Estimating and adjusting for over-coverage when using administrative data: Stats NZ's 2018 Census and future census model

Tracey Savage* – tracey.savage@stats.govt.nz
Patrick Graham – patrick.graham@stats.govt.nz
Anna Lin – anna.lin@stats.govt.nz
Abby Morgan – abby.morgan@stats.govt.nz

Stats NZ, Christchurch, New Zealand

Abstract

Stats NZ's Census Transformation strategy has two aims – to modernise the 2018 Census of Population and Dwellings and to investigate alternative census models for the future. Increasing the use of administrative data is a key enabler for both aims and consistent with Stats NZ's goal to be an administrative-data-first organisation.

In the absence of dwelling and person registers, or unique identifiers for these, we need to combine administrative data from a range of government agencies to enable these transformations. This has resulted in over-coverage of dwellings and people becoming key issues for us to deal with, in addition to under-coverage.

Traditional capture-recapture methods used to produce population estimates from a census rely on the assumption that there is no over-coverage of the target population. Given this, we have developed a new approach to population estimation in the presence of both over-coverage and under-coverage errors.

Keywords: coverage survey; integrated data; dwelling and people coverage; Bayesian inference.

1. Introduction

Stats NZ's Census Transformation strategy has two aims - to modernise the 2018 Census and to investigate alternative ways of producing small area population and social and economic statistics in the future (Statistics NZ, 2012). Increasing the use of administrative data is a key enabler for both aims and consistent with Stats NZ's goal to be an administrative-data-first organisation. The 2018 Census will make innovative use of existing administrative data where possible, particularly in the development of a New Zealand-wide address list and the subsequent enumeration of dwellings (Statistics NZ, 2016a). In addition, after initial feasibility research into a range of different census models, Stats NZ is actively working towards a future census based primarily on government administrative data (Statistics NZ, 2015).

In the absence of dwelling and person registers, or unique dwelling and person identifiers, we have to combine existing administrative data to enable these transformations. Quality issues in the individual data sources and the combined data – such as duplicate records, misclassification and linking errors – lead to over-coverage of the target population. This is a key issue which, alongside under-coverage, we must deal with for dwellings in our 2018 Census, and for both dwellings and people if we move to an administrative-based census model.

Traditional capture-recapture methods used to produce population estimates from a census, rely on the assumption that there is no over-coverage of the target population. When over-coverage is present, it needs to be estimated and adjusted for prior to population estimation – for example with a dependent over-coverage survey as part of the Post-Enumeration Survey (UNSD, 2010). However, concerns regarding the feasibility of implementing a dependent over-coverage survey lead us to consider alternatives.

Our starting point was the theory for modelling data from two lists in the presence of both under- and over-coverage errors, in conjunction with an independent coverage survey, developed by Zhang (2015).

We adapted this theory to apply it to a single list plus independent coverage survey scenario, under certain assumptions. Section 2 of the paper outlines the basic structure of the problem and our theoretical approach. Sections 3 and 4 discuss application of this theory to produce dwelling estimates from the 2018 Census, and population estimates from a future census based primarily on administrative data. Section 5 outlines our conclusions and next steps.

2. Theory

2.1 Basic structure of the problem. In order to describe the basic problem it is helpful to consider the two-way table that would result from cross-tabulating a target population with an overlapping list, as depicted in Table 1.

Table 1: Basic structure for a target population cross-tabulated with a list that overlaps the target population

		List		
		1	0	
Target	1	n_{11}	n_{10}	N_T
	0	n_{01}		
		N_L		

The only directly observable quantity in Table 1 is the total number of units (for example people or dwellings) on the list, N_L. An unknown number, n_{01}, of units on the list are not in the target population. These n_{01} units constitute over-coverage of the list with respect to the target population.

Since the list total N_L is directly observed, if we could determine the over-coverage n_{01} we would immediately be able to determine the number of units both in the target population and on the list as $n_{11}= N_L - n_{01}$. On the other hand, an unknown n_{10} units are in the target population but not on the list. This group represents the under-coverage of the list with respect to the target population. Given estimates of both n_{01} and n_{10} we could obtain an estimate of the target population total N_T as $\hat{N}_T = N_L - \hat{n}_{01} + \hat{n}_{10}$, where we use "^" to denote an estimate.

