Stats NZ
Tatauranga Aotearoa

Methodology for using admin data to count people in the 2023 Census

2023 Census | Tatauranga 2023

Te Kāwanatanga o Aotearoa
New Zealand Government

Citation
Stats NZ (2024). *Methodology for using admin data to count people in the 2023 Census*. Retrieved from www.stats.govt.nz.

ISBN 978-1-99-104984-1

Disclaimer
Stats NZ accessed the data in the IDI for use in the 2023 Census in accordance with security and confidentiality provisions of the Data and Statistics Act 2022. Only people authorised by the Data and Statistics Act 2022 are allowed to see data about a particular person, household, business, or organisation and the results in this paper have been confidentialised to protect these groups from identification. Careful consideration has been given to the privacy, security, and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in Integrated Data Infrastructure: Overarching privacy impact assessment and Privacy impact assessment for the use of admin data in the 2023 Census.

Published in May 2024 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

Methodology for using admin data to count people in the 2023 Census describes the methodology for using admin data to add records (referred to as admin enumerations) to the 2023 Census file when an individual has not provided a response.

This paper is one of a collection of documents summarising the methodology used to combine census responses with admin data. Other papers, including those describing the use of alternative data sources for census attributes, are available from Using a combined census model for the 2023 Census.

Summary of key points

The 2023 Census has implemented a combined model by design, with alternative data sources used to supplement census responses. This includes the use of admin data to add records (referred to as admin enumerations) to the census file when an individual has not responded, and when we are confident in the quality of the admin record.

This combination of census and admin data means the value of overall census outputs is higher than we could achieve from responses alone. Admin enumerations increase coverage of the census file and lead to more representative coverage across sub-populations (such as age groups, ethnic groups, or geographic areas).

We aim to include admin enumerations in the census usually resident population, with a specific usual residence location, when we have confidence the record itself is of high quality and that the inclusion of these records will improve the quality of census outputs.

Where possible, we admin enumerate individuals into a dwelling, as this provides the most potential information and allows these people to contribute to household, family, and extended family outputs. However, even when we do not have sufficient evidence to put the individual into a specific dwelling, there can be value in including individuals in a small geographic area (a meshblock). These admin enumerations in meshblocks still contribute to counts of individuals, including for small geographic areas, and so improve the quality of the census file. Individual counts are also important for a range of uses beyond the census, such as the estimated resident population (ERP) and the determination of electoral boundaries.

The 2018 Census represented a significant change in the census model, with admin records included for the first time. Improvements have been made to the methodology implemented in the 2018 Census. In particular, we have increased the proportion of admin enumerations included within a dwelling from 32 percent in 2018 to 58 percent in 2023. Other improvements mean more individuals are likely to be represented in their correct usual residence location.

The final 2023 Census dataset consists of 89 percent census responses and 11 percent admin enumerations. Differing rates of response mean the distribution of census responses and admin enumerations varies across populations. Admin enumerations make up 21 percent of records for the Māori descent population, higher than the overall admin enumeration rate. The percentage of records of Pacific ethnicity (19 percent) coming from admin enumerations is also higher than the overall rate. Certain areas, particularly on the east coast of the North Island, have similarly high proportions of admin enumeration.

The quality of certain census attributes may be lower in populations with a higher number of admin enumerations, particularly where no alternative data sources are available.

Introduction

Admin enumerations in the 2018 Census

Admin enumerations were first introduced in the 2018 Census, in reaction to the lower-than-expected response, and the observed difference in response rates across sub-populations. These admin enumerations replaced the use of ‘substitute’ records in earlier censuses. Overview of statistical methods for adding admin records to the 2018 Census dataset describes the methodology used.

Overall, this approach was considered to have produced better and more representative counts than was possible through the ‘substitute’ methodology used until the 2013 Census. This was particularly true for populations that had previously been undercounted, including Māori and Pacific populations, and young adults.

Although the use of admin enumerations led to considerable improvements in 2018, approximately 350,000 of the admin records added to the census file had insufficient information to be placed into a dwelling and were instead placed in a meshblock (the smallest statistical geography). These individuals were therefore not included in a household. The inclusion of admin enumerations was also applied in a reactive manner in the 2018 Census, which meant there was less time to develop and to

share the methodology ahead of its implementation. Consequently, the 2023 Census has been developed as a combined model by design.

Admin enumerations in the 2023 Census

The 2023 Census admin enumeration methodology builds on that applied in the 2018 Census. Developments have been aimed at improving the clarity and coherence of the methodology.

Guiding principles for admin enumerations in the 2023 Census

For the 2023 Census, we developed a set of guiding principles for including admin enumerations. *Admin enumeration: Planned approach for the 2023 Census* found on Using a combined census model for the 2023 Census provides more information about the early development of these principles, including reference to the 2023 Census: High Level Design.

These principles were intended to help us decide when the inclusion of admin enumerations would lead to improved quality of census outputs. While quality can be defined in many ways, we generally consider higher quality to include increased coverage (how many records are in the census file), and reduced bias (how coverage differs across sub-populations).

There is a direct link between the number of admin enumerations and coverage. The more admin enumerations included, the higher coverage will be. However, there are always trade-offs. Including additional records is only beneficial if we have confidence we are including the right people in the right place. Otherwise, we risk introducing errors in low-level data that decrease the value of outputs for census users. In addition, including too many records may lead to over-coverage, which is difficult to measure and correct for.

Within this concept of quality, we can consider the inclusion of admin enumerations as two largely independent questions, where we assess both the benefit and risk of adding these records.

1. Will including these admin enumerations improve the overall quality of outputs about individuals (nationally and by key sub-populations)?
2. Will including these admin enumerations in a dwelling improve the overall quality of dwelling, household, family, and extended family outputs?

Taking this two-step approach allows us to make use of the strengths of admin data, without extending this use beyond where it is appropriate. We have existing evidence to suggest it is easier to produce individual population counts from admin

data than it is to produce family and household units. We therefore want to address these questions separately.

The principles described below help us answer these two questions.

Principle 1: Including admin enumerations is more valuable than not including admin enumerations at all.

We consider that, in general, including an individual as an admin enumeration is better than not including them at all. Admin enumerations provide clear value in terms of improving the counts of people, overall and for any sub-population they belong to.

However, we still need to have confidence the individual is a genuine part of the census target population, and that we are including them in the correct usual residence location, to produce high quality aggregate counts.

Principle 2: Including admin enumerations should reduce, rather than increase, biases in the census dataset.

One of the main arguments for including admin enumerations is that they improve the representativeness of the census file. If our methods instead further accentuate differences in coverage then this advantage does not hold.

Principle 3: Including admin enumerations in a dwelling is more valuable than including admin enumerations in a meshblock.

Where possible, we aim to include admin enumerations in specific dwellings. While admin enumerations in meshblocks are likely to improve the overall representation of people within the census file, these people are not attached to a dwelling and cannot contribute to detailed information about households, families, and extended families.

Principle 4: Stronger evidence is required when adding an admin enumeration that would alter information about census respondent(s).

For the 2023 Census, we continue to prioritise census responses. In other words, we do not change any information provided by a respondent unless we have very strong evidence to do so. In the context of admin enumerations, this means the threshold is higher for adding individuals into a household at a responding dwelling than for adding individuals into a non-responding dwelling.

Principle 5: Some error at the individual level is acceptable if overall distributions are improved.

We are similarly willing to accept some error at the individual level if we can show that the overall population distributions are improved. Even census responses have some error, and including admin enumerations only when we have overwhelming

evidence that all information is 'correct' would reduce the overall value we can achieve. Instead, we look for a balance between accuracy at the individual record level and the quality of aggregate outputs.

Principle 6: All use of admin data needs to consider the appropriateness of using this data.

Throughout development of the admin enumeration methodology, we have attempted to consider the appropriateness of the use of admin data, including understanding the benefits and other implications of a given approach. This has involved the use of Ngā Tikanga Paihere, a framework developed to "guide safe, responsible, and culturally appropriate use of data".

More information about the consideration of privacy is described in Privacy impact assessment for the use of admin data in the 2023 Census.

Overview of admin enumeration methodology

The rest of this paper provides technical detail about the implementation of admin enumerations in the 2023 Census, building on the guiding principles.

We first detail the key data sources that contribute to the admin enumeration process. We then describe the methods implemented prior to identifying admin enumerations, including:

- determining the admin resident population on census night (7 March 2023)
- determining relationships between admin individuals
- linking 2023 Census responses to the Integrated Data Infrastructure (IDI)
- determining usual residence address from the available admin sources
- identifying respondents overseas on census night.

Following this, we describe the methods used to add admin enumerations, including:

- adding admin enumerations for absentees (people listed as usually residing in a household, but being elsewhere on census night) in responding private dwellings
- adding admin enumerations in responding private dwellings
- adding admin enumerations in non-responding private dwellings
- adding admin enumerations in non-private dwellings
- adding admin enumerations in meshblocks (a small geographic area).

We finish with a summary of results, including differences in the proportion of census responses and admin enumerations across age, Māori descent, level 1 ethnic groups, and geographic areas.

Appendix 1 shows a flow diagram of the admin enumeration processes.

Data sources

This section describes the key data sources used within the admin enumeration process.

All admin enumeration work for the 2023 Census took place within a secure admin production environment specific to 2023 Census. Access to this environment was strictly controlled, with records de-identified, similar to the measures used for the Integrated Data Infrastructure (IDI).

Privacy impact assessment for the use of admin data in the 2023 Census describes these measures.

2023 Census responses

The 2023 Census target population is everyone in New Zealand on census night (7 March 2023). People who usually live in New Zealand but are overseas on census night are excluded from the census target population. Overseas visitors who are in New Zealand are included. For the purposes of admin enumeration, we are primarily interested only in those who are New Zealand usual residents, rather than overseas visitors.

2023 Census responses were captured both online and on paper. An online household set-up form and paper dwelling form requested that all people present at the dwelling on census night be listed by name and provide their age, gender, and relationship to a reference person within the household. They also asked for people who usually live at the dwelling, but are elsewhere on census night, to be listed as absentees. A separate individual form was to be completed for each person present on census night. However, this does not always occur, and some people listed as being present will not have had an individual form completed. Others will not be included within a census form of any type.

The 2023 Census definition of a response requires that an individual will only be included as a response if they have valid values for two out of three of the following fields: name, date of birth, and meshblock. This definition is the same as for the 2018 Census.

The 2023 Census also aims to count all dwellings within New Zealand. Dwellings are classified as either private or non-private. They are also categorised based on their occupancy status on census night.

