Stats NZ
Tatauranga Aotearoa

Overview of statistical methods for adding admin records to the 2018 Census dataset

New Zealand Government

Crown copyright ©
See Copyright and terms of use for our copyright, attribution, and liability statements.

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

ISBN 978-1-98-858349-5 (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

Purpose and summary 5
Purpose 5
Summary of key points 5
Introduction 6
Data sources 8
Statistical location register and Census dwelling frame 8
Census responses 9
A New Zealand admin resident population 10
Differences between census responses and admin sources 13
Linking census responses to the IDI 15
Methods 17
Dual system estimation population benchmark 17
Early methods development 17
Final methods 18
Summary of admin records 24
Results 25
Contribution of admin enumerations to the final 2018 Census dataset 25
Comparison of final 2018 Census dataset and DSE benchmark populations 29
Sensitivity analysis for key uses 35
Discussion and conclusion 37
References 39
Appendix 1: Admin data sources and time lags 41
Appendix 2: Admin enumerations in responding dwellings 43
Appendix 3: Non-private dwellings – prisons and defence 45
Method 45
Results 47
Usual residence location imputation 48
Appendix 4: Adjusting for missed matches 49
Adjusting for missed matches 50
Results of missed linkage adjustment 51

List of tables and figures

List of tables

1 Results from linking 2018 Census responses to the IDI spine....15
2 Flow for determining admin records to include as admin enumerations....24
3 2018 Census usual resident population count, by unit record source....25
4 Proportion of census responses and admin enumerations, by level 1 ethnic group....27
5 2018 Census and DSE benchmark counts, by age group and selected ethnic group....29
A1 Admin data lags for activity and address information in the September 2018 IDI refresh....42
A2 Household size before and after admin enumeration of people to responding dwellings....44
A3 Census responses and admin enumerations in prisons or penal institutions and defence establishments....47
A4.1 Cross-tabulation of linkage between census forms and IDI-ERT_Sure....50
A4.2 Census responses linked to the IDI-ERP_Sure admin population....51

List of figures

1 Occupancy definitions in census and admin collections....13
2 Progression of records included in the census dataset....19
3 Admin records eligible for inclusion in the census dataset....20
4 Age distribution of census usually resident population count, by unit record source....26
5 Percentage identifying with level 1 ethnic groups, by unit record source....27
6 Percentage of population from each unit record source, by TALB area....28
7 2018 Census and DSE benchmark counts, by sex and single year of age....30
8 2018 Census and DSE benchmark counts, for Māori ethnic group....31
9 2018 Census and DSE benchmark counts, for Pacific ethnic group....31
10 2018 Census and DSE benchmark counts, for Asian ethnic group....32
11 Relative difference of 2013 and 2018 Census from benchmark, for females....33
12 Relative difference of 2013 and 2018 Census from benchmark, for males....33
13 Relative difference of 2013 and 2018 Census from benchmark, by TALB area....34
A4.1 Missed matches between the 2018 Census and admin records....50
A4.2 Distribution of eta values....52
A4.3 Percentage of IDI-ERP_Sure estimated to be missed links, by single year of age and sex....53
A4.4 Linkage adjustment, by single year of age....54

Purpose and summary

Purpose

Overview of statistical methods for adding administrative records to the 2018 Census dataset describes the statistical methods used to determine high-quality administrative (admin) records to include in the 2018 Census dataset.

Summary of key points

For the first time, the New Zealand 2018 Census dataset includes administrative (admin) records to include people who were missed by the census field collection, replacing the use of 'substitute' imputed records in previous censuses. The 2018 Census dataset is made up of census forms received through field collection, supplemented by high-quality administrative records for people from whom we have not received a response. These **admin enumerations** are drawn from a New Zealand resident population that has been derived from administrative data. Members of this New Zealand **admin resident population** are only added to the census dataset if they were in New Zealand on census night, and have not already responded to the census.

Administrative records are included as usual residents in a private dwelling when the administrative household membership information is reliable. In two cases (prisons and defence establishments) administrative records contribute to non-private dwellings. Otherwise, administrative records are not placed in dwellings, but have been counted within the smallest geographic area (a meshblock, up to about 100 people, which serves as a building block for all standard geographies) when evidence suggested this would improve the quality of census small-area data. Some potential administrative records are not included due to limitations in identifying individuals' location. Statistical methods underpin each of these steps and are designed to account for limitations in the quality of administrative sources.

Our understanding of the strengths and limitations of the admin resident population builds on previous work carried out through Stats NZ's Census Transformation programme. The main limitations of the administrative data are some undercoverage and overcoverage relative to the true population, and the reliability of administrative addresses. To overcome these limitations, we developed statistical models to predict how reliable the administrative data are for representing an entire household, and a related model predicts the likelihood that the administrative meshblock reflects a person's true usual residence meshblock. Before considering the administrative records for inclusion in meshblocks, we undertake two main adjustments to ensure that we are not introducing people to the census file who should not be included. We first adjust for potential overcoverage in the admin population. The second adjustment is to account for duplication caused by missed linkages between the received census forms and the administrative data.

The final 2018 Census dataset consists of 89 percent census responses and 11 percent admin enumerations. Admin enumerations make up almost a quarter of the Māori and Pacific populations in the final 2018 Census dataset, compared with 10 percent for the European ethnic group, an indication of lower participation for people of Māori and Pacific ethnicity in the 2018 Census.

The careful use of administrative data has enabled Stats NZ to add real data about real people to the 2018 Census dataset where we are sure they should have been counted but we have not received a census form. The inclusion of admin enumerations also removes the need to add missing census records through statistical imputation processes.

Introduction

Stats NZ is committed to exploring the use of administrative (admin) data for the census in its Census Transformation Strategy (Stats NZ, 2012). The strategy consists of two parallel strands. The first is a short- to medium-term focus on modernising the current full enumeration census, with the first implementation in the 2018 Census of Population and Dwellings. For the second strand, Stats NZ's Census Transformation programme is exploring the feasibility of a census based largely on administrative data, and supported by sample surveys (see The Census Transformation programme). An essential feature of the long-term use of administrative data for census information has been the development of an admin New Zealand resident population derived from linked administrative data.

The two strands are interdependent and the 2018 Census included a goal to increase the use of administrative data. The main use of admin data planned for 2018 was to construct an address frame from admin sources, which has developed some of the organisational infrastructure required for long-term change. The 2018 Census also planned to use alternative sources to improve data quality in the context of missing data for census questions, based on the findings of the longer-term admin data research. The value of this joint strategy has become clear in the light of lower than anticipated response rates in 2018, which have considerably extended the role that admin data now plays in the 2018 Census.

While some aspects of the 2018 Census modernisation have been successful, we faced major challenges when implementing the new collection model, and the overall level of response was lower than expected. 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 the census were disproportionately affected in 2018.

Previous censuses have included an adjustment for non-response, through inclusion of 'substitute' records, a form of unit imputation (Stats NZ, 2014c), and a similar approach was planned for 2018. Once these collection issues became clear in mid-2018, Stats NZ embarked on development of new methods for including people from admin records in the 2018 Census dataset based on the derived admin New Zealand resident population. The new methods have built on the research undertaken by the longer-term census transformation investigations, and fast-forwarded their development for use in the context of a full enumeration census.

Our initial approaches to adjusting the census dataset for non-response were motivated by concerns that the size of the planned unit imputation placed too much risk on the validity of the underlying assumption of households 'missing at random'. We developed models to use households constructed from admin records in place of imputed households, where this could be done reliably. It became apparent that the traditional census approach of only placing people into dwellings where they had been missed would not be sufficient. We then developed models for placing people into meshblocks, rather than into a dwelling. This made full use of the admin resident population and removed the need for any unit imputation.

The inclusion of admin records in a traditional full field enumeration census is unusual internationally. In their 2011 Census, Northern Ireland added records obtained from administrative health data to count people who had been missed by the field collection (Green et al, 2014). Statistics Canada used administrative records in their 2016 Census when major wildfires led to the evacuation of the Fort McMurray area and the suspension of census collection activity very close to census day. Of most relevance here is the United States Census Bureau work investigating the use of high-quality administrative records in the United States 2020 Census as a means of improving efficiency in their field collection process (Keller et al, 2018).

The Statistics Act 1975 and Privacy Act 1993 together provide the legal basis for the use of admin data in the 2018 Census. The primary reason for this is that the Statistics Act requires that the same level of confidentiality is applied to statistics or research as is required by the Privacy Act. Many of the principles within the Privacy Act, such as those related to the purpose of collection, contain exemptions for when information is later used for statistical or research purposes.

There is a natural tension between privacy and providing information that is collected from individuals. This tension is carefully considered whenever census content is developed. The use of admin data respects that same balance between census content and the public good that comes from living in an informed society. Using data with inherent biases can raise significant ethical issues because it can cause disadvantage and harm to particular groups. The use of admin data to improve the quality of the census data is an ethical approach from this perspective. A privacy impact assessment of combining admin data and census form data (Stats NZ, 2019a) concluded that this use of admin data is lawful, safe, and beneficial to New Zealanders.

Here we provide an overview of the methods used to determine which admin records should be added to the census dataset to count people who did not respond to the census field collection. The focus is on the key census demographic counts by age, sex, and ethnicity, and for geographic areas. A methods paper about data sources, editing, and imputation (Stats NZ, in press) describes the use of data from the previous census in 2013 and from linked admin sources for the characteristics of people and dwellings when there is no census response.

We first describe the main data sources for the census and the admin population. The methods section outlines each of the main steps used to add admin records. Appendices give more detail for some processes, and references are provided for more in-depth explanation of statistical methodologies. Results show the contribution of each of the main steps, and compare the final 2018 Census demographic distributions with a benchmark population developed for this purpose. The paper concludes with a discussion.

Data sources

The New Zealand census is a census of both population and dwellings. The census develops a list of all dwellings in New Zealand – the **census dwelling frame** – and then counts people within those dwellings. In the 2018 Census, administrative data have been used as the primary source for constructing the census dwelling frame, supplemented by census field staff. Individuals have been counted in the final census dataset primarily through responses to census forms, supplemented by admin records. We now describe in more detail the census dwelling frame, and the two data sources for individuals that contribute to the final census dataset.

Statistical location register and Census dwelling frame

Stats NZ's statistical location register (SLR) is a central reference list of addresses, address identifiers, and spatial coordinates. The SLR database manages and maintains a relationship between an address and the object it refers to, for example whether it is the address of a dwelling, a business, or a park. A dwelling register forms a component of the SLR. All address information is stored in a standard format, and multiple descriptions of the same 'addressable object' can be aligned to the same address ID. The SLR includes addresses from the 2013 Census and is updated monthly with addresses supplied by Land Information New Zealand and New Zealand Post.

Address-matching software is used to match text address descriptions on other datasets, such as addresses provided by government agencies, to SLR addresses. This ensures that addresses from all sources are consistently matched to the SLR reference address ID.

The SLR provided an initial list of dwelling addresses to the census operation. A national canvassing exercise confirmed or amended these addresses for census mail-out areas (80 percent of all dwellings). For the remaining 20 percent of dwellings, census field staff updated the list of dwellings provided to them when delivering census materials to these areas. The SLR address descriptions are used to populate an 'as you type list' for respondents completing address questions online. Some further changes to the addresses that comprise the census dwelling frame were made based on information from census respondents.

The final census dwelling frame is a list of all dwellings in New Zealand (including private and non-private dwellings), and is a fundamental component of the census. The addresses were used to receipt returned census forms for dwellings, and thus indicate which dwellings have no responses. Every dwelling on the census dwelling frame has a unique identifier as well as other information such as occupancy status, meshblock, dwelling type, and geo location.

The 2018 Census dwelling counts, 1.89 million dwellings in total, are produced from the census dwelling frame. The quality rating for census dwelling counts is 'high'. The 2018 Census total private dwelling count is about 7,000 dwellings (0.4 percent) lower than previously published data for March 2018, which are based on the 2013 Census (see Dwelling and household estimates: March 2018 quarter). The 2018 Census data are likely to reflect a more accurate picture of demolitions since the 2013 Census, particularly of Canterbury dwellings after the 2010 and 2011 earthquakes.

Creating the census dwelling frame for the 2018 Census (Stats NZ, 2019b) provides further details about the census dwelling frame.

Dwellings in the census

Standard terms for measures of dwellings, households and families V1.0.0 provides formal definitions and classification of dwellings as applied in the census. A dwelling means any building or structure, or part thereof, that is used (or intended to be used) for the purpose of human habitation. It can be of a permanent or temporary nature and includes structures such as houses, motels, hotels, prisons, motor homes, huts, and tents. There can be more than one dwelling within a building, such as an apartment building, where each apartment or unit is considered to be a separate dwelling.

As well as permanent dwellings, the census includes people in temporary dwellings, such as tents or camper vans, and the homeless. Households and families information is provided for occupied private dwellings.

In the census, dwellings are classified as Occupied, Unoccupied, or Under construction on census night. Further breakdowns can be made, for example, Unoccupied dwellings may be classified as 'empty' or 'residents away'.

Census responses

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. As no adjustment is made to the census dataset to account for an undercount of overseas visitors, they are of no further interest here. Our focus is on New Zealand residents who are present in New Zealand on census night.

When people respond to the census at a dwelling where they are visiting on census night, they are also asked where they usually live. The most important uses of census data are for the census usually resident population, not where people happen to be on census night.

We note that New Zealand's official population estimates, the Estimated Resident Population (ERP), measure the entire resident population. The ERP is based on the census counts but includes those temporarily overseas on census night, and adjusts for net census undercount based on the Post-enumeration Survey. The ERP series is regularly updated for population change since the 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.

Minimum information requirements for an individual to be included in the census dataset as a response from the field collection are valid values for two out of three of the fields: name, date of birth, and meshblock. This combination of variables provides enough evidence that a unique person is being counted. People whose only information comes from the household listing are included as a census response if they meet these requirements. Individual forms that do not meet these requirements are excluded, for example if both name and date of birth are absent.

Close to 4 million census Individual forms were received, and a further 202,914 people appeared on a household listing, but no Individual form was received for them. This level of response was lower than expected, and was seen across all regions. However, it was clear that non-response was more concentrated in some geographic areas, and that groups which are typically harder to count in the census were disproportionately affected in 2018.

A New Zealand admin resident population

The source of admin records for inclusion in the 2018 Census dataset is a New Zealand resident population derived from the linked administrative data in Stats NZ's Integrated Data Infrastructure.

Integrated Data Infrastructure (IDI)

The IDI is a large research database which 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.

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 et al, 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 that is designed to include nearly all of those who interact 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.

The IDI spine in 2018 included around 10 million individuals found in one or more of these sources. Migrants from Australia, the Cook Islands, Tokelau, and Niue do not require a visa, and are only included in the spine once they are issued a tax number. There is a low level of duplication in the IDI spine where links between these datasets have been incorrectly missed, meaning the same individual is included twice.

Deriving an admin resident population for census day 2018

Because the IDI spine aims to include all those who have ever lived in New Zealand, it also provides the basis for selecting a resident population for a given point in time. We construct an estimate of the New Zealand resident population at a given date from the multiple linked admin data sources in the IDI (Stats NZ, 2017). The resulting New Zealand admin resident population derived from the IDI is called the IDI-ERP,[1] and is currently implemented in the IDI for use by researchers. This IDI-ERP population is also the source of admin records for inclusion in the 2018 Census file.

The IDI-ERP includes those individuals in the IDI spine who have activity in selected admin data sources over a two-year period up to the reference date. Those who have 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

[1] An estimated resident population, or ERP derived from the IDI.

date, for example a resident who migrates to live overseas, or a former resident who goes to a doctor while on a temporary visit to New Zealand (and is thus counted in Ministry of Health data).

Experimental population estimates from linked admin data is an experimental data series of national-level admin population estimates released in 2016, with further releases for a subnational geography time-series in 2017 (Stats NZ, 2017), and inclusion of ethnic groups in 2018 (Stats NZ, 2018). The experimental data series includes estimates at 30 June from 2006 to 2016. Online tables compare the IDI-ERP with official population estimates over the same period. These comparisons are largely encouraging. Often there is close agreement with official figures, and consistency has increased steadily over time. However, there are still marked differences for some age groups and local areas.

