Stats NZ
Tatauranga Aotearoa

Dual system estimation combining census responses and an admin population

New Zealand Government

Citation
Stats NZ (2019). *Dual system estimation combining census responses and an admin population*. Retrieved from www.stats.govt.nz.

ISBN 978-1-98-858345-7 (online)

IDI disclaimer
Stats NZ accessed the data in the IDI for use in the 2018 Census in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation and the results in this paper have been confidentialised to protect these groups from identification.

Careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in Privacy impact assessment for the Integrated Data Infrastructure and Creating the 2018 Census dataset by combining administrative data and census forms data: Our privacy impact assessment.

Published in September 2019 by
Stats NZ Tatauranga Aotearoa
Wellington, New Zealand

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

www.stats.govt.nz

Contents

List of tables and figures

List of tables

List of figures

Purpose and summary

Purpose

Dual system estimation combining census responses and an admin population describes how this estimation process has been used to derive a new estimate of the New Zealand resident population at the time of the 2018 Census, with a focus on meeting key assumptions that underpin the method.

Summary of key points

A lower than expected response rate to the New Zealand 2018 Census led to the development of new methods to add administrative records to the census dataset to count those who had been missed by the census field collection. A reliable estimate of the New Zealand resident population at the time of the census was needed to guide the development of the new methods, and to assess their performance.

Dual system estimation (DSE) combines two independent lists to estimate the size of a population. Census responses and an administrative New Zealand resident population provide two large but incomplete population lists, which have been created through quite different mechanisms. Applying DSE to these two lists overcomes limitations in the available official population estimates for this purpose. However, DSE is vulnerable to biased population estimates if key assumptions that underpin the method are not met. DSE based on census and administrative data raises different challenges from those usually encountered by national statistics offices when applying DSE in the context of a census and a coverage survey. The most critical assumptions to address are erroneous inclusions in the administrative population and missed linkages between the two data sources. In both cases, violation of the assumptions leads directly to an upward bias in the DSE estimates.

We developed a method to remove erroneous records from the administrative data, and a linkage error adjustment was applied to the DSE calculation. Variance of the DSE estimates is very small due to the large size of the contributing datasets. We expect that some bias remains.

Census day benchmark population estimates by age and sex, for key ethnic groups, and the larger sub-national geographies, have been produced.

The DSE results showed very plausible patterns of population structure, and have provided a sound basis for testing different options for non-response mitigation. They have also given an early indication of the patterns of under-coverage remaining in the final 2018 Census dataset, although these cannot be confirmed until the official coverage measurement is available in 2020.

Introduction

A lower than expected response rate to the New Zealand 2018 Census led to the development of new methods to add administrative records to the census dataset to count those who had been missed by the census field collection Stats NZ (2019b). A reliable estimate of the New Zealand resident population at the time of the census was needed to guide the development of the new methods, and to assess their performance. Population estimates were needed for demographic breakdowns by age, sex and ethnic group, and for sub-national geographies.

The available official population statistics for March 2018 were based on the previous census in 2013. Their quality decreases over time, mainly due to uncertainty measuring external and internal migration. In addition, official population estimates are not updated in between censuses for

changes in ethnic populations, apart from indigenous Māori ethnicity. Five years on, the 2013-based estimates were not considered accurate enough for our purpose, while the updated official statistics based on the 2018 Census would not be available until the census dataset was completed.

However, we did have two independently constructed datasets that captured parts of the New Zealand population. The census forms received through the 2018 Census field collection provided a large, although incomplete, section of the population. We had also developed a New Zealand resident population derived from linked administrative data sources which was a good approximation to the New Zealand population, but with unknown levels of under-coverage and over-coverage. Both data sources were independent of the 2013 Census and overcame the limitations of the 2013-based population estimates as they were referenced directly to the March 2018 census date and included ethnicity.

Dual system estimation (DSE), sometimes known as capture-recapture, provides a framework for combining two partial lists of a target population to estimate the total population. DSE is a well-established methodology used in population estimation, often applied by national statistics offices in the context of a census and a coverage survey (for example, Brown et al, 2018; Mule, 2012; Stats NZ, 2014), and widely used in ecology and epidemiology. DSE relies on several assumptions. The main focus of this paper is to describe the approaches we developed to meet key DSE assumptions in the novel context of combining census responses and an administrative-derived population. DSE calculations are also constrained by the amount of detail that can be obtained for ethnic groups and sub-national geographies.

We first describe the two data sources. The methods section outlines the basic premise of DSE, and describes in detail the methods used to remove erroneous inclusions and to adapt the DSE calculation for missed linkages. Results are followed by a discussion and conclusion.

Data sources

We now describe the two population lists that contribute to the DSE.

2018 Census responses

The first list is made up of New Zealand usual residents who completed a 2018 Census form.

The New Zealand census is conducted on a de facto, or 'persons present', basis. The census target population is everyone in New Zealand on census night. People who usually live in New Zealand but are overseas on census night are excluded from the census target population. Overseas visitors who are in New Zealand are asked to complete census forms, but these visitors are not of interest here. When people responded to the census from a dwelling they were visiting on census night, they are asked where they usually live.

The main uses of census data relate to New Zealand residents, and where they usually live in New Zealand, not where they may be on census night. Accordingly, the target population for the DSE is New Zealand usual residents, who are present in New Zealand on census night.

We note that New Zealand's official population estimates, the estimated resident population (ERP), measures the entire resident population. The ERP is based on the census counts but includes those New Zealand residents who are temporarily overseas on census night. It also adjusts for net census undercount. The ERP series is regularly updated for population change since the latest census.

Census forms were completed during the collection phases of the 2018 Census. Forms could be submitted through the online collection system or by completing and returning a paper form. The online collection method was the predominant method used by respondents in the 2018 Census. An online Household Set-up form and paper Dwelling form requested that all people present at the dwelling on census night be listed by name, and provide their age, sex, and relationship to the reference person. A separate Individual form was to be completed for each person listed. This did not always happen, and some people were listed as being present in a dwelling on census night, but no Individual form was received from them.

People whose only information came from the household listing were considered to be a census response. This distinction matters when using the DSE because the quality of linking variables differs in each case. In addition, people who only appear on the household listing do not complete the ethnicity question, an important population variable. For reasons discussed in more detail later, we exclude census responses obtained only from the household listing from the DSE. Only census responses from Individual forms are included in the DSE.

Approximately 4 million Individual forms were received for the 2018 Census, and the individuals who completed these forms comprised the first list for the DSE calculations. While there appeared to be a systematic undercount across the whole population, it was clear that non-response was more concentrated in some geographic areas, and that groups who are typically harder to count in any census were disproportionately affected in 2018.

An administrative New Zealand resident population

The second list is a New Zealand resident population derived from the linked administrative data in Stats NZ's Integrated Data Infrastructure (IDI).

The IDI is a large research database that holds microdata about people and households. Data are gathered from a range of government agencies, Stats NZ surveys and the 2013 Census, and non-government organisations. The data are linked together, or integrated, to form the IDI.

See the Integrated Data Infrastructure for further information.

The basic structure of the IDI consists of a central 'spine' to which the other data collections are linked at the individual level (Black, 2016; Gibb, Bycroft, & Matheson-Dunning, 2016). Broadly, the target population for the spine is all individuals who have **ever** been residents of New Zealand. Thus the spine comprises a reference list designed to include nearly everyone who interacts with the range of data sources that are included in the IDI, with the exception of those who only enter New Zealand as short-term visitors.

The spine is made up of the union of people in three data sources:

- all births registered in New Zealand since 1920
- all visas granted to migrants since 1997 (excluding visitor and transit visas)
- all individuals issued with an IRD (tax) number.

In 2018 the IDI spine included approximately 10 million individuals found in one or more of these sources. Migrants from Australia, Cook Islands, Tokelau and Niue do not require a visa, and are only included in the spine once they are issued with a tax number.

