Stats NZ
Tatauranga Aotearoa

Creating household and family information from admin data: Progress, limitations, and next census

Te Kāwanatanga o Aotearoa
New Zealand Government

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

Disclaimer
Access to the data used in this study was provided by Stats NZ under conditions designed to give effect to the security and confidentiality provisions of the Data and Statistics Act 2022. These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) which is carefully managed by Stats NZ. Integrated data has more information about the IDI.

The results are based in part on tax data supplied by Inland Revenue to Stats NZ under the Tax Administration Act 1994 for statistical purposes. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

Citation
Atharifard, A, Basse, B, Merry, S, Harman, P, and Diener, H (2026). *Creating household and family information from admin data: Progress, limitations, and next census*. Retrieved from www.stats.govt.nz.

ISBN 978-1-991431-30-1

Published in March 2026 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

Summary and key points **5**
Key points 5
How this work informs the next census and future work 6
Introduction to household and family data **8**
Addresses and households **10**
Addresses 10
Household unit construction 12
Household quality 12
Partnerships **14**
Quality of admin partner relationships 14
Relationships beyond partnerships **16**
Data sources 16
Derivation method 17
Admin familial relationships quality 17
Opportunities 21
From households and relationships to families **23**
Family nucleus construction 23
Admin family quality 23
Biases and limitations in our methodology and data **28**
The scope for methodological improvement **31**
Method for using relationship information to improve address selection 31
Potential for future improvements 33
Looking ahead **34**
Survey data 34
Admin data quality evolution 34
Customer-centric and collaborative approach 35
References **37**
Appendix 1: 2023 Census household counts **38**
Appendix 2: Admin households and families excluded from the unit record comparison **39**
Appendix 3: Admin familial relationships by source and relationship type **41**

List of tables and figures

List of tables

1 Consistency rates of admin addresses, households, and families with census for previous investigations and work covered by this paper 9
2 Count of admin and census linked individuals by address agreement 11
3 Count of admin and 2023 Census familial relationships by relationship type 18
4 Agreement/disagreement of family type between admin and 2023 Census, for linked selected families with an identifiable family type 26
5 Household membership agreement for linked admin / 2023 Census households, by level 1 ethnic group 32
6 2023 Census household counts 38
7 Counts of admin households and families excluded from the unit record comparison 40
8 Counts of within-household admin familial relationships by source and relationship type 41

List of figures

1 Constructing households and families from admin data 8
2 Ratio of admin count to 2023 Census count of familial role of individual by overseas-born status 20
3 Count of admin familial relationships (both within-household and across-household) as at 30 June from 2006 to 2024 22
5 Household composition distribution, for admin and 2023 Census 27
6 Family membership agreement for linked admin / 2023 Census families, by level 1 ethnic group 30
7 Constructing households and families from admin data using the new method for address selection 32

Summary and key points

This paper provides an update of our current understanding of how well administrative (admin) data can produce household and family statistics, compared with past censuses. It forms part of Stats NZ's Census Transformation research programme. We also put the work in the context of the next census.

Our focus is on household and family statistics, which rely on two inputs: a list of individuals and their usual residential addresses, and a list of related individuals and the relationships between them. Because the quality of household and family statistics depends on both, we first assess the accuracy of admin addresses and households. We then outline the method used to derive relationships between individuals and estimate the consistency and coverage of these relationships by comparison with 2023 Census data, before producing household and family statistics from the combined address and relationship information.

The methodology has limitations and may introduce biases for some subpopulations: these are discussed in the section Biases and limitations in our methodology and data. Finally, we describe improvements trialled for address selection and suggest directions for future research.

Key points

We have used address and relationship information from admin data to derive household and family information. Key points are:

- Compared with earlier research, the consistency of admin-derived information with the census has improved since the first investigation in 2018, but is roughly unchanged from 2023, when it was used in the third iteration of the experimental administrative population census (APC) and the 2023 Census combined model.
- Households are formed at addresses, and errors accumulate. As a result, household consistency (68.3 percent) is substantially lower than address consistency (86.7 percent).
- Most familial relationships in admin data have high consistency with the 2023 Census, but partnerships and more distant familial relationships are more difficult to capture reliably.
- Families have high consistency with the 2023 Census where they can be formed, but they inherit quality issues from both household formation and partnership identification.
- The overall quality of household and family information from admin data may not reflect all experiences of our communities and data users. Some concepts, such as children not living with their biological parents, require deeper understanding and should be explored separately and in collaboration with researchers, subject matter experts, Te Tiriti partners, and others.
- For the next census, as a baseline solution, household and family data will likely need to be based on the new census annual sample survey. This will be a large survey and, when pooled over five years, will support a wide range of geographic and demographic analyses of households and families. However, this information will be less granular than what users have been accustomed to in the past.
- Future work will focus on how admin data can be used to improve survey estimates. However, this will require the development of new methods.
- Future admin data research will focus on improving the quality of the derived addresses and relationships (for example, by exploring additional data sources for relationships), and understanding their quality and limitations. There is methodological scope to improve the

quality of household and family data, such as the possibility of using addresses and relationships simultaneously to form households.

- Any future work should continue to draw on engagement already occurring, while expanding and deepening collaboration with data users, subject matter experts, Te Tiriti partners, researchers, and other government agencies. Engagement is important both for identifying where current methods fail to reflect lived realities and for helping prioritise areas where additional data sourcing or methodological development would be most valuable.

How this work informs the next census and future work

The 2025 decision to modernise the census means that the next census will be admin-data-first. That is, admin data will be used to enumerate individuals and dwellings and to derive attributes wherever quality is sufficient. Surveys will continue to play an important role, with a new census annual sample survey asking questions of a small percentage of the population to collect information not available from admin data and to help confirm the accuracy of admin data. Modernising the census provides more information about why the census is changing and the decision-making process.

The detailed approach on how information is gathered will depend on the concept: some information will rely on admin data, some on surveys, and others on a combination of both. The analysis in this paper shows that information for households and families, with the current available sources and methodology, will likely need to be based on surveys. This is because the key challenge of an admin-data-based approach is the process of forming households and, by extension, families. Progressing household-and-family-specific survey estimation as a base to build on in the future will therefore be prioritised. It should be noted that, if household and family information for the next census relies on survey estimates **only**, the data available will be less granular than users have been able to access in the past.

However, this paper also shows that there is a wealth of information in admin data about people, their addresses, and their relationships. This richness of information clearly has the potential to improve upon the limited granularity from surveys. To make use of that data, we will need to find approaches that combine admin and survey data. This could either be through complementary products, or an estimation approach that unifies the two. However, this will require the development of new methods.

This in turn means that there is value in continued research on deriving household and family information from admin data in parallel with progressing household-specific survey estimation approaches. The key components underpinning households and families are addresses and relationships. These are of interest in themselves, and any improvement in our ability to use the admin data improves the quality of our data beyond households and families.

Stats NZ has committed to engaging with Te Tiriti partners and other stakeholders as part of modernising the census. This engagement will inform the design of future methodology, ensuring that communities' and users' needs are understood and reflected.

There is still plenty of scope for methodological improvement in using admin data as The scope for methodological improvement shows. There are also more information sources that can be included in future work. By design, this paper does not include 2018 or 2023 Census or any other historical survey as a data source and only uses these as comparators. Relationship information that includes these sources would provide a fuller picture in the short to medium term. Similarly, additional admin data sources, such as detailed visa data including explicit relationships, are also likely to be available for the next census.

Our current admin data environment also has constraints. The Integrated Data Infrastructure (IDI), where our analysis was carried out, has a great number of information sources about individuals but limited information about places. Stats NZ is building the Integrated Statistical Data System for statistical production, which will link information about individuals with information about places and businesses. It is expected that this fuller picture with a richer set of data sources will lead to a better understanding of dwellings and addresses in New Zealand, which could also grow what is possible.

Introduction to household and family data

Household and family information are key statistics produced by Stats NZ. They are of broad public interest and help us understand how we live. They are also extremely important in informing decision-making in a wide range of areas, including local government planning and in setting of government policies such as income support and social housing. Furthermore, they help us understand housing affordability, family violence, child poverty, household crowding, and household net worth.

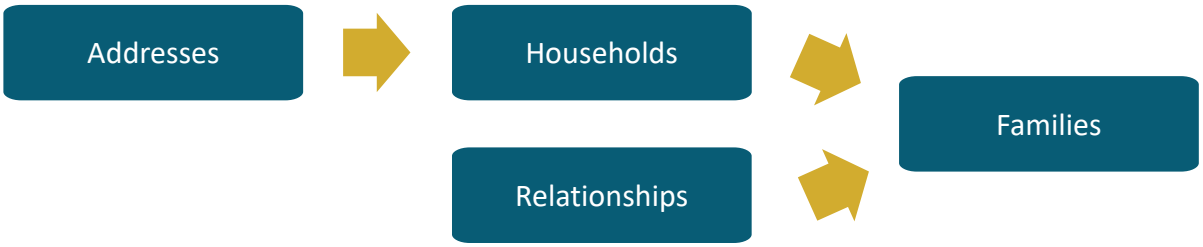
Census and household surveys have traditionally been Stats NZ's main source of household and family data. Census has been crucial to provide household and family information at the most granular level, enabling detailed information for a wide range of subpopulations.

In the census, all household and family data is derived rather than collected directly from a census response (Stats NZ, 2024a). This includes the base units of households, families, and extended families, which are derived from dwelling identifiers and from information about relationships between individuals. To achieve a quality level that is fit for use, the census household and family process involves a significant investment in processing time and resources. Generating the full set of census household relationships is a complex process which involves combining direct responses, values derived from direct responses, imputation, and manual intervention by specialist operators.

Deriving household and family information from admin data involves even more steps than in census. We first need to assign a single admin address to each individual, and to compile a dataset of relationships between pairs of individuals from several admin data sources. Address information is used to construct households, and relationships between individuals in a household are used to construct families. This workflow is shown in figure 1.

Figure 1

Constructing households and families from admin data



As household and family data requires many steps to generate, inconsistencies for individuals accumulate. For example, when comparing two sources, such as census and admin data, we see lower rates of membership agreement for whole households compared with individual addresses. This can be seen in table 1, which summarises consistency rates in the work covered by this paper and previous investigations.