Notice that there are no units in the (0,0) cell. Units neither on the list nor in the target population are irrelevant to the problem of estimating the target population based on the list. Our conceptual starting point for estimation and analysis is the union of the target-population and the list, which is defined by the three occupied cells in Table 1.

Ultimately we would like to produce population estimates at a subgroup level, for example by age, sex, ethnic group, and region. Letting $\mathbf{X}$ denote covariates defining the subgroups, $\mathbf{x}$ a particular setting of these covariates, and Y the cell occupied by an individual unit known to be in the target-list union, a statistical model for the situation is $[Y \mid \mathbf{X}=\mathbf{x}] \sim \text{Multinomial}(1, \boldsymbol{\phi}(\mathbf{x}))$ where $\boldsymbol{\phi}(\mathbf{x}) = \left(\phi_{11}(\mathbf{x}), \phi_{10}(\mathbf{x}), \phi_{01}(\mathbf{x})\right)$ denote the cell probabilities, as shown in Table 2. The cell probabilities sum to one at each setting of the covariates.

Table 2: Underlying cell probabilities for the target-list union at setting x of the covariates

		List	
		1	0
Target	1	$\phi_{11}(\mathbf{x})$	$\phi_{10}(\mathbf{x})$
	0	$\phi_{01}(\mathbf{x})$	

From the three underlying cell-probabilities we can easily obtain probabilities of under- and over-coverage: $\phi^{under}(\mathbf{x}) = \phi_{10}(\mathbf{x}) / \left(\phi_{10}(\mathbf{x}) + \phi_{11}(\mathbf{x})\right)$ is the probability a target population unit with covariates $\mathbf{x}$ is not included on the list and $\phi^{over}(\mathbf{x}) = \phi_{01}(\mathbf{x}) / \left(\phi_{01}(\mathbf{x}) + \phi_{11}(\mathbf{x})\right)$ is the probability that a

unit included on the list is not in the target population. Given these probabilities it is possible to adjust the list for under- and over-coverage.

Since the three cell probabilities must sum to one at each $\mathbf{x}$ it is only necessary to model two of the cell probabilities. In practice it is convenient to model $\phi^{under}(\mathbf{x})$ and $\phi_{01}(\mathbf{x})$; the remaining probabilities can then be obtained as: $\phi_{11}(\mathbf{x}) = \left(1-\phi_{01}(\mathbf{x})\right)\left(1-\phi^{under}(\mathbf{x})\right)$; $\phi_{10}(\mathbf{x}) = \left(1-\phi_{01}(\mathbf{x})\right)\phi^{under}(\mathbf{x})$.

2.2 Inference for coverage probabilities. Suppose we are able to sample from the target population and link the sample to the list without error. Let $\lambda(\mathbf{x})$ denote the probability of inclusion in the sample for a unit with covariates $\mathbf{X} = \mathbf{x}$. If within levels of $\mathbf{X}$, the sample inclusion probabilities do not depend on list inclusion status, the probability structure underlying the sample-list union is as shown in Table 3. Compared to Table 2, it can be seen that the sampling of the target population transfers some people from the (1,1) cell to the (0,1) cell and from the (1,0) cell to the (0,0) cell. The (0,0) cell in the sample-list union is not observable and inference for coverage probability parameters based on the sample-list union must take this into account. The structure of Table 3 is not the same as a traditional capture-recapture population estimation problem, since in the latter situation two or more samplings from the target population are used to make inferences about population size, whereas in our case we have a single sampling that is linked to a list that overlaps the target population.

Table 3: Cell probabilities for the sample-list union assuming sample inclusion probabilities do not depend on list inclusion, within levels of x

		List	
		1	0
Sample	1	$\lambda(\mathbf{x})\phi_{11}(\mathbf{x})$	$\lambda(\mathbf{x})\phi_{10}(\mathbf{x})$
	0	$(1-(\lambda(\mathbf{x}))\phi_{11}(\mathbf{x})+\phi_{01}(\mathbf{x})$	$(1-\lambda(\mathbf{x}))\phi_{10}(\mathbf{x})$

