) to deal with this additional missing data problem.

Some of these lists cover only parts of the population, a common situation in official statistics applications of population size estimation. For the estimation of ethnicity in New Zealand this part coverage is related to the age of individuals. Zwane et al. (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 of van der Heijden et al. (2012), we show under what conditions we may assume collapsibility over such a covariate, and hence, a general result for when the part-coverage of lists may be ignored. This is discussed in Section 3.3.

Section 4 introduces the idea that a different concept of Māori identity is measured in each list because of differences in context or timing, an extension of the idea that the same variable is measured differently in different lists (van der Heijden et al. 2018). The model is extended to include latent classes to represent an underlying concept of Māori/non-Māori. Under this model, a single estimate is obtained for the size of the Māori and non-Māori populations (instead of estimates for each list separately). Section 5 concludes by discussing the effectiveness of these approaches and the corresponding estimates of the size of the Māori ethnic population, and how the differences in the estimates derived from the different lists may be interpreted. It also discusses the sensitivity of the approaches to assumptions of perfect linkage and no overcoverage.

2 | MĀORI ETHNICITY AND RELEVANT LISTS

Ethnicity is the principal measure of cultural identity in New Zealand, and is used across the official statistics system. Identifying the indigenous Māori population is of particular importance due to the partnership and obligations between Māori and the crown established under the Treaty of Waitangi of 1840. Māori are also a key group of policy interest; for example Waldon (2019) discusses the way measurement of ethnicity supports the measurement of health outcomes for indigenous peoples in New Zealand.

Ethnicity is therefore regularly included in statistical and administrative data collections. However, differences in questions or context and changes in perceptions can all affect how ethnicity is recorded in different lists (Simpson et al., 2016). Particularly for indigenous peoples such as the Māori, there is a need to ensure that definitions are created which take into account the ways in which members of these communities perceive themselves (Madden et al., 2019).

Official population estimates and projections for major ethnic groups in New Zealand are based principally on the responses people provide in the five yearly census, adjusted for non-response using a post-enumeration survey. As part of its census transformation programme, Statistics New Zealand is exploring the feasibility of a census based on administrative data (Statistics New Zealand, 2012, 2014). The ability to produce ethnicity data from administrative registers is a key consideration. Ethnicity is collected independently in a number of administrative registers as well as through the census. People do not always report the same ethnicity in each register and sometimes do not report their ethnicity at all, so there is an additional missingness problem to deal with.

A key question is how to combine ethnicity from multiple lists, when information is sometimes conflicting. Reid et al. (2016) compared ethnicity data from the 2013 Census with the ethnicity information collected by administrative registers, for a New Zealand resident population derived from administrative registers. They found that nearly everyone in this administrative data-based New Zealand resident population had ethnicity recorded in at least one administrative register, but that consistency

with census responses varied considerably by register and by ethnic group. The method used to combine these registers has a major impact on the result. Under the assumption that census responses provide the best measure for official statistics purposes, a method that ranks registers based on their consistency with the census was applied. Using administrative data alone was found to produce a time series that reflects expected patterns of increasing ethnic diversity, with the age structure and regional distribution of ethnicity consistently in line with official measures (Statistics New Zealand, 2018). We note that, according to Statistics New Zealand's standard classification of ethnicity, used in many administrative systems, people can and do provide multiple ethnicities. Here we focus on Māori ethnicity so that we have two mutually exclusive categories: Māori (with or without other ethnicities) and non-Māori (everyone else).

Four lists are available, the population census and three administrative registers, each with an ethnicity variable. In this application, we use ethnicity information from the linked lists to explore alternative ways to produce official ethnic population estimates in New Zealand, and also investigate estimates produced from only the administrative data as a possible replacement for the census. In support of this, we analyse a variety of combinations of the census and administrative registers using the approach of Zwane and van der Heijden (2007), with a specific focus on the estimation of the size of the Māori population at the time of the 2013 population census. The analysis requires the extension of the methods to deal with multiple lists and with a variety of different types of missing data.

The population used here is the experimental administrative-based New Zealand resident population known as the 'IDI-ERP' (Statistics New Zealand, 2017a). The IDI-ERP is derived using signs of activity in government sources. Those who have died, or who have moved to live overseas before the reference date are excluded to minimise overcoverage, although some non-residents may remain in the data set. We are interested in the population size, and therefore will implicitly assess the coverage of different sources within IDI-ERP relative to our estimate of the size of the New Zealand population.

The data are probabilistically linked in Stats New Zealand's Integrated Data Infrastructure (IDI). The IDI provides safe access to anonymised 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. Perfect linkage is an essential assumption for DSE. 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.

The three administrative registers are:

- Department of Internal Affairs (DIA) birth registrations data—which includes the ethnicity of the child as reported at registration
- Ministry of Education (MOE) tertiary education enrolment data—which includes ethnicity for students
- Ministry of Health (MOH) National Health Index system, a unified national person list—which includes ethnicity

For a more detailed explanation of these registers, see Reid et al. (2016).

Each of the administrative registers relates to different parts of the population. Birth registrations are for babies born in New Zealand 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 the 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.

The aim is to produce aggregate estimates of Māori and non-Māori ethnicity by combining these four lists: the 2013 Census and the three administrative registers. In Section 3.3. we discuss the problem that the four lists cover different parts of the population.

3 | BUILDING UP THE POPULATION ESTIMATES

3.1 | Two lists

We start by using the two lists with the widest coverage, the Census and the MOH. Being in the Census is denoted by A ($A = 1$ for 'yes' and $A = 0$ for 'no'), and similarly for MOH, denoted by C. The ethnicity variable in the Census is denoted by a ($a = 0$ for non-Māori, $a = 1$ for Māori, $a = -$ for individuals who are in A but did not fill in their ethnicity, and $a =$ x for individuals that are not in A). The ethnicity variable in the MOH is denoted by c and coded similarly to a. In comparison to the methods employed by van der Heijden et al. (2018), the presence of the − level in variables a and c is new, and we first extend the approach to deal with this new level with two lists.

Figure 1 illustrates the form of the data when they are coded in a matrix of individuals in the rows by variables in the columns. In the middle two columns, we depict A and C, that indicate whether individuals are only in A but not in C ($(A, C) = (1, 0)$), in both A and C ($(A, C) = (1, 1)$) or not in A but only in C ($(A, C) = (0, 1)$). At the bottom, we find a horizontal band of 'Individuals missed by both lists', and this refers to $(A, C) = (0, 0)$. This last number has to be estimated to arrive at an estimate of the size of the total population of non-Māori and Māori. The first column stands for ethnicity variable a. When individuals are in A ($(A, C) = (1, 0)$ or $(A, C) = (1, 1)$), there are three types of individuals, namely 0, non-Māori (light grey); 1, Māori (checkerboard pattern); and −, those who have a missing value for ethnicity (gridded pattern). When individuals are not in A but only in C, the ethnicity variable a is automatically not measured and denoted by x (white area). The last column stands for the ethnicity variable c, and it has similar levels to a. Notice that there are three kinds of missing data: there is item missingness − for those individuals that are on a list but did not provide their ethnicity; there is item missingness x for those individuals that are not on one list, and hence have no value on the corresponding ethnicity variable (if only $A = 0$, $a =$ x, and if only $C = 0$, $c =$ x). Last, there is unit missingness for those individuals that are in neither A nor C.

