The Survey Statistician, 2020, Vol. 82, 27-39.

New and Emerging Methods

Exploring developments in population size estimation

James Brown[1], Christine Bycroft[2], Davide Di Cecco[3], Joane Elleouet[4], Gareth Powell[5], Viktor Račinskij[6], Paul Smith[7], Siu-Ming Tam[8], Tiziana Tuoto[9], and Li-Chun Zhang[10]

[1] University of Technology Sydney, james.brown@uts.edu.au
[2] Statistics New Zealand, christine.bycroft@stats.govt.nz
[3] Sapienza Università di Roma, davide.dicecco@uniroma1.it
[4] Statistics New Zealand, Joane.Elleouet@stats.govt.nz
[5] Office for National Statistics (UK), gareth.powell@ons.gov.uk
[6] University of Southampton, vr1v14@soton.ac.uk
[7] University of Southampton, P.A.Smith@soton.ac.uk
[8] Australian Bureau of Statistics and University of Wollongong, stattam@gmail.com
[9] Instituto Nazionale di Statistica, tuoto@istat.it
[10] University of Southampton, L.Zhang@soton.ac.uk

Abstract

This short paper covers some of the new and emerging research in the area of population size estimation in official statistics presented at a workshop in February 2020. It covers work exploring the replacement of traditional census with administrative data sources. These sources typically suffer from both under-coverage and over-coverage and this paper covers the application of dual- and multi-system estimation to tackle coverage errors. These estimation frameworks depend on linkage across different population lists and surveys, and the issue of adjusting for linkage error is also tackled. Finally, some discussion is given to current research in the context of population census for the 2020/21 round and the use of coverage surveys and administrative data to improve census outputs.

Keywords: dual-system estimation, record linkage, over-coverage, administrative data, census

1 Introduction

In February 2020, before the global pandemic, a group of 25 academic and official statisticians meet in University of Technology Sydney to share practical and methodological developments across several countries, as they prepare for the 2020/21 round of population censuses, and the future beyond. The workshop drew from Australia, Ireland, Italy, Netherlands, New Zealand, and UK. In this short paper, we discuss some of the emerging topics presented and discussed during the week.

It complements both the ‘ask the expert’ contribution by van der Heijden and Cruyff (2020) and the ‘new and emerging methods’ contribution by Tam and Holmberg (2020).

One focus is the increasing push to move beyond traditional census and utilise administrative data sources in countries where there is not a strong history of population registers and administrative data use in the compilation of population statistics. This creates a range of methodological challenges. A secondary focus is the desire to utilise administrative data sources within the traditional census framework.

2 Administrative-Based Population Counts

A common theme across several countries is the move for administrative lists to replace a traditional census in countries that do not have a national population register. The Office for National Statistics (ONS) is pursuing the production of Administrative data Based Population Estimates (ABPEs) by linking at a record level several broad coverage administrative data sources and applying rules to count individual records in or out of the ABPE. The focus in the ABPE construction and rules has been to remove as much over-coverage as possible and then to measure the under-coverage in the ABPE with an estimation methodology to produce population size estimates.

ONS plans to run a population coverage survey that will operate in a similar way to the traditional Census Coverage Survey (CCS), which enables the estimation of under-coverage in the traditional Census. A high-quality address frame is being developed which will be used as an address sampling frame for ONS surveys. The survey operation will be like the CCS with an emphasis on collecting variables useful for linkage to the ABPE (name, date of birth, address), ensuring that data are protected by appropriate privacy and security safeguards. However, it will be mixed mode with an online first, self-completion approach, followed by face to face or telephone where necessary. A similar response size to the CCS of around 300,000 addresses is required in order to provide sufficient sample to enable good quality local authority level estimates. The PCS will be a voluntary survey, in common with other social surveys in the UK, and from tests in 2018/19 ONS expect to achieve around 60% response rate. Consequently, the number of sampled addresses would need to be around 500,000 per year to provide the number of responses required to match CCS numbers. This represents a considerable ongoing data collection so the longer-term intention is to integrate the PCS with the Labour Market Survey (LMS) where the PCS questions will appear in a first wave along with some core LMS questions. A sub-sample of Wave 1 responses will be used for subsequent waves for labour market and other survey questions.

