Stats NZ
Tatauranga Aotearoa

Processing and evaluating the quality of 2018 Census data

New Zealand Government

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

ISBN 978-1-98-858352-5 (online)

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

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

www.stats.govt.nz

Contents

List of tables

List of figures

Purpose and summary

Processing and evaluating the quality of 2018 Census data describes how the 2018 Census data was processed and evaluated to produce a final dataset for public release.

The information people submit in their completed census forms needs to be processed and evaluated to produce data that can be used by a range of agencies to understand the current situation of the national population and any changes that may be required. The 2018 Census took place on 6 March 2018 using a 2018 modernised digital-first model, where the focus moved to encouraging the majority of households to complete the census online instead of via paper forms. While information from the online submissions went straight to the processing stage, completed paper forms were scanned, and the handwritten responses were then translated to a compatible electronic format. We needed to develop a new processing system to deal with the new digital-first model as well as the paper responses.

We used automated and manual processes to:

- combine 2018 Census data with 2013 Census and administrative data to improve the final census count
- apply data transformations, add manual data edits, and create metadata
- add data where there were missing values
- create datasets ready for quality evaluation and eventual release to the public.

Strategic goals of the census programme

The work to process and evaluate the 2018 Census data focused on supporting goals 1, 2, 3, 4, 6, and 7 in the 2018 Census strategy:

- Strategic goal 1: Improve data quality while modernising
    - Maintain relevant, coherent, and fit-for-purpose census information with lower processing cost
    - Improve edit and imputation approaches
    - Use IDI data to improve work variable quality, and replace work and income-related responses.
- Strategic goal 2: Reduce the cost of collection operations
    - Build a processing system adapted for census with the Household Processing Platform team.
- Strategic goal 3: Contribute to organisational capability
    - Enable Stats NZ to benefit from census-product expertise
    - Specify census requirements for the Household Processing Platform.
- Strategic goal 4: Increase use of administrative data
    - Replace manual coding of workplace address and industry with administrative data
    - Experiment with administrative data for imputation
    - Use many administrative data sources for evaluation during field operations and traditional data evaluation.

- Strategic goal 6: Adopt test-driven development
    - Use an iterative process of testing early and often, refining the test plan as findings emerge.
- Strategic goal 7: Deliver customer-driven products and services
    - Change data collection, processing, and evaluation to deliver final census data earlier.

Developing a new processing system

We needed to develop a new data processing system to work with the digital-first model that was introduced for the 2018 Census. We developed this processing system with the aims of: increasing automation, reducing manual intervention, and improving editing and imputation sequencing. Continuous processing, together with flexibility to allow for changes, were key principles of this new processing system. We also modernised our evaluations tools and processes to align better with the online data collection system and the inclusion of administrative data (admin data) (see 2018 Census: How we combined administrative data and census forms data to create the census dataset).

During the processing stage, we commonly need to adjust for missing people (units) and missing information (items) from census forms to improve the quality of the final published census information. In previous censuses, we used the statistical method called 'imputation' to fill in the missing data. 2018 is described in the following subsections.

Adjusting for missing people

The participation rate in the 2018 Census was much lower than expected, and the gap in population coverage was too large to rely solely on unit imputation (known as 'substitutes' in previous censuses). We needed to use a different methodological approach to respond to the larger number of missing people. We had to build up the system we were using to deliver data, change our planned processing and evaluations approaches, and limit our reliance on imputation methods.

We used data from alternative data sources to account for people who had not responded to the census count (known as admin enumeration). For example, for ethnicity the admin data sources were: Department of Internal Affairs (births), Ministry of Education (tertiary enrolments), and Ministry of Health (primary health organisation enrolments). Our new processing system enabled us to combine census form data and admin enumeration to create a census dataset with records for approximately 4.7 million people.

Adjusting for missing information

Missing information commonly occurs in census responses where a respondent misses or decides not to provide an answer for a field or they respond but their answer is not usable (for example, their response is illegible or incoherent). Adding admin enumeration or imputations for missing people also results in missing information for those people.