TABLE 1 Summary of Census linked to DIA, MOH and MOE, observed numbers

	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	6045	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

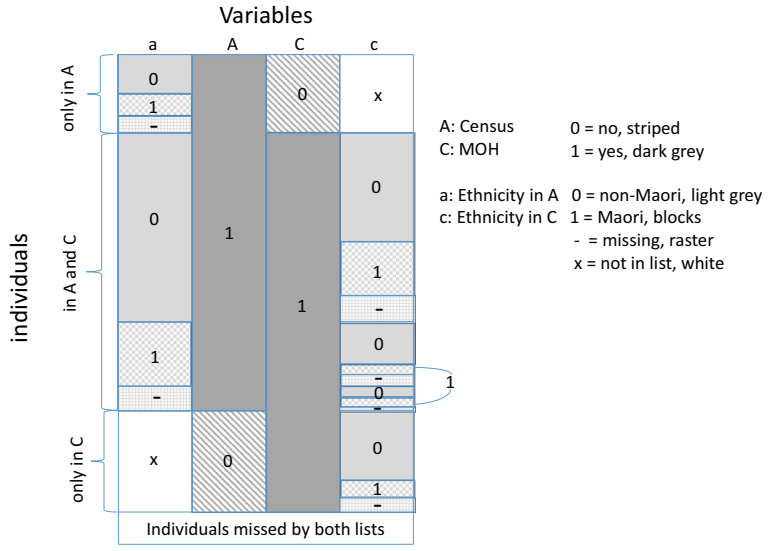

FIGURE 1 Graphical representation of two linked lists [Colour figure can be viewed at wileyonlinelibrary.com]

The problem can also be presented in contingency table format, see Table 2, panel 1. In the rows, we have the combination of variables A and a, with levels $(A,a) = (1, 0), (1, 1), (1, -), (0, x)$, and similarly for the combination of variables C and c in the columns. Thus there is a 4×4 table with a complicated structure. The 3×3 subtable top-left shows the cross-classification for individuals in both A and C. The first two diagonal values are very large, showing that many individuals have the same ethnicity in both the census and MOH. The counts 108,189 and 31,995 refer to the number of individuals for which the information on ethnicity in a contradicts that in c. The counts for $a = -$ show the number of individuals that are in A but whose ethnicity is missing. The aim of our analysis will be to distribute (for example) these 16,512 individuals over the levels $a = 0$ and $a = 1$. It is clear that, even though most of these individuals will be non-Māori ($a = 0$) as they are non-Māori in c, not all of them are non-Māori as for non-Māori $c = 0$ there are still 108,189 individuals that are classified as Māori in a ($a = 1$). Similarly for the counts for $c = -$. The 900 individuals for which the variables a and c are both missing $(a, c) = (-, -)$ will have to be distributed over the four cells $(a, c) = (0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$. When individuals are not in the Census, $A = 0$, then a is automatically missing. In the cross-classification with C and c, the numbers of individuals with $A = 0$ and $a = x$ have to be distributed over $A = 0$, $a = 0$ and $A = 0$, $a = 1$.

The original 15 counts in Table 2, Panel 1, will have to be redistributed over 3 subtables of dimension 2×2 that is, the subtable of size 3×3 has to be reduced to size 2×2, the three values for $A = 0$, $a = x$ have to lead to a subtable of size 2×2 and similarly for the three values for $C = 0$, $c = x$. In a second step, the subtable for $A = 0$, $C = 0$ has to be estimated, and this refers to the individuals that are missed by both lists. Thus two types of missing data are estimated.

Two different mechanisms are used to estimate these two types of missing data that we discuss in a simplified situation. First, ignore the ethnicity variables a and c. Thus, there is the usual DSE problem where the cross-classification of A and C yields three counts in a 2×2 table, and the fourth count, for $(A, C) = (0, 0)$, is to be estimated. The key assumption here is independence of the probabilities for people to be included in A and in C. Thus, we estimate the size of the New Zealand population, but not of the separate ethnicities. Second, ignore the inclusion variables A and C and focus on a and c only. Thus, the levels $-$ and x are merged, and we start off with a 3×3 table. Here the joint distribution of

TABLE 2 Census (A) linked to MOH (C)

		$C = 1$			$C = 0$	Totals
		$c = 0$	$c = 1$	$c = -$	$c = x$	
Panel 1: Observed counts						
$A = 1$	$a = 0$	3,004,335	31,995	150,840	38,634	3,225,804
	$a = 1$	108,189	435,465	12,405	4,368	560,427
	$a = -$	16,512	2769	900	438	20,619
$A = 0$	$a = x$	398,838	146,976	24,636	–	570,450
Totals		3,527,874	617,205	188,781	43,440	4,377,300
		$C = 1$		$C = 0$		Totals
		$c = 0$	$c = 1$	$c = 0$	$c = 1$	
Panel 2: Fitted values under $[Ac][ac][Ca]$						
$A = 1$	$a = 0$	3,170,294.8	33,787.9	38,616.0	411.6	3,243,110.3
	$a = 1$	111,242.5	448,084.8	877.6	3,534.9	563,739.8
$A = 0$	$a = 0$	402,709.4	10,770.8	4,905.2	131.2	418,516.6
	$a = 1$	14,130.7	142,839.1	111.5	1,126.8	158,208.1
Totals		3,698,377.4	635,482.6	44,510.3	5,204.5	4,383,574.8

Ethnicity in A is denoted by a and ethnicity in C is denoted by c, where a and c have levels 0 (non-Māori), 1 (Māori), – (missing) and x (not in list). The data have been randomly rounded to base 3 to protect confidentiality. *Source*: Stats NZ

(a, c) with only values (0, 1) can be estimated under the missing at random (MAR) assumption, where the margins of a (c) are estimated for the individuals only observed in c (a), and the association (odds ratio) between a and c is estimated from the cases where both are observed. This would lead to a 2×2 table of ethnicity in the census and in MOH, but the part of the population missed both by the census and the MOH is ignored.

We now solve the problem of these two types of missing data simultaneously. For the simpler situation that there are no missing data of type – but only of type x, van der Heijden et al. (2018) describe how this result can be found using an EM algorithm with some loglinear model specified by the researcher. Van der Heijden et al. (2018) show that the maximal loglinear model that can be fitted to the data is $[Ac][ac][Ca]$, where the highest order fitted margins are placed between square brackets and all lower-order terms involving the same variables are included. If we denote the levels of A, C, a, c by R, T, r, t, then this model can be denoted by

$$\log \pi_{RTrt} = \lambda + \lambda^{A}_{R} + \lambda^{C}_{T} + \lambda^{a}_{r} + \lambda^{c}_{t} + \lambda^{Ac}_{Rt} + \lambda^{Ca}_{Tr} + \lambda^{ac}_{rt}, \tag{1}$$

with identifying restrictions that the parameters λ, λ^{A}_{1}, λ^{C}_{1}, λ^{a}_{1}, λ^{c}_{1}, λ^{Ac}_{11}, λ^{Ca}_{11} and λ^{ac}_{11} are free, and the other parameters are restricted to be zero. In the maximal model $[Ac][ac][Ca]$, the fitted values are equal to the observed values. However, note that the other two-factor interactions cannot be estimated as there are no data to estimate them; for example, the interaction between A and C cannot be estimated as the joint marginal count for $A = 0$ and $C = 0$ is zero, and similarly for A and a, and for C and c. These are strong assumptions, in particular conditional independence of A and C. However, if model (1) is true, then collapsing over a and c would lead to apparent dependence if all interaction parameters are non-zero, and in this sense model (1) is less demanding than the independence model, where a and c are ignored. Gerritse et al.

(2015) provide the methodology to assess the sensitivity of these estimates. In this paper, we do not pursue this as, by including more than two sources, it will turn out that the assumptions become less demanding.

The EM algorithm can also be applied in this more complicated situation with missing data of type − and x with both types replaced by their expected values in the E-step. The result is given in Table 2, Panel 2. We created our own code (see suppl. materials B and C) instead of the computer program CAT that we used in the past (Schafer et al., 2012; van der Heijden et al., 2012, supplementary materials). Due to the fitted model, in each of the three estimated 2 × 2 subtables the odds ratio between a and c is identical and equal to 377.9, the value from the observed top-left subtable.

