Stats NZ
Tatauranga Aotearoa

Estimating population size of four ethnic groupings in New Zealand

Paul A Smith, University of Southampton

Peter van der Heijden, Utrecht University and University of Southampton

Maarten Cruyff, Utrecht University

Francesco Pantalone, University of Southampton

Hannes Diener, Statistics New Zealand

Kim Dunstan, Statistics New Zealand

Te Kāwanatanga o Aotearoa
New Zealand Government

Citation
Smith, P. A., van der Heijden, P. G., Cruyff, M., Pantalone, F., Diener, H., Dunstan, K. (2023). *Estimating population size of four ethnic groupings in New Zealand*. Retrieved from www.stats.govt.nz.

ISBN 978-1-99-104943-8 (online)

Acknowledgement
Access to the data used in this study was provided by Stats NZ under conditions designed to give effect to the security and confidentiality provisions of the Data and Statistics Act 2022. The results presented in this study are the work of the authors, not Stats NZ or individual data suppliers.

Disclaimer for output produced from the IDI
These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) which is carefully managed by Stats NZ. For more information about the IDI please visit https://www.stats.govt.nz/integrated-data

Published in October 2023 by
Stats NZ Tatauranga Aotearoa
Wellington, New Zealand

Contact
Stats NZ Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4600

www.stats.govt.nz

Contents

List of tables and figures

List of tables

List of figures

This research paper reports on a method for estimating populations by ethnicity, using multiple system estimation and latent class analysis. The method is applied to the broad and overlapping Māori and Pacific ethnic groups to produce an annual time series from 2013 to 2020. The resulting estimates are close to the official ethnic estimated resident populations (ERP).

1 Introduction

The use of administrative data to support estimation in official statistics is becoming widespread. One important area is the matching of data from administrative sources (admin data), which can be treated as multiple captures of a human population and used to estimate the numbers of people missed from those sources.

New *Z*ealand has three administrative sources with wide coverage of the resident population, and the five-yearly Census of Population and Dwellings (we refer to all the sources as registers, although strictly the census is a different type of data collection).

The administrative sources are:

- Department of Internal Affairs (DIA) birth registrations data
- Ministry of Health (MOH) National Health Index system, a unified national person list
- Ministry of Education (MOE) tertiary education enrolment data.

Previous research (van der Heijden *et al*, 2022) by the authors of the current research used admin data from the above three sources that had been matched with census data in the Integrated Data Infrastructure (IDI) by Statistics New *Z*ealand (Stats NZ), to estimate the sizes of the Māori and non-Māori populations in New *Z*ealand. All the registers contain an ethnicity variable, and there appear to be differences in the ways people report their ethnicity in different datasets. Van der Heijden *et al* (2022) estimate the population sizes according to the different definitions of ethnicity based on data from 2013 matched to the 2013 Census, and also introduce latent class multiple system estimation (LCMSE), which combines a latent class model with the usual multiple system estimation to produce a consensus estimate of the split between Māori and non-Māori populations.

Our research extends this approach in two ways:

- by examining how the population size estimates and the latent class model estimates evolve over time. We examine the change from the 2013 Census to the 2018 Census; in both census years all four sources are available. In intervening years, we can still work with four sources, but need to use a census dataset that doesn't correspond precisely with the year of interest, resulting in lower levels of matching. We can also examine the pattern of estimates derived only from the admin data over time.
- by extending the ethnicity analysis to more categories. In principle the methods of van der Heijden *et al* (2022) should work with more categories, but this needs to be demonstrated in practice. The areas that are most challenging are the estimation of the sizes of small populations, and an examination of whether the LCMSE remains interpretable with a larger number of categories.

For the analysis of these features, we use a dataset covering 2013–2020, and use Māori and Pacific ethnicities. These ethnicities are not mutually exclusive, so in fact using these two classifiers generates four ethnic groupings:

- neither Māori nor Pacific
- Māori and not Pacific
- Pacific and not Māori
- both Māori and Pacific.

These four mutually exclusive ethnic groupings give flexibility to sum to various combinations including all those who identify with the Māori ethnicity (by summing “Māori and not Pacific” and “both Māori and Pacific”) and all those who identify with a Pacific ethnicity (by summing “Pacific and not Māori” and “both Māori and Pacific”). The data are described in more detail in section 2.

The estimates presented in this report are adjusted for undercoverage. That is, after linking the four data sources, the models estimate, among others, the number of individuals missed by all data sources.

2 Data

The population used in this research is the experimental administrative-based New *Z*ealand resident population known as the 'IDI-ERP' (Statistics New *Z*ealand, 2017), which was also used in previous research by van der Heijden *et al* (2022).

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 dataset. We are interested in the population size, and therefore will implicitly assess the coverage of different sources within the IDI-ERP relative to our estimate of the size of the New *Z*ealand population.

The data are probabilistically linked in Stats NZ's 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 Health (MOH) National Health Index system, a unified national person list which includes ethnicity
- Ministry of Education (MOE) tertiary education enrolment data – which includes ethnicity of students.

Each of the administrative registers relates to different parts of the population. Birth registrations are for babies born in New *Z*ealand since 1998, or those up to age 14 at the time of the 2013 Census; 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 percent of the IDI-ERP population. Overall, almost 99 percent of the IDI-ERP population have ethnicity information from at least one of these lists, and many people have information from more than one.

The data are derived from linkage within the IDI, as effective at 30 June 2022. The data used in the paper of van der Heijden *et al* (2022) were linked with the 2013 Census. In our new analysis, the three administrative data sources are linked to the 2013 Census in 2013–2017, and then from 2018–2020 are linked to the 2018 Census. The overlap with the census is greatest in census years and becomes smaller with greater distance. This reflects how the data would accumulate in practice – admin data could only be linked to the most recent census, which becomes progressively out of date, and is then replaced by a new census.

Table 1 and table 2 give basic information about the linked data sources by year. We explain the first column for 2013 in table 1. This is the column for the 2013 Census. The other columns, also in other years, can be interpreted in an identical way.

After linking the four data sources the number of individuals that is observed in at least one data source is 4,434,612. The number of individuals not observed in the census is 606,735, so the number of individuals in the 2013 Census is 3,827,877 (4,434,612 minus 606,735), split over four ethnic groupings:

- individuals that report that they are “neither Māori nor Pacific” (observed 3,013,380)
- individuals that report that they are “Māori and not Pacific” (observed 518,691)
- individuals that report that they are “Pacific and not Māori” (observed 228,051)
- individuals that report that they are “both Māori and Pacific” (observed 46,143).

Lastly, there are 21,612 individuals that are observed in the census but do not report an ethnicity.

Table 1

Linked population by ethnic grouping, 2013–2017

2013	Census	DIA[1]	MOH[2]	MOE[3]
Neither Māori nor Pacific	3,013,380	1,229,853	3,351,933	2,393,586
Māori and not Pacific	518,691	421,314	630,555	614,781
Pacific and not Māori	228,051	151,659	303,513	257,823
Both Māori and Pacific	46,143	50,715	38,124	52,317
Ethnicity not provided	21,612	1,552,737	42,069	97,194
Individuals missed	606,735	1,028,334	68,418	1,018,911
Total	4,434,612	4,434,612	4,434,612	4,434,612
2014	Census	DIA	MOH	MOE
Neither Māori nor Pacific	2,955,504	1,268,964	3,398,493	2,443,092
Māori and not Pacific	510,360	432,987	639,462	624,075
Pacific and not Māori	221,778	155,505	307,512	262,290
Both Māori and Pacific	45,210	53,022	40,329	53,763
Ethnicity not provided	20,733	1,526,988	39,702	99,735
Individuals missed	739,668	1,055,787	67,755	1,010,298
Total	4,493,253	4,493,253	4,493,253	4,493,253
2015	Census	DIA	MOH	MOE
Neither Māori nor Pacific	2,903,655	1,310,427	3,457,524	2,501,229
Māori and not Pacific	503,880	445,677	650,247	635,463
Pacific and not Māori	216,651	159,459	312,654	267,186
Both Māori and Pacific	44,601	55,491	42,726	55,452
Ethnicity not provided	20,034	1,502,613	37,776	103,719
Individuals missed	881,754	1,096,908	69,648	1,007,526
Total	4,570,575	4,570,575	4,570,575	4,570,575
2016	Census	DIA	MOH	MOE
Neither Māori nor Pacific	2,857,749	1,351,266	3,523,896	2,562,399
Māori and not Pacific	498,513	459,258	662,463	647,664
Pacific and not Māori	211,983	163,569	318,462	272,598
Both Māori and Pacific	44,187	58,188	45,243	57,336
Ethnicity not provided	19,491	1,479,696	36,468	107,955
Individuals missed	1,026,303	1,146,249	71,694	1,010,274
Total	4,658,226	4,658,226	4,658,226	4,658,226
2017	Census	DIA	MOH	MOE
Neither Māori nor Pacific	2,815,746	1,391,097	3,589,869	2,620,710
Māori and not Pacific	493,458	472,617	674,265	657,762
Pacific and not Māori	207,846	167,541	324,033	277,356
Both Māori and Pacific	43,848	61,089	47,856	59,277
Ethnicity not provided	18,975	1,457,475	35,580	111,924
Individuals missed	1,162,161	1,192,215	70,431	1,015,005
Total	4,742,034	4,742,034	4,742,034	4,742,034

1. DIA = Department of Internal Affairs
2. MOH = Ministry of Health
3. MOE = Ministry of Education

The digitisation of admin data in New *Z*ealand has a finite time horizon, so the DIA covers only births since 1998, and the MOE covers only a similar period (but with a less precise start relative to the age of people included). In our previous research (van der Heijden *et al*, 2022) we used only the data

corresponding with these periods, but our current research includes the people on the DIA register who were born in earlier periods. Ethnicity information was not collected with the admin data for these people, which is why there appear to be large numbers of people without a recorded ethnicity in the DIA register. We return to this point in section 8.