While the IDI-ERP is a good approximation of the NZ resident population, there are some coverage errors:

- overcoverage – people who are not in fact NZ residents at the reference date, but have been wrongly included
- undercoverage – people who are NZ residents and should be included but have not been selected.

We do not currently have good estimates of the size of these coverage errors. In addition, some people may be included more than once due to duplicates in the IDI spine.

Based on this previous Census Transformation work we derived an admin resident population on census night, 6 March 2018, with some adjustments that improve on the methods described previously. The IDI-ERP used for the 2018 Census is constructed as follows.

Inclusions: Retain individuals whose presence is indicated by activity

- For ages five years and over, the spine population is restricted to those individuals who had activity in at least one of the following IDI datasets in the two years before the reference date (6 March 2018):
    - Inland Revenue tax (employer monthly summary of tax paid at source, or annual tax return data; receipt of taxable benefit payments is included)
    - Ministry of Health (pharmaceutical prescriptions, GP enrolment and attendance, hospital admissions, non-admission hospital visits)
    - Ministry of Education (school enrolment, tertiary enrolment or attainment)
    - Accident Compensation Corporation (ACC) claims (ACC is a national no-fault insurance scheme for injuries)
    - Permanent and long-term arrivals (as measured based on stated intentions from arrival cards). This aims to ensure that recent migrants to New Zealand, who might otherwise not have interacted with government agencies, are not omitted from the admin population.
- For ages under five, a New Zealand birth registration or visa approval (excluding visitor or transit visas) before the reference date is sufficient for inclusion in the population. For this age group there is no additional requirement of activity in the previous two years.

Exclusions:

- **Remove those who have left the population.**
    - Linked death records are used to identify individuals with a date of death before the reference date.
    - Linked migration data is used to identify individuals who were not New Zealand residents on the reference date. Individuals are classified as non-residents if the

total length of time spent overseas is at least six of the 12 months spanning the reference date (that is, six months either side of the 6 March census date).

- **Remove duplicate records in the spine.**
  - We remove records identified as duplicates caused by missed links when constructing the IDI spine. Duplicate pairs are defined as those with identical sex, year and month of birth, and address ID. Both records must have only a single spine source, with one coming from birth registrations and one coming from Inland Revenue tax registrations. This combination provides strong evidence that both records refer to the same person and that a link has been missed. A total of 36,700 duplicate pairs were identified, and one of each pair was removed from the admin population.
- **Remove NZ residents temporarily overseas on census night.**
  - For the 2018 Census application, we also remove NZ residents who are temporarily overseas on census night to achieve the equivalent target population to the census. These people are also identified through the linked migration data.

We expect that some undercoverage and overcoverage errors remain in this IDI-ERP constructed for the 2018 Census. The impacts of undercoverage and overcoverage in the IDI-ERP differ. If a person has not responded to the census field collection, and has also not been selected in the IDI-ERP, then they will be missing from the final census dataset. Overcoverage in the IDI-ERP could lead to a person being wrongly included in the final 2018 Census dataset.

Assigning an admin address

The IDI contains address information from a range of admin sources as well as Stats NZ's 2013 Census and household surveys. These addresses have been matched to the SLR address identifiers. Sources available to be used to allocate a usual residence address ID to people in the IDI-ERP are:

- 2013 Census – address of usual residence as reported on the 2013 Census individual form
- New Zealand Transport Agency – motor vehicle registrations
- Ministry of Social Development – address of residence as reported when applying for working-age benefits and superannuation
- Primary health organisation – address of residence as recorded when visiting a general practitioner
- National Health Index – address of residence as recorded when visiting a hospital or outpatient clinic
- Ministry of Education – address of residence as reported when enrolling at primary or secondary school (but not tertiary education)
- Inland Revenue (IRD) – address for the contact residence supplied to IRD
- Accident Compensation Corporation (ACC) – address of residence as reported when filing an ACC claim.

A small number of addresses are also available from Stats NZ surveys (Household Labour Force Survey, Household Economic Survey) and from the Auckland City Mission. Ministry of Social Development postal addresses are also available but are lower quality, and are only used if no other address source is available.

We have multiple address sources, each of varying quality, covering different groups in the population, and with varying incentives for people to update their address information. Each address notification has an associated date, reflecting when that information was provided to an agency. We use this date when selecting addresses for individuals at a given point in time. To assign an address

on census day, we select the most recent address notification prior to census day from any of these sources. If there are conflicting addresses updated on the same day, the sources are ranked in order of preference as shown in the list above.

For children aged up to 17 who have two parents (as defined by Gath & Bycroft, 2018) present in the IDI-ERP at the same address, we assign the child that same address. For any children with fewer than two parents identified in the IDI-ERP, or whose parents are recorded at different addresses, we make no adjustment.

Differences between census responses and admin sources

The use of admin records along with responses from the census collection process means that we are combining data sources with some key differences. Here we discuss the implications of census-night occupancy versus usual residence, and the impact of different timeliness between admin sources and the census collection period.

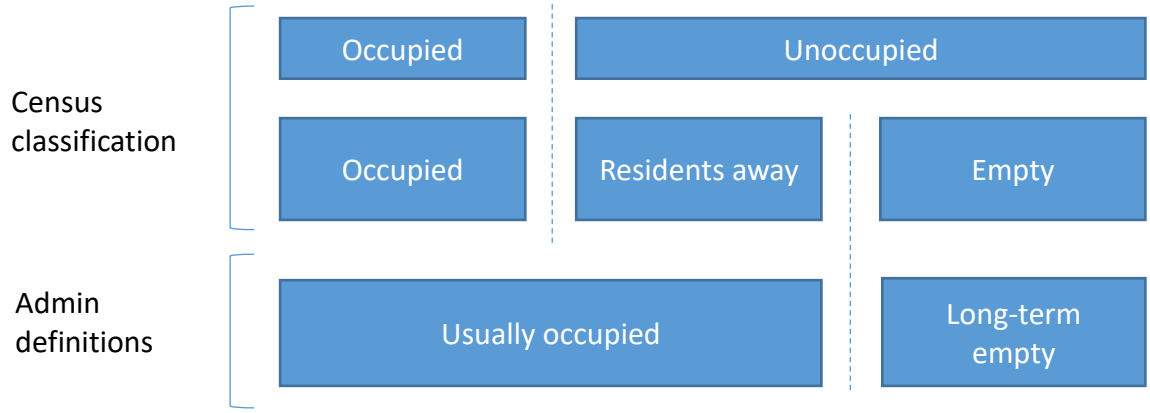
Conceptual differences: usual residence versus census-night location

The census distinction between Occupied dwellings and Unoccupied dwellings (residents away, or empty) is due to the nature of the census field collection, where forms are only collected from dwellings occupied on census night. These operational constraints do not apply to admin sources.

An admin address refers to where someone usually lives. However, in contrast to the census occupancy definitions, this does not tell us whether the person is at home on a specific night. To manage the difference between the admin and census concepts, we define a usually occupied dwelling as a dwelling where people are living around census day. Figure 1 shows how the admin usually occupied category combines the Occupied and Residents Away census categories, cutting across the census high-level occupancy classifications. The remaining admin dwellings that are not usually occupied are defined as long-term empty, equating to the census 'empty' category.

Figure 1

Occupancy definitions in census and admin collections



Consistency of census and admin location

Linked census and admin data can be used to compare the consistency of usual residence location between the two sources. Previous research has shown the IDI-ERP to be a good approximation of the NZ usual resident population, overall and for larger geographic areas, but that accuracy decreases at smaller geographies. In a comparison of the admin address in 2013 with 2013 Census data, 93 percent of IDI-ERP records were in the same territorial authority and Auckland local board (TALB) as the 2013 Census, 84 percent were in the same meshblock, and 82 percent had the same address (Stats NZ, 2017). Consistency had improved in 2018, with 95 percent agreement between census responses and admin records for TALB, 89 percent for meshblocks, and 87 percent for addresses. The recent availability of addresses from motor vehicle registrations and improved address matching to the SLR are likely to have been influential in this improvement.

Individuals can also be grouped into households based on their admin address. Gath and Bycroft (2018) found that 48 percent of admin households had exact agreement with the 2013 Census in terms of membership. Smaller households and households with older adults (aged 65-plus) were more likely to agree.

Timeliness of admin sources

Use of admin data for statistical purposes needs to take account of time lags in the admin sources, as the information will not, typically, have been collected on census day. Timeliness depends on the frequency of interaction of individuals with an agency, and how often data is supplied to Stats NZ. Admin data for the 2018 Census are drawn from the IDI as at September 2018, some six months after the census day reference date in March 2018.

While census data provides point-in-time information that is accurate around the census date, a limitation of admin data is that a delay in its supply to Stats NZ means that the most up-to-date information is not always available for the census date.

For datasets that determine who is included in the IDI-ERP, most, but not all, agencies have provided data that covers the full two-year period prior to the March 2018 census date. While considerable redundancy is built into determining activity, in some cases a person will not be selected into the IDI-ERP because they only interact with a certain activity data source that does not include activity for several months before census day.

Appendix 1 provides full details for each of the data sources used in constructing the IDI-ERP, and the period they cover.

The main impact of admin data lags for whether a person is correctly included in the IDI-ERP for the 2018 Census 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 up to 31 July 2018. This five-month interval after 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. Birth and death registrations up to 27 July 2018 are available. This should include almost all deaths that have occurred before the census reference date, but some births may not have been registered in that time frame. These admin time lags may mean that some of those included in the IDI-ERP were not members of the NZ resident population on the census date, and some newborn babies may have been missed.

The use of admin address information may be more sensitive to time lags. While all but one of the agency sources of address information are available past March 2018, there still may be a lag from

when people change their address to when they notify a government agency of the new address. People who have moved close to the census date may be less likely to have up-to-date admin address information on census day. A lag in the availability of address information would mean that some people could be counted at an old address where they no longer live, resulting in incorrect information about household members and potential impacts on small-area geographic information.

Linking census responses to the IDI

Census responses are linked to the IDI to find people in the admin data who are already counted in the census. Record linkage attempts to find pairs of records from two data sources that refer to the same person. These record pairs for the same person are true 'matches', and other record pairs that refer to different people are non-matches. See for example Fellegi and Sunter (1969) for a formal description.

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[2] between identified census records and an identified population list that is the basis of the IDI spine.[3] Probabilistic linkage based on the models developed by Fellegi and Sunter is applied in a fully automated process. 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. Stats NZ, 2019c, provides details of the linkage process. The main results are shown in table 1.

Table 1

Results from linking 2018 Census responses to the IDI spine

Type of 2018 Census response	Census linking population	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.
Source: Stats NZ

The overall link rate of 97.7 percent is high in the context of other linkage rates in the IDI. The quality of the available variables used to compare records affects our ability to link a census response to the IDI spine. The slightly higher link rate for online responses compared to responses from paper forms reflects difficulties scanning handwritten name and date of birth from paper forms. The lowest link rate of 84 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 not the full date of birth, or country of birth.

[2] Only variables used in the linkage process are available, and access is highly restricted to staff carrying out the linkage. Stats NZ, 2019a gives further details.

[3] 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.

Since the admin NZ resident population is selected from the IDI spine, this linkage process automatically provides links between the census responses and the IDI-ERP admin population used in the 2018 census.

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.

The linkage is designed to minimise false positive linkages. Around 0.6 percent of the linkages are estimated to be incorrect, an estimate derived from manually checking a small sample of linked records. An incorrect link is not a problem in this context if another census response should have been matched to that admin record, since they are still ineligible for selection in the IDI-ERP. If a person did not respond to the census, but their admin record was incorrectly linked to a census response, the incorrect link means the person cannot be included in the census dataset, and the undercount will remain in the final census dataset.

The false negative match rate for Individual forms is estimated as 1.21 percent (Stats NZ, 2019d). If we have not linked a census record to the IDI spine when we should have, the same person could be included in the final census dataset twice. The methods section describes an adjustment made to account for missed linkages.

Methods

The planned approach for non-response in 2018 was a unit imputation model analogous to that used in previous censuses, where full household membership is imputed for occupied but non-responding private dwellings, by borrowing household characteristics from 'nearest neighbour' donor households (Stats NZ, 2014c). This imputation method relies on a missing-at-random assumption, that is, that non-responding households are similar to responding households in the same neighbourhood.

While some improvements in targeting similar households were proposed for 2018, the extent of non-response to the 2018 Census put the validity of the underlying missing-at-random assumptions at risk.

We proposed using the admin resident population instead as an alternative way to count people who had been missed from the census field collection. This additional use of admin data was not planned in the build-up to the 2018 Census, and the project was under considerable time pressure to minimise delays in releasing census outputs.

We tested various approaches, before we determined the final methods. As part of our work on testing approaches, we derived a new population benchmark at the time of the census, which we used to guide the development of the new methods.

Dual system estimation population benchmark

Dual system estimation (DSE) methodology combines two independent lists to estimate the size of a population (see for example, Seber, 1982). DSE combining census responses and an admin population has been used to derive a new estimate of the NZ resident population at the time of the 2018 Census. Stats NZ (2019d) describes how dual system estimation has been applied in this context, with a focus on meeting key assumptions that underpin the method. New methods were developed and applied to remove erroneous records from the admin data, and a linkage error adjustment was applied to the DSE calculation.

Census-day benchmark population estimates by age and sex, for key Māori, Pacific, and Asian ethnic groups, and all TALB subnational geographies have been produced. The DSE population is preferred for our purposes over the available official 2013-based ERP because the contributing data sources are available directly for the March 2018 census date, and so are not affected by errors estimating population change over the five-year period since 2013. The DSE approach is also able to include more ethnic groups for March 2018 than official population estimates.

The DSE results show 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 undercoverage remaining in the final 2018 Census dataset, although these cannot be confirmed until the official coverage measurement is available, in 2020.

Early methods development

The first approaches investigated were aimed at reducing reliance on unit imputation by using admin data to construct households for some of the non-responding dwellings. Several strategies were investigated. These included applying a threshold to only accept admin households with high-quality household membership, compared with including all admin households; and extension of admin

households to include dwellings unoccupied on census night. Smaller mitigations were investigated to use admin data for responding dwellings, and for some non-private dwellings.

Results were assessed against the new DSE benchmark population. We focused on improving population counts (levels) and improving population structure and distributions (proportions) of key demographic variables. We prioritised population structure over placing people into families and households, and prioritised the use of admin data over donor imputation.

While placement of admin records into dwellings did show some improvement over reliance solely on unit imputation, the scale of the non-response meant that the likely census undercount was still too high to provide an acceptable census dataset. There were also a considerable number of people in the admin resident population who had not yet been included in the census dataset. We then extended admin enumeration to count individuals we were confident should be part of the census usually resident population, even if we were unable to place them at a specific dwelling. Donor imputation could no longer be used in this scenario, as it would lead to double counting.