The lower right 2 × 2 table in Panel 2 of Table 2 shows the estimated numbers of people missing from both census and MOH. These numbers are relatively low, due to the large overlap of the two lists. Under independence between A and C conditional on a and c, we find, for $a = 0$ and $c = 0$, the estimate 4,905.2 = 38,616.0 × 402,709.4/3,170,294.8, and this low estimate is due to the large denominator (where $A = 1$ and $C = 1$). Although not many individuals are missed by both lists, approximately one fifth of the missed individuals are Māori.

The parameter estimates are given in Table 3. Their standard errors are very small, which is not surprising given the large observed frequencies. Notice that the estimate 4,905.2 in panel 2 of Table 2 can also be obtained as exp(8.50). The estimated conditional odds ratio between the ethnicity variables a and c of 377.9 can be obtained as exp(5.93). Also notice that the relation between A and c is negative, showing that being in the census, A, goes together with a smaller probability of being recorded as Māori in the MOH (estimated odds ratio is $\exp(-0.92) = 0.40$), whereas being in MOH, C, goes together with a larger probability of being recorded as Māori in the census (estimated odds ratio is $\exp(0.43) = 1.54$).

To estimate confidence intervals, we use a procedure that can be considered, conceptually, to be a hybrid between the non-parametric and the parametric bootstrap. The non-parametric bootstrap draws random samples of size n from the frequencies observed in the sample (including in our example the frequencies of − and x, that is from Table 2, Panel 1). This approach takes the uncertainty due to these missing values into account, but not the uncertainty due to the estimated population size. As a consequence, the non-parametric bootstrap tends to underestimate the variance of $\widehat{N}$ (Buckland & Garthwaite, 1991; Anan et al.

2017). The parametric bootstrap draws random samples of size $\widehat{N}$ from the fitted frequencies (in our example, proportions derived from Table 2, Panel 2), and then sets the counts for cells which would not be observed (e.g. the cells for which $A = 0$ and $C = 0$) to zero. This approach, however,

TABLE 3 Parameter estimates for two lists model

	Estimate	Std. Error	z value	Pr
(Intercept)	8.50	NA	NA	NA
A	2.06	0.002	1248.1	<0.001
c	−3.62	0.006	−585.2	<0.001
C	4.41	0.005	865.1	<0.001
a	−3.78	0.016	−233.7	<0.001
A:c	−0.92	0.003	−268.9	<0.001
C:a	0.43	0.016	27.1	<0.001
c:a	5.94	0.007	903.7	<0.001

does not generate observations of missing values for the ethnicity variables a and c, and therefore also underestimates the variance of $\widehat{N}$ because it excludes the imputation of these missing values.

Therefore, we use a new, hybrid bootstrap procedure here that combines these two properties: we use the observed counts in Panel 1 and supplement these with the fitted value of the total of the subtable for $A = 0$, $C = 0$ in Panel 2. These observed counts and the $A = 0$, $C = 0$ fitted value are used to derive multinomial probabilities (using $\widehat{N}$ as the denominator). Bootstrap samples of size $\widehat{N}$ (rounded) are then drawn with these probabilities, after which the counts in the subtable for $A = 0$, $C = 0$ are set to zero, and the new data analysed with model (1). In this application 2000 bootstrap samples are used to derive a confidence interval using the percentile method.

The estimated total population size for New Zealand is 4,383,575; the 95% confidence interval using the percentile method, estimated with the hybrid bootstrap, is 4,383,404–4,383,736. The census and the MOH differ in which part of this total is Māori, as summarised in Table 4, Panel A. The variable measuring ethnicity with the best validity can be used. If there is no clear preference, a practical solution could be to average the two estimates. Other considerations are discussed in Section 3.4 below. When the number of registers involved is larger than two, we propose to use latent class models to bring the separate estimates into agreement with each other, see Section 4.

To conclude this section, we note what would happen if a classic DSE approach is adopted to the counts in Panel 1 of Table 2, that is, by calculating a non-Māori estimate as 3,527,847*3,225,804/3,004,335 and a Māori estimate as 617,205*560,427/435,465. Thus, there are individuals that are part of both estimates (for example, the count of 108,189 in the sum 3,527,874 and in the sum 560,427), and some individuals not at all (such as 24,636). This is a clear drawback of the classic approach.

TABLE 4 Summary of population size estimates of non-Māori and Māori under the selected models according to the ethnicity classifications in different lists, estimated numbers and 95% confidence intervals

	Māori			**Non-Māori**		
	Estimate	**2.5%**	**97.5%**	**Estimate**	**2.5%**	**97.5%**
Panel A: Two lists (Section 4.1)						
Census	721,948	720,499	723,542	3,661,627	3,660,031	3,663,081
MOH	640,687	639,226	642,233	3,742,888	3,741,370	3,744,338
Panel B: Three lists (Section 4.2)						
Census	729,123	727,440	730,822	3,690,122	3,686,382	3,694,110
DIA	771,217	768,867	773,608	3,648,027	3,643,997	3,652,166
MOH	642,724	641,129	644,307	3,776,521	3,772,866	3,780,461
Panel C: Four lists, restricted model (Section 4.3)						
Census	733,294	731,608	734,947	3,689,668	3,687,698	3,691,559
DIA	775,697	772,236	779,109	3,647,265	3,643,429	3,651,285
MOH	645,112	643,533	646,707	3,777,849	3,775,988	3,779,663
MOE	762,222	760,103	764,323	3,660,740	3,658,421	3,662,854
Panel D: Three lists, administrative data sources only (Section 4.4)						
DIA	804,936	800,904	809,185	3,600,293	3,595,293	3,605,915
MOH	641,495	639,936	643,043	3,763,734	3,760,042	3,768,416
MOE	780,234	777,752	782,557	3,624,995	3,620,873	3,630,055

3.2 | Three lists

We now add the DIA register, denoted B, with ethnicity variable b, and consider data derived from three lists. The observed counts are found in the online supplementary material. Adding a third list has the advantage that we no longer have to assume conditional independence between pairs of lists as in the model $[Ac][ac][Ca]$. Now we can fit models that have pairwise dependence between the three lists, which makes the fitted model more realistic. We note that the increased realism of the model lies in the ability to model associations between the inclusions in the lists, and not from changing the MAR assumption for solving missingness in ethnicity variables, which remains basically the same. The cells of the contingency table with the observed frequencies are denoted with $A, B, C \in \{0, 1\}$ and $a, b, c \in \{0, 1, -, \text{x}\}$, and the cells of the contingency table with the fitted frequencies with $A, B, C, a, b, c \in \{0, 1\}$, since the missing levels $-$ and x have been imputed.

As there are six variables, the potential number of fitted frequencies is $2^6 = 64$. However, there are eight structurally zero cells for the combination of $A = 0$, $B = 0$ and $C = 0$. Also, as in the situation for two lists, it is not possible to fit models where A interacts with a, B with b or C with c. Thus, the maximal model is $[ABc][ACb][BCa][Abc][Bac][Cab][abc]$. This model has 1 (intercept) + 6 (main effects) + 12 (15 two-factor interactions minus the three interactions that cannot be fitted) + 7 (three factor interactions) = 26 parameters. The left graph in Figure 2 shows a graphical model, where interactions between pairs of variables are shown by lines. We describe and use graphical models further in Section 3.3. For some age categories, the probabilities of inclusion in DIA (B) are zero, which means there is an interaction between B and age—the right graph in Figure 2 shows the resulting graphical model. Under certain conditions, which we explain more generally in Section 3.3, the partial coverage does not interfere with the multiple system estimation, and we can ignore in the estimation that DIA is a source not covering the whole population—that is, we do not need explicitly to include age in the fitted model. In this case, we implicitly assume that coverage in *at least one* of the other sources is homogeneous with respect to age, and this is clearly reasonable for MOH. Although the census has some differential response by age, this is much less extreme than the partial population coverage of DIA, and we consider it reasonable to treat this source as homogeneous too. However, even if there is an important dependence of census on age, we can make valid estimates as long as the interaction between Census and MOH is also included in the model.

