Stats NZ
Tatauranga Aotearoa

Data sources, editing, and imputation in the 2018 Census

New Zealand Government

Citation
Stats NZ (2019). *Data sources, editing, and imputation in the 2018 Census*. Retrieved from www.stats.govt.nz.

ISBN 978-1-98-858371-6 (online)

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

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

www.stats.govt.nz

Contents

List of tables and figures

List of tables

List of figures

Purpose and summary

Purpose

Data sources, editing, and imputation in the 2018 Census describes the 2018 Census approach for detecting data errors and for filling in gaps when the characteristics of people or dwellings have not been provided on census forms. We also report on the quality and results of our approach.

Summary of key points

Stats NZ applies edits to detect errors and uses statistical methods to deal with missing census data (non-response) to improve the quality of published census information. These online edits have greatly improved data quality captured from online forms compared with paper forms.

For the first time in a New Zealand census, the 2018 Census used data from alternative sources to fill gaps when the characteristics of people or dwellings have not been provided on census forms. These alternative sources were the previous census in 2013 and a range of administrative (admin) data sources such as birth registrations and tax information. Where high-quality alternatives are available, these provide real information about a person, but given at a different time or different context from the 2018 Census.

The use of statistical imputation for remaining missing data has been extended to a much greater range of variables than in previous censuses.

While both these improvements were planned in the build-up to the 2018 Census with the aim of reducing bias caused by non-response, they have taken on much greater significance in light of the lower than expected response rate to the 2018 census field collection, and the use of administrative records to count people who were missed.

People listed as a member of a household but with no individual form, and people counted through admin enumeration rely on alternative sources and imputation as the source of nearly all individual census characteristics.

The use of admin enumerations has improved the census count over previous censuses for some subgroups of the population, and together with the alternative data sources and imputation, has for many variables maintained (or in some cases improved) the quality of information seen in previous censuses. However, for other variables, high rates of imputation or missing data mean that quality is lower than in previous censuses, and those subgroups with lower response rates to the field collection, such as Māori, Pacific, and young adults are more adversely affected. There may be breaks in the time series due to improvement from previous censuses, or from higher levels of missing data.

Users of 2018 Census data will need to consult the detailed information provided about the data sources and quality ratings when considering the fitness for purpose of the information they wish to use.

Introduction to data sources, editing, and imputation

Background

The New Zealand Census of Population and Dwellings is the official count of how many people and dwellings there are in New Zealand. It provides a snapshot of our society at a point in time and helps to tell the story of its social and economic change.

The New Zealand Government agreed to a Census Transformation Strategy in 2012, with a short- to medium-term focus on modernising the current census model and a longer-term focus on investigating alternative ways of producing small-area population and social and economic statistics.

The 2018 Census strategy (Stats NZ, 2016) sets out an ambitious modernisation programme across all components of census taking. This followed several censuses of minimal content change, and limited innovation. For the first time, the collection was designed to be predominantly online, with paper forms in a supporting role. The strategy also included a goal to increase the use of administrative (admin) data, mainly through an address frame based on admin sources to support census collection, but also to improve data quality. As part of the longer-term research, Stats NZ's Census Transformation programme investigated alternative data sources for many census variables, providing recommendations for the 2018 Census about where admin data could be used to supplement census responses.

While the census aims to collect information directly from all people in New Zealand on census night, it is usual for some people not to fill in census forms, or to submit forms with some questions unanswered. Non-response introduces a bias to the extent that non-respondents differ from respondents for the characteristic of interest. Typically, census respondents are found to differ in important ways from non-respondents. Missing data results in counts lower than in reality and proportions biased towards categories that tend to receive more responses. Stats NZ, in common with other countries, uses statistical processes to compensate for missing data. These processes are designed to reduce the bias caused by differences between the people who do respond and those who do not.

Some aspects of the 2018 modernisation were successful, for example over 80 percent of all received census forms were completed online, and the new approach to compiling the census dwelling frame provided high quality census dwelling counts. However major challenges were faced when implementing the new collection model. The response rate was lower than expected, and more people were listed only as a member of a household but did not return an individual form. New methods based on admin data were developed to compensate. For the first time, the 2018 Census dataset uses admin records to include people who were missed by the census field collection, replacing the use of 'substitute' imputed records in previous censuses. We refer to these records as **admin enumerations**. Methods for admin enumeration are detailed in Overview of statistical methods for adding admin records to the 2018 Census dataset (Stats NZ, 2019a). The final 2018 Census dataset consists of 89 percent counted from census forms, however 4 percent of these records were listed as a member of a household but had no individual form. The remaining 11 percent were counted as admin enumerations.

The census still does not include everyone, and the post-enumeration survey (PES) is run after the census to measure census coverage. The 2013 PES found that net census undercount was 2.4 ± 0.5 percent, with higher rates for young adults, males, and people of Māori and Pacific ethnicity (Stats NZ, 2013). The 2018 PES results are not available until March 2020, however, we

have an indicative estimate of 1.4 percent undercount for the 2018 Census (see Customer update on data quality of 2018 Census).

We define **item non-response** as the situation where a person (or dwelling) is counted in the census, but information about the characteristics of the person or dwelling is missing. Item non-response typically refers to the situation where a survey form was received, but some questions were not answered. Here we also include characteristics information that is missing for people who had been counted in the census through a household listing but had no individual form, or through admin enumerations. For those counted through a household listing, age, sex, census night address, and relationship to reference person are requested for all people present at the dwelling on census night, and information on all remaining characteristics is initially missing. For admin enumerations, there is no census form and all characteristics are initially missing.

Table 1 describes the types of records making up the final census dataset, and the item non-response before there is any mitigation for missing data. Records with no individual form make up 15 percent of the census dataset and rely on alternative data sources and imputation as the source of nearly all individual census characteristics.

Table 1

Description of record types in the census dataset before mitigation of item non-response		
Type of record	**Percent of final census file**	**Description**
Individual form received	85%	• Individual form received for the person • Item non-response when some questions missing
Household listing	4%	• Person listed on online Household set-up form or paper Dwelling form, but no Individual form received • Nearly all individual census characteristics are initially missing.
Admin enumeration	11%	• Person counted through admin enumeration • No individual form received, and not listed on a Household set-up or Dwelling form • All individual census characteristics are initially missing.

Planned changes for the 2018 Census

In line with the 2018 Census strategy, methods were developed to make use of the previous 2013 Census data and admin data for item non-response, and to extend the use of statistical imputation methods to a wider range of variables than in previous censuses. The aim was to improve census data quality by reducing the bias caused by missing data.

2013 Census is data collected in the previous census and is used for variables that do not change much over the five years between censuses.

Admin data is data collected by government or other organisations for non-statistical reasons, such as births, tax, health, and education records.

Imputation is a statistical procedure for entering a value for a specific data item where the response is missing or unusable (OECD, 2013).

When we use the 2013 Census and admin sources, these are linked to the 2018 Census and we are obtaining information for the same individual but at a different time or in a different context. In contrast, imputation is a statistical process that provides an estimated value based on known characteristics of the person, and the pattern of responses from similar people.

Imputation in previous censuses was limited to four variables: age, sex, usual residence and labour force status (Stats NZ, 2014a; note that Māori descent was also imputed for electoral purposes only). The level of item non-response for the remaining variables in the 2013 Census was typically between 5 percent and 10 percent, of which around 5 percent was due to substitute records. The rates of item non-response were expected to reduce in 2018 with greater uptake of the online form. However, the lower than expected response rate for the 2018 Census has meant that the new methods for managing item non-response have taken on greater significance than anticipated. People counted through admin enumeration (11 percent of the final census dataset) are missing values for all attributes, and people counted through a household listing (4 percent of the final census dataset) are missing values for the vast majority of attribute information.

A new processing system was introduced for the 2018 Census that increased automation and reduced manual intervention, and ensured that a consistent, repeatable process was in place. Editing and imputation were part of the processing system, and a modular system was designed to ensure good sequencing between edits and imputation. The aims of the processing system were to provide both efficiency savings and improved data quality. See Processing and evaluating the quality of 2018 Census data (Stats NZ, 2019b) for more information on the processing system.

Aim and scope

This paper focuses on the data and methods used to deal with item non-response (missing characteristics) in the 2018 Census. We also describe the approach to editing to detect invalid or inconsistent responses.

The scope of this paper is all variables describing the characteristics of individuals and dwellings asked on the 2018 census forms, with the exception of current usual residence. The derivation of usual residence address and meshblock from admin data is fully described in Stats NZ (2019a). We also exclude derivation of household and family variables. We note that ‘Don’t know’ (for Māori descent) and ‘Object to answering’ (for Number of children born, and Religion) are valid responses to these questions and are not treated as missing or unusable data in the census.

The new methods for providing values for missing data and the scale of their use represent a significant change in the sources of information for attribute variables from previous censuses. It will also affect time series comparisons between 2018 Census and previous censuses. A summary of data sources and quality ratings for each variable are provided with published outputs. See 2018 Census information by variable and quality in DataInfo+. This paper provides more detailed information on the methodological approach for users of census data. An accompanying document provides further technical details about the methods used to derive census information from admin data sources for each variable (see Summary of admin data used in the 2018 Census dataset).

First, we summarise the editing approach for detecting and resolving errors in the 2018 Census. The following sections describe the alternative sources used for each variable and explain the imputation method. We describe the assessment of quality for each of the data sources. The outcomes are summarised across all census variables in the Results section.

Editing in the 2018 Census

Editing detects and resolves errors to improve the quality of data. Errors can come from respondents, such as marking multiple responses to a single response question, or can be introduced through the capture and processing of data, such as when accidental marks are mistaken for real responses during the scanning of paper forms.

As a largely paper-based and respondent completed survey, editing has been an important part of the census for many years. Editing is closely linked to other processing tasks, such as coding and imputation, and is dependent upon questionnaire design and survey mode. The philosophy of editing in the New Zealand census has changed over time. The 1996 Census involved many complex edits aiming for fully consistent output. The 2001 Census saw a shift to respecting respondents' intentions, and reduced intervention through **micro-edits** applied to individual records. A **macro-edit** stage checked aggregated data for systematic errors. While clearly erroneous data was removed, some apparent inconsistencies were left in the final data. This approach was continued in the 2006 and 2013 censuses.

The 2018 approach to editing was similar to the previous three censuses but aimed to automatically resolve as many erroneous responses as possible and to reduce manual intervention during processing. The online form used edits at point-of-capture which avoids many of the errors found on paper forms.

Types of edits

Edits are designed to detect common types of errors. The following are the types of edits used in the 2018 Census:

- Capture edits – identify respondent errors such as marking multiple responses to a single response question. Where possible, these are corrected to a single response. For example, providing multiple responses to the study participation question. In this case, if both 'full-time' and 'part-time' study tick-boxes are marked but not 'neither of these', the response is coded to 'full-time'. All other combinations of multiple tick-boxes are coded to 'response unidentifiable'.
- Validity edits – identify responses that are not valid. For example, providing a year of arrival in New Zealand that is outside the range of possible years, such as after the census was held.
- Consistency edits – involve comparing responses to different questions. Consistency edits can sometimes result from routing errors. These edits require investigation to identify which response is likely to be erroneous, and fixes are case-specific. For example, the number of years since arrival in New Zealand cannot be greater than a person's age. In this case, the age is assumed to be correct and years since arrival in New Zealand is set to a residual code.