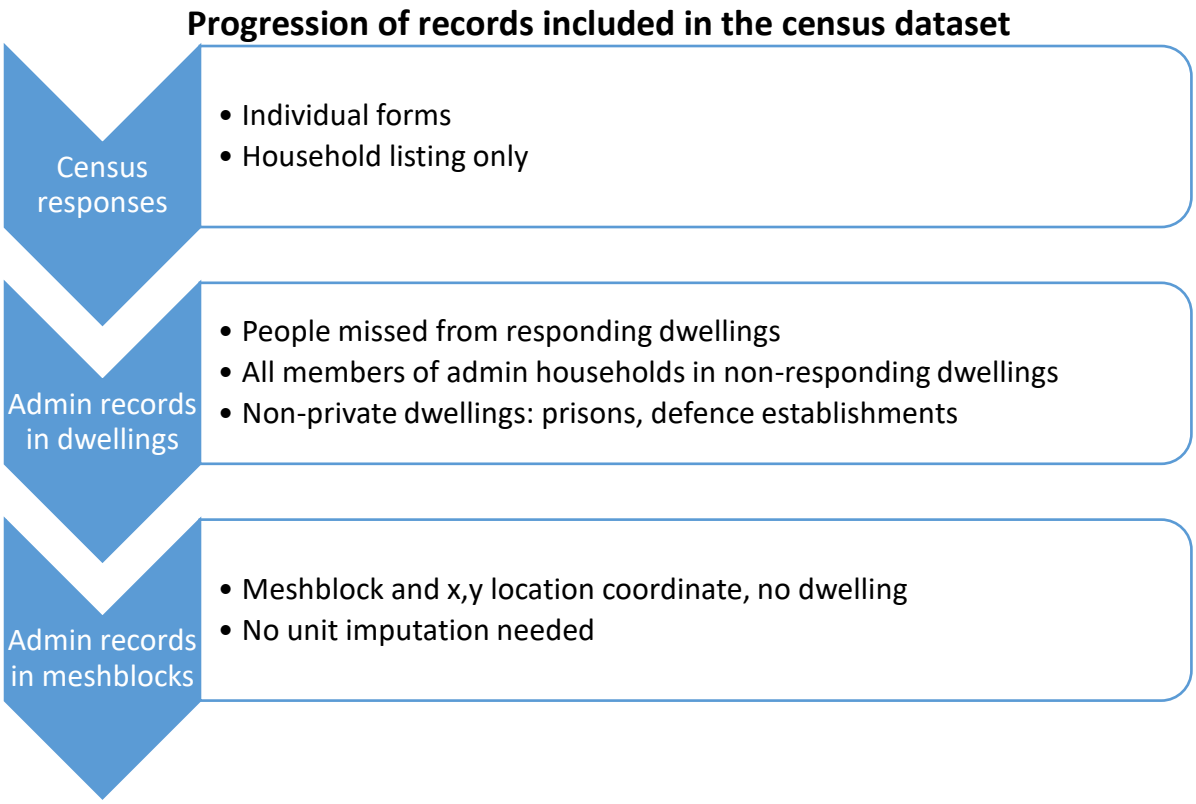
Final methods

The 2018 Census dataset consists of records sourced from census responses, supplemented by high-quality admin records. We now describe the methods applied for the 2018 Census that determine how admin records are added to the census dataset.

These statistical methods have been developed to manage the known limitations of the admin sources and are used to make decisions about which admin records are included in the final census dataset. We are more concerned with avoiding potential overcoverage errors introduced through the use of admin sources than we are to eliminate all undercoverage in the 2018 Census dataset. The primary purpose of the post-enumeration survey is to detect undercount in the census dataset and adjust the ERP official population estimates for this. However, there is little remedy for national or local overcoverage in the census dataset, and this would likely flow through to the ERP.

Figure 2 shows the progressive inclusion of records in the 2018 Census dataset.

Figure 2



Census responses always take priority, and census responses are collected for people within a dwelling. Following the traditional approach for adjusting the census dataset for non-response, admin records are first placed into dwellings. Otherwise, we do not place admin records in dwellings, but within a small geographic area (a meshblock, up to around 100 people), when we can do so while improving the quality of census small-area data.

This extensive use of admin records removes the need for any unit imputation in the census dataset, as has been the practice in previous censuses. The other major difference is in the relationship of individuals and dwellings. In contrast to previous censuses, some 'occupied, but non-responding' dwellings have no residents, while some dwellings with 'residents away' on census night do have usual residents; further, some people are not placed in dwellings at all.

The admin resident population – the IDI-ERP described above – provides an initial admin list of people who can be considered for inclusion in the census dataset. We must first exclude those for whom we already have a census response from the list of candidates for admin enumerations. We also need to recognise limitations in the quality of admin sources when we are determining who should be placed into the census file. When faced with choices of how to implement methods and where to set model parameters, decisions have been conservative. Admin enumerations have been added to the census dataset when they will improve the overall quality of the census data.

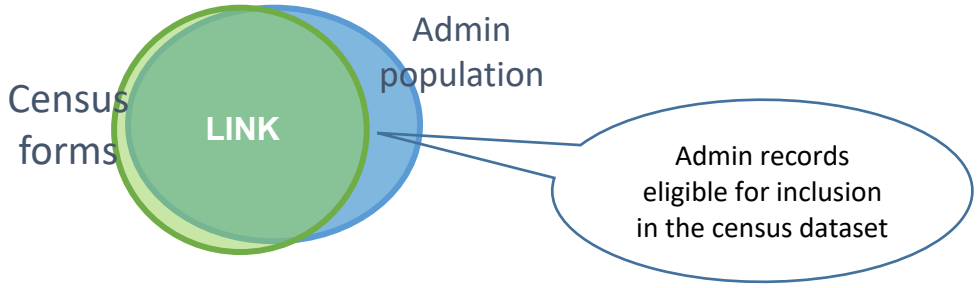
We now outline the statistical methodologies used to include high-quality admin enumerations in the census file.

Step 1. Census responses take priority

Census responses always take priority and include both Individual forms and those listed as belonging to a household, but for whom no Individual form has been received. The link between the census responses and the IDI-ERP identifies those from the admin population who have already responded to the census. These linked IDI-ERP records, as well as residents who were temporarily overseas on census night, are removed from further consideration. Those remaining in the IDI-ERP are then eligible for inclusion in the census dataset. Figure 3 shows a simple illustration of this.

Figure 3

Admin records eligible for inclusion in the census dataset



Step 2. Admin enumerations in dwellings

Previous censuses adjusted for non-response by adding imputed records within dwellings where responses were thought to have been missed. Admin enumerations are included within dwellings in three ways. As each is directed at a different set of dwellings, there is no overlap between them.

Admin enumerations in responding dwellings

Admin records are included in responding private dwellings only within strict conditions. We require strong evidence from the received responses that a person who usually lives there has been missed, and the address and relationship characteristics must be mapped to an eligible admin record.

Examination of responses to the 2018 Census living arrangements question indicated some instances where people reported living with individuals who were not included in the census usually resident household. We used this evidence to infer that one or more members of the household had not responded. We then used the admin address and relationship information to identify these missing household members and add them to the household.

Using these rules, we added 20,643 people from admin records into 13,533 dwellings, with almost half originally appearing as one-person households. Most dwellings (95 percent) had between one and three people added. Nearly all (97 percent) have a complete family relationship grid, a good indication of consistent relationships among household members.

Appendix 2 provides further details.

While households where we have received some census responses may still be missing people, we have not developed a model to predict when admin records ostensibly for the same address should be placed within those responding households.

Admin household enumerations in non-responding private dwellings

The census dwelling frame provides a list of addresses for private dwellings where no census responses have been received. Admin households are complete households constructed from admin data at an address for a private dwelling. Since admin addresses refer to where people usually live, not where they happen to be on census night, admin households may be derived for non-responding census dwellings regardless of whether they were classified as occupied by census field operations or whether residents were determined to be away from home on census night.

This use of admin data to reliably group people within the same dwelling to form households demands high-quality admin address information. We have developed statistical models that predict which households constructed from admin records are likely to have reliable data. The approach is based on methodology developed by the US Bureau of the Census (Keller et al, 2018). While the focus of the US approach is to reduce contacts in their field operations, similar approaches can be used after data collection is completed to use admin records in lieu of imputation.

2018 Census responding households are assumed to represent the truth when training and assessing the models. A model score represents how reliable the admin data is for representing the entire household in a given dwelling, and a cut-off determines which of those households will be added to the census dataset.

The cut-off has been set as a balance between strict criteria of obtaining exactly the same people in the household as we observe in the census, and including admin households that reflect similar adult–child patterns as the census, even if we cannot guarantee that all household members are the same. Making the trade-off in this way means we include relatively more large or complex households than if we had set a more conservative cut-off and makes some allowance for errors in census responding households.

Of the 89,355 admin households that could be derived for non-responding census dwellings, 57,609 (64 percent) were selected into the census dataset. These households contained a total of 141,411 admin records for individuals: 99,159 individuals in households that were assessed as occupied on census night, and 42,252 in households where residents were away on census night. Relatively more households with a single adult (without children) and two-adult households (with or without children) are selected than other categories (single adult with children and three-or-more-adult households). This reflects the level of consistency between census and admin data generally observed for these different household types.

Stats NZ (2019e) provides a full description of the statistical models and further results.

Admin enumerations in non-private dwellings: prisons and defence establishments

Prisons and defence establishments are typically large institutions. They are important in their own right, and have a big impact on the census data for the small areas where they are located. However, there was a substantial census undercount in both types of non-private dwellings. In line with their responsibilities to ensure that census information is provided for those in their establishments, the Department of Corrections and Ministry of Defence provided datasets to Stats NZ containing unnamed unit records for every individual at each of their respective establishments on census night. These data are used to adjust for the gap between census responses and known census-night occupancy.

Department of Corrections data are already included in the IDI and list all those in prisons on census night. This allows us to determine who has already responded to the census. However, as the IDI

data does not state each prison or penal institution, the main use of the Department of Corrections file was to provide the census-night dwelling location for those who did not respond to the census.

As there is no defence data in the IDI, the file supplied by the Ministry of Defence is compared directly with the census records received for each defence establishment. The number required to make up the total for each defence establishment are then selected from the non-responding defence records for inclusion in the census file.

Overall 4,707 admin enumerations were added for prisons or penal institutions, making up just less than half the final total census count in these institutions. For defence establishments, 798, or one-quarter of the 3,150 total census count, are admin enumerations.

Combining census responses and agency-sourced data brought 2018 Census results for prisons and defence establishments much closer to the actual number of occupants on census night.

Appendix 3 provides details of the use of admin data for prisons and penal institutions, and for defence establishments.

Step 3. Admin meshblock enumerations

The remaining IDI-ERP population eligible for inclusion in meshblocks are those who have not been linked to a census respondent and are not already included at a dwelling. We then explicitly adjusted for potential coverage errors that could result from overcoverage in the IDI-ERP, and for duplication caused by missing linkages between the census and the IDI. We have assumed there is a negligible risk of overcoverage and missed matches for the admin records that have been placed in dwellings.

We draw on the new methods developed to remove overcoverage and to adjust for missed matches in the context of the DSE population estimate, based on the same linked census responses and IDI-ERP admin population.

After these adjustments we can be confident that the remaining eligible IDI-ERP people should have been counted by the census. However, we must also consider the accuracy of admin location data, which decreases as geographies get smaller. To limit the errors for small subnational geographies, we removed people who have a low probability of being in the correct admin meshblock.

Adjusting for admin overcoverage

Our approach is designed to remove overcoverage caused by people selected in the IDI-ERP who are not NZ residents at the time of the census. The main source of coverage uncertainty for this census application is likely to be determining migration status on census day, largely due to admin time lags, as we have data for border movements only until July 2018.

As we cannot directly find overcoverage, we will also remove valid records. The goal is to effectively remove as much overcoverage as possible, while also minimising the amount of undercoverage being introduced. We call this subset ‘IDI-ERP_Sure’, to indicate that it includes only people who are ‘sure’ to belong to the NZ 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. Tax and health datasets have high coverage of the population, and requiring an additional activity provides stronger evidence that

people are in fact residents. This approach targets all age and sex groups, combines factors found to predict overcoverage in previous unpublished research, and is easy to apply and explain.

The majority of records removed are young adults, predominantly male, and those approaching the retirement age of 65. The young males are likely candidates for overcoverage. However, the rules may be unnecessarily removing those who take early retirement. Further details are provided in Stats NZ (2019d).

Of the 602,889 IDI-ERP records not already included in the census dataset, 130,311 are excluded to account for overcoverage.

Adjusting for missed matches

In this 2018 Census context we need to avoid double-counting people in the census dataset, which would be caused by missed matches when we link census forms to the IDI spine. Before selecting admin records to place in meshblocks, we first remove overcoverage to create the IDI-ERP_Sure as described above, and then adjust for missed matches.

The rate of missed matches is estimated as 1.21 percent (Stats NZ, 2019d). From this we calculate the number of missed matches to remove from those in the IDI-ERP_Sure who have not been linked to the census. In total, 47,670 records were removed. The adjustment is applied for breakdowns by demographic groups and TALB subnational geographies. The appropriate number of people are removed through random selection within subgroups.

This adjustment does not remove the exact people whose census forms should have been matched to the IDI. Rather it removes a random selection of the right kind of people down to a fine demographic breakdown. This random selection introduces an element of uncertainty into the census dataset.

The estimated rate of missed links makes several assumptions. While these assumptions on the whole appear to be reasonable, they may underestimate the number of missed matches for those listed on the household form, and have not been verified through a clerical review or by other means. The presence of incorrect links introduces a small bias in the other direction, towards overestimation of the number of missed links that have been removed.

The adjustment for missed linkages between census forms and the admin data is designed to prevent the inclusion of too many admin enumerations in the final census dataset. While we cannot completely justify the methods used, they have good face validity, and make a small but important contribution to the use of admin records in the 2018 Census.

Appendix 4 provides a formal description of the methods used to adjust for missed matches.

Predicting reliable admin meshblocks

To limit the errors for small subnational geographies, we also exclude people who have a low probability of a correct admin meshblock (Stats NZ, 2019e). We apply a statistical model adapted from one used to predict reliable admin households.

For all eligible individuals, we compute a model score for the likelihood that their admin meshblock reflects their true meshblock of usual residence. We include all individuals with a probability of greater than 0.5 as meshblock enumerations. This score indicates that their admin meshblock is more likely than not to represent their true usual residence on census night.

The cut-off once again represents a trade-off, this time between maximising the use of admin data to improve national demographic counts, and minimising the number of individuals enumerated in the wrong area. More conservative cut-off values would ensure we are less likely to have incorrect inclusions for a given area but would reduce the overall effectiveness.

By removing overcoverage and making additional adjustments for linkage error, the admin records for individuals eligible for meshblock enumeration are considered highly likely to be part of the usual resident population on census night. We are willing to accept the possibility of some error at the small geography level when it is likely to improve the overall population counts and demographic distributions.

In total, 357,294 people (85 percent of eligible records) are selected as admin meshblock enumerations. Almost all eligible admin records aged 0–17 or 65-plus are included in the census dataset. Considerably more young adults are rejected for having low probabilities of being in the correct meshblock, although the majority are still included as admin meshblock enumerations.

Some 64,815 admin records highly likely to be NZ residents remain, and are not included in the census dataset because their admin meshblock has a low probability of being correct.

Summary of admin records

The previous sections have described how the initial IDI-ERP admin resident population is treated in relation to the 2018 Census. Table 2 summarises each step whereby the admin records are excluded from consideration, or else included as admin enumerations in the final census dataset.

Table 2

Flow for determining which admin records to include as admin enumerations			
	Count		
	Included as admin enumeration	Exclusions	Remaining admin records
Total IDI-ERP at 6 March 2018			4,819,233
Exclude: residents temporarily overseas		67,635	4,751,598
Exclude: census-night prison population[1]		9,978	4,741,620
Exclude: records linked to 2018 Census response		3,976,671	764,952
Include: Admin enumerations in responding private dwellings	20,643		744,306
Include: Admin enumerations in non-responding private dwellings	141,411		602,889
Exclude: IDI-ERP overcoverage		130,311	472,578
Exclude: Adjustment for missed linkages		47,670	424,908
Exclude: Otherwise ineligible records		2,814	422,097
Include: Admin meshblock enumerations	357,294		64,815
Records below meshblock cut-off			64,815

[1] Admin enumerations in defence establishments have been determined separately and are not represented here.

Source: Stats NZ

Results

We first show the characteristics of people added as admin enumerations compared with census responses. We then compare the final 2018 Census dataset with the DSE benchmark population, and lastly summarise sensitivity analyses that have been undertaken for key uses of census data.

Contribution of admin enumerations to the final 2018 Census dataset

This section describes the impact of the admin enumerations on key demographic distributions. We look at the breakdown of the census dataset by census responses and across each of the approaches:

- admin household enumerations
- other dwelling-based admin enumerations, for responding private dwellings and for non-private dwellings (prisons and defence establishments)
- admin meshblock enumerations.