Table 1

Consistency rates of admin addresses, households, and families with census for previous investigations and work covered by this paper

	Investigation				
	Gath and Bycroft (2018)	**APC third iteration**	**2023 Census admin enumeration**	**This paper**	
Comparison	2013 Census	2018 Census	2023 Census	2023 Census	
Method	Rules-based	Current method[1]	Current method[1]	Current method[1]	Exploratory method[2]
Admin address consistency rate (percentage of individuals)	Not derived	89	89.9	86.7	87.3
Admin exact household membership agreement (percentage of households)	47.6	72	N/A	68.3	71.8
Admin exact family membership agreement (percentage of families)	68.2	Not included	Not derived	81.1	83.6

1. This method is based on a machine learning model (XGBoost) followed by a rules-based correction of individuals aged 0–17 years, used for address prediction of admin individuals. This is described in Addresses and households. In the third iteration of the APC this model was trained on 2018 Census data. For both this paper and the 2023 Census admin enumerations, the model was trained on 2023 Census data.

2. This incorporates relationship information and is described in The scope for methodological improvement.

Source: Stats NZ

This paper presents and discusses the findings since the 2023 Census and the third iteration of the APC. It covers:

- an overview of the current state of admin data:
  - Addresses and households – admin addresses and households
  - Partnerships – admin partner relationships
  - Relationships beyond partnerships – other admin relationships
  - From households and relationships to families – admin families
- Biases and limitations in our methodology and data – a discussion of the limitations of the current state of admin household and family data
- The scope for methodological improvement – possible future methodological improvements
- Looking ahead – implications for the next census.

Addresses and households

The first step in deriving household and family data is to construct households. This is done by grouping together individuals who usually reside in the same private dwelling.

Addresses

A person's usual residence is the address of the dwelling where a person usually resides (see concept definition of place of usual residence). For individuals in admin data, we use a dataset of address notifications to determine their usual residence address. An address notification is created when an individual (or someone acting for them) tells an agency about an address linked to that individual. Additionally, it may refer to differing concepts of a person's address, for instance:

- postal or contact address may be recorded
- the usual residential address of a person may differ depending on context, for example, people who split time across different addresses (such as children of separated parents, or individuals who split time across addresses) may report different residential addresses in different situations.

It should also be noted that not every individual will have a usual residential address (for example, people living in some situations like mobile dwellings).

To determine a person's usual residence address from address notifications, an XGBoost machine-learning model is used followed by a rules-based adjustment for children.

Step 1: For each person-address pair where the person has a notification recorded in admin data for that address, an XGBoost machine-learning model assigns a weight between 0 and 1. These weights represent how confident the model is that the address is the usual residential address for the person at the reference date, and are calculated using a wide range of information about the person-address pair. This includes person-level variables such as age, ethnicity, and years since arrival to New Zealand, or notification-level variable such as notification frequency. If a person has multiple addresses assigned to them, the maximum weight address is chosen for that person.

Step 2: A final adjustment is then made for individuals aged 0 to 17 years: If a person has at least two parents and the two parents with the most recently recorded parental relationship have the same predicted admin address (which differs from that selected for the individual), then the individual is reassigned to the parents' admin address. Note that a person may have more than two parents recorded due to step/foster/other caregiver situations (for more details see Derivation method). A discussion on the impacts of this can be found in Biases and limitations in our methodology and data.

The two-step address selection methodology has been used in the third iteration of the experimental Admin Population Census (APC) (Stats NZ, 2023a) and in the 2023 Census (Stats NZ, 2024d). It was also the base for this investigation. We refer to this method as the **current method** to distinguish it from the method described in The scope for methodological improvement. Unless explicitly stated otherwise, when we refer to the current method we are talking about the current method with the XGBoost model trained on 2023 Census data.

To assess the quality of these addresses, admin-derived addresses are compared with the usual residence addresses reported in the 2023 Census, for individuals who both are in the APC and who responded to the 2023 Census. This results in an admin addresses accuracy rate of 86.7 percent,

which is approximately 2 percent lower than in the third iteration of the APC (Stats NZ, 2023a), which compared against the 2018 Census.

While this drop is concerning, there is evidence that, at least partially, it is due to address linking with the 2023 Census. Table 2 below breaks down address disagreements between census and admin data. It shows that, in particular, the number of disagreements with census addresses that do not appear anywhere else in admin data has almost doubled from 2018 to 2023. While some cases, such as field-enumerated rough sleepers, mean there are reasons for addresses not being reflected in admin data, the growth indicates that many of these addresses are missed links. Note that this does not affect the quality of the 2023 Census data itself, only its use as a comparator.

Table 2

Count of admin and census linked individuals by address agreement

		Count of individuals (percentage of total)	
		2018 Census	2023 Census
Same address		3,515,889 (88.3)[1]	3,729,231 (86.7)
Inconsistent address, same meshblock	Census address appears in at least one admin data source	40,023 (1.0)	42,006 (1.0)
	Census address never appears in any admin data source	56,853 (1.4)	114,054 (2.7)
Inconsistent address, inconsistent meshblock		367,089 (9.2)	416,004 (9.7)
Total		**3,979,854 (100)**	**4,301,295 (100)**

1. This differs to the number (89%) reported in Stats NZ (2023a). This is due to a combination of a different IDI refresh used, and slightly different filters used to isolate the population of interest.

Source: Stats NZ

The current methodology improves upon the previous rules-based method (Stats NZ 2017; 2022). However, several structural issues remain:

- **Independent evaluation of person-address pairs**:
  - The XGBoost model treats each person-address pair independently. This can lead to cases where people have a high weight for multiple addresses.
  - Additionally, there is no mechanism for considering household or family members. As a result, while individuals are often assigned to a correct address, their groupings into households or families are less accurate.
- **Address jitter**: Around the time of a move, people may appear to bounce back and forth between addresses across successive months.
- **Fragmented group moves**: Families or groups who live together often appear to move in stages, even when they moved together in reality. This misalignment reflects the lack of relationship information in the XGBoost machine-learning model.

The quality of admin addresses directly affects the quality of admin households and therefore the quality of admin families, extended families, and associated variables. Any improvement in the quality of admin usual residence address for individuals will flow through to improved construction of admin households. Stats NZ has identified the key data sources required for delivering the next census, and we are working with government agencies that can supply this data. Addresses are a priority in this larger effort.

Household unit construction

Census households (see concept definition of households) are constructed by grouping together all census individuals who usually reside in the same census private dwelling. Other Stats NZ household surveys also follow this method. There is currently no explicit dwelling information available from admin data, so instead we construct admin households by grouping together individuals with the same admin usual residence address. Note that some addresses contain multiple dwellings: at these addresses, admin will erroneously create one admin household when there should be many. We include these in our total admin household count but exclude them when analysing household membership.

To form admin households two inputs are needed: an admin list of individuals and the admin usual residence address for each of those individuals. If an admin address contains any non-private dwellings (currently identified by using the census dwelling file), then the individuals at this address will not be part of an admin household unit and no household unique identifier will be assigned.

The above definitions of household do not necessarily reflect the reality of how people live. Examples include:

- children of parents living separately may split their time between two different addresses
- individuals may split their time across addresses within their own family, whānau, or community
- people sometimes share facilities in non-private dwellings.

Household quality

To assess the quality of the constructed admin households, we compared admin household aggregate counts (as at 7 March 2023) with 2023 Census, and we assessed the accuracy of admin household membership by comparing admin/census individuals at 1.3 million linked admin/census addresses where there is one census high-quality household with no admin enumerations (see Appendix 2 for details).

Household counts

In total, we constructed 1,801,383 admin households. This is 5.5 percent lower than the 1,905,561 dwellings on the 2023 Census dwelling frame where it was expected that a household of usual resident(s) existed. This indicates that admin data undercounts households. (Note that the 1.8 million admin households included 27,684 households at addresses with multiple dwellings – this corresponds to 59,415 private dwellings.)

Although the number of admin households is 1.2 percent higher than the 1,780,527 households that 2023 Census identified in occupied private dwellings, there were 125,034 additional 2023 Census households that were outside the output subject population (see Appendix 1 and Stats NZ, 2024a).

The admin household undercount is also evident when we compare the average admin household size (2.78 people per household) with the average 2023 Census household size (2.70 people per household).

Household membership agreement

Of the original 1,801,383 admin households and 1,718,799 census households where we placed at least one usual resident (see Appendix 1), there were 1,278,567 households (71.0 percent of the total admin households) that we were able to link (by address) and compare. Of these 1.3 million selected households, 68.3 percent have exactly the same members as the census household. Our estimate of exact household membership agreement is lower than that for the third iteration of the experimental APC (Stats NZ, 2023a), where 72 percent of linked admin/2018 Census highly comparable households had exact membership agreement.

Household membership inaccuracies are influenced by two factors:

- Admin address misassignment of admin individuals: This may lead to two admin households being wrong – the household that they have incorrectly been placed in and the household that they should have been placed in.
- Admin individuals not able to be linked (by unique identifier) to a 2023 Census individual: These people may be included in an admin household that otherwise matches the 2023 Census household.

There were 522,816 admin households that were not selected for membership comparisons for reasons discussed in Appendix 2.

Partnerships

Once households have been constructed, relationships between all pairs of individuals within households can be used to derive further household and family information. Partner and parent-child relationships are needed to construct family nuclei; additional familial relationship types are needed to construct extended families; and the full set of familial and non-familial relationships is needed to derive variables such as household composition.

The classification of census relationship of usually resident household members lists the following categories: partner, child, parent, sibling, grandchild, grandparent, great-grandchild, great-grandparent, nibling (niece/nephew), pibling (aunt/uncle), other relative (such as cousin or parent-in-law), unrelated.

A familial relationship is defined as a "relationship in which a person is related to another person by blood, registered marriage, consensual union, fostering or adoption," so all the above are familial relationships, except for 'unrelated' which is a non-familial relationship.

Partner relationships include spouses (through marriage), civil union partners, and de facto partners. The other familial relationships include birth/biological, adoptive, foster, and step relationships.

Census relationships have been restricted to those between people who live in the same household. In the 2023 Census, the full set of within-household relationships was found through combining direct relationship responses, relationships inferred from direct responses, admin relationships, and imputed relationships. Manual intervention was used to fill in remaining gaps.

Our general approach for deriving admin relationships is to extract partner, parent and child relationships directly from admin data sources, and then to infer the other relationship types from these. For example, if Person A and Person B are both children of Person C, then Person A and Person B are inferred to be siblings.

We currently do not have any admin data sources that have a positive signal for two individuals being unrelated. Rather, we have to treat the absence of a relationship (direct or inferred) as an indication that they are unrelated. This has implications for the household composition variable, which has a number of categories that describe households in terms of whether the household contains unrelated people.

