Harmonising ethnicity from multiple sources using latent class models

Christine Bycroft, Joane Elleouet, & Hop Tran

christine.bycroft@stats.govt.nz

PopCon2023, Auckland

What is ethnicity?

Ethnicity is a measure of cultural identity in New Zealand and a key social factor used to describe the population.

Ethnicity is collected by multiple agencies

- **Stats NZ: the census** provides the basis for official estimates of the resident population by ethnicity (ERP)
- **Other agencies**: DIA, MoH, MoE, MSD

All these data sources are linked together in the IDI

An individual will often be in multiple sources, and sometimes have conflicting values of their ethnicity.

"Ethnicity is the ethnic group or groups that people identify with or feel they belong to. Ethnicity is a measure of cultural affiliation, as opposed to race, ancestry, nationality, or citizenship. Ethnicity is self-perceived and people can belong to more than one ethnic group."

Statistical standard for ethnicity

Collecting and using ethnicity data is challenging

Legitimate differences in recorded ethnicity between collections because of different contexts and changes in people's perception of their ethnicity over time.

Multiple responses: people can belong to more than one ethnic group

Respondent or other collection and processing errors: in admin sources, but also in census and surveys

180 ethnicities, a hierarchical classification

Level 1 ethnic groups

- European
- Māori
- Pacific
- Asian
- Middle Eastern, Latin American, African (MELAA)
- Other

We need a harmonised view of ethnicity in the IDI

For researchers: many social science applications use ethnicity as a factor in their analysis

For official statistics

- Annual Māori population estimates use IDI ethnicity in external migration
- The census uses IDI ethnicity for non-response

For experimental products

- The administrative population census (APC) derives an admin resident population by ethnicity from admin sources in the IDI

The IDI provides ethnicity for every individual using a deterministic method: '**source ranking**' .

1. Rank sources by their quality (how close to census)
2. Use the highest ranked source available

Does not use all information available for an individual

We have looked at a modelling approach:

Latent class analysis

Motivated by latent class MSE (van der Heijden et al)

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.

Aims and scope

Exploratory work, to promote discussion

Aim: investigate a statistical modelling approach - latent class analysis (LCA) - to predict an individual's level 1 ethnic group from multiple administrative data sources in the IDI.

1. Can latent class modelling be used to predict individual ethnicity in IDI data?
2. What is the consistency between the ethnicity predicted from latent class models and ethnicity from the 2018 Census? Does it improve on the source ranking method?
3. What are the implications of these findings for the use of administrative data to report ethnicity?

Is the census really the highest quality source?

The problem

person	Birth registration child	MoE Tertiary enrolment	Min Health	MoE School enrolment	MSD	2018 Census
A	…	…	European	…	European	European
B	Māori, European	…	European	Māori	…	Māori, European
C	Māori	Pacific, Māori	Pacific, Māori	Pacific	…	Pacific, Māori

Latent class models

A latent class model uses the observed pattern of source responses to classify records into their unobserved underlying latent class.

