Multiple system estimation for the size of the Māori population in New Zealand

Maarten Cruyff[1], Peter G.M. van der Heijden[1,2], Paul A. Smith[2], Christine Bycroft[3], Patrick Graham[3]
[1] Utrecht University, the Netherlands
[2] University of Southampton, UK
[3] Statistics New Zealand

Abstract

We investigate the situation where two or more registers, or lists, of individuals are linked both for the purpose of population size estimation and to investigate the relationship between variables appearing on all or only some of the registers. There is usually no full picture of this relationship because there are individuals that are in only some of the lists, and also individuals that are in none of the lists. These two problems have been solved simultaneously in dual system estimation using the EM algorithm. We extend this approach to four registers (including the population census) to estimate the size of the indigenous Māori population in New Zealand, where the reporting of Māori is not the same in each register and where there is a further missing data problem, with individuals included in one or more registers who did not provide their ethnicity. We consider the implications for estimating the size of the Māori population from administrative data only.

Keywords

dual system estimation, linkage, missing data, register, coverage

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 (eg Brown et al. 1999, 2019), though also in situations involving the use of linked administrative data sources. 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, and linkage of the records from these sources is being widely used to understand and estimate corrections for the under- and over-coverage within them. We will use "registers" as a generic term for all sources containing lists of identifiable units.

When two registers are linked, in general there will be some records from one register which remain unlinked, because there is no corresponding record in the other source. This leads to missing data for any variables which appear in only one register (item missingness). The linked data are used to estimate the size of the population that is not present in either register, and for these unobserved records all the variables are missing (unit missingness). There is an extensive line of research that studies this problem from a missing data perspective, starting with Zwane and van der Heijden (2007), and summarized in van der Heijden et al. (2018). The latter paper concluded that further practical experience with these methods is needed to demonstrate their usefulness in a variety of situations and encourage their wider application.

Here we consider the methods for estimating the size of the Māori population in New Zealand. Ethnicity is the principal measure of cultural identity in New Zealand, and is used across the official statistics system. The 2005 New Zealand statistical standard for ethnicity states that ethnicity is self-perceived and a person can belong to more than one ethnic group. Identifying the indigenous Māori population is of particular importance.

Ethnicity is regularly included in data collections because of its importance in defining groups of policy interest, for example on health outcomes for indigenous people in New Zealand. However, differences in questions, differences in self-perception depending on the context, and changes over time, can all affect how ethnicity is recorded in these data sources (Statistics New Zealand 2005). Ethnicity is collected independently in a number of administrative sources as well as through the census and household surveys. People do not always report the same ethnicity in each source. Also, people do not always report their ethnicity, so there is an additional missingness problem to deal with.

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 NZ is exploring the feasibility of a census based on administrative data (Statistics New Zealand, 2012, 2014). The ability to produce ethnicity data from administrative sources is a key consideration. Using ethnicity information from linked administrative data sources may also improve the current production of official ethnic population estimates.

The aim is to use ethnicity information from linked administrative data to improve official ethnic population estimates in New Zealand. In support of this we analyse a variety of census and administrative sources 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 registers and with a variety of different types of missing data. The methodology falls within the area of data integration of multi-source statistics, see de Waal et al. (2017) and Zhang and Chambers (2018).

Four data sources, the population census and three administrative registers are available, that each have an ethnicity variable. Here we focus on Māori ethnicity in a summarised binary form so that we have two mutually exclusive categories: Māori (with or without other ethnicities) and non-Māori (everyone else). Details of these sources and the procedures which have been used to link them are described in section 2; perfect linkage is an essential assumption for DSE. Then we build up the estimation problem in section 3, starting with two registers, and then four registers, and finally consider using the three administrative sources without the census. Some conclusions are presented in Section 4.

