

Small Domain Estimation of Census Coverage – A Case Study in Bayesian Analysis of Complex Survey Data

*Joane S. Elleouet¹, Patrick Graham¹, Nikolai Kondratev¹, Abby K. Morgan¹, and
Rebecca M. Green¹*

Many countries conduct a full census survey to report official population statistics. As no census survey ever achieves 100% response rate, a post-enumeration survey (PES) is usually conducted and analysed to assess census coverage and produce official population estimates by geographic area and demographic attributes. Considering the usually small size of PES, direct estimation at the desired level of disaggregation is not feasible. Design-based estimation with sampling weight adjustment is a commonly used method but is difficult to implement when survey nonresponse patterns cannot be fully documented and population benchmarks are not available. We overcome these limitations with a fully model-based Bayesian approach applied to the New Zealand PES. Although theory for the Bayesian treatment of complex surveys has been described, published applications of individual level Bayesian models for complex survey data remain scarce. We provide such an application through a case study of the 2018 census and PES surveys. We implement a multilevel model that accounts for the complex design of PES. We then illustrate how mixed posterior predictive checking and cross-validation can assist with model building and model selection. Finally, we discuss potential methodological improvements to the model and potential solutions to mitigate dependence between the two surveys.

1. Introduction

In Aotearoa New Zealand a census is conducted every five years. It is a key input to official population estimates and supports a wide range of social and demographic analyses. Although the census would ideally count all people and their attributes of interest in the country at a given time, it inevitably fails to enumerate the full population. Censuses are expensive undertakings and the performance of the census in enumerating the population is therefore a matter of public interest. Consequently, a post-censal survey (the post-enumeration survey, henceforth PES) is conducted to evaluate the population coverage of the census. As well as providing an evaluation of the census, coverage estimates of the New Zealand census are used to adjust census counts in order to produce population estimates in the form of an estimated resident population (ERP) which is highly disaggregated to geographic and demographic groups. The population estimation system run by Statistics New Zealand (Stats NZ), New Zealand's official statistics agency, therefore requires coverage adjustments at a high level of granularity defined by, at least,

¹Statistical Methods Unit, Stats NZ 8 Willis Street, PO Box 2922, Wellington 6011, New Zealand. Emails: joane.elleouet@gmail.com, patrick.graham.br@gmail.com, nikolai.kondratev@stats.govt.nz, abby.morgan@stats.govt.nz and rebecca.green@stats.govt.nz

Acknowledgments: The views expressed in this article are those of the authors and should not be taken to represent an official view of their affiliated organisation. We thank three anonymous referees and an Associate Editor for thoughtful comments.

combinations of age in single year intervals, sex, ethnicity, 88 local government areas, Māori descent and country of birth (New Zealand or other). As these variables can potentially form hundreds of thousands of observed domains and the PES sample is made of approximately 30,000 individual records, direct estimation meets crucial limitations and the problem is best viewed as a modelling problem in which the objective is to relate the coverage probability to the covariates of interest. We therefore propose a Bayesian multilevel modelling approach to census coverage estimation. Although the official 2018 population estimates were created using a similar method ([Stats NZ 2020b](#)), the data and models used here differ from those used for the official published census coverage estimates and should not be regarded as official statistics.

Many countries with traditional census collections run a post-census coverage survey. Published applications include [Brown et al. \(2019\)](#), [Chipperfield et al. \(2017\)](#), [Hogan \(1993\)](#), [Mule et al. \(2008\)](#), with coverage estimation methods ranging from adaptations of dual systems estimation using a variant of the well-known Lincoln-Petersen estimator ([Brown et al. 2019](#)) to logistic regression of census coverage ([Mule et al. 2008](#); [Chen et al. 2010](#)) followed by inverse coverage probability weighting of the census file to obtain population estimates. Our methods resemble the latter approach, though we use multilevel logistic models to obtain coverage and population estimates at a high level of granularity. Hierarchical Bayes models have also been proposed for estimation of the coverage of the Canadian census ([You and Dick 2004](#)). However, these are area-level models, in contrast to the individual level models discussed in this article. [Elliott and Little \(2000\)](#) developed a Bayesian model for census coverage estimation that incorporates information on population sex ratios, in addition to data from the census and a census coverage survey. However, the data structure assumed in that work differs from the one available for the current analysis.

Modelling complex survey data at the individual level requires attending to the impact of the survey design and nonresponse on inclusion in the data. Whereas the design-based approach to survey inference achieves this through the use of survey weights and variance calculations that respect the survey design, the model-based approach accommodates the impact of survey design and nonresponse on inclusion in the observed data within the model structure. The latter approach is often accompanied by the application of model-derived estimates to benchmark population data to obtain small domain estimates that account for differences in covariate structure between the sample and the target population, as illustrated by the so-called MRP (Multilevel Regression and Post-stratification) method ([Gelman and Little 1997](#); [Lax and Phillips 2009](#); [Si et al. 2020](#)). We cannot use population benchmarks to aid coverage estimation from PES because one of the purposes of PES is to adjust the census data to produce new population benchmarks. Nevertheless, the application of highly disaggregated model-derived estimates from PES to the census to produce estimates of the usually resident population has some parallels with the MRP approach to estimation.

Although the general Bayesian approach to analysis of complex survey data has been well described ([Rubin 1987](#), chap. 2; [Little 2003](#); [Gelman et al. 2014](#), chap. 8), published applications of individual level Bayesian analyses of complex sample surveys remain relatively rare. Some recent applications, unrelated to census coverage, include small area official statistics ([Nandram et al. 2018](#)), political sciences ([Ghitza and Gelman 2013](#);

Shirley and Gelman 2015), and public health (Paige et al. 2020), the latter using simulations to compare design-based to model-based approaches. Bayesian methods, and particularly multilevel Bayesian models, have more commonly been applied to area-level modelling of complex survey data for small domain estimation. In such applications, summary direct estimates with an associated variance estimate are first computed for each area and/or group of interest. Multilevel Bayesian models are then applied to smooth the summary statistics. In the case of complex survey data the direct estimates and variance estimates computed as the first stage of this procedure are usually design-based estimates. Examples include Ghosh et al. (1998), You and Chapman (2006), Molina et al. (2014), and Chen et al. (2014). Reviews of the general approach can be found in Pfeffermann (2013, 45–47) and Rao and Molina (2014, chap. 10). In this approach, design-based estimation is used to deal with the analytical complications of complex sample surveys, freeing the multilevel Bayesian modelling from the requirement to explicitly deal with the survey design.

Application of the area-level approach is problematic in our context where the number of covariate combinations (or domains) exceeds the number of records in the survey data set, so that forming the initial set of domain-level summary statistics is not even possible. Even applying the area-level approach to an aggregated version of the cross-classification of co-variables for which estimates are ultimately required, such that each covariate combination in the aggregated cross-classification occurs in PES, would be difficult unless the degree of coarsening is substantial. In sparse data situations with a binary outcome, conventional design-based variance estimates of proportions can often be zero and this makes subsequent modelling difficult. Consequently, framing the problem as estimation from a model fitted at the level of individual records and from which predictions can then be made seems a logical way forward. However, accounting for a complex survey design complicates the model so that the model fitted to the data is more complex than required for prediction. We illustrate how the model of interest can be, implicitly, recovered from the fitted model by integrating out parameters associated with the survey design but not relevant to the predictions. This article illustrates the potential of Bayesian modelling of complex survey data for challenging small domain estimation problems.

To describe our approach, we first describe the PES design in Section 2. We then present our modelling strategy in Section 3, including model-checking and evaluation. In Section 4, we show results of the model checking procedure. We also include summaries of standardised coverage estimates, by area and by age and ethnic group. The standardisation is achieved by applying the modelled coverage estimates for each group to a common reference population. The reference population used for this estimation is the population estimated by adjusting the census file for under-coverage using the disaggregated coverage estimates obtained from the model. Uncertainty in the estimation of the reference population is automatically incorporated in the posterior distribution for the standardised estimates. Section 5 concludes the article with some discussion of the modelling issues and suggestions for further development.

2. PES and Census Data

The official 2018 census data set comprises census respondents as well as records obtained from administrative sources (Stats NZ 2019). It is subject to under-coverage (eligible

residents missed by the census) and over-coverage (non-eligible individuals mistakenly counted, such as births after the census date and residents temporarily overseas at the time of census). In the 2013 census, over-coverage was approximately 0.7%, in contrast to an under-coverage measure of approximately 3.1% ([Stats NZ 2014](#)). The official ERP is corrected for both types of errors estimated on the full census file ([Stats NZ 2020b](#)). Estimates presented here differ from previously published estimates of the population coverage of the official census file and should not be regarded as official statistics. The main differences with the methodology used for official statistics is that we focus on under-coverage probability estimation and we perform the estimation on the respondent subset of the census file, which excludes administrative enumerations. However, the estimation challenge we describe is similar to the one faced in constructing the official 2018 ERP. Estimates for under-coverage probabilities hold without having to make assumptions about levels of over-coverage. [Stats NZ \(2020b\)](#) addresses over-coverage estimation in a very similar manner to under-coverage, and we refer the reader to this publication for more details on over-coverage estimation.

The 2018 PES used an area-based, stratified two-stage design. For sampling purposes New Zealand was divided into 23,174 small geographic areas (Primary Sampling Units, PSUs) that were grouped into 101 strata, based on a combination of broad geographic region, major urban status, census delivery mode (whether an access code for the online census form was mailed out or a hard copy census form was delivered) and a measure of deprivation. The PSUs were selected using probability proportional to size (PPS), where the size measure was based on historical estimates and included an adjustment for ethnic group proportions. Sampling fractions varied by strata, with urban strata sampled more intensively than non-urban strata, for fieldwork efficiency reasons. PES operated in all strata and a total of 1,365 PSUs were selected for the PES sample. In most PSUs, 11 dwellings were sampled within each PSU using Stats NZ's standard approach in which dwellings within a PSU are grouped into panels of size 11, and one panel is randomly selected. This resulted in a sample of 15,213 households within 15,015 dwellings in the 1,365 selected PSUs. Dwellings refer to the building in which people live, whereas people residing together and sharing facilities within a dwelling constitute a household, and there can be multiple households per dwelling. All usual residents at selected households were eligible for inclusion in the sample. Henceforth, we refer to households and use the terms household effects and household variables when referring to both household and dwelling characteristics. Within the 15,213 visited households, 37,548 people were interviewed. After filtering for refusal, incomplete responses and ineligibility, the final sample included 12,459 households and 31,600 respondents with responses of sufficient quality to be linked to the census and included in the estimation.