Table 3 shows the number of records included from each source. Overall, census responses make up 89 percent of the census dataset, and admin enumeration 11 percent. The admin household enumerations make up 3 percent of the final 2018 Census usual resident population count, and the dwelling-based admin enumerations contribute 0.6 percent, while the largest group is admin meshblock enumerations, which make up 8 percent.

Table 3

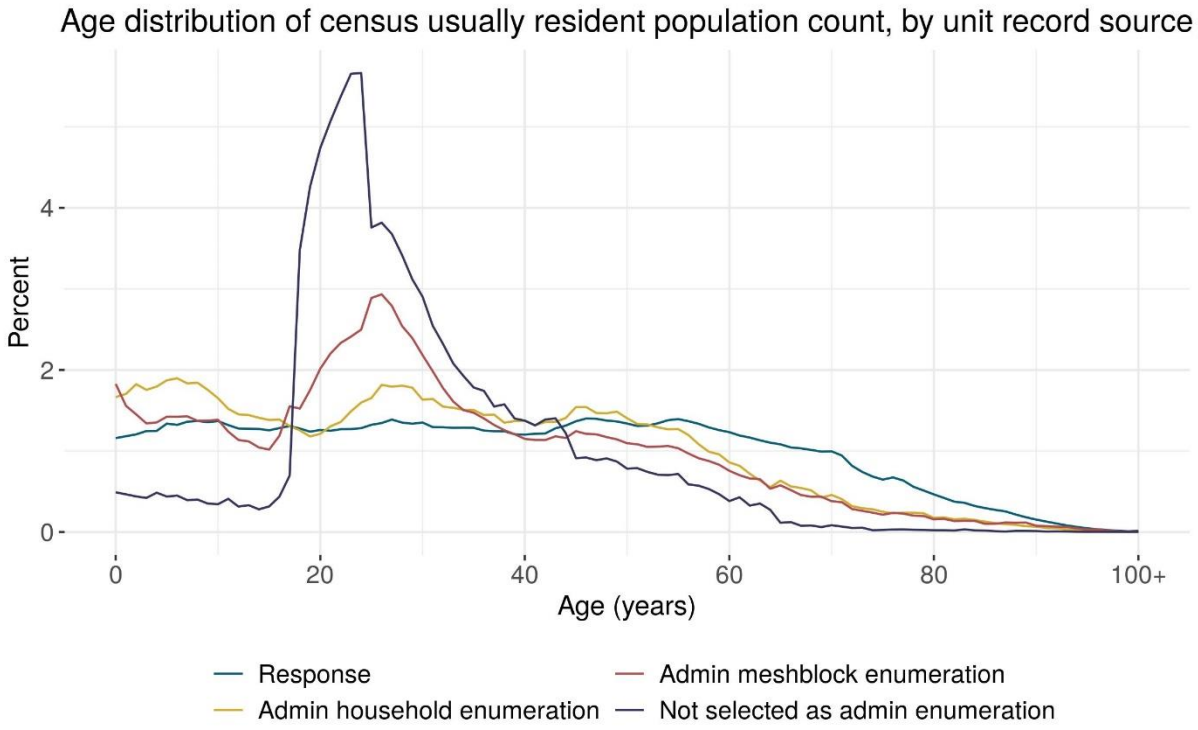
2018 Census usual resident population count, by unit record source				
Unit record source	Count		Percent	
Individual form	3,971,892		84.5	
Individual from household listing	202,914		4.3	
Field-enumerated rough sleeper	99		<0.1	
Census responses		4,174,902		88.8
Admin household enumerations – occupied	99,159		2.1	
Admin household enumerations – residents away	42,252		0.9	
Admin household enumerations		141,411		3.0
Admin enumeration at responding private dwelling	20,643		0.4	
Admin enumeration at prison or penal institution	4,707		0.1	
Admin enumeration at defence establishment	798		<0.1	
Other dwelling-based admin enumeration		26,148		0.6
Admin meshblock enumerations		357,294		7.6
Total		4,699,755		100
Source: Stats NZ				

The age distributions of the admin enumerations are markedly different from received census responses (figure 4). Both the admin household and meshblock enumeration groups tend to include relatively more children and young adults, and fewer people in the older age groups. This trend is most observable for the meshblock enumerations, with 34 percent of these individuals aged 20–34, compared with 20 percent of respondents. This suggests young adults are less likely to have responded to the census than other age groups and are also less likely to have been part of high-quality admin households. In contrast, 16 percent of census respondents were aged 65 or older,

compared with 7 percent for both the admin household and meshblock enumerations, indicating higher census participation for older age groups.

Figure 4 also shows the age distribution of admin records that have been excluded because their admin meshblock was predicted to be low quality. These excluded records are overwhelmingly in the 18-to-24-year age group. This feature is driven by the meshblock prediction model, as this age group tends to have a lower probability of a correct admin meshblock due to poor-quality address information.

Figure 4

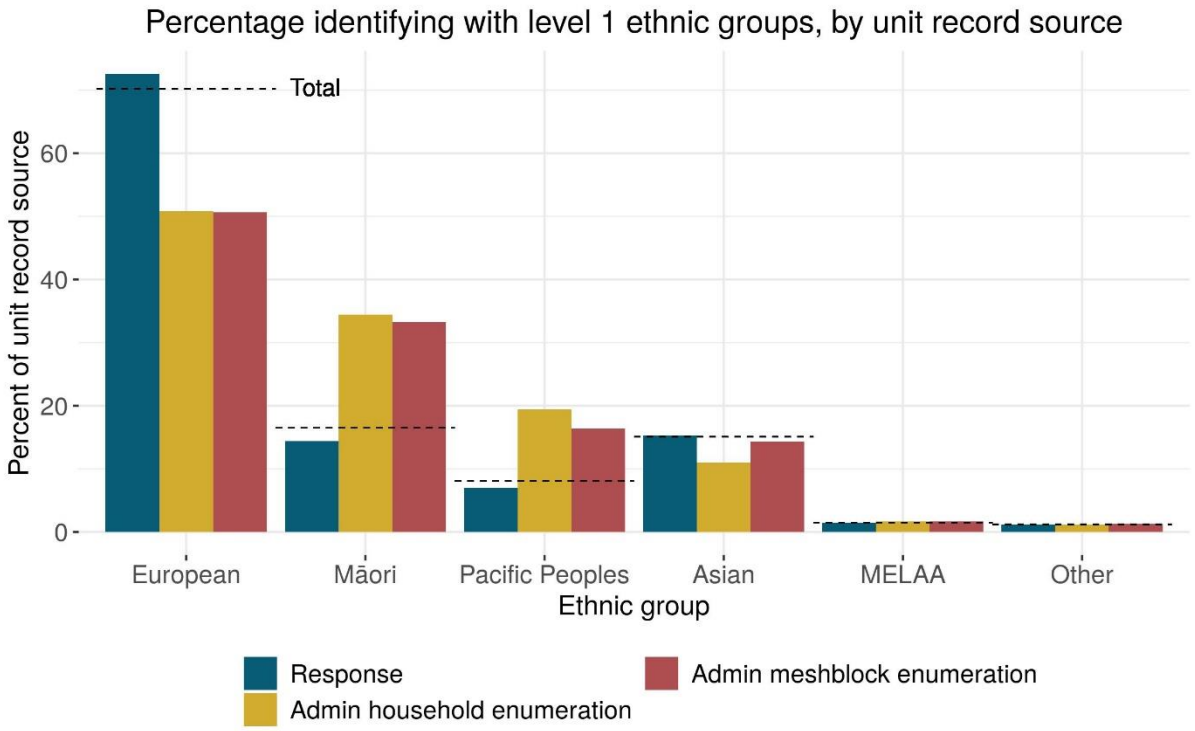


Similarly, figure 5 shows the distribution of level 1 ethnic groups in each source (individuals who identify with multiple ethnic groups are included in multiple categories). The dotted lines show the percentage of the total census usually resident population identifying with each ethnic group.

The admin enumeration categories had considerably higher proportions of people identifying with Māori and Pacific ethnicities than for census respondents, and a lower proportion of people identifying with the European ethnic group. This indicates lower participation for people of Māori and Pacific ethnicity in the 2018 Census.

The ethnic distributions for admin household and meshblock enumerations are reasonably similar, with a slightly higher proportion of admin household enumerations identifying with Pacific ethnicities, and a slightly lower proportion identifying with Asian ethnicities.

Figure 5



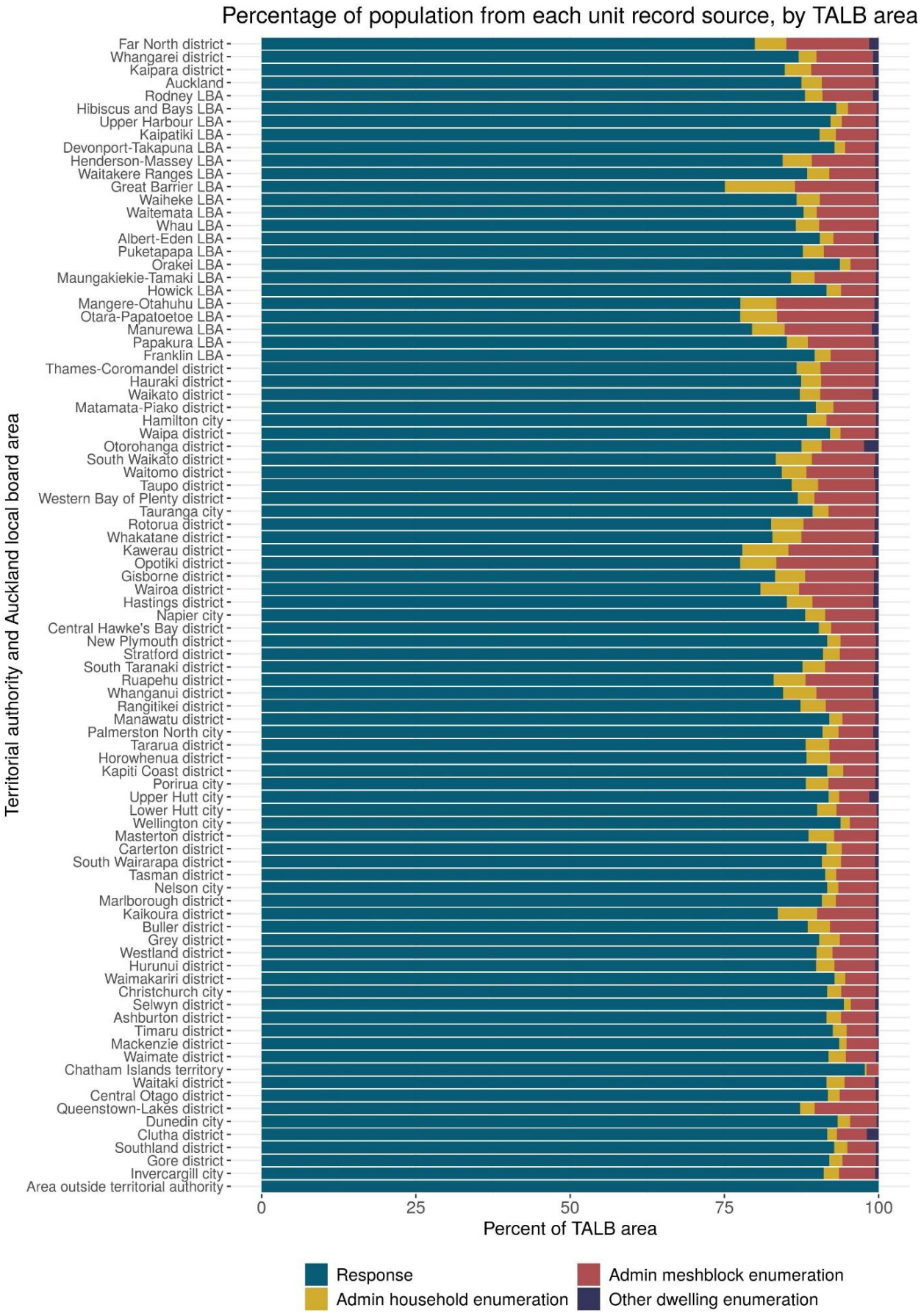
As would be expected from these results, admin enumerations make up a greater proportion of the final census dataset for young adults, and for Māori and Pacific ethnic groups. Almost a quarter of the Māori and Pacific populations are admin enumerations, compared with 10 percent for the European ethnic group (Table 4). These populations have been associated with higher levels of undercoverage in previous censuses (Stats NZ, 2014c). The prevalence of these groups among the admin enumerations is an indication that admin records do include those who are typically harder to count in a census field enumeration. The lower quality of admin address information is the main factor limiting their inclusion in the census dataset.

Table 4

Proportion of census responses and admin enumerations, by level 1 ethnic group				
Ethnic group	Census responses		Admin enumerations	
	Count	Percent	Count	Percent
European	3,029,625	91.9	268,236	8.1
Māori	600,078	77.3	175,761	22.7
Pacific	291,957	76.5	89,688	23.5
Asian	638,514	90.2	69,084	9.8
MELAA	61,626	87.6	8,706	12.4
Other	51,009	87.9	7,041	12.1
Total	4,174,902	88.8	524,853	11.2
Source: Stats NZ				

Figure 6 shows the distribution of unit record source for each territorial authority and Auckland local board (TALB) area. Excluding areas with very small populations, the percentage of the overall usually resident population count coming from census responses ranges from 78 percent for Ōpōtiki district to 94 percent for Selwyn district. Six areas have more than 20 percent of the population count coming from admin enumerations – Ōpōtiki, Kawerau, and Far North districts, and Māngere–Ōtāhuhu, Ōtara–Papatoetoe, and Manurewa local board areas, all areas with higher proportions of Māori and Pacific peoples.

Figure 6


Percentage of population from each unit record source, by TALB area
Territorial authority and Auckland local board area
Far North district
Whangarei district
Kaipara district
Auckland
Rodney LBA
Hibiscus and Bays LBA
Upper Harbour LBA
Kaipatiki LBA
Devonport-Takapuna LBA
Henderson-Massey LBA
Waitakere Ranges LBA
Great Barrier LBA
Waiheke LBA
Waitemata LBA
Whau LBA
Albert-Eden LBA
Puketapapa LBA
Orakei LBA
Maungakiekie-Tamaki LBA
Howick LBA
Mangere-Otahuhu LBA
Otara-Papatoetoe LBA
Manurewa LBA
Papakura LBA
Franklin LBA
Thames-Coromandel district
Hauraki district
Waikato district
Matamata-Piako district
Hamilton city
Waipa district
Otorohanga district
South Waikato district
Waitomo district
Taupo district
Western Bay of Plenty district
Tauranga city
Rotorua district
Whakatane district
Kawerau district
Opotiki district
Gisborne district
Wairoa district
Hastings district
Napier city
Central Hawke's Bay district
New Plymouth district
Stratford district
South Taranaki district
Ruapehu district
Whanganui district
Rangitikei district
Manawatu district
Palmerston North city
Tararua district
Horowhenua district
Kapiti Coast district
Porirua city
Upper Hutt city
Lower Hutt city
Wellington city
Masterton district
Carterton district
South Wairarapa district
Tasman district
Nelson city
Marlborough district
Kaikoura district
Buller district
Grey district
Westland district
Hurunui district
Waimakariri district
Christchurch city
Selwyn district
Ashburton district
Timaru district
Mackenzie district
Waimate district
Chatham Islands territory
Waitaki district
Central Otago district
Queenstown-Lakes district
Dunedin city
Clutha district
Southland district
Gore district
Invercargill city
Area outside territorial authority
0
25
50
75
100
Percent of TALB area
Response
Admin household enumeration
Admin meshblock enumeration
Other dwelling enumeration

The observed differences across age, ethnicity, and geographic area suggest that some population subgroups were more likely than others to respond to the 2018 Census. A key advantage of the admin enumeration approach is that these populations are better represented than they would be based only on imputation using census responses. Admin data is generally of high quality for key demographic variables and so we can be confident the different distribution of admin enumerations compared with census responses is likely to represent an improvement to census counts.