The number of unique individuals in the three linked lists is 4,378,377, the estimated number not in any list is 40,868, and this leads to an estimated population size of 4,419,245 (4,415,848–4,422,929). The estimated number of non-Māori and Māori in the Census, DIA and MOH can be found in Table 4,

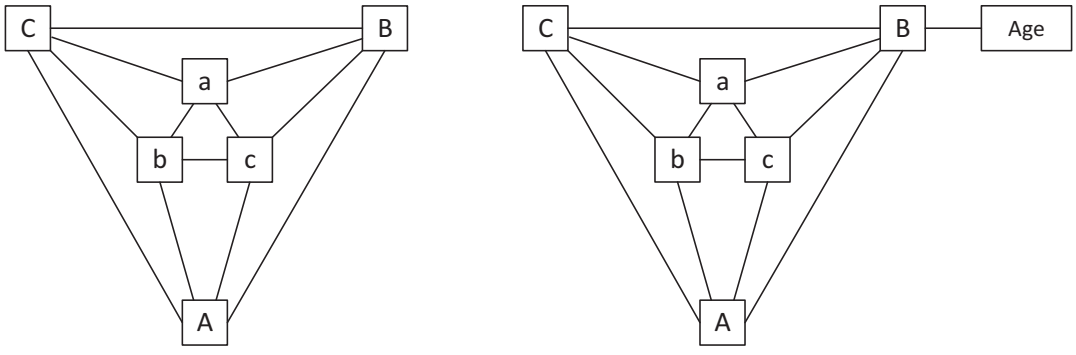

FIGURE 2 Interaction graphs for loglinear model $[ABc][ACb][BCa][Abc][Bac][Cab][abc]$ on the left, extended with a covariate Age on the right

Panel B. An intuitive interpretation of these estimates for DIA is that they are the sizes of the Māori and non-Māori populations under the definition of Māori ethnicity operating in the DIA list. These estimates are valid if the model is true, including that the associations of (ethnicity in) DIA with (ethnicity in) the other registers are the same in the age groups excluded from DIA as in the included age groups. These associations are estimated from the included age groups.

Parameter estimates for the maximal model are found in Table 5. With $n = 4{,}378{,}377$ it is not surprising that it is hard to find more parsimonious models that fit the data, except for the model where the three-factor interaction between C, a and b is set to zero. Using the methodology that we propose it is possible to fit more parsimonious models but in this particular instance, where interpretability of the model is not the main focus, there is not much to gain. For three lists, the maximal model allows for pairwise dependence of lists, whereas with two lists conditional independence had to be assumed. For example, for the pairwise dependence of list A and B we use the parameter estimates for A:B and A:B:c. Then for non-Māori ($c = 0$) the estimated conditional odds ratio between A and B is $\exp(0.332) = 1.4$, and for Māori it is $\exp(0.332-0.160) = 1.2$, so inclusion in A goes together with

TABLE 5 Parameter estimates for three lists model

	Estimate	Std. Error	*z* value	Pr
(Intercept)	10.487	NA	NA	NA
A	0.029	0.050	0.6	0.5592
B	−4.001	0.033	−122.4	<0.001
c	−4.866	0.025	−198.5	<0.001
C	2.254	0.050	45.1	<0.001
b	−4.303	0.183	−23.5	<0.001
a	−5.756	0.201	−28.7	<0.001
A:B	0.332	0.005	67.9	<0.001
A:c	−0.750	0.024	−31.7	<0.001
B:c	1.080	0.014	79.4	<0.001
A:C	2.004	0.050	40.1	<0.001
A:b	0.426	0.086	4.9	<0.001
C:b	0.661	0.182	3.6	<0.001
B:C	2.030	0.032	62.7	<0.001
B:a	1.700	0.059	28.8	<0.001
C:a	0.702	0.201	3.5	<0.001
c:b	4.156	0.032	128.9	<0.001
c:a	5.075	0.022	232.1	<0.001
b:a	5.472	0.275	19.9	<0.001
A:B:c	−0.160	0.008	−19.7	<0.001
A:C:b	−0.857	0.086	−10.0	<0.001
B:C:a	−0.568	0.059	−9.6	<0.001
A:c:b	0.232	0.029	8.1	<0.001
B:c:a	−1.129	0.014	−81.7	<0.001
C:b:a	−0.249	0.275	−0.9	0.3664
c:b:a	−2.136	0.029	−73.0	<0.001

inclusion in B. These relations are much stronger for A and C, and for B and C. For example, for b=0 the estimated conditional odds ratio between A and C is $\exp(2.004) = 7.4$, and for $b = 1$ the estimated conditional odds ratio between A and C is $\exp(2.004-0.857) = 3.2$. Not surprisingly, there are strong relations between the ethnicity variables a, b and c. For parameters a:b, a:c and b:c the estimates are 5.472, 5.075 and 4.156. Interestingly, the estimated three factor interaction parameter a:b:c is negative. This means that, for example, for the non-Māori in c the estimated odds-ratio between ethnicity measures in a and b is $\exp(5.472) = 237.9$, whereas for Māori in c the estimated odds-ratio between the ethnicity measures in a and b is only $\exp(5.472-2.136) = 28.1$. Thus, the negative three-factor interaction estimate a:b:c shows that for non-Māori in one of the three lists the disagreement in the other two lists is smaller than for Māori (i.e. Māori are more likely to be inconsistent between the lists than non-Māori). Table 6 illustrates this for counts marginalised over the lists A, B and C: for example, for $c = 0$ the disagreement is 64,566 + 30,087 (relative to approximately 3,750,000), but for $c = 1$ the disagreement is 26,063 + 18,448 (relative to approximately 625,000).

3.3 | Using registers that cover different parts of the population

Usually, expositions of DSE assume that a person in the population of interest has a chance to be included in each list; in fact, the inclusion probabilities must be assumed to be homogeneous in at least one list, so as a corollary, one list can have partial coverage. Here, the Census and MOH aim to cover the complete population but DIA and MOE aim to cover only subpopulations, so we need to examine what this means for multiple system estimation. Building on earlier work of Zwane et al. (2004) and van der Heijden et al. (2012) we show that we can ignore the fact that DIA and MOE only cover subpopulations when estimating the size of the population. We also show under what conditions this is true in general. An alternative modelling approach is presented by Sutherland et al. (2007), and related work is found in di Cecco et al. (2018), van der Heijden and Smith (2020), and di Cecco et al. (2020).