The third iteration of the APC derived familial relationships and estimated that they had a coverage rate of 89 percent and that 99 percent of admin partner, parent, and child relationships were consistent with the 2018 Census. The quality of the admin familial relationships used for the 2023 Census were estimated to have similar coverage and consistency rates. We think that the consistency of admin partners may have been overstated in these analyses – we explain why we believe this to be the case, and suggest an alternative methodology for measuring consistency in Quality of admin partner relationships.

Our current work (see Relationships beyond partnerships) has focused on the quality of admin familial relationships, excluding 'partner' and 'other related' relationships.

Quality of admin partner relationships

Partner relationships were derived for a previous investigation which analysed the quality of the admin variable 'partnership status in current relationship' (Somerville et al, 2025). This is an

individual variable which describes whether or not an individual is partnered (spouse, civil union, de facto) or non-partnered (separated, divorced or dissolved, widowed or surviving civil union partner, never married and never in a civil union). The investigation analysed the quality of admin partnership status of individuals by comparing admin values with 2018 Census values but, for individuals who were partnered, it did not analyse whether or not the person they were partnered with was correct.

To construct the admin families analysed in this paper (see From households and relationships to families), we used a dataset of admin partners derived as at 7 March 2023. To assess the quality of admin partners, we compared them with 2023 Census relationships.

For pairs of 2023 Census respondents who lived in the same census household and who were both in admin data (and where the census relationship was not an admin relationship or imputed), the ratio of the admin partner count to the census partner count was 0.805. This suggests that admin data misses around 19.5 percent of partner relationships (most of these are de facto relationships).

To estimate the accuracy of admin partnerships, we took admin couples living together in an admin household, where at least one of the partners was in a census household. We restricted census households to those in the household output subject population (an occupied private dwelling, not visitor-only) that comprised census respondents only. The address of the admin household did not have to match the address of the census household. There were three possibilities for the census relationship between the two individuals:

a. The individuals shared a census household and had a census partner relationship. The admin partnership was categorised as accurate.
b. The individuals shared a census household and had a census relationship that was not a partner. The admin partnership was categorised as inaccurate.
c. The individuals did not share a census household. They either lived in different census households (and may or may not have had a census partnership with someone else in that household), or one of the individuals was not in the census dataset. (The point here is that we do not need to know the true partnership status of both individuals in the admin partnership – we only need evidence that the admin partner relationship is wrong for one of them.) The admin partnership was categorised as inaccurate. It is acknowledged that it is possible for partners to live in different households and therefore that some of these may be legitimate partners – however, our investigation suggested that very few were likely to fall into this category.

We excluded census relationships that were imputed, an admin relationship, or derived from an admin relationship. This applied to (a) and (b). We reduced (c) by the proportion of pairs who were filtered out in (a) + (b) to ensure that the contribution of (c) did not affect the overall accuracy rate.

Once these adjustments had been made, accuracy was estimated as:

$$\text{accuracy of admin partner relationships} = \frac{(a)}{(a) + (b) + (c)}$$

We estimated that the accuracy of the admin partnerships was 94.9 percent. This is lower than the 99 percent that was estimated in the third iteration of the APC, and the analysis of admin partnerships used in the 2023 Census – neither of these included (c) in the calculation. The main issue here is that separations (for both de facto and legal partnerships) are challenging to identify in admin data.

Relationships beyond partnerships

Our current work investigates the quality of familial relationships derived from admin data. We excluded partner relationships (discussed and analysed in Partnerships). We also excluded the 'other related' relationship type which groups together a number of relationships, such as cousin, great-nibling, and parent-in-law.

The relationships covered by our investigation were therefore: parent, child, sibling, grandparent, grandchild, great-grandparent, great-grandchild, pibling, and nibling.

In our investigation, we derived familial relationships from admin data independently of household membership. Therefore, admin relationships were identified between people living in the same household and between people living in different households. This is broader than census which restricts relationships to those between people living in the same household.

Data sources

Parent–child relationships were sourced directly from:

- Department of Internal Affairs (DIA) birth register: This records all births in New Zealand, and adoptions in New Zealand of people born overseas. Birth records have been fully digitised from 1998 and mostly digitised from 1976.
- Ministry of Social Development (MSD) benefit data: This information includes details about the partners and children of recipients of working-age benefits, superannuation services, student allowances and loans, and social housing assistance. Note that these are social, but not necessarily legal, relationships between children and their parental figures.
- Working for Families (WFF) data: This contains information about recipients of WFF assistance (tax credits, accommodation supplement, or childcare assistance) since 1999 and includes details about the partners and children of recipients. Note that these are social, but not necessarily legal, relationships between children and their parental figures.
- Ministry of Business, Innovation and Employment (MBIE) immigration data: The MBIE datasets currently available within the IDI contain information on applicants applying for a resident visa or a temporary work, study, or visitor visa. We used joint visa applications to derive parent-child relationships.

2013 Census familial relationships were used, where available, to fill in missing relationships (of any type). We used relationships from 'relationship to reference person' (a question on the dwelling form where the person filling out the form was asked to list all people who usually lived in the dwelling and to describe how each person was related to them) and relationships that could be inferred from them.

Sibling, grandparent–grandchild, great-grandparent–great-grandchild, and nibling–pibling relationships were mainly inferred from parent–child relationships. A small number were sourced directly from the 2013 Census. We did not use 2018 Census data as a historical data source.

Derivation method

Our derivation method comprised four steps:

Step 1: We extracted relationship records from different data sources, applied de-duplication, and then compiled records into a single dataset.

Step 2: We removed records with unlikely age/relationship combinations that were likely to be incorrect (such as a parent younger than a child), and records with an unlikely age for a specific relationship at the time of the notification (such as a parent aged five years at the notification date of the record).

Step 3: We resolved conflicting records by source prioritisation. If two data sources showed different relationships for one pair of individuals, the record from the lower priority source was removed. Based on an analysis of the reliability of the information from each source, sources were ordered as follows: DIA, 2013 Census, MSD, WFF, MBIE.

Step 4: We created additional relationships through a series of inference steps. These records were inferred from a combination of other relationships, for example, the parent of an individual's parent was assigned as the individual's grandparent. Any conflicting records were resolved and the new records fed into the next inference step, so additional great-grandparent–great-grandchild, sibling, and pibling–nibling relationships were also created.

Note that it is possible for someone to have more than two parents (for example, step-parents and biological parents), therefore we retained all known parents if more than two were recorded.

Although we built our method on the method from previous investigations, we made some improvements:

- Our method included a more thorough method of resolving conflicting relationships, where records were prioritised using three levels (by source, then by derivation type, and then by relationship type).
- We used additional criteria to remove records that were likely to be wrong.
- Our method assigned the date of birth of the youngest person of each pair as the start date for each relationship, whereas previous methods used the notification date from the record. We believe that our approach more accurately describes most familial relationships, for example, most parent–child relationships (except for step/foster/adoptive/informal care relationships) start on the date that the child is born.

Admin familial relationships quality

To assess the quality of the derived admin familial relationships, we compared admin counts as at 7 March 2023 with counts from the 2023 Census (7 March 2023), estimated coverage, and analysed consistency with 2023 Census relationships.

Familial relationship counts

We derived 3,722,082 admin familial relationships as at 7 March 2023 between people living in the same admin household, compared with 3,977,391 census familial relationships between people living in the same 2023 Census household.

The sources for the within-household admin relationships (see Appendix 3 for a detailed breakdown) were as follows:

- Parent–child relationships were all sourced directly from admin data. Most of these came from the birth register (83.5 percent), 7.9 percent were sourced from benefit data, 5.7 percent from MBIE, and 2.9 percent from the 2013 Census.
- Other familial relationships were mainly inferred from the admin parent–child relationships (93.9 percent) with a small portion coming from the 2013 Census (6.1 percent).

Overall, 4.1 percent of the within-household familial relationships that we derived were only available from the 2013 Census.

Admin also found 14,075,277 additional familial relationships between people living in different households. These are discussed in Opportunities.

Table 3 shows the breakdown of admin and 2023 Census counts of familial relationships by relationship type.

Table 3

Count of admin and 2023 Census familial relationships by relationship type

Relationship type	Count of admin relationships[1]			Count of 2023 Census within-household[2] relationships
	Within-household	Across-household	Total	
Parent–child	2,392,905	2,450,565	4,843,470	2,571,048
Sibling	997,380	2,523,369	3,520,749	1,028,208
Grandparent–grandchild	188,796	2,557,716	2,746,512	231,966
Great-grandparent–great-grandchild	5,352	275,730	281,082	7,296
Pibling–nibling	137,649	6,267,897	6,405,546	138,873
Total	**3,722,082**	**14,075,277**	**17,797,359**	**3,977,391**

1. The reference date for the admin familial relationships dataset is 7 March 2023 (date of the 2023 Census).

2. For households in the census output subject population (occupied private dwellings, not visitor-only). Both census respondents and admin enumerations were included. All sources of relationships were included (direct responses, and inferred, imputed, and admin relationships).

Note: Fixed random rounding has been applied to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

Source: Stats NZ

Coverage

Coverage is the percentage of the target population for whom we are able to derive a value of the variable under consideration. As we do not have a target population of familial relationships, we are unable to calculate coverage using this method.

We therefore used the following two approaches to understand coverage.

Approach 1: Admin and census proportions of people with a familial relationship

We compared the proportion of admin individuals with an admin familial relationship with the proportion of census individuals with a census familial relationship (note that the latter included

relationships for both census respondents and admin enumerations, and that all sources of relationships were included – direct census responses, and inferred, imputed, and admin relationships).

Of the 5,007,795 individuals living in an admin household, 2,890,599 (57.7 percent) had a familial relationship with at least one other person in the household. Of the 4,743,312 individuals living in a census household, 2,978,814 (62.8 percent) had a familial relationship with at least one other person in the household.

Admin familial relationships were also derived for people living in different households. Of the 5,097,981 total individuals in the APC, 4,571,859 (89.7 percent) had at least one admin familial relationship (any type). We cannot compare this with the 2023 Census, since census relationships are restricted to those between pairs of individuals living in the same household.

Approach 2: Ratio of admin relationships to census relationships

For each familial relationship type, we calculated the ratio of the number of admin relationships to the number of 2023 Census relationships for pairs of census respondents who lived in the same census household and who were both in admin data. (Note that we excluded census relationships if either person in the pair was an admin enumeration, but all sources of relationships were included – direct census responses, and inferred, imputed, and admin relationships.)