Because the IDI spine aims to include anyone who has ever lived in New Zealand, it also provides the basis for selecting a resident population at a given point in time. We constructed an estimate of the New Zealand resident population at a given date from the multiple linked administrative data sources in the IDI (Stats NZ, 2017). This 'admin population' known as the IDI-ERP, includes those individuals in the IDI spine who have activity in selected administrative data sources (health, tax including benefit payments, education enrolment, and Accident Compensation Commission) over a two-year period up to the reference date. Those who died before the reference date are identified by a link to death registrations data and are excluded. International border movements data is used to exclude anyone who was not a New Zealand resident on the reference date, for example a resident who migrated to live overseas, or a former resident who went to a doctor while on a temporary visit.

For the 2018 Census application, we derived the IDI-ERP for census day, 6 March 2018, and made some improvements to previous versions of the IDI-ERP. Stats NZ (2019b). has further details of the IDI-ERP derived for 2018 Census use. We removed New Zealand residents who were temporarily overseas on census night to achieve the equivalent target population to that of the census.

The IDI-ERP is a good approximation to the New Zealand resident population (Stats NZ, 2017, 2018), however there are some coverage errors:

- over-coverage – people who were not in fact New Zealand residents at the reference date, but have been wrongly included
- under-coverage – people who were New Zealand residents and should be included but have not been selected.

DSE accounts for those missing from the admin population but cannot adjust for over-coverage. Over-coverage should therefore be removed before DSE can be applied.

Linking census responses to the IDI

Record linkage attempts to find pairs of records from two data sources that refer to the same person. Record pairs for the same person are true 'matches'. Record pairs that refer to different people are 'non-matches'. A formal description can be seen in Fellegi and Sunter (1969).

As with other data sources that are linked within the IDI, responses from census forms have been linked to the central population list, the IDI spine. The linkage itself is undertaken in a secure linking environment[1] between identified census records and an identified population list that is the basis of the IDI spine[2]. Probabilistic linkage based on the models developed by Fellegi and Sunter is applied in a fully automated process. A one-to-one linkage is enforced, so that it is important to remove duplicate records from both data sources before the linking takes place. The variables used to compare record pairs are: first names and last names, date of birth and age, sex, country of birth, meshblock, and address identifier. Multiple administrative meshblocks are used in the linkage to provide the best opportunity of finding a match with the census meshblock.

The criteria for accepting a census form as a response requires valid values for two of the following three variables: meshblock, date of birth, and name. This criteria supports high quality linkages, since in New Zealand an exact match on two of those variables is highly discriminatory. The linkage process is designed to minimise incorrect links at the expense of missing more true matches. Stats NZ (2019a) provides details of the linkage process. Table 1 sets out the main results of the linkage process.

Table 1

Results from linking 2018 Census responses to the IDI spine

Type of response	Count	Number of links	Link rate (percent)
Online	3,457,410	3,411,924	98.7
Paper	515,175	497,613	96.6
Household listing	202,857	170,163	83.9
NZ residents total	**4,175,445**	**4,079,703**	**97.7**

Note: The 2018 Census figures reflect data available when the files were linked, which differs slightly from final census counts.

The overall link rate of 97.7 percent is very high in the IDI context. The quality of the available variables used to compare records affects our ability to link a census response to the IDI spine. The slightly lower link rate for paper forms compared to online responses reflects difficulties scanning

[1] Only variables used in the linkage process are available, and access is highly restricted to staff carrying out the linkage. See Creating the 2018 Census dataset by combining administrative data and census forms data: Our privacy impact assessment for further details.

[2] The IDI spine is the same admin population list but with direct identifiers such as name and full date of birth removed. The IDI spine records are given a unique anonymised identifier, an snz_uid, that is used as a linking variable throughout the IDI.

handwritten name and date of birth information from paper forms. The lowest link rate of 83.9 percent for census responses taken from the household listing is due to less detailed information being available for the linkage, since sex and age are provided, but the full date of birth, and country of birth are not.

Since the admin New Zealand resident population is selected from the IDI spine, this linkage process automatically provides links between the census responses and admin population used in the DSE.

Two types of errors can occur in a record linkage process:

- false negative matches – records that belong to the same person, but the link has not been made
- false positive matches – when records for different people are incorrectly linked.

Both types of linkage error affect the DSE calculation, and the interaction between them is complex. The presence of false negative matches will inflate the DSE calculation, while false positive matches will tend to decrease the DSE.

Around 0.6 +/-0.3 percent of the linkages are estimated to be incorrect, an estimate derived from manually checking a small sample of linked records. We do not have an estimate of the false positive rate with respect to true non-matches and assume the impact on the DSE is negligible.

An estimated 1.21 percent of Individual forms have not been linked when a true match to the administrative spine record is available. How we measured the rate of false negative matches and adjusted the DSE to account for these is described later in this paper.

Method

This section provides a description of the dual system estimation methodology, together with the approaches developed to meet two key DSE assumptions.

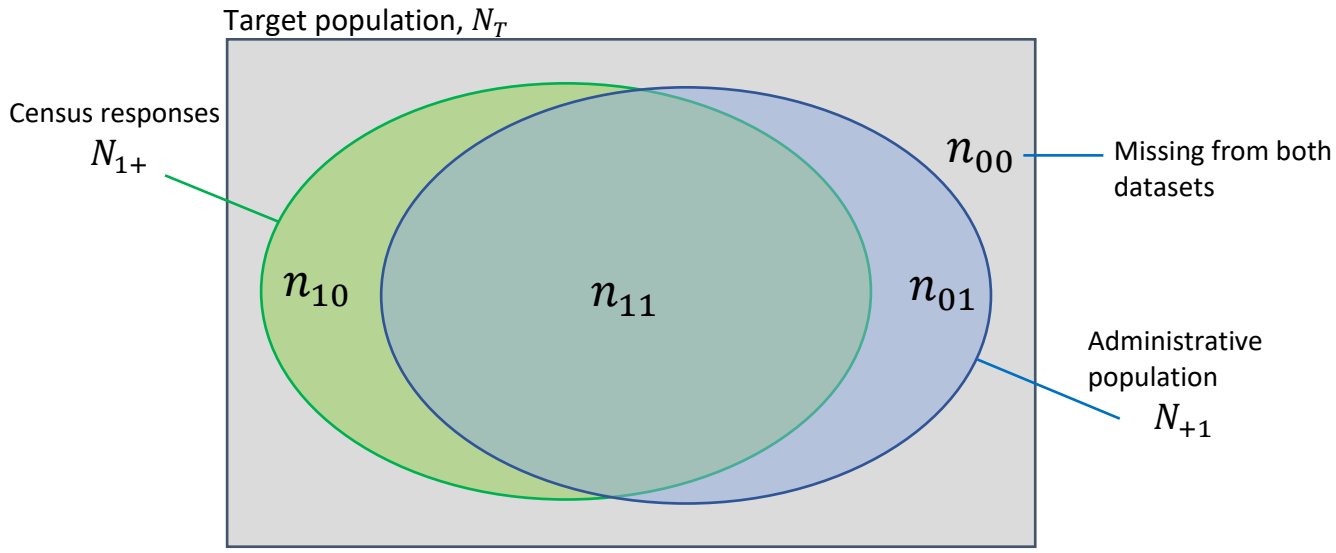
Dual system estimation

The basic premise of the DSE methodology is using the combination of two partial lists, or captures, of a target population to identify units that are included on both lists, and units that are included on only one of the lists. A statistical model can then be applied to estimate the units missed by both lists.

Figure 1 shows the union of the two lists: the received census responses, and the admin population. Those in the target population but missing from both lists are also shown. The target population is the census operational definition – New Zealand usual residents present in New Zealand on census night.

Figure 1

Missed matches between the 2018 Census and admin records



The data problem is typically presented in a contingency table as in table 2, where a subscript 1 infers presence in a list, and a 0 infers a record not observed in the list. The cell n_{00} is not observed and must be estimated to obtain an estimate of the total population N_T.