For the 2018 Census, we intended to use historic and admin data (as well as increased imputation) to fill in missing information. This fitted with the census programme strategic goals of increasing the use of admin data to generate the highest quality data for statistics and to do so in a cost-effective way (strategic goals 4 and 2 respectively). It allowed us to use alternative sources to fill in gaps, but it did not fundamentally change the intended use of this data or alter the information stored in the census datasets from a confidentiality perspective.

The alternative data sources we used

The alternative data sources we used included the 2013 Census, the Integrated Data Infrastructure (IDI) research database, and admin datasets from the Ministry of Defence and Department of Corrections.

Creating the 2018 Census dataset by combining administrative data and census forms data: Our privacy impact assessment and *Data sources, editing, and imputation for the 2018 Census dataset* (Stats NZ, in press) provide more information on how we used admin data in the 2018 Census dataset.

Our new processing system

We ran the new processing system in Pentaho Data Integration (PDI), a Stats NZ enterprise platform. We introduced a module-based processing system that aimed to:

- increase automation
- reduce manual intervention
- improve editing and imputation sequencing.

Continuous processing, together with flexibility around content and specification changes, were key principles of the system. We ran four modules repeatedly to check for and address issues as required. Flexibility allowed us to adjust the data in the final file to account for missing people and information.

While the census data was being collected (the field operation), we used the system to record which households had responded and relay that information back to field operations to assist with prioritising visits to those who had not responded.

Once we had captured the census response data from online and scanned paper forms, we used automated and manual processes to:

- combine it with 2013 Census and admin data (including linking address information) to improve the final census count
- apply data transformations (linking, coding, editing, derivations), add manual data edits, and create metadata (for example, adding 'flags' to the data to identify its source and improve its usefulness)
- add data where there were missing values using the CANCEIS tool – Canadian census edit and imputation system, probabilistic imputation (uses people within the household) and historic 2013 data
- create datasets ready for quality evaluation and eventual release to the public.

Data sources, editing, and imputation for the 2018 Census dataset (Stats NZ, in press) provides more information on the use of other data sources, editing, imputation, and the use of CANCEIS.

The modules

The four modules used in the processing system each involved their own series of processes.