Comparison of final 2018 Census dataset and DSE benchmark populations

The DSE benchmark population is the most suitable estimate of the true census usual resident population available. Comparison with the DSE provides an indication of the coverage patterns for the 2018 Census after admin enumerations have been included. DSE benchmarks are available by single year of age, for Māori, Pacific, and Asian ethnic groups, and for TALB geographies. Further breakdowns are not available due to constraints of the estimation method (Statas NZ, 2019d).

The DSE population total is 4,768,600, compared with the total census count of 4,699,755, 1.4 percent smaller than the DSE. Until the formal post-enumeration survey undercount measurement is available, this 1.4 percent net undercount compared with the DSE provides our best indication of census overall coverage. It compares with a 2.4 +/- 0.5 percent undercount for the 2013 Census (Stats NZ, 2014a).

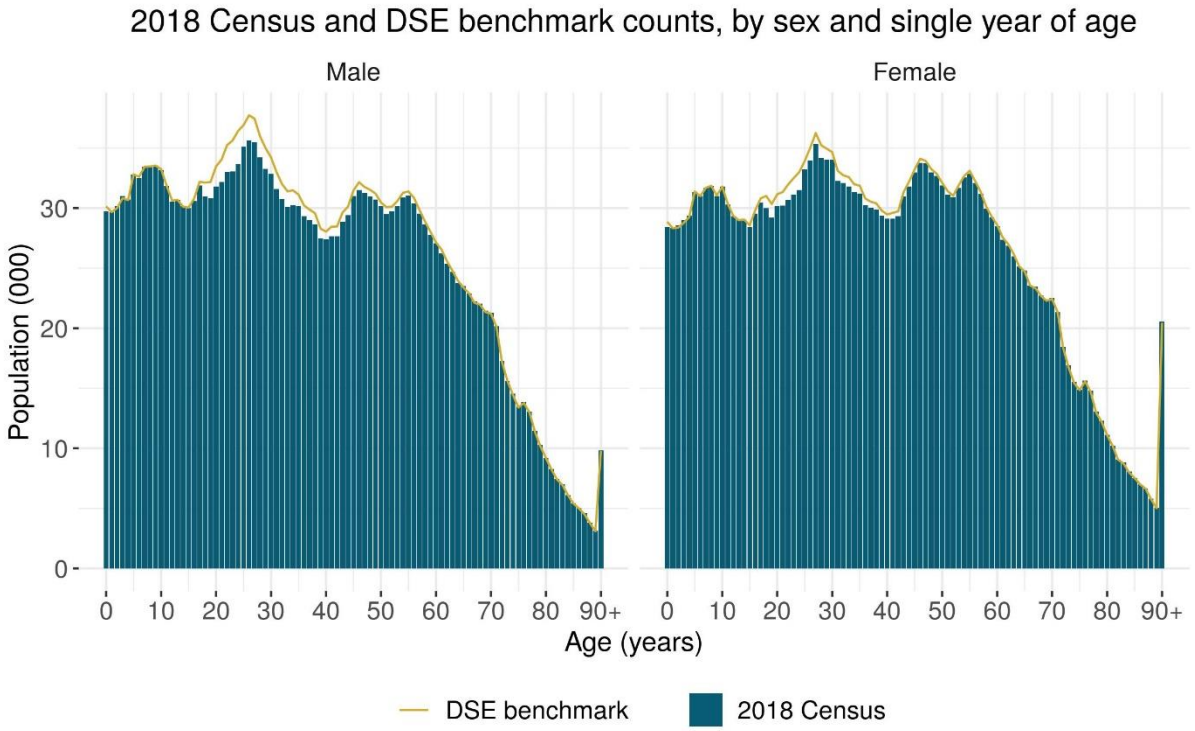
Table 5 shows the DSE population benchmarks compared with the census counts for several key groups. Despite admin enumerations having significantly boosted populations for Māori, Pacific people, and young adults, the undercount is still higher for these groups than the overall average.

Table 5

2018 Census and DSE benchmark counts, by age group and selected ethnic groups			
Population group	2018 Census usual resident population count	DSE benchmark	Difference between 2018 Census and DSE benchmark (percent)
Ages 0–17	1,104,240	1,106,100	-0.2
Ages 18–24	438,381	460,400	-4.8
Ages 25–44	1,248,240	1,285,100	-2.9
Ages 45–64	1,193,724	1,204,800	-0.9
Ages 65+	715,167	712,100	0.4
Māori ethnic group	775,836	807,900	-4.0
Pacific ethnic group	381,642	397,200	-3.9
Asian ethnic group	707,598	727,400	-2.7
Total	4,699,755	4,768,600	-1.4
Source: Stats NZ			

Figure 7 compares the census and DSE distributions by single year of age, for males and females. The census follows the DSE very closely for children and for older adults, while there is some undercount apparent for younger adults. The gap is more marked for younger males than for females. These patterns are consistent with younger adults being more likely to be excluded from the admin enumerations due to lower-quality admin address information.

Figure 7



Age distributions for Māori, Pacific, and Asian ethnic groups are shown in figures 8, 9, and 10 respectively. The markedly younger age profiles for Māori and Pacific people are reflected in the census data; however, the census counts are lower for adults from around 20 years of age to over 40 compared to the DSE. The expected Asian age distribution is also evident in the census data, but with some undercount for adults, centred around age 30. The census counts are very close to the DSE for children and older adults for all three of these ethnic groups.

Figure 8

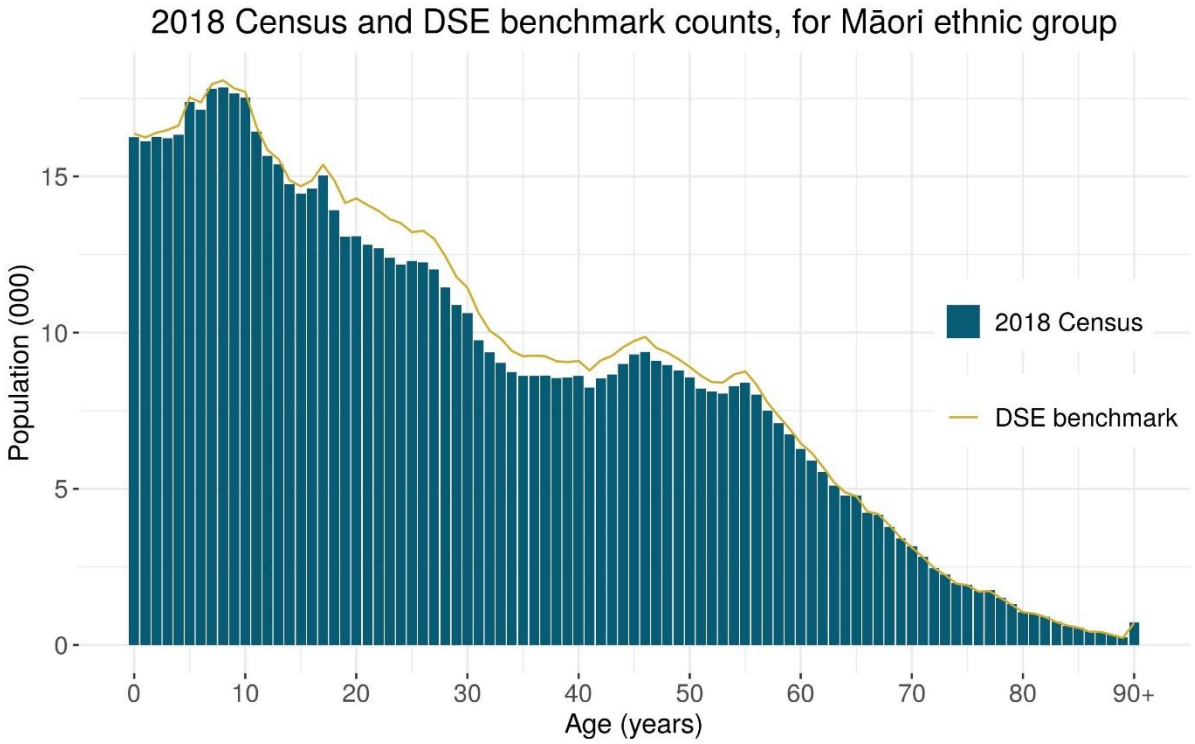


Figure 9

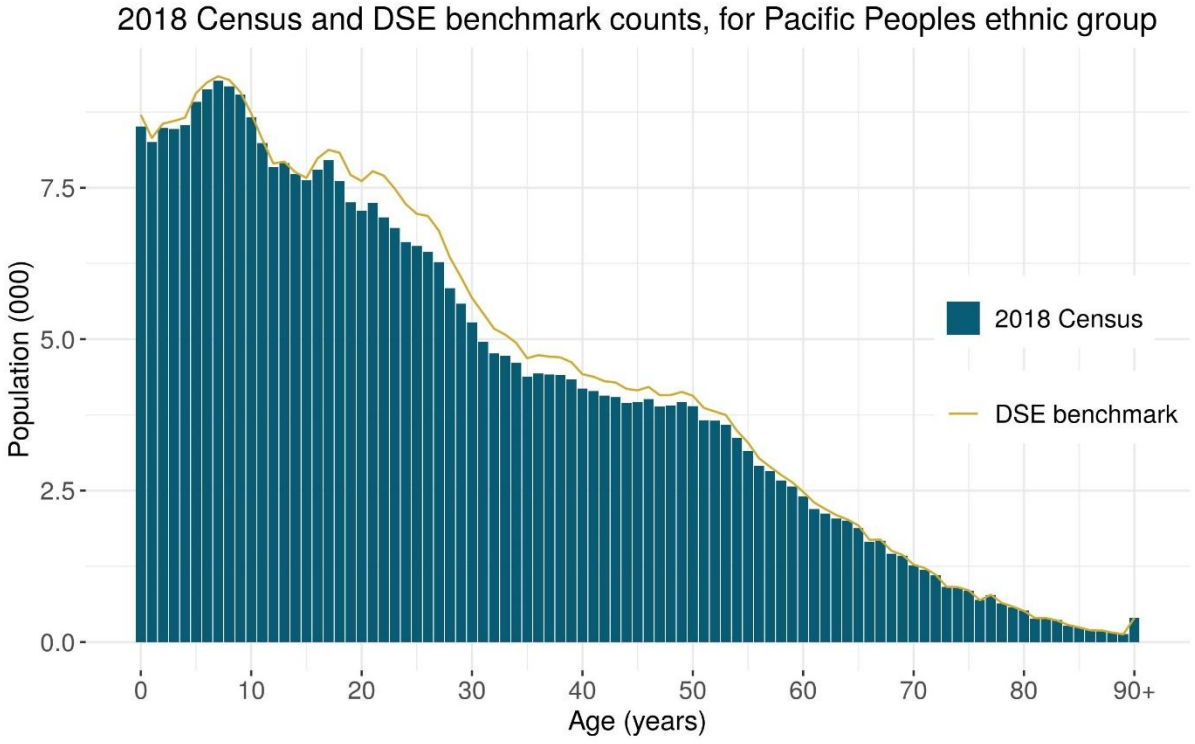
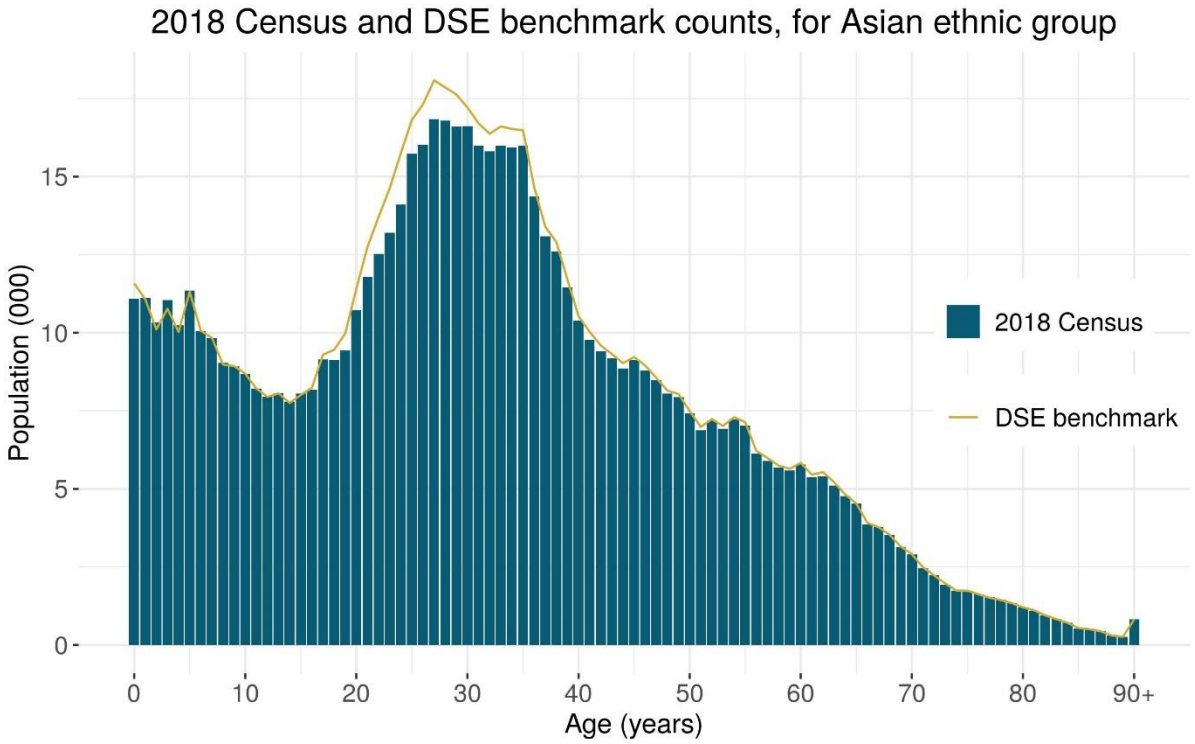


Figure 10



Relative difference provides a more precise comparative measure. Figures 11 and 12 show relative difference between the 2018 Census and the DSE benchmark by single year of age for females and males. There is almost no difference for children aged 0 to 17, and a small overcount for the census compared to the DSE for the oldest ages. Census undercount for younger adults is obvious, and reaches a peak of just over 4 percent for females, and 7.5 percent for males, around the mid-20s.

For the 2013 Census we have net census undercount as measured by the 2013 Post-enumeration survey. The DSE is the closest equivalent we have available at present for 2018. Figures 11 and 12 also include the relative undercount for the 2013 Census. Overall, these age/sex results for 2018 appear to be better than those achieved in 2013, especially for children. Results for females are better in 2018 at almost every age, though there is little difference between the two for ages 20 to 40. For males, the differences are somewhat greater for those aged between about 20 and 40 in 2018 than in 2013.