- Observed values are independent, conditional on the latent class

Model for two latent classes, J sources

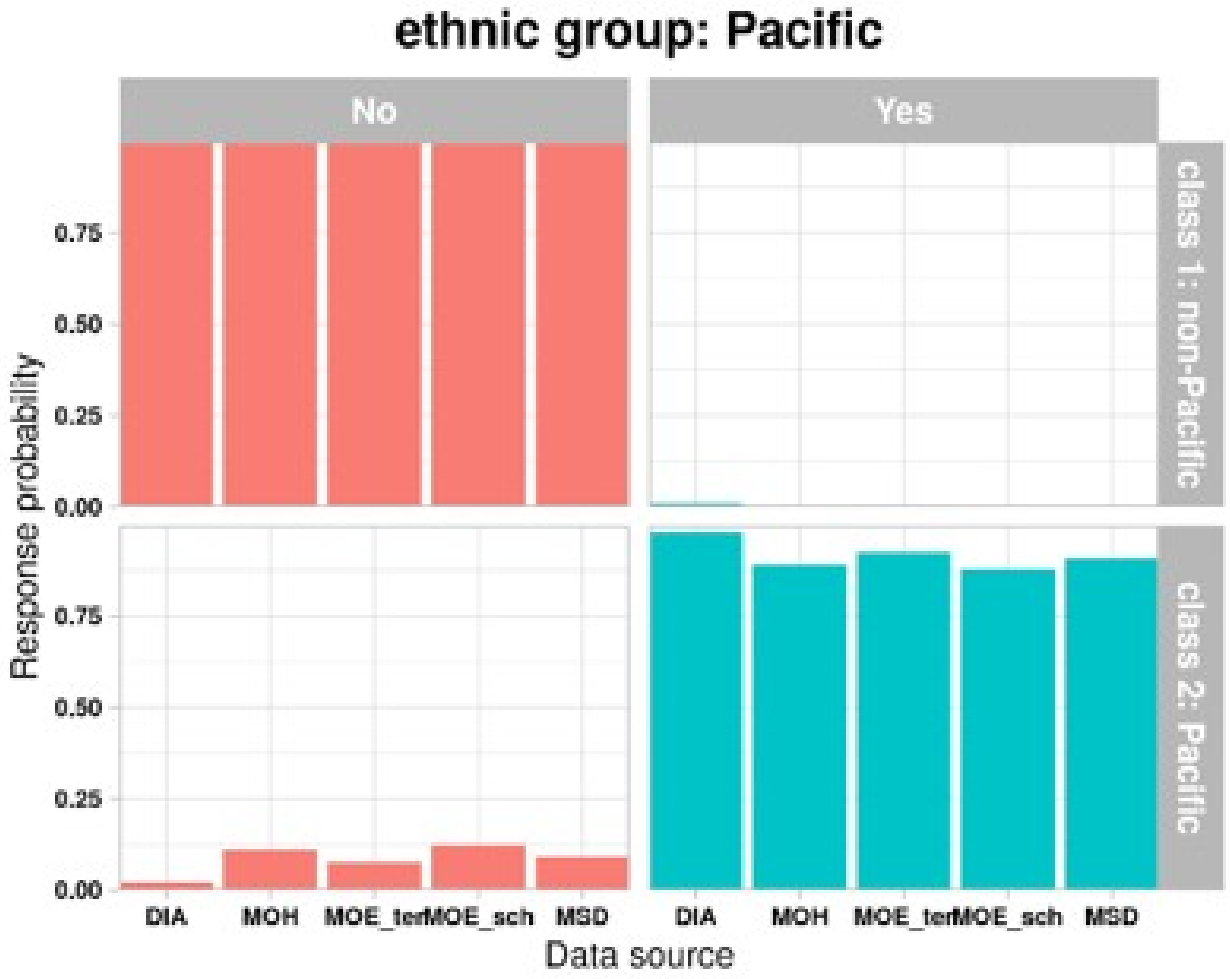
is the vector of probabilities for belonging to the ethnic group in list given latent class,
is the overall rate of class membership.

Latent class models ethnicity application

- Six simple latent class models, one for each level 1 ethnic group separately (European, Māori, Pacific, Asian, MELAA and Other).
- Each level 1 ethnic group is a binary variable (e g Pacific, or not Pacific)
- Set the number of latent classes as two.
- Interpret model classes as ‘belongs’, or ‘does not belong’ to the ethnic group
- Model is run in R poLCA package. Fitted using EM Algorithm.