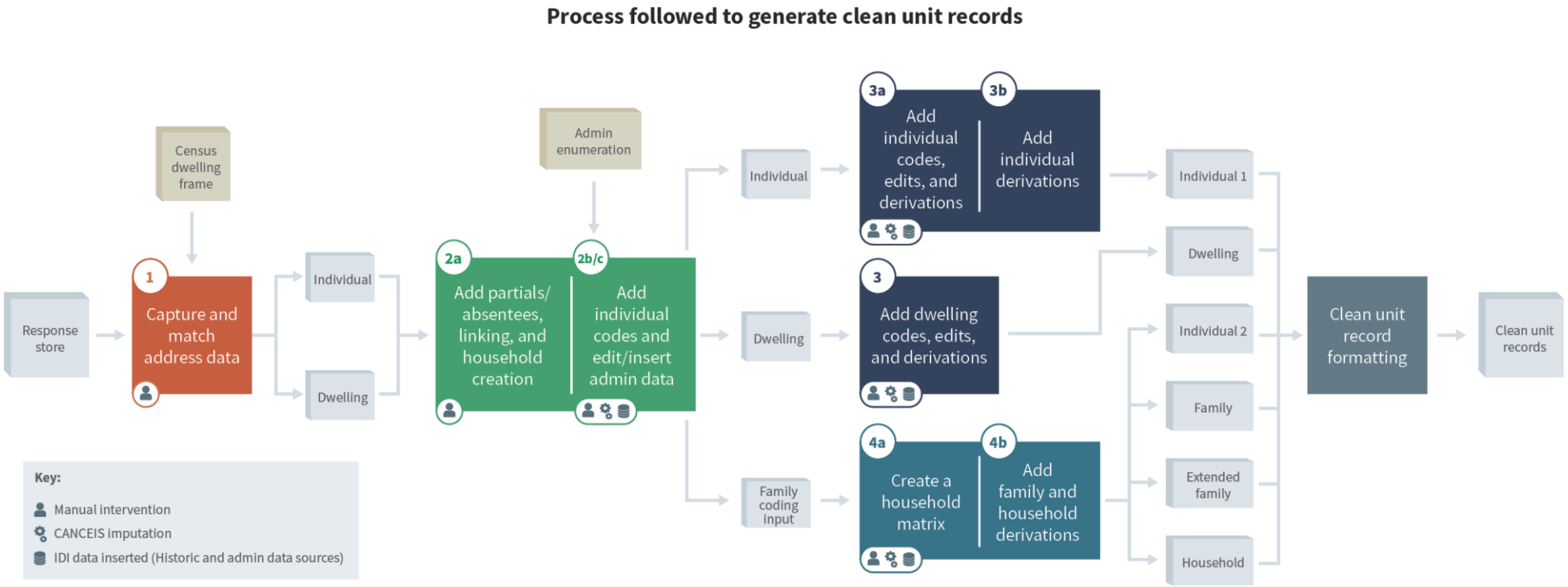
Table 1

Modules in the processing system	
Module	**Process**
1	Capture and match address data
2a	Add partials/absentees, linking, and household creation
2b/2c	Add individual codes and edit/insert admin data
3	Add dwelling codes, edits, and derivations
3a	Add individual codes, edits, and derivations
3b	Add individual derivations
4a	Create a household matrix
4b	Add family and household derivations
5 (not used)	Whole-unit imputation **Note**: This module was decommissioned with the changed methodology to using admin enumeration and was not used in this census.

Further explanation of each module is provided in the next section. Once these modules had been completed, we transformed the raw data into clean unit records, applied confidentiality keys, and produced final datasets.

Figure 1 provides a high-level summary of the process followed through the four modules to generate clean unit records.

Figure 1


Process followed to generate clean unit records
Census dwelling frame
Admin enumeration
Response store
1
Capture and match address data
Individual
Dwelling
2a
Add partials/ absentees, linking, and household creation
2b/c
Add individual codes and edit/insert admin data
Individual
Dwelling
Family coding input
3a
Add individual codes, edits, and derivations
3b
Add individual derivations
3
Add dwelling codes, edits, and derivations
4a
Create a household matrix
4b
Add family and household derivations
Individual 1
Dwelling
Individual 2
Family
Extended family
Household
Clean unit record formatting
Clean unit records
Key:
Manual intervention
CANCEIS imputation
IDI data inserted (Historic and admin data sources)

Key activities for each module

Note: Manual processing edits took place as required across the modules.

Module 1 – Capture and match address data

Module 1 is where information was passed back to the census collection operation to help prioritise field staff visits to households that had not responded.

- Retrieved and reformatted raw response address information from the census dwelling frame (CDF).
- Matched response data to addresses (for census night address, usual residence address, and usual residence one year ago variables).
- Carried out automated and manual processing edits.
- Coded key variables – age and ethnicity.

Module 2a – Add partials/absentees, linking, and household creation

- Linked individual forms to dwelling forms (reconciliation) and then to both census night address and usual residence address (repatriation).
- Added partial responses and absentees to individual dataset.
- Coded key variable – sex.
- Created a household record dataset (initial setup for input for family coding, that is, an initial check that we have the right forms for a dwelling to start building a household).

Modules 2b and 2c – Add individual codes and edit/insert admin data

Module 2b

- Inserted admin data.
- Coded key variables – usual residence and census night address.
- Carried out CANCEIS item imputation.

Module 2c

- Carried out coding for variables in the individual dataset.
- Carried out item imputation, that is, CANCEIS and probabilistic imputation.
- Carried out the final clean-up of family coding inputs.

Module 3 – Add dwelling codes, edits, and derivations

- Carried out coding, imputation, and derivations for dwelling variables.

Modules 3a and 3b – Add individual codes, edits, and derivations

- Carried out coding, imputation, and derivations for individual variables (the majority of derivations occurred in 3b).