A particular challenge in the construction of linked administrative sources is working with hashed data. This is a requirement for some of the potential administrative data sources and creates a challenge from a linkage perspective. A clear recommendation coming from research presented by ONS is that data linkage is carried out using variables ‘in-the-clear’ wherever possible. However, for those instances where this is not possible, ONS presented the development a deterministic matching method, called “Derive and Conquer”, for linking hashed data. This method identifies pairs of matching records using the power of distributed computer processing and makes decisions based on derived agreement variables, which combine information on multiple derived variables. Work is ongoing to assess the quality of this method, but ONS anticipates that the results of linking hashed data will be of poorer quality compared with linking the data in the clear. However, if data suppliers are only willing to give us hashed personal identification data, then it is important to understand the changes in linkage rates, accuracy and bias. This will inform the decision on whether the linked data can be used in the production of official statistics.

Statistics New Zealand (SNZ) is working also towards an administrative-based census in the longer term. An administrative New Zealand resident population has been developed from linked administrative sources in SNZ's Integrated Infrastructure (IDI). Comparison with official population estimates shows that the administrative population is a good approximation (Stats NZ 2017), but includes some under-coverage and over-coverage. Over-coverage is dealt with through the application of rules relating to activity on the component administrative sources but that still leaves the under-coverage to be estimated.

As the work presented by ONS demonstrates, to estimate under-coverage the PCS needs to be very large, and even then, the provision of regular census-like population statistics will be limited by the level of disaggregation supported by the PCS. The concept of *Fractional counting* (Zhang, 2019a) presented at the workshop provides the theoretical foundation for a *system* of statistical data and estimation, which can accommodate all the relevant information and enables one to produce detailed statistics according to different definitions of the target population as required by users. The attraction is the starting point of an extended population dataset (EPD) that contains all possible individuals from linking across multiple administrative data sources, and therefore negligible population under-coverage. Such an EPD is already shown to be feasible in a number of countries, such as Estonia (Tiit and Maasing, 2016), Latvia (CSBL, 2019) and New Zealand (Stats NZ, 2018), but under-coverage is typically induced by the application of rules to remove over-coverage. In this case, rather than inducing under-coverage to target over-coverage, for every individual in the EPD, the following probabilities are applied. (i) the probability of belonging to the target population, (ii) the probability of correct 'official' address in EPD given (i), and (iii) the probability of being present at other addresses shown in the EPD given (i) and not (ii), including 'unknown' as a possible address. These probabilities are the *fractional counters* that are used to produce detailed population statistics, adjusting for over-coverage, rather than the removal of whole records with the risk of induced under-coverage.

3 Exploring the Dual-System Estimation Framework

Use of dual-system estimation is clearly still at the centre of approaches to estimate populations when faced with imperfect population lists, generated by census, administrative systems, and surveys. Below we have the simplest scenario estimating from two lists, A and B.

	In List B	Missed List B	
In List A	N_{11}	N_{10}	N_{1+}
Missed List A	N_{01}	??	
	N_{+1}		$\widehat{N} = \frac{N_{1+}N_{+1}}{N_{11}}$

Zhang (2019b) reminds us of the crucial assumptions behind the dual-system framework but the dual-system approach remains attractive due to its simplicity. There is no model selection for relationships between lists, as independence is the only possible identifiable hypothesis. Even if one source has extremely heterogeneous capture probabilities, no bias is introduced (see Chao *et al*., 2011). The paper by van der Heijden and Cruyff (2020) explained how the framework could deal with missing covariates in its component lists. Contributions to the workshop particularly focused on the issues of linkage error and over-coverage in (administrative) lists.

3.1 Dealing with Linkage Error

Accurate record linkage is crucial for the reliable performance of the dual system estimator. It is arguably the most difficult and resource intensive stage prior to the population size estimation. There is a growing body of work looking at both the estimation of linkage errors (for example Chipperfield *et al*., 2018) as well as developing adjustments for estimated linkage error (for example Tuoto, 2016).

One approach to estimating linkage errors is to sample from the links and non-links. Sampling from the linked pairs, we can estimate the false linkage rate (FLR), and sampling from the non-linked pairs estimate the missing match rate (MMR). In both cases one only needs to verify whether the sampled pairs are true matches or not. However, a practical difficulty of such an approach to MMR is that the number of non-linked pairs can be enormous, while the true matches among them can be very few in comparison, such that the chance of observing a true match by random sampling is almost zero.

Returning to our linkage of list A of size N_{1+} to list B of size N_{1+}, the unknown set of true outcomes are the three cells N_{11} (true matches), N_{10}, and N_{01}; where records in N_{10} can never be correctly linked to records in N_{01}. One approach being developed to estimating the missing links presented at the workshop takes inspiration from the trimmed dual-system estimator (Zhang and Dunne, 2017) by repeating the linkage process under different structures and comparing results.