Integrated Data Infrastructure (IDI)

The Integrated Data Infrastructure (IDI) is a large research database which holds microdata about people and households. Data are gathered from a range of government agencies, Stats NZ surveys and the 2013/2018 Censuses, and non-government organisations. The data are linked together, or integrated, to form the IDI.

The basic structure of the IDI consists of a central 'spine' to which the other data collections are linked at the individual level (Black, 2016; Gibb et al., 2016). Broadly, the target population for the spine is all individuals who have ever been residents of New Zealand. The spine is made up of the union of people in three data sources:

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

The IDI spine in 2023 includes around 11 million individuals found in one or more of these sources. For the 2023 Census, we use data from the October 2023 refresh. This allows for an extended supply of admin data, while still including data specific to census night.

Additional data sources

Additional information was also sourced from government agencies to contribute to the admin enumeration process. These included:

- Department of Corrections data listing individuals in prisons on census night
- Electoral Commission data listing individuals on the electoral roll at 7 March 2023
- Ministry of Social Development data listing individuals receiving civil defence payments in early 2023.

New Zealand Defence Force data was also sourced, but ultimately not used to add individuals to the 2023 Census file due to quality concerns. This represents a change from the 2018 Census, when New Zealand Defence Force data was used to include 798 individuals.

Pre-admin enumeration methodology

This section describes the processes required before we can implement admin enumerations. Table 1 provides an overview of these processes, including changes from the 2018 Census. The rest of this section then provides additional detail about each of these steps.

Table 1: Summary of pre-admin enumeration methodology in the 2023 Census

Process	Description	Changes from 2018 Census
Identify the admin resident population	A set of rules is applied to identify the admin resident population, which represents individuals who we believe were usual residents of New Zealand on census night. The admin resident population serves as the basis for admin enumerations.	Inclusion of additional data sources to improve coverage
Identify relationships between admin individuals	Family relationships are identified between individuals in the admin resident population, using information from previous censuses and admin sources. These relationships can subsequently contribute to family and household characteristics.	• Extended sources • Improved methods for inferring additional relationship types
Link 2023 Census responses to the IDI spine	2023 Census respondents are linked to the IDI spine. These links are used to remove those who have responded from being eligible for admin enumeration.	General improvements to linking methodology
Determine the best usual residence address for admin records	For each individual in the admin resident population, we use a machine learning model to predict a single best usual residence address from the available admin addresses. These addresses, and associated address scores, determine where an individual will be considered for inclusion as an admin enumeration.	Change from a rules-based method to a machine learning model to determine the best address from available admin sources
Identification of respondents overseas on census night	Census respondents who are linked to the IDI and determined to be overseas on census night are reclassified as respondents overseas on census night (ROCNs) and removed from the census usually resident population count.	New for 2023 Census

The admin resident population

We derive an admin resident population (sometimes referred to as the IDI-ERP) as at 7 March 2023, that is used as the starting point for identifying those eligible for admin enumeration.

We begin with the IDI spine, applying a set of rules to determine who was a resident on census day. This follows the same methodology as used in the experimental Administrative Population Census (APC).

Inclusions: retain individuals whose presence is indicated by activity

- For individuals aged five years and over, the spine population is restricted to those who had activity in at least one of the following IDI datasets in the two years before census night.
  - Inland Revenue (IR) tax: employer monthly summary of tax paid at source, or annual tax return data; receipt of taxable benefit payments is included.
  - Ministry of Health (MOH): pharmaceutical prescriptions, GP enrolment and attendance, hospital admissions, non-admission hospital visits, vaccinations.
  - Ministry of Education (MOE): school enrolment, tertiary enrolment, or attainment.
  - New Zealand Customs Service (Customs): recent permanent and long-term arrivals.
  - Department of Corrections: interaction with Corrections system.
  - Ministry of Social Development (MSD): civil defence payments.
- For individuals aged under five, a New Zealand birth registration or visa approval (excluding visitor or transit visas) before the reference date is sufficient for inclusion in the population. For this age group there is no additional requirement of activity in the previous two years.

Exclusions: remove non-residents at the reference date and duplicates

- Linked death records are used to identify individuals with a date of death before census night.
- Linked migration data is used to identify individuals who were not New Zealand residents on census night.
  - Where possible, people are excluded who are not resident according to the 12/16-month rule.
  - Otherwise, individuals are excluded who have not spent at least six months in New Zealand during the period 8 September 2022 to 7 September 2023.

- We remove records identified as duplicates caused by missed links when constructing the IDI spine. Duplicates are identified conservatively as those with identical sex, year and month of birth, and address ID. Both records must have only a single spine source, with one coming from DIA birth registrations and one coming from IR tax registrations. This combination provides strong evidence that both records refer to the same person and that a link has been missed.
- For the 2023 Census application, we also remove New Zealand residents who are temporarily overseas on census night to achieve the equivalent target population to the census. These people are identified through linked migration data. Note this is different to the rules used within the APC, which includes all usual residents, regardless of whether they are currently in New Zealand or not.

A ‘sure’ admin resident population

Within the admin enumeration process we also use a more restricted version of the admin resident population. This ‘sure’ population requires people to have had activity in at least two admin sources (in addition to the rules describe above), with at least one of these sources being Inland Revenue or the Ministry of Health. One exception is for children aged 0, who are included even if they only have a single source of activity.

Requiring multiple sources of activity introduces under-coverage within the population, but also limits the possibility of individuals being included incorrectly.

Admin relationships

Admin family relationships are identified between individuals in the admin resident population, using information from the following sources:

- 2013 and 2018 Censuses
- DIA births, deaths, marriages, and civil unions
- Ministry of Business, Innovation and Employment (MBIE) visa applications
- MSD partnership and child data
- Working for Families (WFF) tax return details.

From each of these sources, we can extract relationship types between pairs of individuals. Some datasets contain a range of family relationship types (such as census data), while others contain specific relationship types (such as DIA Birth and Marriage registers).

We can also infer additional relationships based on the known relationships. For example, if two children share a parent, we infer they are siblings.

The resulting table includes the following relationship types:

- partner
- parent and child
- grandparent and grandchild
- great-grandparent and great-grandchild
- sibling
- pibling (aunt/uncle) and nibling (niece/nephew)
- parent-in-law and child-in-law
- other related (such as cousin or great aunt/uncle).

In partnerships, the status of the relationship and people involved can change over time. We use the start and end dates of different relationship events to select partners that are current on 7 March 2023.

Where we identify more than two parents for a child, we only use the latest two parents.

At multiple steps during this process, we identify and resolve inconsistencies. These inconsistencies include ensuring there are not multiple relationship types between a pair of individuals, and that age differences are as expected (for example, that a parent is not younger than their child). This helps us to maximise the quality and usability of the final output.

Linking census respondents to the IDI spine

2023 Census respondents are linked to the IDI spine. This link enables us to identify admin individuals who have not responded but are part of The admin resident population, and therefore eligible to be added as admin enumerations.

The census linking population includes all New Zealand usually resident records meeting the 2023 Census definition of a response at the time of linking. Overseas visitors are excluded, but absentee records (people listed as usual residents of a dwelling, but who are elsewhere on census night) that have not been linked to an individual form are included.

The IDI linking population includes all individuals in the IDI spine who were born up to 7 March 2023 and have sufficient information for a link to be made.

Both input datasets are first cleaned to remove invalid values and to ensure variables are in a consistent format. For example, additional whitespaces and most punctuation is removed from all name fields, imputed values are removed, and birthplace is converted to the two-character country classification.

Data is then processed through a series of passes, using an automated probabilistic method. Each pass uses a different combination of variables, including first and last names, date of birth and age, sex at birth, birthplace, and usual residence address. For each pass, a cut-off is manually set to determine which record pairs are considered a link. Records that meet the cut-offs within a given pass are considered a link and are excluded from subsequent passes.

For the purposes of linking, we assume all IDI records are unique, but that a small number of duplicates may still exist within 2023 Census responses. As such, each census response can only be linked to a single IDI record, although some IDI records may be linked to multiple census responses.

Overall, approximately 98 percent of census responses are linked to the 2023 Census. While some individuals will be incorrectly linked, or not linked when they should be, we estimate each of these errors to be less than 1 percent.

For detailed linking methodology information see Linking 2023 Census responses to the Integrated Data Infrastructure.

Admin usual residence address

For each individual in the admin resident population, we need to determine their usual residence address to be able to consider them for admin enumeration. For 2023, we have implemented an XGBoost machine learning model to identify the best admin address for each person on 7 March 2023.

The IDI contains address information from a range of admin sources, in addition to the 2013 Census, 2018 Census, and some household surveys. Raw address strings are linked to the Statistical Location Register (SLR), which contains unique address IDs and other information related to the addresses. After this linking, each address notification in the IDI then contains a person identifier, an address identifier, the source agency of the address notification, a notification date (generally representing the date this information was provided to a given agency), and additional geographic information about the address.

There are some limitations to this address information, which include delays in agencies being notified of a change in address, multiple concurrent (and sometimes contradictory) notifications, and the non-uniform distribution of notifications through time (which can be impacted by different reporting schedules across agencies). For

the 2023 Census, we attempt to account for alias addresses, so that all notifications relating to the same physical address are combined.

Our method for identifying an address for each individual has four main steps:

1. Produce a dataset of all person-address pairs from the available admin data.
2. Using an XGBoost model, produce a probability of correctness (defined as agreement with 2023 Census) for each person-address pair.
3. Choose the ‘best’ address for each person, namely the person-address pair with the highest probability.
4. Make an adjustment to align children’s addresses with their parents (in some scenarios).

The final output is a single address for each individual, along with a value relating to our confidence that this address is ‘correct’. We refer to this value as an ‘address score’. Overall consistency between 2023 Census addresses and the predicted admin address for linked census respondents is approximately 90 percent.

Predicting usual residence address from admin data in the 2023 Census describes the model in more detail.

Identification of respondents overseas on census night

When we have strong evidence that a census respondent was not in New Zealand on census night, we reclassify this individual as a respondent overseas on census night (ROCN).

Using links between census respondents and the IDI spine, we can use border movements to indicate whether an individual was in New Zealand or overseas.

We only reclassify individuals as ROCNs when there is evidence of a border movement in the six months either side of census night. This reduces the risk of incorrectly removing people due to incorrect links to, or within, the IDI. While six months is a subjective cut-off, this largely results in excluding those whose most recent border movement was years earlier. In these cases, we are not confident we can trust the border movement data, and therefore err on the side of retaining census respondents in the census usually resident population.