2. Methodolgy

Because a person's reported ethnicity can change over time, and depending on the context, a key question is how to combine ethnicity from multiple sources, when information is sometimes conflicting. Reid, Bycroft, and Gleisner (2016) compared ethnicity data from the 2013 Census with the ethnicity information collected by administrative sources, for a New Zealand resident population derived from administrative sources. They found that nearly everyone in this admin-based New Zealand resident population had ethnicity recorded in at least one administrative data source, but that consistency with census responses varied considerably by source and by ethnic group. The method used to combine these sources 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 sources based on their consistency with the census has been applied. Using administrative data alone was found to produce a time series that reflects expected patterns of increasing ethnic diversity, with age structure and regional distribution of ethnicity consistently in line with official measures (Stats NZ, 2018). The approach however has some limitations, for example it does not allow for reporting errors or conflicts in higher-ranked sources, which may be better managed through a statistical model.

The population used here is the experimental administrative-based NZ resident population known as the 'IDI-ERP' (Stats NZ, 2017). The data are probabilistically linked in Stats NZ's Integrated Data Infrastructure (IDI). The IDI provides safe access to de-identified linked microdata for research and statistics in the public interest.

We use ethnicity data from the 2013 population census and from three administrative sources:

(i) Department of Internal Affairs (DIA) birth registrations data - which includes the ethnicity of the child as reported at registration (ii) Ministry of Education (MOE) tertiary education enrolment data - which includes ethnicity for students (iii) Ministry of Health (MOH) National Health Index system, a unified national person list - which includes ethnicity. For a more detailed explanation of these sources, see Reid et al. (2016).

Each of the administrative sources 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 around the late 1990s, and are mainly for those aged between 18 and 40 years in 2013; both census and health data include all ages, and each has an ethnicity value for around 90 % of the IDI-ERP population. Overall, almost 99 percent of the IDI-ERP population have ethnicity information from at least one of these sources, and many people have information from more than one source.

The aim of the following analyses is to produce aggregate estimates of Māori and non-Māori ethnicity by combining these four independent sources: the 2013 Census and the three administrative sources.

3. Results

3.1 Two registers

We first explain the methodology for two registers and then apply it to four registers. We start by using the two sources with the widest coverage, the Census and the MOH. Being in the Census is denoted by A (A = 1 for ‘yes’, 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 these methods with two registers.

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. The middle two columns 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 is 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 only in A ((A; C) = (1; 0)), there are three types of individuals, namely 0, non-Māori (light grey); 1, Māori (blocks); and ‘-’, those who have a missing value for ethnicity (raster). When individuals are in both A and C ((A; C) = (1; 1)), all three areas are found. 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 ethnicity variable c, and it has similar levels as 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 missed by both A and C.

A second presentation of the problem is in contingency table format, see Table 1, Panel 1. The original 15 counts in Table 1, Panel 1, will have to be redistributed over 3 subtables of dimension 2×2. I.e., 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. Estimates are found using the Expectation- Maximization algorithm. 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 fitted margins are placed between square brackets. The maximal model [Ac][ac][Ca] is saturated in the sense that the fitted values are equal to the observed values. The result is given in Table 1, Panel 2. Due to the fitted model, in each of the three estimated 2×2 subtables the a*c odds ratio is identical and equal to 377.9. The lower right 2×2 table in Panel 2 of Table 1 shows the estimated numbers of people missing from both census and MOH. These numbers there are relatively low, due to the large overlap of the two registers. The estimated total population size for New Zealand is 4,383,613.7. The census and the MOH differ in which part of this