3.3.1 | Partial coverage and collapsibility, two lists

We first introduce the concept of collapsibility in loglinear models for population size estimation described by van der Heijden et al. (2012). Figure 3, taken from their paper, shows four models for DSE, represented by their interaction graphs. 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 M_0 there is no covariate (this is the classical dual system estimator where the interaction is assumed to be zero). In model M_1 there is a covariate X_1 that is related to list A but not to list B, meaning that inclusion probabilities are heterogeneous across the levels of X_1 for A but homogeneous for B. In model M_2 it is the other way around. In model M_3 inclusion probabilities for both A and B are heterogeneous across the levels of X_1. Van der Heijden et al. show that the total population size estimate is identical in models M_0, M_1 and M_2 and

TABLE 6 Contingency table of fitted frequencies for the marginal table of a, b and c

	$c = 0$		$c = 1$	
	$b = 0$	$b = 1$	$b = 0$	$b = 1$
$a = 0$	3,581,229	64,566	18,264	26,063
$a = 1$	30,087	100,639	18,448	579,949

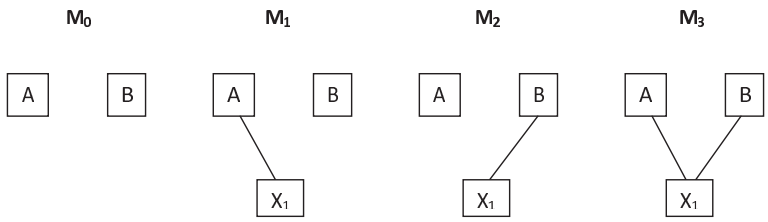

FIGURE 3 Interaction graphs for loglinear models with two lists and one covariate, taken from van der Heijden et al. (2012)

different from the estimate in M_3. Therefore, under models M_1 and M_2 the three-way array of variables A, B and X_1 can be added over the values of X_1, 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 M_3 is not collapsible over X_1, because summing over X_1 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 M_3 the covariate lies on a short path between A and B and therefore one cannot collapse over X_1. In M_1 and M_2 the covariate X_1 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), characterised by two examples, in terms of collapsibility. In Example 1 there are two regions: list A covers the full population, but list B covers only the north. Here, region can be considered as the covariate X_1. If we assume that the inclusion probabilities for A are homogeneous over region, there is no edge from X_1 to A. Yet the inclusion probabilities for B are heterogeneous, as they are positive for the north and zero for the south region, and there is an edge from X_1 to B. This is model M_2 in Figure 3. 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 has three regions, with A covering the north and central regions, and B covering the central and south regions. This situation is described by model M_3: 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 X_1 to A and another from X_1 to B, and therefore X_1*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). For Example 2, Zwane et al. (2004) propose to use missing data methodology to estimate the south observations for register A and the north observations for register B under the assumption that the relation between register A and B found in the middle also holds for the south and the north.

3.3.2 | Extension to more than two lists

In Section 3.2, we have a model for the Census and MOH, that both 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 as on the right of Figure 2 (if *a*, *b* and *c* are ignored). Age is not on a short path, so the model can be collapsed over it; that is, we can treat DIA (*C*) as if it covered the whole population, without affecting the estimated population size. When we add register MOE (in Section 3.4 below), 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.

3.3.3 | General result on the ignorability of partial coverage

We can apply the approach of this section to give a general result. For any number of lists, and any combination of full and partial coverages of the population, the partial coverage will be ignorable (that is, we can obtain the correct results from modelling without taking any special account of the population coverages) if the model is collapsible over the covariate(s) which define(s) the partial coverage(s). Equivalently, the variable(s) defining the coverage(s) do not appear on any short path in the graphical representation of the model.

This means that when a covariate defines the partial coverage for a list that does not cover the full population, we need to assess whether it is on a short path in the model we would like to fit. If it is not on a short path, we may proceed to fit the model to the contingency table formed by the lists, ignoring the covariate, and we will obtain the correct estimated population size (we define this situation as **ignorable partial coverage**). The covariate defining the partial coverage does not need to be included in the model. This latter property is particularly useful in our example where neither age nor the other contextual variables predicting entry to education are available in our data set.

3.4 | Four lists

We now add the fourth list, MOE, denoted *D*, with ethnicity in MOE denoted as *d*, to the analysis. Now the maximal model is [*ABCd*][*ABDc*][*ACDb*] [*BCDa*][*ABcd*][*ACbd*][*ADbc*][*BCad*][*BDac*] [*CDab*][*Abcd*][*Bacd*][*Cabd*][*Dabc*] [*abcd*]. Due to the appearance of four-factor interactions in the model, the assumptions become less and less demanding as more lists are involved. In this model, both DIA and MOE have heterogeneous inclusion probabilities over Age. Applying the reasoning of Section 3.3, Age is not on a short path and we can collapse over it. We conclude that the fact that DIA and MOE do not aim to cover the full population, does not threaten the validity of our estimates.

This model turns out to have some numerical instability in the sense that some log-linear parameters are (nearly) on the boundary or non-identified. We restricted 15 parameters to zero leading to the model [*ABcd*][*AC*][*ADbc*][*BCad*] [*BDac*][*CDa*][*CDb*][*Abcd*][*Bacd*][*Dabc*][*abcd*]. This stabilises the estimates. The deviance of the model is 680.6, which is, with an observed population size of over 4 million, negligible. The population size estimates for the model are in Table 4, Panel C. The loglinear parameter estimates for the maximal model and for the restricted version are in the supplementary materials. The number of unique individuals in the four linked lists is 4,401,990, and the estimated number missed by all lists is 20,972, giving an estimated population size of 4,422,962

(4,421,894–4,424,080). In comparison to the three list solution, by including the fourth list the estimated population size increases by only 4000, mostly Māori.

3.5 | Three registers, ignoring the census

We also present estimates derived only from the three administrative registers, so that we can see what would happen if the census were replaced entirely by an administrative data-based system. The observed number of individuals in at least one of the registers is 4,378,716. We estimate an additional 26,513 individuals missed by all three registers. This leads to a total population size of 4,405,229 (4,401,858–4,409,667). This is somewhat less than the four list estimate that was 4,422,962.

The estimate of 4,405,229 broken down by ethnicity for each of the three lists is presented in Table 4, Panel D. The three-list Māori estimate using the DIA ethnicity concept is larger than the four-list estimate in Table 4, Panel C: 804,936 versus 775,697; for the MOH concept it is very similar and for the MOE concept the three-list estimate is larger as well: 780,234 versus 762,222. This suggests that in the absence of the census, the estimate of the size of the Māori population would be larger. Based on the estimates of measurement error from the latent variable analysis in Section 4, where the census was the most accurate, this suggests an overestimation of the Māori population size.

4 | DEALING WITH MEASUREMENT ERROR IN MĀORI VARIABLES

To arrive at a final estimate of the number of non-Māori and Māori, we describe two approaches, both using the concept of measurement error. Measurement error is taken into account by a latent class model, and in the first, two-step approach we fit the maximal model for four registers and then fit a latent class model on the marginal ethnicity estimates from this model. In the second approach, we fit a one-step overall model.

Deriving a final estimate is related to the field of macro-integration of categorical outcomes, see de Waal et al. (2019, 2020) for a review of the field. Consider the joint margins of the ethnicity variables a, b, c and d of the four lists in Table 7 from the restricted maximal model of Section 3.4. An ad hoc approach is to consider five groups of individuals, according to how many lists record them as Māori. Then a (slightly arbitrary) choice would have to be made which categories would be used to arrive at the estimated numbers of Māori and non-Māori. For example, one could consider individuals who are estimated to be Māori in at least two lists to be Māori, allowing for a measurement error in recording

TABLE 7 Joint margins for variables a (Census), b (DIA), c (MOH) and d (MOE) estimated under the restricted maximal model

		$c = 0$		$c = 1$	
		$d = 0$	$d = 1$	$d = 0$	$d = 1$
$a = 0$	$b = 0$	3,519,852	53,366	10,998	6934
	$b = 1$	55,676	15,600	9686	17,555
$a = 1$	$b = 0$	14,590	21,218	2560	17,747
	$b = 1$	18,443	79,105	28,934	550,697

a non-Māori as a Māori in at most one list. This would give a final estimate of 768,479, corresponding to an estimated probability of 768,479/4,422,962 = 0.174.

A statistical approach to measurement error is to make use of a latent class model (McCutcheon, 1987), a technique that has also been applied to evaluate census-related multiple system estimation by Biemer et al. (2001), see for comparable work Biggeri et al. (1999), Stanghellini and van der Heijden (2004) and di Cecco et al. (2018). See also Boeschoten et al. (2019), for a comparison of the latent class model with the ad hoc approach described in the preceding paragraph. Other recent work making use of the latent class model in official statistics is Boeschoten et al. (2017, 2019). Oberski (2016, 2018) makes a plea for the use of latent class modelling in the context of linked data sources to tackle the measurement error problem.