Admin enumeration methodology

This section describes the methodology for including admin enumerations in the 2023 Census dataset. These steps use the outputs from the pre-admin enumeration processes described in the previous section, as well as additional 2023 Census and admin data.

In line with the guiding principles, we have different quality criteria for different types of admin enumeration. Most of the described methods involve adding admin enumerations to a census dwelling, and so we need to consider how these admin enumerations affect counts of individuals in those dwellings, as well as family and household outputs.

Every individual (unit) in the final 2023 Census file contains an indication of their data source. This unit data source identifies whether the individual was added through admin enumeration, and if so, the specific type of admin enumeration.

Appendix 2 contains the unit data source classification for the 2023 Census.

Table 2 provides an overview of these processes, including changes from the 2018 Census. There is no overlap between the different types of admin enumeration; an individual cannot be included through multiple different steps. Note that members of the admin resident population will not be included as admin enumerations through any of these steps if they do not meet any of the criteria described.

The rest of this section then provides additional detail about each of the described admin enumeration processes.

Table 2: Summary of admin enumeration methodology in the 2023 Census

Process	Description	Changes from 2018 Census
Add admin enumerations for absentees in responding private dwellings	Admin enumerations are included when an absentee record is not linked to a 2023 Census Individual Form but can be linked to an admin record. Those identified as being in New Zealand on census night are included as admin enumerations.	New for 2023. Enables more admin enumerations to be included in dwellings.
Add admin enumerations in responding private dwellings	Admin enumerations are included in responding private dwellings when we have strong evidence that someone has been missed from the household. In line with guiding principle 4, we are more cautious including individuals in responding dwellings than in other scenarios, as this risks contradicting information provided by census respondents.	Extended method for identifying evidence that an individual has been missed
Add admin enumerations in non-responding private dwellings	Admin enumerations are included in non-responding private dwellings when we have a high degree of confidence the combination of admin individuals at a given address are part of a single household.	• Use of an improved machine learning model • Increase in eligible household size from 8 to 15 • Extension to include households in dwellings not originally identified
Add admin enumerations in non-private dwellings (prisons)	Admin enumerations are included in prisons using information provided by the Department of Corrections.	N/A
Add admin enumerations in meshblocks	Admin enumerations are included in meshblocks when they cannot be included in a census dwelling, but we still have confidence they are part of the usually resident population and that their admin address is reliable.	Use of improved machine learning model for determining usual residence address

Process	Description	Changes from 2018 Census
Add admin enumerations in non-private dwellings (other)	Admin enumerations are included in other non-private dwelling types (for example, a residential and community care facility or motor camp) when their best admin address is associated with a non-private dwelling.	• Removal of admin enumerations in defence establishments due to quality concerns • Inclusion of admin enumerations in other NPD types

Admin enumerations for absentees in responding private dwellings

For the 2023 Census we have expanded the census population linked to the IDI to include absentee records. These are people listed on a response as usually living at the dwelling, but who are elsewhere on census night. This step is applied after absentee records are linked to individual forms (a process referred to as repatriation), with only remaining unrepatriated records included in the linking to the IDI.

As these absentees are already listed on a household set-up or dwelling form, we have a high degree of confidence they usual reside within the associated dwelling. We therefore prioritise this step over subsequent admin enumeration categories.

The linking itself occurs during the overall linking census responses to the IDI described above, with absentee records linked at the same time as other usual residents.

We then use border movement information from the IDI to determine the final record type of any linked absentee records:

- For absentees (without an individual form) linked to the IDI and determined **to be** in NZ on census night, we create an admin enumeration record and classify them as an admin-linked absentee in their listed dwelling. These records are therefore able to be included in the census usually resident population count.
- For absentees (without an individual form) linked to the IDI and determined **not to be** in NZ on census night, we code these records as respondents overseas on census night. As with other census records, admin attributes are extracted, and the individuals can contribute to a family and household, where relevant, but they are not included in the census usually resident population count.
- For absentees not linked to the IDI, they remain as an absentee record.

Admin enumerations in responding private dwellings

Admin enumerations are added into responding private dwellings when we have sufficient evidence to believe an individual belongs to a given dwelling and has not responded to the 2023 Census.

An admin record related to an individual is in scope only at their predicted admin address, and only when that admin address corresponds to a dwelling that includes at least one usual resident who has provided a census response. We exclude cases where an address has more than one dwelling.

As this step involves dwellings that already include census responses, we require strong evidence that an admin individual should be included (as per principle 4).

Identifying evidence of individuals missing from responding private dwellings

We begin by identifying evidence that individual(s) could be missing from a responding census dwelling. Two main approaches are used.

The first approach (referred to as ‘unfulfilled living arrangements’) uses ‘living arrangements’ and ‘relationship to reference person’ to identify households where at least one individual has more living arrangements listed than there are other individuals in the household.

The second approach (referred to as ‘strong admin evidence’) is initially based solely on admin data. It identifies admin individuals who have not responded to the census, are related to a census respondent at their predicted admin address, and have a high admin address score (as calculated when predicting the best admin address).

Comparing information from census and admin households

We then compare the information from the admin and census households to identify where there is consistent information that an individual could have been missed. Census households are defined as the combination of usual residents at a given address who have been identified through census responses. Admin households are defined as the combination of individuals in the admin resident population at a given address.

From the unfulfilled living arrangements approach, we compare the specific missing living arrangements with the admin relationships identified between census respondents and admin individuals. An example of a perfect match is when a census respondent indicates they live with a sibling and no matching census respondent can be found, but there is a sibling identified in the admin household. Where there is a

perfect match, we are confident enough to include the matching admin record as an admin enumeration.

For the strong admin evidence approach, we first consider the admin address scores for each person. We do not necessarily have direct information from a census response that they have been missed, and, as per principle 4, we are more cautious of changing existing census households. Although we want to add as many records as possible, we also do not want to lower the quality of certain types of households. We therefore choose to set a reasonably high barrier for an admin enumeration to be added through this approach, and limit potential inclusions to those with an address score greater than 0.85.

This threshold was based on comparisons with available Household Labour Force Survey (HLFS) data, as well as assessments of the type of individuals and households that would be included. This assessment indicated that reducing the address score threshold would risk too many people being included incorrectly, while increasing the address score threshold would lead to a less representative group of people being added (primarily by including higher proportions of those with European ethnicity than with other ethnicities). This choice aligns with principle 2, in that we do not want to introduce bias to family or household outputs through our choice of thresholds.

After setting this threshold, we finally examine whether the inclusion of an individual through this approach would lead to inconsistencies for the overall census household. Specifically, we exclude individuals when:

- they have the same sex at birth, year of birth, and month of birth (or sex and age if full date of birth is missing) as a census respondent in the same household
- they are admin partners with a census respondent who already has a partner
- they are admin parents of a census respondent who already has two parents
- they are being added to a one-person household where the census respondent stated that they live alone
- they are admin partners with a census respondent who stated they do not live with a partner
- their relationships to census respondents are all directly contradictory to provided living arrangements.

It is possible that some individuals removed through these rules are valid members of the household. However, on balance, we determined that their inclusion would be more likely to lower the quality of their associated household. With the exception of

any records identified as being likely duplicates of census respondents, records removed through these rules will still be eligible to be included as meshblock enumerations.

All individuals remaining after applying the above criteria are included as admin enumerations. Family and household information for affected households is then updated to incorporate these added individuals.

Admin enumerations in non-responding private dwellings

Admin enumerations are included in non-responding private dwellings when we have evidence of a high-quality admin household at the relevant address. A household is defined as a group of people who usually live in the same private dwelling.

All non-responding private dwellings were in scope, regardless of whether census collectors classified the dwelling as occupied or unoccupied on census night.

Step 1: Create admin households and census households

We create admin households as the combination of all individuals in the admin resident population with the same usual residence admin address. A range of information is extracted, including:

- characteristics of the individuals in the household (eg, age, ethnicity)
- characteristics of the household (eg, size, type, number of adults/children)
- the minimum address score for individuals within the household and the mean address score across the household
- family relationships between individuals, which are used to derive further characteristics of the household (eg, whether the household contained a family or unrelated individuals, whether an individual's partner lived at the same address)
- characteristics of address notifications for individuals in the household (eg, whether any of the individuals in the household had a non-matching address in a given address source).

There is some overlap with the features described in Admin usual residence address above, but also an additional focus on information that may relate to the overall household.

We also create census households as the combination of all census respondents with the same usual residence census address. We use these as 'true' households,

and we use a model based on these census households and their admin data to judge our confidence in potential admin households.

Step 2: Create separate datasets of households for modelling and enumerating

The admin and census households from step 1 are combined based on the shared address identifiers. Some admin households link in this way to a census household, while other admin households are at addresses without census responses. We then split the data into the following datasets.

- Responding households: admin households at private dwellings that responded to the census. These are further split randomly as follows:
  - training (80 percent of responding households): these are used to train the model
  - validation (20 percent of responding households): these are used to validate the model.
- Subject (households in scope for enumerating): admin households at private dwellings that did not respond to the census. For quality reasons, we restrict eligible admin households to those with a maximum of 15 household members and exclude addresses with multiple dwellings and exclude households that do not contain an individual aged 18 or above. Individuals at an admin address who are identified as having responded to the census with a different usual residence address are removed; household covariates are re-derived for the households that change because of this step.

Step 3: Carry out modelling to predict the probability that an admin household is ‘high quality’

We then carry out XGBoost modelling for the training households (defined in step 2) to build a household model that predicts the probability of an admin household being ‘high quality’. High quality is defined as:

> *At least half the members of the admin household match the members of the census household, and the admin household composition is the same as the census household composition*

where household composition is one of the following categories:

- one adult and no children
- one adult and one or more children
- two adults and no children
- two adults and one or more children

- three or more adults and no children
- three or more adults and one or more children.

This choice of criteria recognises that some slight differences in the household membership may be acceptable if the overall type of household is similar. This also provides more of an opportunity for larger households to be included.

The model results in a household probability for each admin household, where:

$$p\_hhld = Prob(admin\ household\ is\ high\ quality)$$

The performance of the model is assessed using the validation dataset and several criteria, including accuracy, precision, specificity, and sensitivity.

Although a large number of covariates are included in the final model, three are particularly useful in being able to differentiate between high-quality and low-quality admin households:

- the number of adults in the household
- whether or not the household contains at least one person aged 65 or above
- the mean address score for the household.