Table 2

Contingency table for dual system estimation cell counts				
		Admin population		
		1	0	Total
Census responses	1	n_{11}	n_{10}	N_{1+}
	0	n_{01}	n_{00}?	
	Total	N_{+1}		N_T

N_{1+} is the number entering the DSE from the census responses

N_{+1} is the number entering the DSE from the admin population

n_{11} are the linked records, that is, observed in both census responses and the admin population

$n_{10} = N_{1+} - n_{11}$ are records in the census responses but not linked to the admin population

$n_{01} = N_{+1} - n_{11}$ are records in the admin population but not linked to the census responses

The n_{00} cell represents those missing from both lists, and is estimated from the observed cells as

$$\hat{n}_{00} = n_{10}n_{01}/n_{11}$$

The estimated total population, $\widehat{N}_T$, is the sum of the three observed cell counts plus the estimated number missed by both lists

$$\widehat{N}_T = n_{11} + n_{10} + n_{01} + \hat{n}_{00}$$

This is equivalent to the standard Lincoln-Petersen estimator (Petersen, 1896; Lincoln, 1930) which is often expressed as

$$\widehat{N}^{LP} = N_{1+}N_{+1}/n_{11}$$

These calculations can be applied within strata defined by variables present on both lists.

The variables that we wish to produce breakdowns of the population for are used as strata for the estimation. The DSE calculation is run at the level of single year of age, sex, selected ethnic groups, and local authorities (TALB[3]), so that estimates can be summed over these groups to obtain required population estimates.

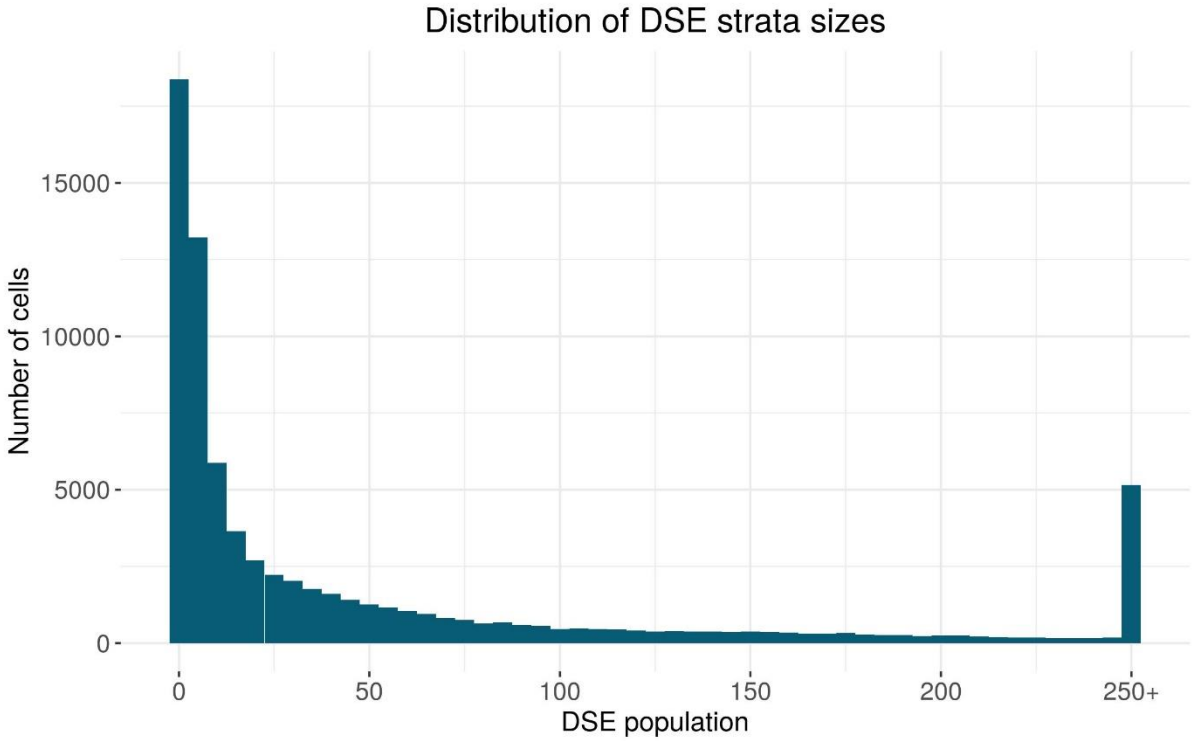
Estimates for geographies smaller than TALB were not considered because the accuracy of administrative location information decreases with small geographies (Stats NZ, 2017). We rely on individuals updating their address information with government agencies to identify where they

[3] Territorial authorities and Auckland local boards (TALB) are the main local administrative areas in New Zealand. They include cities, rural districts, and local boards within the largest city of Auckland. Population sizes range from around ten thousand people to several hundred thousand.

usually live. For many people this produces an accurate location for their usual residence, however for those who move frequently, such as young adults, the information is less reliable.

Applying these cell breakdowns produces up to 126,672 distinct groups: 91 single year of age groups (0 to 90+), 2 sexes (male and female), 8 ethnic groups, and 87 TALBs. In practice, 40 percent (51,094) of the potential combinations have no people in them, and others are very small. The DSE is run for each combination of these groups where individuals are observed. The distribution of strata sizes greater than zero (in groups of five) is shown in figure 2. Of all possible combinations, 61 percent (76,698) have fewer than five observations, and 74 percent (93,572) have fewer than 20 observations. However, these smaller strata contribute only 5 percent of the total observed population.

Figure 2



To account for bias in the DSE introduced by small cell sizes, the Chapman correction (Chapman, 1951) is applied to the Lincoln-Petersen estimator:

$$\widehat{N}_T = \frac{(N_{1+}+1)\,(N_{+1}+1)}{(n_{11}+1)} - 1$$

Wittes (1972) provides a variance estimator of the DSE with Chapman correction as:

$$\frac{(N_{1+}+1)(N_{+1}+1)\,n_{01}n_{10}}{(n_{11}+1)^2(n_{11}+2)}$$

These are calculated for each of the DSE strata and can be summed to obtain estimates of variance by aggregated groups.

The Chapman estimator generally performs better than the Lincoln-Petersen estimator on bias and variance, but may have a negative bias for small strata or when one list has a low probability of

capture (Sadinle, 2008). The difference will accumulate when aggregating over strata. This may lead to a downward bias in the DSE estimates for small sub-groups such as small TALBs, and for the very oldest age groups.

The DSE calculation produces results that are not whole numbers. The decimal places are kept for each strata and results are truncated after aggregating the appropriate cells.

Ethnic group

The New Zealand standard for ethnicity (Stats NZ, nd) defines ethnicity as the ethnic group or groups that people identify with or feel they belong to. Ethnicity is self-perceived, and people can belong to more than one ethnic group. Determining the ethnic groups to apply in the DSE is complicated because the concept of ethnicity allows for multiple responses, while DSE variables must be mutually exclusive so that individuals can be placed in only one cell. Level 1 of the ethnicity classification includes six groups: European, Māori, Pacific, Asian, Middle Eastern/Latin American/African (MELAA), and Other. Mutually exclusive combinations of these level 1 groups must be determined in a way that provides estimates for the key ethnic groups, while not introducing too many very small cells for the DSE calculation.

Estimates of Māori, Pacific and Asian ethnic groups are key performance indicators for the census, and eight mutually exclusive ethnic group combinations were formed to provide DSE estimates for each of them. This means that there are no DSE estimates for the European ethnic group, or for the smaller MELAA and Other groups. The ethnic group combinations are detailed in the appendix.

Missing data and conflicting information

When records are missing values for age, sex, ethnic group, or TALB information, we cannot be certain which DSE cell they belong to.

Age and sex information for the administrative records have no missingness because these are collected from various datasets in the IDI. A small proportion of administrative records are missing ethnicity or address information.

For census records, age, sex, and TALB variables are asked on all census forms, and the small number missing from census responses are statistically imputed in the census dataset. Missing ethnicity is populated from links to 2013 Census and from administrative sources, where possible. The remaining missing ethnicity values are statistically imputed in the census dataset. In the DSE we used ethnicity derived from 2013 Census or administrative sources as this is real data supplied by the same person, though at a different time or different context.

We removed all census records with statistically imputed values, and administrative data with missing information so that these did not enter the DSE calculation. There are limitations to this approach, as we must assume that records with missing stratification values are removed proportionally from the linked and off-diagonal unlinked cells, to avoid introducing bias into the DSE estimates.

Census responses obtained only from the household listing are problematic in this situation. They have no ethnicity information from the census form, and only have an ethnicity from alternative sources if they are linked to the IDI. The linkage rate for these records is lower than for responses obtained from Individual forms, and all unlinked household listing responses have missing ethnicity. For household listing records, most missing ethnicity data will be in the n_{10} unlinked cell. We therefore excluded **all** census responses obtained only from the household listing from entering the

DSE calculation, to avoid disproportionately removing records with missing ethnicity from the unlinked cell.