Figure 11

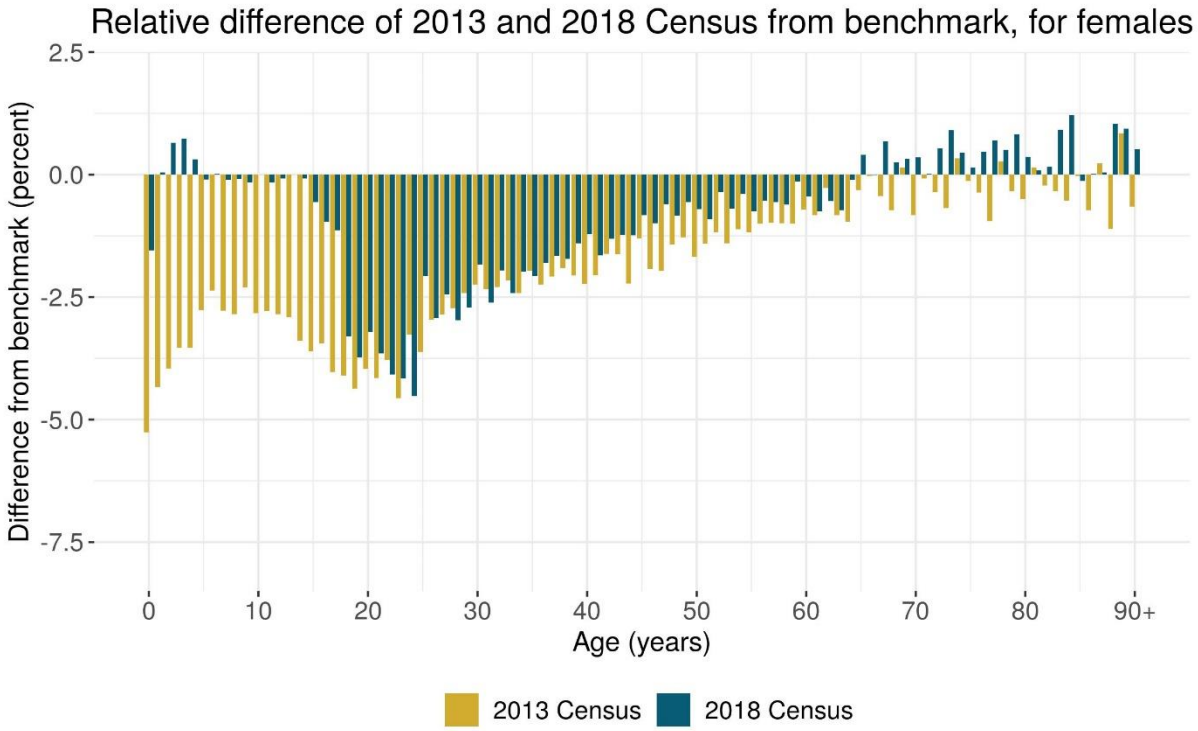


Figure 12

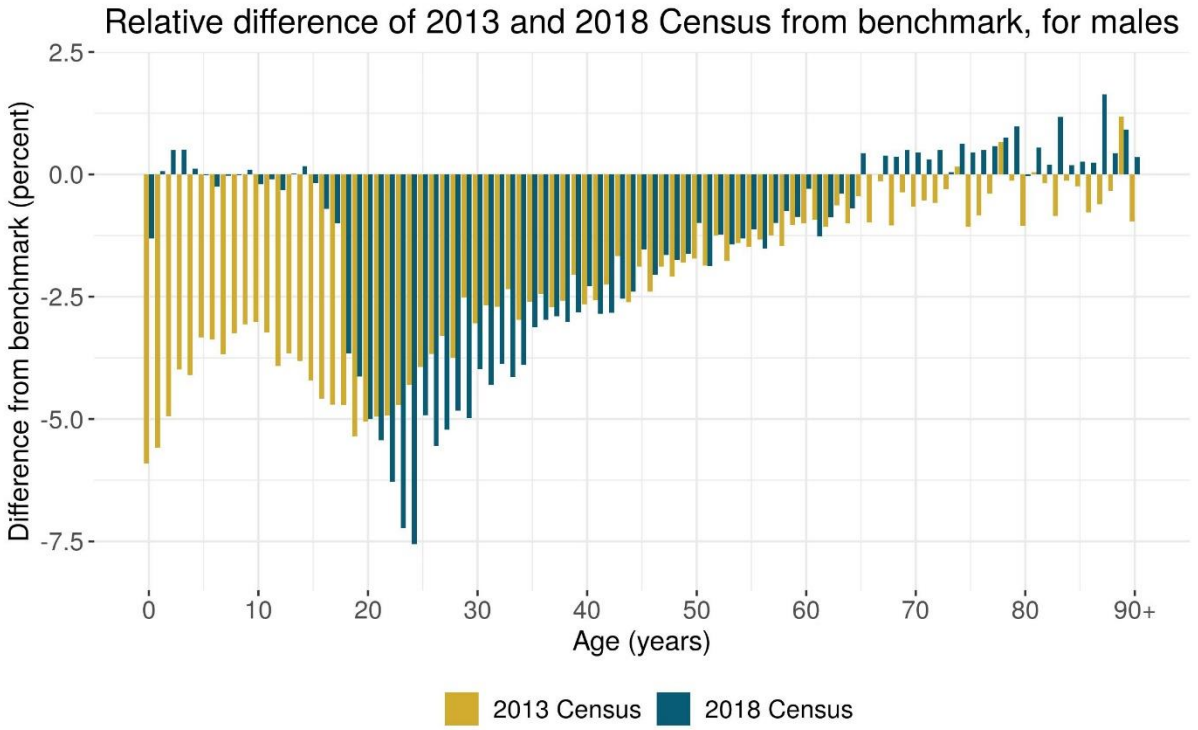
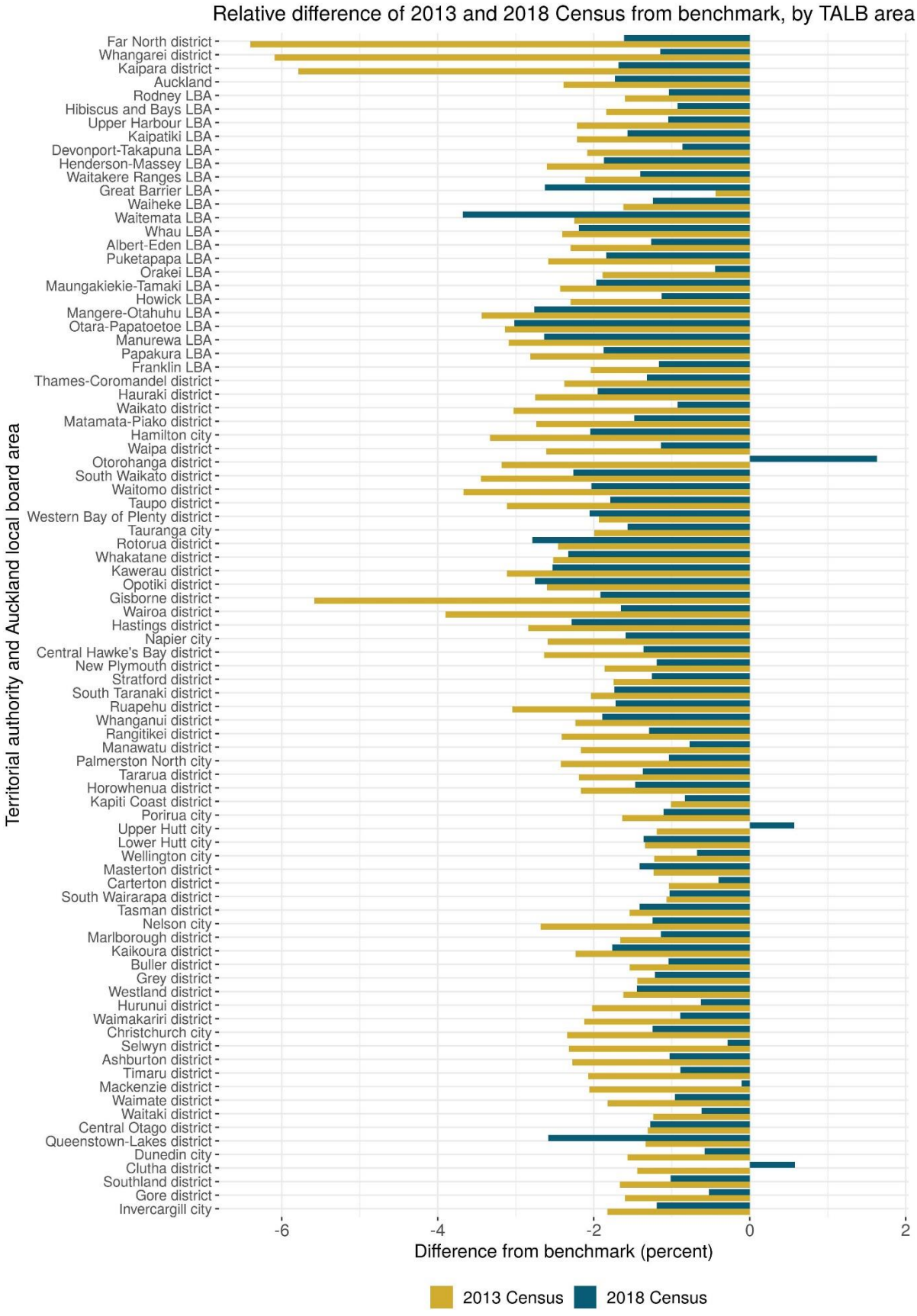


Figure 13 shows the relative differences for TALB geographic areas. In 2018 the relative difference between the census and the DSE benchmark is fairly uniform, between 1 percent and 3 percent lower for most TALBs.

Figure 13



Two exceptions are Queenstown–Lakes District, a high tourist destination, and Waitematā Local Board in central Auckland. Both are areas that tend to include more people with high mobility, for

whom it is less likely that we have an up-to-date admin address. Otherwise, 2018 results appear to be better than in 2013 across all other TALBs. In particular, 2018 results appear much better in the areas with highest undercount in 2013 (Far North, Whāngārei, and Kaipara districts in Northland; Gisborne; and Wairoa, all with census net undercount around 6 percent in 2013).

While we expect the DSE benchmark populations to provide a good comparison for the 2018 Census dataset, they are only indicative of actual coverage patterns. Variance for the DSE is very small. There may be some remaining bias, the direction of which is unclear. For example, the apparent census overcount in the oldest ages may be due to a small bias in the DSE when estimating very small populations. Official estimates of 2018 Census coverage rates will be available in March 2020, when the results of the post-enumeration survey are published.

Sensitivity analysis for key uses

A series of investigations have been completed, primarily by analysts independent of Stats NZ, to assess both the methods used to add people to the 2018 Census dataset and the fitness of the dataset for three important use cases: electoral boundaries, the New Zealand Deprivation Index (NZDep Index), and the district health board funding models.

Analysis of methods

A sensitivity analysis of the methods used to add people to the 2018 Census file was completed by Dot Loves Data. After looking at sources of variation, they found that the threshold for inclusion in meshblocks had the most impact on who was included in the census file. This parameter makes the trade-off between including the right people in the census dataset and the proportion placed confidently in the right meshblock. Their analysis suggested that a sensible balance had been achieved by setting this parameter at 0.5. The best predictor of whether a meshblock has a high level of sensitivity to the parameter settings is the level of deprivation.

Calculations of ethnic proportions are robust to the choice of parameters. Admin enumeration has increased the representation of Māori and Pacific populations, and the analysis found that the choice of cut-offs had minimal impact on the proportions of Māori, Pacific, and other ethnic groups in the census dataset.

Although they are not directly comparable, this analysis found similar results to the 2013 Census post-enumeration survey, in that similar population groups found to be undercounted in the 2013 census have higher sensitivity to the methods used in the 2018 Census for admin enumerations. These groups are consistently found to be harder to count, and include young adults, Māori, and Pacific people.

Electoral boundaries

Further sensitivity analysis undertaken by Dot Loves Data determined that 2018 Census data was robust for the purpose of determining electoral boundaries. Key findings were:

- The number of Māori electorates is the same, regardless of what threshold is chosen.
- The number of North Island general electorates would differ by 1 only if a further 35,000 or more admin records were to be eliminated from the final census file.
- Electorate boundaries drawn using census counts were not likely to be impacted by the choice of threshold for adding admin enumerations at the meshblock level.

Auckland Central was found to be more strongly affected by changes to the threshold than other electorates, partly because new migrants tend to arrive first in Auckland and are less likely to have an up-to-date admin address.

Deprivation Index

Initial analysis of the appropriateness of the 2018 Census dataset for producing an updated New Zealand Deprivation Index) found the census dataset sufficient for creating an updated index (Atkinson, J et al, 2014). Preliminary analyses indicated there were no variables used in the 2013 NZDep Index with an overall proportion of missing values that would make them unable to be used. This is due to the high use of dwelling form data in the index, and admin sourced and imputed values.

District health board funding

Sensitivity analysis was also undertaken for the Population Based Funding Formula (PBFF), which allocates over two-thirds of all health funding in New Zealand. This preliminary analysis by Stats NZ focused specifically on the allocation of funds to Māori and Pacific people.

Findings indicate notable differences in the geographical spread of Māori and Pacific populations between previous population projections for 2018 and the 2018 Census counts. All the census variables that the PBFF uses have a ‘high’ or ‘very high’ census quality rating and it is suspected that the differences in geographical spread of Māori and Pacific populations reflect real-world changes. We have concluded that the census counts would be appropriate for use as a base for the projections provided to the Ministry of Health.

Discussion and conclusion

The 2018 Census dataset consists of people counted through responses to the census field collection plus those counted through admin records. A coherent methodology has been developed for including admin records in the census dataset, based on an understanding of the limitations of the admin data. The inclusion of admin records has been conservative to avoid introducing overcoverage through incorrect admin records, with the consequence that some undercoverage remains in the final 2018 Census dataset. The admin enumerations bring real data about real people into the census dataset for those from whom we do not have a census response. It is clear from our results that the admin resident population does include many people who are typically hard to count through census field collection. The admin enumerations replace the use of unit imputation and make a significant quality improvement over previous census methods for non-response adjustment.

The new DSE benchmark population estimates have proved to be a valuable innovation, giving a sound basis for testing different options for non-response mitigation, and an indication of the coverage patterns in the final 2018 Census dataset.

Based on these results, and the sensitivity analyses for key uses, Stats NZ is now confident that it has compiled a census dataset that will provide census usually resident population counts and electoral counts of acceptable quality.

The value of the census questionnaire component is evident for census variables where no alternative sources are available. These variables consequently have higher levels of missing data in 2018, and those sub-groups with lower response rates to the field collection are more adversely affected. The full extent of these issues is the focus of ongoing work and Stats NZ will provide documentation about the results of its investigations into data quality.

This work was not planned as part of the build-up to the 2018 Census. New methods were needed and were developed under considerable time pressure to minimise delays in releasing census outputs. Consequently, it has not been possible to describe publicly the more extensive use of admin data and the new methods developed to use this data, before the census field collection, as would have been desirable.

Requirements for the effective use of admin data for the 2018 Census included the ability to safely access linked data sources, and statistical models developed for various estimation processes. Essentially the methods have been developed by breaking our problem into its constituent parts and applying the best available solution to each of these problems. The combination of each of these methods are unique but the elements of the solution are all drawn from standard statistical practice. The 2018 Census example is an intermediate step that combines field enumeration and admin data. However, it does not estimate the remaining missing data, nor provide uncertainty measures, which are produced through a coverage survey.

The approaches used have been verified to the extent that was possible in the constrained time frames. The placement of model cut-offs for determining suitable quality for inclusion of admin records involves subjective judgement, aiming for a balance between good household and small-area information, and improving demographic distributions. To some extent every Census involves subjective judgement about inclusion thresholds and we have done more than in previous censuses to test the sensitivity of the choices that have been made in compiling the 2018 Census dataset.

Our experience with the 2018 Census, where we faced unacceptably low response rates, has highlighted the value of undertaking the longer-term research into the use of admin data for the census. Even with the best preparation, a full field enumeration census runs a risk of catastrophic

failure for reasons such as natural disasters. While these may be low probability, impact can be high in some areas. Being well prepared to use admin data to supplement responses, if necessary, provides a welcome risk-management strategy.

The 2018 Census has provided us with greater insight into the properties of our administrative data, and has placed Stats NZ in a good position to further develop the role of administrative data in future censuses and population statistics.

.

References

Atkinson, J, Salmond, C, & Crampton, P (2014). NZDep2013 Index of Deprivation. Retrieved from www.otago.ac.nz/wellington/otago069936.pdf.

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.

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.

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

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

Fellegi, I, & Sunter, A (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Gath, M, & Bycroft, C (2018). The potential for linked admin data to provide household and family information. Retrieved from www.stats.govt.nz.