The PES sample was linked to the census file using a conservative probabilistic linking methodology, followed by clerical checking of all non-linked records and a sample of linked records. Details are described in [Stats NZ \(2020b\)](#). Of all eligible PES person records, 30,397 were linked to a census record (1,300 through manual linking), and the remaining 1,203 PES respondents were not linked to any census record. PES respondents linked to a record in the census respondent file were considered covered by census, whereas unlinked PES records constitute instances of under-coverage.

3. Census Coverage Estimation Under a Bayesian Modelling Framework

3.1. The Bayesian Approach

Fully model-based analysis of complex survey data usually requires multilevel models in order to account for the survey design. Such modelling fits neatly into a Bayesian framework. The Bayesian approach to inference permits coherent assessment of uncertainty for all model parameters and provides a flexible framework for propagation of parameter uncertainty to quantities derived from the model. We exploit this flexibility to obtain posterior distributions for highly disaggregated coverage probabilities (see Subsection 3.4) and for useful summaries of these probabilities (see Subsection 4.4). We generally specify prior distributions to be only weakly informative, in the sense of being open-minded as to the range of parameter values, while guaranteeing that inherent range constraints are respected (e.g., positive variances) and discouraging, but not disallowing, extreme values (Gelman et al. 2008).

In our application, we obtain a Monte Carlo approximation to the joint posterior distribution for all model parameters, by generating a sample from the posterior using Markov Chain Monte Carlo (MCMC) methods. Specifically, the sample is obtained using the program Stan (Stan Development Team 2020b) through the R interface (Stan Development Team 2020a; R Core Team 2019). Stan implements Hamiltonian Monte Carlo, a popular type of MCMC algorithm known to reduce the correlation between successive sampled values and, therefore, efficiently converging to the posterior distribution.

3.2. General Assumptions

A critical assumption of Bayesian analysis of survey data is ignorability (Rubin 1987, chap. 2; Little 2003; Gelman et al. 2014, chap. 8), which in the case of PES, requires conditional independence of inclusion in PES and inclusion in census, given the model covariates and *a priori* independence of the parameters of the models for inclusion in PES and in census. The former assumption is similar to the often invoked “independence” assumption of dual systems population estimation (Chandrasekar and Deming 1949; Brown et al. 2019). When ignorability holds, inference for inclusion in census (that is, census coverage) can proceed without specifying and fitting the model for inclusion in PES. In order to justify the assumption of ignorability, it is usually necessary to include the survey design features in the model, along with other covariates associated with nonresponse. We follow this approach in developing the model for census coverage. The nested geographical clustering of the sample design naturally lends itself to multilevel modelling, and, fortunately, in our case, there is overlap between variables of substantive interest and those predictive of nonresponse. We discuss the ignorability assumptions for our analysis in more detail in Section C of Online Supplemental Material, which tailors the general approach to Bayesian analysis of complex surveys given in Gelman et al. (2014, chap. 8) to the specific case of PES.

As well as conditional independence of inclusion in PES and census, we make the other standard assumptions of dual systems population estimation. We assume no errors in the

linkage of PES to census, and we assume the target population is closed over the operating periods of census and PES.

3.3. General Under-coverage Model

We let $\mathbf{X}$ denote demographic covariates and $\mathbf{x}$ a particular covariate combination. We use the notation TA to denote geographic area, and let $t \in \{1, \dots, 88\}$ indicate a particular TA. To simplify notation in this section we let $\mathbf{V} = (\mathbf{X}, \text{TA})$, so $\mathbf{v} = (\mathbf{x}, t)$ refers to a particular covariate combination $\mathbf{x}$ in TA t . The sample space for $\mathbf{V}$ is the space of all covariate-TA combinations, denoted $\mathcal{V}$.

Introducing the indicators C and Q for inclusion in the census and in the target population respectively, we define the under-coverage probability as

$$p_{\text{under}}(\mathbf{v}, \boldsymbol{\xi}) = \Pr(C = 0 | Q = 1, \mathbf{V} = \mathbf{v}, \boldsymbol{\xi}),$$

where $\boldsymbol{\xi}$ is the parameter vector of the under-coverage model.

The purpose of the model presented here is to estimate $p_{\text{under}}(\mathbf{V}, \boldsymbol{\xi})$. A coverage-adjusted population estimate based on the census can subsequently be obtained by weighting each census record by the inverse of the under-coverage probability:

$$w_i = \frac{1}{(1 - p_{\text{under}}(\mathbf{v}_i, \boldsymbol{\xi}))}, \quad (1)$$

where the subscript refers to the i^{th} census respondent. Using a Bayesian approach enables this adjustment to be applied to each census record for each draw from the posterior distribution for $p_{\text{under}}(\mathbf{v}_i, \boldsymbol{\xi})$ in a Monte Carlo procedure which produces as many simulations of the ERP as needed to obtain precise uncertainty measures (e.g., approximate credible intervals). More details on the Monte Carlo methodology of the ERP production can be found in [Bryant et al. \(2016\)](#) and [Stats NZ \(2020a\)](#).

We let N_h^{ind} , denote the number of usual residents within household h , N_p^{hh} , the number of households in PSU p , N_s^{psu} , the number of PSUs in stratum s , N_t^{strat} , the number of strata intersecting TA t , $N_{\text{tot}}^{\text{strat}}$ the total number of strata and N^{ta} , the total number of TAs. After the linking procedure between PES and census, each record j in household h in the PES data set receives an under-coverage indicator Y_{hj} which states whether the record is present in the census file ($Y = 0$) or absent from it ($Y = 1$). Each record is also characterised by a set of demographic covariates $\mathbf{X}_{hj}^{\text{ind}}$, geographic variables related to the survey design, and local government area, TA. We present the model for census under-coverage in two ways: with a directed acyclic graph (DAG) ([Figure 1](#)), and with the following equations, followed by a description.

$$[Y_{hj} | X_{hj}^{\text{ind}}, p_{\text{under}_{hj}}] \stackrel{\text{indep}}{\sim} \text{Bernoulli}(p_{\text{under}_{hj}}); j = 1, \dots, N_h^{\text{ind}}; h = 1, \dots, N_{\text{psu}[h]}^{\text{hh}}, \quad (2)$$

$$\text{logit}(p_{\text{under}_{hj}}) = \alpha_h^{\text{hh}} + X_{hj}^{\text{ind}'} \boldsymbol{\beta}; j = 1, \dots, N_h^{\text{ind}}; h = 1, \dots, N_{\text{psu}[h]}^{\text{hh}}, \quad (3)$$

$$[\alpha_h^{\text{hh}} | \mathbf{X}_h^{\text{hh}}, \alpha_{\text{psu}[h]}^{\text{psu}}, \boldsymbol{\beta}^{\text{hh}}, \sigma_{\text{hh}}^2] \stackrel{\text{indep}}{\sim} \mathcal{N}(\boldsymbol{\mu} + \alpha_{\text{psu}[h]}^{\text{psu}} + \mathbf{X}_h^{\text{hh}'} \boldsymbol{\beta}^{\text{hh}}, \sigma_{\text{hh}}^2); h : \text{psu}[h] = p; \quad (4)$$

$$p = 1, \dots, N_s^{\text{psu}},$$

$$\left[\alpha_p^{\text{psu}} | \mathbf{X}_p^{\text{psu}}, \alpha_{\text{strat}[p]}^{\text{strat}}, \beta^{\text{psu}}, \sigma_{\text{psu}}^2 \right] \stackrel{\text{indep}}{\sim} \mathcal{N} \left(\alpha_{\text{strat}[p]}^{\text{strat}} + \mathbf{X}_p^{\text{psu}'} \beta^{\text{psu}}, \sigma_{\text{psu}}^2 \right); p : \text{strat}[p] = s; \quad (5)$$

$$s = 1, \dots, N_t^{\text{strat}},$$

$$\left[\alpha_s^{\text{strat}} | \mathbf{W}, \alpha^{\text{ta}}, \sigma_{\text{strat}}^2 \right] \stackrel{\text{indep}}{\sim} \mathcal{N} (\mathbf{W}_s \alpha^{\text{ta}}, \sigma_{\text{strat}}^2), s = 1, \dots, N_{\text{tot}}^{\text{strat}}, \quad (6)$$

$$\left[\alpha_t^{\text{ta}} | \mathbf{X}_t^{\text{ta}}, \beta^{\text{ta}}, \sigma_{\text{ta}}^2 \right] \stackrel{\text{indep}}{\sim} t_3 \left(\mathbf{X}_t^{\text{ta}'} \beta^{\text{ta}}, \sigma_{\text{ta}}^2 \right), t = 1, \dots, N^{\text{ta}}, \quad (7)$$

where the notation $\text{psu}[h]$, and $\text{strat}[p]$ refer, respectively, to the PSU of the h^{th} household and the stratum of the p^{th} PSU. Note $\alpha^{\text{ta}} = (\alpha_1^{\text{ta}}, \dots, \alpha_{N^{\text{ta}}}^{\text{ta}})'$ is a $N^{\text{ta}} \times 1$ vector of TA