Step 4: Carry out ‘test’ enumerations to determine the threshold for household enumerations

We use the validation dataset to investigate different household probability thresholds for enumerating whole households of individuals. Since the validation dataset contains responding households, we can compare the quality of admin households with 'true’ census households. For different household probability thresholds, we ‘enumerated’ the individuals in all households with a household probability above the threshold and compared the quality of enumerations for the different thresholds. Our aim is to enumerate as many households as possible, but to also maintain confidence that the particular households being included are likely to be correct.

As a result of this investigation, we selected a household probability threshold of 0.5. This is the point at which an admin household is more likely than not to be of high quality. This is also the threshold which resulted in the highest accuracy of test enumerations (the proportion of high-quality households enumerated or low-quality households not enumerated). This also aligns with the criteria implemented in the 2018 Census. Appendix 3 contains more information about the distribution of households by household probability.

This threshold represents a balance between including more admin enumerations in households and having confidence that a given household is correct (as described in principles 3 and 5). The value from including individuals in households is considerable, and therefore we are willing to accept some differences in precise household membership.

We use comparatively lower thresholds for these admin enumerations in non-responding dwellings than we do for admin enumerations in responding dwellings. This is because there is not the same risk of altering 2023 Census responses – here we are choosing between having some household information or no household information.

Step 5: Use model on non-responding private dwellings and carry out household enumerations using chosen threshold

We then use the chosen model on the ‘subject’ dataset, which is all valid admin households identified at non-responding private dwellings. The model applied to the data provides us with a household probability for each valid admin household.

All households with a probability greater than the selected household probability threshold of 0.5 are included as admin household enumerations; each of the associated individuals is included as an admin enumeration, with that address as their usual residence address.

If the census dwelling was originally classified as occupied non-responding, these admin enumerations are classified as ‘admin enumeration – occupied, non-responding’. Otherwise, these admin enumerations are classified as ‘admin enumerations – unoccupied, residents away’.

Admin enumerations in non-private dwellings (prisons)

Admin enumerations are added into prisons based on information provided by the Department of Corrections.

Data provided by the Department of Corrections is first linked to the IDI, using shared unique identifiers. This enables links to other data sources within the IDI, as well as to 2023 Census respondents.

Using these links, we first remove individuals who have already responded to the census. We also remove individuals in the Corrections dataset who were not able to be linked to the IDI spine. For these individuals we are not able to identify any other attributes from admin data and cannot determine whether they have already responded to the census.

All remaining individuals in Corrections data are added as admin enumerations, with their census night address set to the prison listed within the supplied data. No attribute information other than census night address is directly sourced from Department of Corrections data.

Determining usual residence address for admin enumerations in prisons

Not all individuals who were in a prison on census night will also be considered a usual resident of the prison. We have two alternatives for the usual residence of these admin enumerations:

1. the prison they were in on census night
2. the address predicted from other admin data sources in the IDI.

A small number (less than 1 percent) of records have no other available admin address. Some individuals also have a predicted admin address matching the prison where they were located on census night. However, for most individuals these two addresses will differ.

For each prison, we calculate the proportion of 2023 Census respondents who also listed that prison as their usual residence address. This proportion is then applied randomly to all admin enumerations within that same prison, in order to roughly preserve the distribution indicated through census responses. Remaining admin individuals are assigned their predicted admin address as their usual residence.

For individuals who are assigned to a usual residence that is not a prison, they will be included in individual counts in the associated usual residence area, but will not be attached to a specific dwelling. This means they will not be included in any family or household outputs.

Admin enumerations in meshblocks

We attempt to include admin individuals who did not respond to the census, and who did not meet the criteria for previous admin enumeration processes, as admin meshblock enumerations.

These meshblock admin enumerations are not placed into a specific dwelling and therefore do not contribute to family or household outputs. While they are less desirable than other types of admin enumeration, they still contribute to improving the count of individuals, overall and for different sub-populations.

For meshblock admin enumerations we reuse the address scores described above in admin usual residence address. These values already represent our confidence

that a given address is correct, although the scores should not be interpreted directly as probabilities.

Step 1: Create dataset of individuals in scope for meshblock enumeration

We first identify individuals in scope to be admin enumerations. This includes everybody who is in the admin resident population, has not responded to the census, and has not already been included as a different type of admin enumeration.

Step 2: Apply adjustment for over-coverage

Over-coverage exists due to some non-residents being erroneously included in the admin resident population or due to incorrect linkages in the IDI (and therefore the admin resident population). To mitigate this, we apply a stricter condition for potential inclusion as meshblock enumerations.

This adjustment reduces the number of records available for admin enumeration, but it means we can be more confident they are genuinely in the target population of the census.

The adjustment requires individuals to have activity in two or more admin data sources over the previous two years, rather than just one. Anyone not meeting these criteria was ineligible to be included as a meshblock enumeration. This has the effect of removing people who could have been incorrectly linked to the spine in a single activity source, as well as reducing the chance of people being included who were incorrectly classified as residents.

Step 3: Apply adjustment for missed links between census and the IDI spine

Occasionally an individual's census response does not link to an admin record in the IDI when it should. In that case, the individual will potentially be counted twice (as both a census respondent and an admin enumeration). We cannot identify these individuals with missed links; any feasible method would be part of the linking methodology. However, we can estimate how many individuals in the admin resident population should have been linked to census responses.

We apply an adjustment to account for missed links between census respondents and the IDI spine. We first identify census respondents we are highly confident exist in the IDI spine (for example, if they say they were born in New Zealand and earned income in the previous 12 months). We then use those who were not linked as an estimate of the number of missed links. Dual system estimation combining census responses and an admin population describes this adjustment in more detail.

We disaggregate this overall proportion into proportions of missed links for sub-populations based on age, ethnicity, and geographic location. We apply these proportions to the corresponding sub-populations in the admin resident population to estimate the total number of individuals who were missed links. Finally, we randomly select the required number of individuals from each sub-population of those in scope for meshblock enumerations and exclude them from being enumerated.

Note that this adjustment does not identify the specific records that are missed links, and therefore likely duplicates. Instead, it aims to remove people with similar demographics to preserve the expected distribution. Using more disaggregated counts may slightly increase the chance of removing the ‘correct’ record, but also leads to combinations with small counts of individuals, which can result in implausible estimates of missed links.

Step 4: Carry out test enumerations to determine the threshold for meshblock enumerations

To determine an address score threshold above which non-responding individuals not yet enumerated will be included as meshblock enumerations, we next carry out test meshblock enumerations using different thresholds and compare the quality of the enumerations for each threshold. Individuals in scope for test meshblock enumerations are those in the validation dataset who were not enumerated as test household enumerations (because they were either out of scope for household enumerations or were below the threshold for being enumerated). Since the validation dataset contains responding individuals, we know whether an individual’s admin meshblock is correct (matches their ‘true’ census meshblock) or incorrect. Our aim was to enumerate as many individuals as possible, but to be reasonably confident that the individual’s meshblock was correct.

We then use the individuals in the validation dataset who were not enumerated as test household enumerations (because they had low household probabilities or were otherwise out of scope for household enumeration) in the previous section on non-responding private dwellings. We investigate the enumerations resulting from different address score thresholds. Since the validation dataset contains responding individuals, we can compare individuals’ admin meshblocks with their 'true’ census meshblocks. In this testing exercise, we enumerated all individuals with an address score above a variety of thresholds and compared the quality of meshblock enumerations for each of these different thresholds. Our aim was to enumerate as many individuals as possible, but to be reasonably confident that the individual’s meshblock was correct.

As a result of this investigation, we selected an address score threshold of 0.4. This was the point at which an individual’s admin meshblock was more likely to be correct than incorrect. It was also the address score at which the accuracy of enumerations

was maximised (the proportion of individuals with a correct meshblock enumerated or with an incorrect meshblock not enumerated). This is the same criteria as was used in the 2018 Census.

As with household enumerations, and as described in principle 5, this threshold leads to a balance between including more admin enumerations and having confidence that a given individual is placed in the correct location. A more conservative threshold would exclude individuals who are likely to have the correct location, while a more lenient threshold could lead to less accurate aggregate counts for sub-populations.

Although we acknowledge that some individuals will not be included in the exact right location, our analysis indicates that a threshold of 0.4 results in reliable aggregate counts at different geographic levels. We also checked that the address score threshold of 0.4 did not adversely affect sub-populations in terms of accuracy or quantity of enumerations, in particular, Māori or Pacific ethnicities. This provides evidence that our choice of threshold is aligned with principle 2 and reduces any impacts of differing response rates across sub-populations. Appendix 4 provides more detail about this choice of threshold.

Step 5: Carry out meshblock enumerations using the chosen threshold

We then apply the selected address score threshold to all in-scope admin non-responding individuals. All individuals with an address score greater than the selected threshold of 0.4 are included as admin meshblock enumerations.

Admin enumerations in non-private dwellings (other)

This section describes the inclusion of admin enumerations in additional non-private dwelling (NPD) types. This approach uses the method discussed above for admin meshblock enumerations.

For any people who would otherwise be included as meshblock enumerations and whose admin address is associated with an NPD, we considering adding them into that NPD, rather than leaving that individual with no usual residence dwelling information.

We include these records when the address corresponds to an NPD and:

- the dwelling type is not a prison or commercial vessel
- there is only a single dwelling at the address
- the dwelling is identified as being occupied on census night.

In all other cases, we retain the individual as a meshblock enumeration instead.

Resulting admin enumeration inclusions and exclusions

The admin enumeration methodology described above can be summarised as beginning with the admin resident population, then applying a set of steps that either exclude certain records from consideration, or include records as admin enumerations in the final census dataset. Table 3 summarises these steps in the order they occur, tallying how many records are included or excluded at each step.

In total, approximately 206,000 records are excluded from being admin enumerations through the adjustments for over-coverage, missed linkages, and low address scores.

Note these counts will not sum to census counts because some census respondents are not linked, or are linked to individuals outside of the admin resident population.

Table 3: Summary of admin enumeration inclusions and exclusions

Step	Count		
	Included as admin enumeration	**Excluded**	**Remaining admin records**
Total admin resident population at 7 March 2023			5,098,095
Exclude: residents temporarily overseas		82,896	5,015,199
Exclude: records linked to 2023 Census response[1]		4,265,442	749,757
Include: records linked to 2023 Census absentee[1]	45,513		704,244
Include: admin enumerations in responding private dwellings	30,279		673,965
Include: admin enumerations in non-responding private dwellings	229,122		444,843
Include: admin enumerations in prisons[1]	4,344		440,499
Exclude: admin resident population over-coverage		76,992	363,507
Exclude: adjustment for missed linkages		65,439	298,068
Include: admin enumerations in meshblocks	228,204		69,864
Include: admin enumerations in other NPDs	6,255		63,609
Exclude: poor quality records (including address score below cut-off)		63,609	0

Source: Stats NZ

1. Some census records in these categories are not part of the admin resident population, so counts are lower than shown elsewhere.