Table 3 shows the number of records within each list with data missing from the census responses and admin data used in the DSE, for each of the variables required for DSE groups.

Table 3

Number of records missing estimation variables in the 2018 Census Individual forms and admin population		
Variable	Census Individual forms	Admin population
Age	2,988	0
Sex	4,752	0
Ethnic group	2,487	22,908
TALB[(1)]	6,429	12,867
1. Territorial authority and Auckland local board area		

We note that an individual may appear twice in the table if that person is missing information for more than one of the variables listed.

In cases where a linked record pair has different values for any of the variables used to define the DSE strata, we assumed the census value was correct, and used the census value to determine the estimation sub-group for the linked record pair.

The values for unlinked administrative records cannot be verified through a census response, and any differences will flow through to the final DSE results. Incorrect administrative TALB values in the unlinked administrative records would tend to induce under-count and over-count when comparing areas.

DSE assumptions

Applications of DSE commonly invoke a number of assumptions:

- a closed target population (no opportunity for people to enter or leave the population of interest)
- independence between the two lists (the likelihood of being recorded on one list has no relationship with the likelihood of being recorded on the other)
- homogeneity of capture of individuals (all individuals have the same likelihood of being captured in a list)
- no erroneous inclusions in either of the two lists (no people included that are not part of the target population)
- perfect linking between the two lists.

If these assumptions are not met, population estimates may be biased. The first three assumptions are relatively straightforward to achieve because of the nature of the lists we have in this application. For the requirement of no erroneous inclusions, we developed an approach for removing over-coverage from the admin population. We were not able to achieve perfect linking, but do assume negligible impact from false positive matches, where a link is made between records that are not a true match. We applied an adjustment to the n_{11} cell to account for incorrectly missed matches.

Each assumption is discussed below in more detail.

Closed target population

A closed population means there is no change in the population due to births, deaths, and migration. We assumed this condition holds for both lists. Those who answered census forms have declared themselves to be residents on census day, and we defined the admin IDI-ERP population as at the reference census date.

Independence

The IDI-ERP is derived independently of participation in the census, so we can safely assume that there is no causal relationship between inclusion on either list. We also assumed that the likelihood of being in the IDI-ERP (ie interacting with government services) has no relationship with the likelihood of returning a census form, though we had no suitable third list available to test this. Given that the census and admin data collection processes are practically independent, we assumed that they are statistically independent and that there is no relationship between inclusion on either list.

Homogeneity

The DSE requires homogeneity of capture of individuals on at least one of the lists. At the total population level, we know that people who return census forms are not homogeneous, and that they vary by such features as age, sex, ethnicity, and location. We also see heterogeneity of capture by demographic variables in the admin population.

We controlled homogeneity through stratification by the same variables that we used to produce estimates of population sub-groups, as described above. We assumed that individuals within each single year of age, sex, ethnic group, and TALB combination have the same likelihood of being captured. This stratification already produces many small strata. We did not consider it feasible to include any further breakdowns that might provide better control for homogeneity, as small cells can bias estimates.

No erroneous inclusions

Erroneous inclusions can be caused by including the same person more than once in the dataset, and by including people who do not belong to the resident population (over-coverage). The DSE is designed to adjust for under-coverage, but assumes no erroneous inclusions in either list. Any erroneous inclusions in the lists will inflate the estimate of the total population because the N_{+1}, or N_{1+} cells will be too large. For example, one percent over-coverage in the IDI-ERP would lead directly to a one percent increase in the total population estimate.

Census processing removed duplicate census forms as far as possible, although up to 8,500 pairs of possible duplicates may remain in the census dataset (Stats NZ, 2019a). The census may have been completed by New Zealand residents who were overseas on census night, for example by filling out the online form, which would introduce over-coverage with respect to the census target population. Around 20,000 census responses may have been filled out by New Zealand residents overseas on census night.

There is a potential for duplicate records in the IDI spine due to missed matches between contributing datasets. The derivation of the IDI-ERP dataset for use in the 2018 Census includes a process to remove likely duplicate spine records, though some may remain. We expect there to be some level of over-coverage in the IDI-ERP based on comparisons of this admin population with official figures. Some age/sex groups show a strong overcount compared with official figures, and there is some clear under-coverage (such as census responses linked to the IDI, but not included in

the selection of records for the IDI-ERP) being compensated for in consistent net results. However, we do not have an estimate of the size of the over-coverage.

The main source of coverage uncertainty for this census application is likely to be the determination of migrant status from border movements data. The September 2018 refresh of the IDI includes information on migration movements that took place until 31 July 2018. This five-month interval since census day in March 2018 should capture the majority of people with activity in government sources, but who do not meet the definition of a usual resident on census day. Deaths registrations up until 27 July 2018 are available, and should include most deaths that have occurred before the census reference date. However, these administrative time lags may mean that some of those included in the IDI-ERP are not members of the New Zealand resident population on the census date.

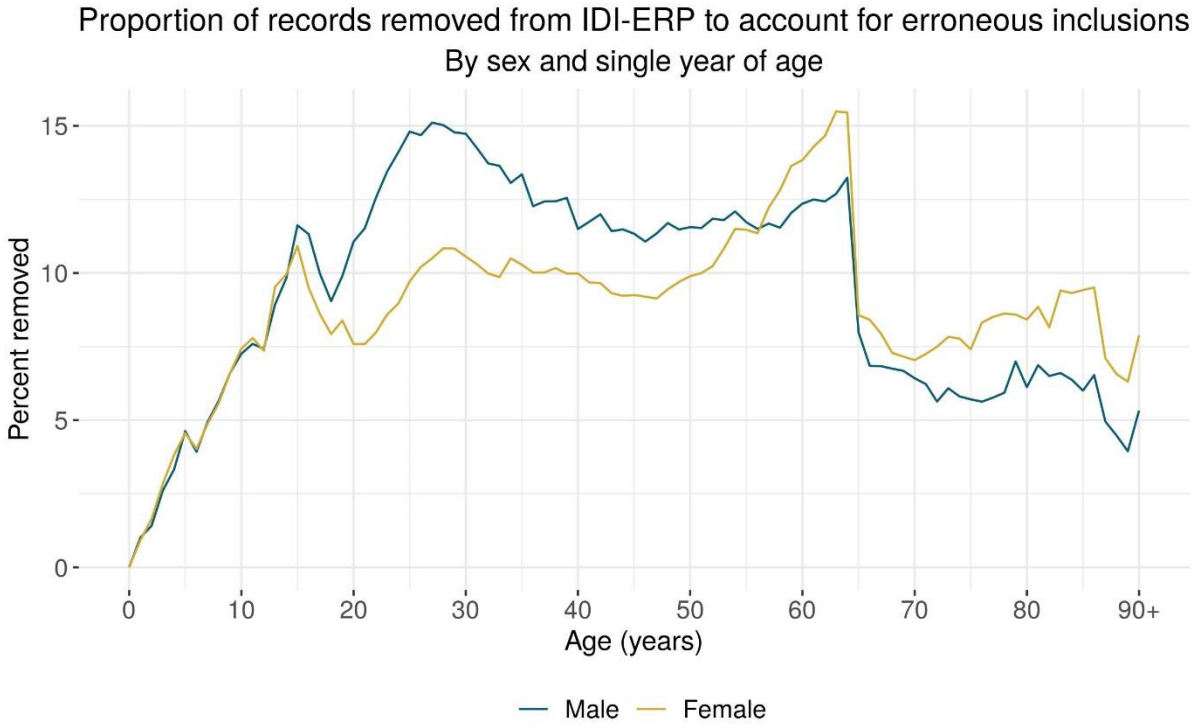
Other reasons for wrongly including a person who is not a New Zealand resident in the IDI-ERP are unclear, but may be due to factors such as false positive linkage errors for the activity datasets in the IDI, and missing deaths linkages.

