Linking 2023 Census responses to the Integrated Data Infrastructure

2023 Census | Tatauranga 2023

Citation

Stats NZ (2024). *Linking 2023 Census responses to the Integrated Data Infrastructure*. Retrieved from www.stats.govt.nz.

ISBN 978-1-99-104982-7 (online)

Published in May 2024 by
Stats NZ Tatauranga Aotearoa
Wellington, New Zealand

IDI 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.

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

Lists of tables and figures

List of tables

List of figures

Purpose and summary

Purpose

Linking 2023 Census responses to the Integrated Data Infrastructure reports how Stats NZ linked responses from the 2023 Census to administrative data (admin data) records in the Integrated Data Infrastructure (IDI), including changes to linking processes since the 2018 Census. It outlines the methods employed for linking the data, along with certain results and analyses, and demonstrates the quality of the links and the linking processes.

A high linkage rate and accurate links are needed to support census use of admin data. Where more responses can be linked to the IDI, it is less likely that Stats NZ is double-counting people. Accurate links ensure that correct attribute information from the admin data is used, and that individual people in IDI data, who have not already been included in census data, are correctly identified and added to the census file.

Summary

The 2023 Census is using a combined census model by design, supplementing census responses with alternative data sources. This includes using admin data to enumerate residents of New Zealand who did not fill out a census form where Stats NZ is confident in the quality of the admin data record. Stats NZ also fills missing data gaps in census attributes, using historical census data, admin data, and statistical imputation.

Using a combined census model for the 2023 Census provides more information.

A key step enabling this use of historical census data and admin data is the linking of 2023 Census responses to the Integrated Data Infrastructure (IDI) data. High-quality linking is essential for the accuracy of downstream processes. As no shared identification number for individual people exists, probabilistic linkage methods are used to join 2023 Census responses to admin data records in the IDI.

The combined census model was first implemented in the 2018 Census, and has had several improvements. Improvements in record linking for the 2023 Census are due to changes in the linking methodology, and more comprehensive and sophisticated processing methods for census data. Stats NZ has changed how it selects linking variables, including associated parameters, and has also improved the structure of the linking process since the 2018 Census. This has resulted in improvements to both the number of links and the overall quality of the linking between census and admin records. Note that the scope of records available for linking has widened since 2018 and absentees have also been included.

The linkage rates for linking 2023 Census responses to the IDI are:

- 97.9 percent for all responses, including census night absentee responses
- 98.2 percent for all responses, where census night absentees are excluded – this is comparable with 97.7 percent in the 2018 Census
- 98.8 percent for individual form responses
- 76.3 percent for responses that only appear on a census household set-up form, dwelling form, or continuation form, with no corresponding individual form (partial responses).

Note that census night absentee responses were responses from dwelling forms and household set-up forms only.

The estimated false positive rate for data linking remained low at 0.8 percent, compared with a rate of 0.6 percent in the 2018 Census. Responses that indicated Māori descent have a linkage rate of 97.3 percent, while the linkage rate for responses that indicated no Māori descent is 98.4 percent.

Overall, the quality of linking in the 2023 Census is similar to that in the 2018 Census, with a small increase in overall linkage rate and similar estimated rates for false positive and false negative errors. The linkage rate has decreased for partial responses, but there are significantly less of these responses than in the 2018 Census. Since partial responses are harder to find high quality links for, the partials that remain comprise an important subsection of the population for whom it is difficult to retrieve information across different sources.

Introduction

The aim of any census is to achieve an accurate count of the population (both people and dwellings) over a narrow time interval. Inevitably some responses are missing some key variables – for instance, address or age – and some individuals are missing altogether. In the 2023 Census, the inclusion of admin enumerations improves the coverage of the New Zealand population, and alternative data sources (including historical census data, admin data, and statistical imputation) improve the quality of census variables.

We use admin data to add individuals not accounted for in received responses, and admin and historical census data to complete missing attribute details. This process relies on linking 2023 Census responses with corresponding records for individuals in the Integrated Data Infrastructure (IDI). This process of supplementing census responses with high-quality admin data records, for people whom we have not received a response, is what we call admin enumeration. Admin enumerations occur after linking.

Methodology for using admin data to count people in the 2023 Census describes the details, techniques, and results of using admin data from the IDI to count people and dwellings in the 2023 Census.

Probabilistic linkage methods involve calculating the likelihood that sets of records match, considering many possible matches. The process of linking the 2023 Census to the IDI is automated through a probabilistic linking software programme. As no common identifier exists in New Zealand, the linkage methods compare various demographic variables (first names and family names, date of birth and age, sex at birth, gender, country of birth, and address information). These methods were developed based on the linkage methods used in the 2018 Census.

Linking 2018 Census respondents to the Integrated Data Infrastructure and Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project has more information about these methods.

Probabilistic linking comes with potential linkage error, for example, linkage errors caused by incomplete or inconsistent data between sources. Linkage error can result in two records being incorrectly linked (false positives) or two records that refer to the same person not being linked (false negatives). If not accounted for in downstream processes, false positive error will underestimate the population, while false negative error will overestimate it. A balance between each type of error must be found as, often, decisions made to reduce one kind of error will increase the rate of the other. For example, setting stricter criteria for two records to be linked makes it more likely that records that are a true match will not be linked.

Input data

This section briefly describes the two key data inputs for the linking process: 2023 Census responses, and the Integrated Data Infrastructure (IDI).

Privacy protection

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. The results in this paper have also been confidentialised to protect individuals from identification.

Before linking 2023 Census responses to the IDI, a privacy impact assessment (PIA) was undertaken. The PIA evaluated the privacy risks associated with linking, and with the wider use of admin data to supplement 2023 Census response data, and described how these risks have been mitigated. Due to the identifying information needed to ensure high quality linkages, the linkage process was carried out in a secure environment, where access was restricted to staff carrying out the linkage, and only variables used in the linking process were available.

Careful consideration has been given to the privacy, security, and confidentiality issues associated with using admin data and survey data in the IDI.

Privacy impact assessment for the use of admin data in the 2023 Census provides more information.

Integrated Data Infrastructure: Overarching privacy impact assessment has further detail.

2023 Census responses

The New Zealand Census of Population and Dwellings is an official count of how many people and dwellings there are in New Zealand. Every five years, Stats NZ traditionally asks everyone in the country to provide information about themselves and their dwellings and households.

The New Zealand census is based on a de facto approach, counting individuals who are physically present in New Zealand on census night. People who usually live in New Zealand but are overseas on census night are excluded from the census target population. Overseas visitors who are in New Zealand are asked to complete census forms.