The overall ratio for familial relationships was 0.909, that is, admin found 90.9 percent of the familial relationships identified in the 2023 Census. In terms of specific relationship types, admin found:

- 92.6 percent of parent–child relationships
- 94.8 percent of sibling relationships
- 68.7 percent of grandparent–grandchild relationships
- 57.2 percent of great-grandparent–great-grandchild relationships
- 63.6 percent of nibling–pibling relationships.

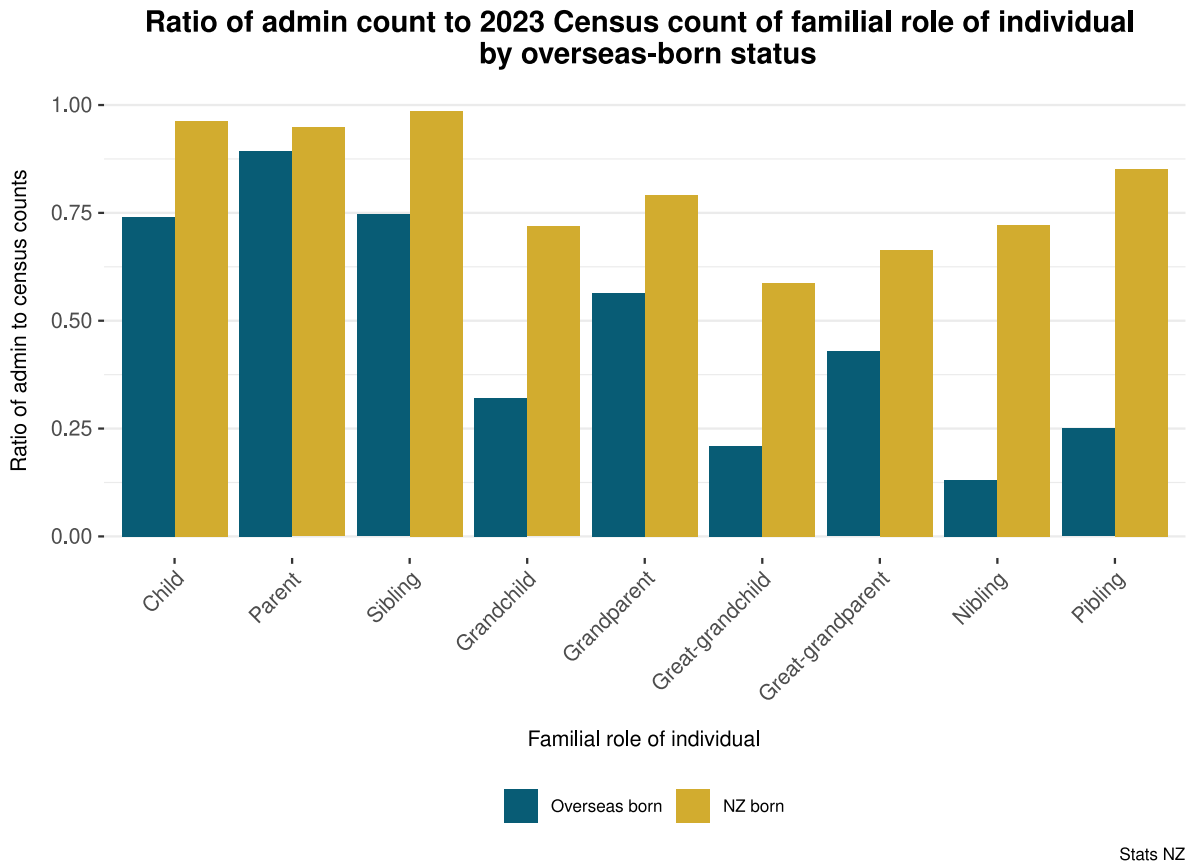
Parent–child relationships are the key relationship type since most other relationships are derived from them. The main groups of parent–child relationships that we are missing are:

- births occurring overseas that are not in the MSD, WFF, MBIE, or 2013 Census datasets – this particularly affects recent immigrants
- New Zealand births prior to the digitisation of records in the birth register (birth records have been fully digitised from 1998, and mostly digitised from 1976) that are not in the MSD, WFF, or 2013 Census datasets
- step and foster parent–child relationships that are not in the MSD, WFF, or 2013 Census datasets.

Being born overseas and age were the primary determinants of missingness:

- Parent–child relationships were missing for 26.0 percent of children born overseas compared with just 3.7 percent of children born in New Zealand. As shown in figure 2, the difference was more marked for more distant relationships.
- Parent–child relationships were more likely to be missing for older individuals.

Figure 2



Consistency

To investigate the accuracy of the admin familial relationships, we assessed the consistency of the admin relationships with the 2023 Census relationships for linked pairs of individuals. (Note that we excluded census relationships if either person in the pair was an admin enumeration, or if the relationship was imputed or an admin relationship.) The overall consistency of admin familial relationships was 98.3 percent.

In terms of relationship type:

- admin parent–child relationships were very accurate (98.8 percent consistency with census relationships)
- admin sibling and grandparent–grandchild relationships were also very accurate (98.4 percent and 96.4 percent consistency, respectively)
- admin great-grandparent–great-grandchild and pibling–nibling relationships were less accurate (72.6 percent and 77.9 percent consistency, respectively) but this is partly due to some inaccurate census responses, for example, great-grandparent–great-grandchild relationships being reported in census as grandparent–grandchild relationships
- ethnic group, Māori descent, and overseas-born status did not affect consistency rates.

In terms of admin data sources:

- Parent–child records from the DIA birth register and MBIE immigration data were extremely accurate (consistency rates above 99 percent).

- Parent–child relationships from MSD benefit data and WFF data were slightly less accurate (96.3 percent and 94.3 percent consistency, respectively).
- Relationships from the 2013 Census (all familial relationship types) had a consistency rate of 95.7 percent.
- Inferred relationships (all relationship types except parent–child) had a consistency rate of 97.4 percent.

Future improvements

We are actively working to source another MBIE dataset which contains explicit (rather than derived) relationship information from visa applications. This will provide comprehensive relationship information for future immigrants and, depending how far back MBIE records go, for some past immigrants.

The 2018 and 2023 Censuses will be added as historical data sources. Some of the current missing familial relationships will be filled using these sources. These will improve coverage in the short term but will provide limited benefits as we move away from the 2023 Census.

We will continue to investigate other possible sources of admin data.

In addition, over time, the coverage of parent–child relationships for people born in New Zealand will improve as the proportion of people whose birth records are missing decreases.

Improved coverage of parent–child relationships will lead to more inferred relationships and therefore higher coverage for all other relationship types. It should be noted, however, that step and foster relationships will continue to be challenging to identify in admin data.

Our analysis demonstrated that parent–child relationships derived from MBIE immigration data were extremely reliable. Therefore, in future work, we will consider reordering the priority levels for admin data sources to see if this helps to improve the quality of admin familial relationships.

Opportunities

An important opportunity is the ability of admin data to identify familial relationships between people living in different households, which census and Stats NZ surveys are currently unable to do. This provides the potential for families, whānau, and extended families across different households to be constructed.

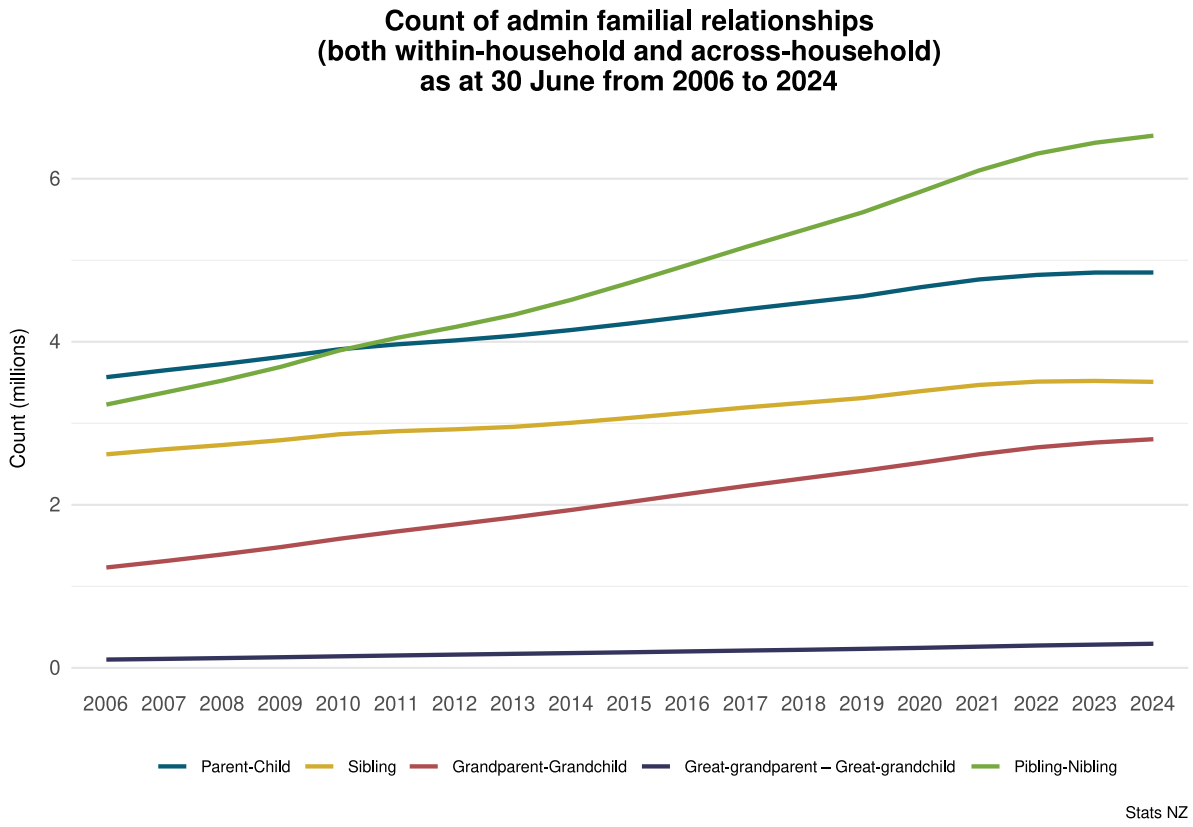
As noted, we derived 14,075,277 familial relationships between people living in the different households as at 7 March 2023. This was many times greater than the number we found between people living in the same household.

To demonstrate that the ability of admin data to identify familial relationships has improved, we analysed an annual time series for each admin familial relationship from June 2006 to June 2024. This is shown in figure 3. Note that both within-household and across-household relationships are included.