Admin enumeration results

This section provides a brief overview of results from the admin enumeration process. We focus on the number of records included through each admin enumeration step, and the distribution of census responses and admin enumerations across sub-populations.

Table 4 shows the number of records included from each source in the 2018 and 2023 Censuses. In both censuses, census responses make up 89 percent, and admin enumerations 11 percent, of the final datasets.

The proportion of admin enumerations associated with a dwelling has increased in 2023, from 32 percent to 58 percent. This is driven by a range of factors, including:

- the introduction of the admin-linked absentee category. In the 2018 Census, it is likely most equivalent individuals were included as meshblock enumerations.
- improvements to the methodology for identifying family relationships, predicting usual residence address, and estimating the quality of households, leading to more admin enumerations within both responding and non-responding private dwellings.

Table 4: 2018 and 2023 Censuses usually resident population counts, by unit data source

Unit data source	Count		Percent	
	2018	2023	2018	2023
Census individual form	3,971,892	4,323,579	84.5	86.6
Individuals on the household listing only	202,914	123,849	4.3	2.5
Field enumerated rough sleeper[1]	99	-	<0.1	-
Census responses	**4,174,902**	**4,447,428**	**88.8**	**89.1**
Admin enumeration (admin-linked absentee)	-	48,171	-	1.0
Admin enumeration (occupied, responding)	20,643	30,279	0.4	0.6
Admin enumeration (occupied, non-responding)[2]	99,159	175,926	2.1	3.5
Admin enumeration (unoccupied, residents away)[2]	42,252	53,196	0.9	1.1
Admin enumeration in private dwelling	**162,054**	**307,569**	**3.4**	**6.2**
Admin enumeration (non-private dwelling, prison)	4,707	4,464	0.1	0.1
Admin enumeration (non-private dwelling, defence)	798	-	<0.1	-
Admin enumeration (non-private dwelling, other)	-	6,255	-	0.1
Admin enumeration in non-private dwelling	**5,505**	**10,719**	**0.1**	**0.2**
Admin enumeration (meshblock usual resident)	357,294	228,204	7.6	4.6
Admin enumeration in meshblock	**357,294**	**228,204**	**7.6**	**4.6**
Total	**4,699,755**	**4,993,923**	**100.0**	**100.0**

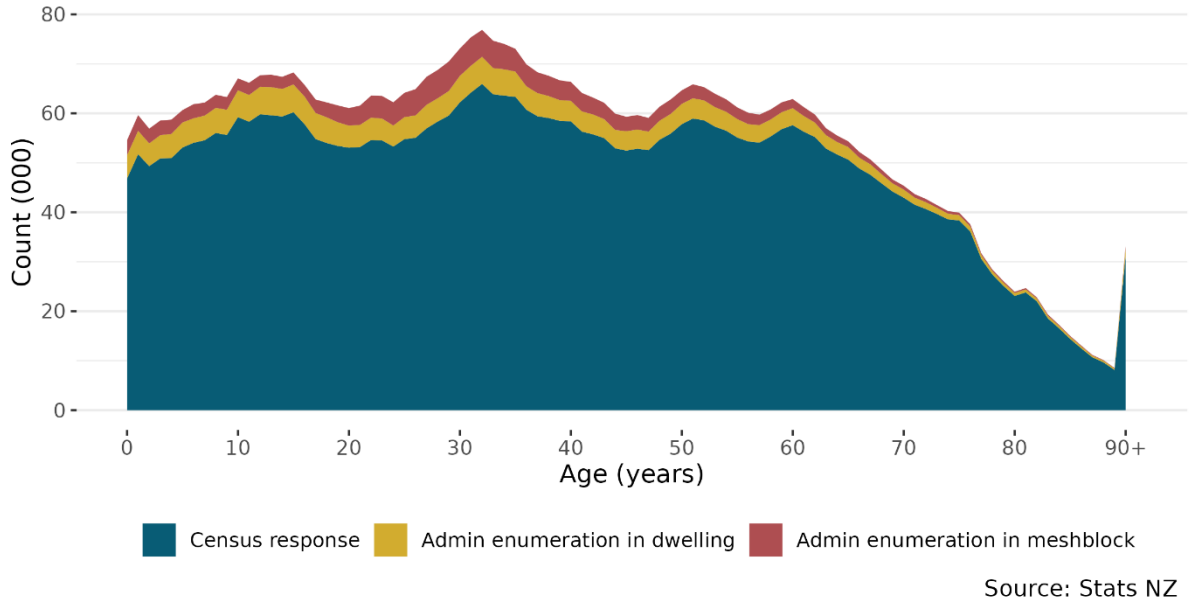
Source: Stats NZ

1. This category was not used in 2023 Census. Similar records are instead classified through other census response categories.
2. These both contribute to admin enumerations in non-responding private dwellings.

The remaining results use the same groupings as table 4, with records categorised as census responses, admin enumerations in dwellings, or admin enumerations in meshblocks.

Figure 1 shows the distribution of census responses and admin enumerations by age. By broad age group, people aged 15 to 29 years have the highest relative proportion of admin enumerations (14 percent), while those aged 65 and older have the lowest proportion of admin enumerations (5 percent).

Figure 1: 2023 Census usually resident population count, by age and unit data source

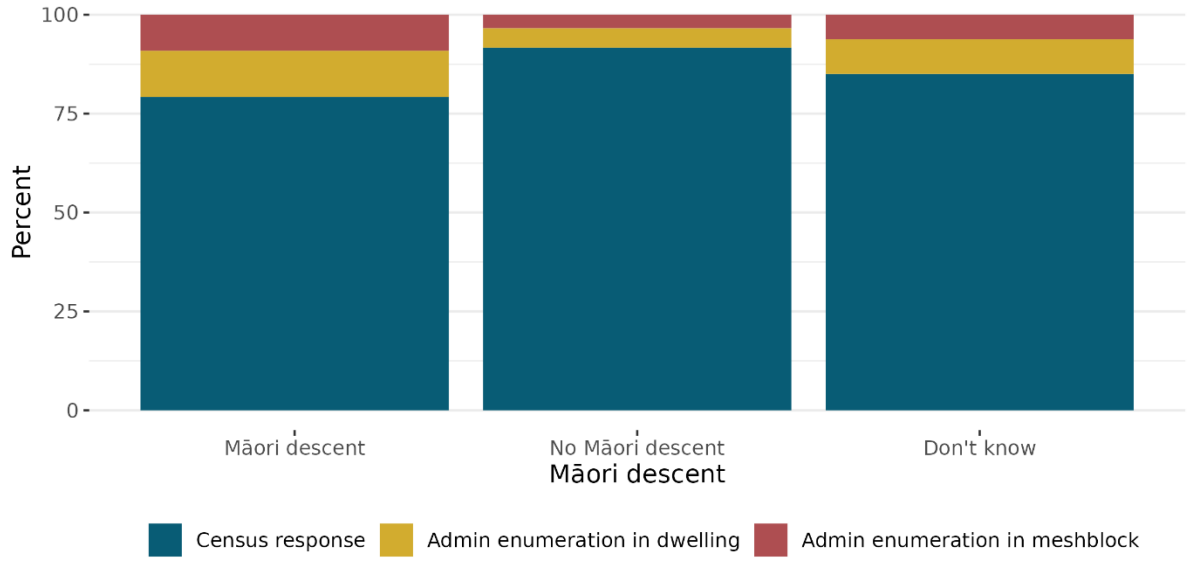


Source: Stats NZ

Figure 2 shows the distribution of census responses and admin enumerations by the Māori descent indicator, and figure 3 shows the distribution of census responses and admin enumerations by level 1 ethnic group.

Approximately 21 percent of people of Māori descent are admin enumerations, compared with 8 percent of those not of Māori descent, and 15 percent of those who do not know if they are of Māori descent.

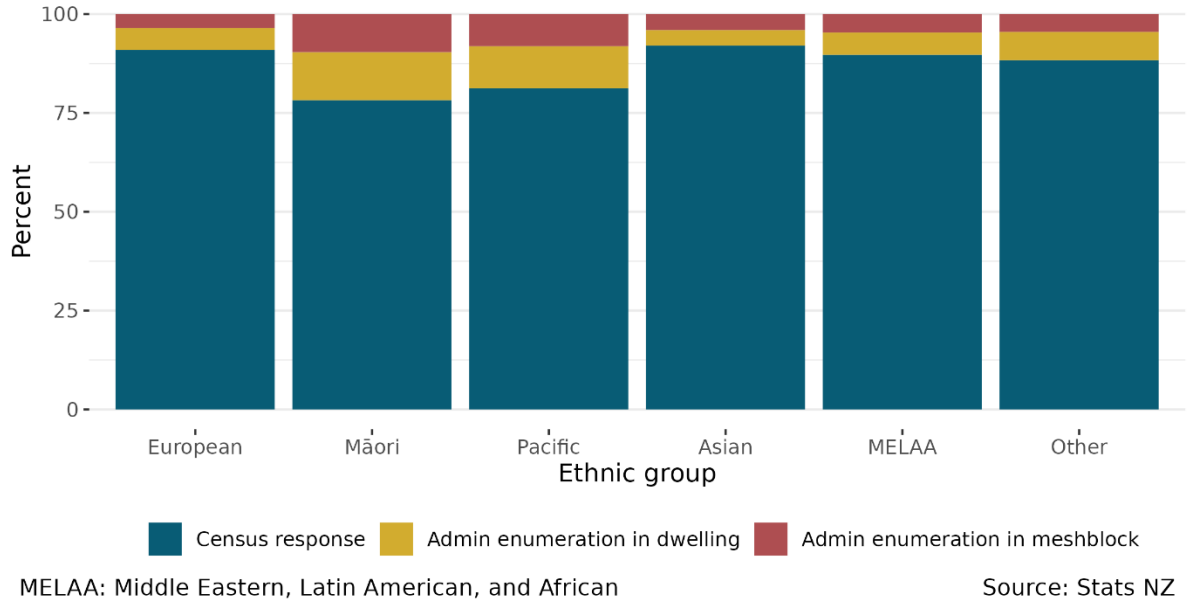
Figure 2: Percentage of population by Māori descent and unit data source



Source: Stats NZ

For ethnic groups, approximately 22 percent of the Māori population are admin enumerations, compared with 19 percent for Pacific peoples, 9 percent for European, and 8 percent for Asian.