Stats NZ imputed ethnicity in some 2018 Census records for the census outputs. These cases have been returned to missing ethnicity for the data used in this project, so all the ethnicity information used is actually observed.

Disclosure protection is applied to the raw data outputs from the IDI such that:

- values < 6 are suppressed, and cannot be distinguished from actual zeros
- values ≥ 6 are randomly rounded such that
  - values that are already multiples of 3 are left unchanged
  - other values are randomly rounded to the nearest multiple of 3 with a probability of ⅔, and to the second nearest multiple of 3 with a probability of ⅓.

Table 2

Linked population by ethnic grouping, 2018–2020

2018	Census	DIA[1]	MOH[2]	MOE[3]
Neither Māori nor Pacific	3,078,318	1,427,292	3,648,579	2,668,656
Māori and not Pacific	483,999	485,241	684,234	665,691
Pacific and not Māori	199,563	171,864	330,213	281,910
Both Māori and Pacific	47,187	63,696	49,554	61,068
Ethnicity not provided	647,754	1,434,051	35,499	114,786
Individuals missed	358,092	1,232,769	66,834	1,022,802
Total	4,814,913	4,814,913	4,814,913	4,814,913
2019	Census	DIA	MOH	MOE
Neither Māori nor Pacific	3,022,869	1,460,334	3,706,665	2,707,614
Māori and not Pacific	478,566	497,613	693,210	671,784
Pacific and not Māori	195,933	176,454	336,888	285,639
Both Māori and Pacific	46,596	66,576	51,033	62,472
Ethnicity not provided	629,643	1,411,518	36,420	116,943
Individuals missed	512,958	1,274,070	62,349	1,042,113
Total	4,886,565	4,886,565	4,886,565	4,886,565
2020	Census	DIA	MOH	MOE
Neither Māori nor Pacific	2,980,548	1,494,033	3,779,628	2,744,556
Māori and not Pacific	474,693	510,735	703,938	677,415
Pacific and not Māori	193,980	181,659	348,870	289,728
Both Māori and Pacific	46,314	69,630	52,683	63,717
Ethnicity not provided	618,963	1,399,071	37,956	117,762
Individuals missed	662,205	1,321,575	53,628	1,083,525
Total	4,976,703	4,976,703	4,976,703	4,976,703

1. DIA = Department of Internal Affairs
2. MOH = Ministry of Health
3. MOE = Ministry of Education

Some investigation (not presented here) with the models in van der Heijden *et al* (2022) demonstrated that this had negligible effect on the results. With the larger number of small cells in the latest dataset (because of the more detailed ethnicity classification), we were concerned that the disclosure protection may have a larger effect on the model fits. We therefore again investigated the effect of the disclosure protection on the model choice and fits, and a summary of the results is given in the appendix.

3 Model

To make the model descriptions sufficiently concise, we need to introduce some notation, extending the notation in van der Heijden *et al* (2022) for the new ethnic groupings in the four data sources, that are census, DIA, MOH, and MOE respectively. The corresponding ethnicity variable in each of the four data sources is denoted with the lowercase letters *a*, *b*, *c*, and *d* respectively, so *a* is the ethnicity in the census. Each of these ethnicity variables has four levels, namely "neither Māori nor Pacific", "Māori and not Pacific", "Pacific and not Māori", and "both Māori and Pacific". To keep the notation simple, we also denote the level of ethnicity by *a*, *b*, *c*, and *d*, and from the context it will be clear whether these letters refer to the ethnicity variable or the levels of the variable.

Previous research by van der Heijden *et al* (2022) made use of the latent class (LC) model. The latent class model assumes the existence of a categorical latent variable, and that the observed variables are independent conditional on this 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 the ethnicity variables *a*, *b*, *c*, and *d*. Let *x* be the latent variable; the number of latent classes needs to be prespecified by the researcher, also, to keep the notation simple, indexed by *x*.

In this paper we investigate latent class models with two, three, and four levels. Let π_x be the probability to fall in latent class *x*. Let $\pi_{a|x}$ be the conditional probability of a census ethnicity given latent class *x*. Then the latent class model is

$$\pi_{abcdx} = \pi_x \pi_{a|x} \pi_{b|x} \pi_{c|x} \pi_{d|x} \quad (1)$$

$$\pi_{abcd} = \Sigma_x \pi_{rabcd} \quad (2)$$

In the earlier research, van der Heijden *et al* (2022) investigated two approaches to model fitting:

- first, a two-stage approach where the latent class model was fitted on the estimates from multiple system estimation (MSE)
- second, a single stage approach in which the latent class and MSE models were combined, using the latent classes to explain some of the interactions in the MSE. We call this the latent class MSE (LCMSE) model.

For the current research, we put most trust in the LCMSE, because these models account simultaneously for all the components, and treat the variability appropriately. We therefore report only on these models in our interpretations below.

4 Results: LCMSE with four ethnic groupings and two latent classes

We carried out latent class analyses using latent class multiple system estimation (LCMSE) (see section 3), where the latent variable is indicated with *Z*. This aligns with the results of van der Heijden *et al* (2022), where the LCMSE models provide estimates of the Māori and non-Māori population sizes that are very close to those published by Stats NZ.

In this section the number of latent classes is two and the number of observed ethnic groupings is four. Also, as the results are very stable, we only discuss the estimates for the year 2013. The purpose of this section is to make an intermediate step from two reported ethnic groupings and two latent classes in the original analysis to four reported ethnic groupings and four latent classes (reported in section 5).

The latent class model has two types of parameters: the latent class sizes, and the conditional probabilities of falling into the levels of the observed variables given that one falls into a latent class. The estimated latent class sizes are shown as proportions of the population in table 3.

Table 3

Estimated latent class sizes of LCMSE with two latent classes

Year	Latent class 1	Latent class 2
2013	0.7618	0.2382
2014	0.7620	0.2380
2015	0.7626	0.2374
2016	0.7632	0.2368
2017	0.7637	0.2363
2018	0.7617	0.2383
2019	0.7613	0.2387
2020	0.7601	0.2399

For 2013, the first class has estimated size 0.7618 and the second class has estimated size 0.2382. We note that this is different from the LCMSE estimates for 2013 in van der Heijden *et al* (2022), which were 0.834 and 0.166. The reason for this difference is that van der Heijden *et al* (2022) distinguish only two ethnic groupings of “not Māori” and “Māori”, where the ethnic grouping of “Pacific and not Māori” was part of the “not Māori” grouping.

The conditional probabilities can be used to interpret the two latent classes, and we focus on the conditional probabilities related to latent variable *Z* for the year 2013 (table 4).

For 2013, given that one is in latent class $Z = 1$, the estimated conditional probabilities for the four census ethnic groupings (variable *a*) to fall in the first latent class are 0.9931, 0.0061, 0.0007, and 0.0001 (figure 1, first pie chart). The first latent class $Z = 1$ is clearly the latent class for “neither Māori nor Pacific” (the majority of people report this ethnic grouping). For 2013, given that one is in latent class $Z = 2$, the conditional probabilities for the four census ethnic groupings to fall in the second latent class are 0.0469, 0.6134, 0.2836, and 0.0561, so this is a mixture of people reporting “Māori and not Pacific” (with probability 0.6134) and “Pacific and not Māori” (with probability 0.2836), with minor probabilities for answering that one is “neither Māori nor Pacific” (0.0469) or “both Māori and Pacific” (0.0561) (see figure 1, second pie chart).

For DIA the estimated conditional probabilities are similar, with 0.9886, 0.0098, 0.0015, and 0.0001 for latent class $Z = 1$ and 0.0367, 0.6371, 0.2486, and 0.0776 for latent class $Z = 2$.

For MOH we also find that the estimated conditional probabilities are similar, with estimates 0.9916, 0.0045, 0.0039, and 0.0000 for the first latent class and estimates 0.0831, 0.5922, 0.2882, and 0.0365 for the second latent class.

And last, for MOE, estimates are 0.9807, 0.0119, 0.0074, 0.0001 for the first latent class and 0.0475, 0.6190, 0.2785, and 0.0549 for the second latent class.

Table 4

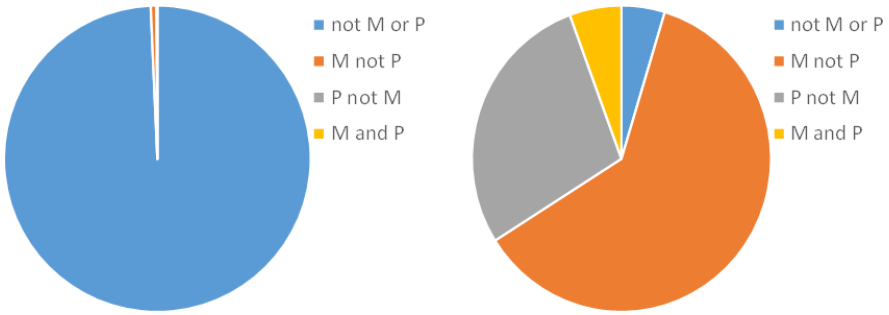
Estimated conditional probability of being in the four ethnic groupings given latent classes *Z*1 and *Z*2 from the two latent class model