Taking linking strategy 1, the set of links D_1 contains L_1 links; while for strategy 2 the set of links D_2 contains L_2 links. The intersection of links, those linked pairs in both D_1 and D_2, contains L_{12}. We have two (under-)estimates of the total number of links that should exist between the two lists, so we can use dual-system estimation to estimate

$$\widehat{N}_{11} = \frac{L_1 L_2}{L_{12}} \tag{1}$$

However (1) is biased if there are false positive links in either L_1 or L_2, and if the two linkage procedures are not independent of each other. But estimating the FLR is possible by, for example, auditing from the linkages L_1 and L_2 and therefore we can adjust the number of linkages down to give $\tilde{L}_1$, $\tilde{L}_2$, and $\tilde{L}_{12}$. We trim the number of links to remove estimated false positives. The result is a revised estimate for the number of true links N_{11} given by

$$\widetilde{N}_{11} = \frac{\tilde{L}_1 \tilde{L}_2}{\tilde{L}_{12}}. \tag{2}$$

Insofar as linkage errors originate from the errors of linkage key variables, one can simply use *different* key variables for the two linkage processes, helping to approximate independence between the linking outcomes. At the workshop, simulations under the following three settings were presented by Zhang and Tuoto to demonstrate the approach:

- Setting 1: D_1 based on Name, Surname and Gender; D_2 based on Day, Month and Year of Birth.
- Setting 2: D_1 based on Name, Surname and Day of Month; D_2 based on Day, Month and Year of Birth, where the common linkage variable Day of Month is *not* affected by errors.
- Setting 3: D_1 based on Name, Surname and Gender; D_2 based on Name, Month and Year of Birth, where the common linkage variable Name is affected by errors.

As expected, estimator (2) of N_{11} is unbiased under setting 1, almost unbiased under setting 2 and biased under setting 3. Ongoing research aims to identify practical diagnostics for independence of linkage procedures, which can provide additional confidence to the prior choice of linkage variables.

Of course, the procedure one applies to link A and B in the end does not need to be the ones used for estimating the number of matches that one is targeting.

Presentations by Statistics New Zealand (SNZ) took a different route to estimating the missed matches. In the assessment of the 2018 Census, SNZ linked the Census to their administrative system. While the link rate is high with almost 98 percent of census responses linked to the administrative system, perfect linkage cannot be assumed. At the workshop, work was presented adjusting for linkage errors within the dual-system framework. Following the simplest model of Ding and Fienberg (1994), we assume no incorrect links (false positives), and estimate the rate of missed links (false negatives) within each stratum. In the case of the 2018 Census, the MMR was estimated by first restricting census returns to those that were very highly certain, based on their characteristics, to be in the administrative file. In other words if the census is list A then N_{10} is defined to be zero for this subset of returns. It is then trivial to estimate

$$\eta = \frac{\Pr(unlinked|match)}{\Pr(linked|match)}, \tag{3}$$

the ratio of linkage probabilities conditional on the true match status using this subset of the census returns. The overall rate of missed matches is estimated as 1.21 percent.

Missed links are added back into the N_{11} cell, and removed from the off-diagonal cells to preserve the marginal totals, leading to another adjusted dual-system estimator

$$\widehat{N} = \frac{N_{1+}N_{+1}}{N_{11}(1+\eta)}. \tag{4}$$

The results based on (4) presented at the workshop showed very plausible patterns of population structure and have given a valuable indication of the patterns of under-coverage remaining in the final 2018 Census dataset in the interim before official population estimates can be released.

An alternative approach to the linkage error problem presented at the workshop explores the theoretical and practical limits of the linkage free dual system estimation (Račinskij *et al*., 2019) The method utilises the parameters of the linkage error model directly in estimation, rather than attempting to resolve the linkage status of all possible record pairs. By exploring the close relationship between dual system estimation and record linkage, it shows that in certain cases the population size estimator can be expressed as a function of the estimated linkage parameters. This relationship allows, at least in theory, dual system estimation without the classification of record pairs into links and non-links. As a result, no one-to-one linkage or extensive clerical classification is required, but naturally, a certain loss of precision occurs, and a substantial modelling effort is needed.