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

Green, B, IJpelaar, J, Marshal, D & Beatty, R (2014). Using an administrative primary care health activity indicator to address under-enumeration in the 2011 Census in Northern Ireland. *Journal of the Statistical and Social Inquiry Society of Ireland, XLIII* .

Keller, A, Mule, V, Morris, D, & Konicki, S (2018). A distance metric for modeling the quality of administrative records for use in the 2020 U.S. Census. *Journal of Official Statistics, 34*(3), 599–624, http://dx.doi.org/10.2478/JOS-2018-0029

Seber, G (1982). *The estimation of animal abundance and related parameters*. New York: Macmillan.

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

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

Stats NZ (2014b). Estimated resident population 2013: Data sources and methods. Retrieved from http://archive.stats.govt.nz.

Stats NZ (2014c). Understanding substitution and imputation in the 2013 Census. Retrieved from www.stats.govt.nz.

Stats NZ (2016). Experimental population estimates from linked administrative data: methods and results. Retrieved from http://archive.stats.govt.nz.

Stats NZ (2017). Experimental population estimates from linked admin data: 2017 release (PDF, 1.4MB). Retrieved from www.stats.govt.nz.

Stats NZ (2018). Experimental ethnic population estimates from linked administrative data (PDF, 1.1MB). Retrieved from www.stats.govt.nz.

Stats NZ (in press). *Data sources, editing, and imputation in the 2018 Census*. To be available from www.stats.govt.nz.

Stats NZ (2019a). Creating the 2018 Census dataset by combining admin data and census forms data: Our privacy impact assessment. Retrieved from www.stats.govt.nz.

Stats NZ (2019b). Creating the census dwelling frame for the 2018 Census. Retrieved from www.stats.govtg.nz

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

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

Stats NZ (2019e). Predicting the quality of admin location information for use in the 2018 Census. Retrieved from www.stats.govt.nz.

Appendix 1: Admin data sources and time lags

While census data provides point-in-time information that is accurate on the census date, a limitation of admin data is that delays mean that the most up-to-date information is not always available for the census date (in this case 6 March 2018).

The admin data available in the IDI for use in the 2018 Census is impacted by data lag, in terms of the delay between the point of collection by government agencies, the supply of that data to Stats NZ, and integration into the IDI. Agencies provide their data to Stats NZ on varying schedules, such as monthly or quarterly. Some datasets are supplied yearly.

The data held within the IDI is updated ('refreshed') with newly supplied information at set points throughout the year. The version of the IDI used for admin enumeration in the 2018 Census was the September 2018 refresh. Use of IDI data available as at September 2018 allowed up to five months after the March 2018 census date for datasets to be received from agencies.

Activity information, border movements, and death registrations identify those people who belong in the admin usual resident population, the IDI-ERP. Address information is used to identify a person's location, whether in a household at a specific address or in a meshblock. The availability of address datasets at census night does not imply that the address provided within the data is up to date. It is simply the latest known address for a person held by the agency at the time the data is supplied.

Although some datasets include information beyond the census date of 6 March 2018, only information collected prior to this date was used for activity and addresses. Migration border movements and death registrations are used up to the latest date they are available, in July 2018.

Table A1 shows the IDI data used to determine admin enumerations in the 2018 Census file and the most recent information available in the September 2018 refresh.

Table A1

Admin data lags for activity and address information in the September 2018 IDI refresh			
Activity data	**Data source**	**End of data coverage**	**Is data available at 6 March 2018?**
Birth registrations	Department of Internal Affairs	July 2018	Yes
Tax data – employer monthly schedule	Inland Revenue	August 2018	Yes
IDI summary table – income tax year summary	Inland Revenue	March 2018	Yes
ACC injury claims	Accident Compensation Corporation	August 2018	Yes
Primary and secondary school enrolments	Ministry of Education	August 2018	Yes
Tertiary education data	Ministry of Education	December 2017	No
Industry training and apprentices	Ministry of Education	January 2018	No
General Medical Subsidies claims	Ministry of Health	June 2017	No
Laboratory claims	Ministry of Health	June 2017	No
National Non-Admitted Patient Collection data	Ministry of Health	June 2017	No
Pharmaceutical data	Ministry of Health	June 2017	No
Primary health organisations	Ministry of Health	November 2017	No
Publicly funded hospital discharges	Ministry of Health	December 2017	No
Death registrations	Department of Internal Affairs	July 2018	Yes
Border movements	Ministry of Business, Innovation and Employment	July 2018	Yes
International travel and migration data	New Zealand Customs Service, Stats NZ	July 2018	Yes
Address data	**Data source**	**End of coverage**	**Data available at 6 March 2018?**
2013 Census	Stats NZ	March 2013	No
Driver licence and motor vehicle registers	NZ Transport Agency	August 2018	Yes
ACC claims	Accident Compensation Corporation	August 2018	Yes
Primary and secondary school enrolments	Ministry of Education	August 2018	Yes
Tax data	Inland Revenue	July 2018	Yes
Primary health organisations	Ministry of Health	June 2018	Yes
National Health Index	Ministry of Health	July 2018	Yes
Benefits and social services	Ministry of Social Development	July 2017	No
Household and Labour Force Survey	Stats NZ	June 2018	Yes
Household Economic Survey	Stats NZ	January 2018	No
Auckland City Mission data	Auckland City Mission	January 2018	No

Source: Stats NZ

Appendix 2: Admin enumerations in responding dwellings

The census living arrangements question is a multi-response question that asks: 'Who lives with you at your usual address?' (see Living Arrangements – New Zealand Standard Classification 1999V1.2.0).

We use the answer to the living arrangements question conjunction with the relationship to reference person on the household form to determine the relationships between everyone in the household.

Our examination of responses to the 2018 Census living arrangements question indicated some instances where people reported living with individuals who were not included in the census usually resident household. We used this evidence to infer that one or more members of the usually resident household were non-responding. We then used relationship information available in the IDI to identify these missing household members and added them to the household. People were only added to responding households if admin data was available to confirm the missing relationship identified via the census living arrangements question.

We first identified the responding census households where individuals' stated living arrangements indicated that there were more people in the household than currently included. For example, if a person indicated that they lived with a partner and child, then we would expect that household to include at least three people. There are a range of situations used to infer a non-responding household member:

- We can reasonably expect that there is one person missing from the household because the living arrangement gap is for a partner.
- We can reasonably expect that there are multiple people missing from the household because there is more than one living arrangement category with a gap.
- We can reasonably expect that there is at least one person missing from the household due to a living arrangement gap, but the number of people missing is unknown (eg when people report living with a child, we do not know how many children this refers to).

We set a rule to only identify households as containing non-responding members where at least 30 percent of the household members indicated extra people. This means for example in a household with three responses, one of those would need to identify extra people.

Once we had identified households that contained non-responding members, we used admin data to add people to these households if there was evidence of the missing relationship from admin sources.

Within the IDI, relationship information is available from a number of sources (Gath & Bycroft, 2018). For people in the IDI-ERP, we have the potential to identify their partner, their parents, and their children. From this information we can deduce the relationships of parent, child, partner, in-laws, siblings, grandparents, and grandchildren, if they are in the same household. However, many people in the IDI-ERP have minimal or no relationship information in the IDI.

We used a set of rules to add people from the IDI-ERP (who had not already been counted in the 2018 Census) into responding households using admin data. We used the census living arrangements gap together with the admin relationships data to identify a person who is:

1. at the same address as the census household
2. related to a census respondent in that household
3. confirmed by admin data to have a relationship that aligns with the missing living arrangement relationship.

It was possible to verify the third point most of the time. However, in the few cases where there were several people in the household that indicated multiple missing living arrangements, we have not been able to verify this final point.

Using these rules, we added 20,643 people into 13,533 dwellings. As a result, approximately 97 percent of responding households with people added through this method ended up with a complete family relationship grid. The vast majority (95 percent) of dwellings had between one and three people added. Table A2 provides a comparison of household size based only on responding members with the final household size following the addition of people through admin enumeration based on the methods outlined above.

Table A2

Household size before and after admin enumeration of people to responding dwellings

Household size responding members only	Household size following admin enumeration									
	2	3	4	5	6	7	8	9	10	11
1	6,948	1,530	864	333	147	39	24	6	0	0
2	0	2,019	759	228	72	21	6	6	0	0
3	0	0	330	93	33	9	9	0	0	0
4	0	0	0	33	12	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0

Source: Stats NZ

Appendix 3: Non-private dwellings – prisons and defence

We describe here in more detail how administrative records were used in the 2018 Census dataset for two non-private dwelling types: prisons and penal institutions, and defence establishments.

The Department of Corrections has provided data to Stats NZ containing unnamed unit records for every individual in each prison or penal institution on census night, 6 March 2018. This file also contained a Ministry of Justice identifier that enabled these agency-supplied records to be linked to the Corrections data already provided in the IDI.

The Ministry of Defence similarly provided a dataset of unnamed individuals in defence establishments on census night, 6 March 2018. This file did not contain any admin identifiers that enabled linking to the IDI.

Variables provided by the agencies are the location of the establishment, and the age, date of birth, sex, and ethnicity for each individual at that establishment on census night. For prisons, the agency-provided file is linked to Corrections data already in the IDI, so that characteristics can be completed from the person's 2013 Census response or admin sources where possible. For defence establishments, admin record enumerations will be added to the census file with demographic characteristics as provided by the agency.

Not everyone present in these dwellings are usual residents. Since the agency provides only census-night location, we impute a usual residence for an appropriate proportion of the admin records added to the census dataset.

Iwi was also provided, but not used for the census dataset due to quality concerns. Other characteristics not available from alternative sources will be either imputed or missing, depending on the variable.

Method

Prisons and penal institutions

For prisons, we have three sources of information:

- census responses for some of the occupants on census night
- a list of all occupants on census night provided by the Department of Corrections
- Corrections data in the IDI.

Information about which prison or penal institution a person was at on census night was not available from the IDI Corrections data, so the agency-provided file was the only source of the census-night dwelling location.

Our aim was to supplement the census responses with additional people as indicated by the agency-provided admin records. The method used was as follows:

- Link agency-provided records to the IDI Corrections data based on the shared Ministry of Justice identifiers. This is done via a concordance between encrypted identifiers in the IDI and unencrypted identifiers provided in the agency file.

- Remove records if they have already been linked to the 2018 Census (whether recorded at a prison or otherwise), or have been included in the census as other forms of admin enumeration.

- Remove records that correspond to missed links between the census and IDI spine.

    - Take census prison responses that were not linked to the IDI spine.
    - Take agency-provided records that were not linked to the 2018 Census.
    - Where we get an exact match on location and date of birth, remove the record from the agency file.

- Create census admin records for the remaining individuals for each institution.

    - Add all the relevant characteristics information from the 2013 Census and admin sources where available.
    - Otherwise use ethnicity from the agency-provided file.
    - If appropriate, any missing characteristic information is imputed.

Note that the final count for each institution may not exactly match that provided by the Department of Corrections because responses for some people have been received from other locations.

Defence establishments

For defence establishments we have two sources of information for each dwelling:

- census responses for some of the occupants on census night

- a list of all occupants on census night provided by the agency.

Our aim is to supplement the census responses with as many admin records as are needed to reach the total number of occupants for each dwelling as indicated by the agency. The method used is as follows:

- Determine the number of admin records required to reach the agency total occupancy count for each dwelling. This is simply the admin total count minus total census responses, calculated for each dwelling.

- Conduct a consolidation exercise between the list of admin occupants and the census responses for each establishment. This is done via a comparison of age, sex, date of birth, and dwelling to find exact matches between a census response and an admin occupant, though not all census responses are able to be matched. This comparison determines those admin records where we clearly have a census response. The remaining admin records are eligible for inclusion in the census file.

- New census records are created from the list of admin occupants for whom there is no matched census response. Since not all census records have been matched, a random selection of the required number of the non-responding admin occupants is selected for inclusion in the census file.

Results

Table A3 provides the number of census responses received, the number of census responses matched to the agency-provided files, and the number of admin records added following the methods outlined above. Overall 4,707 admin enumerations are added for prisons or penal institutions, making up just less than half the final total census count. Of these admin enumerations, most have used the IDI linkages to obtain relevant characteristics information from the 2013 Census and admin sources, with just 99 reliant on characteristics provided by the Department of Corrections.

For defence establishments 798, or one-quarter of the 3,150 total census count, are admin enumerations. All of these defence admin enumerations use the demographic characteristics as supplied by the Ministry of Defence.

Table A3

Census responses and admin enumerations in prisons or penal institutions and defence establishments

Non-private dwelling subtype	Agency total	Census responses	Census matches[1]	Admin enumerations	Final census total	Expected (pre-census)
Prison or penal institution	10,700	4,917	4,000	4,707	9,624	~10,000
Defence establishment	1,800	2,352	970	798	3,150	~3,200

1. Records in agency-provided files that could be linked to a census response

2. Includes vessels excluded from scope of Ministry of Defence-supplied data

Source: Stats NZ

Records from the file provided by Corrections could be linked to:

- 4,000 census respondents with a census-night location of a prison or penal institution
- 1,100 census respondents with a census-night location that was not a prison or penal institution
- 50 admin enumerations from other mitigations.[4]

To avoid including individuals twice, we do not include any of these agency records linked to a census form as prison admin enumerations. We also do not change the census-night location for any of these records, which leads to some apparent undercount in the 2018 census-night population count for prison and penal institutions (they are still included in the census file, but at a different dwelling type).

Census forms were received from a further 1,000 people reported as being in a prison on census night but were not linked to the Corrections file. These were approximately balanced by agency-

[4] Due to a change from the initial methodology, prison enumeration occurred after all other forms of admin enumeration, so some people later determined to be in prison on census night had already been admin enumerated through an alternative mitigation.

provided records that were not linked to the 2018 Census but were not included as admin enumerations. Of these, 500 have the same census-night prison location and date of birth as a census respondent, providing evidence that they were already included as census respondents. An additional 300 records were unable to be linked to the IDI and were not added to the census file.

Usual residence location imputation

The admin data sources provided a list of those present on census night but did not indicate whether they usually live at their census-night address, or usually live elsewhere. While we do not know what criteria are used by census respondents, in general usual residence is defined as living, or intending to live, somewhere for 12 months or more. Admin enumerated individuals were assigned a usual residence indicator to achieve the same overall proportions as observed in the 2013 Census. There is an assumption that the proportion of usual residents observed in the 2013 Census would be more accurate than that of the 2018 Census respondents.

For prisons, in 2013, 52 percent of prisoners stated that they usually lived within the prison. This resulted in just over half of the admin enumerations being allocated a usual residence location within the prison.

For defence establishments, 2013 Census responses indicated 61 percent of the population usually resided within the dwelling. To maintain this proportion, all admin enumerated defence establishment occupants had their usual residence location assigned to the census-night dwelling.

Appendix 4: Adjusting for missed matches

Administrative records are added to the 2018 Census dataset as a mitigation for non-response to the census field collection. Linking the 2018 Census data to the IDI spine determines which admin records are those who have already responded to the census, and identifies those who have not responded. Our key concern here is when people in the IDI spine who have **not** been linked to a census response are candidates for inclusion in the census dataset. If we have not linked a census record to the IDI spine when we should have, the same person could be included in the final census dataset twice.

First, we require an estimate of the rate of missed matches. Details are provided in Stats NZ (2019d). Briefly, missed matches are estimated from a subset of the census responses that, on the basis of certain questions, are highly likely to be in the IDI spine. Respondents in this subgroup that were not in fact linked to an IDI spine record are very likely to be missed matches, and enable us to estimate the false negative match rate. The IDI spine is made up of people who were born in NZ, have entered NZ as a migrant on a visa, or are registered to pay tax. Consequently we formed the subset of census respondents highly likely to be in the IDI spine on the basis of responses to census questions that closely coincide with these criteria. We obtained an overall estimate of 1.21 percent who should have been linked to the IDI spine, but were not. Missed link rates are estimated for fine demographic breakdowns by five-year age group, sex, TALB, and mutually exclusive level 1 ethnic groups.

