May 2020

# Constructing households from linked administrative data

An attempt to improve address information in the IDI

SOCIAL WELLBEING AGENCY | TOI HAU TĀNGATA

New Zealand Government

## Authors

Simon Anastasiadis[†], Akilesh Chokkanathapuram[†], Craig Wright[†]

## Acknowledgements

Megan Gath[‡], Christine Bycroft[‡], Sharon Snelgrove[‡], Miranda Devlin[•], and Marianna Pekar[†] all provided review of this paper.

Robbie Batley[°] and Matthew Bray[°] identified flaws in an earlier version of the analysis.

Special thanks to the Stats NZ Integrated Data Team, without whom we would not have the valuable resource that is the IDI.

† = Social Wellbeing Agency. ‡ = Stats NZ. • = Ministry of Housing and Urban Development. ° = Taylor Fry.

## Creative Commons Licence




## Integrated Data Infrastructure disclaimer

The results in this paper are not official statistics, they have been created for research purposes from the Integrated Data Infrastructure managed by Statistics New Zealand. The opinions, findings, recommendations and conclusions expressed in this paper are those of the author(s) not Statistics NZ, or other government departments.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation and the results in this paper have been supressed to protect these groups from identification.

Careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the Integrated Data Infrastructure. Further detail can be found in the Privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes.

Any person who has had access to the unit record data has certified that they have been shown, have read, and have understood section 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the Integrated Data Infrastructure for statistical purposes and is not related to the data's ability to support Inland Revenue's core operational requirements.

## Liability

While all care and diligence has been used in processing, analysing and extracting data and information in this publication, the Social Wellbeing Agency gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.

## Citation

Social Wellbeing Agency 2020. Constructing households from linked administrative data: An attempt to improve address information in the IDI. Wellington, New Zealand.

ISBN 978-0-473-52420-3 (online)

**Published in May 2020 by**
Social Wellbeing Agency
Wellington, New Zealand

# Summary: Further work is needed to construct households in the IDI

Good household information in linked administrative data is highly desirable because of the significant links between individuals' experiences and the experiences of other people in their household. Examples include childcare, exposure to second-hand smoke, and household income. However, to date, the address information available in the Integrated Data Infrastructure (IDI) has not been accurate enough for researchers to be confident using it to identify households.

We attempted to improve the quality of address information in the IDI by applying a series of detailed cleaning rules to remove patterns of incorrect or untrusted information. The performance of our resulting table was validated against Census 2013 but showed no improvement over the existing address table in the IDI.

The main point of this publication is **not** to suggest that we now have methods for improving household information in the IDI. Rather it is to warn other IDI researchers not to duplicate our work, and to provide our learnings so that other efforts might be more successful.

We anticipate that IDI researchers may find our paper useful for:

- Updated estimates of household accuracy (to the September 2019 refresh).
- Demonstration of how individual address accuracy varies with age.
- Estimates of the theoretical limit on household accuracy.
- Discussion of the challenges of working with address information.

# Table of Contents

# Households are seldom straightforward to infer from linked administrative data

Administrative data, though collected for non-statistical, non-research purposes, are now being used for conducting research. Linked administrative data, identifying where the same individual appears across multiple datasets, can provide a more complete view of an individual than relying on a single dataset.

One challenge that arises from linked data is synthesising a single consistent set of information from different data sources (Hand, 2018). This challenge is obvious when seeking to determine an individual's address as many organisations collect this information but the accuracy and frequency with which address information is updated varies.

Address information is often the starting point for the identification of households. For the purposes of this paper, we consider a household to be the group of people who share the same dwelling.[1] Given that constructing a consistent dataset of individuals' residential addresses is not straightforward with linked data, working across individuals to identify households is even more challenging. This has implications for any analysis that seeks to look beyond the individual.

International interest in linked data often focuses on the use of government administrative data as an alternative to a traditional census. This includes effort in the Abu Dhabi Emirate (Statistics Centre, 2015), Australia (Mowle and Watmuff, 2018), Canada (Trépanier et al., 2014), India (Pronab, 2008) and a range of European countries (Valente, 2010). However, an administrative census is restricted by the topics and definitions that are collected (Tønder, 2008). Unless household membership is collected directly, researchers will need to construct it from individual information.

In New Zealand, the Integrated Data Infrastructure (IDI) is a significant repository of administrative government data, linked for research purposes. At least nine government organisations collect address information and provide it into the IDI. Work has been done to combine these different sources into a single table: The existing approach uses a person's most recent address as the best estimate for their residence at a point in time and allows less trusted sources to be superseded by more trusted sources (Stats NZ, 2015a). Several subsequent applications have identified limitations in the accuracy of the existing table (Stats NZ, 2017 & 2019, Gath and Bycroft, 2018) and to date the available address table has not been accurate enough that researchers can use it to identify households with confidence.

The Social Wellbeing Agency constructed a new address table within the IDI in an attempt to provide better identification of households. This followed conversations with staff from a range of organisations (including The Treasury, Oranga Tamariki, Ministry of Social Development, Accident Compensation Corporation, University of Auckland, and University of Otago) who expressed a need for improved address information.

[1] This differs from, but has significant overlap with, the definition used for official statistics in New Zealand: "One person who usually resides alone or two or more people who usually reside together and share facilities (such as eating facilities, cooking facilities, bathroom and toilet facilities, a living area)" (Stats NZ, 1999a & 1999b).

Our approach was to apply detailed rules to remove misinformation by identifying trusted and non-trusted address records. These rules look for patterns of address information over time and across people associated with the same household. The table constructed by our analysis was validated against a national census. While every rule we developed improved overall accuracy, our approach failed to provide a significant improvement over the existing table.[2]

This paper covers the data, method and accuracy of our constructed table, and discusses the challenges and opportunities for further improvement. It is intended primarily to enable researchers working with linked administrative data to build on our work to create high-quality address and household information.

# Administrative data contains address notifications not address changes

Residential address is often recorded as a person's present address at the time of their interaction with the recording organisation. Depending on the organisation, a second residential address and/or a postal address may be optionally recorded. We consider these types of records to be address notifications – notice of a person at an address at a specific time.

Address change information records that a person changed to a new address and when they changed. There are five key ways that address notifications differ from address change information:

1. Address notifications can occur even if the person has not changed address. A common example of this is where an organisation asks you to confirm your address every year.
2. An organisation may not have address notifications for every address a person resides at. This is more likely for people who move often or who interact with the organisation infrequently.
3. Where an address change has a corresponding address notification, there will often be a delay between them, so the date of the notification does not match the date of the address change.
4. Address notifications do not always distinguish between residential and non-residential (e.g. postal) addresses. This can be observed where young adults who live away from home while studying provide organisations with their home address as a postal address instead of their residential address; and where investors provide the address of their accountant for tax purposes instead of a personal address.
5. Address notifications can occur for purely administrative reasons that do not reflect a person informing an organisation of their address. Automated reporting, changes in software systems, and staff under time pressure not updating addresses can all result in notifications that do not reflect the address of the person at the reported time.

Address information in the IDI takes the form of address notifications and contains examples that differ from address changes in each of the five ways identified above. It follows that the goal of

[2] Our calculation of accuracy may vary from previously published numbers due to differences in validation methodology. Unless noted otherwise, all results were produced during our analysis and hence reflect a consistent validation approach.

our analysis is to construct the best possible estimate of address changes given a collection of address notifications.

In some contexts, information may be provided or used as intervals of time (address spells) not as points of time (notifications or changes). However, as intervals of time are defined by their start and end points, the middle of an interval provides no additional information. Hence analysis of addresses can focus on those points in time when address notifications or changes take place.

## There are challenges when using linked administrative data for addresses

When working with address information from a single organisation the best estimate for an individual's address at any point in time is given by the last notified address. While this is a strong starting point it is not enough by itself to provide a best estimator for address change when working with linked administrative data.

Combining information from different data sources increases the amount of information available and this means the correct information is more likely to exist in a combined data collection. However, as the amount of information increases so too does the possibility for misinformation or conflicting information. Comparisons between different data sources and between people across different data sources are key to realising the value of address information in linked data.