Module 4a – Create a household matrix

- Created a relationship matrix – all people in a household were put into a matrix of relationships to others in the household.
- Carried out a CANCEIS imputation for missing relationships.
- Carried out manual intervention edits (that is, where manual operators had fixed or completed a household matrix) and the records go out for family coding.

Module 4b – Add family and household derivations

- Took a relationship matrix as key input and created derivations of household and family variables.

Clean unit record formatting

- Applied confidentiality keys.
- Created the final datasets.

Manual processing

A key aim of the 2018 Census processing was to reduce the amount of manual processing and therefore reduce the number of manual operators required. Manual processing was used to investigate and manually repair (where possible) any data issues that could not be resolved by automated processing.

Manual operators checked forms with quality flags to determine a respondent's intentions and/or resolve edit failures, particularly where complex decisions were required, for example around building family relationships (family coding).

A manual processing operation ran from May to August 2018, involving more than 50 full-time staff. Manual operators fixed items across census variables. They prioritised items by considering the variable's priority rating (priority level 1, 2, or 3), the number of records affected, and the ability to repair the issue.

Data quality assurance for 2018 Census provides more information on the quality rating scale and the quality framework.

When they had completed the manual corrections, we loaded the data back into the dataset to be included in the ongoing automated processing operation.

Adding metadata flags to the data

The composition of the 2018 dataset is more complex than previous censuses due to the use of several different data sources and methodological approaches, as described previously. We added metadata flags to the data to give internal and external users clarity around the sources that we had used to determine the combined output for a census variable.

Table 2 and 3 outlines the levels and codes applied to the individual and item data sources.

Table 2

Individual unit record source, 2018 Census					
Level 1		**Level 2**		**Level 3**	
Code	**Descriptor**	**Code**	**Descriptor**	**Code**	**Descriptor**
1	Response	11	2018 Census individual form	111	2018 Census individual form
1	Response	12	Individuals on the household listing only	121	Individuals on the household listing only
1	Response	12	Individuals on the household listing only	122	Field enumerated rough sleeper
2	Admin enumeration	21	Admin enumeration	211	Admin enumeration (occupied, not responding)
2	Admin enumeration	21	Admin enumeration	212	Admin enumeration (non-private dwelling, prison)
2	Admin enumeration	21	Admin enumeration	213	Admin enumeration (non-private dwelling, defence)
2	Admin enumeration	21	Admin enumeration	214	Admin enumeration (unoccupied, residents away)
2	Admin enumeration	21	Admin enumeration	215	Admin enumeration (occupied, responding)
2	Admin enumeration	21	Admin enumeration	216	Admin enumeration (meshblock usual resident)

Table 3

Item source indicators, 2018 Census			
Level 1		**Level 2**	
Code	**Descriptor**	**Code**	**Descriptor**
1	2018 Census	11	2018 Census form
1	2018 Census	12	2018 Census (missing from individual form)
2	2013 Census	21	2013 Census
3	Admin data	31	Admin data
4	Imputation	41	Within household donor
4	Imputation	42	Donor’s 2018 Census form
4	Imputation	43	Donor’s 2018 Census (missing from individual form)
4	Imputation	44	Donor’s response sourced from 2013 Census
4	Imputation	45	Donor’s response sourced from admin data
4	Imputation	46	Donor’s response sourced from within household
5	No information	51	No information

Evaluating the data collection process

Once we had developed the process for collecting the census data, we needed to evaluate it before we released it for use to check it provided quality data that would be fit for purpose. We evaluated the data quality by census variables (for example, age, sex, ethnicity, religious affiliation).

Data warrants of fitness

We used a warrant of fitness (WOF) process to evaluate each census variable. WOFs help ensure we provide consistent information about the quality of each census variable. While we don't publish WOF documentation, we use it to produce the 'Information by variable' documentation (see Information by variable).

During the WOF process, we:

- analyse the data – check the data quality as listed in the variable specifications. This includes conducting:
    - time series checks
    - checks against expectations
    - checks at lower levels of the classification
    - checks at lower levels of geography
    - consistency checks
    - checks of key cross-tabs with other variables
- assign a quality rating to each variable, with a breakdown of the data sources and non-response rates (Note: Data quality assurance for 2018 Census provides more information on the quality rating scale.)
- complete an outline of the edits, including data edits
- make recommendations for using the data (including recommendations for the next census).