These edits are applied at different phases of the census, and edits could be implemented and resolved in different ways.

The online form applies micro-edits at the point of capture. The online form did not permit capture errors such as multiple responses to a single response question and enforced some validity edits and consistency edits. The 'allowed range of values' feature meant that a respondent could not give an invalid response, for example, a date of birth aged over 120 years. As-you-type functionality allows a respondent to select from a list based on what has been typed so far and this helps to limit invalid responses, although free text could still be invalid. The automated routing meant that respondents were filtered past questions they were not required to answer based on an earlier response. For example, there are fewer questions for children than for adults. Some key questions were made

mandatory (date of birth, sex, and Māori descent) and avoided item non-response at capture. Consistency checks between questions were kept to a minimum on the online form to avoid frustrating respondents.

Having online edits such as these reduces the amount of editing required later, though it does not eliminate all errors. The continued use of a paper form meant that all types of edits were still required during processing.

Micro-edits were specified and coded into the processing system as a sequential set of rules. Consistency edits were applied between variables for the same person, but not between people in the same household. The software tool Canadian Census Editing and Imputation System (CANCEIS) was part of the processing system and was used to implement imputation. While CANCEIS also includes a sophisticated system for census editing it was not used in the 2018 Census.

The outcome of an edit failure could be resolved automatically by determining a valid value or setting to a residual category. Otherwise some edit failures were sent to manual operators to determine respondents' intentions and/or resolve edit failures. In 2018, most manual intervention was for edits that affected family coding. Edit logs recorded the number of times an edit was applied in the processing system. This information was used for monitoring and to guide the future development of edits.

Macro-edits were applied to identify any major data issues in the aggregated results. Checks are typically made for inconsistent combinations or for unusual values that are not defined as micro-edits. Changes were made to the data to resolve problems for affected records. For example, all weekly rent amounts greater than $2,000 were coded to a residual so that the values could be imputed. Issues were prioritised and not all problems found by macro-edits were able to be resolved due to time constraints. Remaining issues are noted in the published information about variables.

Responses that could not be given a valid value in the editing process were coded to residual responses such as 'Response Unidentifiable' or 'Response Outside Scope'. These then enter the item non-response mitigation processes as detailed in the following sections.

Micro-edits were designed for errors typically found in received census forms, and imputed values were constrained to valid values that met consistency edits. However, the processing system did not apply these micro-edits after missing values were replaced by alternative sources (2013 Census or admin data). Capture and validity edits should not be required since the 2013 Census data has already passed edits applied in the previous census and the admin data already only includes valid values. There is potential for these alternative sources to introduce inconsistencies. The macro-edits applied to aggregate results would identify inconsistencies introduced from any source.

For some variables, admin data may be more reliable than census responses. For example, a respondent may incorrectly indicate that their landlord is Housing NZ Corporation. Or respondents may omit benefits or ACC as an income source, when it is clear from tax data that they have received income from those sources. However, in keeping with the approach of respecting respondents' intentions, there were no edits between admin data and census responses, and no changes made to census responses if they differed from admin values. This approach could be reviewed in future.

Item non-response methods in the 2018 Census

Item non-response occurs when values for specific variables for a person, household or dwelling are missing or unusable. For item non-response in the 2018 Census, alternative data sources were used to fill in the missing characteristics when they were available and of good quality. Statistical imputation was only used if alternative data was not available, and only for variables where there was a sound imputation model.

Our approach to filling in missing characteristics involved the following sources:

- Historic data (2013 Census) – information sourced from responses to the 2013 Census for that person or dwelling
- Admin data – information sourced from admin data for that person or dwelling
- Statistical imputation.

In general, 2013 Census responses are prioritised over admin data. Both 2013 Census and admin data are always prioritised over statistical imputation, given that these alternative sources represent information about the actual individual (or dwelling), even though it was not collected directly for the 2018 Census. The pattern of sources differs by variable with combinations of one, two, or all three sources (2013 Census, admin, and statistical imputation) being used, or none. Residual codes only remain in the data for variables that did not have statistical imputation applied.

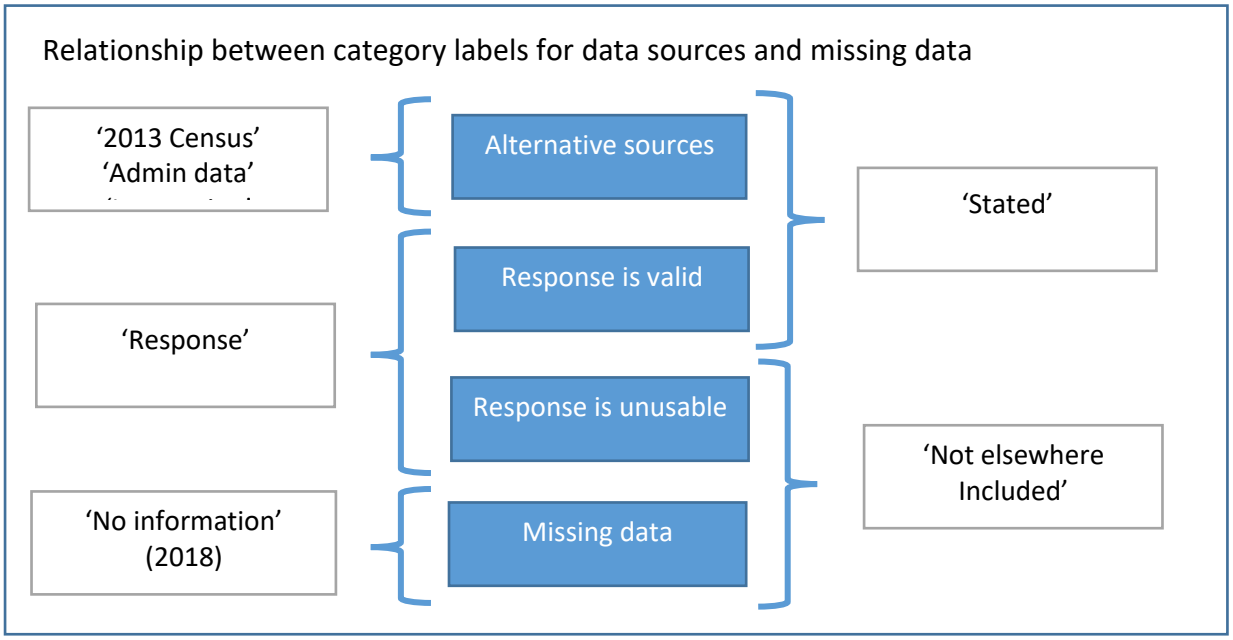
The following sections describe this approach. We first cover the use of the Integrated Data Infrastructure (IDI) to extract information from the 2013 Census and admin data sources. We then describe the CANCEIS software and the statistical imputation process.

Describing source and output categories

Labels for the data source categories in the final dataset are shown in the left-hand side of figure 1. Alternative sources are the 2013 Census, admin data, or imputation. A response from a census form is either a valid response or determined to be unusable through the editing process. ‘Missing’ data is labelled as ‘No information’ in 2018 and ‘Not stated’ in 2013.

When considering data for output purposes, a different distinction is made. Only valid categories provide useful information and are labelled as the total ‘Stated’. Unusable responses are grouped with missing data and labelled as ‘not elsewhere included’. The percent ‘not elsewhere included’ is the main indicator of item non-response in the final dataset.

Figure 1


Relationship between category labels for data sources and missing data
'2013 Census'
'Admin data'
'Response'
'No information' (2018)
Alternative sources
Response is valid
Response is unusable
Missing data
'Stated'
'Not elsewhere Included'

Data sources

In this section, we describe the alternative sources used to fill in missing item information in the 2018 Census.

The Integrated Data Infrastructure (IDI)

The 2013 Census and almost all admin data used to replace missing data and residual responses in the 2018 Census was sourced from the linked data in Stats NZ's Integrated Data Infrastructure (IDI). The IDI is a large research database that holds microdata about people, households, and dwellings. Data is 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. Tax records provide a link from a person's record in the IDI to Stats NZ's Business Register.

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. 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.

All datasets contained in the IDI are then linked to this central spine.

IDI data from the September 2018 refresh is the source of all admin data used in the 2018 Census. The 2018 Census respondents were linked to the September 2018 IDI spine (Stats NZ, 2019c). An overall linkage rate of 97.7 percent was achieved. This means that information provided in the previous census or available in admin sources can be accessed for nearly all 2018 Census respondents. The main exceptions are around 30,000 people only listed on household forms who were unable to be matched to the IDI spine. Since all admin enumerations originate from the IDI spine, they are able to access linked 2013 Census and data from other admin sources.

Dwellings in the 2018 Census have an address ID from the reference list held by Stats NZ's Statistical Location Register (SLR). IDI addresses are also matched to an SLR address ID, and this provides the link for dwelling characteristics information from the IDI. Some addresses contain multiple dwellings, meaning we could not be sure which information related to which census dwelling. Therefore, we did not use 2013 Census or admin values for addresses associated with multiple dwellings.

Workplace variables are obtained from Stats NZ's Business Register. Tax information in the IDI provides a link from an individual to their employer and to the information about the employer held on the Business Register.

Admin data for people in prisons and defence establishments on census night

In addition to the admin data available in the IDI, admin datasets were provided directly to Stats NZ for use in the 2018 Census dataset for two non-private dwelling types: prisons and penal institutions, and defence establishments.

The Department of Corrections 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 were completed from the person's 2013 Census response or admin sources whenever possible. For 100 individuals, the values provided by the agency were used in preference to imputation. For defence establishments, admin record enumerations will be added to the census file with demographic characteristics as provided by the agency. See Overview of statistical methods for adding admin records to the 2018 Census dataset for more information about prisons and defence establishments (Stats NZ, 2019a).

Historic census information used in the 2018 Census

The 2013 New Zealand Census of Population and Dwellings was held on 5 March 2013. Results from the post-enumeration survey (Stats NZ, 2014b) indicated a response rate of 92.9 percent of the estimated population of New Zealand residents in the country on census night. Data from the 2013 Census was used to fill in missing values for 18 variables in the 2018 Census. These included individual variables (for example, birthplace, languages spoken) and dwelling variables (for example, dwelling type, number of bedrooms). The full list of variables is provided in table 2.

The use of historic census data allows us to use people's previous responses to census questions to fill in missing information. For variables that do not change much over time, such as birthplace, we would expect 2013 responses to be a highly accurate source of information for the 2018 Census. For other variables, such as ethnicity, smoking behaviour, and tenure of household, some change over time can be expected, and the use of 2013 responses means we are unable to capture change occurring in the five years between censuses. Despite this, the 2013 census information was expected to provide better results than the alternatives of admin data (for ethnicity) and imputation (smoking behaviour) or leaving a missing value (tenure of household).