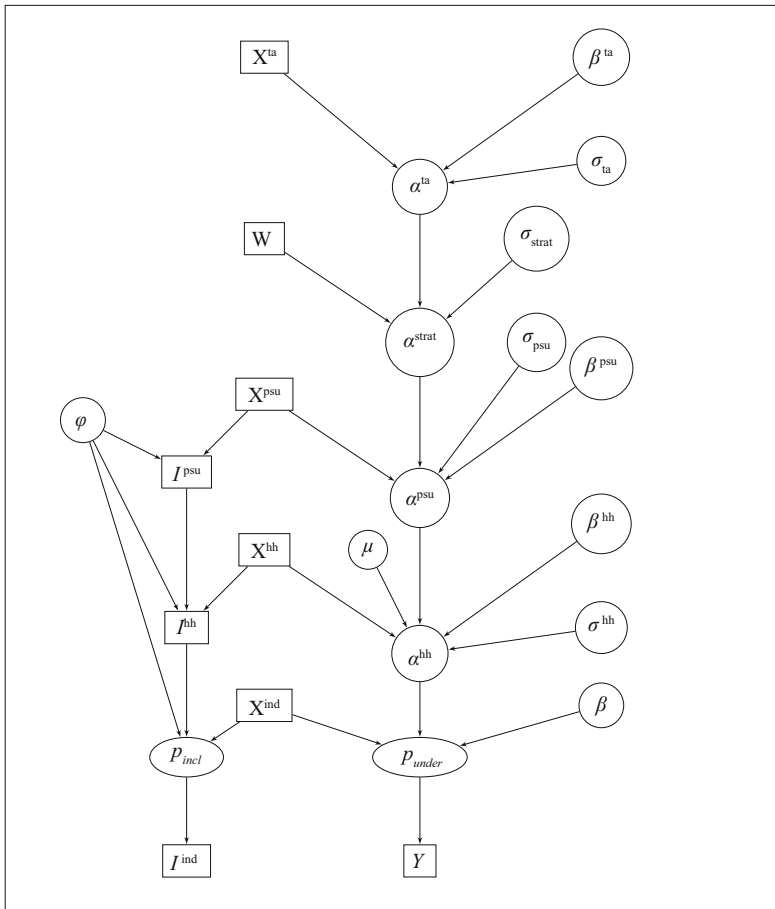


Fig. 1. Directed acyclic graph (DAG) representing the structure of the model for census coverage (Y). Rectangles represent observables and circles represent model parameters. p_{incl} and p_{under} are probabilities obtained as deterministic functions of observables and parameters. I^{psu} , I^{hh} , I^{ind} are indicators for PSU, household and individual inclusion in PES. The inclusion model depends on the parameter vector, ϕ . The inclusion indicators may depend on the model covariates but, because they are assumed conditionally independent of census coverage and parameters of the census coverage model, given the covariates, and because the inclusion and coverage model parameters are assumed to be a priori independent, inclusion in PES is ignorable. Under these assumptions, modelling of census coverage using PES data can proceed without modelling inclusion in PES.

effects. The notation t_3 in Equation (7) corresponds to a Student- t distribution with three degrees of freedom.

We model the under-coverage indicator Y_{hj} for individual j in household h using a Bernoulli distribution with probability $p_{underhj}$ (2). A logistic regression is specified for $p_{underhj}$ with individual covariates $\mathbf{X}_{hj}^{\text{ind}}$ and household-specific varying intercept α_h^{hh} (3). Equations (4)–(6) show how the varying intercept α_h^{hh} contains an overall average μ and all levels of the hierarchy reflecting the PES sampling design: it is modelled as a normal distribution, and the mean of this distribution is the result of a regression with a PSU-level effect and household covariate effects in Equation (4). These household covariates are a “hard-to-find” binary variable (HTF) which accounts for the variation in dwelling enumeration success between areas (details in Section D of the Online Supplemental Material), and potential individual demographic variables summarised at the household level. The PSU-level effect α_p^{psu} is itself a varying effect that we model with a normal distribution, and the mean of this distribution is the result of a regression with a stratum-level effect term and PSU covariate effects in Equation (5). These covariates are the PSU sampling variables used in the PES sampling design (see Table 1). The stratum-level effect α_s^{strat} is a varying effect modelled with a normal distribution whose mean is a weighted mean of TA-level effects from TAs present in the stratum (6). We add a TA level to the model, as this is the geographic resolution required for the publication of the ERP.

As each of the 101 strata generally spans several TAs, the relationship of strata to TAs is described by an occurrence matrix $\mathbf{W}$ where each row corresponds to a stratum and each column corresponds to a TA. Each matrix cell $\mathbf{W}(s,t)$ therefore represents the proportion of TA t included in stratum s . These cell proportions are estimated based on individual counts within small geographical units in the augmented census file, which includes administrative records in addition to census responses (Stats NZ 2019). We let $\mathbf{W}_s$ denote the s^{th} row of $\mathbf{W}$. Finally the TA effects α_t^{ta} are modelled through covariates $\mathbf{X}^{\text{ta}}$, which correspond to four socio-economic predictors of TA effects that are calculated from NZ Deprivation indices (Atkinson et al. 2019). We choose a t_3 distribution at the TALB level because it has more mass in its tails than the normal distribution, which helps avoid over-shrinkage at higher levels of the hierarchical model. A detailed description of the individual covariates included in $\mathbf{X}^{\text{ind}}$ as well as higher level covariates $\mathbf{X}^{\text{hh}}$, $\mathbf{X}^{\text{psu}}$ and $\mathbf{X}^{\text{ta}}$ is given in Table 1.

In practice, incorporating the group covariates $\mathbf{X}^{\text{psu}}$ and $\mathbf{X}^{\text{hh}}$ at the individual level of the model (3) by allocating all individuals in a group (household or PSU) the covariate values for that group gives an equivalent formulation to Equations (2–7) and makes subsequent predictions easier to compute. We let $\mathbf{X} = (\mathbf{X}^{\text{ind}'}', \mathbf{X}^{\text{hh}'}', \mathbf{X}^{\text{psu}'}')'$. A demonstration of the equivalence of the two approaches is detailed in Section B of the Online Supplementary Material.

The choice of individual covariates used in Equation (3) is largely guided by New-Zealand post-enumeration surveys from previous years (Table 1). For instance, age, ethnic group and sex are known to affect census inclusion in distinct ways, so we include these variables and interactions in all models we examine. The four ethnic indicators are Māori, Pacific, Asian and Other. They are mutually non-exclusive, allowing individuals to belong to multiple ethnic groups. We added interaction terms between ethnicities representing two common profiles of people with multiple ethnicities: Māori-Other, and Māori-Pacific.

Table 1. Covariates used in the coverage model. “n. param” shows the number of parameters estimated for each of the covariates and covariate interactions. Covariates only used in the second model are depicted in grey.

Variable	Coding	Description	n. param
Individual covariates			
sex	binary	0 = male, 1 = female	1
age	10 splines	quadratic age splines with knots at ages 10, 20, 30, 40, 51, 61, 71, and 81	10
Māori	binary	Māori ethnicity indicator	1
Pacif	binary	Pacific ethnicity indicator	1
Asian	binary	Asian ethnicity indicator	1
Other	binary	indicator for “other” ethnicities	1
NZ born	binary	0 = born abroad, 1 = born in New Zealand	1
Māori descent	binary	0 = non-Māori descent, 1 = Māori descent	1
Individual covariate interactions			
Māori * Other	binary		1
Māori * Pacif	binary		1
sex * age	binary	sex and all 10 age splines	10
Asian * NZ born	binary		1
ethnicity * age	Binary	5 first age splines with each ethnicity and with Māori* Other (3-way)	25
Household covariates $\mathbf{X}^{hh}$ (model 2 only)			
Māori	binary	presence of Māori	1
Pacif	binary	presence of Pacific	1
Asian	binary	presence of Asian	1
Other	binary	presence of Other	1
Female	binary	presence of females	1
Māori descent	binary	presence of people of Māori descent	1
NZ born	binary	presence of people born in New Zealand	1
HTF	binary	hard-to-enumerate area	1
Household covariate interactions (model 2 only)			
between ethnicity indicators	binary		6
ethnicity * female	binary		1
ethnicity * NZ born	binary		1
PSU covariates $\mathbf{X}^{psu}$			
Pacif prop	continuous	proportion of Pacific adults	1
PSU size	categorical	S(< 50 dwellings)/M(50-100)/L(>100)	2
TA covariates $\mathbf{X}^{ta}$			
communication	continuous	Prop. of people with no access to internet at home	1
income	continuous	Prop. of people living in households with income below the poverty threshold	1
qualification	continuous	Prop. of people aged 18-64 without any qualifications	1
internet response	continuous	proportion of online census responses	1

For individual-level variables that are available but whose effect on coverage is less obvious, and for more subtle interactions between covariates, we compute several models differing in their covariates and interactions. Careful examination of resulting parameter

posterior distributions and predictions as well as out-of-sample deviance calculations are the used to guide model selection.

All individual covariates except age are binary variables. The challenge with modelling age is that it is inherently an ordered categorical variable with potentially more than 100 categories. Treating this variable as such creates the challenge of estimating a large number of parameters, and dividing the sample into excessively small categories. One solution is to create broader categories such as five-year age groups, but this solution does not reflect the continuous character of age and its effect on census coverage. It also introduces the additional issue of subjectively selecting categories, and creates breaks among contiguous years that may share extreme values. Another solution, implemented here, is to apply a spline transformation to the original variable. We model age using ten quadratic splines defined by eight internal breakpoints (see Table 1). Figure S1 (Online Supplemental Material) illustrates the transformation by showing the spline values for each age present in the census. One can see that at any given age, a maximum of three splines contribute to describing the underlying age. This stems from our choice of quadratic polynomials rather than higher-degree polynomials, in order to limit the smoothing of patterns that would result from highly overlapping spline curves.

We select *Cauchy* (0, 2.5) as a prior for μ , which is a standard prior recommended in Gelman et al. (2008). Covariate effect parameters β , β^{hh} , β^{psu} and β^{ta} are drawn from independent $\mathcal{N}(0,1)$ distributions. This is not unduly restrictive yet places low prior probability on extreme values. As a reference point, after converting to the odds ratio scale, a $\mathcal{N}(0,1)$ prior for a logistic regression parameter corresponds to the 95% prior interval $\exp(\pm 1.96)$, implying a $\exp(3.92) \approx 50$ – fold range of prior variation for the effect in question. Group-level variances σ_{ta}^2 , σ_{strat}^2 , σ_{psu}^2 , and σ_{hh}^2 are drawn from independent *Cauchy*⁺ (0, 2.5) distributions, where *Cauchy*⁺ refers to the Cauchy distribution truncated to positive values.