Tam and Holmberg (2020) gives a method to integrate big data with survey data but this depends on high quality linkage. In the framework presented here, U is the population of interested with known size N. For units $i \in U$, y_i is the outcome of interest; and for a (large) subset of the population units $i \in B \subset U$ of size N_B, y_i is observed. Membership of this alternative data source is identified by the indicator $\delta_i = 1$. When you over-lay a sample A selected from U with inclusion probabilities $1/d_i$, Tam and Holmberg (2020) suggest one strategy is an approximately design-unbiased post-stratified estimator given by

$$\hat{Y}_{PS} = \sum_{i \in U} \delta_i y_i + (N - N_B)\frac{\sum_{i \in A} d_i(1-\delta_i)y_i}{\sum_{i \in A} d_i(1-\delta_i)}. \tag{5}$$

The population membership indicator δ_i for the units found in B is determined by a linkage process to a population frame, which under-pins the sample A. This facilitates the detection of duplicates in B as well as the removal or records that do not belong to U. However, in the presence of linkage errors, we observe $\hat{\delta}_i$ as the outcome of the linkage process and replacing δ_i with $\hat{\delta}_i$ in (5) results in a bias.

At the workshop, work presented the adjustment

$$\hat{Y}_{PS} = \sum_{i \in U} \delta_i y_i + (N - N_B) \frac{\sum_{i \in A} d_i (1 - \delta_i) y_i - (1 - R) \sum_{i \in B} y_i}{\sum_{i \in A} d_i (1 - \delta_i) - (1 - R) N_B}, \quad (6)$$

where $R = Pr(\hat{\delta}_i = 1 | \delta_i = 1)$. This returns (6) to an approximately unbiased estimator provided inclusion in the sample A is independent of inclusion in B, and further inclusion in the sample A is independent of the outcome of the linkage process given the true status of inclusion in B. Simulation results demonstrated R can be estimated using bootstrapping, using the methods outlined in Chipperfield *et al*. (2018).

3.2 Over-coverage with Two Lists

SNZ has been developing an alternative Bayesian model to estimate coverage of an administrative list using an Administrative Census Coverage Survey (ACCS), with early results presented at the workshop. The novelty of the proposed approach is in the simultaneous estimation of over-coverage and under-coverage of the administrative list in an extension of the dual-system estimation framework.

Specifically, the system in which the model operates is the union of the target population and the administrative list. Let N_U be the union size. Individuals are then either in both the target and the list (N_{11}), in the undercount of the target population (N_{10}) or the overcount (N_{01}), with $N_U = N_{11} + N_{10} + N_{01}$. Transforming counts into inclusion probabilities φ = (φ$_{11}$, φ$_{10}$, φ$_{01}$), which depend on individual covariates X (demographic and geographic characteristics), we have the following table of individual cell probabilities for a given x Є X:

		Admin list	
		1	0
target	1	φ$_{11}$(x)	φ$_{10}$(x)
	0	φ$_{01}$(x)	0

We now conduct the ACCS selecting areas from the target population, where λ(x) is the sample inclusion probability of an individual (given that the individual's target population area was selected in the ACCS), which we assume varies with X. Linking between the coverage survey and the administrative list, we now have the following cell probabilities:

		Admin list	
		1	0
ACCS	1	λ(x)φ$_{11}$(x)	λ(x)φ$_{10}$(x)
	0	φ$_{01}$(x) + (1−λ(x)) φ$_{11}$(x)	(1−λ(x))φ$_{10}$(x)

In the inevitable case where λ(x)≠1 (due to both among- and within-dwelling non-response), the distribution of λ over X is unknown. Simulation experiments presented at the workshop showed that the model and data as currently set up provide no information to estimate λ simultaneously with φ.

Therefore, potential options to address the inherent limitations of 2-list population estimation in the presence of over-coverage still rest on refining rules and administrative record quality for inclusion in the administrative list to reduce over-coverage to a negligible level. The trimmed dual-system approach (Zhang and Dunne 17) offers a strategy provided multiple rules can be defined for implementing trimming from the administrative list.

3.3 Over-coverage with Multiple Lists

When more than two lists are available, log-linear models are the common choice to extend the dual-system framework, as they allow us to manage different capture probabilities in different lists, and to model the dependencies among captures of a same individual in different lists. The inclusion of a categorical latent variable constitutes a simple extension to account for unobserved (latent) individual heterogeneity in the capture probabilities, and log-linear latent class models are indeed successfully used in capture-recapture literature (see for example Biggeri *et al*., 1999; Stanghellini and van der Heijden, 2004; Thandrayen and Wang, 2010).