In some cases, the 2013 Census and admin sources are complementary. The 2013 Census is able to provide information from earlier time periods that is not available from the admin sources, while the admin sources are available if needed for more recent periods and are the only source for the years since the 2013 Census. Examples are years since arrival in New Zealand, number of children born, and highest qualifications.

Usual residence five years ago is unique as there is no question on the 2018 Census form for this variable. The categories 'not born five years ago' and 'overseas five years ago' are derived from the variables age and years since arrival in New Zealand respectively. Otherwise, usual residence five years ago data was sourced from a link to the 2013 Census when available.

Table 2

Historic census data used for attribute information	
Data source	**2018 Census variables**
Individual characteristics	
2013 Census (Stats NZ)	Birthplace Years since arrival in NZ Māori descent Ethnicity Number of children ever-born Highest secondary school qualification Post-school level Post-school field of study Languages spoken Religious affiliation Smoked – ever Smoked – regular Usual residence five years ago Usual residence address
Dwelling characteristics	
2013 Census (Stats NZ)	Dwelling type Number of rooms Number of bedrooms Tenure of household

Sources of admin information used in the 2018 Census

Census transformation investigations into alternative sources for census variables described the coverage of admin sources and how well they aligned with the statistical standards used by the census and measured the consistency of these admin data sources with 2013 Census responses. This research informed our decisions around those variables that were suitable to use as alternative sources of information in the 2018 Census. Appendix A provides links to the research papers relevant to the admin data used in the 2018 Census.

Admin data from the IDI that was used in the 2018 Census came from the following sources: Department of Internal Affairs (DIA), Ministry of Business, Innovation and Employment (MBIE), Ministry of Education (MOE), Ministry of Health (MOH), Ministry of Social Development (MSD), Inland Revenue (IR), Housing New Zealand (HNZ) and Stats NZ's Business Register. A summary IDI table derived from a range of sources was used for age and sex information. See table 3 for a list of admin information used to mitigate item non-response. Admin data was also used to identify a usual residence address for people who were counted through admin enumeration. For completeness, table 3 indicates the admin sources used to determine a usual residence address for admin enumerations. Also provided in the table is an indication of whether admin data for each source was available up to census night.

Table 3

Admin data sources used for attribute information		
Data source	**2018 Census variables**	**Data available up to 6 March 2018?**
Individual characteristics		
Multiple sources (IDI summary tables)	Age Sex	Yes
Department of Internal Affairs Birth registrations	Birthplace Māori descent Ethnicity Number of children ever-born	Yes
Ministry of Business, Innovation, and Employment Border Movements data	Birthplace Years since arrival in NZ	Yes
Ministry of Health	Ethnicity Usual residence address	Yes
Ministry of Education	Highest secondary school qualification Post-school level Post-school field of study Ethnicity Study participation Usual residence address	Some, but not all
Inland Revenue	Personal income Sources of income Usual residence address	Some, but not all
Business Register (Inland Revenue and Stats NZ)	Industry Sector of ownership Workplace address	Yes
Ministry of Social Development	Usual residence address	No
Accident Compensation Corporation	Usual residence address	Yes
NZ Transport Agency	Usual residence address	Yes
Auckland City Mission	Usual residence address	No
Dwelling characteristics		
Housing New Zealand	Sector of landlord Tenure of household Number of bedrooms Weekly rent paid by households	Yes
Ministry of Business, Innovation, and Employment Tenancy Bonds data	Dwelling type Sector of landlord Tenure of household Number of bedrooms Weekly rent paid by households	Yes
Source: Stats NZ		

As can be seen from table 3, most agencies provide information about several variables, and for some variables, several data sources are used. Where several data sources are available for a single variable, they are used in a specific priority order.

Method for using alternative data sources

Information was extracted from the IDI database for the listed variables, based on derivations described in previous research papers (appendix A) where possible. We created datasets for individuals and for dwellings containing information for each relevant variable derived from linked 2013 Census and/or admin data.

The information in these datasets were securely extracted from the IDI environment into the 2018 Census data store. The values were then transferred to the processing system and used to replace missing information (Stats NZ, 2019b).

For the majority of variables using information from both 2013 Census and admin sources, 2013 Census responses were prioritised. The exceptions were:

- highest secondary school qualification and post-school level, for which the maximum value from either the 2013 Census or MOE was taken, and
- the dwelling variables of dwelling type, number of bedrooms, and tenure of household, for which admin data was prioritised over 2013 Census data.

The specific method for deriving values differed for each variable. Three examples are provided as illustration.

Methods example: Ethnicity

Where there was no ethnicity response sourced from a 2018 Census form the following sources were used (in priority order):

1. 2013 Census
2. DIA birth registrations
3. MOE tertiary enrolments
4. MOH.

Ethnicities were used from each source in prioritised order based on quality assessments of the admin sources (Stats NZ, 2018). Only valid values from these sources were used, not residual categories. MOE data can include multiple records for each person, each representing a given year. We select only the latest available year, that is, closest to the 2018 census date.

DIA collects ethnicity to level 4 of the classification, and comparisons between the 2013 Census and birth registrations show strong consistency between the two sources at levels 1 and 2. While consistency is still good for MOE and MOH at levels 1 and 2, there is a tendency for these agencies to include fewer people with multiple ethnic groups. In the 2013 comparison, Māori was reported less frequently in MOH data compared with the 2013 Census.

The available MOE ethnicity is coded at level 3, and MOH ethnicity is mainly coded at level 2. These codes have been mapped to level 4 and may contribute to ‘not further defined’ categories for larger groupings that cannot be coded to smaller level 3 and level 4 ethnicities.

Where a value was not available from alternative sources, statistical imputation was applied.

Methods example: Years since arrival in New Zealand

Where there was no 'years since arrival in New Zealand' response sourced from a 2018 Census form, the following sources were used (in priority order):

1. 2013 Census (with five years added)
2. MBIE migration data.

Where available, we take the response from the 2013 Census, adding five years to account for the gap between 2013 and 2018 Censuses. Otherwise, if someone appears in the MBIE movements data, we calculate the number of years between their first recorded arrival in New Zealand and 6 March 2018.

The earliest available border movements data are from June 1997. While we are more likely to have 2013 Census data as a source for earlier migrants, when border movements are used for an individual who first moved to New Zealand prior to 1997, we will only pick up their first travel post-1997. We only use admin values up to 18 years in New Zealand in the census, that is, border movements from the year 2000, as data from 1998 and 1999 was unexpectedly high. This may be because earlier migrant arrivals were more likely to be included in the first years of the MBIE travel data. We also make no attempt to identify the first long-term movement. Some individuals may have had a short-term visit before the permanent movement that will be picked up by our approach as their first arrival. See Gath and Das (2019) for more information about admin data on years since arrival in New Zealand.

Methods example: Workplace variables using the Business Register

Workplace variables Industry and Sector of ownership are both derived from the Business Register for the business enterprise. For census respondents, the workplace is identified on the Business Register from information about the place of work provided on the census form. When this is not available, the employer is found through tax information for the individual. In priority order, this is:

1. Employer monthly schedule (EMS) for January 2018 to March 2018 (mainly for wage and salary earners and some self-employed)
2. IR3 forms for the March 2018 year (for self-employed).

We extract a list of all employers for each individual from the EMS for the latest month with available information. For people with more than one employer, we select the one from which they have the highest income.

Issues with alternative data sources

The appropriate use of alternative data sources to fill gaps, using information about the same person, is considered a clear improvement over previous censuses when for most variables item non-response was left as 'missing' in the final dataset. However, a previous census, or admin data, is not the same as a survey response collected at the time of the census, and we acknowledge that there remain limitations in the use of admin data.

Conceptual differences

Information in admin data has not been collected primarily for statistical purposes, and there can be conceptual differences between the information collected, and what might be asked in the census. The admin data used in the census is chosen because concepts map closely to what the census aims to measure. However, ethnicity is an example where responses may differ depending on the context.

Ethnicity is self-perceived, and a person can belong to more than one ethnic group. However, some admin data sources include ethnicity information reported by people other than the individual (for example, reported by a parent), and sources differ in the number of ethnicities they allow people to report.

Data lags

The admin data used from the IDI was taken from the September 2018 IDI update or 'refresh'. Refreshes are scheduled throughout the year to update the IDI with any newly supplied information. Lags in the availability of data can occur between the point of collection by government agencies, the supply of that data to Stats NZ, and integration to the IDI. Adjustments have been made to the derivation of admin values to account for data lags.

Table 3 indicates whether information for each alternative data source was available up to census night. Two sources did not have complete information available through to census night for the variables considered here. Data from MOE covered education enrolments and qualifications gained up until December 2017. This meant study participation could not be directly derived for census night and was instead based on education enrolment information available up to the end of 2017. While the lag in qualifications is short, it does have a specific impact since secondary school qualifications gained from the 2017 school year are not available by December. This will mainly impact the group of students aged 16 to 19 years for whom we rely on admin data for their highest school qualifications.

Similarly, not all tax information for the self-employed was available through to census night. Therefore, total income and sources of income for the self-employed were derived using a combination of information from the 2017 and 2018 tax years.

As a self-identified characteristic, ethnicity can change over time. If a respondent has not had a recent interaction with an agency it may mean that their ethnicity may be out of date. Weekly rent from tenancy bonds is another example as the amount is from the start of the tenancy and is not updated if rents are increased for the same tenancy. In these cases, a person's information held in the IDI may therefore not match what they would have responded at the time of the 2018 Census.

Changes in classification / level of detail

Not all data in the IDI is collected at the same level of detail or uses the same classification as in the 2018 Census. Ethnicity from MOE was available to level 3 of the classification and from MOH was available at level 2 of the classification, both containing less detail than the complete level 4 classification used in the census. As a result, some individuals could be included in a less-precise category. In contrast, income from tax information is more precise than that collected in the census and is coded to census income bands.

There have also been changes in classification, such as for religious affiliation, that mean a small number of 2013 Census responses might not map directly to a category in 2018.

Statistical imputation

Statistical imputation is the process for entering a value for a specific data item where the response is missing or unusable. We use the term 'imputation' to refer to values that result from a statistical process, in contrast to our methods of deriving values from real information about a person or dwelling from the previous census or admin sources.

For the 2018 Census, three types of statistical imputation methodology were used, in order of priority:

- Within-household donor imputation – the person closest in age in the respondent's usual residence household is selected as a donor (used only for ethnicity, Māori descent, religious affiliation and language).
- Deterministic imputation – the characteristic is derived from other variables (only used for sex and Māori descent).
- Donor imputation – based on the nearest-neighbour imputation methodology (NIM).