Year	*Z*1_*a*1	*Z*1_*a*2	*Z*1_*a*3	*Z*1_*a*4	*Z*1_*b*1	*Z*1_*b*2	*Z*1_*b*3	*Z*1_*b*4
2013	0.9931	0.0061	0.0007	0.0001	0.9886	0.0098	0.0015	0.0001
2014	0.9931	0.0061	0.0007	0.0001	0.9888	0.0096	0.0015	0.0001
2015	0.9931	0.0061	0.0007	0.0001	0.9890	0.0095	0.0015	0.0001
2016	0.9931	0.0061	0.0007	0.0001	0.9892	0.0093	0.0014	0.0001
2017	0.9931	0.0060	0.0008	0.0001	0.9893	0.0092	0.0014	0.0001
2018	0.9901	0.0084	0.0013	0.0002	0.9895	0.0090	0.0014	0.0001
2019	0.9900	0.0085	0.0013	0.0002	0.9891	0.0094	0.0015	0.0001
2020	0.9900	0.0085	0.0013	0.0002	0.9885	0.0098	0.0015	0.0001
Year	*Z*1_*c*1	*Z*1_*c*2	*Z*1_*c*3	*Z*1_*c*4	*Z*1_*d*1	*Z*1_*d*2	*Z*1_*d*3	*Z*1_*d*4
2013	0.9916	0.0045	0.0039	0.0000	0.9807	0.0119	0.0074	0.0001
2014	0.9916	0.0044	0.0039	0.0000	0.9807	0.0118	0.0074	0.0001
2015	0.9917	0.0044	0.0039	0.0000	0.9808	0.0117	0.0074	0.0001
2016	0.9917	0.0043	0.0039	0.0000	0.9810	0.0115	0.0074	0.0001
2017	0.9918	0.0043	0.0039	0.0000	0.9811	0.0114	0.0074	0.0001
2018	0.9922	0.0037	0.0040	0.0000	0.9822	0.0103	0.0074	0.0001
2019	0.9924	0.0036	0.0040	0.0000	0.9823	0.0102	0.0074	0.0001
2020	0.9924	0.0036	0.0040	0.0000	0.9825	0.0101	0.0073	0.0001
Year	*Z*2_*a*1	*Z*2_*a*2	*Z*2_*a*3	*Z*2_*a*4	*Z*2_*b*1	*Z*2_*b*2	*Z*2_*b*3	*Z*2_*b*4
2013	0.0469	0.6134	0.2836	0.0561	0.0367	0.6371	0.2486	0.0776
2014	0.0475	0.6151	0.2814	0.0561	0.0363	0.6378	0.2476	0.0784
2015	0.0480	0.6161	0.2798	0.0561	0.0358	0.6380	0.2464	0.0798
2016	0.0484	0.6168	0.2785	0.0563	0.0353	0.6384	0.2450	0.0812
2017	0.0488	0.6177	0.2770	0.0565	0.0347	0.6396	0.2424	0.0833
2018	0.0521	0.6164	0.2689	0.0627	0.0374	0.6381	0.2392	0.0852
2019	0.0529	0.6166	0.2681	0.0624	0.0366	0.6377	0.2384	0.0872
2020	0.0539	0.6145	0.2697	0.0619	0.0360	0.6366	0.2384	0.0891
Year	*Z*2_*c*1	*Z*2_*c*2	*Z*2_*c*3	*Z*2_*c*4	*Z*2_*d*1	*Z*2_*d*2	*Z*2_*d*3	*Z*2_*d*4
2013	0.0831	0.5922	0.2882	0.0365	0.0475	0.6190	0.2785	0.0549
2014	0.0816	0.5923	0.2879	0.0381	0.0489	0.6177	0.2780	0.0555
2015	0.0803	0.5921	0.2880	0.0397	0.0498	0.6169	0.2772	0.0561
2016	0.0788	0.5917	0.2883	0.0412	0.0507	0.6157	0.2767	0.0568
2017	0.0778	0.5917	0.2877	0.0428	0.0516	0.6146	0.2761	0.0578
2018	0.0806	0.5914	0.2845	0.0435	0.0533	0.6140	0.2740	0.0586
2019	0.0818	0.5892	0.2849	0.0441	0.0537	0.6124	0.2746	0.0593
2020	0.0828	0.5842	0.2885	0.0445	0.0540	0.6090	0.2771	0.0599

Note: The column indicator *Z*1_*a*1 stands for P ($a = 1 \mid Z = 1$). The ethnic groupings in registers are denoted *a* in Census, *b* in DIA, *c* in MOH, and *d* in MOE. Interpretation of ethnic groupings are 1 – neither Māori nor Pacific, 2 – Māori and not Pacific, 3 – Pacific and not Māori, and 4 – both Māori and Pacific.

Figure 1

Composition of two fitted latent classes according to ethnicity information observed in the 2013 Census



M = Māori ethnicity, P = Pacific ethnicity
Note: We interpret that the first latent class (on the left) represents the "neither Māori nor Pacific" grouping, and the second latent class (on the right) represents a mixture but mainly the "Māori and not Pacific" and "Pacific and not Māori" groupings.

We conclude the following:

- The two latent classes do correspond with the ethnic groupings in the input data.
- The first latent class is "neither Māori nor Pacific". Given that one falls in this latent class, one seldom indicates having a Māori or Pacific ethnicity.
- The second latent class is mainly a mix of "Māori and not Pacific" and "Pacific and not Māori", with smaller probabilities for "both Māori and Pacific" and "neither Māori nor Pacific".
- The estimated conditional probabilities for the second latent class fluctuate a bit over the four data sources. For example, for 2013, the conditional probability to answer that one is "Pacific and not Māori" (ethnic grouping 3 in table 4) is 0.2836 in the census (the value for $Z2\_\alpha3$), 0.2486 for DIA, 0.2882 for MOH, and 0.2785 for MOE. This does not stand in the way of a clear interpretation.
- The estimates of the latent class sizes are very stable across the years, with estimates 0.7618 for the first latent class (the class for the "neither Māori nor Pacific") and 0.2382 for the second latent class (mainly the class for "Māori and not Pacific" and "Pacific and not Māori").

5 Results: LCMSE with four ethnic groupings and four latent classes

We now investigate whether it is possible to estimate models for four latent classes, one for each of the four ethnic groupings. As in section 4, we describe the results provided by the LCMSE models.

The purpose of this section is to make the step from two latent classes and four reported ethnic groupings to four latent classes and four ethnic groupings. The main question is whether the LCMSE model can assign each of the four latent classes to separate ethnic groupings.

A latent class model with four binary lists and three or more latent classes is not identified (that is, the parameter estimates are not unique) (Goodman, 1974). However, we use a model with eight variables (four for list membership and four for ethnicity) with a complex structure of missingness and structural zeroes. It is difficult to assess identifiability analytically for this kind of situation, so instead we test by rerunning the algorithm multiple times from different starting values for the parameters. In the four latent class model, this sometimes fails to converge or gets stuck at a local maximum, but when the global maximum is reached, it is unique[1]. So, in this case the model appears to be identifiable and we can make inference on the parameters.

Table 5

Estimated latent class sizes of LCMSE with four latent classes

Year	Latent class 1	Latent class 2	Latent class 3	Latent class 4
2013	0.7605	0.1540	0.0697	0.0158
2014	0.7608	0.1536	0.0696	0.0160
2015	0.7615	0.1528	0.0694	0.0163
2016	0.7621	0.1521	0.0692	0.0165
2017	0.7627	0.1515	0.0689	0.0169
2018	0.7605	0.1533	0.0687	0.0175
2019	0.7603	0.1530	0.0689	0.0178
2020	0.7592	0.1525	0.0701	0.0182

Table 5 shows the latent class sizes. The first latent class of the latent class model with four latent classes, $Z1$, is very similar to the first latent class of the latent class model with two latent classes $Z1$. It is again the latent class for "neither Māori nor Pacific". In the model with four latent classes, the estimated class size of this latent class is, for 2013, 0.7605, whereas it was estimated as 0.7618 in the model with two latent classes.

In the model with two latent classes, the second latent class had estimated size 0.2382, whereas in the model discussed in this section, with four latent classes, this second latent class is now split over three latent classes, $Z2$, $Z3$, and $Z4$. As can be derived from the conditional probabilities of reported ethnicity given the latent class:

- the second latent class is for "Māori and not Pacific", having estimated class size 0.1540 in 2013
- the third is for "Pacific and not Māori", having estimated class size 0.0697 in 2013
- the fourth is for "both Māori and Pacific", having estimated latent class size 0.0158 in 2013.

[1] Several random starting values are used. The best fitting of these is provisionally the global maximum. Further random starting values are then generated and the model refitted; we did not detect any better fits than the provisional global maximum among these cases, so we interpret that it is indeed the global maximum.

Over the years the latent class sizes are stable, with minor fluctuations. The estimated conditional probabilities provide the meaning of the latent classes, and are split into two tables for readability: one for latent classes 1 and 2 (table 6a), and one for latent classes 3 and 4 (table 6b).

Table 6a

Estimated conditional probability of being in the four ethnic groupings given latent classes *Z*1 and *Z*2 from the four latent class model