When processing census responses, duplicate records and facetious responses are detected and removed. All the 2023 Census counts in this paper are based on the census linking population available at the point of linkage, which differ slightly from the final counts because there are additional steps after linking.

Census linking population

The census linking population includes responses from New Zealand usual residents (individuals who respond with a usual residence address in New Zealand). The census responses include:

- **individual forms** submitted online or using a paper form
- **partial responses** that only appear on a household set-up form, a dwelling form, or a continuation form (with no corresponding individual form).

The census linking population excludes responses where usual residence is recorded as overseas but includes absentee responses (someone who usually lives there, but who was not present on census night). Absentee responses from an online household set-up form, or a paper dwelling form, are subsequently assessed for their presence in New Zealand on census night, using admin data.

To be counted as a response in the census linking population, a record must have valid responses for two of these three variables: name, date of birth, and geographic meshblock.

The census linking population contained approximately 4,476,660 records, including 3,654,114 adults and 822,546 children. Of these 4,476,660 records, there were approximately 4,350,378 individual forms and 126,282 records from partial responses only. Table 1 shows the number of census records available for linking, by record type and unit data source.

Table 1: Census linking population (excludes absentees)

Record type	Unit data source	Number of records	Percentage of records
NZ adult	Census individual form	3,559,176	79.5%
NZ adult	Partial responses	94,938	2.1%
NZ child	Census individual form	791,202	17.7%
NZ child	Partial responses	31,344	0.7%

Source: Stats NZ

IDI spine

The IDI is a large research database holding microdata about people, dwellings, households, and businesses. Data is gathered from a range of government agencies, Stats NZ surveys and 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). The target population for the spine is all individuals who have ever been residents in 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 visas and transit visas)
- all individuals issued with an IRD (tax) number.

The IDI spine is rebuilt (or refreshed) on a quarterly basis. All datasets contained in the IDI are then linked to this spine. IDI data from October 2023 is the source of all the admin data used in linking with 2023 Census data. The IDI spine in October 2023 was the most recent IDI spine data available at the time of linking and included approximately 11 million individuals found in one or more of the spine sources.

The IDI spine attempts to keep one record per person. For linking purposes, we include additional records per person, using address information from other admin sources (various government agencies and Stats NZ surveys) which are not used in the construction of the IDI spine. This is because address information can vary between individuals, and it can be difficult to keep this information up to date. This additional address information is often in multiple records for the same individual in the IDI spine, where only the address variable changes between records which refer to the same individual. The 11-million-record IDI spine becomes approximately 45 million ‘person-address’ records.

Linking variables

In New Zealand, we do not have unique personal identification numbers available to link the 2023 Census data to the IDI. Therefore, the following personally identifying variables have been used to compare records:

- first names
- family name
- first initials of first names – the first letter of each word of a person's first names
- last characters of first names – the last two letters of the first names
- last characters of family name – last two letters of the family name
- sex at birth
- gender
- day, month, and year of birth
- age
- country of birth
- address ID – a unique address identifier obtained from matching an address string to a reference address in Stats NZ’s Statistical Location Register
- meshblock – a small geographic area corresponding to the address ID. Meshblock is the smallest geographic unit for which statistical data is collected and processed by Stats NZ.

Table 2: Linking information for records appearing on an individual census form versus those that only appear on a household set-up or dwelling form

Individual forms	Household set-up forms and dwelling forms
First names	First names
Family name	Family name
Age	Age
Day of birth	
Month of birth	
Year of birth	
Sex at birth	
Gender	Gender
Usual residence address	Usual residence address
Usual residence meshblock	Usual residence meshblock
Country of birth	

Source: Stats NZ

Meshblock and address

In the census file, we use a person's usual residence – where a person considers themselves to usually reside – to look up the address ID and meshblock associated with it. The census address ID is from the census dwelling frame, which is based on the Stats NZ statistical location register.

Addresses in the IDI are held in an address notification table, which includes address information from various government agencies, and from Stats NZ surveys such as the Household Economic Survey, 2013 Census, and 2018 Census. Individuals can have more than one address recorded in this dataset.

In the IDI table used for linking, each person can have multiple records for each address they have provided to admin data sources. To improve the linking result, we developed a method for selecting the most recent address for each person. The most recent address is the most recent notification, dated on or before census day, which was received into the IDI for the refresh used for linking. This address is used in earlier linking passes (rounds or iterations of sending unlinked data through the IBM QualityStage software used for probabilistic linking) to ensure a link is made to the most recent address, before other addresses are considered in later passes. This method was also used for the 2018 Census. This contribution of the address ID variable during linking means that we can allow for slightly larger differences across other variables, or allow for partial information.

There are also other variables common to both census and the IDI that are not used as linking variables. This includes qualification, and workplace address. Variables like this have been excluded, primarily because their values tend to be less consistent across the two linking datasets.

Linking methodology

This section describes the methodological approach used for linking 2023 Census responses to the IDI. We updated the 2018 Census methodology to accommodate changes in census data and to increase linking quality.

Linking pre-processing

The variables used for linking are first cleaned and standardised. In summary, the key steps are:

- cleaning names
- removing punctuation
- cleaning dates (for instance, standardising the formatting)
- coding variables so they match the classifications and concepts in the IDI spine.

Name cleaning refers to the removal of special characters, digits, and duplicate words from names. Names are also converted to uppercase to be consistent with name format in the IDI spine. Only information coming directly from a census form has been used in the census data side of the linking process. Census linking data arrives with operational and manual fixes, but some reactive cleaning also occurs where there are unexpected quality concerns with this data.

Linking procedures

To link census responses to the IDI, a probabilistic linkage method based on the Fellegi-Sunter method is applied in an automated process (Fellegi & Sunter, 1969). The IBM QualityStage software performs this linking process.

Linking 2018 Census respondents to the Integrated Data Infrastructure has more detailed methodology information.

Calculation of linkage weights

A link is determined by the total weight. If that weight is at or above a determined threshold, then we accept that two records are a match. The weight, for each single linking variable, is calculated from two probabilities:

- The *m* probability (reliability) is a measure of the trustworthiness of the data. It is the probability of two values agreeing, given that they refer to the same unit.

  m probability = Pr (two values agree | records are a match)

Another way of thinking about this is, given that two records should match, how likely is it that an issue with the data makes the values different? The *m* probability is the likelihood that they do match when they should. The initial set of *m* probabilities is estimated from 2018 Census data, and a new set of *m* probabilities is calculated using a SAS program after the first linking results are returned.