If the sample inclusion probabilities, $\lambda(\mathbf{x})$, can be assumed known, a conditional likelihood for the coverage probability parameters can be constructed from the cell probabilities given in Table 3, after conditioning on being in one of the three observable cells. The conditional likelihood is given by:

$$L^{cond}(\phi) = \prod_{i:\tilde{Y}_i=(1,1)} \frac{\lambda(\mathbf{x_i})\phi_{11}(\mathbf{x_i})}{1-\left(1-\lambda(\mathbf{x}_i)\right)\phi_{10}(\mathbf{x}_i)} \prod_{i:\tilde{Y}_i=(1,0)} \frac{\lambda(\mathbf{x_i})\phi_{10}(\mathbf{x_i})}{1-\left(1-\lambda(\mathbf{x}_i)\right)\phi_{10}(\mathbf{x}_i)} \prod_{i:\tilde{Y}_i=(0,1)} \frac{(1-\lambda(\mathbf{x_i}))\phi_{11}(\mathbf{x_i})+\phi_{01}(\mathbf{x_i})}{1-\left(1-\lambda(\mathbf{x}_i)\right)\phi_{10}(\mathbf{x}_i)} \quad (1)$$

where $\tilde{Y}_i$ denotes the cell-location for the i^{th} unit in the observed sample-list union. The conditional likelihood can be maximised with respect to $\boldsymbol{\phi}$ or combined with a prior for $\boldsymbol{\phi}$ to yield a full posterior distribution for Bayesian inference.

In practice, survey samples designed to assess list coverage are likely to be obtained using multi-stage area based designs. Space limitations preclude a full exposition of inference for the coverage probabilities in this case, but in principle a structure such as that depicted in Table 3 can be considered to hold for each primary sampling unit (PSU), with hierarchical models used to pool information over the PSUs. Inclusion probabilities will typically vary by PSU, but depending on the design may be independent of covariates within PSUs.

3. Application 1 - 2018 Census: Dwelling coverage

The 2018 Census will make innovative use of existing administrative data where possible, particularly in the enumeration of dwellings (Statistics NZ, 2016). Land Information New Zealand (LINZ) and New Zealand Post data will be combined with 2013 Census dwelling addresses to develop a New Zealand-wide address list. This address list will form the basis of a large scale address canvassing exercise to clean the administrative list and create a list of dwellings for mail-out. Internet access codes will be mailed to approximately 80% of dwellings and their occupants, a major change from the 2013 Census

where forms were delivered to all dwellings by a collector. Combined with a 70% target for self-response (on-line or mail-back) in 2018, collectors will have much less direct contact with respondents. As a result, the level of misclassification of dwellings (dwelling / non-dwelling, and private / non-private dwellings) is expected to increase, as is the risk of duplicates and alternative addresses referring to the same dwelling. Each of these factors can cause over-coverage of dwellings. 2018 Census will include processes such as field follow up of non-responding dwellings to mitigate these risks and to minimise the over-coverage of dwellings.

In response to the potential over-coverage of dwellings in the 2018 Census, we are planning to adapt the field operation and statistical processes in our 2018 Post-Enumeration Survey (PES) to enhance detection of over-coverage and misclassification. We will also explore alternatives to the traditional dual system estimation method, which does not allow for over-coverage and misclassification errors.

In the 2018 PES we will conduct an independent enumeration of dwellings in a sample of contained geographical areas (PSUs) across New Zealand. We will then link this list of dwellings with the equivalent list of dwellings from the census. We will take a conservative approach to automatic linking in order to minimise the risk of false positive links, followed by manual linking to resolve false negative links. We will then attempt to identify or estimate the amount of coverage error (both under and over) in the PES. We are unable to assume there will be no coverage error in the PES enumeration of dwellings because we know from experience and international research that field enumeration processes are prone to coverage error (Eckman & Kreuter, 2011; Thompson & Turmelle, 2004), particularly those that have low or no contact with the public. Unless accounted for in our estimation processes, any coverage error in PES will be reflected as census coverage error (for example, a dwelling missed by PES but found by census will incorrectly reflected as census over-coverage).

We are currently considering two approaches to accounting for coverage error in the PES list of dwellings when estimating the coverage error in the census, both of which use the outlined theory to varying degrees. In both approaches the target population is all permanent private dwellings in New Zealand (with some minor exclusions for operational practicalities).