Year	*Z*1_*a*1	*Z*1_*a*2	*Z*1_*a*3	*Z*1_*a*4	*Z*1_*b*1	*Z*1_*b*2	*Z*1_*b*3	*Z*1_*b*4
2013	0.9938	0.0052	0.0008	0.0001	0.9893	0.0090	0.0015	0.0002
2014	0.9938	0.0052	0.0008	0.0001	0.9895	0.0088	0.0015	0.0002
2015	0.9938	0.0053	0.0008	0.0001	0.9896	0.0087	0.0015	0.0002
2016	0.9937	0.0053	0.0008	0.0001	0.9898	0.0085	0.0015	0.0002
2017	0.9937	0.0053	0.0009	0.0001	0.9899	0.0084	0.0015	0.0002
2018	0.9909	0.0074	0.0014	0.0003	0.9902	0.0082	0.0015	0.0002
2019	0.9908	0.0075	0.0015	0.0003	0.9898	0.0085	0.0016	0.0002
2020	0.9907	0.0075	0.0015	0.0003	0.9893	0.0089	0.0016	0.0002
Year	*Z*1_*c*1	*Z*1_*c*2	*Z*1_*c*3	*Z*1_*c*4	*Z*1_*d*1	*Z*1_*d*2	*Z*1_*d*3	*Z*1_*d*4
2013	0.9922	0.0041	0.0037	0.0000	0.9813	0.0110	0.0076	0.0001
2014	0.9922	0.0040	0.0037	0.0000	0.9814	0.0109	0.0076	0.0001
2015	0.9922	0.0040	0.0037	0.0001	0.9814	0.0108	0.0076	0.0001
2016	0.9922	0.0040	0.0038	0.0001	0.9815	0.0107	0.0076	0.0001
2017	0.9923	0.0039	0.0038	0.0001	0.9816	0.0106	0.0076	0.0001
2018	0.9928	0.0034	0.0038	0.0001	0.9827	0.0095	0.0077	0.0001
2019	0.9929	0.0033	0.0037	0.0001	0.9829	0.0094	0.0076	0.0001
2020	0.9929	0.0033	0.0038	0.0000	0.9830	0.0093	0.0076	0.0001
Year	*Z*2_*a*1	*Z*2_*a*2	*Z*2_*a*3	*Z*2_*a*4	*Z*2_*b*1	*Z*2_*b*2	*Z*2_*b*3	*Z*2_*b*4
2013	0.0486	0.9464	0.0012	0.0039	0.0478	0.9448	0.0010	0.0064
2014	0.0489	0.9461	0.0012	0.0039	0.0471	0.9456	0.0010	0.0062
2015	0.0492	0.9458	0.0012	0.0039	0.0464	0.9464	0.0010	0.0061
2016	0.0494	0.9454	0.0012	0.0039	0.0456	0.9472	0.0011	0.0061
2017	0.0497	0.9452	0.0012	0.0040	0.0447	0.9483	0.0011	0.0060
2018	0.0486	0.9442	0.0015	0.0056	0.0492	0.9436	0.0011	0.0060
2019	0.0491	0.9438	0.0015	0.0056	0.0483	0.9443	0.0012	0.0062
2020	0.0496	0.9433	0.0015	0.0056	0.0474	0.9449	0.0012	0.0065
Year	*Z*2_*c*1	*Z*2_*c*2	*Z*2_*c*3	*Z*2_*c*4	*Z*2_*d*1	*Z*2_*d*2	*Z*2_*d*3	*Z*2_*d*4
2013	0.1122	0.8843	0.0011	0.0024	0.0587	0.9334	0.0008	0.0070
2014	0.1103	0.8862	0.0011	0.0024	0.0602	0.9319	0.0008	0.0071
2015	0.1085	0.8880	0.0011	0.0024	0.0612	0.9308	0.0009	0.0071
2016	0.1066	0.8899	0.0011	0.0024	0.0622	0.9297	0.0009	0.0072
2017	0.1051	0.8913	0.0011	0.0024	0.0633	0.9286	0.0009	0.0073
2018	0.1098	0.8868	0.0011	0.0023	0.0664	0.9260	0.0009	0.0068
2019	0.1111	0.8854	0.0011	0.0024	0.0669	0.9254	0.0009	0.0068
2020	0.1128	0.8836	0.0011	0.0024	0.0673	0.9251	0.0009	0.0067

Note: The column indicator *Z*1_*a*1 stands for P ($a = 1 \mid Z = 1$). The ethnic groupings in registers are denoted *a* in Census, *b* in DIA, *c* in MOH, and *d* in MOE.
Interpretation of ethnic groupings are 1 – neither Māori nor Pacific, 2 – Māori and not Pacific, 3 – Pacific and not Māori, and 4 – both Māori and Pacific.

Table 6b

Estimated conditional probability of being in the four ethnic groupings given latent classes *Z*3 and *Z*4 from the four latent class model

Year	*Z*3_*a*1	*Z*3_*a*2	*Z*3_*a*3	*Z*3_*a*4	*Z*3_*b*1	*Z*3_*b*2	*Z*3_*b*3	*Z*3_*b*4
2013	0.0607	0.0008	0.9351	0.0034	0.0249	0.0010	0.9660	0.0081
2014	0.0620	0.0008	0.9337	0.0035	0.0250	0.0010	0.9660	0.0080
2015	0.0630	0.0009	0.9326	0.0035	0.0253	0.0009	0.9657	0.0081
2016	0.0639	0.0009	0.9317	0.0036	0.0254	0.0009	0.9655	0.0082
2017	0.0645	0.0009	0.9310	0.0036	0.0255	0.0009	0.9652	0.0084
2018	0.0770	0.0022	0.9156	0.0051	0.0258	0.0007	0.9648	0.0086
2019	0.0786	0.0023	0.9139	0.0052	0.0254	0.0008	0.9648	0.0090
2020	0.0805	0.0025	0.9118	0.0051	0.0252	0.0008	0.9645	0.0096
Year	*Z*3_*c*1	*Z*3_*c*2	*Z*3_*c*3	*Z*3_*c*4	*Z*3_*d*1	*Z*3_*d*2	*Z*3_*d*3	*Z*3_*d*4
2013	0.0382	0.0082	0.9488	0.0047	0.0318	0.0042	0.9504	0.0136
2014	0.0377	0.0080	0.9497	0.0047	0.0328	0.0044	0.9492	0.0136
2015	0.0371	0.0078	0.9503	0.0048	0.0335	0.0046	0.9480	0.0138
2016	0.0364	0.0077	0.9511	0.0048	0.0341	0.0049	0.9471	0.0140
2017	0.0362	0.0076	0.9515	0.0047	0.0343	0.0049	0.9466	0.0142
2018	0.0370	0.0074	0.9509	0.0046	0.0345	0.0050	0.9469	0.0137
2019	0.0380	0.0074	0.9499	0.0047	0.0347	0.0049	0.9468	0.0136
2020	0.0385	0.0074	0.9494	0.0047	0.0345	0.0048	0.9473	0.0134
Year	*Z*4_*a*1	*Z*4_*a*2	*Z*4_*a*3	*Z*4_*a*4	*Z*4_*b*1	*Z*4_*b*2	*Z*4_*b*3	*Z*4_*b*4
2013	0.0116	0.0632	0.0981	0.8272	0.0063	0.0686	0.0808	0.8443
2014	0.0118	0.0635	0.0978	0.8269	0.0061	0.0666	0.0800	0.8473
2015	0.0119	0.0640	0.0974	0.8267	0.0058	0.0646	0.0782	0.8514
2016	0.0117	0.0642	0.0965	0.8275	0.0059	0.0627	0.0761	0.8553
2017	0.0117	0.0643	0.0965	0.8275	0.0055	0.0606	0.0739	0.8600
2018	0.0149	0.0655	0.0778	0.8419	0.0057	0.0645	0.0738	0.8560
2019	0.0160	0.0663	0.0783	0.8394	0.0053	0.0628	0.0729	0.8589
2020	0.0169	0.0671	0.0786	0.8375	0.0052	0.0620	0.0723	0.8604
Year	*Z*4_*c*1	*Z*4_*c*2	*Z*4_*c*3	*Z*4_*c*4	*Z*4_*d*1	*Z*4_*d*2	*Z*4_*d*3	*Z*4_*d*4
2013	0.0353	0.3246	0.1340	0.5060	0.0199	0.2348	0.0842	0.6611
2014	0.0340	0.3141	0.1310	0.5208	0.0205	0.2367	0.0858	0.6569
2015	0.0331	0.3044	0.1283	0.5342	0.0209	0.2390	0.0861	0.6541
2016	0.0322	0.2962	0.1251	0.5465	0.0213	0.2407	0.0864	0.6516
2017	0.0309	0.2880	0.1229	0.5583	0.0221	0.2381	0.0876	0.6522
2018	0.0318	0.2918	0.1220	0.5544	0.0223	0.2360	0.0876	0.6541
2019	0.0325	0.2941	0.1221	0.5514	0.0225	0.2337	0.0873	0.6565
2020	0.0330	0.2955	0.1238	0.5477	0.0225	0.2314	0.0872	0.6589

Note: The column indicator *Z*3_*a*1 stands for P (*a* = 1 | *Z* = 3). The ethnic groupings in registers are denoted *a* in Census, *b* in DIA, *c* in MOH, and *d* in MOE. Interpretation of ethnic groupings are 1 – neither Māori nor Pacific, 2 – Māori and not Pacific, 3 – Pacific and not Māori, and 4 – both Māori and Pacific.

The following discussion interprets the estimates for the 2013 Census. The estimates for other years are very similar:

- In the first latent class (table 6a), the estimated probability to report "neither Māori nor Pacific" in 2013 is 0.9938 (and the other estimated conditional probabilities in 2013 are very small as the four estimated conditional probabilities must sum to 1).
- In the second latent class, the estimated conditional probability to report "Māori and not Pacific" is largest, being 0.9464, with "neither Māori nor Pacific" being second largest with 0.0486.
- In the third latent class (table 6b), the estimated conditional probability to report "Pacific and not Māori" is largest, being 0.9351, with reporting "neither Māori nor Pacific" being second largest with 0.0607.

- In the fourth latent class, the estimated conditional probability to report "both Māori and Pacific" is largest, being 0.8272, with reporting "Māori and not Pacific", with 0.0632, and reporting "Pacific and not Māori", with 0.0981, both sizeable.

For DIA, MOH, and MOE, results are similar, with some notable differences for the fourth latent class for "both Māori and Pacific" (table 6b). To show this, we report the four estimated conditional probabilities of the four data sources census, DIA, MOH, and MOE as sequences:

- For the first latent class, "neither Māori nor Pacific", the estimated conditional probabilities over the four data sources are 0.9938, 0.9893, 0.9922, and 0.9813 – very high and stable estimates over the four data sources.
- For the second latent class, "Māori and not Pacific", the estimated conditional probabilities are 0.9464, 0.9448, 0.8843, and 0.9334, where the conditional probability of reporting "Māori and not Pacific" is a bit lower for MOH than for the other three data sources. At the same time, there is a relative increase in the conditional probability for "neither Māori nor Pacific", having a conditional probability of 0.1122.
- For the third latent class, "Pacific and not Māori" the estimated conditional probabilities are 0.9351, 0.9660, 0.9488, and 0.9504.
- For the fourth latent class, "both Māori and Pacific", the estimated conditional probabilities are 0.8272, 0.8443, 0.5060, and 0.6611, where the conditional probabilities for reporting "both Māori and Pacific" is lower for MOH and MOE. For MOH it is only 0.5060, where the estimated conditional probability for reporting "Māori and not Pacific" is 0.3246 and for reporting "Pacific and not Māori" is 0.1340. For MOE it is 0.6611, where the estimated conditional probability for reporting "Māori and not Pacific" is 0.2348 and for reporting "Pacific and not Māori" is 0.0842. This is an example where the fourth latent class has a more mixed pattern (figure 2).