We run three HMC chains for each model, with the first half used as warm-up. We set the target average proposal acceptance probability to 0.9 and let all other algorithm parameters be set at their default value. For each model, we determine chain length experimentally by increasing it until convergence is reached. We ensure convergence by using the potential scale reduction factor, $\hat{R}$ (Gelman et al. 2014, 284–285) and by visually assessing chain profiles. We also monitor the effective Monte Carlo sample size to ensure appropriate post-convergence Monte Carlo sample size (Gelman et al. 2014, 286–287).

We explore potential models in two stages. We first focus on individual covariates and their interactions. Group-level covariates at the household, PSU and TA level as well as the stratum level are present in the varying intercept to account for the PES sampling design. The basic model therefore involves all individual covariates as well as the group-level covariates pertaining to the sample design.

However, results associated with this approach (see Section 4) suggest that individual covariates cannot fully account for variation in census coverage. As the census interview process is dwelling-based, households are an important component of the survey design and this is accounted for in the model through the first level of the varying intercept, α^{hh} . It is likely that census response is partly driven by household-level characteristics that are unobserved. It is also possible that an individual's response or nonresponse is influenced by another individual in the household. For instance, it is reasonable to suggest that

children's response to census is dependent on the parents or caregivers they live with. In such cases, we expect nonresponse of the former to depend on nonresponse of the latter, therefore bringing nonresponse at the household level. This is inconsistent with the model structure, which implicitly assumes that the household-level intercept and individual predictors are independent. To allow for correlation between household and individual characteristics, we follow the solution described in [Gelman and Hill \(2006, 506–507\)](#): we create versions of the individual covariates aggregated at the household level. Therefore, in a second stage, we experiment with the creation of many household-level covariates calculated from all individual covariates. The new covariates are included in $\mathbf{X}_h^{\text{hh}}$ in Equation (4), and described in [Table 1](#). The outcomes from including these additional household covariates in the model are addressed in Sections 4 and 5.

3.4. Predicting Under-Coverage Probabilities of Census Records

To produce the ERP, coverage probabilities are required for each combination of covariates occurring in the census file. Household- and PSU-level covariates are included in the ERP production but not individual household or PSU effects. The geographic level for application of coverage probabilities is the TA level. While other choices could have been made, these settings provided a compromise between computational tractability and granularity of estimation. Below we describe how the model can be used to generate coverage probabilities at the desired level of demographic and geographic detail.

After fitting the multilevel logistic model, 1,000 samples are extracted from the posterior distribution. Each of the 1,000 draws from the posterior can be used to predict the undercoverage probability associated with each combination of covariate values that exist in the census. Parameters related to the sampling design (household, PSU, and stratum effects) are integrated to obtain a posterior prediction for each covariate-TA combination. For each combination of TA and individual, household and PSU level covariates, $\mathbf{v} = (\mathbf{x}', t)'$ where $\mathbf{x} = (\mathbf{x}^{\text{ind}'}, \mathbf{x}^{\text{hh}'}, \mathbf{x}^{\text{psu}'})'$, and for each draw from the posterior of ξ , we require

$$p_{\text{under}}(\mathbf{v}, \xi) = \Pr(Y = 1 | \mathbf{V} = \mathbf{v}, \xi)$$

$$= \sum_{s=1}^{N_t^{\text{strat}}} \Pr(Y = 1 | \mathbf{X} = \mathbf{x}, \text{strat} = s, \xi) \Pr(\text{strat} = s | \text{TA} = t) \quad (8)$$

$$= \sum_{s=1}^{N_t^{\text{strat}}} \left(\int (\Pr(Y = 1 | \alpha^{\text{hh}}, \mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}, \text{strat} = s, \xi) \times \right. \quad (9)$$

$$\left. p(\alpha^{\text{hh}} | \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{strat} = s, \xi) \right) d\alpha^{\text{hh}} \times \Pr(\text{strat} = s | \text{TA} = t))$$

$$= \sum_{s=1}^{N_t^{\text{strat}}} \left(\int \expit(\alpha^{\text{hh}} + \mathbf{x}^{\text{ind}'} \beta) \mathcal{N}(\alpha^{\text{hh}} | \mu + \mathbf{x}^{\text{hh}'} \beta^{\text{hh}} + \mathbf{x}^{\text{psu}'} \beta^{\text{psu}} + \alpha_s^{\text{strat}}, \sigma_{\text{hh}}^2 + \sigma_{\text{psu}}^2) d\alpha^{\text{hh}} \right. \quad (10)$$

$$\left. \times \Pr(\text{strat} = s | \text{TA} = t) \right),$$

where $\expit()$ is the inverse logit function and $\mathcal{N}(\cdot | \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2 . Writing $\mathbf{X} = \mathbf{x}$ instead of $\mathbf{V} = \mathbf{v}$ in the first component of Equation (8) follows from the assumptions of the model given by Equations (2–7). TAs affect census under-coverage only via strata, so after conditioning on strata, conditioning

on TAs becomes unnecessary. Similarly, we write $\mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}$ instead of $\mathbf{X} = \mathbf{x}$ in the first element in the integral in Equation (9) because conditioning on α^{hh} means we do not need to condition on $\mathbf{X}^{\text{hh}}$ and $\mathbf{X}^{\text{psu}}$. The normal density for the household effects in the integrand in Equation (10) follows from the model Equations (4) and (5) since, by the mixture property of the normal distribution (Gelman et al. 2014, 577), we have

$$\begin{aligned} p(\alpha^{\text{hh}} | \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{strat} = s, \xi) \\ &= \int p(\alpha^{\text{hh}} | \alpha^{\text{psu}}, \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{strat} = s, \xi) p(\alpha^{\text{psu}} | \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \alpha_s^{\text{strat}}) d\alpha^{\text{psu}} \\ &= \int \mathcal{N}(\alpha^{\text{hh}} | \mu + \mathbf{x}^{\text{hh}'} \boldsymbol{\beta}^{\text{hh}} + \alpha^{\text{psu}}, \sigma_{\text{hh}}^2) \mathcal{N}(\alpha^{\text{psu}} | \mathbf{x}^{\text{psu}'} \boldsymbol{\beta}^{\text{psu}} + \alpha_s^{\text{strat}}, \sigma_{\text{psu}}^2) d\alpha^{\text{psu}} \\ &= \mathcal{N}(\alpha^{\text{hh}} | \mu + \mathbf{x}^{\text{hh}'} \boldsymbol{\beta}^{\text{hh}} + \mathbf{x}^{\text{psu}'} \boldsymbol{\beta}^{\text{psu}} + \alpha_s^{\text{strat}}, \sigma_{\text{hh}}^2 + \sigma_{\text{psu}}^2). \end{aligned}$$

$\Pr(\text{strat} = s | \text{TA} = t)$ is estimated using an occurrence matrix constructed using the same data as $\mathbf{W}$, that is the official census file, which augments the census respondent file with administrative records.

The integral in Equation (10) produces predicted coverage probabilities that are marginalised with respect to household and PSU effects. That is, they are not predictions that are relevant to particular households, but are expectations over the distribution of household effects among households with covariate values $\mathbf{x}^{\text{hh}}$ in PSUs with covariates $\mathbf{x}^{\text{psu}}$. An alternative, conditional prediction, could be obtained by setting the household and PSU effects to zero (or some other value) but such predictions are tied to households and PSUs with the specified effect and would not be appropriate for application to the census file for which the desired notion is that of an unknown household with particular household covariate values in an unspecified PSU with particular PSU covariate values. Further discussion on the marginal and conditional predictions can be found in Skrongdal and Rabe-Hesketh (2009) and Pavlou et al. (2015) and some more details on the derivation of Equation (10) are given in Section B of the Online Supplemental Material. In our application we use Monte Carlo integration to approximate the integral in Equation (10).

4. Results

4.1. Using Mixed Predictive Checks to Assist With Model Assessment

Models were run with three HMC chains of sufficient length (11,000–12,000 iterations) to ensure convergence. Stan run times with parallel chains were 8.0 hours for the initial model (model 1), and 17.9 hours for the model with household covariates (model 2, see Subsection 4.2). After discarding the first half of each chain as warm-up period, the $\hat{R}$ convergence diagnostic (Gelman et al. 2014, 284–285) was less than 1.01 for all monitored parameters for both models. We assess the quality of the models using posterior predictive checking focusing on marginal predictions for two different groupings: demographic categories formed by all binary demographic covariates, and TAs. Results for all checks performed on two models are presented in Figure 2 for predictions on demographic groupings, and in Figure 3 for TA-level predictions. Note that only the first 42 TAs of the North Island are shown in Figure 3 (See Online Supplementary Material Figure S2 for all other TAs). For the

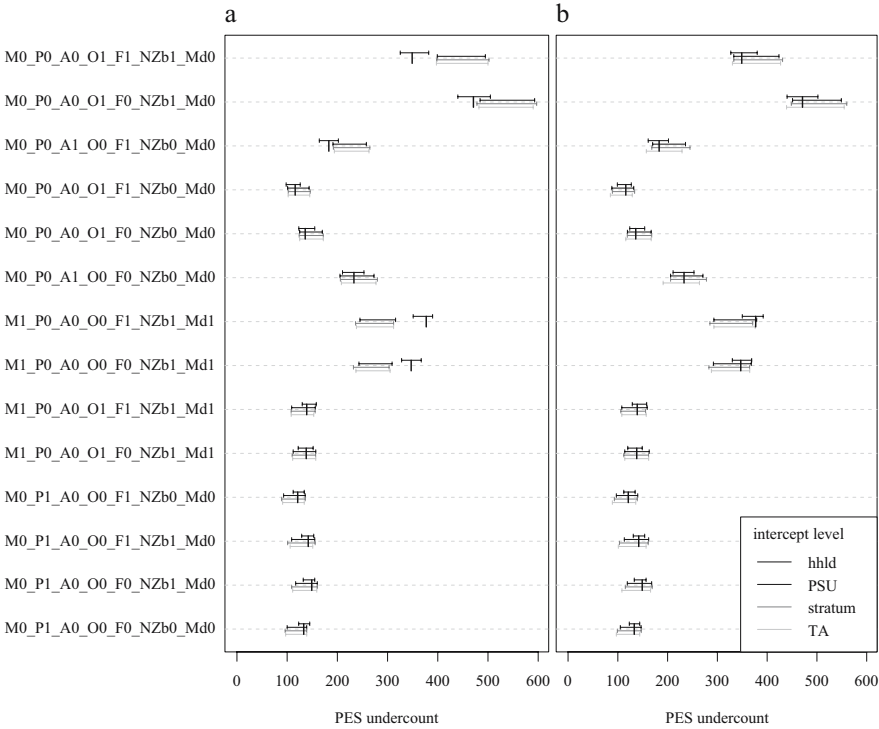


Fig. 2. Posterior predictive checks (PPCs) for model 1 (without household covariates, left) and model 2 (with household covariates and their interactions, right). Only the 14 most common demographic categories are represented, out of 81 demographic categories, all geographic areas pooled. The vertical black bars represent the PES observed counts for the categories displayed on the left axis. PPC results for different levels of integration are represented by horizontal 90% credible intervals. For each category, the top interval corresponds to PPCs performed on the raw output from the logistic regression, and lower intervals (different shades of grey) correspond to PPC results after Monte Carlo integration of household parameters, household and PSU parameters, and household, PSU, stratum parameters, respectively. The labels on the left are to be interpreted as a combination of demographic variables (letters) with whether or not the category comprises individuals corresponding to the demographic variable (0 = no, 1 = yes). M = Māori, P = Pacific, A = Asian, O = Other, F = Female, NZb = New-Zealand born, Md = Māori descent. For instance the top category corresponds to females of Other ethnicity only, who were born in New Zealand and are not of Māori descent.