The preferred approach is to establish the PES list of private dwellings for a given PSU as the source of truth (the target population). This is achieved by manually resolving the true address type of any PES or census dwellings that are unable to be linked, or records that are linked on address but have different dwelling type classifications (private/non-private). The manual resolution process will use a range of available information (including administrative data, internet searching, and field intelligence) to establish the true state of the address, if it represents a permanent private dwelling and if it should be in the PES 'source of truth' list. This will allow us to identify if records in PES and not in census represent PES over-coverage or census under-coverage, and vice versa. The benefit of this approach is that it enables us to not only estimate the rates of over- and under-coverage, but it also supports the estimation of misclassification error as a contributing factor to the coverage errors.

Applying the theory of section 2 at PSU level, we have $\boldsymbol{\lambda} \equiv 1$, and so the situation simplifies back to that shown in Table 2 where census is the list and the PES is the target population for a given PSU. In this situation the likelihood simplifies to:

$$L^{cond}(\phi) = \prod_{i:\tilde{Y}_i=(1,1)} \phi_{11}(\mathbf{x_i}) \prod_{i:\tilde{Y}_i=(1,0)} \phi_{10}(\mathbf{x_i}) \prod_{i:\tilde{Y}_i=(0,1)} \phi_{01}(\mathbf{x_i}). \quad (2)$$

The overall likelihood is then obtained as the product of the likelihoods for each PSU.

The second approach being considered is to use the sub-sample of dwellings selected for PES interview to learn about dwelling coverage error in the PES enumeration of dwellings. If field intelligence identifies under-coverage in the PES list, the estimated sample inclusion probability $\boldsymbol{\lambda}$ will be less than one. Adjusting for over-coverage is more challenging as the theory presented in section 2 assumes no over-coverage in the sample (PES list in this case). Further work is required to determine whether the theory can be extended to support this approach.

4. Application 2 – Future administrative census: People coverage

For a future census based primarily on administrative data, one of the key research questions we need to answer is: Can linked administrative sources, with a coverage survey and statistical model, produce estimates of the New Zealand resident population, to a standard that will meet key customer requirements? The target population in this scenario is individuals who are resident in New Zealand at a specific reference date.

There is no person register, nor a unique national identifier for people in New Zealand. Instead, we are using Stats NZ's Integrated Data Infrastructure (IDI) as a test environment for our research. The IDI is a large research database containing de-identified microdata about people and households, from a range of government and non-government sources (Statistics NZ, 2016c). The IDI uses probabilistic linking to combine these sources – with a union of New Zealand birth, tax, and visa data as its central 'spine'– to create a broadly ever-resident population (Black, 2016).

To produce population estimates from the IDI, we first use activity recorded in the administrative data (for example income tax payments, school enrolment) to identify individuals who are resident in New Zealand at a given date. Comparing this administrative-based population list with the official estimated resident population (ERP) as at 30 June 2013 shows the administrative-based population agrees closely with the ERP at the national level. However, there is evidence of coverage errors at some ages – for example under-coverage of children, and over-coverage of males aged 19-55 (Statistics NZ, 2016b).

We therefore need to estimate and adjust for coverage errors in the administrative-based population list to produce accurate population estimates. Our research is currently focused on applying the theory described in section 2 using a single sample from the target population (with a complex multi-stage area-based survey design) to assess list coverage. We assume that this coverage survey sample is linked without error to the population list, to create the sample-list union as illustrated in Table 3.

We are investigating two different approaches to implementing the theory described in section 2: (i) Directly using the multinomial likelihood of equation (1), and (ii) an approximate approach involving separate estimation of under-coverage and the probability for the $\tilde{Y}=(0,1)$ cell in the sample-list union (Table 3), followed by back-calculation of the latter probability to recover an estimate of $\phi_{01}(\mathbf{x})$. For the first approach we parameterise the model in terms of $\phi^{under}(\mathbf{x})=\phi_{10}(\mathbf{x})/\left(\phi_{11}(\mathbf{x})+\phi_{10}(\mathbf{x})\right)$ and $\phi_{01}(\mathbf{x})$ and specify hierarchical Bayesian logistic models for these probabilities. The hierarchical models account for the complex survey design by including PSU and stratum indicators in the model, using a similar approach to that described in Bryant et al (2016). While this modelling approach is theoretically correct, it is computationally demanding.