Synthesising a single consistent set of information from different sources of address notifications requires a range of challenges to be overcome. These include:

1. Standardised coding of addresses. Where different sources record addresses differently, a common coding is required against which every source can be mapped.
2. Resolving simultaneous notifications. Simultaneous notifications occur where the same source records multiple addresses for the same person on the same day. This could be due to mistakes entering an address – where the error and its correction are both recorded as notifications – or due to a lack of distinction between primary residential, secondary residential, and postal addresses.
3. Information may not be synchronised across organisations. This can lead to situations where two organisations have different addresses for the same person at the same time.
4. Members of the same household, who change address together, have subsequent notifications with different organisations at different times. This means that a naïve interpretation of the data would suggest a period in the middle when people who belong to the same household were living in separate addresses.
5. Some household members interact with data collecting organisations less frequently than others leading to a lack of notifications. This is more common for young children because they are less likely than adults to interact with a government organisation, as they do not work, study, receive government benefits, or drive.

Within the IDI, the first two challenges listed above are compounded by the de-identification process. Because address information is considered personally identifying, and the IDI is de-identified to protect individuals' privacy, all addresses have been replaced by numeric

placeholders we refer to as address IDs. This means that while a researcher can be confident that the same address ID represents the same address, there is no way to check via these codes whether two addresses should be treated as equivalent. Other forms of privacy protection, such as perturbating geospatial coordinates (Culhane, 2016), may be used in other data environments and have their own challenges.

The extent to which de-identification or privacy protections are a concern depends on the quality of the address matching software, and the reference address list that text addresses are compared with. Addresses are complex and idiosyncratic, and a reference address list must be able to associate different descriptions of the same 'addressable object' with a single address ID. Address matching in the IDI has improved recently with the introduction of a Statistical Location Register and new address matching software.

A further challenge in the IDI is the distinction between dwellings and addresses. In a small number of cases, a single address ID can refer to multiple dwellings with each dwelling containing its own household. This lack of distinction can result in situations where the occupants of multiple dwellings with the same address ID appear to form a single household. Such situations are easier to identify in larger or more formal contexts (such as an apartment block) but can easily go unnoticed in smaller or informal contexts (such as a private property subdivision).

All of the above assumes we can link administrative datasets together by determining where the same person appears across multiple data sources. This is a challenging problem in its own right. However, linking data sources or adjusting for incorrectly linked identities is beyond the scope of this paper and our work takes linked identities as given in the IDI.

# We used every source of address information available

Linked administrative data lets us combine address notifications from a variety of sources. Although the potential for conflicting information increases with the number of data sources, if each source provides more information than misinformation, then increasing the number of sources is likely to improve the accuracy of the results.

Assessing the accuracy of a data source (and of any constructed, cross-source table) can not be done by comparing administrative data against administrative data. Some source of truth, that includes household membership, is required to validate against.

Thankfully, data linking is not restricted only to administrative data, but can also include other sources of information, such as censuses and household surveys. We consider such sources to be of high accuracy, and hence suitable to validate administrative sources against. Within the IDI, linked data includes a national census and multiple waves of three different household surveys

conducted by Statistics New Zealand (Stats NZ). The high quality of Census 2013 (Stats NZ, 2014) [3], and the in-person collection of the survey responses we use (Stats NZ, 2015b)[4], give us confidence using these sources as the truth against which to validate administrative data.

Table 1 and Table 2 list the sources of validation and administrative data (respectively) used in our analysis. The data drawn from these sources were expressed as address notifications.

**Table 1: Validation address data sources**

| Source | Description |
|---|---|
| **Census 2013** | Usual residential address as reported by the individual (or their caregiver) in Census 2013 (5 March 2013). |
| **General Social Survey (GSS)** | Observed address during an interview for the New Zealand General Social Survey. Combines information from the 2008, 2010, 2012, 2014, and 2016 survey waves. |
| **Household Economic Survey (HES)** | Observed address during an interview for the Household Economic Survey. Combines information from the 2007/08 wave through to the 2016/17 wave. |
| **Household Labour Force Survey (HLFS)** | Observed address during an interview for the Household Labour Force Survey. Combines information from the 2007 wave through to the 2018 wave. |

All three surveys and the census record household membership in addition to individual details. Census 2013 was the only data source at the time of our analysis that explicitly distinguished between address and dwelling.

**Table 2: Administrative address data sources**

| Source | Description |
|---|---|
| **Accident Compensation Corporation (ACC) claims** | Address given when submitting a claim for accident cover. |
| **Housing New Zealand Corporation (HNZC) tenancies** | Address by tenancy in state social housing as captured by regular snapshots of the social housing database. |
| **Inland Revenue Department (IRD) applications** | Address from tax records, dated by when it became valid. |
| **Inland Revenue Department (IRD) timestamps** | Addresses from tax records, dated by database timestamp. Records that duplicate the address-date pairs from the IRD applications are excluded. |
| **Ministry of Education (MoE) enrolments** | Enrolee address for attendance at primary, intermediate and secondary school. |
| **Ministry of Health (MoH) National Health Index (NHI)** | Address from an individual's National Health Index (NHI) record. The NHI is a unique, health identifier used across the health sector. |

[3] For Census 2013, paper survey forms were posted to every address and people filled out the forms that they received at their address. The census asked people to confirm whether they were completing the form at their usual residence, and if not to provide the address of their usual residence.

[4] The three Stats NZ surveys used (GSS, HES and HLFS) are conducted via a face-to-face computer assisted interview by a Stats NZ staff member.

| | |
|---|---|
| **Ministry of Health (MoH) Primary Healthcare Provider (PHO)** | Address provided to a patient's Primary Health Organisation (PHO). This is most often a General Practice (GP) doctor's office. |
| **Ministry of Social Development (MSD) residence** | Residential address reported by (primary) benefit recipient. For some benefits, where both members of a couple qualify to receive the benefit one partner is designated the primary recipient. |
| **Ministry of Social Development (MSD) postal** | Postal address reported by (primary) benefit recipient where differs from residential address. |
| **Ministry of Social Development (MSD) partner** | Residential address reported by the primary benefit recipient applied to their partner, for couples where both qualify to receive a benefit. |
| **Ministry of Social Development (MSD) child** | Residential address reported by a benefit receiving adult, applied to the children they report themselves as caring for. |
| **New Zealand Transport Authority (NZTA) driver licensing** | Address given at registration, endorsement, or renewal of a driver license. |
| **New Zealand Transport Authority (NZTA) vehicle licensing** | Address given at latest registration of a vehicle, primarily motor vehicles, but includes some trailers, chasses, and mobile machines. |

Each of the organisations that provide address information into the IDI do so in different formats. Before our analysis, all the administrative address notifications were combined in a standard format that captures four key details:

1. Source
2. Cross-source person identifier
3. Address ID
4. Date of notification

In addition to address notifications, for our work in the IDI, we used three pieces of non-address information. First, date of birth is available from a range of sources and lets us estimate a person's age at each notification.[5] Age is important because the frequency with which people change address varies over the life course. Second, birth records allow us to link children with their birth parents. This link helps us ensure children are resident with their guardians. Third, Stats NZ provides an estimated residential population as at 30 June each year derived from the linked administrative sources in the IDI (Stats NZ 2017 & 2018). The estimated residential population is used to focus our validation on New Zealand residents, excluding tourists, temporary visitors, and identities with incomplete linking.

[5] As the IDI has been de-identified, day-of-birth is not available. Hence, we can only estimate a person's age based on year and month of birth.

# Each source was checked for stability over time

Administrative address records can occur for reasons other than address notifications. Some of these causes can be identified by examining how the number of addresses changes over time for a given source. For example, a hardware migration could result in a record for every individual at the time of the migration. Under this scenario, the isolated increase in the number of records per day provides an indication that something unusual has taken place.

As part of selecting the data for our analysis, each data source was checked for days where the number of records was significantly higher or lower than usual. This identified several days in the IRD data where the number of records was 10 to 100 times larger than the number of records on equivalent days. Documentation supported our concern that these records did not reflect notifications, therefore all records from this source on these dates were excluded from the analysis.

Stability of information over time also impacts the quality of results. If the number of records or sources differs over time, then an analysis designed around a specific point in time may not generalise well to other points in time. This was important for our work as validation for the whole population was only possible on a single date: the date of Census 2013.