It is helpful to be able to assess the fuzziness of the latent classes and we believe that the averages of the conditional probabilities reported in the bullets above are an easily interpretable way to achieve this. The averages of the conditional probabilities by latent class are shown in table 7, and they hardly change between years. They show that the first three latent classes are clearly defined, although within this the "Māori and not Pacific" class shows a greater chance of inconsistency. The fourth latent class, for "both Māori and Pacific" is more likely to show differences from the reported ethnicity, but still more than 70 percent of cases agree with the latent class, so the interpretation seems sound.

Table 7

Average probabilities (over the four data sources) that reported ethnicity is the same as the consensus ethnicity derived from the interpreted latent classes, 2013–2020

	$Z = 1$	$Z = 2$	$Z = 3$	$Z = 4$
2013	0.9892	0.9272	0.9501	0.7097
2014	0.9892	0.9275	0.9497	0.7130
2015	0.9893	0.9278	0.9492	0.7166
2016	0.9893	0.9281	0.9489	0.7202
2017	0.9894	0.9284	0.9486	0.7245
2018	0.9892	0.9252	0.9446	0.7266
2019	0.9891	0.9247	0.9439	0.7266
2020	0.9890	0.9242	0.9433	0.7261

Note: Probabilities are conditional on being in latent class *Z* and are averaged over the four data sources Census, DIA, MOH, and MOE. Interpreted ethnic groupings for the latent classes *Z* are 1 – neither Māori nor Pacific, 2 – Māori and not Pacific, 3 – Pacific and not Māori, and 4 – both Māori and Pacific.

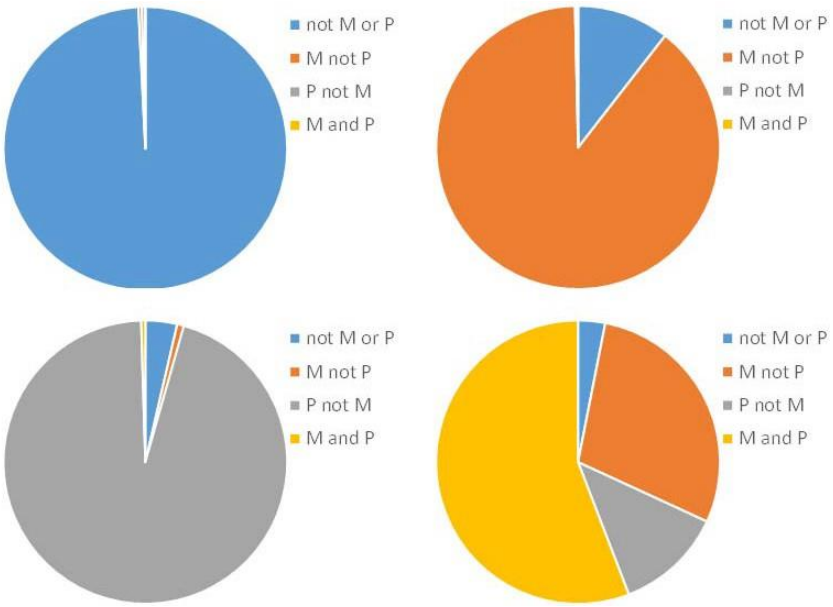
The proportions of the latent classes are very stable across years, which is a strong indication of the stability of the approach, since each year has been modelled and fitted separately.

We can see some small differences, the most consistent of which is the effect of changing to the new census in 2018, which adjusts the proportions in latent classes 1 and 2 back to values similar to the ones produced in 2013 at the previous census.

Latent class 3 ("Pacific and not Māori") seems to reach a minimum in 2018 and then increase faster than it had previously declined, while latent class 4 ("both Māori and Pacific") shows a steady increase across the whole time period.

Figure 2

Composition of four fitted latent classes according to ethnicity information observed in the 2017 MOH register



M = Māori ethnicity, P = Pacific ethnicity
Note: We interpret that the first latent class (top left) represents the "neither Māori nor Pacific" grouping; the second latent class (top right) represents mainly the "Māori and not Pacific" grouping; the third latent class (bottom left) represents mainly a "Pacific and not Māori" grouping; and the fourth latent class (bottom right) represents a mixture but mainly the "both Māori and Pacific" and "Māori but not Pacific" groupings.

5.1 Interpreting the quality of the registers as sources of ethnicity information

Using the probabilities of having different ethnicities in the contributing registers, given that records belong to the latent classes from the LCMSE of van der Heijden *et al* (2022), Smith *et al* (2021) deduce that the census has the best quality ethnicity information overall using the data for Māori and non-Māori ethnicity and the two latent class models. This is based on how similar the ethnicity information collected in the census is to the fitted latent classes, which are interpreted as the underlying true (or consensus) values.

We consider this specific notion of 'misclassification', the difference between the recorded value and the latent class estimate. We acknowledge that the recorded value is in some sense true for the ethnicity concept of the related source, but here we compare with a consensus estimate across the sources. The Smith *et al* (2021) analysis implicitly gave equal weight to the two latent classes.

In the analysis here with four latent classes, the assessment can be made in essentially the same way by examining the probability for the ethnic grouping recorded in each source to correspond with the value interpreted for the latent class to which it belongs (table 8).

Table 8

Probabilities that ethnic grouping as recorded in the four data sources corresponds with latent class containing those records, for 2013

	P(*a* match | *Z*)	P(*b* match | *Z*)	P(*c* match | *Z*)	P(*d* match | *Z*)
Z = 1	0.9938	0.9893	0.9922	0.9813
Z = 2	0.9464	0.9448	0.8843	0.9334
Z = 3	0.9351	0.9660	0.9488	0.9504
Z = 4	0.8272	0.8443	0.5060	0.6611
Mean	0.9256	0.9361	0.8328	0.8816
Weighted mean	0.9798	0.9785	0.9649	0.9667

Note: Probabilities are extracted from tables 6a and 6b. The ethnic groupings in registers are denoted *a* in Census, *b* in DIA, *c* in MOH, and *d* in MOE. Ethnic groupings are 1 – neither Māori nor Pacific, 2 – Māori and not Pacific, 3 – Pacific and not Māori, and 4 – both Māori and Pacific.

Taking a simple mean, which gives each ethnic grouping the same weight, we find that the ethnic grouping *b* in DIA (births) is the best, with the census information *a* close behind. DIA is not as good as census with the largest group, "neither Māori nor Pacific", but is better with the smaller populations. Note that DIA has missing ethnicity for many people, so there are many imputed values, which will be derived from the model, which in turn is based on the information in the other datasets. This may affect the comparison.

The MOH and MOE are poorer at identifying the "both Māori and Pacific" grouping. The MOH ethnic grouping shows greater differences for the "Māori and not Pacific" and "neither Māori nor Pacific" than the other sources. This reflects known differences in how the ethnicity information is gathered by MOH, such as ethnicity being assigned rather than self-identified, and much lower rates of multiple ethnicity identification compared with sources such as census (Harris *et al* 2022 and references within, Neuwelt *et al* 2014). These differences are not repeated for the "Pacific and not Māori" grouping, which seems to be recorded with similar quality in all the sources.

We can also consider a weighted mean (table 8) to show in which source the recorded ethnicity is most often right across all latent classes. This swaps the order of the first two sources, with the census marginally better than DIA, and with a smaller difference between these and the remaining sources which are less affected by the poor performance of the "both Māori and Pacific" latent class because of its small population size.

The probabilities of having a given ethnicity conditional on appearing in a latent class are so stable across the years (table 6a and table 6b) that this conclusion hardly changes. In 2018 the census (mean 0.9232) is not so close to DIA (0.9387) as it was in 2013, but the difference is quite small, and in the weighted mean the DIA is marginally better than census.

5.2 Summary of results

We conclude the following:

- The four latent classes do correspond with the ethnic groupings in the input data.
- The first latent class is the latent class for “neither Māori nor Pacific”. Given that one falls in this latent class, one seldom indicates having another ethnicity. For this latent class the results of the four-latent class solution are very similar to the results of the first latent class of the two-latent class solution.
- Compared with the two-latent class solution, in the four-latent class solution the remaining three latent classes now differentiate between “Māori and not Pacific”, “Pacific and not Māori”, and “both Māori and Pacific”.
- The latent classes for “Māori and not Pacific” and “Pacific and not Māori” can be interpreted from the conditional probabilities in a straightforward way.
- The fourth latent class, the latent class for “both Māori and Pacific”, is somewhat more difficult. For the census and DIA, the reported conditional probabilities indicate clearly that this is the latent class for “both Māori and Pacific”, but for MOH and MOE, the conditional probabilities to report “both Māori and Pacific” are only 0.5019 and 0.6549 respectively, with large conditional probabilities for “Māori and not Pacific” and “Pacific and not Māori”. So, even though the interpretation of the fourth latent class is not as straightforward as the interpretation of the first three latent classes, the difficulties in interpretation can be well understood.
- The estimates of the latent class sizes and conditional probabilities are very stable across the years.
- The DIA appears to have the highest ethnicity quality based on the four latent class model, closely followed by the census.