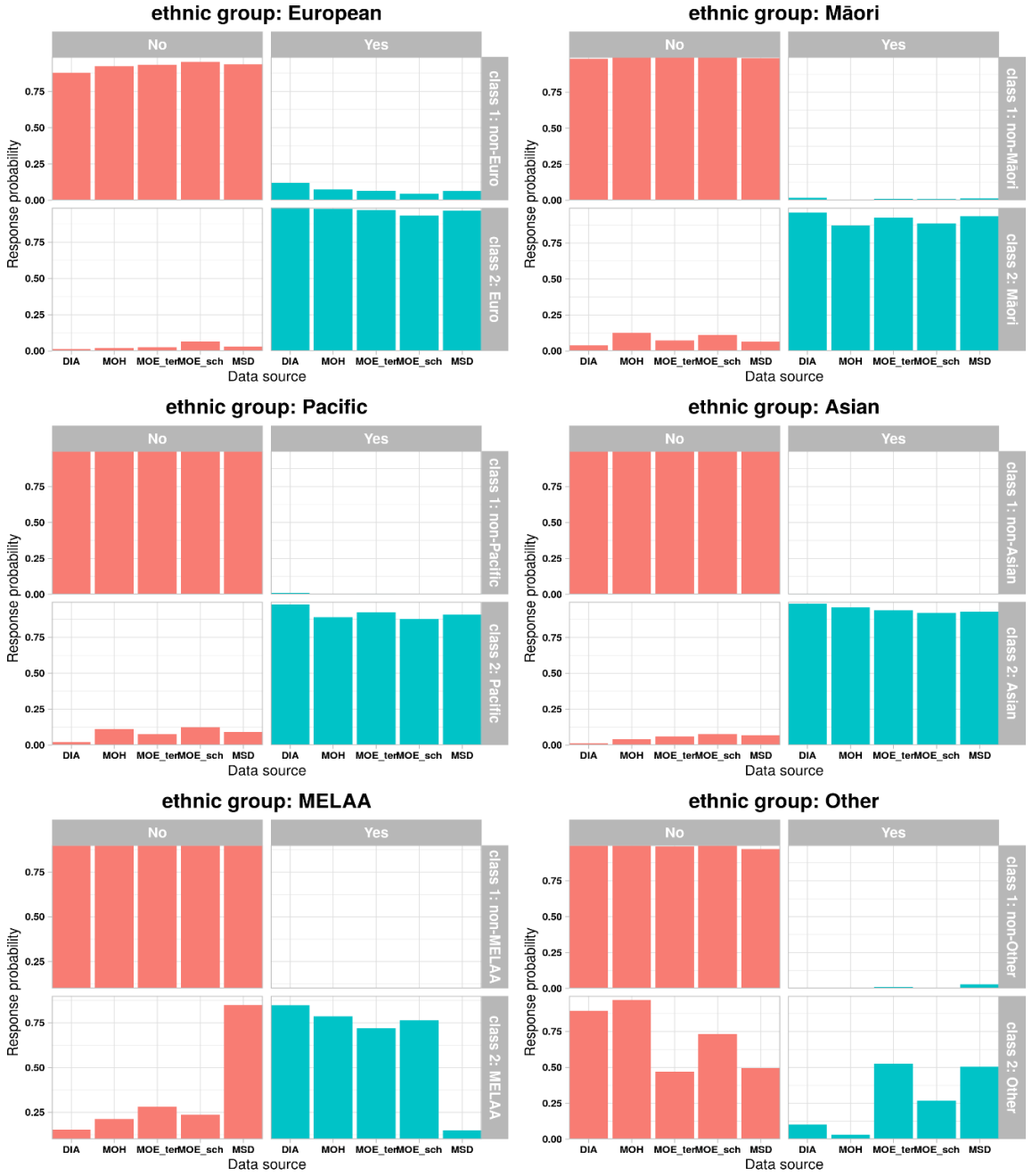
Validating the model

Conditional probabilities

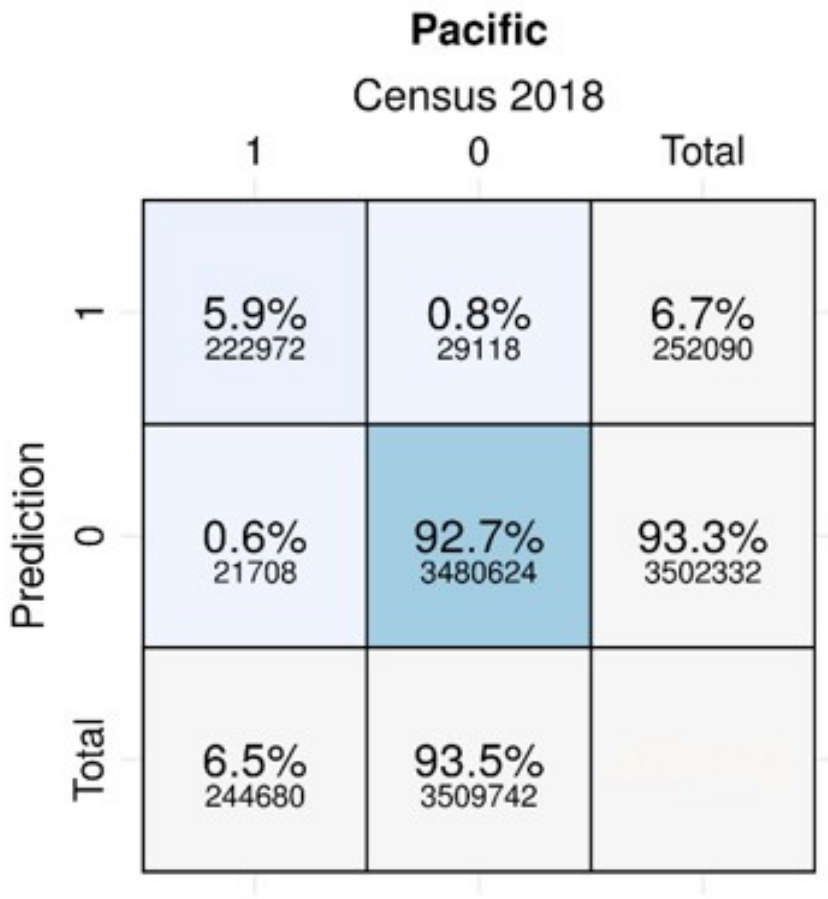


The probability of a ‘Yes’ response given latent class 2 is very high, and the probability of a No response, given latent class 2 is low.

Validating the model



Comparing with 2018 Census



The confusion matrix.
2018 Census questionnaire responses are taken as the best proxy for the ‘true’ ethnicity

Comparing with 2018 Census

Sensitivity of ethnic group classification for source ranking and LCM, with and without 2013 Census			
Ethnic group	Source ranking	LCM admin only	LCM with 2013 Census
European	0.955	0.947	0.96
Māori	0.864	0.884	0.916
Pacific	0.906	0.911	0.938
Asian	0.928	0.953	0.968
MELAA	0.72	0.792	0.846

Sensitivity

2018 Census within full LCM

- The best LC model is one that uses all sources, including 2018 Census.
- 2018 Census is the highest quality source
- 2018 Census vs model may indicate a level of ‘natural variation’ in reporting ethnicity

Sensitivity of sources in relation to best model predictions

Ethnic group	DIA	MOH	MOE tertiary	MOE school	MSD	2013 Census	2018 Census
European	0.985	0.974	0.958	0.916	0.953	0.977	0.985
Māori	0.940	0.863	0.907	0.862	0.918	0.941	0.952
Pacific	0.966	0.887	0.912	0.854	0.902	0.943	0.934
Asian	0.975	0.945	0.919	0.897	0.913	0.974	0.980
MELAA	0.846	0.837	0.673	0.732	0.147	0.822	0.840

Sensitivity