**Figure 1: Notifications by source over time**

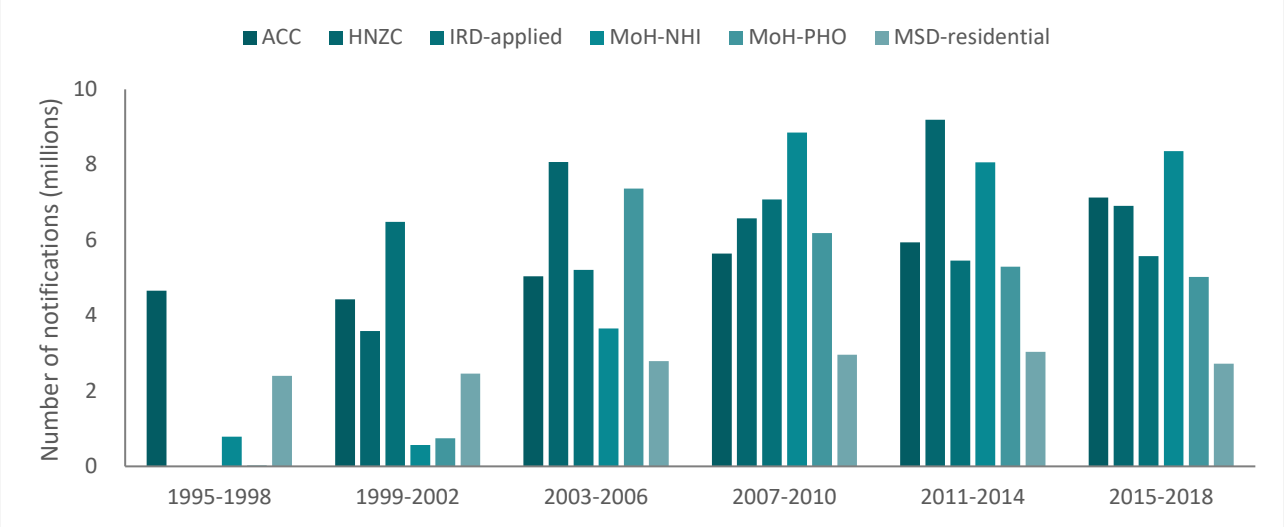



Figure 1 shows the number of notifications in four-year bands for the six largest administrative sources. This shows that, while there is some variation, data after 2007 is more stable than earlier data. This is consistent with findings from Stats NZ (2017 & 2019). So, in line with this work, we focus our analysis on the period where notifications are most stable over time.

# More notifications do not guarantee higher accuracy

For a single organisation, it is reasonable to assume that regularly asking customers to update their address improves the accuracy of address information. When working across different sources, obtaining more address notifications could be thought of as equivalent to asking customers to update their address more often. However, this reasoning ignores two features of linked administrative data:

1. Consistency between sources is not guaranteed. So, the same address recorded by two different sources may appear differently in linked data, and people may provide a residential address to one organisation and a postal address to another if type of address is not specified.
2. Address notifications in administrative data often occur as a reaction to changes of address. So, people who move more often will have more notifications, but are also quicker to move away from addresses they have given notice of.

Table 3 gives an overview of each data source including accuracy and number of notifications. Accuracy is defined as the proportion of individuals for whom the last notified address ID (from any administrative source) at the time of Census 2013 matches the Census 2013 address ID. Defined this way, accuracy for each source is partly dependent on notifications from other sources – a non-matching notification from one source that is replaced by a more recent notification from a second source will improve accuracy for the first source when measured this way.

**Table 3: Accuracy and age-range by administrative source**

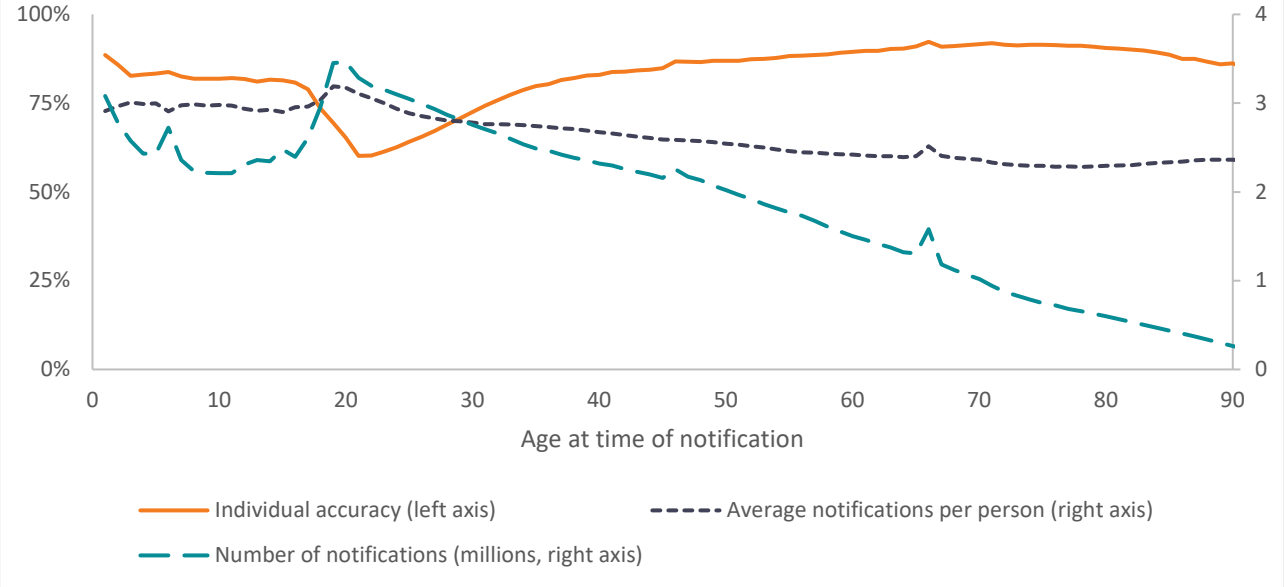
| Source | Age range | Number of notifications | Average notifications per person | Accuracy |
|---|---|---|---|---|
| **ACC** | Birth to 100+ years | 34,884,723 | 6.2 | 82.4% |
| **HNZC** | Birth to 100+ years | 35,279,364 | 57.1 | 90.5% |
| **IRD-applications** | Birth to 100+ years | 32,832,552 | 3.5 | 78.1% |
| **IRD-timestamps** | 16 to 75 years | 2,893,953 | 1.3 | 58.1% |
| **MoE** | Birth to 60 years | 1,344,450 | 1.0 | 81.2% |
| **MoH-NHI** | Birth to 100+ years | 30,941,499 | 4.1 | 84.7% |
| **MoH-PHO** | Birth to 100+ years | 24,988,050 | 4.2 | 83.5% |
| **MSD-residence** | Birth to 100+ years | 18,476,316 | 3.7 | 82.1% |
| **MSD-postal** | 16 to 100+ years | 1,968,099 | 1.6 | 55.1% |
| **MSD-partner** | 17 to 70 years | 490,749 | 2.1 | 79.8% |
| **MSD-child** | Birth to 18 years | 3,789,891 | 4.3 | 73.0% |
| **NZTA** | 15 to 100+ years | 8,741,064 | 1.8 | |

Accuracy of NZTA data is omitted from the table above because, at the time of writing, only the latest registration for each vehicle is loaded into the IDI each refresh. This means that address history can not be reconstructed from vehicle registration notifications, so it is not a fair comparison to access the accuracy of NZTA data against Census 2013. When comparing the accuracy of notifications against addresses from recent household surveys (such as HLFS 2018), NZTA notifications are one of the most accurate sources.

Gath and Bycroft (2018) demonstrated that accuracy of address notifications varies with the age of an individual at the time of the notification. This is to be expected as the frequency with which individuals move or update their address varies across different life-stages.

Figure 2 gives the total number of notifications and their accuracy against age. Accuracy is calculated at the individual-level by comparing the last notified address ID (from any administrative source) against Census 2013. The figure provides a clear picture of how the number of notifications increases and the accuracy of notifications decreases between the ages of 18 and 35 – the life-stage where New Zealanders tend to be more mobile.