A different use of the latent variable discussed at the workshop is to model specific subpopulations, in the case of tackling over-coverage identifying the class out-of-scope units. The idea of using a binary latent variable to model over-coverage originates in various works of Biemer (see chapter 6.3 of Biemer *et al*., 2011), and has been explored in detail in Di Cecco *et al*. 2018 and Di Cecco (2019). We need at least four sources for any latent class model to be identifiable, but, in population count, it is not common to have more than three sources available. In cases where a certain number of administrative sources are used to produce a single statistical register, a simple way to increase the number of lists is to treat separately some of the sources composing the register. Another trick consists in including in the model some of the covariates such as gender, age and nationality of the units. In this way, we increase the number of manifest variables and we can tune the dependencies among all the variables as an alternative to simple stratification. Finally, we can include as capturing lists the so-called "signs of life", such as individuals' workplace or place of study, which will obviously come with a lot of over-coverage, but would provide useful information and contribute to the total number of lists and of identifiable models. In this case, some units have null probability of capture such as people outside working age. A way to overcome this problem is to treat the capture profiles of these units as partially observed, that is, to treat the information about the capturing status of certain units as missing at random (Di Cecco *et al*., 2018).

These approaches are being explored in the context of the Italian Permanent Census Strategy. The overall strategy integrates population registers and sample surveys so as to obtain the usual resident population counts at the reference date, while removing coverage errors from the register. Two sample surveys are carried out to integrate with the population register: an area sample survey mainly designed to estimate the under-coverage, and a register based sample survey mainly planned to integrate information not recorded in the population register. Dual-system estimation is used to evaluate the usual resident counts, after adjusting for people erroneously included in the register (over-coverage). The register-based sample provides the adjustment for over-coverage but this relies heavily on the field contact data, where non-response must be classified as either genuine non-response by a usual resident or non-response because the record on the register no longer relates to a current usual resident.

Research at the workshop presented a pilot using a latent class model, with full details reported in Fortini *et al*. (2020). The model integrates the 2018 register based survey data, affected by both missing values and response errors, with administrative signals of employment, school attendance and pension payments (signs of life) in order to estimate over-coverage rates for relevant profiles of

statistical units for the 14 largest Italian municipalities. Manifest variables used for estimating the model parameters are: A, Age classes (4 classes); C, Citizenship (2); H, Household size (2); M, Municipality code (14); S, Sign of life (10) from administrative data and F, Field contact result (3). The first category of Field contact includes people confirmed as resident by interviewers during the survey, either as a respondent or direct non-respondent (refusal, unable). The second category consist of people checked as not resident, such as those moved or deceased, using proxy information collected by interviewers in the field or by municipal officers through local auxiliary data. The third class encompasses all other sampled individuals lacking evidence from the field to be classified. The Signs of life variable S collects information from many education, employment and welfare administrative sources in 10 classes with the aim to rank people according to whether or not their usual residence corresponds to their legal residence.

A two class latent variable (X) is defined to classify usual vs not usual residents and the following path analysis model with latent variables is used to fit the multivariate table of observed variables.

$$\Pr(F,S,H,A,C,M) = \sum_X \Pr(F|M,X)\Pr(S|A,X)\Pr(H,A,C,M,X). \quad (7)$$

Apart from conditional independence relationships induced by the factorization in (7) on the full, not observable table, this model relies on the following further assumptions:

1. the structural part of the model $\Pr(H,A,C,M,X)$ accounts for relationship between X and the other variables except for indicators S and F;
2. structural model assumes a hierarchical interaction configuration $\{XHM, XAM, XCM, HACM\}$;
3. in the observed table, structural zeros are considered in every cell where H*E combination is "0-19 years age" and "single-parent household" respectively, to consider that (almost) all Italian people dwelling in a household and falling in age class ‘0-19’ actually live with their parents;
4. the measurement part of the model $\Pr(F|M,X)$ and $\Pr(S|A,X)$ consists in two logit functions between indicators F and S with the latent variable X respectively;
5. logit models regarding indicators F and S also consider variables M and A as predictors in addition to latent variable X, assuming only marginal effects to ensure model identification;
6. marginal probability $\Pr(F=1|X=1)$ to be classified as usual resident in the household by the interviewer during field contact (F=1) is fixed to zero given that person is actually a not usual resident (X=1). This assumption is made to force X into providing usual residence status, assuming that the status assigned during the interview is by far the most reliable information collected by interviewers. In this way, about 90% of the sample in assigned to usual residence (X=2) without error.

The presented model utilises 538 parameters leaving 6181 degrees of freedom.