We achieved the assumption of no erroneous inclusions by removing records from the IDI-ERP through applying more rigorous selection rules for inclusion. The goal was to effectively remove as much over-coverage as possible, while also minimising the amount of under-coverage being introduced. We called this subset 'IDI-ERP_Sure', to indicate that it includes only people who are 'sure' to belong to the New Zealand resident population.

The IDI-ERP includes people who have engaged with government in at least one data source. The selection criteria for inclusion in the IDI-ERP_Sure is more strict and requires that people must have activity in at least two data sources – tax or health, plus one other. An exception is made for babies under one year old, who have no additional requirements for inclusion as they may only appear in the births data. Tax and health datasets have high coverage of the population, and requiring an additional activity provides stronger evidence that they are in fact residents. This approach targets all age-sex groups, combines factors found to predict over-coverage in other unpublished Stats NZ research, and is easy to apply and explain.

About 10 percent (453,579) of the IDI-ERP file was removed when creating the IDI-ERP_Sure subset. Figure 3 shows the records removed from the IDI-ERP as a percent of the total in each single year of age. The majority of records removed were young adults (predominantly male) and those approaching 65 years (retirement age). The young males are likely candidates for over-coverage, however the rules may be unnecessarily removing those who take early retirement. We have not been able to validate this approach to removing erroneous inclusions in the admin population, and it seems likely that at least some erroneous records remain.

Figure 3



We also investigated 'trimmed DSE' approaches based on work by Zhang and Dunne (2017). Trimmed DSE attempts to remove erroneous records from a list to meet the assumption that there is no over-coverage in the sources used in the DSE. Records are systematically removed from the file according to some criteria. A score is devised as a diagnostic which can be used to indicate when records being trimmed are no better than random selection. Several trimmed DSE examples were attempted with New Zealand data, however these became very complicated with different approaches needed for different age groups. The trimmed DSE also removed many more records compared with the activity rules. Over-trimming risks introducing bias if the proportion of genuine residents removed does not remain constant between the linked n_{11} cell and those only in the admin population, the n_{01} cell. Further work would be needed to apply the trimmed DSE approach in this context.

After removing over-coverage based on the refined activity rules, the admin population entering the DSE calculation became the IDI-ERP_Sure dataset.

Perfect linking

DSE assumes perfect linkage and is sensitive to departures from this assumption. With our large-scale automated linkage process we were not able to enforce perfect linking. The emphasis on minimising false positive matches typically results in a higher rate of missed matches. We assumed that false positive matches had a negligible impact on the DSE, and adjusted the calculation for incorrectly missed matches.

Adjustment for missed matches

To account for linkage error, we made an adjustment to the internal DSE cell counts to correct for missed matches when a census record failed to link to its true corresponding record in the admin population.

Ding and Fienberg (1994) describe a method for adjusting for linkage error in the context of population estimation. Following Ding and Fienberg's simplest model, we assumed that the true matched pairs will be linked with probability a. That is,

$$a = \Pr(linked|match)$$

Ding and Fienberg also describe a model that allows for an adjustment to the DSE due to false positive matches, ie $\Pr(linked|nonmatch)$. We did not apply the more complete error model, and assumed that the probability of true non-matches being linked was negligible. Under that assumption, linkage error occurs with probability $1 - a$.

The number of linked records in the n_{11} cell was adjusted based on probabilities of being linked given the true match status.

Once we have a value for a , the new estimate for the linked cell, $\hat{n}_{11}$ is

$$\hat{n}_{11} = n_{11}/a$$

Or equivalently, expressed in additive form

$$\hat{n}_{11} = n_{11}(1 + \eta)$$

where $\eta = \frac{\Pr(unlinked|match)}{\Pr(linked|match)}$ i.e. η is the ratio of linkage probabilities conditional on the true match status.

The number added to the linked cell, n_{11}, must then be removed from the two unlinked cells n_{10} and n_{01} to preserve the marginal totals. The resulting cross tabulation for the DSE adjusted for linkage error is shown in table 4.

Table 4

Contingency table for dual system estimation cell counts adjusted for missed matches				
		Admin population		
		1	0	Total
Census responses	1	$n_{11}(1+\eta)$	$N_{1+} - n_{11}(1+\eta)$	N_{1+}
	0	$N_{+1} - n_{11}(1+\eta)$		
	Total	N_{+1}		

Linkage errors may vary by demographic characteristics. For example, string comparison of names tends to perform better for European names, while Polynesian and Asian names may be more difficult to match correctly. The adjustment is made for each combination of the estimation strata.

The interaction between both types of linkage error is complex. The presence of false positive matches means that the original linked cell, n_{11}, is too large, and will tend to decrease the DSE. Ding and Fienberg (1994) confirm empirically an earlier conclusion (Ding, 1990) that due to high capture probabilities in census applications, the matching bias is dominated by the false negative match rate when the false negative match rate and false positive match rate are about the same magnitude. As

that may perhaps be the case here, it supports a view that the adjustment for missed linkages captured most of the bias due to linkage error.

Estimating the missed matches

Ding and Fienberg (1994), and Di Consiglio and Tuoto (2015) describe methods for measuring linkage error in the context of population estimation, but both rely on a clerically checked sample or other source of truth. We are matching approximately 4 million census records to an IDI spine consisting of nearly 10 million records and did not have the time or resources for checking even a sample to estimate missed matches. Chipperfield and Chambers (2015) take quite a different modelled approach using bootstrap simulations, however we were not able to implement this within the constraints of our record linkage software and IT environment.

We estimated false negative matches based on an approach developed by Choi, 2019. Rather than relying on a sample or gold standard, missed matches were estimated from a subset of the census forms that met the criteria for inclusion in the IDI spine.

The IDI spine is made up of people who were born in New Zealand, have entered New Zealand as a migrant on a visa, or are registered to pay tax. We estimated missed matches on the basis of responses to census questions that closely coincide with these criteria for inclusion in the spine. A subset of the census Individual forms, M^*, was created through applying strong requirements for membership in two or more of the datasets that make up the IDI spine. To be exact, M^* includes people who filled out an Individual form and indicated that they:

- were born in New Zealand and
    - have reported taxable income, or
    - are aged 14 years or less
- were not born in New Zealand and
    - arrived in New Zealand after 1997 and have reported taxable income, or
    - arrived in New Zealand after 1997 and are aged 14 years or less (excluding those children who were born in countries that do not require a visa, namely Australia, Cook Islands, Niue and Tokelau).

Those in the M^* subset should all have been linked to the IDI spine. Those who were not linked are assumed to be missed true matches. The subset M^* consists of 3,311,928 people or 83 percent of the 3,971,892 census Individual forms. From this we obtained an overall estimate for $1 - a$ of 1.21 percent of the M^*subset who should have been linked to the IDI spine, but were not. This translates to an overall linkage probability ratio, η, of 0.0122. Missed links and η values were estimated for each of the strata used in the DSE: single year of age, sex, TALB, and mutually exclusive level 1 ethnic groups.

The M^* subset provides the estimates of a and η used to adjust the DSE calculation, but required some assumptions about the applicability of using these linkage error rates for the datasets entering in the DSE.

First, we assumed the same rate of missed matches between the census forms and the IDI spine applies to other census responses that were not in the M^*subset. This appears reasonable for the Individual forms used in the DSE, as we assume that the quality of the linking variables is independent of the criteria for inclusion in M^*, for example whether taxable income was received or not, and whether a migrant entered New Zealand before or after 1997.

We also assumed that selection in the IDI-ERP, and the more restricted conditions required for the IDI-ERP_Sure dataset, was independent of whether a match between the census and the IDI spine

was incorrectly missed. This appears to be a reasonable assumption, since selection of the IDI-ERP was independent of participation in the census, and the propensity of obtaining a match to the census was partially dictated by the quality of the census response information used in the linking.