first predictive check, we use each sample from the joint posterior distribution of parameters to replicate the PES data under the logistic model described in Equations (2)–(3). We compare the 1,000 simulated data sets with the observed data. We summarise the aggregated undercount distributions from simulated data sets using 90% posterior predictive intervals and assess whether observed undercounts fall within these intervals (Figures 2a and 3a, top intervals). This self-consistency check allows us to confirm that the model fits the data: all observed aggregated undercounts fall within the 90% posterior predictive intervals from simulated data, for both demographic and geographical groupings.

The PES model is designed to predict under-coverage for census records. Census records can be considered as “new observations” that we need the model to output predictions for. We therefore need to assess not only the fit but also the predictive ability of the model when confronted with new observations. This is especially important as these observations do not fit into the hierarchy of households, PSUs, and strata that was solely defined to account for

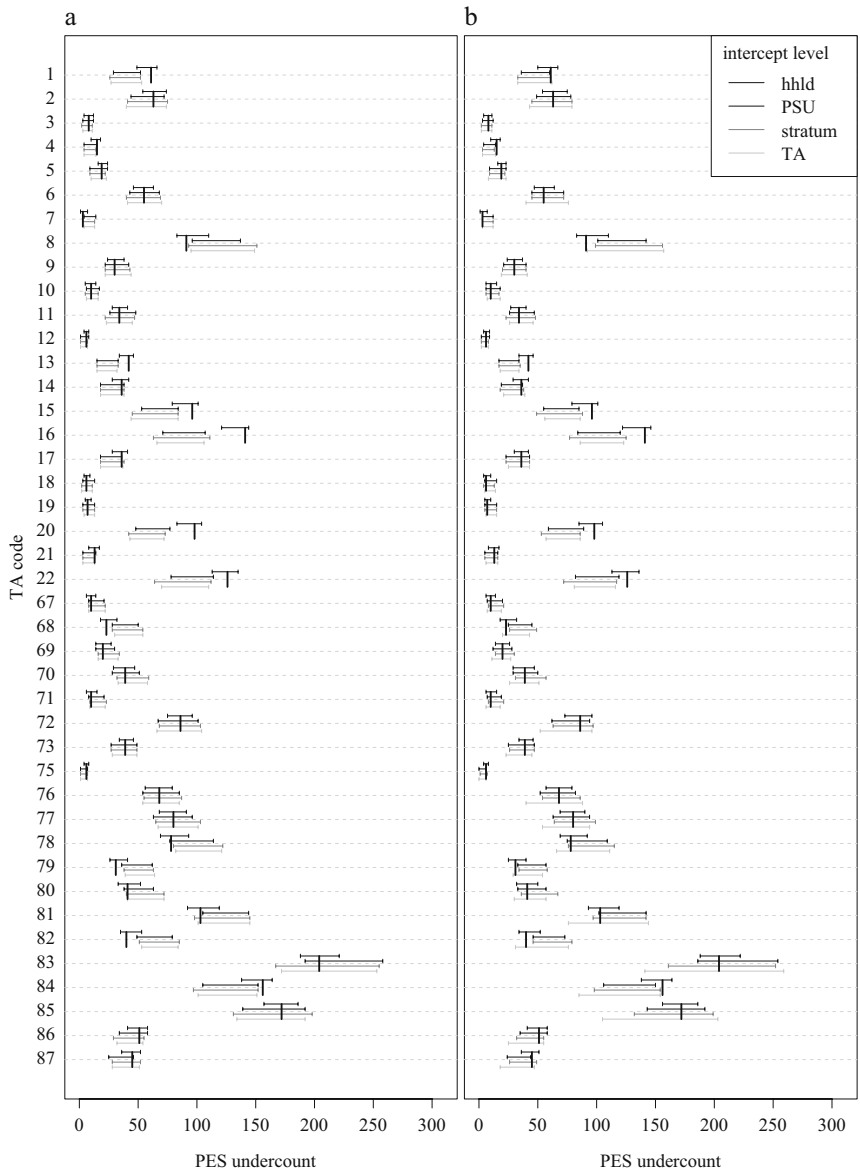


Fig. 3. Posterior predictive checks (PPCs) for model 1 (without household covariates, left) and model 2 (with household covariates and their interactions, right). Undercounts by TAs are represented for the first 42 TAs of the North Island, all demographic categories pooled. See Figure 2 caption for further details.

the PES sampling design. Therefore, we need to determine how good the model is at estimating under-coverage probabilities for census records, which are characterised by the same TA and demographic information as PES records but are for the most part not included in the households and PSUs selected for PES. This can be done using mixed predictive checking, whereby predictions are performed for new individuals (outside of the PES sampling frame) with exactly the same demographic and group-level predictors as PES individuals (Gelman et al. 1996). In our case, it amounts to applying Equation (10) to all PES

records, drawing from the Bernoulli process to simulate the under-coverage indicator, and aggregating the results to the same groupings as previous posterior predictive checks (TA and demographic categories). The results are displayed as light grey 90% credible intervals (bottom interval) on [Figures 2a and 3a](#). With results aggregated to demographic groups, the model shows a substantial misfit for five out of the 14 most common demographic groups, which is a higher proportion than the 10% roughly expected under the assumption that the model is adequate. The TA grouping also shows widespread misfit, with almost a third of all TA under-coverage counts lying outside of the predicted 90% credible intervals.

When a misfit is observed after integration of the sampling design parameters, it is useful to investigate what level of the hierarchy causes the problem, especially in models with more than two levels. In our case, unaccounted for variation could be present at the individual, household, PSU, or stratum level. To assess the problematic level, we compute mixed predictive checks where some but not all of the grouping levels in the hierarchy are integrated. We first perform predictions from the PES data considering that the stratum and PSU of each individual are known but the household is new, therefore sampling the varying intercept from the population distribution for households with their given covariates. For an individual in PSU p , in a household with household covariate values $\mathbf{x}^{\text{hh}}$ and with individual covariate value $\mathbf{x}^{\text{ind}}$, this means calculating the following under-coverage probability:

$$q^{\text{psu}}(\mathbf{x}^{\text{ind}}, \mathbf{x}^{\text{hh}}, \boldsymbol{\xi}, \alpha_p^{\text{psu}}) = \int \text{expit}(\alpha^{\text{hh}} + \mathbf{x}^{\text{ind}'} \boldsymbol{\beta}) \mathcal{N}(\alpha^{\text{hh}} | \mu + \mathbf{x}^{\text{hh}'} \boldsymbol{\beta}^{\text{hh}} + \alpha_p^{\text{psu}}, \sigma_{\text{hh}}^2) d\alpha^{\text{hh}}.$$

The results are displayed as dark grey 90% credible intervals (second from the top) on [Figures 2a and 3a](#). We repeat the procedure where both households and PSUs are new, which in practice consists in sampling the intercept from the population distribution for households, integrated over PSUs. In this case, the under-coverage probability in stratum s is calculated as follows:

$$q^{\text{strat}}(\mathbf{x}^{\text{ind}}, \mathbf{x}^{\text{hh}}, \mathbf{x}^{\text{psu}}, \boldsymbol{\xi}) = \int \text{expit}(\alpha^{\text{hh}} + \mathbf{x}^{\text{ind}'} \boldsymbol{\beta}) \mathcal{N}(\alpha^{\text{hh}} | \mu + \mathbf{x}^{\text{hh}'} \boldsymbol{\beta}^{\text{hh}} + \mathbf{x}^{\text{psu}'} \boldsymbol{\beta}^{\text{psu}} + \alpha_s^{\text{strat}}, \sigma_{\text{hh}}^2 + \sigma_{\text{psu}}^2) d\alpha^{\text{hh}}.$$

The results of these predictions are displayed as grey 90% credible intervals (third from the top) on [Figures 2a and 3a](#).

4.2. Adding Household Covariates Improves the Model for Demographic Groups

The four performed predictive checks, with integration occurring at different grouping levels, clearly show that a major misfit arises when predicting under-coverage of individuals in new households. This misfit is visible for both demographic groupings ([Figure 2a](#)) and TA groupings ([Figure 3a](#)), but does not grow larger when predictions are calculated with unknown PSUs and/or strata. This suggests that the model did not estimate an adequate population distribution for households. Further graphical investigations (not shown) suggest some dependence between the varying intercept and individual covariates. As noted in Subsection 3.3, the logistic model described in Equation (2)–(7) assumes independence between intercept and individual-level predictors. Following [Gelman and Hill \(2006, 506\)](#), we address

this inconsistency by aggregating the individual covariates suspected to cause the dependency to the hierarchical level in question, and introduce the new variables as group-level covariates. We first test different ways of aggregating demographic covariates (ethnicity, age, sex, and New-Zealand born) at the household level. For continuous variables, it makes sense to average values across individuals in a group. However, several ways to aggregate categorical variables at the group level can be considered. For each individual categorical covariate, we test the following aggregation methods: (1) binary variable indicating presence/absence of household occupants with the demographic characteristic, (2) continuous variable of proportion of household occupants with the demographic characteristic, and (3) binary variable indicating a majority of household occupants with the demographic characteristic. We find that overall model performance is best when using (1) for all group-level covariates. We therefore only present results with these covariates (see Table 1 for the final list of household covariates and their interactions). We apply the same four predictive checks as we applied to the original model, and present the results in Figures 2b and 3b.