- The *u* probability (commonness) is a measure of how likely it is that two values will agree by chance. It is the probability of two values agreeing given that the records do not relate to the same unit.

  u probability = Pr (two values agree | records are not a match)

In other words, this is a measure of relative frequency. The more common a value is, the more likely two unrelated records are going to contain such values. The *u* probabilities are calculated by IBM QualityStage.

From these two probabilities, we calculate a weight for each single linking variable. The calculation used depends on whether the two values for each single linking variable agree. If they do agree, a positive weight is generated. If they do not agree, a negative weight is generated. The weight's size measures the evidence these values provide about each record pair being a match.

The two calculations are:

agreement weight = log (m / u)

and

disagreement weight = log (1 – m / 1 − u)

The total weight for each record pair is calculated from the sum of the weights for all linking variables.

Data integration manual has more information about these probabilities.

Selecting thresholds

After weights are calculated for each possible record pair, links are determined by selecting cut-off weights. Only record pairs with composite weights on or above the set cut-off value are deemed links.

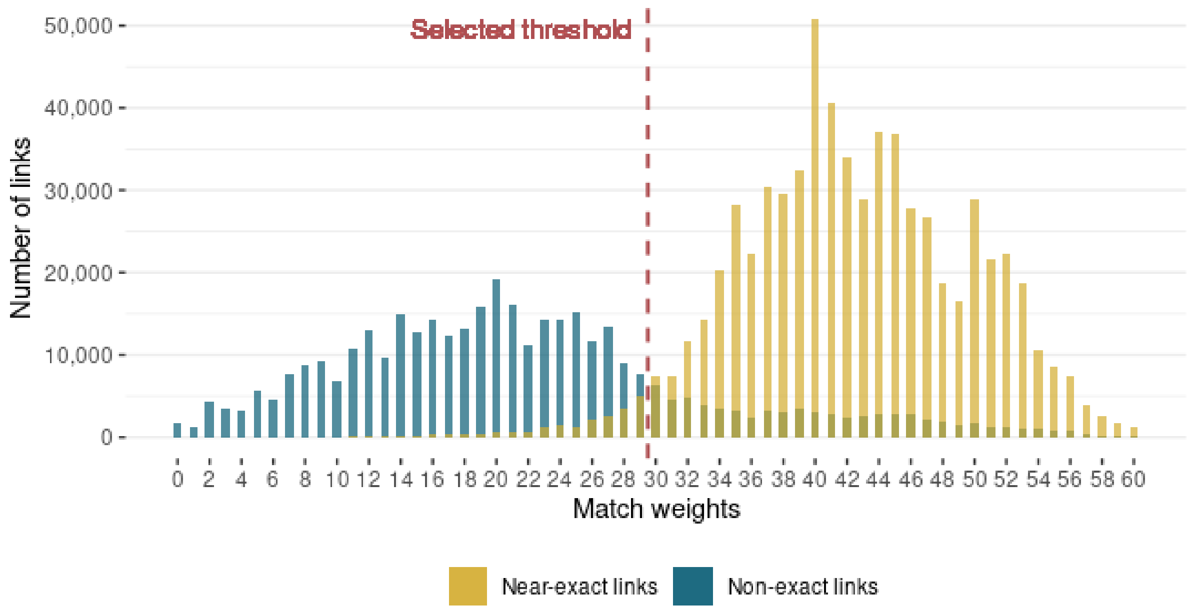
Possible values for cut-off thresholds are generated using a SAS program developed for linking projects in the IDI. To set the value of the cut-off thresholds, links are classified into near-exact links and non-exact links. Only links that closely agree on first names, family name, and date of birth are classified as near-exact links. These rules are outlined in the [Appendix: Near-exact link conditions](#).

There are multiple rounds, or passes, of sending unlinked data through the IBM QualityStage program. For each pass, the distribution of near-exact link and non-exact link counts by weight are plotted on a frequency distribution graph, and (typically) appear as two overlapping distributions. Non-exact links mostly have lower weight score values than near-exact links.

The aim is to select a cut-off threshold that prioritises near-exact links. This is done by selecting the weight score at the junction of the non-exact and near-exact weight score distributions, where near-exact links begin to greatly outnumber non-exact links. In figure 1, this would likely be at a weight score of ≈ 30. Census records with match weights above and including the threshold level are then removed from subsequent passes.

We manually review the links at the distribution junction, to inform the selection of a suitable cut-off threshold value. We consider the linkage rate, false positive rate, and false negative rate, in addition to these graphs, to select threshold values.

Figure 1: Non-exact and near-exact links, by match weight (example)



Linking pass structure

The data is processed through a series of passes in the IBM QualityStage software. After each pass, only the remainder of records in both the census and admin data will be eligible for subsequent passes. For example, records already linked in pass one will not go to pass two for linking.

Each pass can have a different combination of blocking variables, linking variables, and cut-off weight. A blocking variable is defined in IBM QualityStage as a variable where both records must have exactly matching values, otherwise that record pair is excluded from consideration in that pass. Records with missing values for the blocking variable are also excluded. Using blocking variables reduces the number of comparison pairs that need to be considered in each pass, which makes the overall task of linking more efficient. The linking passes and cut-off thresholds contain the variables used for the Fellegi-Sunter method in that pass.

Order of linking passes

We arrange the linking passes so that the strictest blocking occurs in the earliest passes, making the links which are highest quality and easiest to link first. Between passes, census records that have linked are excluded from consideration in subsequent passes. As the number of census records for possible linking reduces, the strictness of the blocking variables is also reduced. At the end of the linking process, if two census records link to the same person in the IDI, then only the link from the earliest pass is kept. If this occurs in the same pass, then only the link with the highest linkage weight is kept.

In 2023 Census linking, blocking variables in later passes are generally less restrictive and have more unique response options, for example, using age instead of date of birth, or using a single blocking variable rather than two. The principle is to ease the strictness of blocking variables across the passes. The exception to this principle is pass five. The focus of pass five is on linking records with similar names that have the same address value. Although pass five has a less restrictive blocking variable than earlier passes, the linkage weight threshold is set higher than usual. This illustrates that, in practice, there is more to consider when structuring the linking passes than just the restrictiveness of the blocking variables. For instance, table 3 shows that linking variables are not the same between passes. The precise construction of the passes is an iterative process of modification and testing of the quality of the links made.