Within-household donor imputation is used for ethnicity, Māori descent, religious affiliation, and language. For this type of imputation, we find the person within the household who is closest in age to the census respondent with missing information and has a valid response. We then copy their values, provided they have them. Within-household donor imputation is used for these cultural variables as we assume that people are more similar to those within their household and of similar age than otherwise. This might produce a small bias towards undercounting multi-cultural households, but the assumption is more likely to apply than not.

Deterministic imputation uses a set of rules to derive a value for a given characteristic. Deterministic imputation is used for two variables – sex and Māori descent. Sex was imputed based on name for a small percentage of people for whom we had name but no sex. An external R-package (see Gender, an R package) was used to determine the likely sex based on US Social Security Administration data. Whichever sex was identified as more likely was selected.

Māori descent was imputed from iwi, where a valid iwi was used to determine that a person was of Māori descent. In other words, if a person had not responded to the Māori descent question or had provided unusable information but had also indicated affiliation to an iwi listed in the iwi classification, Māori descent was coded to 'Yes'. Deterministic imputation of Māori descent from iwi came after within-household imputation and before donor imputation.

The main imputation has been undertaken using CANCEIS software. The following sections describe CANCEIS and the implementation of donor imputation.

CANCEIS

CANCEIS is a software system developed by Statistics Canada to perform donor imputation based on the nearest-neighbour imputation methodology (NIM; Statistics Canada, 2015). Nearest-neighbour imputation is a standard approach to imputation, and CANCEIS has been used by other national statistics institutes to perform imputation for census data (Bankier, 1999; Guertin et al, 2014), and has been used for Stats NZ household surveys. The 2018 Census is the first time CANCEIS has been used for a New Zealand census. The imputation is designed to correct, as far as possible, for distributional bias caused by differential non-response, that is, when non-respondents are different from respondents. It is not expected to provide exactly the same values for each individual and will increase uncertainty.

To impute characteristics of people, we used CANCEIS at the unit of an individual. CANCEIS finds respondents similar to the person with data needing imputing (the donee) by using matching variables such as age and sex, and other related variables, and searching in close geographic areas. A distance function defined for the matching variables ensures that the potential donors and the donee are similar. After searching for potential donors, CANCEIS provides the closest 10. We selected the closest match as the donor, and the required information is copied from the donor to the donee.

For the 2018 Census, CANCEIS was run in topic-based modules so that groups of related variables would be imputed together, allowing a single donor for a group of variables that a donee required. A donee may receive a different donor for another module. Variables grouped together within a CANCEIS module were imputed together to ensure correlation between the variables was preserved. The modules are run sequentially, which allows them to build upon one another. For example, age, sex, and higher-level usual residence were imputed in Individual module 1, then those variables (including the imputed values) were used as matching variables for Individual module 2 where more detailed usual residence was imputed.

Some respondents may have had variables filled in by 2013 Census data, admin data, within-household donor imputation, or deterministic imputation before the CANCEIS donor imputation process. These respondents were also able to be selected as donors. Level 2 of the item source classification used to denote the source of item information in the 2018 Census includes separate categories for each type of donor.

Consistency edits are included in CANCEIS so that donors are not selected that create implausible combinations of data for the donee. For example, a person aged 15 years or younger with a missing value for main means of travel to education cannot be given a donor that drove to their place of education, as the donee cannot legally drive in New Zealand. This edit is also applied earlier to the input data.

The CANCEIS modules

Table 4 describes the CANCEIS modules, including the variables imputed in each module and the matching variables used to identify donors.

Table 4

CANCEIS imputation modules		
Module: Subject population	**Variables imputed**	**Matching variables**
Individual module 1: **Aggregated usual residence geography and sex and age** *Census night population*	Age, sex, and high-level usual residence geographies (SA2, TALB, region)	Age, sex, usual residence variables, census night variables, ethnicity, dwelling type
Individual module 2: **Small area usual residence address** *Census night population*	Detailed usual residence geographies (X,Y coordinates, meshblock, SA1)	Age, sex, usual residence variables, census night variables, ethnicity
Individual module 3: **Census night address** *Census night population*	Detailed census night address variables (X,Y coordinates, meshblock, SA1, SA2, TALB, region)	Age, sex, usual residence variables, census night variables, ethnicity
Individual module 4: **Cultural variables** *Usually resident population*	Ethnicity, Māori descent, language, religion, and study related variables	Age, sex, usual residence variables, ethnicity, Māori descent, study related variables, birthplace
Individual module 5: **Work and Income variables** *Usually resident population aged 15 or older*	Income, smoking and employment related variables	Age, sex, usual residence variables, ethnicity, total income, smoking variables, employment related variables, highest qualification, dwelling type
Household and dwelling module *All dwellings*	Dwelling record type, dwelling type, number of rooms and bedrooms (for private dwellings), sector of landlord, tenure of household, weekly rent paid by household and rent period (for occupied private dwellings)	Location variables, dwelling record type, dwelling type, number of rooms and bedrooms, sector of landlord, tenure of household, weekly rent paid by household
Family coding modules *Households with two or more residents*	Relationships for each member of the household. Each household size has a separate module (size 2 to size 8)	Relationship variables and living arrangements

Notes: SA1 = statistical area 1; SA2 = statistical area 2; TALB = territorial authorities and Auckland local boards.
A full discussion of the family coding system is beyond the scope of this paper.
Source: Stats NZ

The parameters of CANCEIS are configurable so each module may be different. For example, the number of times a donor could be used was restricted to five times for individual data. For the Household/Dwelling and Family Coding modules, donors could be used more times because there were considerably fewer units in those populations. Other configurable settings include distance functions (how to determine how similar a potential donor is to the donee) and weight (how to determine which matching variables are most important).

Donor selection

Distance is a measure of the space between the record missing a response and a potential donor. CANCEIS offers considerable flexibility in defining distance parameters for different types of variables. When values for a variable are exactly the same, the distance equals 0; when they are totally dissimilar, for example, male and female, the distance equals 1. When the values are similar but not exactly the same, for example, ages 28 and 30, a distance between 0 and 1 is assigned. The premise for nearest neighbour imputation is distance values closer to 0 are more favourable (see Guertin et al, 2014).

The search for potential donors can be computationally intensive for large datasets like census. CANCEIS shortlists the best potential donors using the distance functions. Each potential donor encountered is evaluated against the shortlist and either accepted as a closer match or rejected as less similar than those already on the list. The donor search is divided into stages for efficiency, and at the end of each stage a decision is made whether there is a sufficient shortlist from which to choose a donor or whether to continue to search the next stage. CANCEIS uses a ripple search method looking for potential donors above and below moving out from the record missing data. The approach means the sorting and ordering of the dataset variables is critical. Geographic variables were correlated with most of the variables census imputed, so the datasets were sorted geographically. Once a shortlist of potential donors has been assembled, CANCEIS chooses one record to be the actual donor. Parameters were set so that CANCEIS selected the donor with the smallest distance measure, which meant that the same donor would be selected if the imputation was rerun. This was important because we were aiming for consistency in the iterative processing system. Monitor codes are produced for each variable imputed so that information about the imputation is preserved.

Limitations

With the exception of the Family Coding modules, information about other members of the household is not included in the matching variables, so each missing individual record is imputed independently of those within the same dwelling. This is a potential area of improvement for future applications of CANCEIS within Stats NZ and has the potential to improve data quality. Although it would be more complex to apply, CANCEIS does have the functionality to work with the household as a unit. This would mean that cultural variables could be imputed from a similar household, allowing for more variation within households than is possible with the current within-household imputation. Edits such as those between age and relationship could be applied and resolved, thus improving household and family relationship data.

The imputation methodology using CANCEIS was developed before we were aware of the lower than expected response rate for 2018 Census. Testing of the CANCEIS methodology was performed using 2013 Census responses and assuming a 10 percent non-response rate. We have not tested how well CANCEIS performs with an imputation rate higher than 10 percent.

Census night location for hotels, motor camps, and public hospitals

This section describes the specific case of assigning a census night location within certain non-private dwellings (NPDs) to people with an unknown location on census night. People counted through admin enumeration at dwellings classified as ‘Unoccupied – Residents Away’ only have a known usual residence and not a census night location, as they were identified as away on census night. Some of these people were assigned a census night location within the NPD subtypes of hotels/motels, motor camps, and public hospitals.

This was done to adjust the substantial census night undercount of New Zealand usual residents in these types of NPDs. These NPDs typically consist mainly of people who are visiting on census night but usually live elsewhere. Using data available from other sources, we estimated the expected number of people present in these NPD types on census night. For hotels/motels and motor camps, a combination of information from the March 2018 Accommodation Survey and 2013 Census was used to estimate the expected number of New Zealand domestic residents on census night. This was done for each dwelling that could be linked to at least one of these sources. For hospitals, the public hospital discharge dataset in the IDI provided the number of usual residents in each hospital on census night. As the 2018 hospital discharge data was not available in time for census processing, an average of the three previous years occupancy on the equivalent date was used.

For each dwelling type, demographic distributions of the census night occupants were obtained from the 2013 Census. The difference between the estimated census night occupants and census responses gives the expected undercount by five-year age group, sex, and territorial authorities and Auckland local boards (TALB).

Based on these estimates, admin enumerations without a census night location (from 'Unoccupied – Residents Away' dwellings) were assigned to these NPDs on census night, effectively imputing a census night location. The allocation process completes each dwelling as far as possible, while meeting the overall expected demographic distributions. Note that this process was not used to add additional people to the census file, but only to assign a census night location to people known to be away from home on census night but with no known census night location.

Of the 2,795 NPDs requiring extra census night occupants, 2,647 received at least one person. Table 5 shows the final counts of New Zealand usual residents in NPDs on census night. The final figures are very close to expected for public hospitals, and slightly lower than the estimated figures for hotels and motor camps. For hotels, while the final count is more than twice the number indicated by census responses, the remaining estimated undercount of more than 9,000 is due to a lack of suitable candidates from the admin enumerations. For motor camps, with just 540 people added, we are being conservative to avoid over-counting.

Table 5

Counts of census night occupants by dwelling type for New Zealand usual residents

NPD subtype	Census responses	Admin enumerations imputed to NPD	Final total count	Estimated total
Hotel, motel, or guest accommodation	21,798	26,820	48,615	58,000
Motor camp / camping ground	6,378	540	6,918	8,400
Public hospital	4,578	3,138	7,716	7,600

Source: Stats NZ

There remain more than 11,000 usual residents of admin households in 'Unoccupied – Residents Away' dwellings who have not been assigned a census night dwelling. They were not included in these NPDs because they did not fit the demographic profile we aimed to match. These remaining people are given a census night location via CANCEIS item imputation. Imputation provides a census night meshblock for these people and does not place them at specific dwellings on census night.

Variables with no adjustment for item non-response

There were some variables for which no adjustments were made for missing or unusable information. In some cases, no admin data was available or the census question had not been asked in the 2013 Census. Variables were assessed early on in the development process and some of the reasons why no adjustments were applied include:

- the variable was low priority
- a good-quality admin source was not available or research had not been completed
- the information varied too much over time
- operational limitations.