The latent class model assumes the existence of a categorical latent variable, and that the observed variables are independent conditional on the latent variable. Thus, the latent variable 'causes' the responses to the observed variables, and explains the interactions between the observed variables. Let π_{abcd} be the joint probability for variables a, b, c and d, with levels indexed by r, s, t and u respectively, with 0 for non-Māori and 1 for Māori. Let X be the latent variable, with two levels indexed by x ($x = 1, 2$). Let π_x^X be the probability to fall in latent class x. Let $\pi^a_{r|x}$ be the conditional probability for variable a to fall in $r = 0, 1$ given latent class x. Then the two-class latent class model is

$$\pi_{rstux} = \pi_x^X \pi^a_{r|x} \pi^b_{s|x} \pi^c_{t|x} \pi^d_{u|x} \tag{2}$$

with

$$\pi_{rstu} = \Sigma_{x=1,2} \pi_{rstux}. \tag{3}$$

As there are 15 probabilities in Table 7, and nine parameters in Equation (2) (i.e. eight independent conditional probabilities and one independent class size parameter), there are 6 degrees of freedom.

We fit the latent class model with two latent classes to the data in Table 7. This gives a two-stage procedure—first fitting a model to deal with the missing data (both − and x), and then applying a second, latent class, model to the resulting estimates. We find the estimates in Panel 1 of Table 8. In this latent class model, the first latent class is to be interpreted as the class for non-Māori, and the estimated probability of falling in this class is 0.827. The estimated probability for the Māori class, 0.173, corresponds to an estimated Māori population size of 770,677. Estimated conditional probabilities of being Māori for each latent class are also shown in Panel 1 of Table 8; they are consistently low for the non-Māori latent class and high for the Māori latent class. These estimated conditional probabilities can be interpreted as measurement error estimates under the assumption that the model is correct. For

TABLE 8 Estimates of latent class models with two latent classes

		Census	DIA	MOH	MOE
	π_x	$\pi^a_{r=1\,\vert\, x}$	$\pi^b_{s=1\,\vert\, x}$	$\pi^c_{t=1\,\vert\, x}$	$\pi^d_{u=1\,\vert\, x}$
Panel 1: Estimates for Table 7					
Class 1	0.827	0.004	0.016	0.003	0.015
Class 2	0.173	0.937	0.937	0.826	0.922
Panel 2: Estimates for LCMSE					
Class 1	0.834	0.007	0.014	0.005	0.016
Class 2	0.166	0.957	0.958	0.847	0.959

example, with $1-0.937 = 0.063$ the census ~~has~~ and DIA have the smallest measurement error for Māori: given that the true status of someone is Māori, he or she has an estimated probability of 0.063 to say he or she is non-Māori in the census or DIA. For the non-Māori the MOH has the smallest measurement error, with an estimate of 0.003, closely followed by the census. The measurement errors are much larger in the latent class of the Māori than in the class of the non-Māori. The population size of the Māori latent class of 770,677 may be considered high when we compare it with the estimates for the Census, DIA, MOH and MOE of 733,167, 761,545, 643,429 and 770,047, respectively. However, we have to take into account that in this Māori latent class individuals have relatively large measurement error, making them report regularly that they are non-Māori. The reverse measurement error, that individuals in the non-Māori latent class report that they are Māori, is much smaller. We note that a three-class latent class model is not identified if the number of observed variables is four, even though the number of cells is larger than the number of parameters (Vermunt & Magidson, 2004).

The parameters in Equation (2) can also be used to derive posterior probabilities $\pi_{x|rstu} = \pi_{rstux}/\pi_x$. The posterior probabilities turn out to be very close to either 1 or 0; the probability of being in the non-Māori class is close to 1 when Māori ethnicity is reported in at most one of the lists, and close to 0 when Māori is reported in two or more lists. When we assign individuals to either the non-Māori or the Māori class using these probabilities rounded to 0 or 1, the sum of the number of Māori is 768,479, very close to the estimate of 770,677 reported above (see Supplementary materials D for details).

We have just fitted a latent class model with two classes on the margins for variables a, b, c and d. It is also possible to define overall models. For this purpose, we first rewrite the latent class model with two latent classes as a loglinear model for the observed variables a, b, c and d and the latent variable X (Hagenaars, 1993). The latent class model can be denoted as the loglinear model $[aX][bX][cX][dX]$ for the probabilities π_{rstux}. Written as a loglinear model using λ-parameters, we have

$$\log\pi_{rstux} = \lambda + \lambda^a_r + \lambda^b_s + \lambda^c_t + \lambda^d_u + \lambda^X_x + \lambda^{aX}_{rx} + \lambda^{bX}_{sx} + \lambda^{cX}_{tx} + \lambda^{dX}_{ux}. \quad (4)$$

We incorporate this loglinear model in the maximal model for variables A, B, C, D, a, b, c and d in the following way. To start with, we define probabilities $\pi_{rstuRSTUx}$ that allow us to define a model for the eight observed variables and the latent variable. These probabilities that include a latent variable are related to the probabilities for the observed variables only by

$$\pi_{rstuRSTU} = \Sigma_x \pi_{rstuRSTUx}. \quad (5)$$

In the section on four lists we saw that the maximal model is $[ABCd]$ $[ABDc][ACDb][BCDa]$ $[ABcd][ACbd][ADbc][BCad][BDac][CDab][Abcd]$ $[Bacd][Cabd][Dabc][abcd]$. The restrictive features of this maximal model are (1) that variables denoted by a lower case letter cannot appear in the same term as variables denoted by capitals, so that, for example, variables a and A cannot appear in the same interaction term, and (2) the four capitals A, B, C and D cannot appear together in the same term. We are particularly interested in the joint marginal counts of a, b, c and d.

We now modify the maximal model to include a latent variable. In the maximal model, there are three groups of terms. First, there is the last term $[abcd]$, that we replace by the latent class model $[aX][bX][cX][dX]$. Thus, we are formulating a model for observed variables and the latent variable. Second, we eliminate the terms $[ABcd]$, $[ACbd]$, $[ADbc]$, $[BCad]$, $[BDac]$, $[CDab]$, $[Abcd]$, $[Bacd]$, $[Cabd]$ and $[Dabc]$ as the joint appearance of two or three of the lower case letters for variables a, b, c and d in one single term is in conflict with the latent class assumption. For example, consider $[ABcd]$. The interaction between c and d is explained by the latent variable X. One could argue that, although it is in conflict with the latent class model to include the c, d-interaction, one could

still include the interactions for A, c, d, for B, c, d and for A, B, c, d. However, this would result in a non-hierarchical loglinear model and these are difficult to interpret, so we eliminate these terms completely. Third, we can retain the terms [$ABCd$], [$ABDc$], [$ACDb$], [$BCDa$]. Thus, a latent class model where Equation (4) is extended with the list variables A, B, C and D, is [$ABCd$][$ABDc$] [$ACDb$][$BCDa$][aX][bX][cX][dX]. We refer to this model as LCMSE, short for latent class multiple system estimation.

Our code for the LCMSE model can be found in the supplementary materials B and C. The LCMSE model was also fitted using lEM (Vermunt, 1997); however, lEM does not provide an easy way to calculate estimates of the missing part of the population.