IBM QualityStage ‘matching functions’

In each isolated pass, linking occurs based on selected variables that are expected to reveal the highest number of true matches. Each linking variable has a specific matching function applied to it in IBM QualityStage, to perform pairwise matching. Three types of matching functions were used in IBM QualityStage for the 2023 Census linking:

- CHAR: Compares values on a 'character-by-character basis'. It is often used to catch spelling mistakes.
- MULT_UNCERT: Compares all words in one variable with all words in the same variable of a second record by using a 'string comparison algorithm'.
- ABS_DIFF: Compares the absolute difference between two numbers to a specified value.

These and other comparisons can be seen in IBM documentation for IBM QualityStage software – Match comparisons.

Table 3: Blocking and linking variables for each pass

Pass	Blocking variables	Linking variables	IBM matching functions	Pr (*m*)	Linkage weight cut-off
1	Date of birth, Most recent meshblock				2
		First names	MULT_UNCERT	0.95	
		Family name	MULT_UNCERT	0.93	
		First initials	CHAR	0.95	
		Sex at birth	CHAR	0.99	
2	Date of birth, Sex at birth				9
		First names	MULT_UNCERT	0.95	
		Family name	MULT_UNCERT	0.93	
		First initials	CHAR	0.95	
		Country of birth	CHAR	0.95	
		Most recent meshblock	CHAR	0.85	
3	Most recent meshblock				21
		First names	MULT_UNCERT	0.95	
		Family name	MULT_UNCERT	0.93	
		First initials	CHAR	0.95	
		Country of birth	CHAR	0.95	
		Gender	CHAR	0.98	
		Age	ABS_DIFF	0.98	
		Day of birth	CHAR	0.85	

Pass	Blocking variables	Linking variables	IBM matching functions	Pr (*m*)	Linkage weight cut-off
		Month of birth	CHAR	0.85	
		Year of birth	CHAR	0.85	
4	Meshblock (full address list)				20
		First names	MULT_UNCERT	0.95	
		Family name	MULT_UNCERT	0.93	
		First initials	CHAR	0.95	
		Country of birth	CHAR	0.95	
		Gender	CHAR	0.98	
		Age	ABS_DIFF	0.98	
		Day of birth	CHAR	0.85	
		Month of birth	CHAR	0.85	
		Year of birth	CHAR	0.85	
5	Most recent address ID				13
		First names	MULT_UNCERT	0.95	
		Family name	MULT_UNCERT	0.93	
		First initials	CHAR	0.95	
		Country of birth	CHAR	0.95	
		Gender	CHAR	0.98	
		Age	ABS_DIFF	0.98	
		Last characters of first names	CHAR	0.90	
		Last characters of family name	CHAR	0.90	
6	Age, Family name				10
		First names	MULT_UNCERT	0.95	
		First initials	CHAR	0.95	
		Gender	CHAR	0.98	
		Last characters of first names	CHAR	0.90	

Source: Stats NZ

Changes for the 2023 Census methodology

Inclusion of absentee records

An absentee is an individual who is listed on an online household set-up form, or a paper dwelling form, as an absentee – someone who usually lives at the address associated with the form, but who was not present at that address on census night. In the initial stages of processing census data (before linking), individuals listed as absentees from a household are associated with individual forms at their census night address. Absentee records not processed in this manner are subsequently considered for linking.

Any absentees who are linked or deduplicated during earlier processing stages are not considered absentees for linking, but retain an absentee flag in the census file. In the 2018 Census, absentee records were out of scope for linking. For the 2023 Census, these records were included in linking and added to a new type of admin enumeration, admin-linked absentees. Like other partial responses, absentees do not have date of birth information, instead they only have name, age, gender, and usual residence address.

If a person's absentee record is linked to the IDI, and the person is determined to be in New Zealand on census night, then we include that person as an admin enumeration.

Methodology for using admin data to count people in the 2023 Census has more information.

Sex at birth and gender

The 2023 Census introduced separate questions about sex at birth and gender. The gender question has three options: male, female, and another gender. Sex at birth has two options: male and female. On the individual form, respondents are presented with both the gender and sex at birth questions. On the online form, a response to at least one of these questions is mandatory. Household listings (household set-up form, dwelling form, and continuation form) only ask for gender, so partial responses only have gender values, and not sex at birth values. Previous censuses asked one question relating to these concepts with two options: male and female.

In the IDI, there is not currently a clear distinction between gender, sex, and sex at birth information. However, better admin measures of these concepts should become available (and therefore the quality of gender as a linking variable should improve) as admin data sources move toward using the Data standard for gender, sex, and variations of sex characteristics.

Methodologies for filling gaps in gender and sex at birth concepts for the 2023 Census discusses how, of the sex at birth and gender census attributes, sex at birth tends to more closely agree with the combined sex/gender variable in the IDI. In most prior linking projects at Stats NZ, the sex variable has typically been used as a blocking variable as well as a linking variable, as it is considered consistent between sources. In 2023 Census linking, gender cannot be used for blocking, as it would exclude all individuals who provided ‘another gender’ responses.

2023 Post-enumeration Survey: Linking design reports that this process demonstrates no significant reduction in linking performance by including gender as a linking variable instead of sex. The 2023 Post-enumeration Survey (PES) happens after 2023 Census collection, and the results are linked to census responses, historical census data, and admin data using a similar data linking methodology. PES only collects gender, so there is no option to use sex at birth in PES.

To link 2023 Census data to the IDI, we found that a combination of both gender and sex at birth worked well. While individual forms receive both gender values and sex at birth values, partial responses only have gender values and do not have sex at birth values. This method allows more partial responses to link than otherwise would.

Short names

A general problem faced when performing data integration at Stats NZ are cases where individuals with names less than three characters long are linking to incorrect records, where the name is a subset of characters in a longer name. For example, ‘Lin’ might be matched with ‘Linda’.

For linking in the 2023 Census, two additional approaches to linking names have been used. The initials of the first names are extracted and used as a linking variable. In addition, we use the last two letters of the first names, and the family name, as linking variables. This approach using final letters is only included in the final two passes and has a minor effect on final linking.

There may be a combination of these methods that does not affect overall linkage quality, but allows better quality links for people with short names. However, there are too many potential combinations of methods to investigate exhaustively. This remains an area of interest for future linking quality research at Stats NZ.

Minor adjustments to the structure of linking processes

In the census file used for linking, each person has a unique record. In the IDI table used for linking, there can be several address records for an individual with the same identifying code. Each of these pairs is a ‘person-address’ record from the IDI.

In 2018 Census, records that linked in one pass were removed from both the 2018 Census table and the IDI table in subsequent passes.