Figure 3: Percentage of population by level 1 ethnic group and unit data source



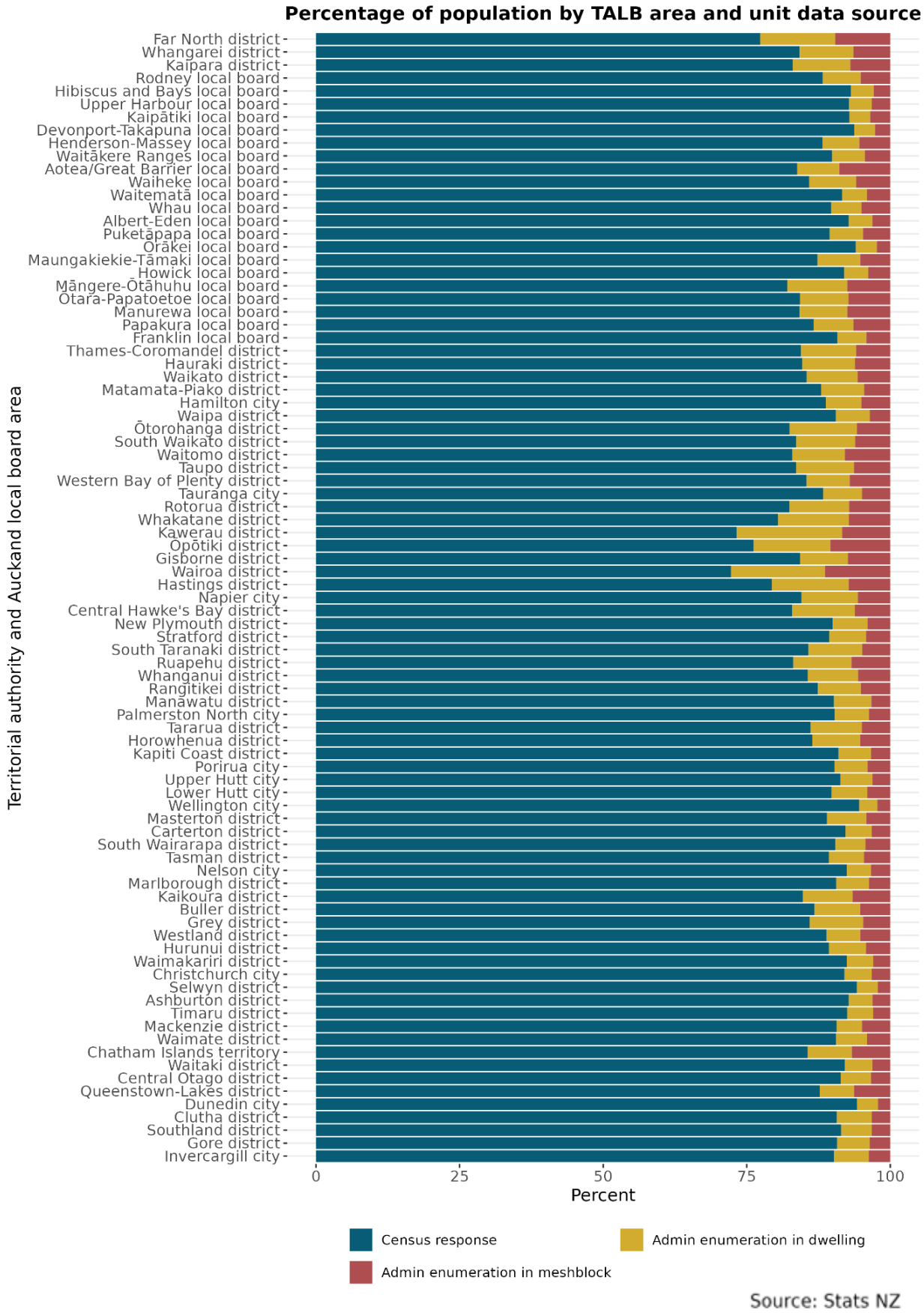
MELAA: Middle Eastern, Latin American, and African

Source: Stats NZ

Figure 4 shows the proportion of admin enumerations across territorial authority and Auckland local board (TALB) areas. Of all territorial authority areas, Wairoa district (28 percent) and Kawerau district (27 percent) had the highest proportion of admin enumerations, while Wellington city (5 percent) and Dunedin city (6 percent) had the lowest proportion.

Early 2023 involved several weather events in the North Island, with large flooding events followed by Tropical Cyclone Gabrielle. Certain areas, including the North and East Coast of the North Island, were heavily affected by these events, and we can see evidence of relatively more admin enumerations in these areas compared with regions that were less affected.

Figure 4: Percentage of population by TALB area and unit data source



Discussion

The 2023 Census builds on the model first implemented in the 2018 Census. Admin records are combined with census responses to produce counts of the New Zealand usually resident population. Combining these two sources means we can produce population counts that have higher coverage, and more consistent coverage across sub-populations, than would be possible from census responses alone. Because response rates differ across populations, higher proportions of admin enumerations are visible across some age groups, ethnic groups, and geographic areas.

We use a range of methods to ensure as many people as possible can be included, and included in the correct place. Different considerations are required depending on whether they are at a private dwelling or a non-private dwelling, and the level of response received for that dwelling. In general, we follow the same framework as in 2018 by balancing the importance of correctness at an individual level with the impact on aggregate census counts. That is, some error is considered acceptable at unit record level if it still results in more useful outputs overall.

Although the 2018 Census admin enumeration methodology was applied in a reactive manner, additional development for the 2023 Census has validated many of the decisions made in 2018. The overall framework and the criteria for determining when admin enumerations should be included have remained broadly consistent.

We have implemented some improvements for the 2023 Census. Absentee records have been linked to the IDI, allowing them to be included as admin enumerations. Improvements to the methods for deriving admin addresses and relationships mean we can have more confidence in the information about the individuals and their associated households. We have also expanded the criteria for including admin enumerations in both responding and non-responding private dwellings. As a result, more admin enumerations are included in dwellings, which means they can contribute towards family, household, and extended family outputs.

However, the current model has limitations. Census responses and the admin resident population measure slightly different concepts. We must account for this when combining the two populations. For the 2023 Census, responses take priority, which means we apply certain rules to reduce the risk of including too many people. Alternative census models might apply different priorities to the available data sources. Further improvements to the data and methods will continue to provide more options in combining the available data sources.

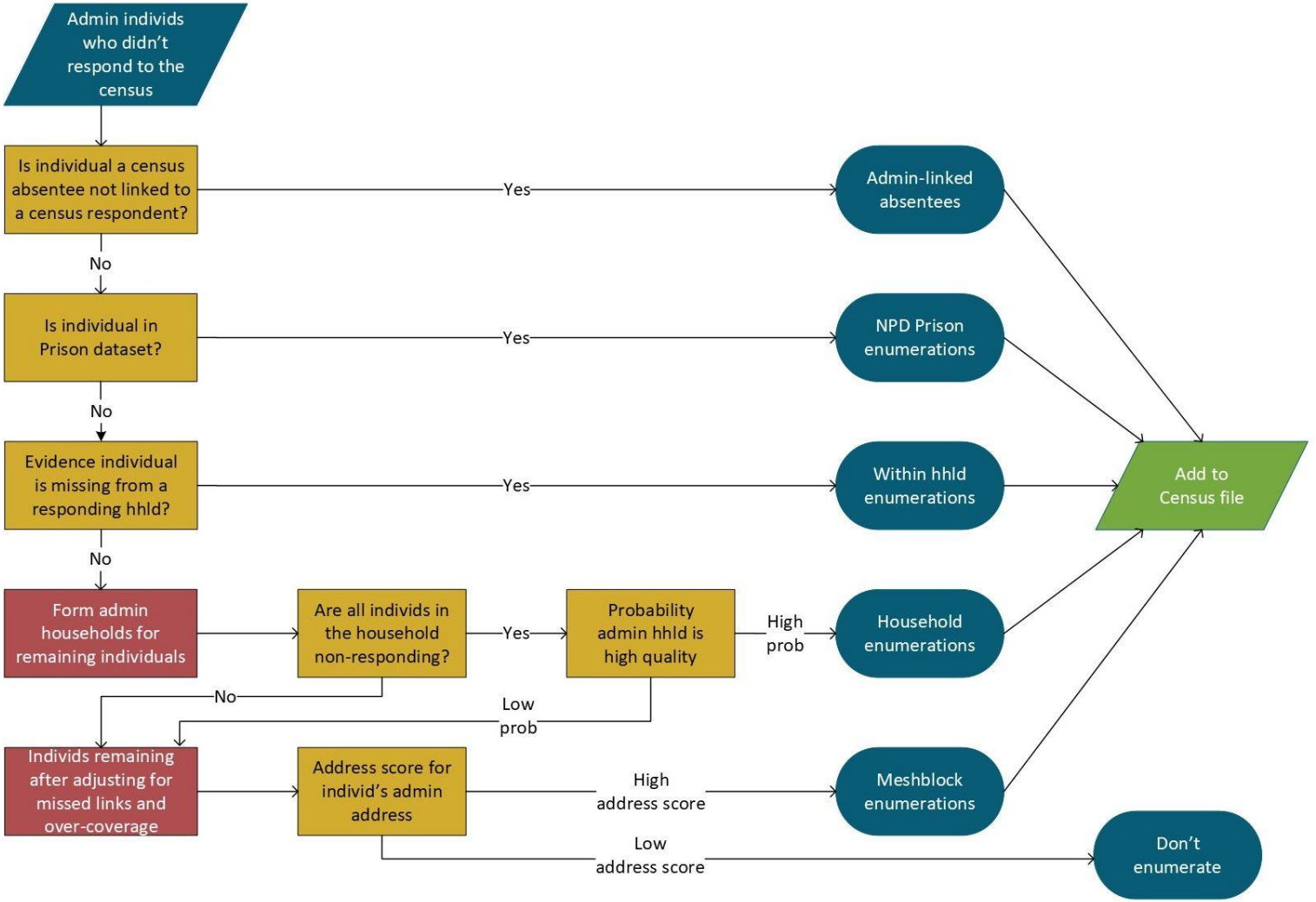
References

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

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

Appendix 1: Flowchart of admin enumeration process

Figure 5: Flowchart of admin enumeration process



Note: individ - individual; hhld - household; NPD - non-private dwelling; prob - probability

Appendix 2: Unit data source classification

Table 5: Unit data source classification

Census individual unit data source classification	
Code	**Descriptor**
111	Census individual form
121	Individuals on the household listing only
211	Admin enumeration (occupied, non-responding)
213	Admin enumeration (non-private dwelling, defence)
214	Admin enumeration (unoccupied, residents away)
215	Admin enumeration (occupied, responding)
216	Admin enumeration (meshblock usual resident)
217	Admin enumeration (admin-linked absentee)
218	Admin enumeration (non-private dwelling, other)
Source: Stats NZ Unit data source	

Appendix 3: Thresholds for admin enumerations in non-responding private dwellings

This section provides more information about the choice of thresholds for admin enumerations in non-responding dwellings. We first look at test households (those present in both 2023 Census and admin data), then assess the impact on households that did not respond to the 2023 Census.