Overall, the model provides coherent and reasonable results. This approach can help in understanding the strength and limitations of administrative signals and how they can be used to support statistical surveys. A key result is that the risk of over-coverage defined by the latent class model computed for each large municipality is often lower than the corresponding estimates obtained by simply leaving out people lacking field contact evidence (F=3), under the ‘missing at random’ assumption for missing information. Concerning the capability of Field contact (F) to identify over-coverage cases, while class 1 (people resolved as usual resident) is error free by design, class 2 and 3 show a probability to truly identify over-coverage between 50-60%, with little difference

between them. This fact suggests that proxy information is in fact not much better than no contact for identifying over-coverage.

This application demonstrates that as we include more sources, we can use more complex models, but, in general, we face more methodological challenges. In particular, model selection is a notoriously challenging problem as different models with similar goodness of fit can lead to extremely different estimates of the population size. Results from both simulated data and real data shows that model misspecification can produce severely biased estimates, and classic tools for model selection like different information criteria (Bayesian, Akaike, Deviance) can indicate completely different models as preferable. However, a Bayesian approach to this class of models is relatively simple as outlined in Di Cecco *et al*. (2020). It allows the introduction of prior knowledge (on the capture probabilities of each source, or on the population size to be estimated) and leads to straightforward construction of interval estimates (HPD) of the population size, as well as of any other quantity of interest. In addition, it allows the implementation of model averaging techniques, which provides a natural way to address the problem of model selection, providing at the same time a robustification of the estimates to model misspecification (Di Cecco, 2020).

The use of a latent variable to model a specific subpopulation is not new, and is often open to criticism, as it is necessary to get some validation on its interpretation. A clear example of validation is given by inspecting the posterior probabilities of belonging to the two latent classes for those units captured in all lists. In the context of identifying over-coverage, if they have an almost equal posterior probability of belonging to the two classes, our interpretation will not be defensible. Conversely, if they belong with almost certainty to one class, that constitutes evidence in favour of our model. This approach is strengthened when one of the lists is a survey that explicitly identifies usual residents as in the application above.

4 Enhancements to the 2020/21 Round of Population Census

There is considerable ongoing work in preparation for upcoming censuses that looks to develop the coverage assessment of census, and utilise administrative data throughout the census process.

4.1 Census Coverage

Statistics New Zealand presented their work assessing the coverage of the 2018 Census by comparison to their Administrative Population (Stats NZ, 2019). Over-coverage in the administrative list is removed by applying stricter rules for being included in the administrative population, at the expense of removing true members of the population. However, increased under-coverage is not a problem in the dual-system context. The approach to dealing with potential heterogeneity is a classic post-stratification into small sub-populations based on an individual's location and characteristics. The dual-system estimator is calculated within strata defined by single year of age, sex, ethnic group, and an intermediate sub-national geography[1]. A limitation of this approach is that many small cells are created by these strata. One consequence is that estimation by ethnicity is restricted to only three ethnic groups: Maori, Pacific, and Asian. The Chapman correction factor[2] is applied to avoid a zero denominator and reduce bias due to small numbers.

[1] Territorial authority and Auckland Local Boards

[2] The Chapman correction factor $\widehat{N}_T = \frac{(N_{1+}+1)\,(N_{+1}+1)}{(N_{11}+1)} - 1$

ONS are exploring developing their 2021 coverage estimation to go beyond classic post-stratification in Brown *et al*., 2019 based on geography and age-sex for controlling heterogeneity. This is possible because the census data for the entire country will be available quicker than it was in 2011. This will result from increased electronic collection and less reliance on scanning for data capture. No batching into Estimation Areas (EA) is needed and the estimation can be carried out either at the national or at the regional levels, similar to the approach used for example in the 2010 US Census (US Census Bureau, 2008; US Census Bureau, 2012). A sufficiently large sample of data permits the census coverage for a reasonably large set of characteristics to be modelled using logistic or mixed effects logistic regression models as proposed in Alho (1990) and Alho *et al*. (1993). Estimated census non-response weights (reciprocals of estimated census response probabilities) can then be applied to each census observation with the corresponding characteristics (US Census Bureau, 2012). Similar to the concept of fractional counting for an administrative file, summing up all weighted census observations with the characteristic of interest will produce an estimated population size of units with the characteristic.