The following variables did not have any alternative data sources or imputation used to replace missing characteristics and residual responses. We include two variables – usual residence one year ago, and years at usual residence – where the only other source is the use of admin age for the category age less than one year.

Individual variables:

- Usual residence one year ago
- Years at usual residence
- Disability/activity limitations
- Legally registered relationship status
- Individual home ownership
- Unpaid activities.

Dwelling variables:

- Main types of heating/fuel types used to heat dwellings
- Access to telecommunications systems
- Number of motor vehicles
- Dampness indicator
- Mould indicator
- Access to basic amenities.

Measuring the quality of data sources

As with the 2013 Census, the 2018 Census used a quality assurance framework to indicate whether the quality of the final dataset was fit for purpose. The framework consists of three metrics related to different aspects of quality, that contribute to the overall rating of a variable. Some aspects of 2013 quality measurement have been revised for 2018 as a result of the new data sources and change in methodologies. Data quality assurance for 2018 Census has a detailed explanation of the quality rating scale. The quality ratings for 2018 Census variables are published in the 2018 Census information by variable and quality in DataInfo+ along with relevant commentary.

The quality rating for metric 1, 'data sources and coverage', is a quantitative score used to assess each variable on the overall quality of the coverage and data sources used. The rating for a valid census response is defined as 1.00, and a missing value is rated as zero. Ratings for other sources are the best estimates available of their quality relative to a census response. While metric 1 is calculated as a specific number, there is however an inherent uncertainty to the score. For example, we recognise that census responses will include some errors due to, for example, respondent misunderstanding or census processing errors.

Quality of alternative data sources

The historic 2013 Census and admin source ratings reflect measured consistency with the 2018 Census. The ratings have been derived by comparing 2013 Census and admin values to 2018 Census responses for linked individuals or dwellings. Only valid responses from a received 2018 Census form are included. This comparison provides an indication of the level of consistency between the sources for a group of individuals with both sets of information available. An assumption is made that this provides a good indication of the quality of the alternative sources for those whose census data is missing.

The ratings for alternative sources may be conservative in some situations where the previous census or admin variable is more accurate than the 2018 Census response. In some cases, the administrative value is given a rating of 1 where it is known to be as good, or better quality than survey responses (for example, age and sex, and Industry and Sector of Ownership from the Business Register).

For income and highest qualifications, the available admin data is high quality, but some data will be missing. For example, the available IR tax data does not include some investment income, nor does it include non-taxable income. Therefore, tax data is unlikely to over-estimate total income, but may be an under-estimate. When calculating the quality rating score, we assume that if the respondent's income is in a lower band than the admin, then the admin is correct. However, if the respondent reports a higher income, we assume the census response is correct. Similarly, for highest qualifications, if there is a formal record of a higher qualification than is reported by the respondent, we assume the admin qualification is correct. Otherwise a higher qualification reported by the respondent is assumed to be correct.

We apply the same priority ranking in these comparisons as is used in the derivation process. For example, when deriving birthplace, we prioritise historic 2013 Census records over both sources of admin records. To derive the 2013 Census rating, we compare 2013 Census responses to 2018 Census responses. To derive the admin rating, we compare admin responses to 2018 Census responses only for individuals who did not also have a 2013 Census response.

The quality rating is intended to give a good indication of the overall quality of these other sources for the main uses of census data. Where reasonable, we assess the level of exact agreement for these responses, that is, how many of the 2013 Census or admin values were the same as the 2018 Census response (for example, the same country of birth, or Māori descent). For some variables, (personal income, weekly rent, and years since arrival in New Zealand) requiring an exact match is considered unnecessarily precise, and scores are calculated from agreement with one band, or one year.

For variables with detailed hierarchical classifications, a decision is made about which level of the classification to use. Ethnicity is evaluated at level 2 and languages at level 3 of the classification. For some variables we also needed to allow for multiple responses. For ethnicity, we calculate the proportion of 2018 Census level 2 ethnicity responses that were also observed in the 2013 Census or admin data. Similarly, for languages and income sources the comparison is the proportion of each 2018 Census category that was also observed in 2013 Census or admin responses. For individuals with multiple ethnicities (income sources or languages spoken), each category found in 2018 is compared separately.

For religious affiliation, the comparison is exact agreement for a combination of level 1 religious affiliations. Some changes in classification between 2013 and 2018 Censuses and fluctuation between levels of specificity made lower level comparisons less relevant for religion.

For variables where one category dominates other categories (for example, most people speak English) the quality ratings will reflect consistency of the majority category more than the smaller categories.

Table 6 provides the quality ratings for 2013 Census data and admin data for individual variables, and table 7 provides the quality ratings for dwelling variables.

Table 6

Quality ratings for 2013 Census data and admin data – individual variables			
Variable	**Quality rating score**		
	2013 Census	**Admin data**	**Derivation of rating score** Comparisons are with 2018 Census
Age	-	1.00	Admin values are high quality
Sex	-	1.00	Admin values are high quality
Usual residence meshblock	-	0.84	Mean of predicted meshblock probabilities for admin enumerations (Stats NZ, 2019a)
Address one year ago	-	1.00	Age 0 derivation only. Same rating as age.
Years at usual residence	-	1.00	Age 0 derivation only. Same rating as age.
Census night meshblock	-	0.77	Rating for usual residence meshblock multiplied by proportion of people at usual residence on census night
Birthplace	0.99	0.92	Exact agreement
Years since arrival in New Zealand	0.92	0.70	Agreement within one year
Māori descent (census)	0.95	0.92	Exact agreement. 'Don't Know' responses included in comparisons.
Ethnicity	0.91	0.76	Percent agreement for level 2 responses
Number of children ever born	0.96	0.97	Exact agreement
Partnership status	-	0.75	Exact agreement on partnership status
Study participation	-	0.89	Exact agreement for not studying
Highest post-school level	0.86	0.56	Within one qualification level
Highest secondary school qualification	0.82	0.57	Exact agreement
Highest qualification	0.83	0.83	Derived value equal or higher than census
Educational institution address	-	0.50	Estimated
Personal income	-	0.84	Admin value equal or higher, or one band lower than census
Sources of income	-	0.72	Percent agreement with each income source category
Sector of ownership	-	1.00	Admin values from Business Register
Industry	-	1.00	Admin values from Business Register
Workplace address	-	0.50	Estimated
Languages spoken	0.93	-	Percent agreement with each level 3 category
Religious affiliation	0.84	-	Exact agreement for combination of level 1 religious affiliations
Cigarette smoking behaviour: ever smoked	0.93	-	Exact agreement
Cigarette smoking behaviour: regular smoker	0.93	-	Exact agreement

Note: Only variables sourced from 2013 Census or admin data are presented in the table.

Source: Stats NZ

Table 7

Quality ratings for 2013 Census data and admin data – dwelling variables			
Variable	**Quality rating score**		**Derivation of rating score**
	2013 Census	**Admin data**	Comparisons are with 2018 Census
Dwelling type	0.94	0.80	Exact agreement separate/joined dwellings
Number of rooms	0.79	-	Within one room
Number of bedrooms	0.79	0.85	Exact agreement
Tenure of household	0.77	0.93	Exact agreement
Sector of landlord	-	0.96	Exact agreement
Weekly rent paid by household	-	0.89	Within one band
Source: Stats NZ			

Tables 6 and 7 show that the consistency of 2013 Census data and admin data with 2018 Census responses was generally high, and where both sources are available, is higher for 2013 Census than the admin sources. The exceptions are two dwelling variables – number of bedrooms and tenure of household – where the admin sources (HNZ Corporation and Tenancy Bonds) were more consistent with 2018 Census. The admin data was given priority for both variables.

The following variables have relatively lower quality ratings from admin data:

- Highest post-school level and highest secondary school qualification: The quality rating for highest qualification is higher once these two input variables are combined.
- Educational institutional address: Missing values are mostly assigned to 'Regional council not further defined' based on usual residence meshblock. The quality rating reflects that even though the regional council is very likely to be correct, this does not provide full information as to the specific educational institution.
- Workplace address: Some missing values are assigned to 'Regional council not further defined' based on usual residence meshblock. The quality rating reflects that even though the regional council is very likely to be correct, this does not provide full information as to the specific workplace.

Quality of statistical imputation

As part of the quality assurance framework for the 2018 Census, a quality rating system was devised for imputed values. Within-household imputation and deterministic imputation were given a flat rating of 0.7. Ratings for variables receiving CANCEIS imputation have been calculated using a base rating that is multiplied by the quality rating of the donor data source (see Data quality assurance for 2018 Census). Base ratings were informed by assessing the consistency between CANCEIS imputations and 2013 Census responses for a subset of responses during development and testing of the software. The high-level process was to take the 2013 Census usual resident adult population, remove some responses and compare the imputed results to the original results. People created by unit imputation in 2013 (called 'substitute records' at the time) were ignored. It should be noted that only 10 variables receiving CANCEIS imputation in the 2018 Census were included in the investigation.

These tests showed that while consistency between true and imputed values differed by variable, at the national level, the distributions were closely preserved when using the imputed values. At least in the relatively small test examples used, the imputation was shown to be working well.

For the 2018 quality ratings, three base rating are used: low = 0.5, mid = 0.6, and high 0.7. Any variable that was not included in the analysis was given the mid-level base rating by default. We have based the imputed scores on consistency between imputed and real values, in the same way as we have approached ratings for alternative data sources. However, in contrast to the admin and 2013 Census sources, imputed values are modelled values. The main purpose of the imputations is to correct, as far as possible, for distributional bias caused by differential non-response, that is, when non-respondents are different from respondents. To this end, the key indicator of a successful imputation is improving the distributions and maintaining relationships between variables, rather than providing exactly correct predictions for individuals (although the success of imputation will also depend on what the data is used for). The use of only three ratings (0.5, 0.6, and 0.7) emphasises the fact that we do not have any better way of estimating the quality of the imputation for each variable, and we don't wish to imply more precision than we actually have. Table 8 provides the quality ratings for statistical imputation for all imputed variables.

Table 8

Quality ratings for statistical imputation

Rating	Variables	Included in comparative analysis
Low (0.5)	Personal income	Yes
	Sources of income	Yes
	Industry	Yes
	Occupation	Yes
	Main means of travel to work	Yes
Middle (0.6)	Census night meshblock	No
	Māori descent (Census)	No
	Ethnicity	No
	Languages spoken	No
	Religious affiliation	No
	Main means of travel to education	No
	Weekly rent paid by household	No
High (0.7)	Age	No
	Sex	No
	Work and labour force status	Yes
	Status in employment	Yes
	Sector of ownership	Yes
	Cigarette smoking behaviour: ever smoked	Yes
	Cigarette smoking behaviour: regular smoker	Yes

Source: Stats NZ

Results

This section provides results of our item non-response adjustments, including the percentage of each subject population with information coming from an alternative data source or from imputation.