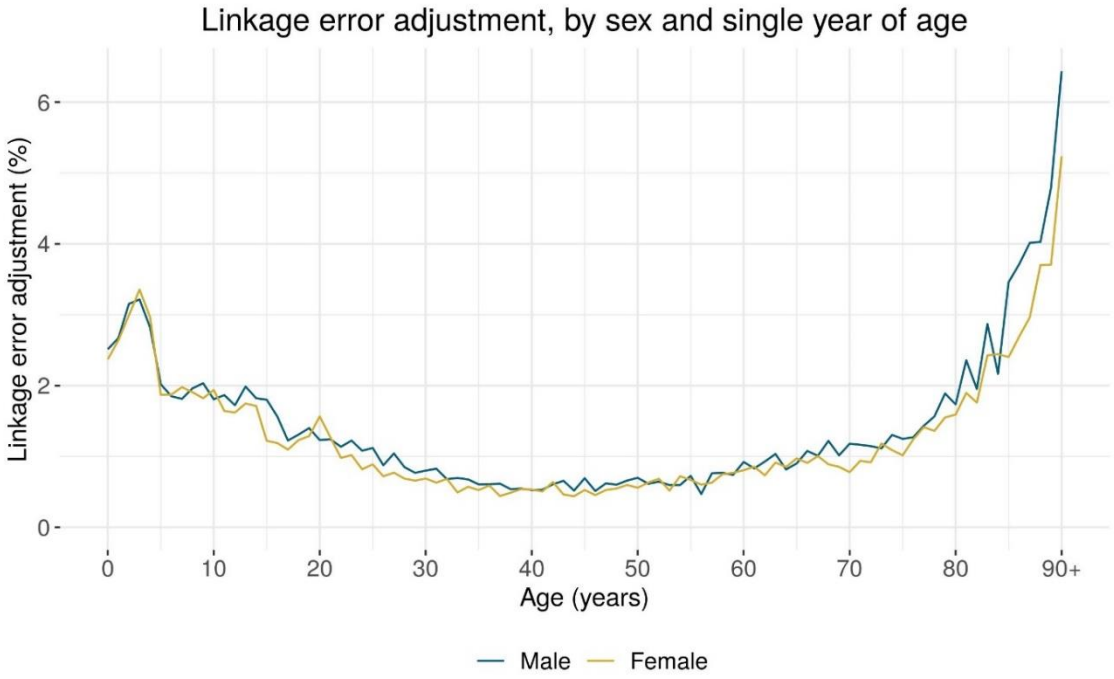
For very small strata there may be no records in the M^* subset of census that can be linked to the IDI spine and η will be undefined. In that scenario we used the overall estimate of η calculated using all the links between M^* and the IDI spine.

Because the size of the census and admin population lists are fixed, the marginal totals need to be preserved. To ensure that the adjusted $\hat{n}_{11}$ cell count is not larger than the number of records available for linking, in practice the linkage-error adjustment applied is:

$$\hat{n}_{11} = \min(n_{11}(1+\eta), N_{1+}, N_{+1})$$

The linkage error adjustment added 41,900 to the linked n_{11} cell and removed the same number from the off-diagonal cells. Rates for missed matches, $1-a$, were higher for children, at around two percent, and for older adults aged above 80 years, where rates ranged from two percent to six percent (figure 4). These patterns closely reflected the inverse of linkage rates (Stats NZ, 2019a).

Figure 4



Results

The DSE contingency table

Table 5 shows the national totals in each of the cells before the adjustment for missed matches. Table 6 provides the adjusted internal cells once the missed links are added into the linked $\hat{n}_{11}$ cell and subtracted from the off-diagonals. The estimated total population from the DSE is 4,768,600 ± 800, though we note that this is not a direct calculation from the numbers shown, but was calculated for each of the DSE group breakdowns and aggregated to the national total.

There are 88,900 people estimated by the DSE to not be included in either the census Individual forms or the admin IDI-ERP_Sure population.

Table 5

Dual system estimation contingency table before adjustment for missed matches				
		Admin population: IDI-ERP_Sure		
		1	0	Total
Census Individual forms	1	3,507,400	449,500	3,956,900
	0	764,600		
	Total	4,272,000		

Table 6

Dual system estimation contingency table after adjustment for missed matches				
		Admin population: IDI-ERP_Sure		
		1	0	Total
Census Individual forms	1	3,549,300	407,600	3,956,900
	0	722,700	88,900	
	Total	4,272,000		4,768,600

As some bias may remain, we cannot determine exactly how accurate the DSE population is. Comparisons with two other population estimates provide some means of quality assurance: the original IDI-ERP admin population, and a revised 2013-based ERP (both of which were for census day, and were adjusted to remove residents temporarily overseas).

The original IDI-ERP population and the DSE overlapped by 722,700 people from the DSE n_{01} cell 'not found in the census, in the admin population'. However, the two estimates can be considered as relatively independent since the DSE is dominated by the census responses. The IDI-ERP will be affected by any admin over-coverage and under-coverage, which is removed from the DSE.

The revised 2013-based ERP for national age, sex distributions was first published by Stats NZ in August 2019, and applies an improved measure of international migration from 2013 through to 2018. While external migration estimates should be more accurate, the series will reflect any errors in the official 2013 base ERP.

Table 7 shows DSE totals and 95 percent confidence intervals for the national population, and estimates for the three ethnic groups. The DSE uncertainty intervals are very small (less than 0.1 percent) and reflect the large size of the populations contributing to the DSE. Comparisons are provided for the IDI-ERP and revised 2013-base ERP. The total population estimates are all within 42,000.

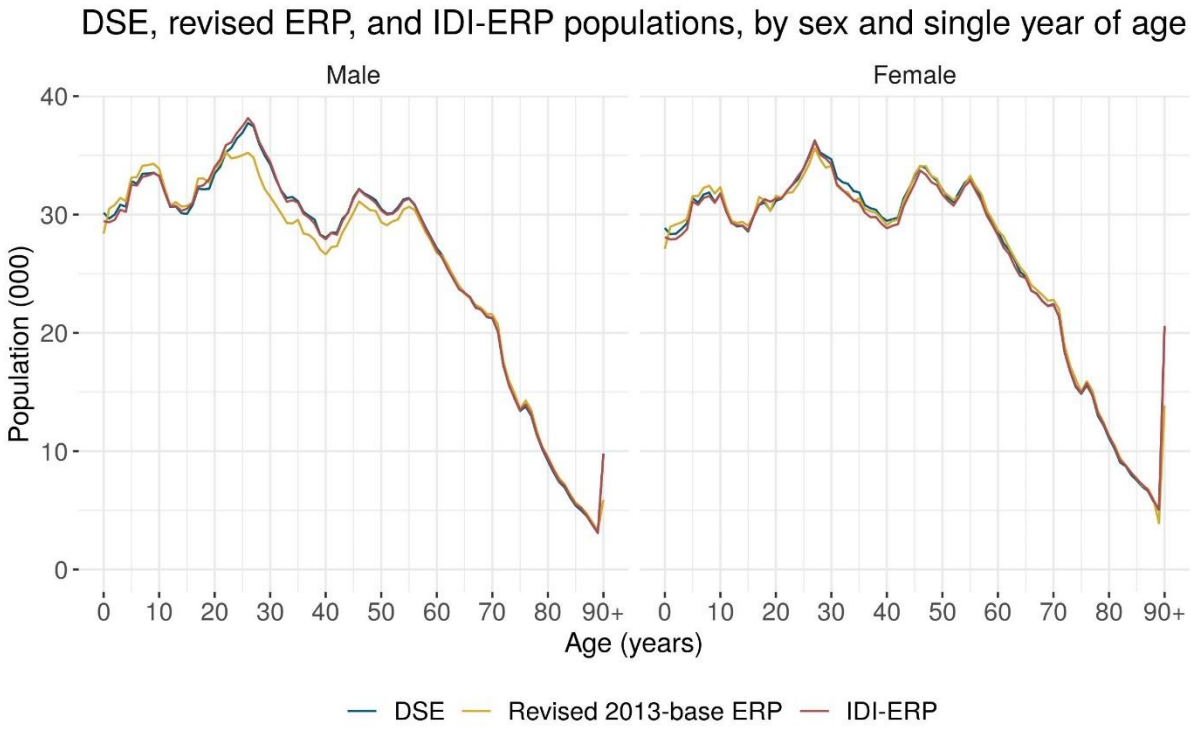
Table 7

Estimated totals and confidence intervals for national population and for ethnic groups, as at 6 March 2018			
Population group	Dual system estimation population	2013-base revised ERP	IDI-ERP
Māori	807,900 ± 370	-	797,580
Pacific	397,200 ± 360	-	399,381
Asian	727,400 ± 480	-	674,385
Total	4,768,600 ± 800	4,727,800	4,751,598