Warrant of fitness quality processes

The change in methodology for the 2018 Census meant that we had to scrutinise the data quality even more rigorously. To identify and fix data issues, our WOF process involved:

- the use of internal variable experts and census topic owner analysts
- raising and presenting data problem reports to an internal advisory group, who:
    - weighed up the cost/benefit of verifying any issues with the data
    - recommended a resolution option.

The WOFs individually and collectively went through a quality assurance process to check:

- each WOF was coded correctly (for example, the correct data was being assessed)
- the analysis was correct
- the data assessment was accurately reflected in commentaries and the final quality rating
- consistency of analysis and quality ratings across the suite of WOFs.

The completed WOFs were then used to inform the quality assessment framework analysis. Data quality assurance for 2018 Census provides more information on the quality assessment framework analysis.

Final dataset

Once we had evaluated the data collection process, we ran the system, applied the confidentiality process, set the final population group numbers, and produced a final dataset for public use.

Glossary

admin enumeration – the use of administrative data to add people to the usually resident census population when a census response has not been received.

administrative data (admin data) – data collected by government or other organisations for non-statistical reasons, such as births, tax, health, and education records.

Canadian census editing and imputation system (CANCEIS) – the editing and imputation tool created by Statistics Canada for use with their census. We have configured CANCEIS for Stats NZ census data.

census dwelling frame (CDF) – A list of all private and non-private dwellings in New Zealand, used during the census.

clean unit records – the de-identified dataset of individual people, groups (for example, families), or dwellings used for outputting census or survey information.

coding – the process by which a description of an item or activity supplied as a survey response is matched to the code of a classification category. The coded categories are defined in standard and census-specific classifications.

confidentiality – the process of keeping data from, and about, individuals and organisations private and ensuring that data is not made available or disclosed without authorisation. Authorisation should ideally be given by the person providing the data but may also be through legislation.

confidentiality key – a random number assigned to each census record to enable random rounding of counts and to add noise (random irregularities) to measures to protect against disclosure.

derivation – a variable that is created or calculated from one or more other variables, for example, a person's age is derived from their date of birth.

edit – detecting, and in some cases resolving, wrong and/or illogical responses to census fields.

family coding – either a couple, a couple with children, or a solo parent with children. The family coding process identifies these groups within households and others in the household who are related to them and creates output derivations for demographics about the groups.

historic data – data sourced from 2013 Census.

imputation – in statistics, the process of replacing missing data with estimated values through statistical methods. For the 2018 Census, the method for estimating values was the nearest-neighbour imputation methodology (NIM), which finds similar respondents with a response to the

variable in question. The processing system then finds the closest match to the respondent with missing or unidentifiable data ('donor respondents') and imputes the donor respondent's response. This method of imputation uses the CANCEIS (Canadian census edit and imputation system) developed by Statistics Canada and used in their census. See also 'item imputation' and 'unit imputation'.

Integrated Data Infrastructure (IDI) – a large research database that holds microdata about people and households. The data in the IDI comes from government agencies, Stats NZ surveys, and non-governmental organisations (NGOs). The data is linked, or integrated, to form the IDI.

item imputation – the statistical process of determining a value for a variable where a response was missing (item non response) or not usable (for example, response unidentifiable). Item imputation is carried out for individuals, dwellings, households, and families. See also 'unit imputation'.

reconciliation (or forms reconciliation) – the process of determining if all forms have been received for an occupied dwelling (or, if not, how many forms are outstanding).

repatriation – transferring all records of individuals who are away from home on census night into the record on their dwelling of usual residence.

unit imputation (previously 'substitutes') – the statistical process of creating a unit record, where data indicates that there is a unit present but no response has been received. Unit imputation in processing occurs for individuals and households but not for dwellings. We had planned to use unit imputation in the 2018 Census (in processing module 5) but used admin enumeration instead.