Mitigation of non-response

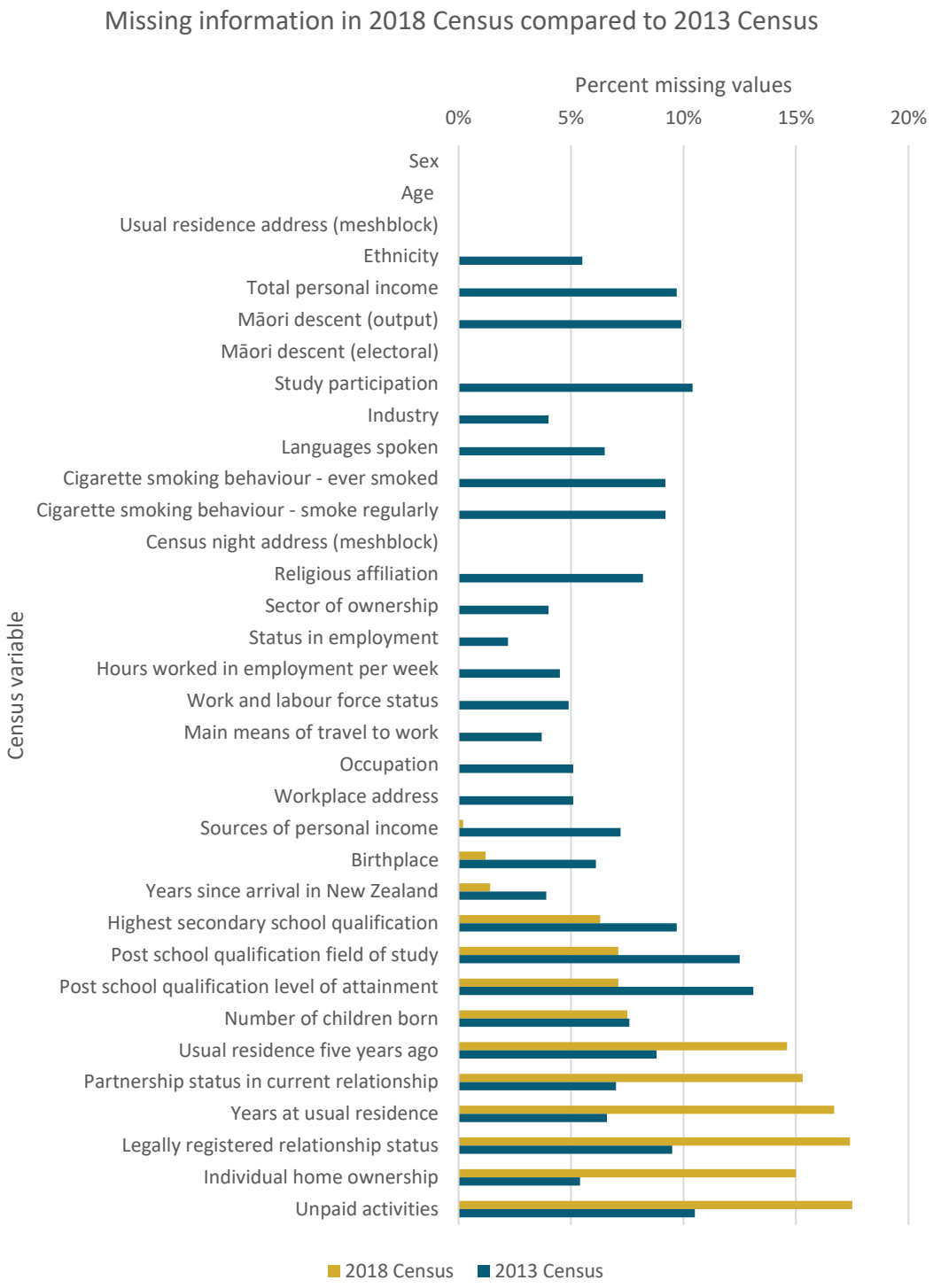
The use of alternative sources and imputation has removed or minimised item non-response for most census variables. For variables where these mitigations were not used, the level of non-response is around 15 percent higher for individual variables due to the lack of individual census forms. This contrasts with 2013 Census where nearly all variables have item non-response of between 1 and 5 percent from census individual forms plus close to 5 percent missing information from substitute records.

For dwelling variables, item non-response has similarly been removed or minimised for variables that had non-response mitigations applied. For variables where mitigations were not used, the level of non-response ranges from 8 to 11 percent. In the 2013 Census, the non-response for dwelling variables ranged from 4 to 8 percent.

Figure 2 shows the distribution of ‘not elsewhere included’ (not stated plus any residual categories) for individual variables for the 2018 Census and 2013 Census. Comparisons are only provided for variables included in both 2013 Census and 2018 Census.

Figure 2 highlights where our non-response methods have reduced the proportion of not elsewhere included information. Variables at the top of the graph are those with alternative data sources and/or imputation, and it can be seen that missing information was reduced to zero for many variables (an improvement from the 2013 Census). Moving down the list of variables in the graph are those where less alternative information was available and fewer non-response methods were implemented. For some of these variables, the amount of missing information in 2018 is much higher than in 2013 (for example, years at usual residence, and unpaid activities).

Figure 2


Missing information in 2018 Census compared to 2013 Census
Percent missing values
0%
5%
10%
15%
20%
Census variable
Sex
Age
Usual residence address (meshblock)
Ethnicity
Total personal income
Māori descent (output)
Māori descent (electoral)
Study participation
Industry
Languages spoken
Cigarette smoking behaviour - ever smoked
Cigarette smoking behaviour - smoke regularly
Census night address (meshblock)
Religious affiliation
Sector of ownership
Status in employment
Hours worked in employment per week
Work and labour force status
Main means of travel to work
Occupation
Workplace address
Sources of personal income
Birthplace
Years since arrival in New Zealand
Highest secondary school qualification
Post school qualification field of study
Post school qualification level of attainment
Number of children born
Usual residence five years ago
Partnership status in current relationship
Years at usual residence
Legally registered relationship status
Individual home ownership
Unpaid activities
2018 Census
2013 Census

Contribution of each source by variable

Figure 3 illustrates the relative contribution of data sources to each individual variable and the percent missing (coded to 'No information'). Census responses are the major source for all variables, (with the exception of usual address five years ago). The graph is ordered by the proportion with no information, and secondly by the proportion imputed. The value of the alternative data sources in reducing the amount of imputation, and the level of missing data is evident from top to bottom of the graph. Figure 4 presents the same information for dwelling variables. Appendix B provides the data sources used for item non-response mitigation (2013 Census, admin or imputation) and the percentage each source contributes to individual variables.

Figure 3

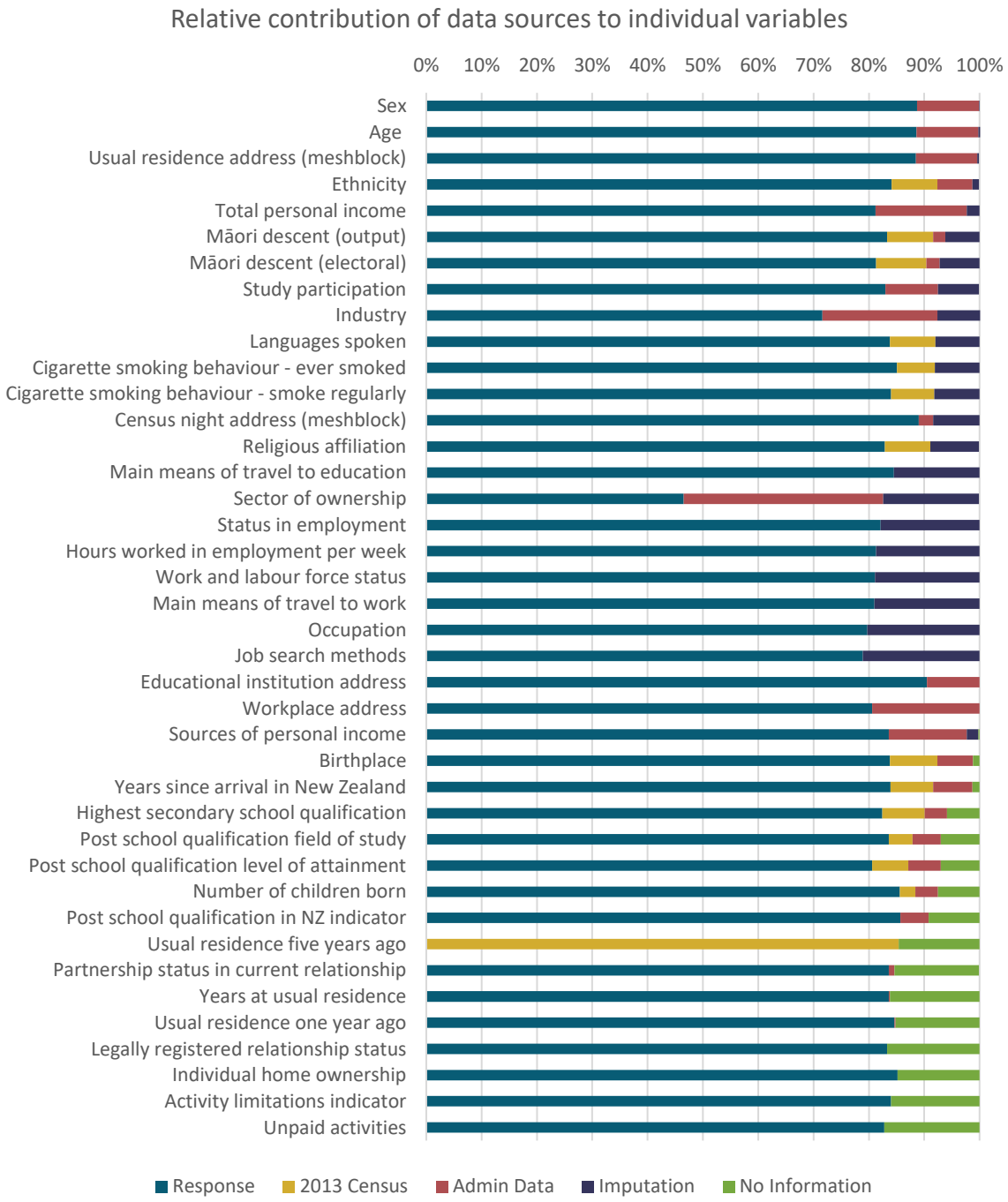
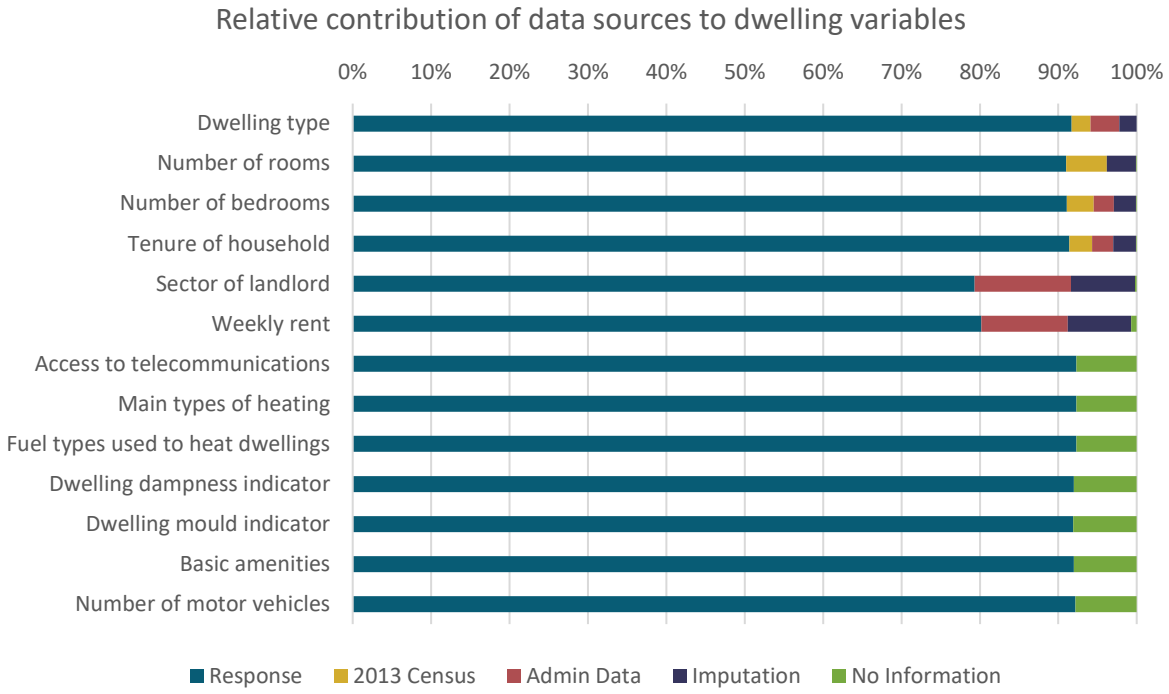