Overall, this graph shows a rising trend for all types of relationship, with the total number of familial relationships increasing from 10,747,797 in 2006 to 17,986,632 in 2024. As well as an increase in the availability of relationships from admin data over this period, this reflects the growth in the population of Aotearoa New Zealand.

The rate of increase differs across relationship types. The number of parent–child relationships increased by 36.0 percent from 2006 to 2024. More distant relationship types have a greater increase – note that this is expected as a single birth generally results in two parent–child relationships, four grandparent–grandchild relationships and eight great-grandparent–great-grandchild relationships.

Figure 3



From households and relationships to families

The next step in deriving household and family information is to construct family nuclei. This is done by grouping together partners, parents, and children who live in the same household.

Family nucleus construction

Census families (see concept definition of families) are constructed by combining census household members into related groups of people. These groupings involve parents, partners, and children, and are assigned a family type: couple without children; couple with child(ren); one parent with child(ren). A child within a family can be of any age but cannot have a partner and/or children of their own.

Admin families are constructed similarly, from two inputs: household members and their relationships (parents, partners, and children). We have also constructed admin family nuclei in the cases where there are: more than two partners without children; two unpartnered parents with child(ren) (that is, two people who share child(ren) but do not have a partner relationship); more than two parents (partnered or unpartnered) with child(ren). We included these in the admin family count because we know that some of these less usual family types exist. We wanted to estimate how many of these families were authentic and how many were due to errors in admin data.

In census households where there were child(ren) but no parents, a census family was constructed if there was another adult who could be placed in a parent role. We did not construct an admin family in households where there are child(ren) but no parents (that is, we did not place any other suitable adults in a parent role).

Note that the above definitions of census and admin family do not necessarily reflect the reality of how people live. For example, children of parents living separately may split their time between two different addresses; other extended family and/or household members may have a parent role; individuals may split their time across addresses within their own family, whānau, or community; families live in non-private dwellings.

Admin family quality

The quality of the constructed admin families was assessed similarly to admin households (see Addresses and households). For family membership agreement we also excluded admin and census addresses with multiple families (see Appendix 2).

For aggregate family counts we also compared admin families to Stats NZ family projection (2018-base, medium B, published December 2021). The projections have, as a starting point (base), an estimate of households and families which are currently derived every five years. Because the 2023-base estimate (at 30 June 2023) will not be derived until late 2026, in this investigation we use the 2018-base projections: Family and household projections: 2018(base)–2043 (published December 2021). Projections are produced with variants: low, medium A, medium B, and high. Here we have taken the medium B series, a more recent and up-to-date projection based on the 2018 Census data, which accounts for changing living arrangements compared to the earlier medium A. For more detailed information about the methodology see 2018-base family and household projections: Data and methods.

Family counts

In total, we were able to construct 1,160,004 admin families (from 1,801,383 admin households) as at 7 March 2023. This is:

- 10.4 percent lower than the 1,294,503 families (constructed from the 1,718,799 census households where we placed at least one usual resident) in the 2023 Census – note, however, that there was an undercount of 2023 Census households and families due to people with insufficient information unable to be placed in households (Stats NZ, 2024a)
- 22.1 percent lower than the Stats NZ family projection (2018-base, medium B, published December 2021) of 1,488,400 families as at 30 June 2023.

This undercount is mainly due to missing admin relationships and incorrect admin addresses. A missing admin partnership will lead to a family nucleus not being formed for a couple who do not have children. Similarly, an incorrect admin address for one partner in a couple will lead to a family nucleus not being formed if the couple does not have children. In addition, as mentioned above, and if possible, census assigns a suitable parent (such as a grandparent) to children in a census household with no parent. This leads to some census families being formed. However, in this investigation we have not replicated this methodology for admin families because it requires knowledge of extended family members.

The admin family count is an improvement on a previous investigation (Gath & Bycroft, 2018), where the admin family count was 41 percent lower than the 2013 Census family count.

Family membership agreement

Of the original 1,160,004 admin families and 1,294,503 census families, there were 740,265 families (63.8 percent of the total admin families) that we were able to link (by address) and compare. Of these 0.74 million selected families, 81.1 percent have exactly the same members as the census families. This is an improvement on the findings in Gath and Bycroft (2018), who used a slightly different (but comparable) methodology to show that 68 percent of linked individuals had exact family membership.

There were 419,739 admin families that were not able to be selected for membership comparisons. The reasons for their exclusions are discussed in Appendix 2.

Family type distribution

Figure 4 compares the admin family type distribution with 2023 Census and 2018-base family projections (medium B, published December 2021) as at 30 June 2023 for families with an identifiable family type. Admin over-counts ‘one parent with child(ren)’ families, and underestimates ‘couple without children’ families. These over-counts and under-counts are generally due to missing admin relationships and incorrect admin addresses.

There were 28,260 families (2.4 percent) that fell outside the family type standard classification categories. When we investigated these, we concluded that most contained either admin relationship and/or address errors, and should have been classified as one of the three standard family types.

Figure 4

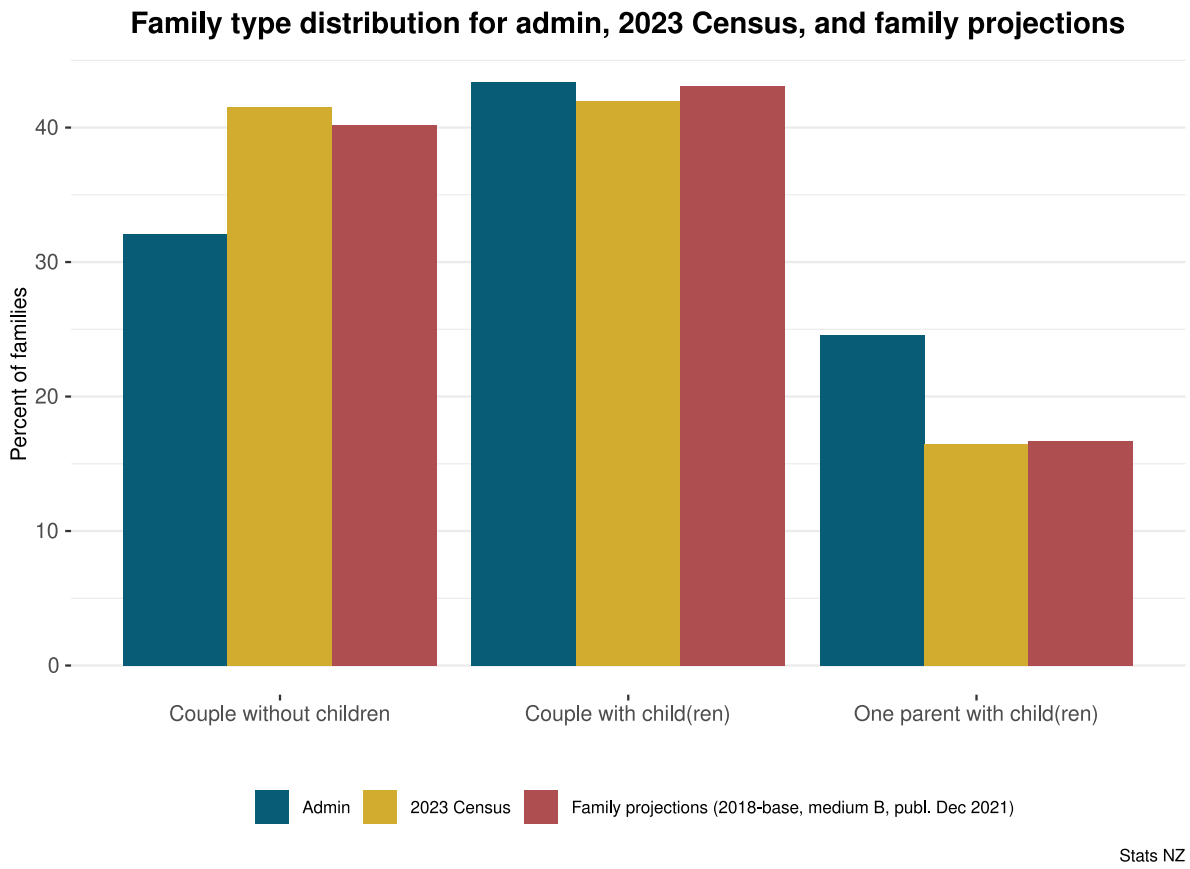


Table 4 investigates the family type agreement (shaded) and disagreement (unshaded) for the 722,838 linked admin / 2023 Census families which have an identifiable admin family type. Overall, 88.8 percent of families had a family type value that was consistent with the 2023 Census. Family type consistency/agreement was similar (between 88 and 90 percent) across the different family types.

Disagreements in family type consistency are mainly due to incorrect or missing admin relationships (especially missing partnerships) and incorrect addresses. For example:

- cases where the couple in the admin household was the same as the census couple, but the admin household included children aged 17 or above who, in the census dataset, were at a different address. Admin addresses are particularly poor for young adults who tend to move more frequently (Stats NZ, 2023b).
- cases where both parents in the census family were in the admin household but the partnership between them was missing. As noted earlier, we estimate that around 19.5 percent of partnerships are missing in admin data (see Partnerships for methodology).
- cases where an ex-partner (now living at a different address) was included in the admin household, so the admin family-type error was caused by both an incorrect address and an incorrect partner relationship. Somerville et al (2025) noted that the 'separated' partnership category is challenging to derive from admin data, and that around half of people who are separated are still recorded as partnered in admin data.

Table 4

Agreement/disagreement of family type between admin and 2023 Census, for linked selected families with an identifiable family type				
	Family type of the admin family			
	Couple without children	Couple with child(ren)	One parent with child(ren)	**Total number of families**
Family type of the census family	Count of linked families (percentage of row)			
Couple without children	255,162 (89.9)	22,908 (8.1)	5,769 (2.0)	283,839 (100)
Couple with child(ren)	10,008 (2.9)	303,705 (87.8)	32,382 (9.4)	346,095 (100)
One parent with child(ren)	600 (0.6)	9,525 (10.3)	82,779 (89.1)	92,904 (100)
Total number of families	**265,770**	**336,138**	**120,930**	**722,838**

Note: Fixed random rounding has been applied to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

See Appendix 2 for information on how families were selected for comparison.

Percentages are the fraction of the census response for each family type of admin family and add to 100 percent for each row.