**Figure 2: Number of notifications and accuracy against age**




We repeated the analysis shown in the above figure for each of the different data sources. While each data source has its own pattern in the number of notifications (for example, MoE has a spike in notifications about age 5 for primary school enrolment), and sources vary in their average level of accuracy, the dip in accuracy between ages 18 and 35 can be observed in every data source.

# Our analysis uses a rule-based approach

Given a range of address notifications from linked administrative data some method is required for combining these into a consistent set of address notifications. The approach we used for our analysis was to impose a series of rules to separate notifications into trusted information and misinformation.

Sometimes known as an “expert system”, this approach requires a researcher or expert to identify patterns of concern in the data and rules that correct the pattern. This approach has the advantage that the rules are easy to describe and that new rules can be added as new patterns of concern are identified. However, its main disadvantage is that it is dependent on experts’ ability to identify and correct patterns of misinformation.

For many of the cleaning rules we needed to identify not just the pattern but also the timescale that the pattern happens across. These cleaning rules were parameterised during development and the parameters were tuned afterwards. Tuning is the process of adjusting parameters to ensure the best performance – we tuned to maximise the number of people with consistent administrative and validation addresses.

A total of seven cleaning rules were applied: four rules at the individual-level – improving consistency for each person separately – and three rules at the household-level – improving consistency between people. These rules are described in the next two sections in the order they were applied. For both the individual-level and household-level rules an overview is first provided before each rule is discussed and illustrated. Readers requiring the precise technical definitions are advised to see the project code published on GitHub.[6]

## Consistency in address information was improved for individuals

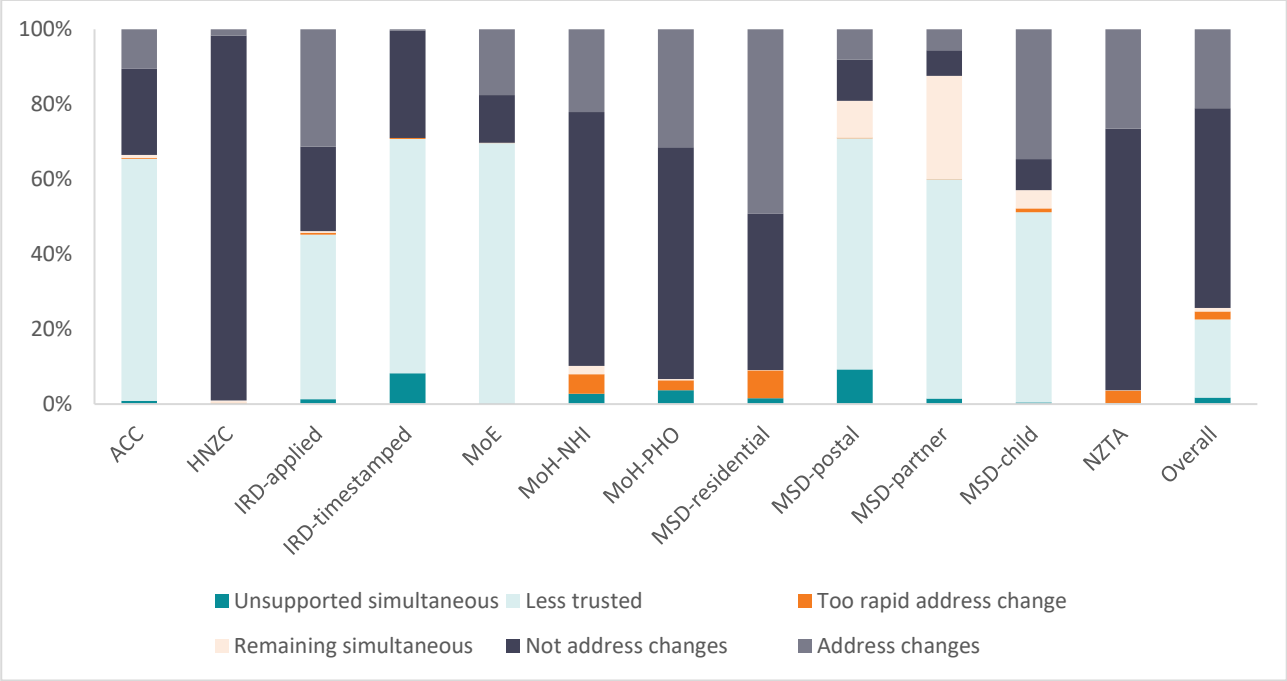
Producing accurate household information, with consistency between different household members, first requires accurate address information for individuals. Inspection of the cross-source-collection of address notifications revealed a range of concerning patterns at an individual level. An overview of these patterns and the rules that were designed to correct for them follows, with further details for each rule in the corresponding numbered subsection further below.

1. Simultaneous notifications – multiple notifications on the same day, but for different address IDs – are likely to represent misinformation: for example, mistyped addresses or a lack of distinction between residential and postal addresses. Simultaneous notifications without other notifications to support them were removed.
2. Only some sources distinguish between residential and non-residential addresses at collection. Those sources that do not make this distinction could not be used with confidence. Notifications from these non-distinguishing sources – unless supported by notifications from a distinguishing source – were removed.
3. Rapid movements into and out of the same address are likely to represent misinformation caused by different sources recording an address change at different times. Unsupported notifications whose removal eliminates a rapid change of addresses were removed.
4. To ensure no person has more than one notification on any day, where there are simultaneous notifications that persist following the resolution of the three previous rules, one address ID is chosen arbitrarily.

Following these rules, and before improving consistency between individuals, we discarded all notifications that did not imply a change of address (so given two consecutive notifications, if both have the same address ID, then the second notification is discarded). Figure 3 gives the proportion of records from each source that were removed during each stage of the analysis. Note that the two categories “Not address changes” and “Address changes” are all the trusted notifications.

[6] https://github.com/nz-social-wellbeing-agency/enhanced_IDI_address_and_household

**Figure 3: Notifications removed during individual-level cleaning by source**




It is clear from the figure that several sources were less trusted than others, given the proportion of notifications that were cleaned away. We tested the complete removal of several of these sources but found that our results worsened without them. Further details on each rule are given below.

## (1) Unsupported simultaneous notifications were removed

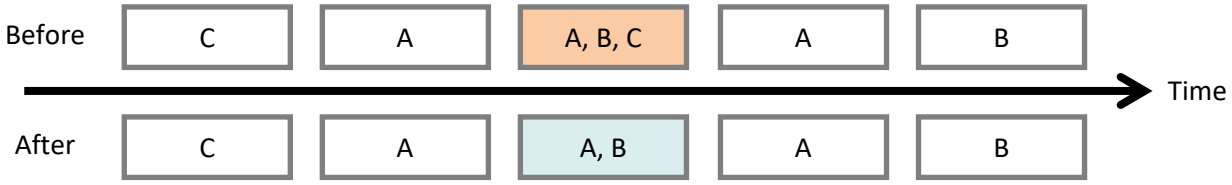
People can update their address with different organisations on the same date or with the same organisation multiple times on the same date. If the resulting address notifications were time-stamped, then it is straightforward to keep only the last notification on each day. However, as notifications were only date-stamped (and lack the time of day) some other method is required to resolve these simultaneous notifications.

Within the IDI, a significant proportion of the population had at least one date upon which they had multiple notifications (at least two, but as many as five were common) for different address IDs. Consultation of the data documentation did not eliminate the possibility that these simultaneous notifications could include mistypes or past addresses that were submitted to the database and then corrected and resubmitted.

In response, we imposed the rule that any simultaneous notification where there are no notifications for the same person at the same address ID within the following 360 days are removed. A parameter was used to define the duration of this period. Parameter tuning suggested 360 days as the value that gave the highest accuracy.

Figure 4 provides a graphical representation of a hypothetical person's address notifications before and after this rule is applied. Each box represents a notification date, letters representing address IDs, orange highlighting a concern, and blue its resolution.

**Figure 4: Visualisation of rule to resolve simultaneous notifications**




Note that this rule can result in all the simultaneous notifications on the same date being discarded if the individual never has any future notifications at any of the address IDs. In the same way, all the simultaneous notifications on the same date can be kept if the individual has future notifications at all the address IDs.