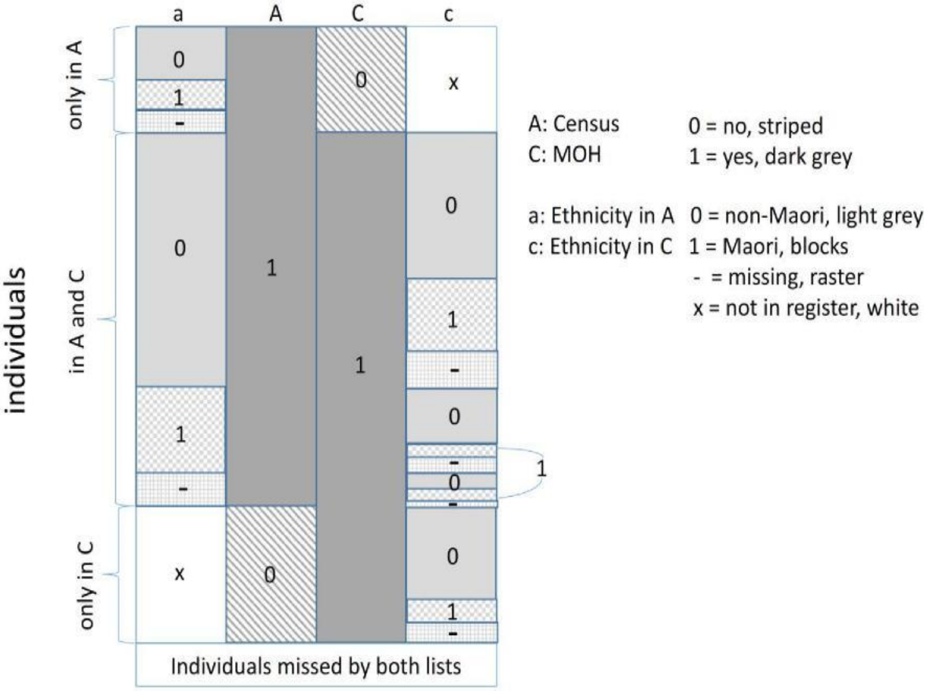
a
A
C
c
only in A
in A and C
only in C
individuals
0
1
-
x
1
Individuals missed by both lists
A: Census 0 = no, striped
C: MOH 1 = yes, dark grey
a: Ethnicity in A 0 = non-Maori, light grey
c: Ethnicity in C 1 = Maori, blocks
- = missing, raster
x = not in register, white

Figure 1: Graphical representation of two linked registers

total is Māori. For the Census the estimated number is 721,971. For the MOH it is 640,711.

3.2 Four registers

We now add the other two registers, DIA and MOE, to the analysis. Now the maximal model is [ABCd][ABDc] [ACDb][BCDa] [ABcd][ACbd] [ADbc][BCad] [BDac][CDab] [Abcd][Bacd] [Cabd] [Dabc][abcd]. Notice that, as for the two registers, a capital variable label cannot be in the same interaction term as a lower case variable label, as these interactions cannot be estimated from the data. Notice that the assumptions become less and less demanding as more registers are involved. The number of unique individuals in the four linked registers is 4,401,282, and the estimated number missed by all registers is 25,939, giving an estimated population size of 4,427,221.

The estimated numbers of Maori are displayed in Table 2. 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. Consider the margins of the ethnicity variables a; b; c and d of the four registers. A statistical approach to measurement error is to make use of a latent class model (McCutcheon, 1987). See Table 3. 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 0.826. The probability for the Māori class corresponds to an estimated Māori population size of about 770,000. Estimated conditional probabilities of being Māori for each latent class are also shown in Table 3; they are consistently low for the non-Māori latent class and high for the Māori latent class.

Panel 1: Observed counts

			C = 1		C = 0	Totals
		c = 0	c = 1	c = -	c = x	
A = 1	a = 0	3,004,329	31,998	150,855	38,640	3,225,822
	a = 1	108,192	435,468	12,402	4,377	530,439
	a = -	16,512	2,769	894	435	20,160
A = 0	a = x	398,838	146,985	24,642	-	570465
Totals		3,527,871	617,220	188,793	43,452	4,377,336

Panel 2: Fitted values under [Ac][ac][Ca]

		C = 1		C = 0		Totals
		c = 0	c = 1	c = 0	c = 1	
A = 1	a = 0	3,170,298.4	33,791.2	38,619.1	411.6	3,243,120.3
	a = 1	111,244.8	448,084.6	879.3	3,541.9	563,750.6
A = 0	a = 0	402,713.4	10,772.5	4,905.7	131.2	418,522.8
	a = 1	14,131.1	142,848.3	111.7	1,129.2	158,220.3
Totals		3,698,387.7	635,496.6	44,515.8	5,213.9	4,383,613.7

Table 1: Census *(A)* linked to MOH (C). Covariate 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 register). Observed counts have been randomly rounded to protect confidentiality. Source: Stats NZ.