6 Population estimates

Table 9 provides population estimates based on the LCMSE model. These population estimates are adjusted for undercoverage, that is, for the individuals that are missed by all four data sources:

- The first column gives the number of individuals observed at least once in one of the four data sources in that year.
- The column "Missed" provides the estimated number missed by all four data sources.
- The next column provides population estimates for New Zealand.
- The final four columns are split between the ethnic groupings of this study: "neither Māori nor Pacific", "Māori and not Pacific", "Pacific and not Māori", and "both Māori and Pacific".

From 2013 to 2020 the total population of New Zealand grew 12 percent from approximately 4.4 million to 5.0 million. This growth of approximately 12 percent is also seen in the first three ethnic groupings. In the last and smallest ethnic grouping, "both Māori and Pacific", the percentage growth is much higher at approximately 30 percent.

Table 9

Population estimates from the four-class LCMSE models for 2013–2020

Year	Observed	Missed	Estimated population				
			Total	Neither Māori nor Pacific	Māori but not Pacific	Pacific but not Māori	Both Māori and Pacific
2013	4,410,624	23,842	4,434,466	3,372,460	682,880	309,207	69,918
2014	4,467,237	29,611	4,496,848	3,421,165	690,697	312,942	72,044
2015	4,542,948	41,490	4,584,438	3,491,021	700,634	318,107	74,676
2016	4,628,844	55,311	4,684,155	3,569,978	712,329	324,326	77,522
2017	4,712,850	62,140	4,774,990	3,641,984	723,643	328,871	80,492
2018	4,793,973	23,304	4,817,277	3,663,720	738,597	330,870	84,091
2019	4,864,143	25,206	4,889,349	3,717,482	747,844	337,013	87,010
2020	4,956,681	24,963	4,981,644	3,782,237	759,775	349,065	90,567

Note: Owing to rounding, the estimated ethnic grouping populations may not sum to the total estimated population.

For the 2018 Census year, the 30 June 2018 official estimates of the resident population[2] (ERP, Statistics New Zealand, 2020), equivalent to the groupings in table 9 and table 10 are:

- total population 4,900,600
- neither Māori nor Pacific 3,753,500
- Māori and not Pacific 739,400
- Pacific and not Māori 330,600
- both Māori and Pacific 77,100.

We notice a relatively large discrepancy for the total population size, that is estimated in table 9 as 4,817,277, and this discrepancy is largely due to the "neither Māori nor Pacific" grouping, that we estimate as 3,663,719.

Another source of ethnic group estimates is the experimental Administrative Population Census (APC), which is part of Stats NZ's census transformation programme looking at the potential for a future census based on admin data supported by sample surveys. For more details of the APC see

[2] The ERP is based on 2018 Census counts, together with a coverage survey, some dual system estimation, and some births/deaths and migration data.

Statistics New *Z*ealand (2022). The APC gives total responses for Māori and Pacific groupings, so Māori in APC should be compared with the sum of the groupings "Māori and not Pacific" and "both Māori and Pacific" (and similarly for the Pacific grouping).

From APC, Information by variable, Ethnicity, (Statistics New *Z*ealand, 2022, table 3):

- 2018 total responses Māori 787,317 in APC and in our estimates 822,685
- 2018 total responses Pacific 408,105 in APC and in our estimates 414,959.

Table 10 provides for each of the ethnic groupings in each year a 95 percent confidence interval estimated with the percentile method of the bootstrap. The confidence intervals show that the population estimates are very stable.

Table 10

Population estimates and 95 percent confidence intervals for the four ethnic groupings, 2013–2020

	Neither Māori nor Pacific			Māori and not Pacific		
Year	Population estimate	2.5%	97.5%	Population estimate	2.5%	97.5%
2013	3,372,459	3,370,208	3,374,841	682,879	681,291	684,507
2014	3,421,164	3,418,473	3,423,975	690,696	689,024	692,345
2015	3,491,020	3,487,220	3,494,867	700,633	698,993	702,329
2016	3,569,977	3,565,238	3,575,035	712,328	710,677	713,993
2017	3,641,983	3,636,279	3,648,153	723,642	721,874	725,364
2018	3,663,719	3,660,967	3,666,505	738,595	736,904	740,269
2019	3,717,481	3,714,678	3,720,525	747,843	746,221	749,489
2020	3,782,236	3,779,440	3,785,353	759,774	758,127	761,369
	Pacific and not Māori			Both Māori and Pacific		
Year	Population estimate	2.5%	97.5%	Population estimate	2.5%	97.5%
2013	309,206	307,732	310,780	69,918	69,380	70,474
2014	312,941	311,362	314,649	72,043	71,467	72,595
2015	318,106	316,413	320,024	74,676	74,067	75,292
2016	324,325	322,405	326,367	77,522	76,926	78,119
2017	328,870	327,132	330,963	80,491	79,918	81,124
2018	330,869	329,499	332,209	84,090	83,474	84,769
2019	337,012	335,794	338,275	87,009	86,376	87,680
2020	349,064	347,842	350,262	90,567	89,900	91,227

7 Estimates when the census is not used

We now study the estimates when the census is not used at all. Thus, we only make use of the three administrative registers DIA, MOH, and MOE. As it turns out, the latent class model becomes unidentified if we try to estimate the LCMSE with four latent classes (compare with section 5). However, we can fit the LCMSE with three and two latent classes. We focus on 2013 results, as the other years provide very similar estimates.

In the three-class LCMSE model estimates the first latent class is taken by the grouping of "neither Māori nor Pacific", see table 11 and table 12. There are large conditional probabilities of 0.9852, 0.9927, and 0.9807 for grouping "neither Māori nor Pacific" given one is in latent class 1, for DIA, MOH, and MOE respectively.

The second latent class is taken by the grouping of "Māori and not Pacific", as there are large conditional probabilities of 0.9204, 0.8871, and 0.9244 for grouping "Māori and not Pacific" given one is in latent class 2, for DIA, MOH, and MOE respectively.

The third latent class is a mix of the groupings that involve Pacific, that is "Pacific and not Māori" and "both Māori and Pacific". In latent class 3 the conditional probabilities for these two groupings are 0.8208 and 0.1560 in DIA, 0.8527 and 0.0819 in MOH, and 0.8442 and 0.1096 in MOE.

Therefore, in comparison to the four-class LCMSE where each grouping had its own class, in the three-class LCMSE "Pacific and not Māori" and "both Māori and Pacific" are grouped together. This is also evident from the comparison of the latent class sizes of the four-class solution in table 5 with the latent class sizes of the three-class solution in table 11: the probabilities in the three-class solution are very similar to the probabilities of the four-class solution, where in the three-class solution the latent class sizes of "Pacific and not Māori" and "both Māori and Pacific" are grouped together.

Not using the census gives estimates of Pacific from latent class 3, but with the major limitation that we cannot obtain estimates for the entire Māori grouping, which is comprised of latent class 2 and part of latent class 3. Therefore, an alternative fourth source would be needed in the absence of a census; a large-scale survey might be a suitable alternative, but this is a topic for future research.

Table 11

Estimated latent class sizes of LCMSE with three latent classes without the census

Year	Latent class 1	Latent class 2	Latent class 3
2013	0.7622	0.1573	0.0805
2014	0.7623	0.1571	0.0806
2015	0.7625	0.1568	0.0807
2016	0.7628	0.1564	0.0808
2017	0.7629	0.1561	0.0811
2018	0.7625	0.1560	0.0815
2019	0.7617	0.1562	0.0821
2020	0.7600	0.1563	0.0837

Table 12

Estimated conditional probability of being in three ethnic groupings given latent classes *Z*1–*Z*3 from the three latent class model without the census

Year	*Z*1_*b*1	*Z*1_*b*2	*Z*1_*b*3	*Z*1_*b*4	*Z*1_*c*1	*Z*1_*c*2	*Z*1_*c*3	*Z*1_*c*4	*Z*1_*d*1	*Z*1_*d*2	*Z*1_*d*3	*Z*1_*d*4
2013	0.9852	0.0127	0.0020	0.0001	0.9927	0.0044	0.0029	0.0000	0.9807	0.0129	0.0063	0.0001
2014	0.9856	0.0124	0.0019	0.0001	0.9928	0.0043	0.0030	0.0000	0.9807	0.0128	0.0064	0.0001
2015	0.9859	0.0121	0.0019	0.0001	0.9928	0.0042	0.0030	0.0000	0.9808	0.0127	0.0065	0.0001
2016	0.9862	0.0119	0.0019	0.0001	0.9929	0.0041	0.0030	0.0000	0.9809	0.0126	0.0065	0.0001
2017	0.9864	0.0117	0.0018	0.0001	0.9931	0.0040	0.0030	0.0000	0.9810	0.0124	0.0065	0.0001
2018	0.9863	0.0118	0.0018	0.0001	0.9931	0.0039	0.0030	0.0000	0.9812	0.0122	0.0065	0.0001
2019	0.9860	0.0120	0.0019	0.0001	0.9933	0.0038	0.0029	0.0000	0.9814	0.0120	0.0065	0.0001
2020	0.9856	0.0124	0.0019	0.0001	0.9933	0.0037	0.0029	0.0000	0.9817	0.0118	0.0064	0.0001
Year	*Z*2_*b*1	*Z*2_*b*2	*Z*2_*b*3	*Z*2_*b*4	*Z*2_*c*1	*Z*2_*c*2	*Z*2_*c*3	*Z*2_*c*4	*Z*2_*d*1	*Z*2_*d*2	*Z*2_*d*3	*Z*2_*d*4
2013	0.0334	0.9204	0.0023	0.0439	0.0981	0.8871	0.0014	0.0134	0.0469	0.9244	0.0010	0.0277
2014	0.0330	0.9214	0.0022	0.0433	0.0961	0.8887	0.0013	0.0139	0.0490	0.9228	0.0010	0.0272
2015	0.0326	0.9223	0.0022	0.0430	0.0943	0.8901	0.0013	0.0143	0.0504	0.9218	0.0011	0.0267
2016	0.0321	0.9231	0.0022	0.0427	0.0924	0.8916	0.0013	0.0147	0.0518	0.9208	0.0011	0.0263
2017	0.0315	0.9248	0.0021	0.0416	0.0912	0.8931	0.0012	0.0145	0.0534	0.9198	0.0010	0.0257
2018	0.0309	0.9253	0.0020	0.0418	0.0916	0.8929	0.0012	0.0143	0.0544	0.9188	0.0011	0.0257
2019	0.0302	0.9246	0.0020	0.0432	0.0936	0.8908	0.0012	0.0144	0.0551	0.9179	0.0011	0.0260
2020	0.0294	0.9239	0.0020	0.0446	0.0964	0.8879	0.0012	0.0145	0.0554	0.9173	0.0011	0.0262
Year	*Z*3_*b*1	*Z*3_*b*2	*Z*3_*b*3	*Z*3_*b*4	*Z*3_*c*1	*Z*3_*c*2	*Z*3_*c*3	*Z*3_*c*4	*Z*3_*d*1	*Z*3_*d*2	*Z*3_*d*3	*Z*3_*d*4
2013	0.0205	0.0026	0.8208	0.1560	0.0449	0.0205	0.8527	0.0819	0.0324	0.0138	0.8442	0.1096
2014	0.0206	0.0026	0.8166	0.1602	0.0437	0.0204	0.8502	0.0857	0.0335	0.0149	0.8396	0.1120
2015	0.0208	0.0026	0.8110	0.1656	0.0428	0.0204	0.8474	0.0894	0.0345	0.0160	0.8349	0.1147
2016	0.0209	0.0025	0.8055	0.1711	0.0418	0.0206	0.8445	0.0930	0.0353	0.0171	0.8304	0.1172
2017	0.0206	0.0025	0.7962	0.1807	0.0411	0.0212	0.8397	0.0980	0.0359	0.0183	0.8251	0.1207
2018	0.0204	0.0026	0.7910	0.1860	0.0417	0.0218	0.8363	0.1002	0.0361	0.0187	0.8222	0.1230
2019	0.0197	0.0026	0.7864	0.1913	0.0433	0.0223	0.8332	0.1011	0.0360	0.0186	0.8212	0.1242
2020	0.0193	0.0025	0.7817	0.1964	0.0449	0.0224	0.8318	0.1010	0.0356	0.0184	0.8220	0.1241