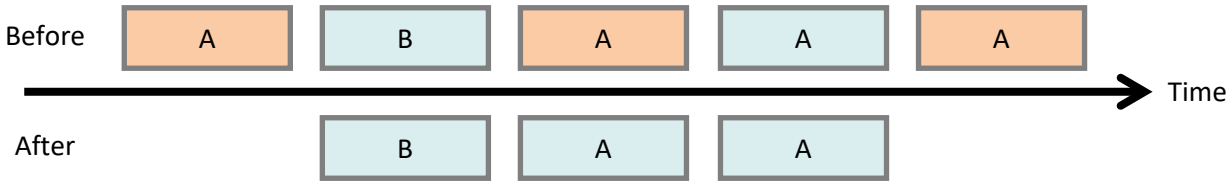
## (2) Non-residential notifications were removed

Some sources of address information do not distinguish between residential and non-residential addresses when collecting information. Such sources are less trusted to provide accurate information about individuals' residential addresses. While we could discard all information from these less trusted sources, in some cases the less trusted sources may record an address change before a more trusted source.

Within the IDI we identified HNZC, MoH, MSD residential, and NZTA as more trusted sources that distinguish between residential and non-residential addresses (this is consistent with Stats NZ 2019). All other sources were considered less trusted. For each person, we identified cases where notifications from less trusted sources gave the same address as a notification from a more trusted source that occurred at most 150 days after it. These cases were considered evidence that the less trusted source recorded an address earlier than a more trusted source. These earlier notifications were kept, but all other notifications from less trusted sources were discarded.

Figure 5 provides a graphical representation of a hypothetical person's address notifications before and after this rule is applied. Each box represents a notification date, letters represent address IDs, and orange and blue highlighting denoting notifications from less and more trusted sources respectively.

**Figure 5: Visualisation of rule to resolve notifications from less trusted sources**




A parameter was used to define how close notifications from less trusted sources needed to be to notifications from more trusted sources. Parameter tuning suggested 150 days as the value that gave the highest accuracy.

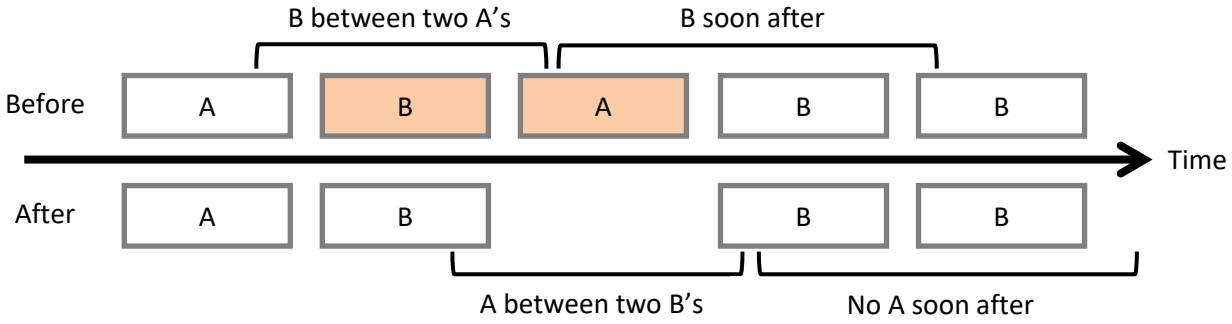
## (3) Notifications for rapid changes of address were removed

Different sources update their address information at different times. This can result in some sources holding information for past addresses unaware that other sources have updated addresses. Where past address information is reported in notifications, this produces patterns of notifications that suggest unusually rapid changes in address.

While some individuals may move with high frequency, certain patterns are more likely to be due to old address information than genuine movements. The simplest variant is where notifications suggest an individual moved out of, back into, and then out of again the same address within a short space of time. We identified such situations as two notifications recorded with the same address ID close to one another with a notification for a different address ID between them. Unless there is another notification for the different address ID soon after (suggesting it is not an old address) then the notification with the different address ID is discarded.

Figure 6 provides a graphical representation of a hypothetical person's address notifications before and after this rule is applied. Each box represents a notification date, letters represent address IDs, and orange highlighting notifications that might need removal. Note that a naïve interpretation of the before notifications suggests three address changes (A → B, B → A, A → B), while the after notifications suggest only one address change (A → B).

**Figure 6: Visualisation of rule to resolve too rapid address changes**




For this rule, two parameters were used to define (1) how close together the notifications of the same address should be, and (2) how soon after a supporting notification is required. These were initially set at 140 and 130 days respectively, but parameter tuning suggested 360 and 210 days[7] provided greater accuracy.

## (4) Remaining simultaneous notifications are resolved

Our methodology requires that each person has only one residential address at each point in time. Therefore, where there are simultaneous notifications (one date with notifications for more than one address ID) a person can only be resident at one of them. In the absence of other information, and consistent with Gath and Bycroft (2018), one address ID was chosen arbitrarily (the one with the maximum ID number).

**Figure 7: Visualisation of rule to resolve any remaining simultaneous notifications**


[7] These long lengths of time may be due to the persistence of past addresses. If some sources provide address notifications that are more than a year out-of-date, then a longer interval is needed here to capture and remove these notifications.

Figure 7 provides a graphical representation of a hypothetical person's address notifications before and after this rule is applied. Each box represents a notification date, letters represent address IDs, orange highlighting a concern, and blue its resolution.

# Consistency in address information was improved for households

Producing accurate household information requires analysis to consider interactions between individuals who share the same address. Comparisons between individuals who had at least one address ID in common revealed a range of concerning patterns, most of which are due to the difference between address notifications and address changes. An overview of these patterns and the rules that were designed to correct for them follows, with further details for each rule in the corresponding numbered subsection further below.

5. Members of a household who make the same move are likely to have moved at the same time, even though they have notifications on different dates. Where several people make the same move together, we give every person the earliest address notification date.
6. Children often interact with organisations less frequently than their parents, resulting in cases where both parents have notifications for an address change, but the child does not. Where a child is living with their parents and both parents change address without the child, we add notifications for the child to change to the same address.
7. At occupancy transitions, notifications for the new household moving in can occur before notifications for the old household moving out. This produces a short period of overlap where two households appear to be sharing the same dwelling. Where this occurs, we adjust the notification date for the household moving out to be before the household moving in.

Note that rules 5 and 7 update the timing of records so they do not change the total number of address notifications. In contrast, rule 6 adds new notifications. [8] Further details on each rule are given below.

## (5) Members of the same household have the same move dates

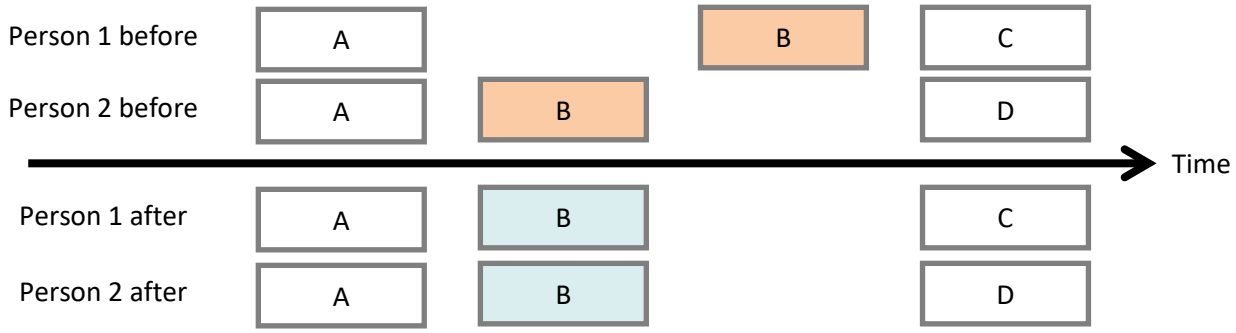
Just as address notifications are not synchronised between organisations (rule 3), notifications are also not synchronised between people within a household. This means that household members who change address together will have address notifications at different points in time. Therefore, when we observe groups of individuals who make the same move with notifications at a similar time, we can infer a common date for the address change.