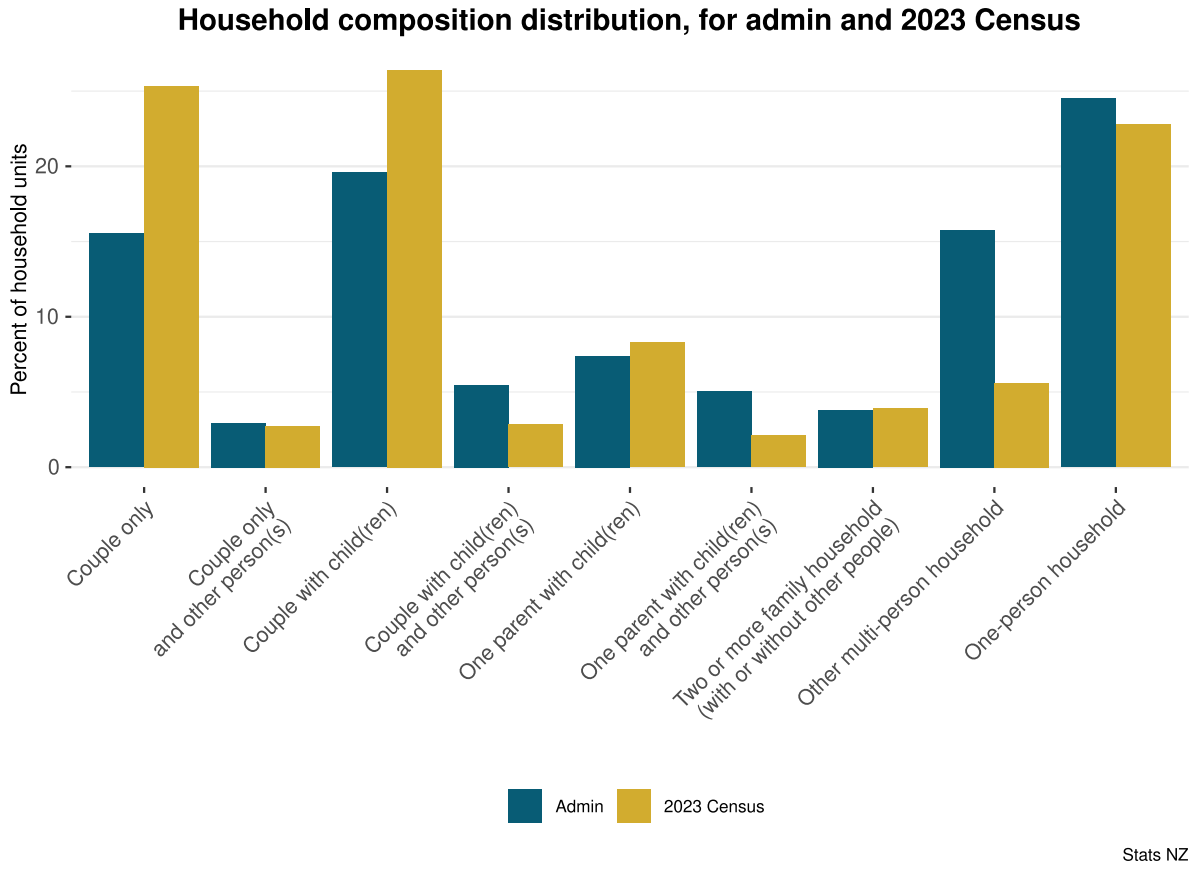
Source: Stats NZ

Household composition distribution

Household composition (see concept definition of household composition) classifies households according to the relationships between usually resident people. Figure 5 compares the admin household composition distribution with 2023 Census for households with an identifiable household composition. Note that two- and three or more family households have been combined.

Admin overestimates households where there are non-family members, particularly multi-person households, and couples or one person with children and other people. To a lesser extent, admin overestimates one-person households. Admin underestimates households with couples (both with and without children). These under and overcounts are generally due to missing partnership relationships in the admin data and/or incorrect admin addresses for individuals.

Figure 5



Households containing children aged under 15 years without a parent

Census aims to ensure that all children aged under 15 years are part of a family nucleus by assigning a parent from within the household to any child aged under 15 years without a parent. Assigned parents need to be 18 years or over, and the order of assigning a parent is: grandparent, great-grandparent, pibling, sibling, other related person, unrelated person. The relationship between an assigned parent and the child is then changed to a parent–child relationship.

For admin households, there are 89,118 children (in 69,507 households) aged under 15 years who do not live with a parent. Of these, there are 22,896 children (25.7 percent) who would have someone suitable in the household that could be assigned as an ‘other adult in a parent role’. For 2023 Census households, there were 23,145 children (in 15,972 households) aged under 15 years who did not live with a parent. Of these, there were 20,874 children (90.2 percent) who had someone assigned as an ‘other adult in a parent role’.

There are a comparatively large number of admin households where everyone is aged under 15 years (10,035 admin households compared with 471 census households).

Biases and limitations in our methodology and data

The methods described in this paper are designed to maximise overall agreement with census measures of households and families. However, summary quality measures can obscure important limitations for particular population groups, living arrangements, or relationship types. This section highlights several methodological and data-related limitations that may influence the interpretation of the results and suggests areas where further investigation is required.

Address selection methodology for children

XGBoost is used to select the most likely usual residential address for an individual. This is followed by a correction to addresses assigned to individuals aged 0 to 17 years: if a person has at least two parents, and the two parents with the most recently recorded parental relationship have the same predicted admin address (which differs from that selected for the individual), then the individual is reassigned to the parents' admin address. This adjustment was made based on evidence of common occurrences of address updates for children lagging behind those of their parents after a move.

This method is used in:

- the admin address selection method considered in this paper
- the third iteration of the APC for household assignments
- 2023 Census admin enumerations.

In our implementation, the adjustment was applied to 4.7 percent of admin children. We estimated, based on linked admin/2023 Census comparisons, that it corrected the child's address in 74.6 percent of cases but moved 17.3 percent of children away from their true address. Around 10.4 of children moved by this rule were moved to a two-parent family when 2023 Census data indicated this is not representative of their true living situation.

While this approach improved overall address agreement it does mean some less common living arrangements may be missed. Children living with relatives acting as parents may be reassigned back to birth parents and, where more than two parents are recorded, the two most recently recorded parents are used, which may also misplace some children.

Any biases in 2023 Census data due to this adjustment will be negligible. In 2023 Census, 1.1 percent of individuals aged between 0 and 17 are moved due to this rule, and a vast majority of the time they will correctly be assigned to a two-parent family.

Biases in relationship methodology

Removing records with unlikely age-relationship combinations may lead to under-representation of families with less common structures. Assigning a single relationship for each pair of individuals is necessary for producing household and family statistics but may not fully reflect lived realities. Because parent–child relationships combine legal records (such as birth registrations) with social relationships recorded in administrative programmes, caregiving relationships may be obscured when a different relationship from a higher-priority source exists. For example, a child raised as whāngai by a pibling may appear only as a nibling–pibling relationship in the data, and the caregiving role may not be recognised.

Consistency measures

Consistency of admin-derived households, families, and relationships is assessed by comparison with 2023 Census responses, which are treated as the most reliable available benchmark. 2023 Census admin enumerations were excluded to avoid comparing admin data with values that were themselves derived from admin sources. While this improves the independence of the benchmark, restricting the comparison to census respondents may introduce bias.

Census respondents may differ systematically from individuals who did not respond to the census. For example, people who complete a census form may also be more likely to keep their admin information up to date, which could increase agreement between census and admin addresses.

Response rates also vary across population groups. In the 2023 Census the overall response rate was 87.6 percent, compared with 76.5 percent for Māori ethnicity and 79.2 percent for Pacific peoples ethnicity (Stats NZ, 2024c). Additionally, linkage rates between 2023 Census and the IDI spine vary across groups with European ethnicity having a 99.0 linkage rate and Pacific people having the lowest linkage rate at 97.9 percent (Stats NZ, 2024b).

Because consistency measures are calculated using 2023 Census/admin linked respondents only, differences in response and linkage rates may limit the meaningfulness of comparisons across groups.

Additional restrictions were applied when constructing the comparison dataset for household membership agreement to ensure a reliable benchmark. Specifically, we excluded households with nine or more members, households classified as residual, and households where some members were admin enumerations. While these restrictions improve the quality of the comparison, they may under-represent some household compositions. Larger households also have more opportunities for address misassignment, which may influence observed agreement rates.

Consistency measures presented in this paper should therefore be interpreted as indicators of agreement with census responses under the comparison conditions used, rather than as direct estimates of the true accuracy of admin-derived households and families.

Biases in 2023 Census household and family distributions

The 2023 Census count of households in the output subject population (households in occupied private dwellings, not visitor-only) included 61,728 ‘blank’ households where we did not place any usual residents (see Appendix 1 for more information). Census household attributes, such as household composition, were only derived for the households where we placed at least one usual resident; therefore, there will be a bias if the blank households differ from these.

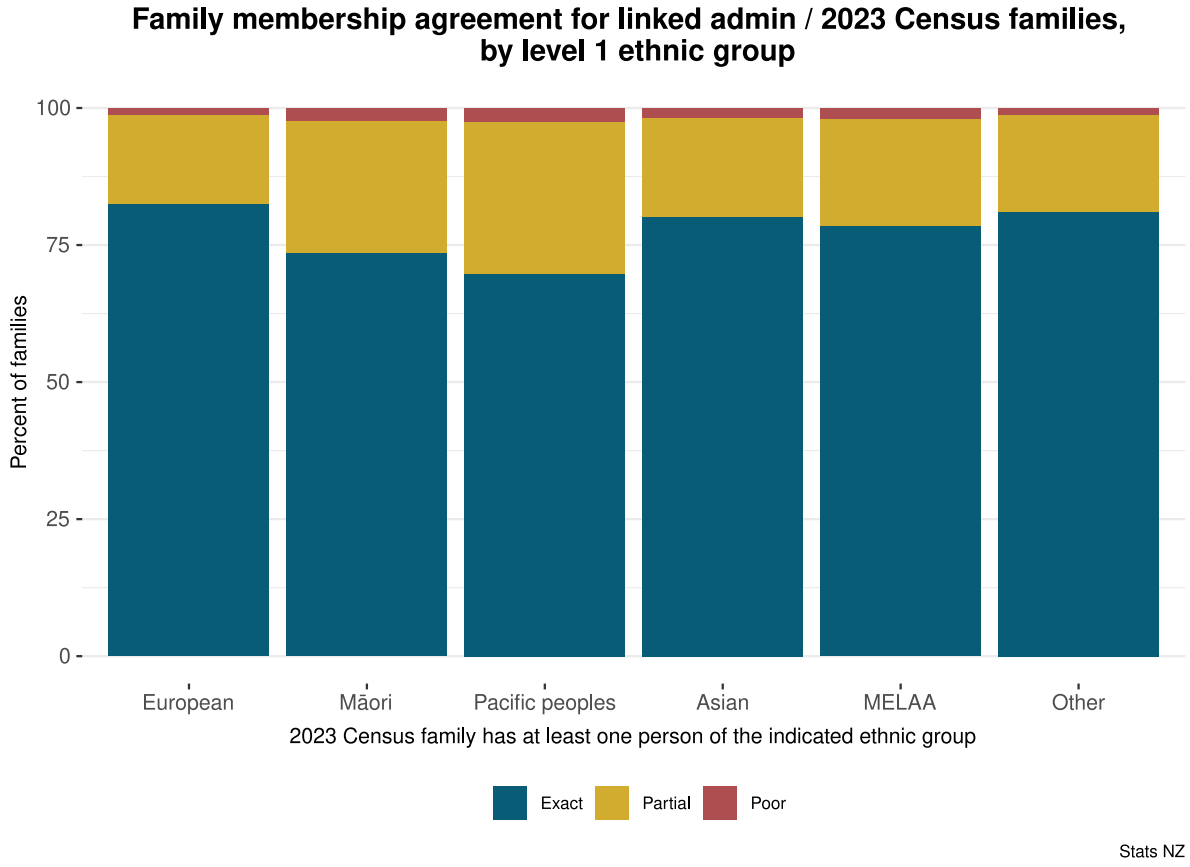
There is a flow-on from this for families which were constructed from households where we placed at least one usual resident: there is an undercount of census families (those that belong to the blank households) and a corresponding potential bias if the missing families differ from the families that were constructed.