Research indicates that weights based on mixed effects logistic regression can achieve higher overall accuracy both at the national and subnational levels. Modelling census coverage this way, allows controlling for effects other than age, sex, hard to count defined on geographic areas; and therefore has the potential to deal with heterogeneity bias more efficiently. Research shows that, in the 2011 Census methodology, the lack of control for variables such as ethnicity and tenure could result in relative bias of at least -0.2% at the national level (with a target of the absolute relative bias at the national level not to exceed 0.5%). A second gain is that the coverage modelling naturally embeds different geographic levels such as Local Authority (LA) level, avoiding a direct synthetic assumption as in Baffour *et al*. (2018), that patterns of under-coverage for each LA within an EA were constant after just controlling for age, sex, and hard to count.

For estimated (net) under-coverage, ONS will again adjust the final census database. The adjustment will use a two-stage approach for 2021: 1) Impute missed households (and persons within them) and 2) Impute characteristic variables for the persons and households imputed in stage 1. In 2011, missed persons were also first imputed into counted households. While the 2011 Census adjustment methodology based on Steele *et al*. (2002) worked well to provide a census database that agreed with coverage estimates, the implementation was challenging. An evaluation of the methodology concluded that alternative approaches should be explored for the 2021 Census focusing on the selection of donor households for imputation and their placement within the adjusted database.

Combinatorial Optimisation (CO) (Voas and Williamson, 2000) is an integer programming method which involves finding the best (an optimal) combination from a finite set of combinations for a problem. In the context of the adjustment problem, CO involves the selection of a combination of households from the existing census database that best fit the coverage estimate benchmarks. As part of the 1st stage, the CO selection of donor households are placed in a geographic location. Administrative data sources are being considered here as they may be able to provide additional information about census dummy forms (forms for addresses where unable to enumerate, but only collecting basic information), which are used as placeholders for donors. Coverage estimates of the census population will be provided in more detail in 2021, so CO will constrain its selection to a wider range of key demographic characteristics at a lower geographical level, which would not be possible with the calibration approach implemented in 2011.

4.2 Improving Census Returns

Several presentations at the workshop looked at the integration of administrative data to enhance the the traditional census process. For example, Australian Bureau of Statistics (ABS) presented work exploring the use of administrative sources to help enhance fieldwork processes, as well as the potential to repair a census return. ONS are exploring variable replacement. For example, a new admin-based number of rooms variable will be produced from linked Valuation Office Agency (VOA) data, which similar to all other census variables will need to undergo edit and imputation. Recent research has therefore focused on whether the assumptions underpinning nearest-neighbour donor-based imputation, typically used for item imputation in census, are still valid when using linked administrative data. For example, although the reasons for missingness will be different to traditional surveys, missingness must still be predictable from the observed data for the method to be valid. Research so far supports the use of donor-based methods, so this will be implemented with linked admin-census data in 2021 for the first time.

The experience in New Zealand in 2018 demonstrates clearly that the common move towards a "digital-first" census impacts on census returns, and therefore on the design of the edit and imputation strategy. Research has shown that characteristics and non-response rates differ between online and paper respondents, which introduces a risk of bias during imputation. Imputation will therefore need to be conditioned by response mode to avoid propagation of characteristics from one mode to another. ONS are exploring the use of linked administrative data to allow for searching the census data for donors conditional on age, even when age is itself missing. The method involves using an 'admin age' variable from linked administrative data as an additional matching variable when searching census returns for donors during imputation. Proof of concept has so far been successful but further work is required around the practical implementation of this method before it can be applied in 2021.

5 Concluding Remarks

There is extensive research being undertaken across the world of offcial statistics to better enhance the use of administrative data in the efficient production of timely official statistics. The current pandemic only reinforces that need. The work presented at the workshop focused particularly on the production of population estimates in both the context of administrative data only and census enhamced with administrative data; demonstrating a range of exciting approaches, especially in the areas of linkage and over-coverage.

References

Alho, J. (1990) Logistic Regression in Capture-Recapture Models. *Biometrics*, **46**, 623-635.

Alho, J., Mulry, M., Wurdeman, K., & Kim, J. (1990) Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation. *Journal of the American Statistical Association*, **88**, 1130-1136.

Baffour, B., Silva, D., Veiga, A., Sexton, C., & Brown, J. J. (2018) Small area estimation strategy for the 2011 Census in England and Wales. *Statistical Journal of the IAOS*, **34**, 395-407.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.). (2011) *Measurement errors in surveys*. John Wiley & Sons.

Biggeri, A., Stanghellini, E., Merletti, F., & Marchi, M. (1999) Latent class models for varying catchability and correlation among sources in Capture-Recapture estimation of the size of a human population. *Statistica Applicata*, **11**, 563-576.