Within the IDI, where two people have notifications that suggest they moved from and to the same address within 90 days, the notification for the second mover is replaced so that both people have the earlier move date. Figure 8 provides a graphical representation of this rule, with each box

[8] In our construction, rules 5 and 7 result in changes to 11.0 and 1.0 million records respectively; rule 6 adds 30,900 records. Combined these rules affect 32% of all 37.3 million notifications in our final address table.

representing a notification date, letters representing address IDs, orange highlighting a concern and blue its resolution.

**Figure 8: Visualisation of rule so household members have same address changes**




Two variants of this rule were also used: one for people who moved from different address IDs to the same address ID, and one for people who moved from the same address ID to different address IDs. The 90-day interval was reduced to 70 and 60 days respectively for these two variants. These three interval sizes were defined as parameters, and parameter tuning suggested these values as the ones that give the highest accuracy. A further variant of this rule that handled single-dwelling and multi-dwelling addresses differently was also considered, but not used.

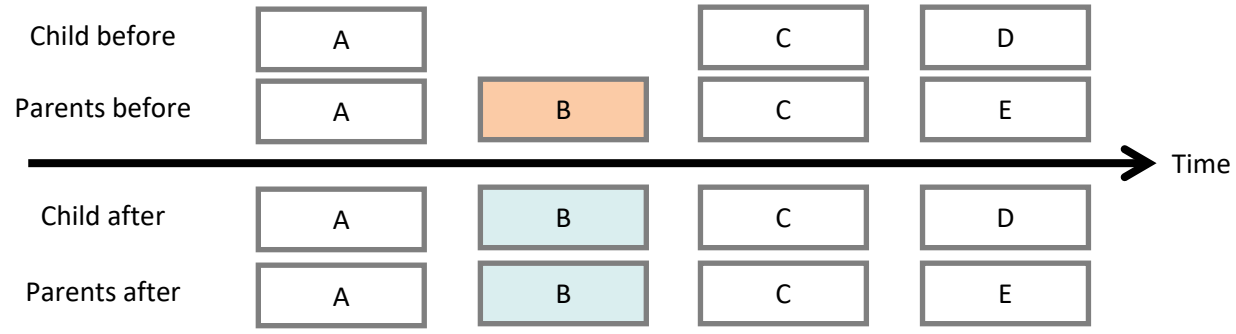
## (6) Children move when both parents live together

Due to different frequencies with which children and adults interact with government organisations, we may observe cases where a lack of notifications for the child means that a change of address is not observed for them when their parents move. As parents are unlikely to leave a child behind when they move, we can infer that the child moves as well.

We impose the rule that where a child has the same address ID as both biological parents, the parents' notifications show a change of address ID, and the child's notifications show a change to a different ID (but one shared with the parents in the future), then an additional notification is generated so that the child changes address with the parents. This rule is consistent with Stats NZ (2019) that suggests where both of a child's parents live together but the child has a different address ID then assigning the child the same address ID as the parents improves overall accuracy.

Figure 9 provides a graphical representation of this rule, with each box representing a notification date, letters representing address IDs, orange highlighting a concern, and blue its resolution.

**Figure 9: Visualisation of rule for children to move with parents**



Note that this rule uses birth records to identify biological parents. The presence of both biological parents was chosen as a more conservative version of this rule. Other forms of legal guardianship were not considered. Children are defined to be 15 years of age or younger.[9]

## (7) Households move out before new households moves in

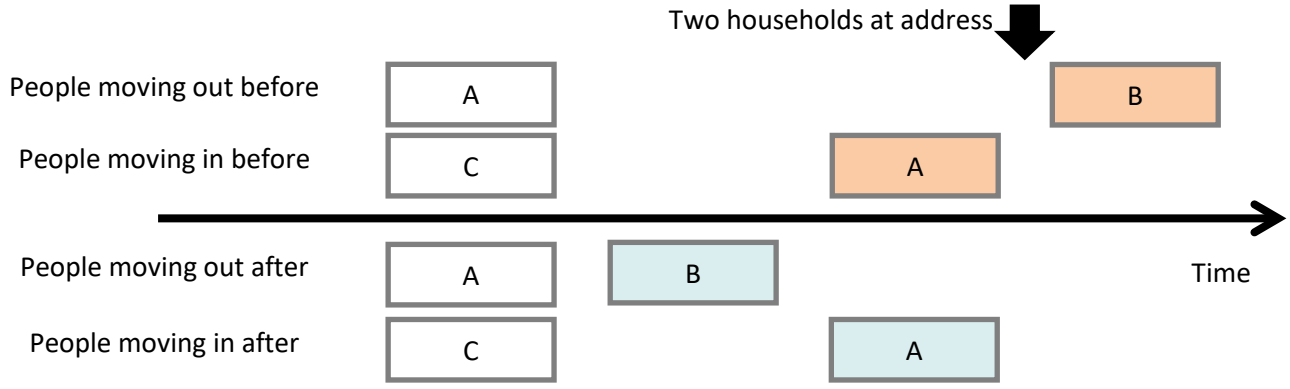
An occupancy transition occurs when all the occupants of a dwelling move out and a new group of occupants move in. Due to the difference between address notifications and address changes, there will be transitions where the notifications for the group moving in occur before the notifications for the group moving out. Hence there is an interval during which the dwelling appears to be occupied by two households.

To correct for this, we identify cases where people are moving out of an address close to the time that others are moving in. If the moving in occurs before the moving out, then the notification date for the people moving out is replaced with a notification date that is one day before the move in. As the initial version of this rule resulted in some unusual behaviour where people spent only a short duration at an address, we restrict the rule to require the people moving out and in all spend a minimum of 120 days at the address.

A parameter was used to define how close together the move in and move out dates needed to be. Parameter tuning suggested 35 days as the value that gave the highest accuracy.

Figure 10 provides a graphical representation of this rule, with each box representing a notification date, letters representing address ID, orange highlighting a concern and blue its resolution. The transition address ID is A, with one group moving out of address A and the other moving into address A.

**Figure 10: Visualisation of rule for household move out to precede new move in**




Note that given the complexity of identifying entire groups moving out and in together, this rule does not require all the people at an address to move together. If only one person is moving out and only one person is moving in the rule can still be applied. Consequently, it depends on the previous two rules to ensure consistency between household members.

[9] Several other ages were considered, both younger and older. Fifteen was chosen for its accuracy and its consistency with legislation (in New Zealand from the age of 16 a person can choose to leave home without their parents' consent).

# Validation of our constructed table produces mixed results

Combining linked administrative data to create a table of address information from which household composition can be accurately identified requires significant analysis. Assessing the accuracy of any resulting table is an important stage of the analysis, so we can determine whether any improvement has been made and, if so, how much.

## The constructed table has equivalent household accuracy to the existing table

To determine the accuracy of household membership, we follow the classification approach described by Gath and Bycroft (2018).[10] The key points in this process are as follows:

1. We select our validation population as the people in the estimated residential population who have an address ID in both the administrative and validation sources. This ensures that identities with incomplete linking and temporary visitors are excluded.
2. For each address ID in the validation dataset, we determine the proportion of household members who have the same address ID in both the validation and administrative data.
3. This proportion is used to classify the quality of the match for each household: Perfect, Partial and Poor where 100%, at least 50%, and less than 50% of household members have the same address ID respectively.
4. The previous two steps are repeated for each address ID in the administrative dataset. This gives two classifications for each household into Perfect, Partial and Poor. The first starting from validation data and the second starting from administrative data.
5. The two classifications for the match are combined. The final classification of the match for each household is the worst of the two assigned in the previous steps.

The result of this process is that where there are no differences in membership between the validation and administration data, the match is classified as Perfect; where there are small differences, the match is classified as Partial; and where there are significant differences, the match is classified as Poor. A worked example of the classification process can be found in the appendix.

Note that this is a point-in-time validation method. Consistency of household membership and transitions in membership over time requires a different collection of validation data and a different methodology than we have used here.