Introducing household covariates and their interactions considerably improves the fit of predicted undercounts of demographic groups to the observed data (Figure 2), with only one observed value sitting just outside the 90% credible intervals from the TA-level prediction. However, the modification only partially improves the fit to TA counts (Figure 3). While the model including household covariates fits most TAs well, seven TAs on Figure 3 still show predictive credible intervals that contain the true value when simple posterior predictive checks are performed but do not encompass it when performing any of the mixed predictive checks. For these TAs, the estimated household population distribution seems wrong, and we hypothesise that some unknown household characteristics cause unaccounted-for heterogeneity. Consistent with this hypothesis, between household variation is the largest component of unexplained variation and did not shrink after the inclusion of household level covariates (Table 2).

4.3. Complementary Cross-validation Tests

Posterior predictive checks give insight into the fit of the model to the data. They are a first “sense-check” of an analysis. The three additional mixed predictive checks, simulating data with the same covariates as the data but different hierarchical groups, constitute a step further in assessing not only the fitness but also the predictive power of the model, and where its limitations lie. However, these checks still use the exact same covariate values in the simulated data sets as in the observed data sets. One way to further determine the

Table 2. Marginal posterior quantiles of variance parameters for model 1 (without household covariates) and model 2 (with household covariates)

	Model 1			Model 2		
	2.5%	50%	97.5%	2.5%	50%	97.5%
σ^{hh}	4.10	4.35	4.62	4.18	4.45	4.74
σ^{psu}	0.29	0.76	1.05	0.08	0.63	0.98
σ^{strata}	0.01	0.18	0.45	0.01	0.15	0.42
σ^{ta}	0.01	0.13	0.43	0.01	0.12	0.43

predictive power of the models is cross-validation. For all tested models, and to assist with model selection, we calculated approximate leave-one-out-cross-validation scores using Pareto-smoothed importance sampling (PSIS). We show the results for the two main models presented in Subsections 4.1 and 4.2 in Table 3. Lower leave-one-out cross-validation scores (or their importance sampling approximation, LOOIC) indicate a lower out-of-sample deviance, and therefore more accurate predictions to new data. Model 2, with additional household covariates, has a lower LOOIC value than model 1, although the difference is of the order of one standard error. While LOOIC is the most appropriate measure of predictive accuracy for complex hierarchical models, the Pareto-k diagnostic values for both models but especially model 2 suggest the error in the LOO approximations might be high and the LOOIC values might understate predictive accuracy (Vehtari et al. 2017). This is typical of flexible hierarchical models where some groups have very few observations. Both model 1 and 2 are like this: the lowest level of the hierarchy, households, often contains only one or two observations (individuals).

4.4. Standardised Estimates

The PES model output gives estimates of $p_{under}(\mathbf{v}, \mathbf{g})$ and gives us insight into how different geo-demographic groups respond to census. From a demographer’s point of view and for the sake of planning future censuses, it is also valuable to know what factors are actually driving nonresponse patterns. For instance, in a TA with a high estimated census undercount, it can be of interest to know if nonresponse is due to the demographic composition of the TA, or if there are there intrinsic difficulties associated with operating a large-scale survey in this area. If demographic effects are predominant, then we can assume nonresponse is driven by behavioural patterns in the respondents, whereas area effects would suggest potential issues with incomplete address registers or other operational pitfalls. Insight into the relative impact of the different covariates on census coverage can be gained by calculating under-coverage probabilities across the categories of the variable of interest for a standardised distribution of all other covariates. For instance, one can obtain area-level estimates where differences due to their demographic composition are statistically removed, leaving only differences pertaining to intrinsic area characteristics. The same standardisation logic can be applied to other variables, for instance one can obtain estimates by ethnicity, standardising

Table 3. Out-of-sample deviance diagnostics for the model without household covariates (w/o hh covar.) and with household covariates (w/ hh covar.). $elpd_{loo}$: expected log pointwise predictive density. p_{loo} : effective number of parameters. LOOIC: Pareto-smoothed importance sampling leave-one-out cross-validation approximation. Values in brackets correspond to standard error estimates. The Pareto-k distribution section bins estimates of importance for all data records into categories ordered by decreasing quality.

model	$elpd_{loo}$	p_{loo}	LOOIC	Pareto-k distribution			
				$(-\infty, 0.5]$	$(0.5, 0.7]$	$(0.7, 1]$	$(1, 1)$
W/o hh covar.	- 5,747.1 (100.6)	2,858.6 (60.3)	11,494.3 (201.2)	65.1%	24.0%	9.9%	0.9%
W/ hh covar.	- 5,639.8 (99.4)	2,814.0 (59.7)	11,279.7 (198.8)	62.1%	25.5%	11.3%	1.1%

areas and all other demographic covariates. Following the example of TA-level standardised estimates, we can define, for a given TA t :

$$p_{\text{under}}^{\text{std}}(t, \xi) = \sum_{\mathbf{x}} p_{\text{under}}(\mathbf{x}, t, \xi) \Pr^{\text{std}}(\mathbf{X} = \mathbf{x} | \xi), \quad (11)$$

where $\Pr^{\text{std}}()$ refers to the covariate probabilities from some standard distribution. Note that the standard distribution is allowed to depend on the model parameters. This is not usual but suits our situation because a natural choice of standard population is the corrected version of the census file, based on the under-coverage probabilities, estimated from the model. Thus, if the inverse under-coverage probability for the i^{th} census record corresponding to a particular setting of parameter values ξ is $w_i(\xi)$ we can define $\Pr^{\text{std}}(\mathbf{X} = \mathbf{x} | \xi)$ as

$$\Pr^{\text{std}}(\mathbf{X} = \mathbf{x} | \xi) = \frac{\sum_{i: \mathbf{X}_i = \mathbf{x}} w_i(\xi)}{\sum_i w_i(\xi)}, \quad (12)$$

where the summations are over records in the census file. With the standard probabilities defined as in Equation (12), standardised under-coverage probabilities can be obtained for each TA and repeating this for each draw from the posterior for ξ will produce a sample from the joint posterior for the standardised under-coverage probabilities by TA. Credible intervals and other summaries, including for contrasts between TAs, can be computed from the posterior sample. The standardised coverage probability, given by Equation (11) can be contrasted with the marginal TA under-coverage probability which is

$$\Pr(Y = 1 | \text{TA} = t, \xi) = \sum_{\mathbf{x}} p_{\text{under}}(\mathbf{x}, t, \xi) \Pr(\mathbf{X} = \mathbf{x} | \text{TA} = t, \xi). \quad (13)$$

Comparing Equations (13) and (11) it can be seen that they differ only in the covariate distribution, with the standardised probabilities using the covariate distribution of the chosen standard population in place of the TA-specific covariate distribution used to obtain the marginal coverage probability. By definition, the standardised probabilities are all based on the same covariate distribution, so differences in standardised TA under-coverage probabilities reflect genuine geographic differences in census under-coverage.

We focus on standardised under-coverage probability estimates for TAs and for age-sex profiles, using the output from the model with household covariates (model 2). [Figure 4](#) shows results across TAs. Although uncertainty bounds are wide and overlap among most TAs, some TAs seem to have higher nonresponse probabilities than the majority, all else being equal. As our mixed predictive checks have identified some inaccuracies in the undercoverage predictions for some TAs (Subsection 4.2), we are cautious about drawing conclusions on our TA-level standardised estimates. [Figure 5](#) shows that Māori and Pacific people generally have higher nonresponse levels than other ethnic groups. People in their twenties have the highest nonresponse levels across all ethnic groups. A secondary under-coverage peak around age 50 is also visible, although it is more pronounced in Pacific people than in other ethnic groups.

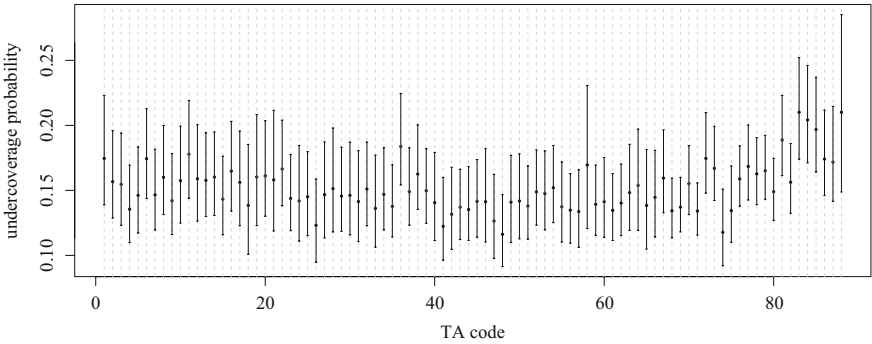


Fig. 4. Standardised estimates of under-coverage probabilities for each TA. Points correspond to posterior medians and error bars correspond to 90% credible intervals.