The second approach estimates under-coverage (φ^{under}) using a hierarchial logistic model and a separate logistic model for the probability for the (0,1) cell in the observed sample-list union – that is we model $\tilde{\phi}_{01}(\mathbf{x})=\Pr(\tilde{Y}=(0,1)\,|\,\mathbf{X}=\mathbf{x},\tilde{Y}\neq(0,0))$. Equating the estimated probabilities obtained from this model to the (0,1) cell probability in Table 3 yields:

$$\phi_{01}(\mathbf{x})=\frac{(1-\lambda(\mathbf{x}))\phi_{11}(\mathbf{x})+\phi_{01}(\mathbf{x})}{1-(1-\lambda(\mathbf{x}))\phi_{10}(\mathbf{x})} \qquad (3)$$

After writing $\phi_{11}(\mathbf{x})=\left(1-\phi_{01}(\mathbf{x})\right)\left(1-\phi^{under}(\mathbf{x})\right)$; $\phi_{10}(\mathbf{x})=\left(1-\phi_{01}(\mathbf{x})\right)\phi^{under}(\mathbf{x})$, and substituting for estimated values of $\phi^{under}(\mathbf{x})$ obtained from the under-coverage model, we can obtain estimates of $\phi_{01}(\mathbf{x})$ by solving (3) for $\phi_{01}(\mathbf{x})$, assuming the sample inclusion probabilities are known at each setting of $\mathbf{x}$. In practice, these inclusion probabilities may also need to be modelled and the resulting estimates substituted in (3). This second approach takes much less computing time than option one described above, but is not guaranteed to return estimates of $\phi_{01}(\mathbf{x})$ that fall in the (0,1) interval.

As a first step, we are testing these approaches for a simplified scenario of perfect linking between the administrative-based population list and the coverage survey, and no misclassification error in the administrative data (for example in geographic location, or ethnic group). This involves simulating an administrative-based population containing both under- and over-coverage error, and a coverage survey, using 2013 Census data as the source for both.

5. Conclusions

We have developed a new population estimation theory that, under certain assumptions, can adjust for both under- and over-coverage errors in a population list, yet requires only a single sample from the target population. To date, this theory appears very promising, but we need to do a significant amount of work to further extend the theory, implement and test it, and determine how best to apply it in practice to produce 2018 Census dwelling estimates, and population estimates from a future administrative based census. The next steps for our work are as follows:

(i) extend the theory to deal with misclassification of variables on the list and linkage error between the survey and list, and explore strategies to improve the efficiency of our computations

(ii) confirm how we will proceed for the 2018 PES, then explore implementation of the theory in the context of dwelling coverage and start defining the estimation model

(iii) complete testing for an administrative based census, and publish an experimental series of modelled population estimates from the IDI, using Stats NZ's Household Labour Force Survey as a proxy coverage survey.

References

Black, A. (2016). The IDI prototype spine's creation and coverage. (Statistics New Zealand Working Paper No 16–03). Retrieved from www.stats.govt.nz.

Bryant, J., Dunstan, K., Graham, P, Matheson-Dunning, N., Shrosbee, E., & Speirs, R. (2016). Measuring uncertainty in the 2013 base estimated resident population (Statistics New Zealand Working Paper No 16–04). Retrieved from www.stats.govt.nz.

Eckman, S., & Kreuter, F. (2011). Confirmation bias in housing unit listing. *The Public Opinion Quarterly*, *75*, 1, 139–150.

Statistics NZ (2012). Transforming the New Zealand Census of Population and Dwellings: Issues, options, and strategy. Retrieved from www.stats.govt.nz.

Statistics NZ (2015). Census transformation – a promising future. Cabinet paper (redacted). Retrieved from www.stats.govt.nz.

Statistics NZ (2016a). 2018 Census Strategy. Retrieved from www.stats.govt.nz.

Statistics NZ (2016b). Experimental population estimates from linked administrative data: methods and results. Retrieved from www.stats.govt.nz.

Statistics NZ (2016c). Integrated Data Infrastructure. Retrieved from www.stats.govt.nz.

Thompson, G., & Turmelle, C. (2004). Classification of address register coverage rates – a field study. In *Proceedings of the section on Survey Research Methods, American Statistical Association*, 4477–4484. Retrieved from ww2.amstat.org.

Zhang, L-C. (2015). On modelling register coverage errors. *Journal of Official Statistics, 31, (3)*, 381–396. Retrieved from www.degruyter.com.

United Nations Statistics Division (UNSD) (2010). Post enumeration surveys: Operational guidelines – Technical Report. Retrieved from http://unstats.un.org.