DSE population distributions

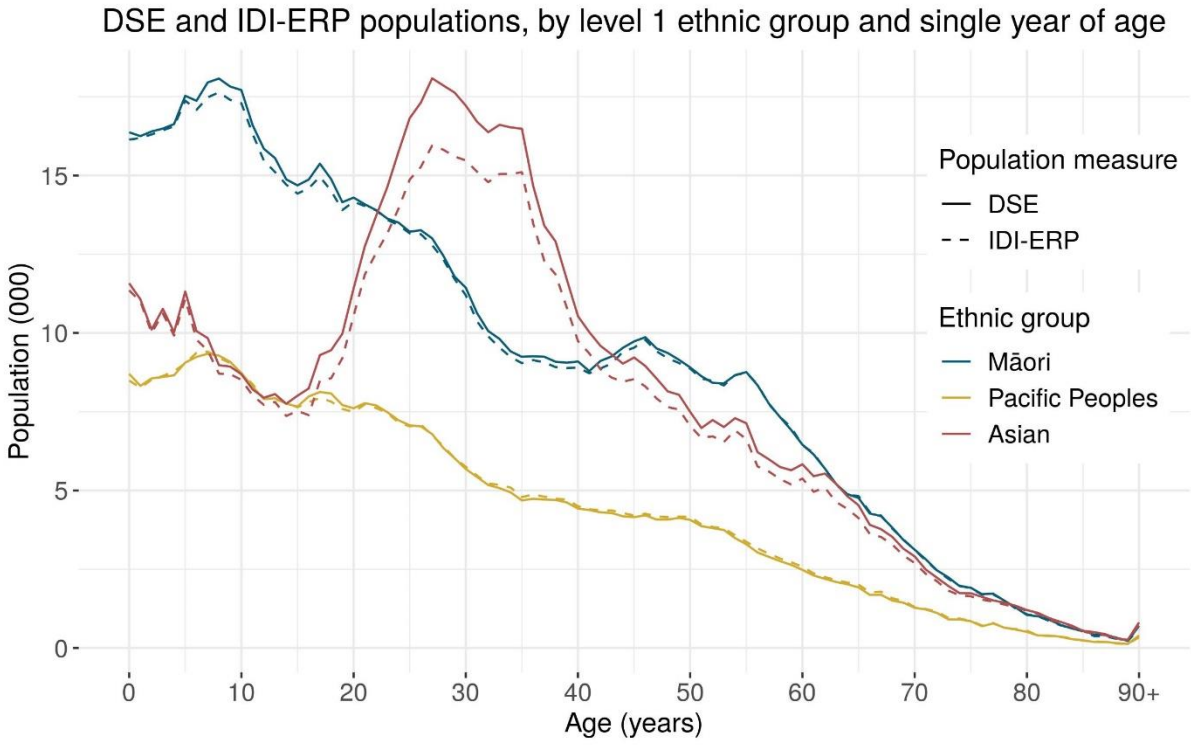
Figure 5 shows the national DSE estimates by single year of age and sex compared with the original IDI-ERP admin population and the revised 2013-based ERP. For females we see a highly consistent distribution across all three estimates. For males the DSE and IDI-ERP are largely consistent, while the revised 2013-based ERP is lower from ages in the early 20s through to about 50 years of age. These results are encouraging for both the DSE, and the IDI-ERP estimates.

Figure 5



Ethnic population distributions by single year of age for the DSE are shown in figure 6. These follow expected patterns, showing the younger age profiles for Māori and Pacific. The higher number of people between 20 and 40 years in the Asian age distribution reflects recent migration patterns.

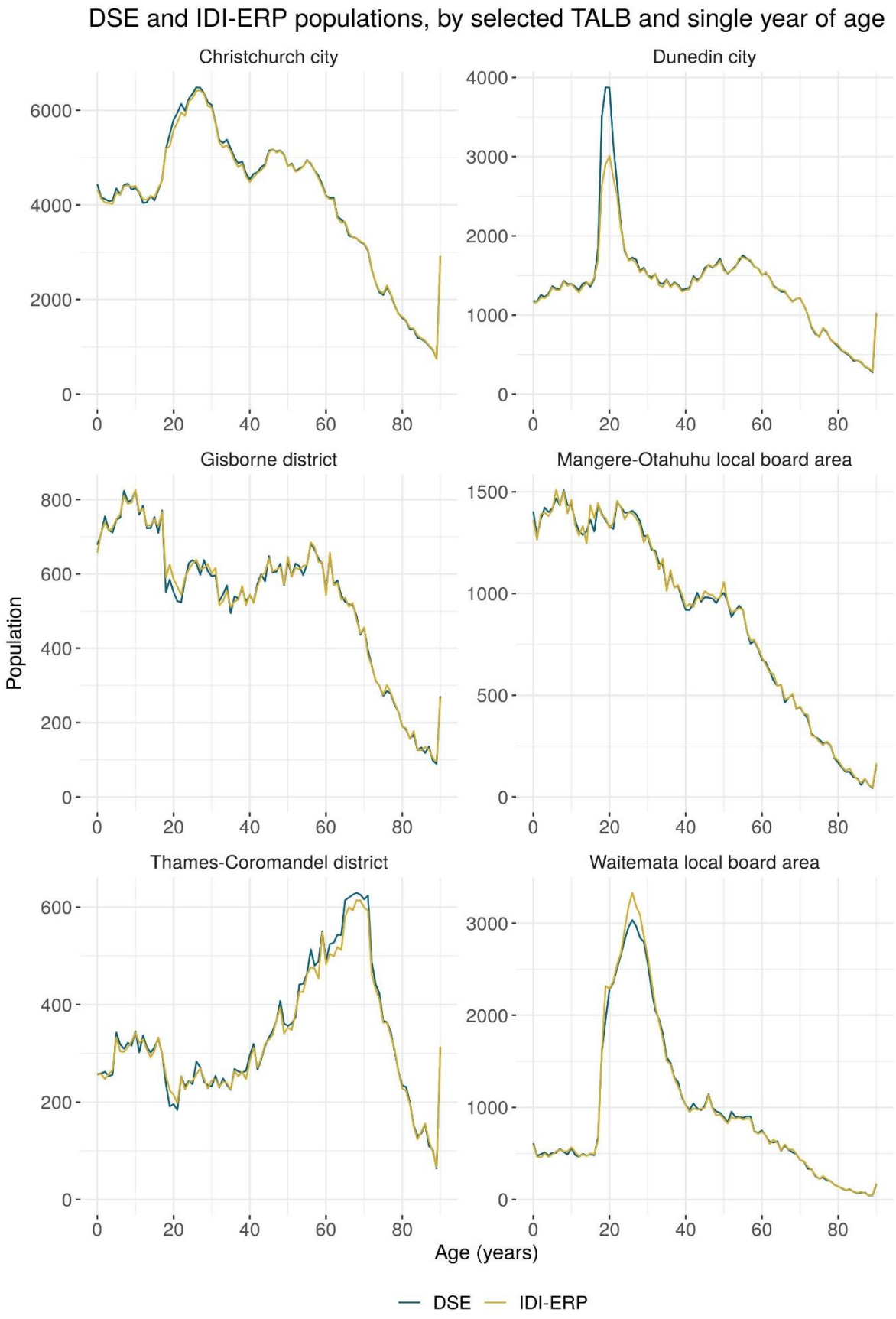
Figure 6



In contrast to the 2018 IDI-ERP admin population, the DSE uses census respondent ethnicity values for the 3.5 million people in the linked cell, and relies much less on administrative ethnicity values. The DSE also does not have the over-coverage or under-coverage we expect to be reflected in the IDI-ERP ethnic group distributions. The comparison in figure 6 shows that the DSE and IDI-ERP age distributions are very close for Māori and Pacific ethnic groups, again an encouraging result for both estimates. For the Asian ethnic group, the IDI-ERP is lower than the DSE estimate through adult working ages.