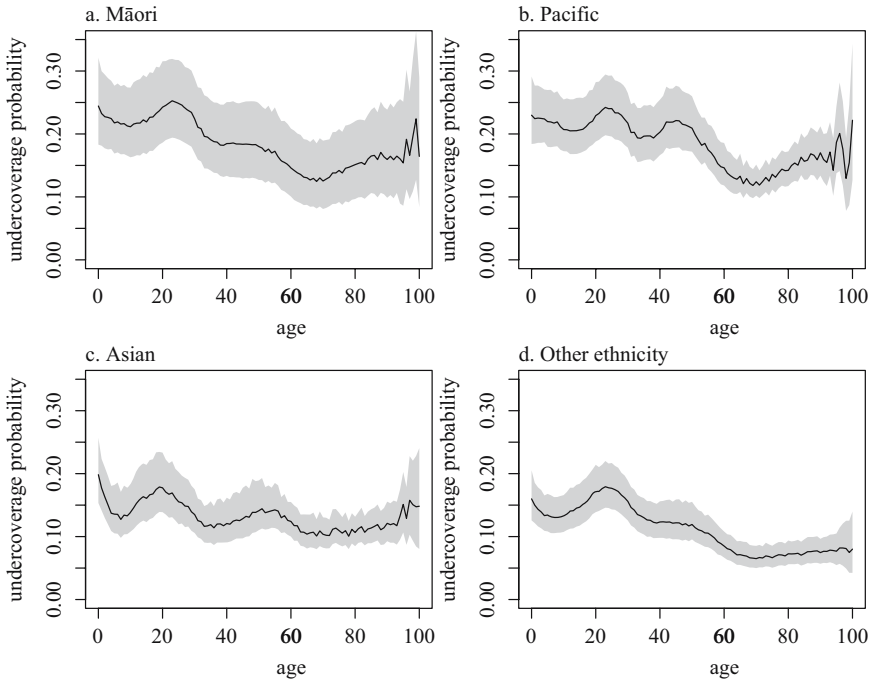


Fig. 5. Standardised estimates of under-coverage probabilities by age, for each ethnicity. Black lines correspond to posterior medians and grey shading corresponds to 90% credible intervals.

4.5. Effects of Coverage Adjustment on Census Counts

As stated in Subsection 3.3, census counts by demographic and geographic categories are corrected using the posterior values for $p_{\text{under}}(\mathbf{v}_i, \boldsymbol{\xi})$, following Equation (1). Although producing adjusted counts is out of scope for this article, it is useful to provide a sense of the scale of the correction to the census counts implied by the estimated under-coverage probabilities. Assuming model 2 is chosen, we find that overall average undercoverage of census responses is 10.9%. For a hypothetical census data set comprising four million responses this would mean adding about 490,000 individuals to the census respondent population. However, some demographic groups are better represented in census than

others. We find that census undercoverage in young male adults of Pacific and Māori populations can reach 35%. This means that the population for such categories is about 1.5 times larger than in the census responses, though in the official census file this effect is offset by the inclusion of administrative records.

5. Discussion

5.1. Posterior Predictive Checks Help Understand and Select the Multilevel Model

Through the present analysis, we have illustrated how census coverage can be quantified using a modelling approach to the analysis of complex survey data, thereby allowing insights at high levels of granularity in demographic and geographic attributes. To estimate undercoverage of the New Zealand 2018 census, we have fitted a binomial model with four nested geographic levels reflecting the complex sampling design of the post-censal survey. Especially, we have shown that multilevel models are a promising approach to analyse survey data with complex sampling designs. We reiterate that our results are experimental results that reflect a different analysis to the one used to output official 2018 census coverage estimates ([Stats NZ 2020b](#)).

We have also illustrated how performing extensive posterior predictive checking can assist in model selection in the case of a complex multilevel model structure. Combining posterior predictive checking, mixed predictive checking and cross-validation allowed us to assess the fit and predictive limitations of competing models as well as identifying aspects of the models that require modifications. Especially, performing mixed predictive checking at all levels of the hierarchy allowed us to identify the lack of fit at a specific level (here, the household level). We could assess the improvement associated with the subsequent addition of household-level covariates using additional mixed predictive checks and comparing cross-validation results across models.

Even after the addition of household-level covariates, posterior predictions did not always fit under-coverage data at the TA level. Some TA under-coverage counts were over-estimated, with 10 (of 88) TAs having PES-observed census under-coverage counts below the predicted 90% credible intervals. Other TA counts were under-estimated, with 12 TAs having observed under-coverage counts over the predicted 90% credible intervals ([Figures 3 and S2](#)). Mixed posterior predictive checks at each level of geographic parameter integration allowed us to attribute most of this misfit to the household level. The lack of fit is unlikely to be related to the demographic attributes of household occupants, as most of these attributes have been accounted for as household-aggregated level variables, and posterior predictive checks for demographic groups show no apparent bias. Subsequent investigations have failed to identify commonalities between TAs with similar misfit patterns.

The only noticeable pattern in this result is the relationship between observed undercoverage proportion of a TA and the direction of the estimation error: TAs which tend to be under-estimated are the ones with high observed under-coverage, whereas over-estimated TAs tend to have a low observed under-coverage proportion. Taken at face value this result suggests the model may be over-regularising more extreme estimates. In multilevel modelling, one expects small hierarchical groups with extreme observations to have predictions shrunk towards their expected value under the model. As the number of

levels increases, we can expect shrinkage to increase too for predictions made at higher levels. However, if overshrinkage, *per se*, was the main reason for the TA-level misfit, we would expect the issue to primarily affect smaller areas, whereas several of the TAs that the model fits poorly are large urban areas with a relatively large PES sample size. Further, if over-shrinkage was the primary reason for the lack of model fit in some areas, we would expect predictions to gradually show more shrinkage as we move from household-level predictions towards TA-level predictions, instead of a single jump from adequate fit at the household level to misfit for prediction at all other levels. We experimented with replacing the normal distributions for household, PSU and stratum effects by the heavier tailed t distribution with three degrees of freedom, but this had no impact on results. If the lack of fit in some TAs was due to the normal models tending to over-shrink extreme values we would have expected to see some improvement in fit when t_3 priors were adopted. The most plausible explanation for the lack of fit in some TAs is that one or several important factors related to geography were not included in the model. If this explanation holds, it follows that estimates are, in some cases, being shrunk towards expectations that do not exhibit the appropriate amount of geographic variation because a geographically varying covariate has been omitted from the model. In this specific sense, the estimates for some TAs may be exhibiting the effects of over-shrinkage. As noted above, given our posterior predictive checking results, the omitted covariate(s) seems likely to be a household-level variable, such as an aspect of dwelling construction (e.g. freestanding versus in an apartment block) that varies by area and is related to census coverage (e.g., census enumeration may be more difficult in apartment blocks). A natural next step for future model improvement would involve attempting to identify the missing covariate(s). If the missing covariates cannot be identified or sourced, a potential alternative is to specify the problematic group distribution as a mixture of several distributions. This may allow the model to recover unobserved categories within groups and improve model fit. However moving to mixture distributions at one or more of the model levels introduces additional computational complexity.

5.2. Individual Demographic Characteristics Drive Census Coverage Patterns

Standardised estimates give insight into the role played by different demographic and geographic attributes in driving coverage differences between groups. Though common practice in epidemiology and demography, the use of standardisation to adjust for differences in co-variate distributions has been less common in official statistics. In our case it provided a simple way to present comparative results from a complex model. Comparing [Figures 4 and 5](#) suggests high census under-coverage patterns are in general driven more by individual demographic attributes than by geographic ones. For instance, Māori and Pacific people as well as people in their early twenties tend to respond to census less than other demographic groups, independently of where they live. Although most TAs do not seem to intrinsically drive census under-coverage, clusters of TAs with higher under-coverage propensity can be identified. In this case standardised estimates can be used in the planning of future census operations. For instance, incentivisation and follow-up efforts could be allocated more heavily in TAs where under-coverage propensity has historically been high.

5.3. *Design-Based versus Model-Based Approach*

The most common approach to analysing complex surveys has traditionally been through a design-based method, where individual sampling weights are calculated from the sampling frame and subsequently adjusted for nonresponse. This approach has limitations when survey nonresponse is difficult to track. For instance, we do not know the number of occupants in nonresponding households nor the number of nonrespondents in a responding household. Sampling weights are often adjusted for nonresponse using benchmark population data to ensure that weighted sample distributions are close to known population distributions. However, for PES, such benchmark population data is not available, because PES is used in conjunction with census to estimate a new benchmark population. A further challenge to the application of design-based methods in PES is the absence of an accurate count of dwelling numbers by PSU at the time of the PES fieldwork, which complicates the computation of selection probabilities and hence sampling weights. Moreover, a sample size of about 15,000 households does not allow precise design-based estimates at the required level of geographic and demographic disaggregation. In this regard, the modelling approach seems natural, especially when geographical attributes are treated in a hierarchical fashion. Multilevel modelling facilitates pooling of information across areas and is desirable for small area and small domain estimation problems.

Modelling of survey data is, of course, possible from a design-based perspective, though there appear to be efficiency gains through explicitly modelling the survey design structure rather than dealing with the impact of survey design through sampling weights (Lumley and Scott 2017). Design-based multilevel modelling is challenging, because the pseudo-likelihood methods commonly used for design-based fitting of single-level models are more difficult to apply in the case of multilevel models. Pseudo-likelihood estimates of multilevel models are potentially sensitive to the scaling of survey weights, even when design and analysis clusters are identical (Rabe-Hesketh and Skrondal 2006). Methods based on pairwise composite likelihood are a promising alternative to pseudo-likelihood methods for fitting design-based multilevel models but require knowledge of joint selection probabilities (Rao et al. 2013; Yi et al. 2016). In the PES analysis the geographic clusters of analytical interest are the TAs, which were not part of the sample design and this further complicates the application of design-based methods to multilevel modelling (Lumley and Scott 2017).

5.4. *Mitigating the Ignorable Inclusion Assumption*

One of the fundamental assumptions of the PES model is the independence between inclusion in PES and inclusion in census, conditional on design features and covariates included in the model. This means that the list of dwelling addresses used for census and the PES sampling frame need to be built independently, a requirement sometimes difficult to satisfy. Another challenge to the conditional independence assumption is respondent behaviour. For instance, a respondent's negative experience with census might influence whether they open the door to PES interviewers. Such behaviour would lead to non-ignorable nonresponse and complicate the analysis by requiring that the model for inclusion in PES be explicitly formulated and included in the model fitting. Pfeiffermann et al. (2006) develops a conditional likelihood approach to incorporating non-ignorable

nonresponse in multilevel modelling of survey data. Extending the PES model to deal with non-ignorable nonresponse may be a worthwhile direction for future development of the model. In Figure 1, this would result in additional edges between one or more of the inclusion indicators and the coverage indicator, illustrating the need to explicitly specify the model for inclusion and to estimate the inclusion model jointly with the coverage model. Alternatively, it may be possible to incorporate external information that allows the assumption of conditional independence between PES and census inclusion to be weakened (Elliott and Little 2000; Brown et al. 2019).