The latent class parameter estimates can be found in Panel 2 of Table 8. They are similar to those in Panel 1. The deviance of the model is 10,922.25. A proper evaluation of this value should take the observed population size of 4,401,990 into account, and therefore we divide the deviance by 4,401.99, so that we have an impression of the deviance if the population size had been 1000. Thus, we find a normed deviance of 2.5, showing a good fit. The population size estimate for this model population is 4,447,071, with a 95% confidence interval of 4,435,301–4,465,050. The estimated proportion of Maori in the LCMSE is 0.166, lower than in the two-stage model, and the non-Māori proportion is correspondingly higher. It is as if the LCMSE approach is more likely to treat individuals with uncertain status as part of the non-Māori latent class. This reduces the estimated measurement errors in the Māori latent class, and increases them in the non-Māori. We prefer the integrated approach of the LCMSE because it deals with all the variability in the data appropriately, whereas the two-step approach means that the latent class estimates depend on the fitted values of the first-stage model. On the other hand, the price to pay for this is that the LCMSE model involves an apparently substantial change in the formulation of the model.

As a second model, we adjust the LCMSE model by including a latent variable Y that explains the relations between the variables A, B, C and D. Theoretically, such a model allows for heterogeneity of inclusion probabilities, where in one class inclusion probabilities for the four lists are higher and in the other class they are lower, see Stanghellini and van der Heijden (2004) for a medical example. So if there were two such subpopulations, the latent class model would reveal this. A model for this is [AY] [BY][CY][DY][aX][bX][cX][dX] (now excluding all of the original interactions as these are explained by the latent class variables). This model has a deviance of 291,464.1, and a normed deviance of 66.2. Thus, we reject this model. We also considered extending the model with an interaction between X and Y, giving [AY][BY][CY][DY][aX][bX][cX][dX][XY]. The fit improves but is still not good: the normed deviance is 43.7.

We conclude that the LCMSE model is the best model for these data. Interestingly, the estimated probability of the Māori latent class of 0.166 in this model is close to the estimated probability of Māori of 16.5 for the IDI-ERP found by Statistics New Zealand. Also, the tables analysed are for Census day 5th March 2013. The official ERP figure for March 2013 is 4,436,000 +/− about 0.5% (https://www.stats.govt.nz/topics/population). This is very close to our four list estimate of 4,422,962 and the LCMSE estimate of 4,447,071.

5 | DISCUSSION

Van der Heijden et al. (2018) presented an approach for estimating the margins of auxiliary variables in the DSE framework. They suggested that more experience with applications of this methodology was needed to be able to judge its usefulness. Here this approach is extended to multiple system estimation with four lists, and a more complicated missing data structure.

The data used here were probabilistically linked in Stats New Zealand's IDI, which means that the linkage has not been subjected to a clerical review stage. This increases the risk that there will be some linkage error which is not accounted for in our models, and since dual and multiple system estimation using loglinear models is critically dependent on the lack of linkage error (Biemer & Stokes, 2004; Gerritse et al., 2017; Wolter, 1986), there is a risk that estimates will be inflated through matching error. The data have also been randomly rounded to base 3 to protect confidentiality, but we expect this to have a negligible impact on the fitted values as the number of individuals in the linked data set is larger than 4 million.

There is also a risk of overcoverage (also known as list inflation) in administrative registers, through duplicate records and the inclusion of people who are not part of the target population, which is the usual resident population in the case of the census as considered here. It is likely that some duplicates remain in the IDI-ERP despite attempts to remove as many of them as possible. There is as yet no estimate of the remaining overcoverage in the IDI-ERP. However, among the sources considered here the DIA birth register is less likely to suffer from duplicates because it does not obtain repeating data—although for the same reason it may suffer from overcoverage when people emigrate. However, the sequence of population size estimates from models with two to four lists shows only small increases, suggesting that the overcoverage is small in the added registers.

We assume that the inclusion probability is homogeneous in at least some of the lists (see Section 3.2). This assumption seems most at risk for the census source with respect to age, though as we discussed previously this does not invalidate our analysis because we have homogeneity in other sources; at worst it may restrict our choice of models to ensure that age does not appear on a short path between registers. However, even in the case that age was important in *all* the lists, it would still be possible to use missing data methodology to estimate the unobserved parts of the lists under the assumption that the relations in parts of the population where the lists overlap also hold in the other parts - a missing at random assumption (Zwane et al., 2004).

The modelling process to choose the most parsimonious loglinear model to use in population size estimation and marginal total estimation for auxiliary variables such as Māori/non-Māori is complicated in our application by the size of the data. For most of the models except for the model for four lists, since the census and the MOH register cover most of the population of New Zealand, the entries in the contingency tables (for example Table 2, Panel 1) are very large. Therefore, any term added to the model has plenty of data for estimation, and will almost certainly be significant. Therefore, we have ended up with saturated models, except for the model for four sources where the saturated model is numerically unstable, and the latent class model, that is not saturated by definition. The estimated sizes of the Māori population are plausible, but a different value is created according to the definition in each list (in this sense the Māori identifiers act as alternative variables, as in van der Heijden et al., 2018, Section 4). A preferred version of the variable, or some pragmatic combination of estimates, must be chosen in order to obtain a final estimate.

The latent class analysis of the same data deals with these different definitions, and provides a consolidated single estimate, and, as a consequence, also estimates of the measurement errors in each source. In the two models for which estimates are presented, the measurement errors for Māori status in the census are among the lowest (Table 8, Panel 1 and 2), which provides some support for the intuition in the statistical system that the purpose-designed census collection gives a better estimate of the size of the Māori population than one based on administrative registers which only collect this variable as a by-product of their main purpose. The preferred LCMSE model (Table 8, Panel 2) shows almost the same estimated measurement error for the Māori in census, DIA and MOE, with MOH showing a substantially greater measurement error. MOH has the lowest estimated measurement error for non-Māori, but only a little better than the Census (under either latent class approach).

Despite these differences, the estimate using only the three administrative registers without the input of the census is reasonable, although there are some differences from methods based on the census.

We conclude that the methods of van der Heijden et al. (2018) provide stable results that allow for detailed interpretation of the processes of inclusion in the lists considered, and of recording Māori status. The conditions for treating the partial coverage of the lists as ignorable are met in this study, so that the modelling can be applied directly, and we show that they can be extended to deal with different forms of missing data. The latent class approach provides a principled method to produce a common estimate accounting for differences in the definition of Māori status among the data sources, and also provides estimates of the measurement error in the different definitions which can be used to understand the quality of the administrative sources.

DISCLAIMER

The results in this paper are not official statistics, they have been created for research purposes from the Integrated Data Infrastructure (IDI) managed by Stats New Zealand. The opinions, findings, recommendations and conclusions expressed in this paper are those of the authors, not Stats New Zealand. Access to the anonymised data used in this study was provided by Stats New Zealand in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation and the results in this paper have been confidentialised to protect these groups from identification. Careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the Privacy impact assessment for the Integrated Data Infrastructure (Statistics New Zealand, 2017b).

ORCID

Peter G. M. van der Heijden http://orcid.org/0000-0002-3345-096X
Paul A. Smith http://orcid.org/0000-0001-5337-2746

REFERENCES

Anan, O., Böhning, D. & Maruotti, A. (2017) Uncertainty estimation in heterogeneous capture–recapture count data. *Journal of Statistical Computation and Simulation*, 87, 2094–2114. https://doi.org/10.1080/00949655.2017.1315668

Bakker, B., van der Heijden, P. & Scholtus, S. (2015) Preface to special issue on coverage problems in administrative sources. *Journal of Official Statistics*, 31, 349–355.

Biemer, P.P. & Stokes, S.L. (2004) Approaches to the modeling of measurement error. In: Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. & Sudman, S. (Eds.) *Measurement error in surveys*. New York: Wiley.

Biemer, P., Woltman, H., Raglin, D. & Hill, J. (2001) Enumeration accuracy in a population census: an evaluation using latent class analysis. *Journal of Official Statistics*, 17, 129–148.

Biggeri, A., Stanghellini, E., Merletti, F., & Marchi, M. (1999) Latent class models for varying catchability and correlation among sources in capture–recapture estimation of the size of a human population. *Statistica Applicata*, 11, 563–576.