Brown, J. J., Sexton, C., Abbott, O., & Smith, P. A. (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 IAOS*, **35**, 481-499.

Chao, A., Tsay, P. K., Lin, S. H., Shau, W. Y., & Chao, D. Y. (2001) The applications of capture-recapture models to epidemiological data. *Statistics in medicine*, **20**, 3123-3157.

Chipperfield, J., Hansen, N., & Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. *International Statistical Review*, **86**, 219-236.

CBSL (2019). Method Used to Produce Population Statistics. Central Statistical Bureau of Latvia. https://www.csb.gov.lv/sites/default/files/data/15_04_ 2019_Iedz_Metodologija_ENG.pdf.

Di Cecco, D., Di Zio, M., Filipponi, D., & Rocchetti, I. (2018) Population size estimation using multiple incomplete lists with overcoverage. *Journal of Official Statistics*, **34**, 557-572.

Di Cecco, D. (2019). Estimating population size in multiple record systems with uncertainty of state identification. *Analysis of Integrated Data*, 169-196. Chapman and Hall/CRC.

Di Cecco, D. (2020); Bayesian Model Averaging for Latent Class Models in Capture-Recapture, *50th Scientific Meeting of the Italian Statistical Society*, University of Pisa, 22-24 June 2020.

Di Cecco, D., Di Zio, M., & Liseo, B. (2020). Bayesian latent class models for capture-recapture in the presence of missing data. Submitted.

Ding, Y., & Fienberg, S. E. (1994) Dual system estimation of Census undercount in the presence of matching error. *Survey methodology*, **20**, 149-158.

Fortini, M., Bernardini A., Caputi M., & Cibella N. (2020) Combining "signs of life" and survey data through latent class models to consider over-coverage in Capture-Recapture estimates of population counts. *50th Scientific Meeting of the Italian Statistical Society*, University of Pisa, 22-24 June, 2020.

Račinskij, V., Smith, P. A., & van der Heijden, P. (2019) Linkage Free Dual System Estimation. arXiv:1903.10894.

Stanghellini, E., & van der Heijden, P. G. (2004) A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account. *Biometrics*, **60**, 510-516.

Stats NZ (2017) Experimental population estimates from linked administrative data: 2017 release. https://www.stats.govt.nz/experimental/experimental-population-estimates-from-linked-administrative-data.

Stats NZ (2018) Integrated Data Infrastructure. https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure.

Stats NZ (2019) Dual system estimation combining census responses and an admin population https://www.stats.govt.nz/methods/dual-system-estimation-combining-census-responses-and-an-admin-population.

Steele, F., Brown, J., & Chambers, R. (2002) A controlled donor imputation system for a one-number census. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **165**, 495-522.

Tam, S. M., & Holmberg, A (2019). New data sources for official statistics – a game changer? *The Survey Statistician*, **81**, 21-35.

Thandrayen, J., & Wang, Y. (2010). Capture–recapture analysis with a latent class model allowing for local dependence and observed heterogeneity. *Biometrical Journal*, **52**, 552-561.

Tiit, E.-M., & Maasing, E. (2016). Residency index and its applications in censuses and population statistics. *Eesti statistika kvartalikri* (Quarterly Bulletin of Statistics Estonia), **3/16**, 41-60.

Tuoto, T. (2016) New proposal for linkage error estimation. *Statistical Journal of the IAOS*, **32**, 413-420.

US Census Bureau (2008). 2010 Census Coverage Measurement Estimation Methodology. https://www.census.gov/coverage_measurement/pdfs/2010-E-18.pdf.

US Census Bureau (2012). 2010 Census Coverage Measurement Estimation Report: https://www.census.gov/coverage_measurement/pdfs/g10.pdf.

Van der Heijden, P., & Cruyff, M. (2020) Wider applications for dual and multiple estimation. *The Survey Statistician*, **81**,16-20.

Voas, D., & Williamson, P. (2000) An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata. *International Journal of Population Geography*, **6**, 349-366.

Zhang, L. C., & Dunne, J. (2017). Trimmed dual system estimation. In *Capture-recapture methods for the social and medical sciences*, 237-257. CRC Press.

Zhang, L. C. (2019a) *On provision of UK neighbourhood population statistics beyond 2021.* Report for ONS.

Zhang, L. C. (2019b) A note on dual system population size estimator. *Journal of Official Statistics*, **35**, 279-283.