Test enumerations

To determine the threshold for household enumerations, we carried out test enumerations using responding 2023 Census households.

A household probability threshold of 0.5 was chosen for two reasons. This was the point at which an individual admin household was more likely to be high quality than low quality (see figure 6). It was also the household probability threshold that maximised the accuracy of test enumerations (the proportion of high-quality households enumerated plus low-quality households not enumerated). While we considered a range of other quality metrics, accuracy was ultimately determined to be the most relevant for balancing the guiding principles described above.

Figure 6: Number of test households, by house probability and agreement between admin and census households

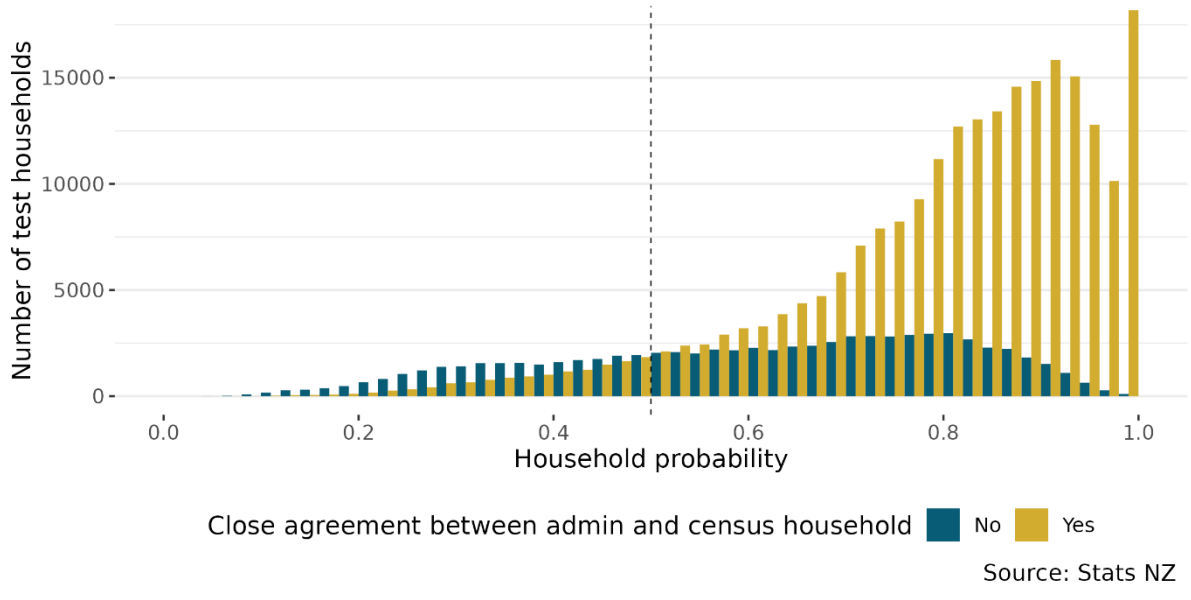


Table 6 shows the number of test households and individuals that would be enumerated with a household probability threshold of 0.5. Around 88 percent of test households were enumerated and, of these, 81 percent were high-quality households. In terms of individuals in the test households, 88 percent were

enumerated and, of these, 94 percent had the correct admin address. The higher precision for individuals indicates that producing an entire household is more difficult than getting a single address correct.

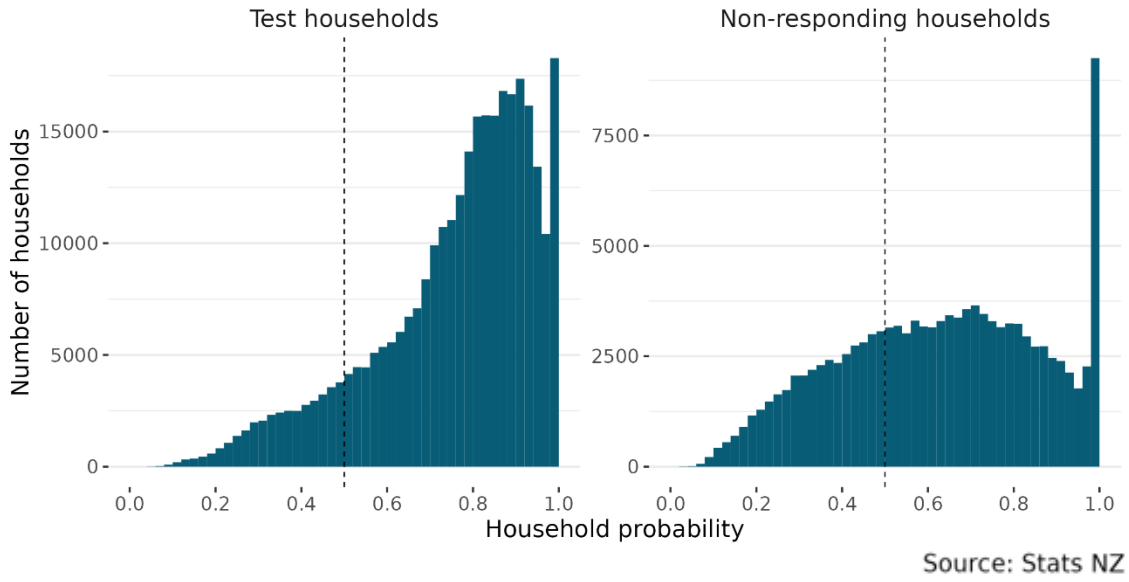
Table 6: Test household enumerations using a household probability threshold of 0.5

Measure	Percentage
Percentage of test households enumerated	*88.0*
Of the above, percentage of admin households that were high quality	80.8
Percentage of test individuals enumerated	*87.6*
Of the above, percentage of individuals with the correct admin address	94.3

Actual enumerations

It should be noted that responding and non-responding households may differ. Figure 7 shows that the household probability distribution for responding (test) households is different from that for non-responding households. The test households tend to have higher household probabilities than non-responding, although in both cases there is a spike at probabilities close to 1.

Figure 7: Number of test and non-responding households, by household probability



While we cannot know the true quality of non-responding households, these household probabilities represent our best estimate of this quality. Using the selected 0.5 threshold, 69 percent of eligible households and 79 percent of eligible individuals are included as admin enumerations. Both of these are lower than the numbers for

test households above, which reiterates that non-responding households tend to be harder to measure correctly through admin data.

Figures 8 and 9 show the percentage of eligible individuals who are included as admin enumerations in non-responding households, by level 1 ethnic group and by age. In both cases, the dashed line represents the overall percentage enumerated. The percentages are relatively similar across ethnic groups, with slightly more people of European ethnicity, and slightly fewer people of Māori ethnicity, enumerated. By age, older age groups have the highest relative proportion that are enumerated. People in these age groups tend to be less reliable, have higher quality admin address information, and to be in smaller households.

Figure 8: Percentage of eligible individuals included as household enumerations, by level 1 ethnic group

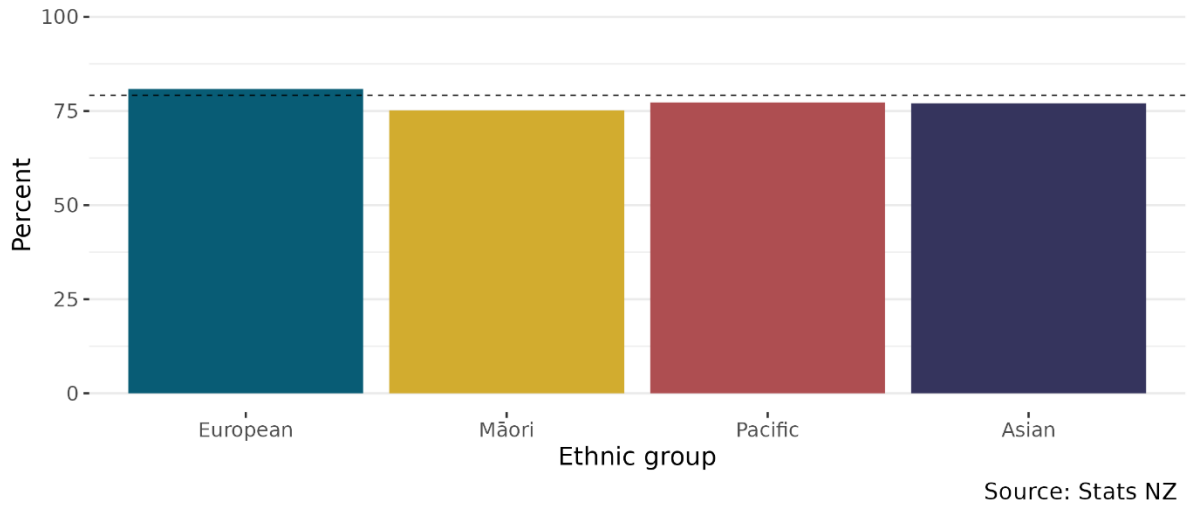
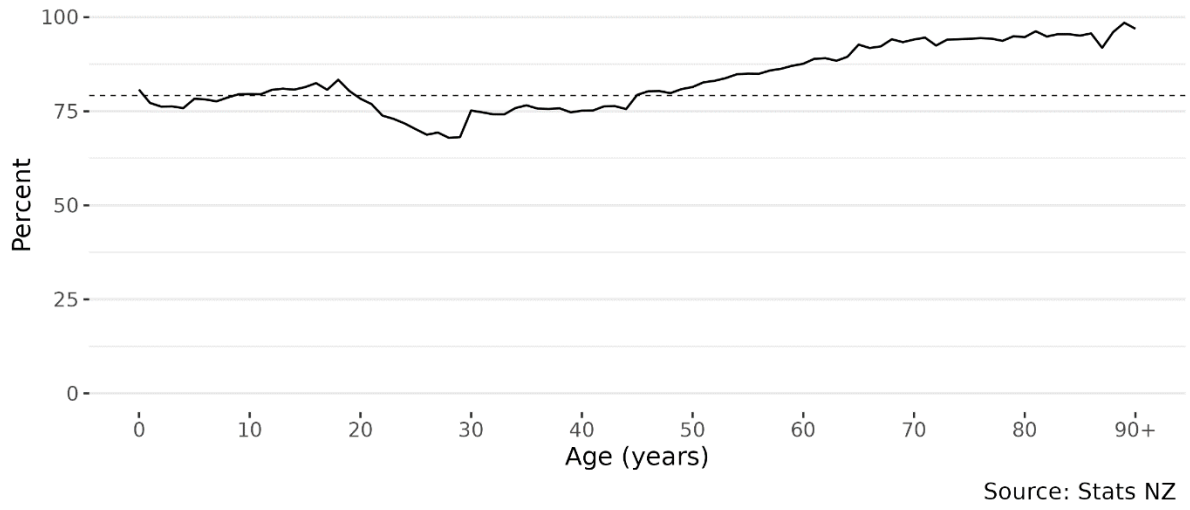


Figure 9: Percentage of eligible individuals included as household enumerations, by single year of age



Appendix 4: Thresholds for admin enumerations in meshblocks

This section provides more information about the choice of thresholds for admin enumerations in meshblocks. We first look at test individuals (those present in both 2023 Census and admin data who would not have been added as a household enumeration), then assess the impact on individuals that did not respond to the 2023 Census, and who are in scope for meshblock enumeration.