Figure 4



Unequal impacts of missing data

The use of admin enumerations has improved the census count for some subgroups of the population over previous censuses, and together with the alternative data sources and imputation, has for many variables maintained (or in some cases improved) the quality of information seen in previous censuses. However, for other variables high rates of imputation or missing data mean that quality is lower than in previous censuses, and those subgroups with lower response rates to the field collection, such as Māori, Pacific, and young adults are more adversely affected. Some geographies will be more affected by missing data than others. There may be breaks in the time series due to both improvement from previous censuses, and from higher levels of missing data.

For example, the use of alternative data sources and statistical imputation means that everyone in the census usually resident population has ethnicity information, and all employed people have an occupation (compared with 5.5 percent having missing ethnicity and 5.1 percent having missing occupation information in the 2013 Census). Occupation information has not been sourced from either the 2013 Census or admin data, so all values are either from 2018 Census responses or CANCEIS imputation.

Figure 5 shows the distribution of response and imputed values for each level 1 ethnic group. For the European ethnic group, 85 percent of occupation values are taken from a census response, compared with only 66 percent for Māori and 59 percent for Pacific ethnicities. All records from a household set-up form and all admin enumerations can only have imputed values for occupation. The Māori and Pacific populations are relatively overrepresented in these groups, resulting in higher levels of imputation for variables such as occupation.

Figure 5

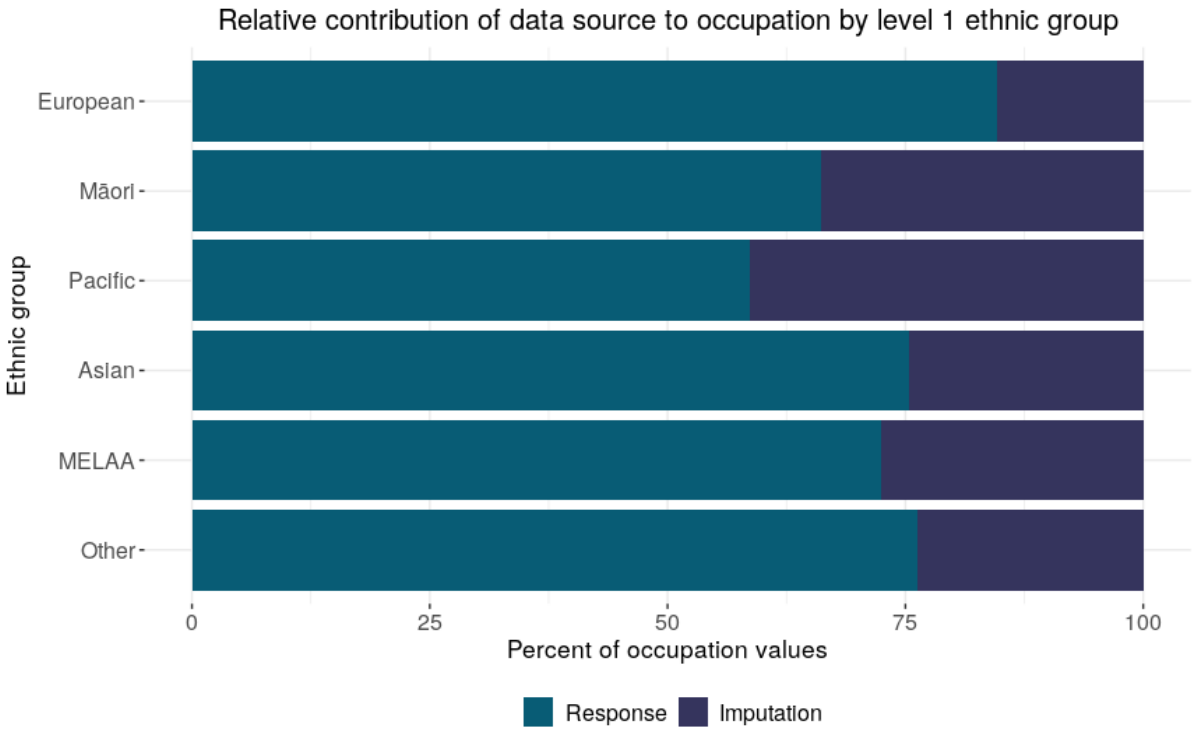
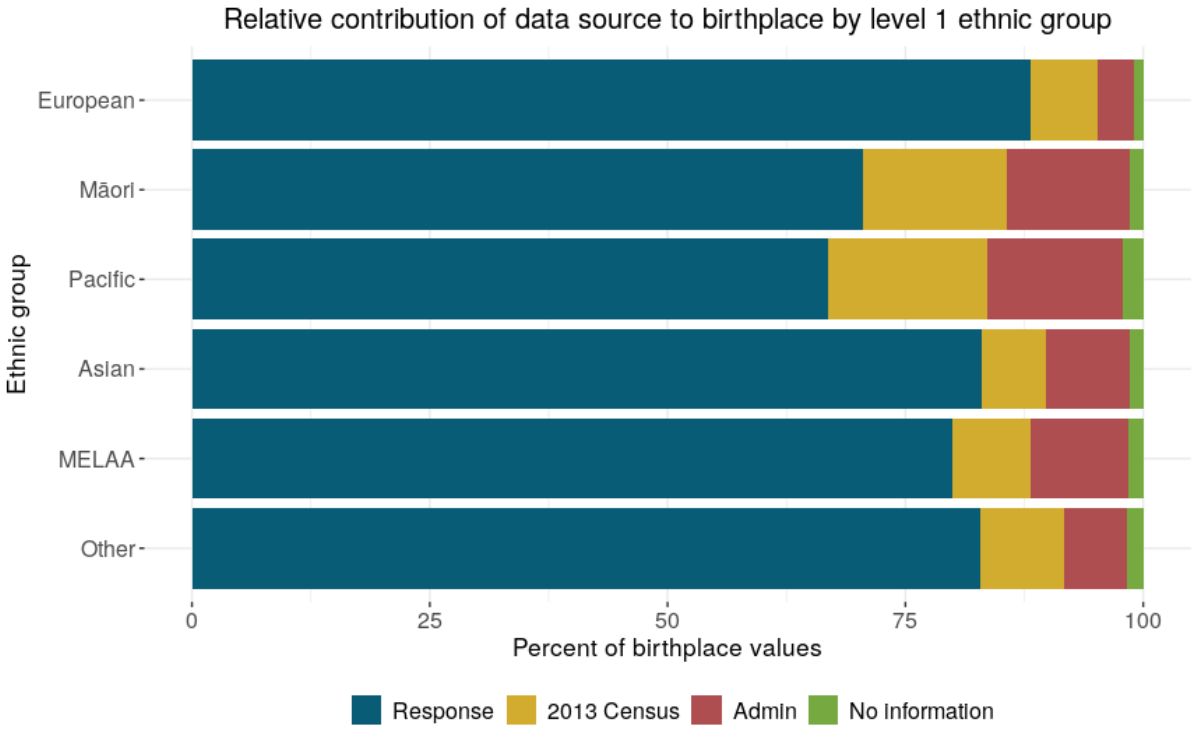


Figure 6 shows a similar comparison for birthplace by level 1 ethnic group. A similar pattern is observed with fewer responses for the Māori and Pacific ethnicities. However, in this example alternative sources are available, meaning most of the remaining values are sourced either from the 2013 Census or admin data.

Figure 6



Conclusion

The 2018 Census has implemented new methods for editing census data and for adjusting for item non-response. Use of online edits has greatly improved data quality captured for online forms compared with paper forms. Item non-response has been successfully removed or reduced to levels below those seen in previous censuses for a majority of variables. This has been achieved despite around 15 percent of the census dataset having no information from Individual forms.

For the first time in a New Zealand census, the 2018 Census used data from alternative sources to fill gaps in the census. We used these alternative data sources wherever possible to fill in missing or unusable information for census respondents and to provide all characteristics for admin enumerations. Our alternative data sources are 2013 Census responses and admin data. Use of these alternative sources depends on each variable and whether any alternative source is available and is known to be of high quality.

For about half the variables, we use donor imputation to fill in any remaining missing or unusable information. Donor imputation has been achieved mainly through a nearest-neighbour donor imputation methodology implemented in the CANCEIS software, which copies values from people identified as most similar to those with missing information. The methodology itself is unbiased and designed to preserve distributions rather than optimising for finding the true value for a given individual. However, the use of imputation increases the uncertainty of the census data, and caution should be exercised when imputation rates are high.

The 2018 Census was the first time the CANCEIS software had been used in a New Zealand census. It enabled the imputation of a wider range of variables than in the past using a common methodology. However, there is potential for further improvement in future censuses to implement the powerful ‘minimum change’ editing features offered by the software, and to use the household as a unit for editing and imputation.

These new methods and the scale of their use represent a significant change in the sources of information for attribute variables from previous censuses. Populations with lower response rates to the field collection and thus greater use of admin enumerations (including Māori, Pacific, and younger adults) are more reliant on alternative sources and imputation, and more adversely affected by remaining missing data.