Note: The column indicator *Z*1_*b*1 stands for P (*b* = 1 | *Z* = 1). The ethnic groupings in registers are denoted *b* in DIA, *c* in MOH, and *d* in MOE. Interpretation of ethnic groupings are 1 – neither Māori nor Pacific, 2 – Māori and not Pacific, 3 – Pacific with or without Māori.

We also compare the two-class LCMSE solution without the census with the three-class LCMSE solution without the census. The estimates in table 13 and table 14 show that now the three ethnicity categories “Māori and not Pacific”, “Pacific and not Māori”, and “both Māori and Pacific” are grouped together in the second class.

We conclude the following:

- A latent class corresponding with “neither Māori nor Pacific” is clearly interpretable in each model, with almost the same probabilities.
- With three latent classes the “Māori and not Pacific” grouping is also in its own class, with the other two groupings together in the final class. With only two latent classes all the remaining ethnic groupings fall together in the second class.
- The probabilities are consistent with aggregating the classes from the four register and four latent class model, which suggests that essentially the same results can be derived for up to three latent classes without the census.

Table 13

Estimated latent class sizes of LCMSE with two latent classes without the census		
Year	Latent class 1	Latent class 2
2013	0.7621	0.2379
2014	0.7622	0.2378
2015	0.7624	0.2376
2016	0.7626	0.2374
2017	0.7626	0.2374
2018	0.7622	0.2378
2019	0.7614	0.2386
2020	0.7596	0.2404

Table 14

Estimated conditional probability of being in two ethnic groupings given latent classes *Z*1 and *Z*2 from the two latent class model without the census

Year	*Z*1_*b*1	*Z*1_*b*2	*Z*1_*b*3	*Z*1_*b*4	*Z*1_*c*1	*Z*1_*c*2	*Z*1_*c*3	*Z*1_*c*4	*Z*1_*d*1	*Z*1_*d*2	*Z*1_*d*3	*Z*1_*d*4
2013	0.9849	0.0132	0.0018	0.0001	0.9927	0.0046	0.0027	0.0000	0.9811	0.0140	0.0048	0.0001
2014	0.9853	0.0129	0.0017	0.0001	0.9927	0.0045	0.0027	0.0000	0.9812	0.0138	0.0049	0.0001
2015	0.9856	0.0126	0.0017	0.0001	0.9928	0.0044	0.0027	0.0000	0.9813	0.0136	0.0050	0.0001
2016	0.9859	0.0124	0.0017	0.0001	0.9929	0.0043	0.0027	0.0000	0.9813	0.0135	0.0051	0.0001
2017	0.9861	0.0122	0.0016	0.0001	0.9931	0.0042	0.0027	0.0000	0.9815	0.0133	0.0052	0.0001
2018	0.9860	0.0123	0.0016	0.0001	0.9932	0.0042	0.0027	0.0000	0.9817	0.0131	0.0051	0.0001
2019	0.9857	0.0126	0.0016	0.0001	0.9933	0.0041	0.0026	0.0000	0.9820	0.0129	0.0051	0.0001
2020	0.9852	0.0130	0.0017	0.0001	0.9934	0.0040	0.0026	0.0000	0.9822	0.0127	0.0051	0.0001
Year	*Z*2_*b*1	*Z*2_*b*2	*Z*2_*b*3	*Z*2_*b*4	*Z*2_*c*1	*Z*2_*c*2	*Z*2_*c*3	*Z*2_*c*4	*Z*2_*d*1	*Z*2_*d*2	*Z*2_*d*3	*Z*2_*d*4
2013	0.0048	0.0001	0.0295	0.6401	0.0813	0.5922	0.2900	0.0366	0.0427	0.6162	0.2858	0.0553
2014	0.0049	0.0001	0.0293	0.6407	0.0795	0.5928	0.2895	0.0382	0.0445	0.6149	0.2848	0.0558
2015	0.0050	0.0001	0.0291	0.6409	0.0780	0.5930	0.2893	0.0398	0.0458	0.6141	0.2836	0.0565
2016	0.0051	0.0001	0.0288	0.6413	0.0764	0.5931	0.2891	0.0413	0.0470	0.6130	0.2828	0.0572
2017	0.0052	0.0001	0.0283	0.6421	0.0753	0.5931	0.2886	0.0430	0.0483	0.6117	0.2820	0.0581
2018	0.0051	0.0001	0.0278	0.6424	0.0757	0.5918	0.2888	0.0437	0.0490	0.6101	0.2820	0.0590
2019	0.0051	0.0001	0.0270	0.6417	0.0775	0.5891	0.2892	0.0442	0.0494	0.6086	0.2825	0.0596
2020	0.0051	0.0001	0.0264	0.6402	0.0797	0.5835	0.2923	0.0445	0.0496	0.6056	0.2849	0.0599

Note: The column indicator *Z*1_*b*1 stands for P (*b* = 1 | *Z* = 1). The ethnic groupings in registers are denoted *b* in DIA, *c* in MOH, and *d* in MOE. Interpretation of ethnic groupings are 1 – neither Māori nor Pacific, 2 – Māori and/or Pacific.

8 Methodological discussion

In section 2 we saw that the records in IDI that belong to periods before digitisation of administrative records give rise to many cases with missing ethnicities, particularly in DIA. In van der Heijden *et al* (2022) these records were excluded, leading to a need for methods to assess whether the partial coverage of some of the register sources affected the population size estimates from the MSE. In the current research, we do not have to concern ourselves with partial coverage, but instead use the same methods as van der Heijden *et al* (2022) to assign estimated ethnicities to cases where the ethnicity is missing.

The LCMSE model is used both to impute ethnicities in observed cases where ethnicity is missing, and to estimate the number of unobserved cases by ethnicity, in both cases according to the pattern of ethnicity in the observed cases.

If we take the LCMSE model structure and parameters as fixed, then it should make no difference whether a case is missing or present with missing ethnicity − the estimates should be the same. We therefore tried changing all the cases with missing ethnicity into missing cases and fitting the same models.

Although the results are qualitatively the same, there are some quantitative differences (results not shown), which suggests that the parameter estimates are affected by this process. We need to do more work to understand why this happens. This is closely linked to the work on invariant population size estimation and partial coverage reported in van der Heijden *et al* (2022).

9 Conclusions

As indicated in the Introduction, the aim of our research was to extend analyses presented earlier in van der Heijden *et al* (2022).

The aim was to extend this approach in two ways:

- by examining how the population size estimates and the latent class model estimates evolve over time
- by extending the ethnicity analysis to more categories.

It was already well known that the population of New Zealand grew over 2013–2020, and this is reflected in both the input datasets (derived from the population census and administrative sources) and the population size estimates that account for undercoverage. The latent class model estimates presented here are remarkably stable over time, with only small changes in the proportion of the population in each class.

Also, the extension of the ethnic groupings from two to four was successful. From the latent class conditional probabilities it is evident that the four ethnic groupings nicely fall in separate latent classes, where the misclassifications (as defined in section 5.1) for the "neither Māori nor Pacific" are very small, and the misclassifications for the other ethnic groupings are a bit larger.

Future work could include extending the ethnic groupings even further, and including gender, age, and region.

In principle, an additional source is needed for each new ethnicity in the latent class model, but it may be possible to extend the models with additional covariates instead of additional sources, and/or to restrict the parameters to make the models identifiable. This is a topic for future research.

References

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*(2):215–231.

Harris, R. B., Paine, S., Atkinson, J., Robson, B., King, P. T., Randle, J., Mizdrak, A., and McLeod, M. (2022). We still don't count: the under-counting and under-representation of Māori in health and disability sector data. *New Zealand Medical Journal*, 135:54–78.