We assume that these probabilities for missed links between census responses and the IDI spine are a reasonable approximation when applied to adjusting for missed links between the census responses and IDI-ERP_Sure records.

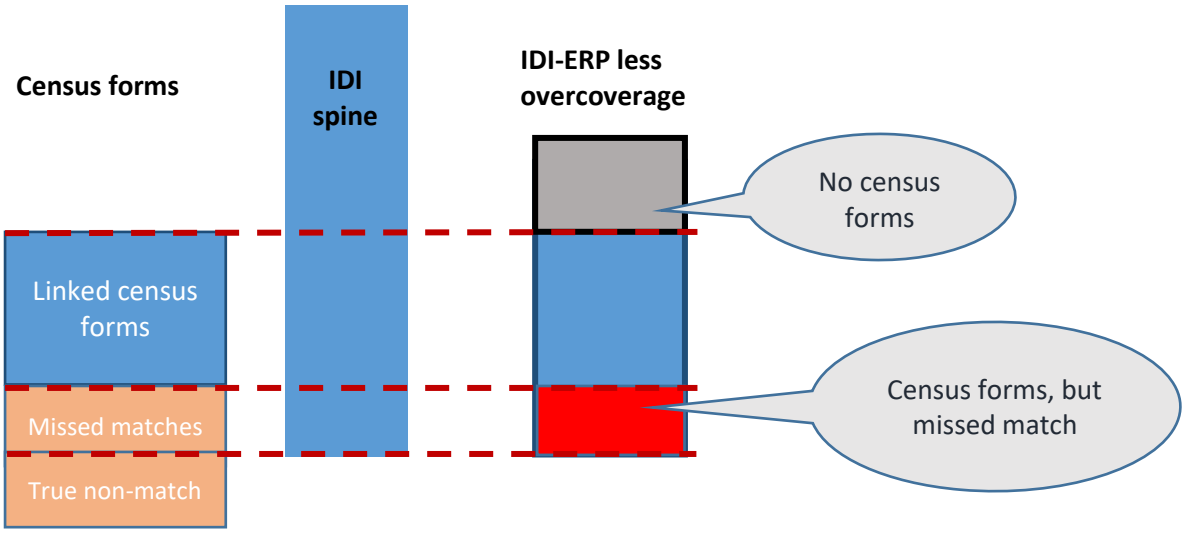
Figure 14 illustrates these linkages. The eligible admin records are people in the IDI-ERP_Sure who have not already been included in the census, which we observe through the link made to the census. The IDI-ERP_Sure consists of people who were linked to the census forms (shown in blue), and the remainder of the IDI-ERP_Sure, who are a combination of two groups:

- people who did not respond to the census (grey), and
- those who did respond to the census but where the match was missed (orange).

We estimate the number of missed matches in the IDI-ERP_Sure, shown as the red box in figure A4.1.

Figure A4.1

Missed matches between the 2018 Census and admin records



Adjusting for missed matches

More formally, we have an estimate of the probability of being linked, or not linked, given that the census form is a true match to an admin record. Let

$$\Pr(ul|m) = \Pr(unlinked \mid match)$$
$$\Pr(l|m) = \Pr(linked \mid match)$$

The results of the linkage between the census forms and the IDI-ERP_Sure can be represented as a cross-tabulation, as shown in table A4.1.

Table A4.1

Cross-tabulation of linkage between census forms and IDI-ERP_Sure				
		Admin population: IDI-ERP_Sure		Total
		1	0	
Census responses	1	n_{11}	n_{10}	N_{1+}
	0	n_{01}		
Total		N_{+1}		

N_{1+} is the number of census responses.
N_{+1} is the number of records in the admin IDI-ERP_Sure admin population.
n_{11} are the linked records, that is, in both the census and the admin population.
$n_{10} = N_{1+} - n_{11}$ are records in the census 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.

Incorrectly missed matches means that the observed n_{11} cell is too small, and the pairs of records that should have been matched are still in the n_{01} and n_{10} cells.

The observed links, n_{11} , are the true number of matched records multiplied by the probability of being linked, given a true match. We can estimate the true matched records, $\hat{n}_{11}$, by:

$$\hat{n}_{11} = n_{11} / \Pr(l|m).$$

Or more conveniently in an additive form:

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

where $\eta = \Pr(ul|m) / \Pr(l|m)$ is the ratio of probabilities conditional on the true match status.

This provides us with the number of missed matches to remove from the observed unlinked admin records, the n_{01}cell.

Our adjusted estimate of n_{01}, $\hat{n}_{01}$ is

$$\hat{n}_{01} = n_{01} - n_{11} * \eta$$

We calculate η and apply the adjustment for breakdowns by demographic groups and TALB subnational geographies. The appropriate number of people are removed through random selection within subgroups.

This calculation makes no allowance for false positive matches. These incorrect links inflate the number in the n_{11} cell and bias the adjustment towards removing too many records.

Results of missed linkage adjustment

Table A4.2 shows results for the link between 2018 Census responses and the IDI-ERP_Sure. In total, 3,662,000 links were made, representing 89 percent of the 2018 Census responses and 86 percent of the IDI-ERP_Sure.

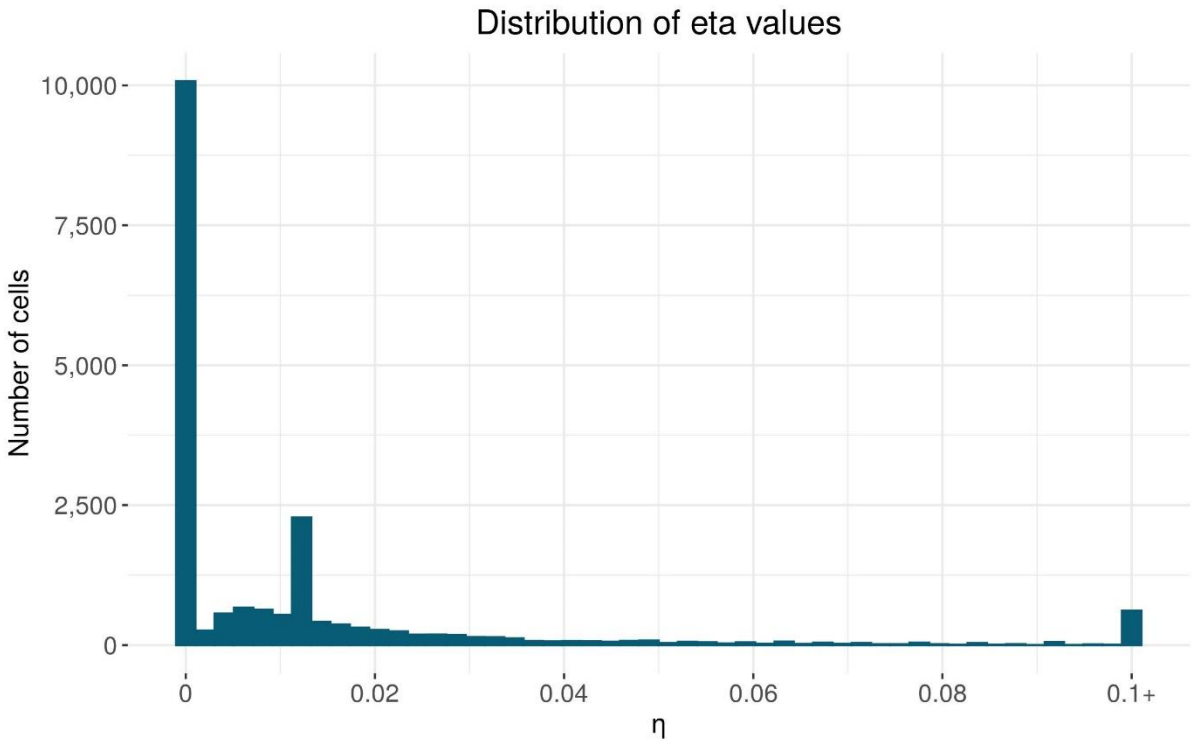
Table A4.2

Census responses linked to the IDI-ERP_Sure admin population				
		IDI-ERP_Sure		
		1	0	Total
Census	1	3,662,000	472,500	4,134,500
	0	610,000		
	Total	4,272,000		

The overall estimate of missed matches of 1.21 percent (that is, the proportion of census responses that should have been linked to the IDI spine, but were not) translates to an overall ratio, η, of 0.0122.

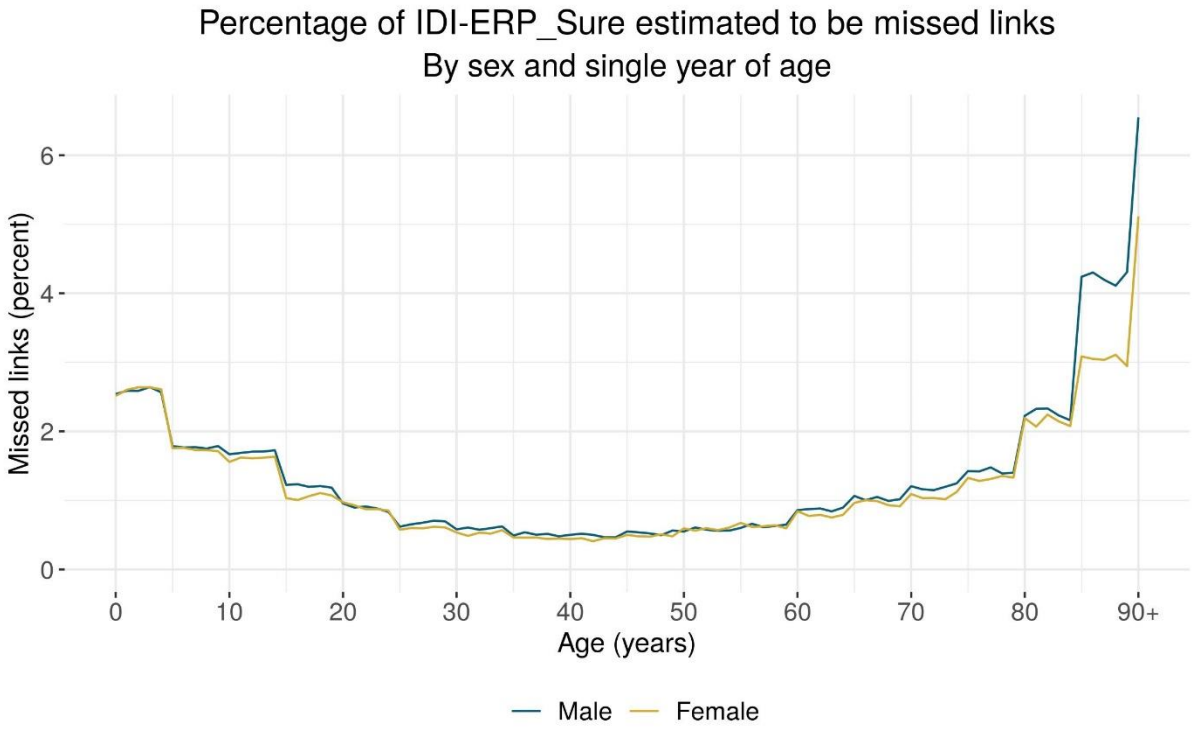
We estimated η for each combination of territorial authority or Auckland local board area (TALB), sex, five-year age group, and ethnicity (using indicators for Māori, Pacific, and Asian ethnicities) – a total of 26,448 cells. Figure A4.2 shows the distribution of η values. Note that 7,274 cells had zero records from either input and are not included. Cells with no census responses available to calculate the missed match rate were assigned η calculated from the overall population (0.0122), resulting in the larger number of cells with η between 0.012 and 0.014. Almost half of the cells have η values of 0, indicating all census responses used to calculate the missed match rate were linked to IDI-ERP_Sure, although only 5 percent of the population was included in these cells.

Figure A4.2



The number of missed links were then estimated by single year of age, applying the same η within each five-year group. We apply these rates at the single year of age level to ensure consistent numbers of links are removed, and to reduce the impact of differences in the population counts across each age. In total, 51,500 records were estimated to be missed links, meaning they should be removed from the n_{01} cell. Figure A4.3 shows the estimated percentage of IDI-ERP_Sure that is estimated to be missed links by single year of age and sex. The rate of incorrectly missed links is higher for children, at around 2 percent, and for those aged over 80 where it ranges from 2 percent to 6 percent.

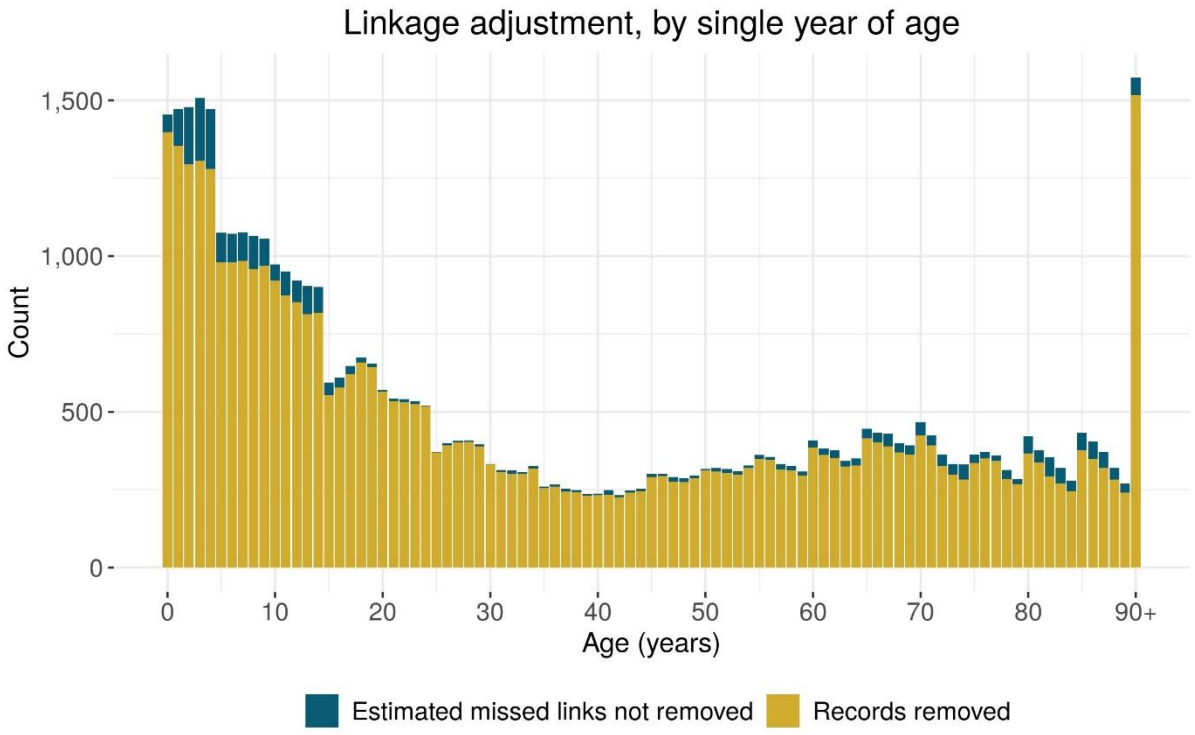
Figure A4.3



We assumed that no records already included as admin enumerations through the household-based mitigations were missed links, leaving 472,578 IDI-ERP_Sure records. The records to be removed were sampled from within each subgroup. In a small number of cases, there were fewer IDI-ERP_Sure records available than were estimated to be missed links. This could suggest the estimated number of missed links was too high for these groups, or that some false negatives were included in earlier admin enumerations. However, given the large number of cells, and the small differences observed, our estimates appear to be reasonable.

In total, 47,670 records were removed, slightly fewer than the estimated number of missed links. However, the differences are small and spread across all ages, and are considered unlikely to significantly impact the final population counts. Figure A4.4 shows the estimated number of missed links, and the number of sampled records that were removed, by single year of age.

Figure A4.4


Linkage adjustment, by single year of age
1,500
1,000
500
0
Count
0
10
20
30
40
50
60
70
80
90+
Age (years)
Estimated missed links not removed
Records removed