Boeschoten, L., Oberski, D. & de Waal, T. (2017) Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (MILC). *Journal of Official Statistics*, 33, 921–962.

Boeschoten, L., de Waal, T. & Vermunt, J.K. (2019) Estimating the number of serious road injuries per vehicle type in the Netherlands by using multiple imputation of latent classes. *Journal of the Royal Statistal Society, Series A*, 182, 1463–1486.

Brown, J., Diamond, I., Chambers, R., Buckner, L. & Teague, A. (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. https://doi.org/10.1111/1467-985X.00133

Brown, J., Sexton, C., Abbott, O. & Smith, P. A. (2019) The framework for estimating coverage in the 2011 Census of England and Wales: combining dual-system estimation with ratio estimation. *Statistical Journal of the IAOS*, 35, 481–499. https://doi.org/10.3233/SJI-180426

Buckland, S. & Garthwaite, P. (1991) Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics*, 47, 255–268. https://doi.org/10.2307/2532510

di Cecco, D., di Zio, M., Filipponi, D. & Rocchetti, I. (2018) Population size estimation using multiple incomplete lists with overcoverage. *Journal of Official Statistics*, 34(2), 557–572.

di Cecco, D., di Zio, M., & Liseo, B. (2020) Bayesian latent class models for capture–recapture in the presence of missing data. *Biometrical Journal*, 62, 1–13.

Gerritse, S.C., van der Heijden, P.G.M. & Bakker, B.F.M. (2015) Sensitivity of population size estimation for violating parametric assumptions in loglinear models. *Journal of Official Statistics*, 31, 357–379. https://doi.org/10.1515/jos-2015-0022

Gerritse, S.C., Bakker, B.F.M. & van der Heijden, P.G.M. (2017) *The impact of linkage errors and erroneous captures on the population size estimator due to implied coverage*. Den Haag: Statistics Netherlands, discussion paper 2017–16.

Hagenaars, J.A. (1993) *Loglinear models with latent variables*. Newbury Park: Sage.

Hand, D.J. (2018) Statistical challenges of administrative and transaction data (with discussion). *Journal of the Royal Statistical Society, Series A*, 181, 555–605. https://doi.org/10.1111/rssa.12315

Madden, R., Coleman, C., Mashford-Pringle, A. & Connolly, M. (2019) Indigenous identification: past, present and a possible future. *Statistical Journal of the IAOS*, 35, 23–27.

McCutcheon, A. (1987) *Latent class analysis*. Newby Park: Sage.

Oberski, D. (2016) Estimating error rates in an administrative register and survey questions using a latent class model. In: Biemer, P.P. et al. (Eds.), *Total survey error in practice: improving quality in the era of big data*. New York: Wiley, pp. 633–670.

Oberski, D. (2018) A research programme for dealing with most administrative data challenges: data linkage and latent variable modelling—discussion on 'statistical challenges of administrative and transaction data' by David J Hand. *Journal of the Royal Statistical Society, Series A*, 181, 555–605.

Reid, G., Bycroft, C. & Gleisner, F. (2016) *Comparison of ethnicity information in administrative data and the census*. Christchurch: Statistics New Zealand.Retrieved from https://www.stats.govt.nz/assets/Research/Comparison-of-ethnicity-information-in-administrative-data-and-the-census/comparison-of-ethnicity-information-in-administrative-data-and-the-census.pdf

Schafer, J.L., Harding, T. & Tusell, F. (2012) Package 'cat'. Retrieved from https://CRAN.R-project.org/package=cat

Simpson, L., Jivraj, S. & Warren, J. (2016) The stability of ethnic identity in England and Wales 2001–2011. *Journal of the Royal Statistical Society: Series A*, 179, 1025–1049.

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(2), 510–516. https://doi.org/10.1111/j.0006-341X.2004.00197.x

Statistics New Zealand. (2012) Transforming the new zealand census of population and dwellings: issues, options, and strategy. Christchurch, Statistics New Zealand. Retrieved from https://www.stats.govt.nz

Statistics New Zealand. (2014) An overview of progress on the potential use of administrative data for census information in New Zealand: census transformation programme. Christchurch, Statistics New Zealand. Retrieved from https://www.stats.govt.nz

Statistics New Zealand. (2017a) Experimental population estimates from linked administrative data: 2017 release. Christchurch, Statistics New Zealand. Retrieved from https://www.stats.govt.nz

Statistics New Zealand. (2017b) Integrated data infrastructure: overarching privacy impact assessment. Christchurch, Statistics New Zealand. Retrieved from https://www.stats.govt.nz/assets/Uploads/Integrated-data-infrastructure/idi-overarching-pia.pdf

Statistics New Zealand. (2018) Experimental ethnic population estimates from linked administrative data. Christchurch, Statistics New Zealand. Retrieved from https://www.stats.govt.nz

Sutherland, J., Schwarz, C. & Rivest, L.-P. (2007) Multilist population estimation with incomplete and partial stratification. *Biometrics*, 63, 910–916. https://doi.org/10.1111/j.1541-0420.2007.00767.x

Van der Heijden, P.G.M. & Smith, P.A. (2020) On estimating the size of overcoverage with the latent class model. A critique of the paper "population size estimation using multiple incomplete lists with overcoverage" by di Cecco, di Zio, Filipponi and Rocchetti (2018, JOS 34 557–572). (arXiv:2005.05452v1)

Van der Heijden, P.G.M., Whittaker, J., Cruyff, M., Bakker, B. & Van der Vliet, R. (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, 831–852. https://doi.org/10.1214/12-AOAS536

Van der Heijden, P.G.M., Smith, P., 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.

Vermunt, J.K. (1997). *LEM 1.0: A general program for the analysis of categorical data*. Tilburg: University, Tilburg. https://jeroenvermunt.nl/#Software.

Vermunt, J.K. & Magidson, J. (2004) Latent class analysis. In: Lewis-Beck, M.S., Bryman, A.A. & Liao, T. (Eds.) *The sage encyclopedia of social sciences research methods*. Thousand Oaks, CA: Sage Publications, pp. 549–553.

de Waal, T., van Delden, A. & Scholtus, S. (2019) Quality measures for multisource statistics. *Statistical Journal of the IAOS*, 35, 179–192.

de Waal, T., van Delden, A. & Scholtus, S. (2020) Multi-source statistics: basic situations and methods. *International Statistical Review*, 88, 203–228.

Waldon, J. (2019) Identification of indigenous people in a otearoa-New Zealand–Ngāmata o taku Whenua. *Statistical Journal of the IAOS*, 35, 107–118.

Wolter, K.M. (1986) Some coverage error models for census data. *Journal of the American Statistical Association*, 81(394), 337–346. https://doi.org/10.1080/01621459.1986.10478277

Zhang, L.-C. & Chambers, R.L. (2019) *Analysis of integrated data*. Boca Raton: CRC Press.

Zhang, L.-C. & Dunne, J. (2018) Trimmed dual system estimation. In: Böhning, D., Van der Heijden, P. & Bunge, J. (Eds.) *Capture–recapture methods for the social and medical sciences*. Boca Raton: CRC Press, pp. 229–235.

Zwane, E. & van der Heijden, P.G.M. (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. https://doi.org/10.1002/sim.2577

Zwane, E., Van der Pal-de Bruin, K. & Van der Heijden, P.G.M. (2004) The multiple-record systems estimator when registrations refer to different but overlapping populations. *Statistics in Medicine*, 23, 2267–2281. https://doi.org/10.1002/sim.1818

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: van der Heijden PG, Cruyff M, Smith PA, Bycroft C, Graham P, Matheson-Dunning N. Multiple system estimation using covariates having missing values and measurement error: Estimating the size of the Māori population in New Zealand. *J R Stat Soc Series A*. 2022;185:156–177. https://doi.org/10.1111/rssa.12731