Table 4 gives the overall household level accuracy from our analysis – both the accuracy of the initial table and the output table after cleaning rules were applied are given. For comparison, we

[10] An exact duplication of Gath and Bycroft's validation method was not possible given the absence of published code and their use of some datasets we did not have access to. We replicated their method as best we could from the description in their paper and conversations with both authors.

also provide the accuracy measures for the existing IDI address table using the June 2017 and September 2019 IDI refreshes, and the theoretical best case for our dataset.

**Table 4: Accuracy of household membership**

| Address table | Number of addresses | | | Percent of Addresses | | |
|---|---|---|---|---|---|---|
| | Perfect | Partial | Poor | Perfect | Partial | Poor |
| **June 2017 refresh** | 591,300 | 477,822 | 173,142 | 47.6% | 38.5% | 13.9% |
| **Sept. 2019 refresh** | 839,559 | 387,870 | 106,158 | 63.0% | 29.1% | 8.0% |
| **This paper – input data** | 801,570 | 426,222 | 109,902 | 59.9% | 31.9% | 8.2% |
| **This paper – output after rules** | 829,620 | 395,646 | 103,710 | 62.4% | 29.8% | 7.8% |
| **Theoretical best case** | 1,007,394 | 246,342 | 46,752 | 77.5% | 18.9% | 3.6% |

Notes:

- The June 2017 refresh results are those published by Gath and Bycroft (2018).
- The September 2019 refresh table is based on the original work by Gath and Bycroft but includes some adjustments (Stats NZ 2019).
- The results for this paper use data from the September 2019 refresh but our own process as described above.
- The theoretical best case assumes all individuals with an administrative notification that matches their validation address, before the validation date, are assigned to the correct address regardless of any non-matching notifications they may also have.
- Within the IDI, the September 2019 refresh table is most commonly referred to as the [address_notification] table, and the input data for this paper is most similar to the [address_notifications_full] table.

Several comparisons can be drawn from these results: First, the difference between the June 2017 refresh results (47.6% of households perfect) and the input data to this paper (59.9% perfect) reflects significant improvements by Stats NZ in the quality of address information and identity linking. Second, the difference between the input data (59.9% perfect) and output from this paper (62.4% perfect) show that our cleaning rules do improve the quality of address information.[11] Third, the lack of difference between the September 2019 refresh (63.0% perfect) and the output from this paper (62.4% perfect) show that our cleaning rules are not able to improve the quality of address information beyond that already obtained. Finally, the difference between all results and the theoretical best case (77.5% perfect) shows that significant improvements in address quality are still possible.

# Improvements in individual accuracy vary with age

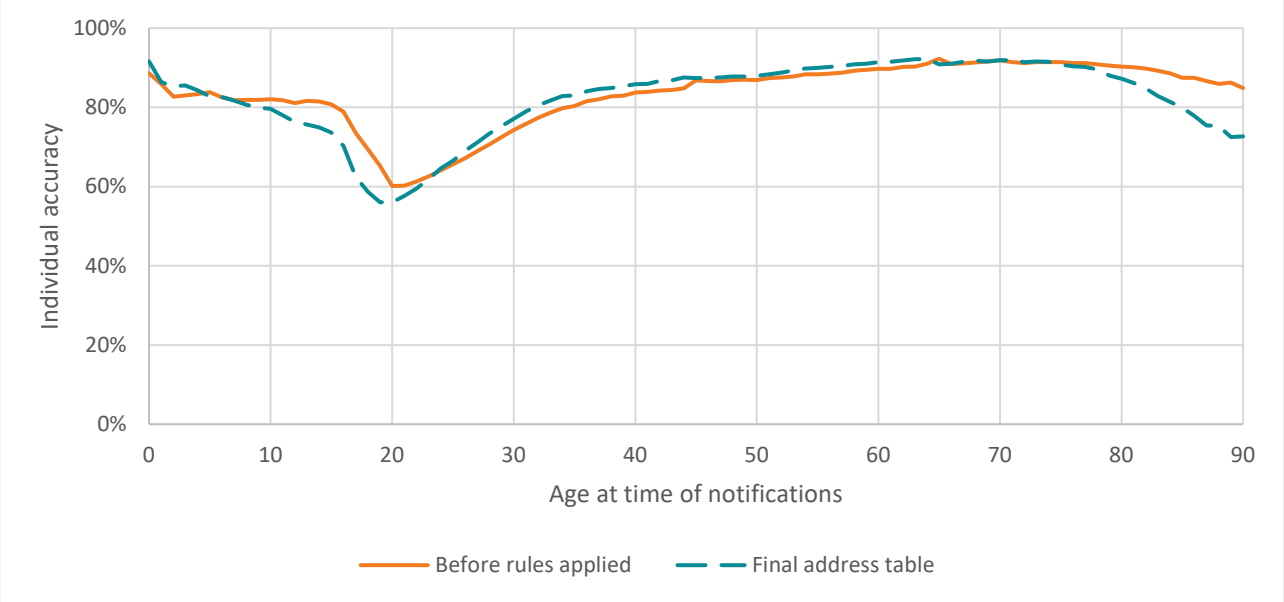
Though our motivation has been improving the accuracy of household membership, a corresponding improvement in accuracy of individuals' address records should also be expected. However, our analysis produced no meaningful change in individual level accuracy.

[11] Given that the overall improvement in accuracy is small, we omit the more detailed results showing the improvement made by each rule.

While disappointing, it is important to note that further improvements are difficult: For 11% of individuals their address ID from Census 2013 never appears in any of their administrative records before the census. This means that individual-level accuracy of more than 89% is impossible. Furthermore, for 8% of individuals their address ID from Census 2013 is reported by only one administrative source before census. This means that improving individual-level accuracy beyond 81% is very difficult.

Figure 11 gives accuracy against age using all administrative data sources before any of our rules are applied (identical to Figure 2), and our final address information. It shows that while our analysis improves average individual accuracy it does not make universal improvements: accuracy for some age groups worsens under our analysis.

**Figure 11: Comparison of accuracy against age (Census 2013)**




Note that the two age groups for whom individual accuracy worsens are those with the greatest and least number of notifications per year as observed in Figure 2. A likely cause is that because our rules have been tuned using the entire population, they assume an average number of notifications per year. Therefore, they perform poorly when the number of notifications differs from the average.

# Validation against survey sources appears better

All the validation so far has been done in comparison to Census 2013. In Table 1 we listed the three main household surveys by Stats NZ (in addition to the Census) as other sources of validation data. Validation was done against these sources as well. However, the results are mixed, with significant variation in accuracy over time. Table 5 provides an overview of the validation measures for each validation source.

**Table 5: Validation results against different sources of truth**

| Validation source | Years & Average respondents per year | Percent of household matches classified as perfect | | Percent of individuals with an address match | |
|---|---|---|---|---|---|
| **Census** | 2013<br>3,663,000 respondents | 62.4% | | 82.3% | |
| **GSS** | 2008-2016<br>(biannual – 5 waves)<br>20,000 respondents | Max:<br>Mean:<br>Min: | 81.9%<br>80.9%<br>80.2% | Max:<br>Mean:<br>Min: | 85.3%<br>84.3%<br>81.5% |
| **HES** | 2007/08-2016/17<br>(annual – 10 waves)<br>8,500 respondents | Max:<br>Mean:<br>Min: | 84.0%<br>81.8%<br>79.2% | Max:<br>Mean:<br>Min: | 86.8%<br>85.3%<br>82.3% |
| **HLFS** | 2007-2018<br>(annual – 12 waves)<br>145,000 respondents | Max:<br>Mean:<br>Min: | 78.3%<br>70.5%<br>68.0% | Max:<br>Mean:<br>Min: | 86.2%<br>83.2%<br>79.6% |

The strongest conclusion we can draw from Table 5 is that the accuracy of our constructed table varies over time. A likely cause of this is the parameters were tuned using Census 2013, and hence have not been selected for consistent accuracy across time. While later years tend to have higher accuracy than earlier years, we are reluctant to draw conclusions from this as there is more address information for later years to work from.

The difference in household-level accuracy between the different validation sources can be attributed to differences in sample size between census and the surveys. The validation sources with the smallest sample size have the highest household-level accuracy. This is because the validation algorithm considers only those individuals with an address in both the administrative and validation sources, so validation sources that cover a larger population have more scope for error.