Age distributions for selected sub-national geographies are shown in figure 7. These reflect TALBs with smaller and larger populations, and with different age profiles. We compare the DSE with the IDI-ERP. Overall, age distribution patterns are very similar in both sources. We expect the IDI-ERP to be more affected by inaccuracies in the administrative location, particularly for young adults. We do see clear differences for young adults around age 20 that suggest the DSE has improved over the IDI-ERP admin estimates. For example, in Dunedin City, a university town, the DSE estimates around age 20 years are higher than the IDI-ERP, while in Gisborne District, a rural area, DSE shows fewer at this age.

Figure 7


DSE and IDI-ERP populations, by selected TALB and single year of age
Christchurch city
Dunedin city
Gisborne district
Mangere-Otahuhu local board area
Thames-Coromandel district
Waitemata local board area
Population
Age (years)
DSE
IDI-ERP

Discussion and conclusion

Our aim was to produce high quality estimates of the New Zealand population at the time of the New Zealand census in March 2018. Population estimates for key demographic and sub-national geographic breakdowns were needed to guide development of new methods for adding administrative records to the 2018 Census dataset, as a mitigation for a lower than expected response to the census field collection. This work was not planned as part of the build-up to the 2018 Census, and the project was under considerable time pressure to minimise delays in releasing census outputs.

The available official population estimates, based on 2013 Census data, were impacted by uncertainty in the measurement of population change due to external migration and internal migration accumulated over the five-year period between censuses. The lack of breakdowns by all the main ethnic groups was another limitation. We had available two data sources for the March 2018 date that did not require estimates of change due to migration, and that do include ethnicity: the 2018 Census responses, and a New Zealand resident population constructed from administrative data. Both are large, but incomplete, lists of the New Zealand population as at March 2018, and dual system estimation provides a methodology for estimating the total population, once the data are linked.

While DSE is a standard and widely used methodology, it rests on several core assumptions. We faced two main challenges in this relatively novel application: removing over-coverage from the administrative list, and measuring linkage error and adjusting the DSE calculation to account for it. We developed approaches to managing both issues that attempted to minimise bias in the DSE. These approaches are the best that could be achieved in the constrained time frames, and we have not been able to verify the success of removing over-coverage, or the estimates of linkage error.

Variance of the DSE estimates is very small due to the large size of the contributing datasets. The net effect of any remaining bias is unclear. Failure to remove over-coverage from the administrative data, and any over-coverage in the census responses, will inflate the DSE populations, while the presence of false positive matches has the opposite effect. Applying the DSE to very small strata may lead to a downward bias for small sub-groups. While we expect that some bias remains, we do not know what direction it might be. Official population estimates based on the 2018 Census dataset and adjusted for net under-coverage will be released in early 2020 and will provide a comparison for the DSE benchmark populations described here.

The results show very plausible patterns of population structure and have provided March 2018 benchmark national population estimates by age and sex, and for the three ethnic groups and larger sub-national geographies that could be estimated. They have proved their value as we have developed our approach to using administrative records to supplement the census dataset in the face of unanticipated low response rates.

The DSE estimates have provided a sound basis for testing different options for non-response mitigation, and have also given an indication of the patterns of under-coverage remaining in the final 2018 Census dataset.

References

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

Brown, JJ, Sexton, C, Abbott, O, & Smith, PA (2018). 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, vol. Pre-press, no. Pre-press, 1–19. Retrieved from https://content.iospress.com.

Chapman, DG (1951). Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Publications in Statistics 1*, 131–160.

Choi, H (2019). Adjusting for linkage errors to analyse coverage of the administrative population. *Statistical Journal of the IAOS*, *35(2)*, 253-259.

Chipperfield, J & Chambers, R (2015). Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. *Journal of Official Statistics*, *31(3)*, 397–414.

Di Consiglio, L & Tuoto, T (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, *31(3)*, 415–429.

Ding, Y (1990). *Capture-recapture Census with uncertain matching*. PhD dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania.

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

Fellegi IP & Sunter AB (1969). A theory of record linkage. *Journal of American Statistical Association*, *4*, 1183–1210.

Gibb, S, Bycroft, C, & Matheson-Dunning, N (2016). Identifying the New Zealand resident population in the Integrated Data Infrastructure (IDI). Retrieved from http://archive.stats.govt.nz.

Lincoln, FC (1930). Calculating waterfowl abundance on the basis of banding returns. *United States Department of Agriculture, Circular 118*, 1–4.

Mule, T (2012). 2010 Census coverage measurement estimation report: summary of estimates of coverage for persons in the United States. *DSSD 2010 Census Coverage Measurement Memorandum Series* #2010-G-01. http://www.census.gov/coverage_measurement/pdfs/g01.pdf.

Petersen, CGJ (1896). The yearly immigration of young plaice Into the Limfjord from the German Sea. *Report of the Danish Biological Station* (1895) *6*, 5–84.

Sadinle, M (2008). *On the performance of dual system estimators of population size: A simulation study.* Conflict Analysis Resource Centre – CERAC and Departamento de Estadistica Universidad Nacional de Colombia.

Stats NZ (2014). Coverage in the 2013 Census based on the New Zealand 2013 Post-enumeration Survey. Available from http://archive.stats.govt.nz.

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

Stats NZ (2018). Experimental ethnic population estimates from linked administrative data. Retrieved from www.stats.govt.nz.

Stats NZ (2019a). Linking 2018 Census responses to the Integrated Data Infrastructure. Retrieved from www.stats.govt.nz.

Stats NZ (2019b). Overview of statistical methods for adding admin records to the 2018 Census dataset. Retrieved from www.stats.govt.nz.

Stats NZ (nd). Statistical standard, Ethnicity V1.0.0. Retrieved from http://aria.stats.govt.nz.

Wittes, J (1972). 331. Note: On the bias and estimated v0ariance of Chapman's two-sample capture-recapture population estimate. *Biometrics, 28*(2), 592–597. https://doi.org/10.2307/2556173

Zhang, L & Dunne, J (2017). Trimmed Dual system estimation. In: *Capture Recapture Methods for the Social and Medical Sciences*, eds. D Bohning, PGM Van der Heijden, Bunge, J, pp239–259, CRC Press, Boca Raton, Florida.

Appendix: Ethnic groups used in the DSE

The concept of ethnicity in New Zealand allows for a person to belong to more than one ethnic group. The six ethnic groups in level 1 of the classification are each stored as a single variable with a 1 or 0 indicating a person's membership of that group. The following eight combinations of mutually exclusive ethnic groups were formed to derive estimates of the total Māori, Pacific, and Asian populations using DSE:

1. Māori only (where the Māori indicator is equal to 1, Asian and Pacific indicators are equal to 0 and all other ethnic indicators can be 1 or 0)
2. Asian only (where the Asian indicator is equal to 1, Māori and Pacific indicators are equal to 0 and all other ethnic indicators can be 1 or 0)
3. Pacific only (where the Pacific indicator is equal to 1, Māori and Asian indicators are equal to 0 and all other ethnic indicators can be 1 or 0)
4. Māori and Asian (where the Māori and Asian indicators are equal to 1, the Pacific indicator is equal to 0 and all other indicators can be 1 or 0)
5. Māori and Pacific (where the Māori and Pacific indicators are equal to 1, the Asian indicator is equal to 0 and all other indicators can be 1 or 0)
6. Asian and Pacific (where the Asian and Pacific indicators are equal to 1, the Maori indicator is equal to 0 and all other indicators can be 1 or 0)
7. Māori, Asian and Pacific (where the Māori, Asian and Pacific indicators are equal to 1 and all other indicators can be 1 or 0)
8. Other (where the Māori, Asian and Pacific indicators are equal to 0 and at least one other indicator is 1)

The indicators for European, MELAA and Other ethnic groups are not considered if either the Māori, Asian or Pacific indicator is 1. For example, if an individual has the Māori and European ethnic indicators equal to 1, their ethnic group would be Māori. If an individual only has the European ethnic indicator equal to 1, their ethnic group would be Other. Because of this, estimates for the Other ethnic group category will not be accurate under this approach, neither will estimates for European or MELAA be available; however creating the Other ethnic group is required to retain all other records. This is a limitation to this approach.

To obtain estimates for Māori, Pacific and Asian ethnic groups, the DSE estimates can be summed over the groups that contain the ethnic group of interest. For example, to obtain estimates of Māori, estimates of Māori only; Māori and Asian; Māori and Pacific; and Māori, Asian and Pacific will be aggregated.