Neuwelt, P., Crengle, S., Cormack, D., McLeod, M., and Bramley, D. (2014). General practice ethnicity data: evaluation of a tool. *Journal of Primary Health Care*, 6:49−55.

Smith, P. A., van der Heijden, P. G., and Cruyff, M. (2021). Measurement errors in multiple systems estimation. In Porzio, G. C., Rampichini, C., and Bocci, C., (Eds.), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group, Firenze, September 9-11 2021,* (pp. 211–214). Firenze University Press.

Stats NZ (2017). Experimental population estimates from linked administrative data: 2017 release. Available from www.stats.govt.nz.

Stats NZ (2020). Estimated resident population 2018: Data sources and methods. Available from www.stats.govt.nz.

Stats NZ (2022). Experimental administrative population census: Data sources, methods, and quality (second iteration). Available from www.stats.govt.nz.

van der Heijden, P. G., Cruyff, M., Smith, P. A., Bycroft, C., Graham, P., and Matheson-Dunning, N. (2022). Multiple system estimation using covariates having missing values and measurement error: Estimating the size of the Māori population in New Zealand. *Journal of the Royal Statistical Society: Series A*, 185:156–177.

Appendix – Evaluation of rounding impact

In this appendix we investigate the impact of the rounding process on the final estimates. As seen in section 2, the data outputs from the IDI have been through disclosure control:

- values smaller than six are suppressed and cannot be distinguished from actual zeros
- values greater than or equal to six are randomly rounded such that:
    - values that are already multiples of three are left unchanged
    - values that are not multiples of three are randomly rounded to the nearest multiple of three with a probability of ⅔ and to the second nearest multiple of three with a probability of ⅓.

Our evaluation of the rounding impact here focuses on the model introduced in section 5, that is, a LCMSE with four latent classes, four lists, and four ethnic groupings.

1. We evaluated the impact of the rounding via Monte Carlo simulation, where at each replicate we generate a "new" dataset from the original dataset of 2013 (we believe the analysis holds regardless of the year used) by means of reversing the rounding process just described, and then the model is fitted.
2. We then looked at the Monte Carlo sampling distributions of the coefficients to evaluate the impact of the rounding process. Figure 3 plots the coefficients β obtained by the original model against the Monte Carlo expected values of the coefficients under the rounding process. The points are scattered around the 45-degree line, showing that the Monte Carlo sampling distributions are indeed centred around the coefficients obtained by the original model.
3. To assess the variability, we observed:

    (i) the boxplots of the sampling distributions of the coefficients in figure 4, which outlines the presence of some "outliers" for the higher-order interactions

    (ii) the entire sampling distributions for the main effects of the model in figure 5, which shows the presence of some limited variability around the model coefficients (the interaction terms have similar behaviour and are omitted here for brevity).

In conclusion, the results suggest that the impact of the rounding process is limited. Indeed, the Monte Carlo distributions are centred around the original coefficients obtained by the model fitted on the original dataset. There is a little variability, which was expected due to the complexity of the model and the presence of a large number of small cells, but overall, the results are reassuring. Therefore, no additional step is required in analyses at this level to address the impact of disclosure control on the conclusions.

Figure 3

Coefficients β obtained by the original model plotted against the Monte Carlo expected values of coefficients, $E[\beta_{MC}]$, under the rounding process

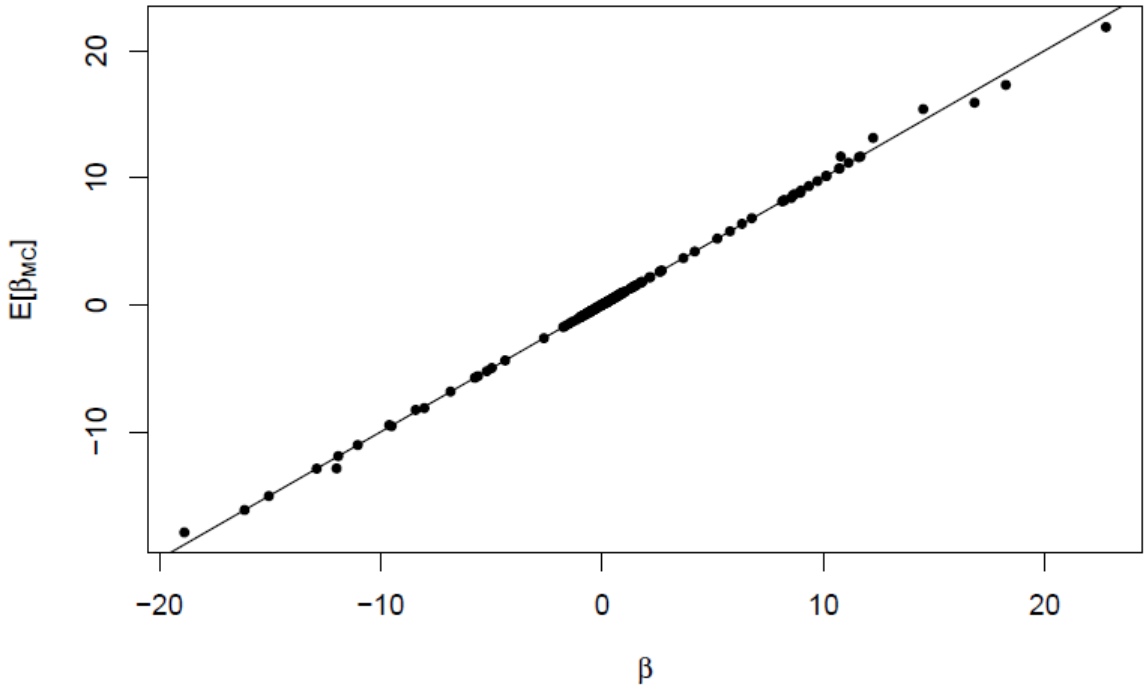


Figure 4

Boxplots for the sampling distributions of the coefficients, ordered from smallest to largest

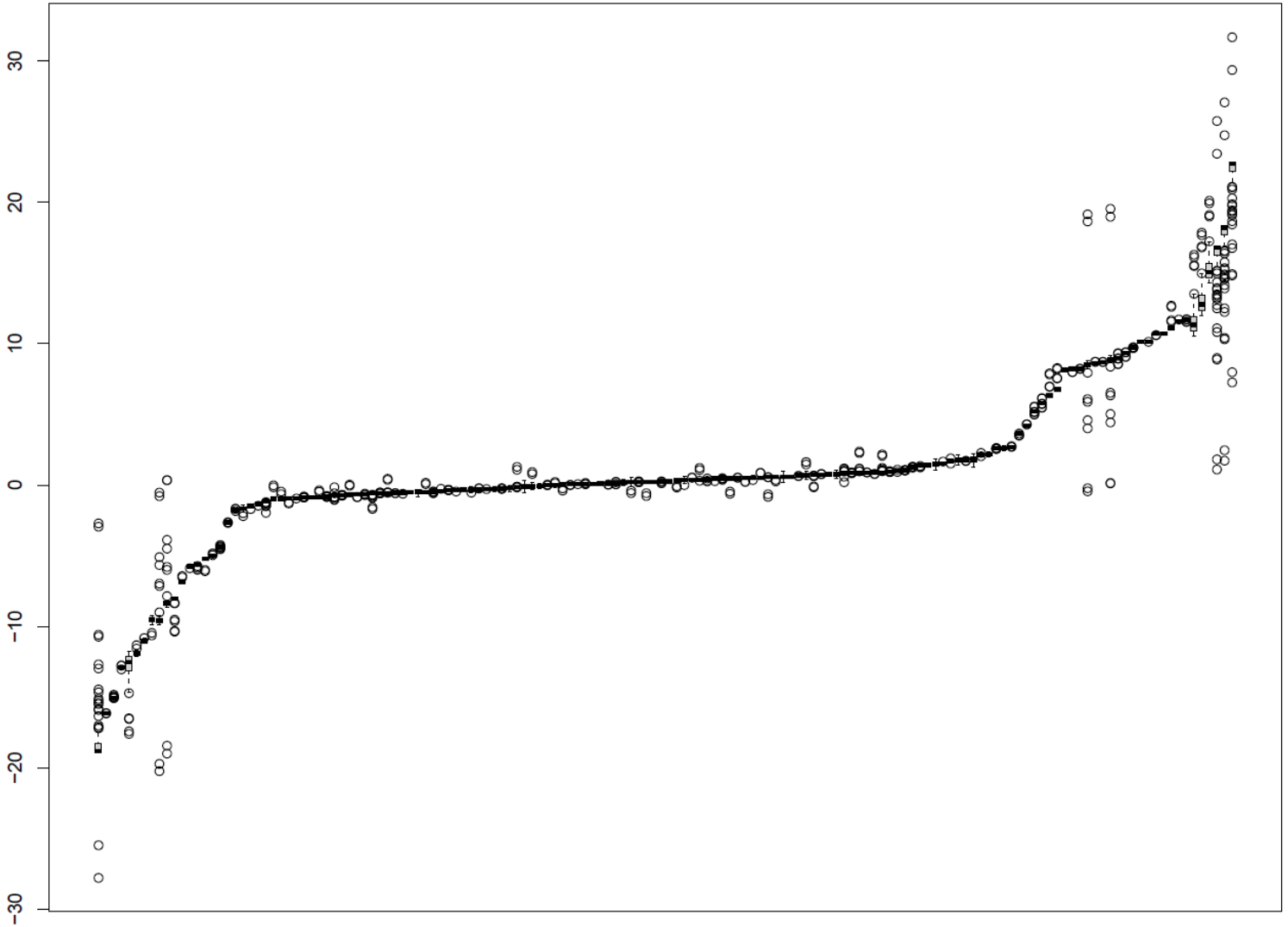


Figure 5

Monte Carlo distributions for the main effects of the model