# Our work is open to support future research

Creating a single consistent set of information from linked administrative data will continue to be both challenging and necessary for research. By building on the progress of existing work further improvements can be made.

As part of this, alongside our technical documents, the Social Wellbeing Agency makes a practice of releasing code on our GitHub page once projects are complete.[12] The repository for this work includes the data preparation, application of cleaning rules, and the validation process.[13] Tuning of

[12] https://github.com/nz-social-wellbeing-agency

[13] https://github.com/nz-social-wellbeing-agency/enhanced_IDI_address_and_household

parameters to produce the highest accuracy result was not automated and hence is not available in the repository. Please contact us by email (info@swa.govt.nz) if you want to know more.

## The construction could be further extended

During our research we considered a range of other rules, and variants of rules, that we could not incorporate into the final table. The following provides a short list of ideas should future research extend the existing approach.

- **Make rules responsive to age or frequency of notifications.** As Figure 11 illustrates, the rules used in our analysis improve the average accuracy but worsen accuracy for specific age groups. Hence adapting the rules so they account for differences in age or the local frequency of notifications would correct for this.
- **The addition of non-residency indicators.** When people die, they can no longer be resident at an address, have an address notification or change. Hence death records could be used as an indicator of non-residency. In a similar way, while a person is overseas or in prison they are not resident in a New Zealand address – though if this period lasts for only a short length of time then they may still be considered part of the household.
- **Supplementing address information with region information.** Where a source gathers region, but not address, we could use this information to focus on within-region-addresses. Consider for example a student attending classes in-person (not remotely or by correspondence) who provides an out-of-region postal address. The region of the education provider would let us identify that such an address notification is unlikely to be residential.
- **Focus on address changes within each source.** Our analysis infers a change of address where two consecutive notifications report different address IDs regardless of the source of those notifications. However, a change of address recorded by a single source is more likely to represent an address change because concerns about linking and consistency between sources do not apply. Hence, we might wish to treat within-source changes differently from across-source changes.
- **Adjusting for density of notifications.** Sources differ in the density of their notifications (how many notifications per person per unit of time). A high-density source can produce a lot of incorrect notifications in a short space of time. Hence, we may wish to treat notifications from such a source differently from lower density sources.
- **Adapt for children living without both biological parents.** Rule 6 requires that a child lives with both biological parents. This is a highly restrictive condition that does not allow for single parenting, mixed families, and non-parent guardianship. This rule could be generalised by considering ‘all adults a child lives with’ instead of ‘both biological parents’.
- **Allow for concurrent addresses.** Our analysis has assumed that people have a single residential address, but some people may live across more than one address. For example, consider a child who moves every few days between parents that have separated. Unless analysis allows for multiple concurrent addresses, accuracy for people with two residential addresses will be approximately half the accuracy of people with one residential address.

Note that regardless of the rules added, successive improvements in accuracy will be more difficult. This is because the people who still lack accurate address information have, in general,

less consistent address information to begin with compared to the people for whom we already have accurate information.

## Other approaches merit consideration

Developing a set of rules is one approach to creating a consistent collection of address and household information. A complete solution may also require improvements in how data is collected, prepared, linked, as well as analysed. The following provides a list of several such improvements.

- **Collecting address change date.** Organisations often collect current address without any indicator of when an address became current. Collecting new information on how long ago the address change took place would help reduce the difference between address notifications and address changes.
- **Standardisation of addresses.** Where address information is provided in different formats, establishing a consistent format before use is essential. Without standardisation of address information, we may incorrectly infer a change of address where there has been a change of format. Work here is already underway, with Stats NZ is leading the development of Data Content Standards across government that include addresses.[14]
- **Using address to improve linking of identities.** Any effort to establish consistent address information will be ineffective without correct linking of identities between sources. Some of the loss of accuracy in our constructed table will be due to linkage error (incorrect linking of identities between sources). When people have the same address recorded with two different organisations (even if it is not a current address) this can provide evidence to strengthen our confidence that both sources are referring to the same person.

Regardless of the future approach taken, we recommend that subsequent research also examine the raw address strings, and the process by which raw address strings are converted to address ID values. As a starting point we suggest examining those individuals who indicated in their response to Census 2013 they had lived at the same address for five years of more. Using only address IDs, 18% have administrative address notifications in the five years preceding Census 2013 with different addresses than those reported in Census 2013. Given the stability of this subpopulation, and the substantial proportion with non-matching address IDs, it will likely provide a fruitful starting point to investigate consistency between sources of address information.

[14] https://www.data.govt.nz/manage-data/data-content-standards/

# References


Culhane, Dennis. 2016. The potential of linked administrative data for advancing homelessness research and policy. *European Journal of Homelessness*, 10(3).

Gath, Megan and Christine Bycroft. 2018. *The potential for linked administrative data to provide household and family information*. Statistics New Zealand.

Hand, David. 2018. Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3), pp. 555-605.

Mowle, James and Ross Watmuff. 2018. The changing role of the Census in Australia's integrated data landscape. Paper prepared for the 16th Conference of IAOS. OECD Headquarters, Paris.

Sen, Pronab. 2008. "Challenges of Using Administrative Data for Statistical Purposes – India Country Paper." National Statistical Commission, India.

Statistics Centre – Abu Dhabi. 2015. *Manual of Statistical Quality Standards and Procedures for Administrative Records*. Abu Dhabi Emirate, United Arab Emirates.

Statistics New Zealand. 1999a. *Statistical Standard for Dwelling Occupancy Status.*

Statistics New Zealand. 1999b. *Statistical Standard for Occupied Dwelling Type.*

Statistics New Zealand. 2014. *Coverage in the 2013 Census based on the New Zealand 2013 Post-enumeration Survey.*

Statistics New Zealand. 2015a. *Metadata – Geospatial information in the IDI*. Unpublished guidance.

Statistics New Zealand. 2015b. *IDI Data Dictionary: Household Labour Force Survey (July 2015 edition)*.

Stats NZ. 2017. *Experimental population estimates from linked administrative data: 2017 release*.

Stats NZ. 2018. *Experimental ethnic population estimates from linked administrative data.*

Stats NZ. 2019. *Overview of statistical methods for adding admin records to the 2018 Census dataset*.

Tønder, Johan-Kristian. 2008. The Register-based Statistical System – Preconditions and Processes. *In Shanghai: International Association for Official Statistics Conference Shanghai.* pp. 14-18.

Trépanier, Julie, Jean Pignal, and Don Royce. 2014. Administrative Data Initiatives at Statistics Canada. *In proceedings for the Federal Committee on Statistical Methodology Research Conference. Washington, D.C.*

Valente, Paolo. 2010. Census taking in Europe: How are populations counted in 2010? *Population & Societies*, (467), 1.

# Appendix

We provide the following example to illustrate how the quality of the match between administrative and validation data is classified for each household into Perfect, Partial and Poor.

Table 6 gives an example of five individuals (ID numbers 1-5) who are recorded at two different address IDs (labelled A and B). This is how the data appears before the classification process takes place. The classification of the match for each address ID as Perfect, Partial, or Poor is shown in Table 7. This is the output from the classification process.

**Table 6: Example address data for accuracy classification**

| Individual ID | Validation address ID | Admin address ID | Match |
|---|---|---|---|
| **1** | A | A | Yes |
| **2** | A | A | Yes |
| **3** | B | A | No |
| **4** | B | A | No |
| **5** | B | B | Yes |

**Table 7: Example accuracy classification for each address ID**

| Address ID | Classification from validation | Classification from admin | Final classification |
|---|---|---|---|
| **A** | Perfect | Partial | Partial |
| **B** | Poor | Perfect | Poor |

Explanation:

- As all the individuals at address ID A in the validation data are also at address ID A in the administrative data the classification-from-validation for address ID A is Perfect. However, as only half of the individuals at address ID A in the administrative data are also at address ID A in the validation data the classification-from-administration for address ID A is only Partial. The final classification for address ID A is therefore Partial.
- For address ID B, less than half the individuals with address ID B in the validation data have the same address ID in the administrative data so the classification-from-validation for address ID B is Poor. This means that, even though all the individuals with address ID B in the administrative data have the same address ID in the validation data, the final classification for address ID B is Poor.