In the 2023 Census, we changed this to allow for IDI person-address records that linked with a census record to be included in subsequent linking passes. This is a minor change because, as already stated, the IDI has many person-address records for the same person – an average of five records per person – which will be available for linking in later passes, even if one of those person-address records is linked in an earlier pass. This linking method change from the 2018 Census to the 2023 Census better reflects the structure of the IDI person-address table and is therefore a more appropriate linking design.

The admin enumeration process that occurs after linking requires the final links to be one-to-one links (person-to-person links), but currently the linking process returns person to person-address links. Inevitably, we have cases of multiple census records linking to the same spine identifier. Some of the records removed in this process are genuine duplicates, others will be false positives with limited linking information. In the 2023 Census, immediately after linking, only non-unique spine identifiers with the highest quality links are kept – those with the lowest pass value and the highest match weight. The result is that a table of records with only unique identifiers for census records and IDI records is created.

This adjustment to the structure of the linking allows for more records to link in earlier passes (excluding the first pass), which are higher quality passes with stricter matching requirements around address variables. It also accounts for links in later passes, which may be missed or linked incorrectly in earlier passes. This resulted in small improvements to the overall linking of approximately a few thousand records.

Results

This section presents results of linking 2023 Census responses to the IDI spine. We include linkage rates, as well as other linkage quality measures.

Overall linking rates and linkage error

Linkage rates

The overall linkage rate is 97.9 percent. This value includes all records that are in scope for linking – New Zealand adult, New Zealand child, Absentee adult, and Absentee child. The linkage rate is the number of records linked, divided by the number of records in the census table used for linking.

The 2023 Census had a programme performance measure (PPM) of 97 percent of census responses to be linked to the IDI. The PPM only reflects the linkage rates for the New Zealand adult and New Zealand child record types. This PPM was achieved with a linkage rate of 98.2 percent for these record types. Note that the PPM was set before the decision was made to admin enumerate absentees.

Linking passes

Table 4 presents the results for all records (including absentees) involved in linking, for each linking pass. Most links (94.9 percent) were made in the first three passes. The remaining 3.0 percent of links were made in the final three passes. Partial responses are ineligible to link in pass 1 and 2, and most of this record type linked in pass 3. Note that figures in table 4 include absentees and are not calculated in the same way as the PPM.

Table 4: Number of links for each pass (includes absentee records)

Pass	Number of all records linked	Percentage of all records linked
1	3,479,265	75.9%
2	652,929	14.2%
3	222,234	4.8%
4	88,809	1.9%
5	29,919	0.7%
6	14,961	0.4%
Not linked	96,309	2.1%
Total	**4,488,117**	**97.9%**

Source: Stats NZ

False positives

A false positive refers to a census response that has incorrectly linked with a record in the IDI spine, whereas a true positive is a census response that has correctly linked with a record in the IDI spine. In this instance, false positives could add incorrect admin data attribute information, or incorrect historical census attribute information, to a census record.

We calculated the false positive rate for individual forms to be approximately 0.8 percent through an independent clerical review process. This compared with 0.6 percent in 2018.

False negatives

A false negative refers to a census response that has a true match in the IDI spine, but was not linked to any records. We calculated the average false negative rate for individual forms to be approximately 0.94 percent.

To enable the number of false negatives to be determined, we first need a method of estimating whether an individual record is in the IDI spine, and therefore should be linked.

Dual system estimation combining census responses and an admin population describes an approach to identify false links between census responses and the IDI spine. This method was developed as a more efficient way of estimating the scale of missed linkages than clerical linking.

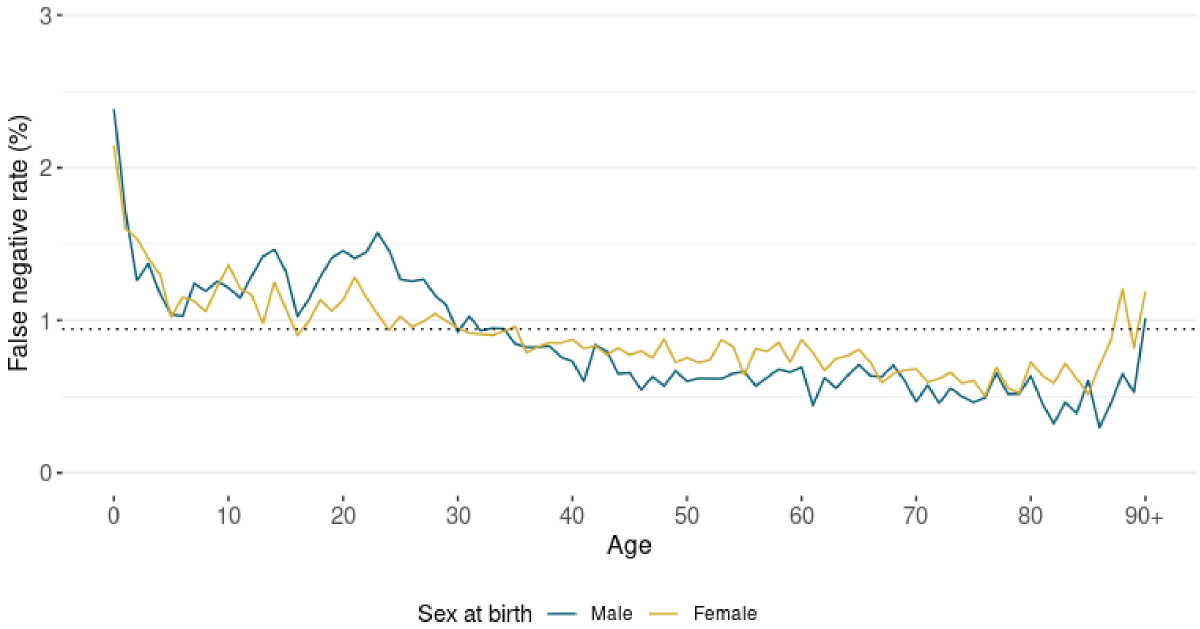
For this approach, we identify a subset of census responses, subset M*, that we are confident should be in the IDI spine (Choi, 2019). This subset is defined based on its census responses. We include responses that indicate someone was born in New Zealand and meets at least one of two criteria:

- records with reported taxable income
- records which were aged 14 years or less.

We also include responses indicating that the respondent was born outside of New Zealand, and in countries that do not require a visa – Australia, Cook Islands, Niue, and Tokelau – and who arrived in New Zealand after 1997 and before 2023, and who meet at least one of the two criteria above. This population is used to estimate the rate at which links are missed. We assume this rate can then be applied to those people not within the M* subset.

In figure 2, the estimated false negative rate for male responses is slightly higher than for female responses, for individuals below 30 years of age. However, for responses above 30 years of age, males show a slightly lower false negative rate than females. The dotted line represents the average estimated false negative linkage rate for all records (0.94 percent).