Test enumerations

To determine the threshold for meshblock enumerations, we carried out test enumerations using responding individuals not included in test household enumerations.

An address score threshold of 0.4 was chosen for several reasons. This was the point at which an individual's admin meshblock was more likely to be correct than incorrect (see figure 10). It was also the address score at which the accuracy of test enumerations was maximised (the proportion of individuals enumerated who had a correct admin meshblock plus those who were not enumerated and had an incorrect admin meshblock). It was noted that a threshold of 0.4 did not adversely affect sub-populations, in particular, Māori or Pacific peoples. It was also shown that, although some test individuals enumerated using a threshold of 0.4 were placed in the wrong meshblock, aggregate counts for sub-geographies were reasonably accurate.

Figure 10: Number of test individuals, by address score and agreement between admin and census meshblocks

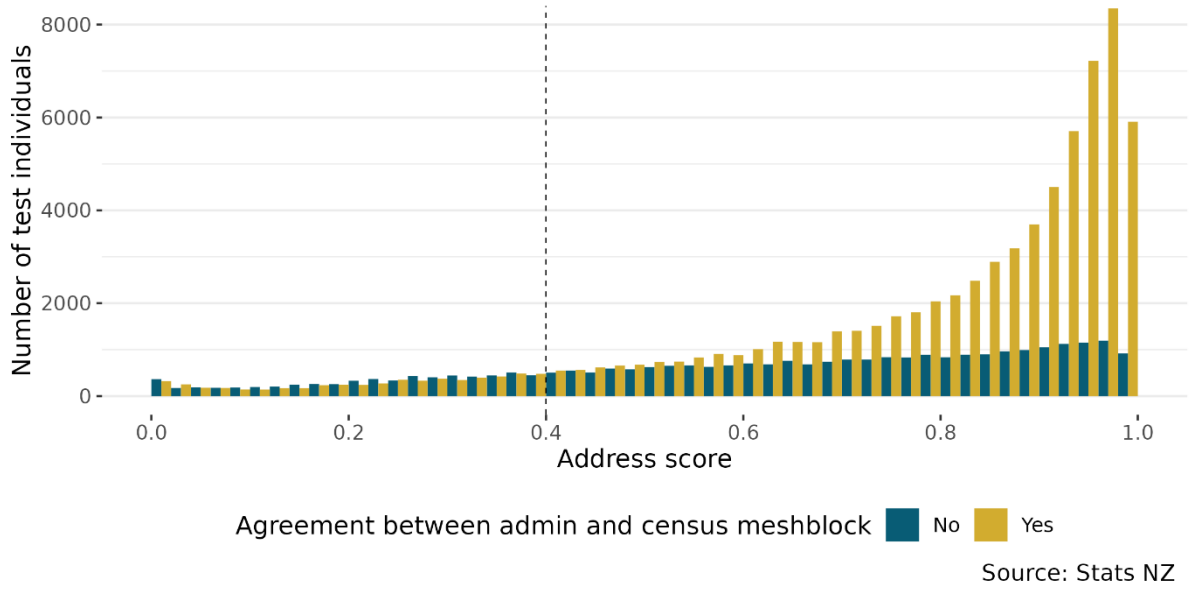


Table 7 shows the number of test individuals that would be enumerated with an address score threshold of 0.5. Around 88 percent of remaining test individuals were enumerated and, of these, around three quarters had the correct admin meshblock. Of those not enumerated, a much lower proportion (just under half) had the correct admin meshblock. These numbers suggest many of the records with address scores below 0.5 are still correct. However, on balance we believe it is better to exclude these records to reduce the risk of introducing over-coverage, both nationally and for smaller areas.

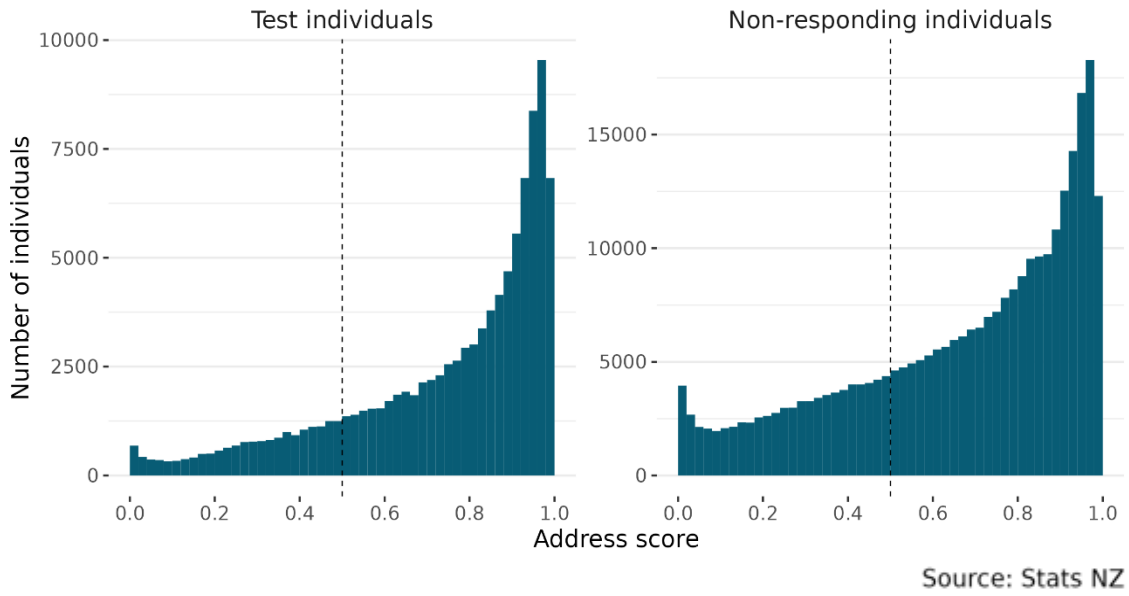
Table 7: Test meshblock enumerations using an address score threshold of 0.5

Measure	Percentage
Percentage of test individuals enumerated	*88.3*
Of the above, percentage of individuals with the correct admin meshblock	74.1
Percentage of test individuals not enumerated	*11.7*
Of the above, percentage of individuals with the correct admin meshblock	47.2

Actual enumerations

Figure 11 shows the distribution of address scores for responding and non-responding individuals. The overall distributions are similar, although non-responding individuals tend to have lower address scores on average.

Figure 11: Number of test and non-responding individuals, by address score



Using the selected 0.4 address score threshold, 81 percent of eligible individuals are included as admin enumerations. This is slightly lower than the numbers for test individuals above.

Figures 12 and 13 show that the percentage of eligible individuals who are enumerated is similar across ethnic groups and age groups. The dashed line shows the percentage for the full population. Slightly higher proportions of young children and older aged individuals are enumerated, but the differences are smaller than for household enumerations.

Figure 12: Percentage of eligible individuals included as meshblock enumerations, by level 1 ethnic group

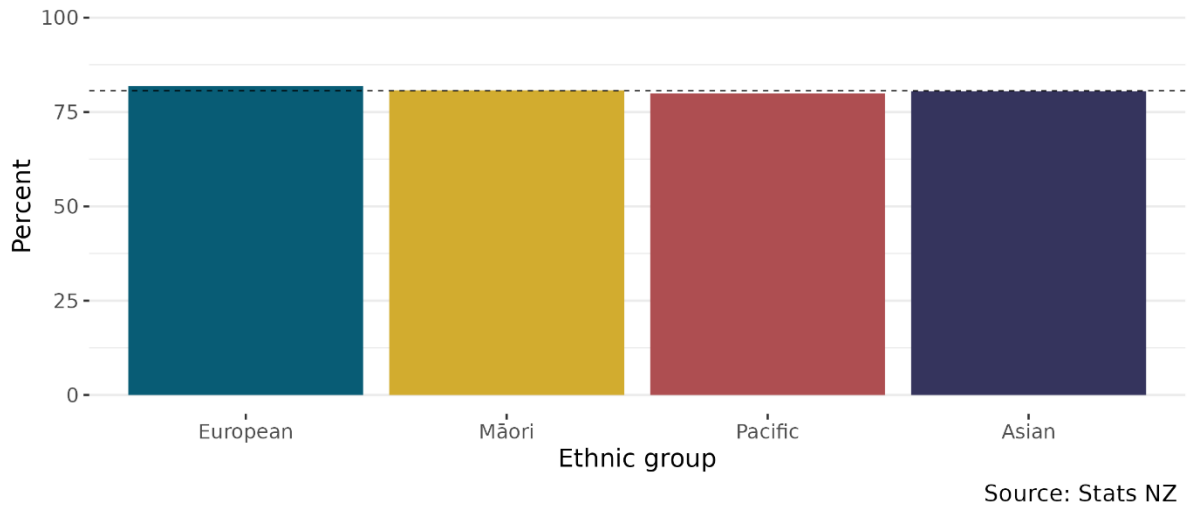


Figure 13: Percentage of eligible individuals included as meshblock enumerations, by single year of age