Family membership agreement by level 1 ethnic group

As shown in figure 6, family accuracy was not the same for all level 1 ethnic groups. Families with at least one person belonging to the Pacific peoples ethnic group has the lowest exact membership agreement (69.7 percent), followed by families with at least one person belonging to the Māori ethnic group (73.6 percent). Families with at least one person belonging to the remaining level 1 ethnic groups (European, Asian, MELAA, Other) have exact membership agreement rates closer to

the overall rate of 81.1 percent. Further investigation is required. For example, 2023 Census results show that there are differences in household size, family size, and family type structure across different ethnicities (see How we live – a look at families and households from the 2023 Census and Aotearoa Data Explorer) and this could be a contributor to the differences in family accuracy reported here.

Figure 6



The scope for methodological improvement

The work above shows that the creation of admin households and families is limited by the quality of admin address information available for a person. The current approach for assigning one admin address to a person treats all individuals independently. In reality, people in a household often move between addresses together. Also, individuals have a higher probability of living with people they share a relationship with. This leads to the natural idea (for selecting individual addresses and forming households) of linking people to multiple addresses via address notifications and linking people together via relationships. Note that while this method shows promise it is still in the early stages of development.

Statistics Estonia has successfully worked to combine address and relationship information to improve household and family statistics (United Nations Economic Commission for Europe, 2023). Motivated by this, we developed a methodology to improve address selection by incorporating relationship information. The preliminary results from this method show small improvements, with address agreement increasing by 0.6 percentage points and exact household membership agreement by 3.5 percentage points. It is likely that as more admin relationship data becomes available the performance of the method will improve.

Method for using relationship information to improve address selection

Our new (exploratory) method involves combining the following two XGBoost models:

1. An XGBoost model that outputs weights for each person-address pair where the person has an address notification available for that address.
    a. This model is currently used as Step 1 in the current method (described in Addresses and households).
    b. Each person-address pair is evaluated separately by XGBoost – meaning that the sum of all person-address weights for a given person may sum to more than one.
2. A separate XGBoost model is run on pairs of related people. This model outputs weights representing the probability that the two people live together. We refer to this as a ‘cohabitation weight’.

We then use the following method to combine the information from both models:

1. A logit transform is applied to the weights produced by each XGBoost model. Different transforms were considered, but logit gave optimal results and is likely to remain consistent over time.
2. People are assigned to addresses in a way that globally maximises the sum of the within-household weights.
    a. A household is the address and all people assigned to that address.
    b. Within-household weights are defined as the weights between household members and weights between household members and the household address.

Conceptually this moves away from the current approach of treating addresses and relationships separately (see figure 1) to combining both pieces of information to construct households as shown in figure 7.

Figure 7

Constructing households and families from admin data using the new method for address selection

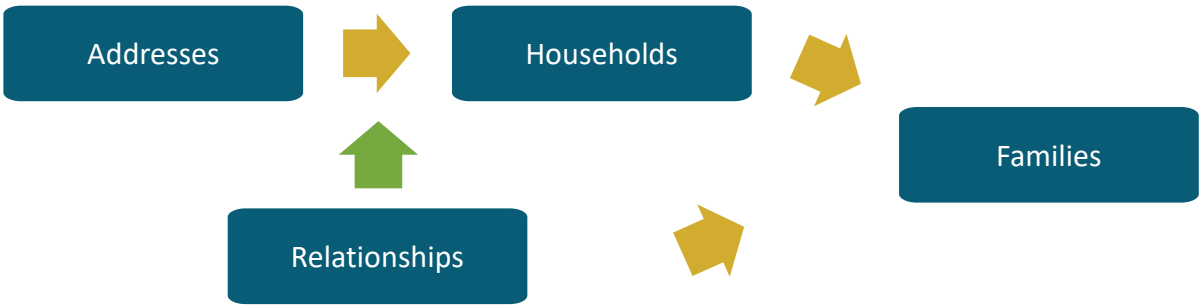


Table 5 summarises membership agreement for households created using the current method and the exploratory method. Overall, there is a 3.5 percentage point increase in exact household membership agreement when relationship information is included. The biggest improvement is seen for Pacific peoples households (that is, households containing at least one person of Pacific peoples ethnicity) with a 5.2 percentage point increase in exact membership agreement, and lowest for European households with a 3.3 percentage point increase in exact membership agreement.

Table 5

Household membership agreement for linked admin / 2023 Census households, by level 1 ethnic group

	Current method			Exploratory method		
	Exact	Partial	Poor	Exact	Partial	Poor
Ethnic group[1]	Percentage of households					
Overall	68.3	26.0	5.7	71.8	22.8	5.4
European	71.2	23.9	4.9	74.5	20.8	4.6
Māori	56.4	35.3	8.3	60.9	31.3	7.8
Pacific peoples	50.3	41.2	8.5	55.5	36.5	8.0
Asian	58.4	33.7	7.9	62.9	29.6	7.5
MELAA	57.4	34.2	8.4	62.0	30.1	7.9
Other	68.7	26.6	4.7	72.6	23.0	4.4

1. For the linked admin / 2023 Census households, the 2023 Census household has at least one person of the indicated ethnic group.

Note: Percentages are the fraction of linked admin / 2023 Census households eligible for comparison for each level 1 ethnic group.

Due to rounding, totals for each ethnic group may not add to 100 for each experiment.

MELAA – Middle Eastern/Latin American/African.

Source: Stats NZ

While generally including relationship information leads to improved address assignments and household membership agreement, this method creates too few households (1.78 million in the

exploratory method compared to 1.8 million using the current method, and to 1.91 million in the 2023 Census).

Potential for future improvements

We have identified several areas of research that could enhance the robustness of the address selection methodology.

Introducing a mechanism to avoid the creation of too many implausible households

The methodology described in this document currently has no mechanism to discourage over-saturated addresses or otherwise improbable households.

- A small adjustment could be applied to addresses with more than a fixed number of people assigned, so that where two assignments are almost as good as each other, the assignment with more realistic address sizes is chosen even if it has slightly lower overall weight.
- A mechanism could be applied to reduce the number of implausible households created by our methodology. The plausibility of a household could be found through another trained XGBoost model.

If these suggestions are implemented, it will be important that we ensure that our definitions of 'implausible' and 'over-saturated' are evidence-based and are able to be changed over time. This would be monitored through comparison with survey benchmarks.

Expanding and/or refining the methodology of using person–person relationship information

This work is still in the early stages and many avenues for improvement are yet to be explored. In this section we provide suggestions for further exploration.

1. Work is needed to reduce impacts of summing dependent cohabitation weights (for example, 'both parents live at this address' pulling a child who may have left home to that address, or 'a parent and child live at this address' pulling a separated parent to the address). To mitigate this, we could:
    a. Require the maximum contribution from cohabitation weights per person to be capped.
    b. Incorporate negative weights when two people are unlikely to live together. Examples of cases where negative person–person weights may improve outcomes include dissolved relationships, or young adults who we believe to have recently moved from their family home. This would need careful monitoring to ensure it does not adversely affect people who live together intermittently.
2. Currently, cohabitation weights only exist for individuals with relationships that can be found using data available in the IDI. There is potential to train person–person cohabitation weights using overlapping address notifications or other information available about interactions between people (such as sharing a child).
3. Introducing longitudinal checks on household memberships over time could help distinguish persistent living arrangements from short-term noise. Care would need to be taken with this to ensure that what we categorise as 'noise' is not a result of short-term moves.

Looking ahead

The analysis in this paper shows both the potential and the limitations of using admin data to derive household and family information. Admin data provides rich information about individuals, addresses, and relationships. However, the work also shows that, unless significant improvements are made, admin data on its own is not sufficient to produce basic household and family statistics at the quality required for the next census.

Survey data

The new census annual sample survey will need to play a fundamental role in collecting household and family information. When pooled over time, this survey will support a broad range of household and family analyses by geographic and demographic characteristics. However, this information will be less granular than the household and family outputs users have been accustomed to from past censuses.

The most important next step is to develop a detailed estimation strategy for household and family information derived from the new census annual sample survey. This can be developed using data from existing household surveys and from the census survey design and readiness test.

This should not be interpreted as meaning that admin data contains no information on households and families. The findings in this paper suggest that admin data can provide a substantial amount of useful structural information. The challenge is that this information does not currently support a direct replacement of census-style household and family outputs.

Therefore, the next step after the design of survey-only estimation is to design and develop a combined approach for household and family information. In such an approach, survey data would provide the direct information needed to measure household and family concepts, while admin data could be used to support estimation, increase coherence across outputs, and help explain or model differences across population groups and living arrangements.

Developing such an approach will require research into how household and family concepts can be estimated using both survey and admin sources in a coherent way, and how uncertainty and bias can be assessed when the two sources differ. There is an opportunity to explore, with our communities, Te Tiriti partners, and stakeholders, how household and family concepts could evolve to better reflect their reality and meet their information needs.