Figure 2: Estimated false negative rate, by individual year of age and sex at birth



Source: Stats NZ

Bias in linkage rates

We compared the linkage rates for certain groups, to determine if the linking is unbiased and working well for different groups in the population. Results for relevant variables of interest – age, sex at birth, gender, ethnic group, territorial authority area – all include values from alternative data sources. This is done to provide a more accurate representation of linkage rates across these groups.

Partial responses

Partial responses (records that only appear on household listings, household set-up forms, dwelling forms, or continuation forms) are more difficult to link to the IDI, compared with those that appear on individual forms. This is because household listings have less linking information available, as demonstrated in table 2. Partial responses are also often proxy responses – information about one person but submitted by someone else. Consequently, we expect these records may be more likely to have some incorrect information than other responses. While we cannot account for errors in linking information provided by respondents, such errors will reduce the overall linkage quality for these records.

Online and paper

The census file contained a higher number of paper forms in the 2023 Census than the 2018 Census due to changes in the collection strategy. The counts and link rates of these response types are shown in table 5.

Table 5: Counts of linking population, by response type

Response type	2018 Count	Percentage of 2018 count linked	2023 Count	Percentage of 2023 count linked
Individual form (online)	3,411,924	98.7%	3,313,734	99.0%
Individual form (paper)	497,613	98.6%	957,723	98.2%
Partial response (online)	..	88.1%	65,625	78.2%
Partial response (paper)	..	76.8%	29,283	72.5%

Symbol: .. figure not available

Source: Stats NZ

Figure 3 shows the linkage rate by response type and single year of age. Records sourced from individual forms achieved much higher linkage rates than records sourced from partial responses. While records from individual forms have consistent linkage rates across the age range, partial responses have a higher degree of variance across the age range, with a noticeable peak at 90 years of age. While the linkage rates for partial responses can be 10 to 15 percent lower than the linkage rates for records from individual forms, partial responses only make up about 2.8 percent of the census records used for linking to the IDI spine. Therefore, the overall linkage rate remains high.

Figure 3: Linkage rate, by unit source, response mode, and individual year of age

Source: Stats NZ

Māori descent

In table 6, the linkage rates for those of Māori descent are marginally lower than those without Māori descent.

The Māori descent population includes those who identified as having Māori descent on a 2023 Census form, as well as those identified as being of Māori descent through alternative data sources.

Data sources and imputation for Māori descent in the 2023 Census has more details on how Māori descent values are sourced through alternative data sources.

Table 6: Summary of links aggregated by Māori descent

Māori descent	Number of records linked	Percentage of records linked
Māori descent	765,519	97.3%
No Māori descent	3,530,016	98.4%
Don’t know	118,743	97.2%

Source: Stats NZ

Ethnicity

Linkage rates were generally similar across different ethnic groups, as shown in table 7.

Ethnic group responses often include multiple ethnic groups. A respondent is listed for each ethnic group they have identified in either their census response or from alternative data sources. For example, someone who has noted their ethnicity as being both Māori and Asian, as a multiple response, would be counted in both the Māori and Asian populations.

Ethnicity – 2023 Census: Information by concept has more information on ethnic group responses.

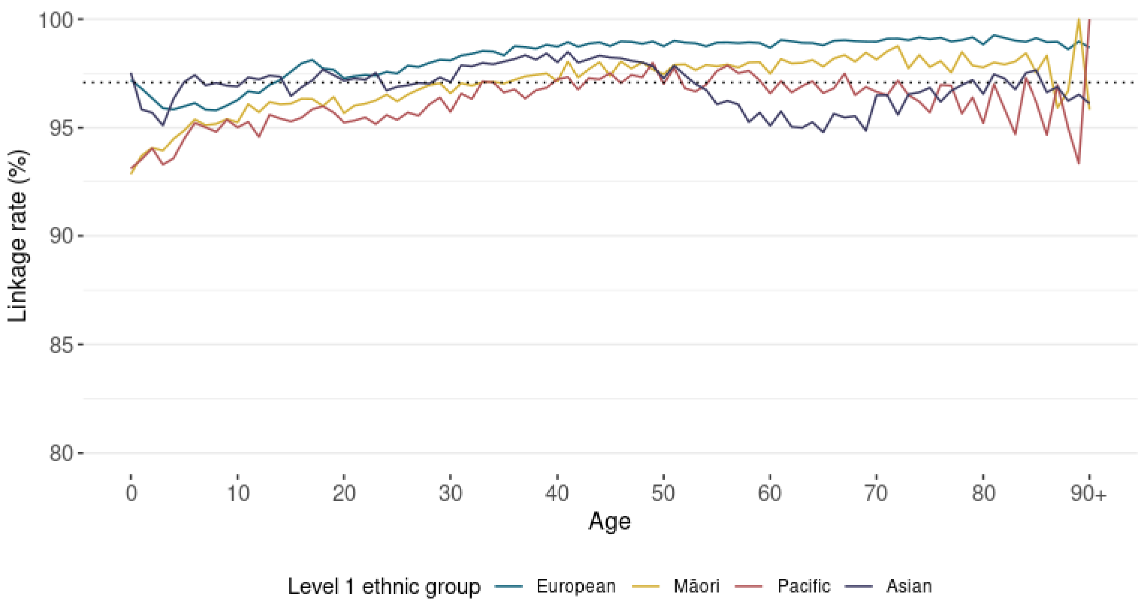
Table 7: Summary of links aggregated by level 1 ethnic group

Level 1 ethnic group	Number of records linked	Percentage of records linked
European	2,996,577	99.0%
Māori	645,915	98.3%
Pacific peoples	331,515	97.9%
Asian	774,378	98.6%
Middle Eastern/Latin American/African	80,070	98.3%
Other	47,850	98.7%
Total	**4,271,454**	**98.8%**

Source: Stats NZ

The overall trend in linkage rate by age is similar across ethnic groups. Responses in the European group show the highest linkage rates across ages, while linkage rates for Pacific responses were consistently lower, across ages. There is also a small but noticeable dip in the linkage rates of responses in the Asian ethnic group above 50 years old shown in figure 4.