	Census	DIA	MOH	MOE
non-Māori	3,690,913	3,668,349	3,782,239	3,665,099
Māori	736,308	758,872	644,983	762,122

Table 2: Summary of Census linked to DIA, MOH and MOE, estimated numbers

		census	DIA	MOH	MOE
	π_x	$\pi^a_{r=1\|x}$	$\pi^b_{s=1\|x}$	$\pi^c_{t=1\|x}$	$\pi^d_{u=1\|x}$
Class 1	0.826	0.004	0.012	0.003	0.014
Class 2	0.174	0.939	0.930	0.824	0.924

Table 3: Estimates of latent class model with two latent classes

3.3 Three registers without the Census

We also made estimations for three registers without the Census, see Table 4. We also present estimates derived only from the three administrative data sources, 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 4,377,573. We estimate an additional 24,058 individuals missed by all three registers. This leads to a total population size of 4,401,631. This is somewhat less than the four register estimate of 4,427,221.

	DIA	MOH	MOE
non-Māori	3,599,611	3,760,211	3,625,453
Māori	802,020	641,421	776,179

Table 4: Summary of DIA, MOH and MOE, ignoring census, estimated numbers.

4. Discussion and Conclusion

Van der Heijden et al. (2018) presented an approach for estimating the margins of auxiliary variables in the dual system estimation 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 registers, and a more complicated missing data structure. 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 registers considered, and of recording Māori status.

References

1. Brown, J.J., C. Sexton, O. Abbott, and P.A. Smith (2019, in press). 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.*
2. Brown, J.J., I.D. Diamond, R.L. Chambers, L.J. Buckner, and A.D. Teague (1999). A methodological strategy for a one-number census in the UK. *Journal of the Royal Statistical Society: Series A*, 162(2), 247-267.
3. De Waal, T., A. van Delden and S. Scholtus (2017). Multi-source Statistics: Basic Situations and Methods. The Hague, Statistics Netherlands, discussion paper 2017/12.
4. McCutcheon, A.L. (1987) *Latent class analysis*. Sage, Newby Park.
5. Reid, G., C. Bycroft, and F. Gleisner (2016). Comparison of ethnicity information in administrative data and the census. Statistics New Zealand, Christchurch. 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.
6. Statistics New Zealand (2005). Statistical standard for ethnicity. Christchurch, Statistics New Zealand. http://archive.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards/ ethnicity.aspx.
7. Statistics New Zealand (2012). Transforming the New Zealand Census of Population and Dwellings: Issues, options, and strategy. Christchurch, Statistics New Zealand. URL retrieved from https://www.stats.govt.nz.
8. 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. URL retrieved from https://www.stats.govt.nz.
9. Statistics New Zealand (2017). Experimental population estimates from linked administrative data: 2017 release. Christchurch, Statistics New Zealand. URL retrieved from https://www.stats.govt.nz.
10. Statistics New Zealand (2018). Experimental ethnic population estimates from linked administrative data. Christchurch, Statistics New Zealand. URL retrieved from https://www.stats.govt.nz.
11. Van der Heijden, P.G.M., P.A. Smith, M. Cruyff, and B. 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 Official Statistics, 34, 239-263.
12. Zhang, L.-C., and R.L. Chambers (Eds) (2018). Analysis of integrated data. Boca Raton: CRC Press.

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