Admin data quality evolution

The current findings should be understood in the context of a changing data environment. Admin data sources and linkage systems continue to evolve, and the quality of address and relationship information available to Stats NZ is not fixed. Some of the current limitations will lessen over time as new sources become available and methods are refined. Work is underway to improve admin data quality and the frequency with which it is collected across the government data system. Stats NZ has identified the key data sources required for delivering the next census, and we are working with government agencies to supply this data. Addresses are part of this effort together with other concepts such as ethnicity. A focus is to improve data for younger and older people, and recent immigrants to New Zealand.

Almost all people have a notification for their true usual residential address. However, some people also have notifications for other addresses, and our address selection methodology may erroneously select one of these as the person's usual residential address. There is scope to improve address assignment by using address and relationship information jointly rather than treating each person independently. The preliminary work described in The scope for methodological improvement suggests that this direction is promising, even though it is at an early stage.

There is also scope to improve relationship coverage through the inclusion of additional sources as these become available, particularly those that provide explicit relationship information for immigrants and for family formations that are currently difficult to observe. Additionally, for the next census, relationship information from the 2018 and 2023 Censuses will be available as historical sources. However, partnerships will benefit less from this than other familial relationships due to the fact that they change over time. Together with information from the new census annual sample survey and other household surveys, these sources will improve coverage of familial relationships that are currently missing from admin-only derivations. Note that partnerships will continue to be the most challenging relationship to derive accurately from admin data.

The work presented in this paper has been undertaken in the IDI, which contains a rich set of information about individuals but relatively limited information about places. Future development of the Integrated Statistical Data System is expected to provide a stronger foundation for linking information about people, places, and businesses. This richer data environment will improve our ability to understand dwellings, addresses, and residential arrangements, and therefore reduce some of the current limitations in household formation.

For these reasons, the quality of admin-derived household and family information should not be viewed as fixed. It will be necessary to reassess this quality as data sources, linkage quality, and methods change.

Customer-centric and collaborative approach

Future work in this area should not be seen as only a technical exercise. Household and family information is used in a wide range of policy, planning, and research settings, and different users are interested in different aspects of household and family life. Some users are primarily interested in standard census-style outputs, while others may be particularly concerned with concepts that are not well captured by standard admin definitions, including whānau structures, shared care arrangements, and social parenting roles.

This means that future work should be informed by engagement with data users, subject matter experts, Te Tiriti partners, researchers, and other government agencies. Such engagement is important both for identifying where current methods fail to reflect lived realities and for helping to prioritise areas where additional data sourcing or methodological development would be most valuable.

Methodological transparency will also remain important. Clear communication about what admin data can and cannot currently support will be necessary if household and family outputs are to remain credible and useful. This includes transparency about trade-offs, about groups for whom outputs may be less accurate, and about the distinction between legal, admin, and lived family relationships.

Overall, the work described in this paper should be understood as part of a broader system transformation. The move to a modernised census is not simply a matter of replacing one data

source with another. It requires the development of new methods, new combinations of data, and new ways of assessing and communicating quality. Within that transformation, household and family information remain particularly challenging concepts.

The findings in this paper indicate that admin-only household and family outputs are not yet of sufficient quality for the next census. However, they also show that admin data has an important role to play in improving future household and family statistics when used in combination with survey data and ongoing methodological development.

References

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

Somerville, W, Harman, P, and Simpson, K (2025). Assessing the quality of partnership status attributes from admin data. Retrieved from www.stats.govt.nz.

Stats NZ (2017). Experimental population estimates from linked administrative data: 2017 release. Retrieved from www.stats.govt.nz.

Stats NZ (2022). Experimental administrative population census: Data sources, methods, and quality (second iteration). Retrieved from www.stats.govt.nz.

Stats NZ (2023a). Experimental administrative population census (third iteration): Data sources, methods and quality for household information. Retrieved from www.stats.govt.nz.

Stats NZ (2023b). Experimental administrative population census (third iteration): Information by variable. Retrieved from www.stats.govt.nz.

Stats NZ (2024a). Families and households in the 2023 Census: Data sources, methodology, and data quality. Retrieved from www.stats.govt.nz.

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

Stats NZ (2024c). Post-enumeration survey: 2023. Retrieved from www.stats.govt.nz.

Stats NZ (2024d). Predicting usual residence address from admin data in the 2023 Census. Retrieved from www.stats.govt.nz.

United Nations Economic Commission for Europe (2023). Estonian 2021 census: determining households and dwellings from registers using graphs. Retrieved from unece.org.

Appendix 1: 2023 Census household counts

Table 6 provides more detailed information about 2023 Census households than the counts shown in Addresses and households.

Table 6

2023 Census household counts		
	Count of households	**Description**
Households in the census output subject population (occupied private dwellings, not visitor-only)		
Households with at least one usual resident	1,718,799	• Dwelling is private • Dwelling has a household with usual resident(s) • Some or all usual residents in the dwelling on census night • At least one person responded to the census, or we placed at least one admin enumeration
Blank households	61,725	• Dwelling is private • Dwelling has a household with usual resident(s) • Some or all usual residents in dwelling on census night • No-one responded to the census, and we did not place any admin enumerations
Total	1,780,527	Total number of households in the output subject population
Households outside the census output subject population		
Households with at least one usual resident	50,430	• Dwelling is private • Dwelling has a household of usual residents • No usual residents in dwelling on census night – there may or may not be visitors in the dwelling on census night • Usual residents placed in the dwelling via repatriation or admin enumeration
Blank households	74,604[1]	• Dwelling is private • Dwelling has a household with usual resident(s) • No usual residents in dwelling on census night – there may or may not be visitors in the dwelling on census night • No usual residents placed in the dwelling via repatriation or admin enumeration
Total	125,034	Total number of households outside the output subject population
Grand total	**1,905,561**	**Total dwellings where we would expect some usual residents**

1. This includes 10,932 households in occupied, visitor-only private dwellings where we did not place any usual residents.

Note: Fixed random rounding has been applied to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

Source: Stats NZ

Appendix 2: Admin households and families excluded from the unit record comparison

Of the original 1,801,383 admin households and 1,718,799 census households in which at least one usual resident was placed, there were 1,278,567 households (71.0 percent of the total admin households) that could be selected for membership agreement comparison. The remaining 522,816 admin households were excluded from the comparison for prioritised reasons described in table 7.

For households, we firstly wanted to compare one admin household with one census household at the same address. We therefore had to exclude admin addresses that could either not be found in census, and/or contained multiple census dwellings. Secondly, we removed admin households that were linked to a census household outside the census household subject population (see Appendix 1). We then removed admin households that were linked to a census household with admin enumerations because we did not want to compare admin with admin. Finally, we excluded admin households linked to poor-quality census households because we were trying to isolate admin household errors.

Of the original 1,160,004 admin families and 1,294,503 census families, there were 740,265 families (63.8 percent of the total admin families) selected for membership agreement comparison. The remaining of 419,739 admin families were excluded from the comparison for prioritised reasons described in table 7.

Table 7

Counts of admin households and families excluded from the unit record comparison

Prioritised list for removal of households/families from unit record comparison	Count of admin households removed from comparison (percentage of column)	Count of admin families removed from comparison (percentage of column)
Households/families that could not be linked (by address) to census because an address match could not be found[1]	175,239 (33.5)	113,823 (27.1)
Households/families that were linked to an address with multiple dwellings[2]	21,018 (4.0)	20,097 (4.8)
Households/families that were linked to a census household/family that was not in the census household/family output subject population (occupied private dwelling – not visitor only)	22,419 (4.3)	9,444 (2.2%)
Households/families that were linked to a census household/family that included an admin enumeration[3]	94,845 (18.1)	47,166 (11.2)
Families that were at an admin address where either the admin household and/or the linked census household had more than one family	N/A	138,558 (33.0)
Households/families that were linked to a census household/family that was unlikely to be correct[4] due to any of the following conditions: • not everyone in the census household had an individual form • for households, the census household had been classified as residual[5] • for families, the census household contained imputed relationships • the number of people in the census household was ≥ 9	209,298 (40.0)	90,651 (21.6)
Total	**522,816** **(100)**	**419,739** **(100)**

1. Possible reasons why an admin address could not be found in census included mobile and temporary dwellings where no address existed; address matching error; incorrect dwelling-type classification in either the APC or the census.

2. These are households where admin data created one household but where there should be several households.

3. These were out of scope for the comparison because we did not want to compare admin with admin.

4. We chose to exclude these poor-quality census households from the comparison because when the admin and census household membership differed, we could not be sure whether it was the admin household or the census household that was incorrect.

5. Households with quality issues that had not been resolved by the end of census processing.

Note: Fixed random rounding has been applied to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

Source: Stats NZ

Appendix 3: Admin familial relationships by source and relationship type

As mentioned in 'Relationships beyond partnerships', parent-child relationships were all sourced directly from admin data and most other familial relationships were created through inference.

Table 8 shows a full breakdown of within-household admin familial relationships by source and relationship type.

Table 8

Counts of within-household admin familial relationships by source and relationship type

	Count of within-household admin relationships[1] (percentage of column)					
Source of admin relationship	Parent-Child	Sibling	Grandparent-Grandchild	Great-grandparent - Great-grandchild	Pibling-Nibling	**Total**
DIA birth register	1,997,958 (83.5)	0 (0)	0 (0)	0 (0)	0 (0)	**1,997,958 (53.7)**
MSD benefit data	112,527 (4.7)	0 (0)	0 (0)	0 (0)	0 (0)	**112,527 (3.0)**
WFF data	75,981 (3.2)	0 (0)	0 (0)	0 (0)	0 (0)	**75,981 (2.0)**
MBIE immigration data	136,494 (5.7)	0 (0)	0 (0)	0 (0)	0 (0)	**136,494 (3.7)**
2013 Census	69,945 (2.9)	56,520 (5.7)	19,056 (10.1)	432 (8.1)	5,040 (3.7)	**150,993 (4.1)**
Inferred	0 (0)	940,860 (94.3)	169,740 (89.9)	4,920 (91.9)	132,609 (96.3)	**1,248,129 (33.5)**
Total	**2,392,905 (100)**	**997,380 (100)**	**188,796 (100)**	**5,352 (100)**	**137,649 (100)**	**3,722,082 (100)**

1. The reference date for the admin familial relationships dataset is 7 March 2023 (date of the 2023 Census).

Note: Fixed random rounding has been applied to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

Percentages are the fraction that each source contributed to a relationship type, and add to 100 percent for each column.

DIA – Department of Internal Affairs; MSD – Ministry of Social Development; WFF – Working for Families; MBIE – Ministry of Business, Innovation and Employment.

Source: Stats NZ