Figure 4: Linkage rate, by individual year of age and level 1 ethnic group (all records)

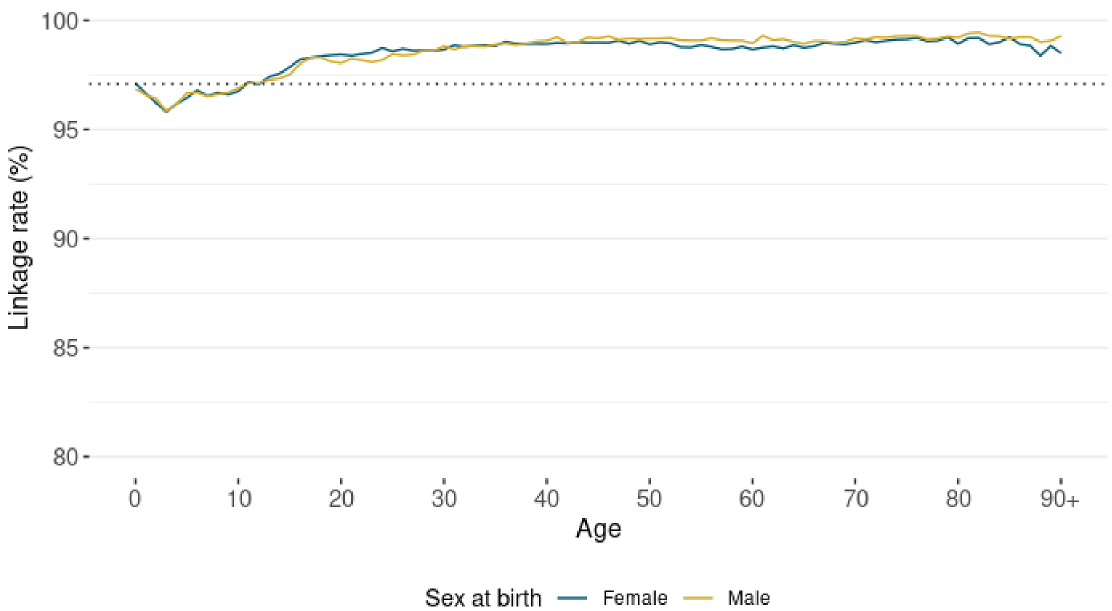


Source: Stats NZ

Sex at birth and gender, by age

As shown in figure 5, linkage rates remain fairly stable with age, but there is slight under-representation for those under 18 years old. Some of this may be due to undercoverage in the IDI for this group. The difference between male responses and female responses is negligible.

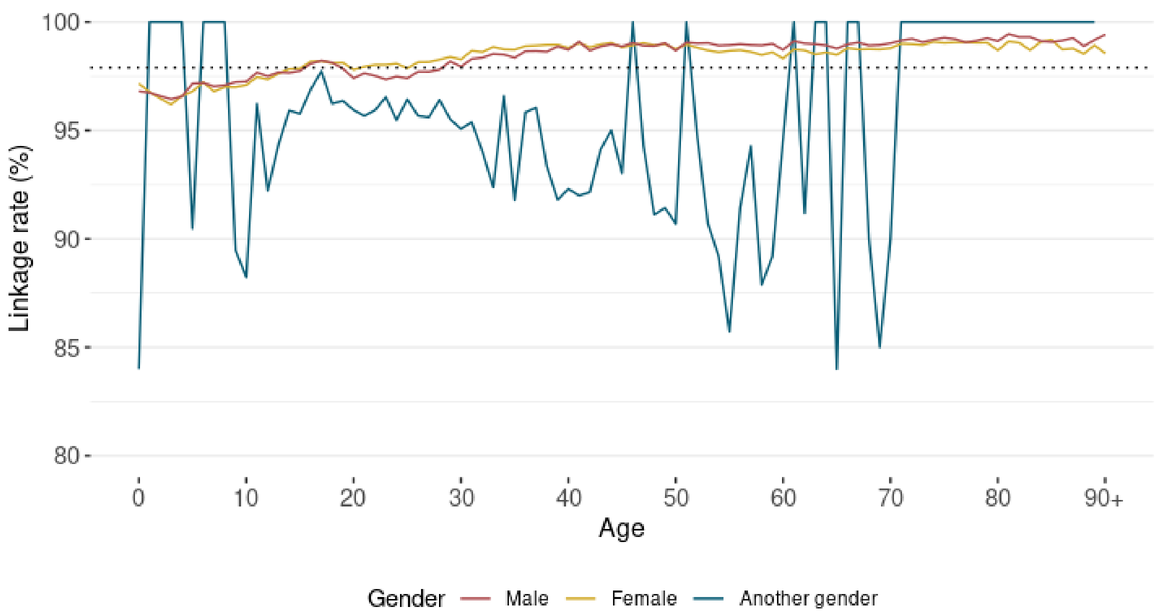
Figure 5: Linkage rate, by individual year of age and sex at birth (all records)



Source: Stats NZ

The linkage rate by gender seen in figure 6 is marginally lower than for male and female categories when compared with sex at birth. The Another gender category has high degrees of variability across age, possibly due to relatively low numbers of records in this category.

Figure 6: Linkage rate, by individual year of age and gender (all records)



Source: Stats NZ

Subnational area

Linkage rates are stable across regional council areas (table 8), ranging from a 97.5 percent linkage rate in Gisborne to 98.8 percent in Taranaki and Southland.