Time series data is inevitably affected. The admin enumerations mean that the 2018 Census includes more people who are typically ‘hard to count’ than previous censuses. For variables where no non-response mitigation is available, the level of missing data will be higher than in the past. However, quality will be improved over previous censuses for variables where good alternative sources are available and imputation rates are relatively low. For these variables, 2018 Census counts across all categories of a variable will be higher because there is no longer missing data, and distributions will be less affected by non-response bias.

References

Bankier, M (1999). *Experience with the new imputation methodology used in the 1996 Canadian Census with extension for future censuses*. Proceedings of the Workshop on Data Editing, UN/ECE, Rome 2–4 June 1999.

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.

Gath, M & Das, S (2019). Potential for admin data to provide country of birth and years since arrival in New Zealand 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.

Guertin, L, Bureau, M, & Morel, J (2014). Editing the 2011 Census data with CANCEIS and options considered for 2016. Retrieved from www.unece.org.

OECD (2013). Glossary of statistical terms: imputation. Retrieved from www.stats.oecd.org.

Statistics Canada (2015). *CANCEIS user's guide version 5.2*. Ottawa: CANCEIS Development Team, Social Survey Methods Division, Statistics Canada.

Stats NZ (2013). Post-enumeration survey: 2013. Retrieved from www.stats.govt.nz.

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

Stats NZ (2014b). Post-enumeration survey: 2013. Retrieved from www.stats.govt.nz.

Stats NZ (2016). 2018 Census strategy. Retrieved from www.stats.govt.nz

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

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

Stats NZ (2019b). Processing and evaluating the quality of 2018 Census data. Retrieved from www.stats.govt.nz.

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

Appendix A: Census Transformation research papers

The Census Transformation Programme researched alternative data sources for many census variables before 2018 Census. The published papers and census variables that they apply to are provided in table 9 below.

Table 9

Published Census Transformation research papers	
Census Transformation paper	**Variables**
An initial investigation into the potential for admin data to provide census long-form information: Census Transformation programme.	All
Quality of geographic information in the Integrated Data Infrastructure	Usual residence address (meshblock)
Identifying Māori populations using administrative data: A comparison with the census	Māori descent Ethnicity (for Māori) Languages spoken (for te reo Māori)
Comparison of ethnicity information in administrative data and the census	Ethnicity
Comparing education and training information in administrative data sources and census	Study participation Post-school qualification (level of attainment) Highest secondary school qualification
Comparing income information from census and administrative sources	Total personal income Sources of personal income
Comparing housing information from census and tenancy bond data	Number of bedrooms Tenure of household Sector of landlord Weekly rent paid by household
Potential for admin data to provide country of birth and years since arrival in New Zealand information	Birthplace Years since arrival in New Zealand
Comparing 2013 Census and admin data for number of children born	Number of children born
Source: Stats NZ	

Appendix B

Table 10

Percent of subject population with information from alternative data sources or imputation – individual variables						
Variable	**Subject population**	**2018 Census**	**2013 Census**	**Admin data**	**Imputation[(1)]**	**No information**
Age	Census usually resident population	88.6%	-	11.2% 11.05% IDI personal details 0.10% Corrections <0.1% Ministry of Defence	0.3%	0.0%
Sex	Census usually resident population	88.7%	-	11.2% 11.05% IDI personal details 0.10% COR, prisons <0.1% MoD, defence	0.1%	0.0%
Usual residence address (meshblock)	Census usually resident population	88.5%	-	11.1% 11.1% IDI addresses (Table 3)	0.3%	0.0%
Usual residence one year ago	Census usually resident population	84.6%	-	0.2% 0.2% IDI personal details	-	15.2%
Years at usual residence	Census usually resident population	83.6%	-	0.2% 0.2% IDI personal details	-	16.2%
Census night address (meshblock)	Census night population	89.0%	-	2.6% 2.6% IDI addresses (Table 3)	8.4%	0.0%
Birthplace	Census usually resident population	83.8%	8.6%	6.4% 4.3% DIA, births 2.1% MBIE, migration	-	1.2%

Percent of subject population with information from alternative data sources or imputation – individual variables						
Variable	**Subject population**	**2018 Census**	**2013 Census**	**Admin data**	**Imputation[(1)]**	**No information**
Years since arrival in New Zealand	Overseas-born census usually resident population	83.9%	7.7%	7.1% 7.1% MBIE, migration	-	1.3%
Māori descent (output)	Census usually resident population	83.3%	8.3%	2.2% 2.2% DIA, births	6.2%	0.0%
Māori descent (electoral)	Census usually resident population	81.3%	9.1%	2.4% 2.4% DIA, births	7.2%	0.0%
Ethnicity	Census usually resident population	84.1%	8.2%	6.2% 2.02% DIA, births 2.32% MoE, qualification enrolments and course 1.92% MoH, cohort demographics <0.1% COR, prisons <0.1% MoD, defence	1.2%	0.0%
Number of children born	Female census usually resident population aged 15 years and over	85.6%	2.8%	4.1% 4.1% DIA, births	-	7.5%
Partnership status in current relationship	Census usually resident population aged 15 years and over	83.6%	-	1.0% 1% business rules applied to: DIA births, MSD benefits, WFF tax credits, MBIE migration	-	15.3%
Study participation	Census usually resident population	83.0%	-	9.5% 9.5% MoE, course completions, TEC IT learners, targeted training, and student qualifications	7.4%	0.0%

Percent of subject population with information from alternative data sources or imputation – individual variables						
Variable	**Subject population**	**2018 Census**	**2013 Census**	**Admin data**	**Imputation[1]**	**No information**
Post-school qualification (level of attainment)	Census usually resident population aged 15 years and over	80.6%	6.5%	5.9% 5.9% MoE, course completions, TEC IT learners, targeted training, and student qualifications	-	7.0%
Post-school qualification (field of study)	Census usually resident population aged 15 years and over	83.6%	4.3%	5.1% 5.1% MoE, course completions, TEC IT learners, targeted training, and student qualifications	-	7.0%
Post-school qualification in NZ indicator	Census usually resident population aged 15 years and over	85.7%	-	5.1% 5.1% MoE, course completions, TEC IT learners, targeted training, and student qualifications	-	9.2%
Highest secondary school qualification	Census usually resident population aged 15 years and over	82.4%	7.7%	4.0% 4.0% MoE, qualification enrolments and course	-	5.9%
Educational institution address	Studying census usually resident population	90.5%	-	9.5% 9.5% IDI addresses (table 3)	-	0.0%
Total personal income	Census usually resident population aged 15 years and over	81.2%	-	16.5% 16.5% IR, tax	2.3%	0.0%
Sources of personal income	Census usually resident population aged 15 years and over	83.6%	-	14.1%	2.1%	0.2%

Percent of subject population with information from alternative data sources or imputation – individual variables						
Variable	**Subject population**	**2018 Census**	**2013 Census**	**Admin data**	**Imputation[1]**	**No information**
Sector of ownership	Employed census usually resident population aged 15 years and over	46.5%	-	14.1% IR, tax 36.1% 36.1% IR, tax and EMS	17.3%	0.0%
Industry	Employed census usually resident population aged 15 years and over	71.6%	-	20.8% 20.8% IR, tax and EMS	7.7%	0.0%
Workplace address	Employed census usually resident population aged 15 years and over	80.6%	-	19.4% 14.0% IR, tax and EMS 5.4% IDI addresses (table 3) with business rule	-	0.0%
Languages spoken	Census usually resident population	83.8%	8.2%	-	8.0%	0.0%
Religious affiliation	Census usually resident population	82.9%	8.2%	-	8.8%	0.0%
Cigarette smoking behaviour: ever smoked	Census usually resident population aged 15 years and over	85.1%	6.8%	-	8.1%	0.0%
Cigarette smoking behaviour: regular smoker	Census usually resident population aged 15 years and over	84.0%	7.8%	-	8.1%	0.0%
Usual residence five years ago[2]	Census usually resident population	0.0%	85.4%	-	-	14.6%
Main means of travel to education	Studying census usually resident population	84.5%	-	-	15.5%	0.0%
Main means of travel to work	Employed census usually resident population aged 15 years and over	81.0%	-	-	19%	0.0%

Percent of subject population with information from alternative data sources or imputation – individual variables

Variable	Subject population	2018 Census	2013 Census	Admin data	Imputation[1]	No information
Work and labour force status[3]	Census usually resident population aged 15 years and over	79.9%-84.0%	-	-	16.0%-21.2%	0.0%
Status in employment	Employed census usually resident population aged 15 years and over	82.1%	-	-	17.9%	0.0%
Occupation	Employed census usually resident population aged 15 years and over	79.7%	-	-	20.3%	0.0%
Hours worked in employment per week	Employed census usually resident population aged 15 years and over	81.3%	-	-	18.7%	0.0%

1. Includes within-household donor imputation, deterministic imputation, and CANCEIS donor imputation combined.

2. At the time of publishing, this variable was under review to investigate alternative data sources.

3. Work and labour force status is a derived variable with a number of components. The range of imputation for all components of this variable is presented.

Notes: DIA = Department of Internal Affairs; IR = Inland Revenue; EMS = Employer Monthly Schedule; MoE = Ministry of Education; MoH = Ministry of Health; COR = Department of Corrections; MoD = Ministry of Defence; MBIE = Ministry of Business, Innovation, and Employment; WFF = Working for Families

Variables are not included in the table if they had no values sourced from 2013 Census, admin data, or imputation.

Source: Stats NZ

Table 11

Percent of subject population with information from alternative data sources or imputation – dwelling variables						
Variable	**Subject population**	**2018 Census**	**2013 Census**	**Admin data**	**Imputation[(1)]**	**No information**
Dwelling type	Occupied dwellings	91.7%	2.4%	3.7% 3.7% MBIE, tenancy bonds	2.2%	0.0%
Number of rooms	Occupied private dwellings	91.1%	5.2%	-	3.7%	0.1%
Number of bedrooms	Occupied private dwellings	91.1%	3.4%	2.6% 0.9% HNZC 1.8% MBIE, tenancy bonds	2.8%	0.1%
Tenure of household	Households in occupied private dwellings	91.5%	2.9%	2.7% 0.8% HNZC 1.8% MBIE, tenancy bonds	2.9%	<0.1%
Sector of landlord	Households in rented occupied private dwellings	79.4%	-	12.3% 2.7% HNZC 9.5% MBIE, tenancy bonds	8.2%	0.2%
Weekly rent paid by household	Households in rented occupied private dwellings	80.2%	-	11.0% 4.1% HNZC 6.9% MBIE, tenancy bonds	8.1%	0.7%

1. Includes within-household donor imputation, deterministic imputation, and CANCEIS donor imputation combined.

Notes: MBIE = Ministry of Business, Innovation, and Employment; HNZC = Housing New Zealand Corporation
Variables are not included in the table if they had no values sourced from 2013 Census, admin data, or imputation.

Source: Stats NZ