Comparing with ERP

Estimated ethnic proportions in the admin resident population for LCA and source ranking, compared to official ethnic population estimates			
Ethnic group	Source ranking (% of admin population)	LCM (% of admin population)	ERP (% of estimated total population)
European or Other	69.8	68.0	70.2
Māori	16.2	16.5	16.7
Pacific	8.5	8.5	8.3
Asian	14.7	15.0	15.7
MELAA	1.5	1.7	1.6

Estimates for NZ usual resident population

Discussion

- Latent class models are a valid approach to determine underlying ethnicity from observed values
- The deterministic source ranking method does work pretty well
- LCM gives better results for some individuals
- LCM confirms that the 2018 Census is the best source among those we used.
- LCM advantages
  - Uses all the information available
  - Provides misclassification measures for all contributing sources.
  - Flexible: easy to include other data sources (e g birth registration parents)
  - Flexible: the models accommodate varying quality among data sources for different ethnicities
  - Potential
    - improve model with addition of covariates (e g age, Māori descent, birthplace)
    - use for level 2 or level 3 ethnicity
    - include time in the model for inter-ethnic mobility

Bycroft, C, Elleouet, J, and Tran, H (2023). *Harmonising ethnicity from multiple administrative data sources using latent class modelling. Retrieved* from www.stats.govt.nz.