6. References

- Atkinson, J., C. Salmond, and P. Crampton. 2019. *NZ Dep 2013 index of deprivation interim research report*. Technical report, Department of Public Health, University of Otago, Wellington. Available at: <https://www.otago.ac.nz/wellington/otago730394.pdf>. (accessed February 2022).
- Brown, J.J., C. Sexton, O. Abbott, and P.A. Smith. 2019. "The framework for estimating coverage in the 2011 census of England and Wales: Combining dual-system estimation with ratio estimation." *Statistical Journal of the International Association of Official Statistics* 35: 481–499. DOI: <https://doi.org/10.3233/SJI-180426>.
- Bryant, J., K. Dunstan, P. Graham, N. Matheson-Dunning, E. Shrosbree, and R. Speirs. 2016. *Measuring uncertainty in the 2013-base estimated resident population* (Statistics New Zealand Working paper: 16–04). Wellington, New Zealand: Statistics New Zealand. Available at: <https://www.stats.govt.nz/> (accessed February 2022).
- Chandrasekar, C. and W.E. Deming. 1949. "On a method of estimating birth and death rates and the extent of registration." *Journal of the American Statistical Association* 44: 101–115. DOI: <https://doi.org/10.1080/01621459.1949.10483294>.
- Chen, C., J. Wakefield, and T. Lumley. 2014. "The use of sampling weights in Bayesian hierarchical models for small area estimation." *Spatial and Spatio-Temporal Epidemiology* 11: 33–43. DOI: <https://doi.org/10.1016/j.sste.2014.07.002>.
- Chen, S.X., C.Y. Tang, and V.T. Mule Jr. 2010. "Local post-stratification in dual system accuracy and coverage evaluation for the US census." *Journal of the American Statistical Association* 105: 105–119. DOI: <https://doi.org/10.1198/jasa.2009.ap08404>.
- Chipperfield, J., J. Brown, and P. Bell. 2017. "Estimating the count error in the Australian census." *Journal of Official Statistics* 33: 43–59. DOI: <https://doi.org/10.1515/jos-2017-0003>.
- Elliott, M.R. and R.J. Little. 2000. "A Bayesian approach to combining information from a census, a coverage measurement survey, and demographic analysis." *Journal of the American Statistical Association* 95 (450): 351–362. DOI: <https://doi.org/10.1080/01621459.2000.10474205>.
- Gelman, A., J. Carlin, H. Stern, D. Dunson, D. Vehtari, and A. Rubin. 2014. *Bayesian Data Analysis*. Boca Raton, FL.: CRC Press.
- Gelman, A., and J. Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical models*. Cambridge: Cambridge university press.

- Gelman, A., A. Jakulin, M.G. Pittau, and Y.-S. Su. 2008. "A weakly informative default prior distribution for logistic and other regression models." *Annals of Applied Statistics* 2: 1360–1383. DOI: <https://doi.org/10.1214/08-A0AS191>.
- Gelman, A. and T.C. Little. 1997. "Poststratification into many categories using hierarchical logistic regression." *Survey Methodology* 23: 127–135. Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997002/article/3616-eng.pdf?st=76F1g34m> (accessed July 2022).
- Gelman, A., X.-L. Meng, and H. Stern. 1996. "Posterior predictive assessment of model fitness via realized discrepancies." *Statistica Sinica* 6: 733–760.
- Ghitza, Y., and A. Gelman. 2013. "Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups." *American Journal of Political Science* 57: 762–776. DOI: <https://doi.org/10.1111/ajps.12004>.
- Ghosh, M., K. Natarajan, T. Stroud, and B.P. Carlin. 1998. "Generalized linear models for small-area estimation." *Journal of the American Statistical Association* 93: 273–282. DOI: <https://doi.org/10.1080/01621459.1998.10474108>.
- Hogan, H.P. 1993. "The 1990 post-enumeration survey: Operations and results." *Journal of the American Statistical Association* 88: 1047–1060. DOI: <https://doi.org/10.1080/01621459.1993.10476374>.
- Lax, J.R., and J.H. Phillips. 2009. "How should we estimate public opinion in the States?" *American Journal of Political Science* 53: 107–121. DOI: <https://doi.org/10.1111/j.1540-5907.2008.00360.x>.
- Little, R.J. 2003. "The Bayesian approach to sample survey inference." In *Analysis of Complex Surveys*, edited by R. Chambers and C. Skinner: 49–57. John Wiley and Sons.
- Lumley, T., and A. Scott. 2017. "Fitting regression models to survey data." *Statistical Science* 32: 265–278. DOI: <https://doi.org/10.1214/16-STS605>.
- Molina, I., B. Nandram, and J. Rao. 2014. "Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach." *Annals of Applied Statistics* 8: 852–885. DOI: <https://doi.org/10.1214/13-A0AS702>.
- Mule, T., T. Schellhamer, D. Malec, and J. Maples. 2008. "Using continuous variables as modeling covariates for net coverage estimation." In *JSM Proceedings: Section on Survey Research Methods: 1941–1948*. Denver. Available at: <http://www.asasrms.org/Proceedings/y2008/Files/301279.pdf> (accessed February 2022).
- Nandram, B., L. Chen, and B. Manandhar. 2018. "Bayesian analysis of multinomial counts from small areas and sub-areas." In *JSM proceedings: Section on Survey Research Methods: 1140–1162*. Vancouver. Available at: <http://www.asasrms.org/Proceedings/y2018/files/867100.pdf> (accessed February 2022).
- Paige, J., G.-A. Fuglstad, A. Riebler, and J. Wakefield. 2020. "Design-and model-based approaches to small-area estimation in a low and middle income country context: comparisons and recommendations." *Journal of Survey Statistics and Methodology*. DOI: <https://doi.org/10.1093/jssam/smaa011>.
- Pavlou, M., G. Ambler, S. Seaman, and R.Z. Omar. 2015. "A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes." *BMC Medical Research Methodology* 15: 1–6. DOI: <http://doi.org/10.1186/s12874-015-0046-6>.

- Pfeffermann, D. 2013. "New important developments in small area estimation." *Statistical Science* 28: 40–68. DOI: <https://doi.org/10.1214/12-STS395>.
- Pfeffermann, D., F.A.D.S. Moura, and P.L.D.N. Silva. 2006. "Multilevel modelling under informative sampling." *Biometrika* 93: 943–959. DOI: <https://doi.org/10.1093/biomet/93.4.943>.
- R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/> (accessed February 2022).
- Rabe-Hesketh, S., and A. Skrondal. 2006. "Multilevel modelling of complex survey data." *Journal of the Royal Statistical Society*: 169: 805–827. DOI: <https://doi.org/10.1111/j.1467-985X.2006.00426.x>.
- Rao, J. and I. Molina. 2014. *Small-area estimation*. Hoboken, NJ: John Wiley & Sons, Inc.
- Rao, J., F. Verret, and M.A. Hidirolou. 2013. "A weighted composite likelihood approach to inference for two-level models from survey data." *Survey Methodology* 39(2): 263–282. Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/12-001-x2013002-eng.pdf?st=LsDJmvSV> (accessed July 2022).
- Rubin, D.B. 1987. *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: John Wiley & Sons.
- Shirley, K.E., and A. Gelman. 2015. "Hierarchical models for estimating state and demographic trends in US death penalty public opinion." *Journal of the Royal Statistical Society* 178: 1–28. DOI: <https://doi.org/10.1111/rssa.12052>.
- Si, Y., R. Trangucci, J.S. Gabry, and A. Gelman. 2020. "Bayesian hierarchical weighting adjustment and survey inference." *Survey Methodology* 46: 181–214. Available at: <https://www150.statcan.gc.ca/n1/pub/12-001-x/2020002/article/00003-eng.htm>. (accessed July 2022).
- Skrondal, A., and S. Rabe-Hesketh. 2009. "Prediction in multilevel generalized linear models." *Journal of the Royal Statistical Society*: 172: 659–687. DOI: <https://doi.org/10.1111/j.1467-985X.2009.00587.x>.
- Stan Development Team. 2020a. *R Stan: the R interface to Stan. R package version 2.21.2*. Available at: <http://mc-stan.org/> (accessed February 2022).
- Stan Development Team. 2020b. *Stan Modeling Language Users Guide and Reference Manual, version 2.25*. Available at: <http://mc-stan.org/> (accessed February 2022).
- Stats NZ. 2014. *Coverage in the 2013 Census based on the New Zealand 2013 Post-enumeration Survey*. Wellington: Statistics New Zealand. Available at: <https://www.stats.govt.nz/>. (accessed February 2022).
- Stats NZ. 2019. *Overview of statistical methods for adding admin records to the 2018 Census dataset*. Wellington, NZ: Statistics New Zealand. Available at: <https://www.stats.govt.nz/> (accessed February 2022).
- Stats NZ 2020a. *Estimated resident population 2018: Data sources and methods*. Wellington, NZ: Statistics New Zealand. Available at: <https://www.stats.govt.nz/> (accessed February 2022).
- Stats NZ. 2020b. *Post-enumeration survey 2018: Methods and Results*. Wellington, NZ: Statistics New Zealand. Available at: <https://www.stats.govt.nz/> (accessed February 2022).

- Vehtari, A., A. Gelman, and J. Gabry. 2017. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". *Statistics and Computing* 27: 1413–1432. DOI: <https://doi.org/10.1007/s11222-016-9696-4>.
- Yi, G.Y., J. Rao, and H. Li. 2016. "A weighted composite likelihood approach for analysis of survey data under two-level models." *Statistica Sinica* 26: 569–587. DOI: <https://doi.org/10.5705/ss.2013.383>.
- You, Y., and B. Chapman. 2006. "Small area estimation using area level models and estimated sampling variances." *Survey Methodology* 32: 97–104. Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9263-eng.pdf?st=a4rH5VTf> (accessed July 2022).
- You, Y. and P. Dick. 2004. "Hierarchical Bayes small area inference to the 2001 census undercoverage estimation." In *JSM Proceedings: Section on Government Statistics*: 1836–1840. Available at: <http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000377.pdf> (accessed February 2022).

Received July 2021

Revised December 2021

Accepted March 2022