Table 8: Linkage rate, by region of usual residence

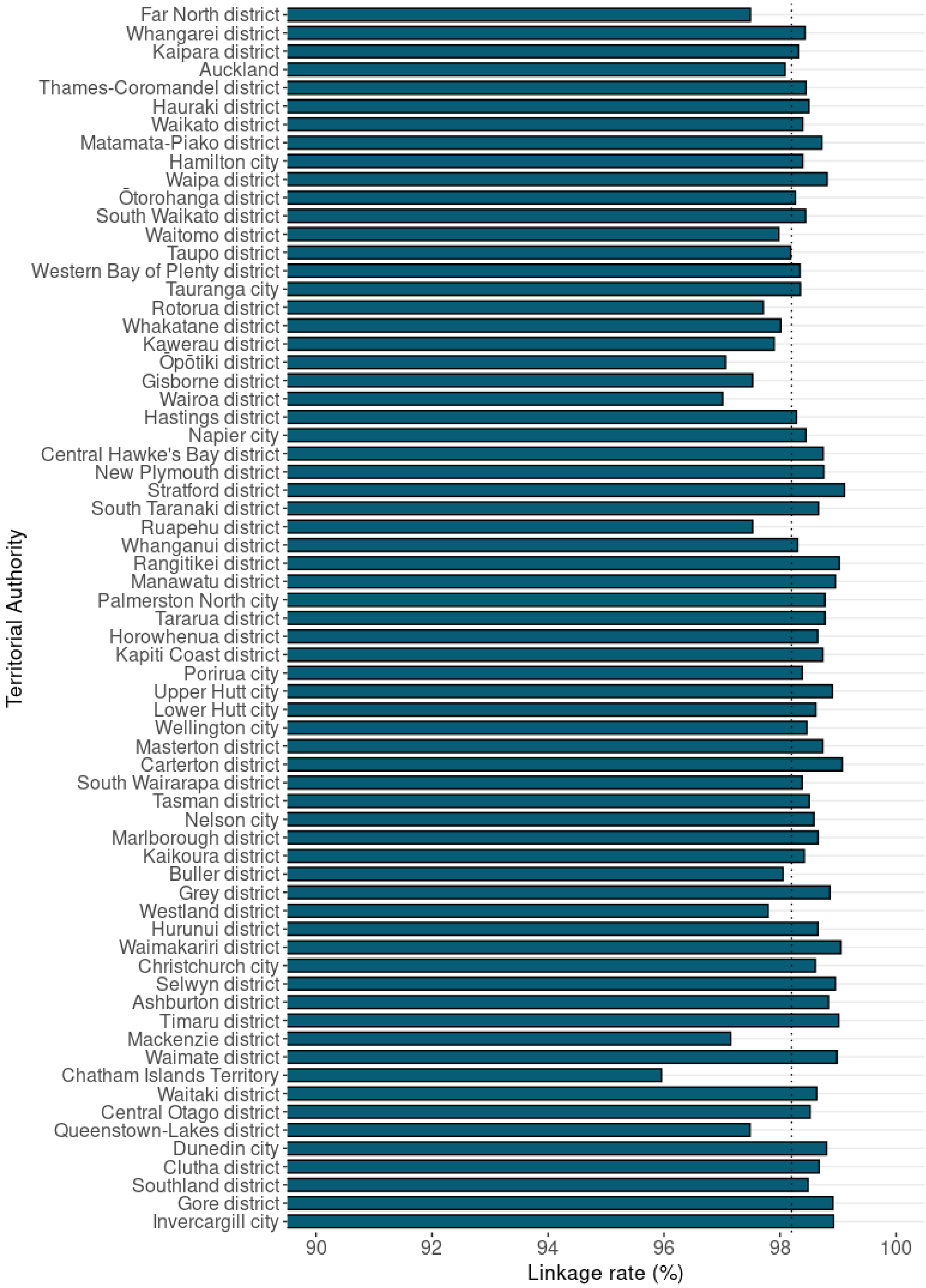
Region of usual residence	Number of records linked	Percentage of records linked
Northland	156,228	98.1%
Auckland	1,467,387	98.1%
Waikato	430,176	98.5%
Bay of Plenty	281,001	98.1%
Gisborne	42,360	97.5%
Hawke’s Bay	140,868	98.3%
Taranaki region	111,672	98.8%
Manawatū-Whanganui	219,771	98.7%

Region of usual residence	Number of records linked	Percentage of records linked
Wellington	475,641	98.6%
Tasman	51,210	98.5%
Nelson	48,312	98.6%
Marlborough	44,421	98.7%
West Coast	28,590	98.3%
Canterbury	597,192	98.7%
Otago	220,125	98.5%
Southland	90,228	98.8%

Source: Stats NZ

Linkage rates across territorial authority areas were also consistent. In figure 7, the linkage rate across territorial authorities is largely stable. The Chatham Islands Territory had the lowest linkage rate at 96.5 percent.

Figure 7: Linkage rate, by territorial authority area (all records)



Source: Stats NZ

Conclusion

Linking 2023 Census responses to the IDI is crucial for enabling the use of alternative data sources, allowing us to fill gaps in the census file with historical census and admin data. Small improvements have been made to the linking that increase the quality of census outputs.

Census responses counted through partial responses (people listed in a household form, but with no individual form) had the lowest linkage rates, due to the limited identifying and demographic information available about these people in their census responses. For this group, there was a lower linking rate than for the 2018 Census (76.3 percent in 2023, compared with 83.9 percent in 2018). However, there were fewer such records available for linking in the 2023 Census, and therefore this group could be revealing valuable insights about hard-to-reach populations. Hence, this could be a valuable area to investigate for future census programmes to improve census data quality even further.

Linkage rates were consistently high across all demographic subgroups investigated, although slightly lower rates were observed for children aged 2 to 4 years, and people aged 90 and older.

Overall, our linking methodology resulted in a high linkage rate between 2023 Census responses and records in the IDI. This high linkage rate, combined with very low false positive and false negative rates, gives us confidence in the linking results. It confirms our ability to obtain accurate information about people from historical censuses and admin data available in the IDI, while also ensuring we will not be double-counting people through the admin enumeration process.

Future work beyond the 2023 Census

Creating the best linkages possible to the IDI data will continue to be important for data quality considerations. The large linking population in the census presents more challenges than in linking sampled survey populations. Without the support currently offered in sample surveys that Stats NZ performs, the quality of variables which can be used for linking is likely to be lower than those from sample surveys where an interviewer can support individuals to answer questions. We anticipate that future work will continue to make further improvements.

As mentioned in the Changes for the 2023 Census methodology section, name variation is still a problem in identifying people, as it is not easy to determine whether a name variation is a different spelling of the same name, or a name for a different person. The available string functions in IBM QualityStage are not sufficient to deal with name variation when comparing short names. Name variation is more of a problem in the census than it is in surveys, due to the larger linking population.

Adopting a more accurate name matching algorithm is an area for future research to improve linkages between census responses and the IDI.

There is also some evidence of linkage error when children are indirectly linked to 2013 Census or 2018 Census responses during admin enumeration processes. For example, if someone aged under five is linked to a 2018 Census response, this suggests an incorrect link, either between the 2018 Census and the IDI spine or the 2023 Census and the IDI spine. More work could be undertaken to understand these scenarios and resolve them where possible.

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.

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

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

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: Near-exact link conditions

Links are classified into near-exact and non-exact links based on how well key variables agree.

For first names to agree, we allow:

- one of the following:
  - insertion
  - deletion
  - replacement
  - double
  - single
  - swap
- two of the following:
  - truncations
  - appendments.

For a family name to agree, we allow one of the following:

- insertion
- deletion
- replacement
- double
- single
- swap
- truncation
- appendment.

For date of birth to agree, we allow one of the following:

- replacement
- swap
- transposition of the day and month, providing the year is the same
- census date of birth is blank, providing the age is the same.

A near-exact link must have agreement as defined above for first name, family name, and date of birth. A non-exact link is then defined as any link that does not satisfy those near-exact conditions.