Experimental population estimates from linked administrative data

2017 release

Census Transformation

Stats NZ

New Zealand Government

Disclaimer
The results in this paper are not official statistics. They were created for research purposes using the Integrated Data Infrastructure (IDI) managed by Stats NZ.

Access to the anonymised data used in this study was provided by Stats 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. The results in this paper were confidentialised to protect these groups from identification.

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

See Privacy impact assessment for the Integrated Data Infrastructure for more information.

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

ISBN 978-1-98-852833-5 (online)

Published in October 2017 by
Stats NZ Tatauranga Aotearoa
Wellington, New Zealand

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

www.stats.govt.nz

Contents

List of tables and figures

List of tables

List of figures

Introduction

Census Transformation in New Zealand

In March 2012 the New Zealand Government agreed to a Census Transformation strategy. This strategy has two strands:

- a short-to-medium term focus on modernising the current census model and making it more efficient
- a longer-term focus on exploring the feasibility of a census based on administrative data (Stats NZ, 2012, 2014a).

Investigations into the long-term direction for census are focused on understanding future census information requirements and the ability of administrative data to meet those requirements.

Census transformation – a promising future (a 2015 Stats NZ Cabinet paper) recommended that Stats NZ work actively towards a future census based primarily on Government's administrative data, supported by redevelopment of its household surveys. See Census Transformation in New Zealand for more information.

About this paper

For an administrative-based census, we must be able to identify individuals who are resident within New Zealand at a given point in time without relying on a full-enumeration census. We also need to identify where in New Zealand these individuals live.

In 2016 we released our first experimental series of national-level population estimates derived from linked administrative data in the Integrated Data Infrastructure (IDI). In this second release we extend those national estimates. We also include population estimates for subnational estimates for the first time.

This published data series includes estimates at 30 June 2007–16 for:

- New Zealand by single year of age and sex
- territorial authority and Auckland local board (TALB) areas by five-year age group and sex
- area unit total populations.

This paper describes the method used to produce these series, including improvements made since the first experimental series release. We compare the results with official population estimates, finding that the results are largely encouraging. Often there is close agreement with official figures, and we have seen results improving steadily over time. However, there are still marked differences for some age groups and local areas. We summarise the possible factors contributing to any observed differences, and will continue to explore approaches to overcome potential sources of error.

We invite your feedback on any of the methods and results covered in this report in order to support our future development. To submit feedback, please complete this form.

Note that the experimental population estimates are not official statistics. Rather, they are published as output from research into a different methodology than that currently used in the production of official estimates.

Background

Estimates of the New Zealand resident population are the most critical output based on the census. The Census Transformation project needs to answer the question: Can linked administrative sources, combined with a coverage survey and statistical model, produce estimates of the New Zealand resident population and dwellings to a standard that will meet key customer requirements?

Gibb and Shrosbree (2014) and Gibb, Bycroft, and Matheson-Dunning (2016) describe the development of a method for constructing a resident population using linked administrative data sources available in Stats NZ's Integrated Data Infrastructure (IDI). Comparisons with the official estimated resident population (ERP) suggested enough promise for further investigation.

In September 2016, an experimental series of population estimates was released for New Zealand by single year of age and sex (Stats NZ, 2016a). Overall, these results generally agreed closely with the official ERP. However, there was evidence of both undercoverage and overcoverage for some groups in the population. Potential sources of coverage error were identified, such as linkage errors in the IDI, the incorrect classification of migrants, and individuals not selected because they were not active in administrative sources.

The quality of address information in administrative data was also assessed in Gibb and Das (2015). They found that coverage was high, with most people having an address in at least one source, but that further improvements were required to increase the accuracy of the available addresses. They also showed that selecting the most recently updated address, regardless of source, was better than any of the individual sources on their own.

Aims and scope

This paper updates the methods for producing population estimates from linked administrative data in the IDI. We compare the resulting estimates, overall and by subnational area, against the official ERP. We then discuss the potential factors contributing to any observed differences.

This paper accompanies the release of a second experimental series of population estimates produced from linked administrative data – see Experimental population estimates from linked administrative data on our Innovation Site for more detail. We will soon be releasing an interactive tool for visualising the published estimates and viewing detailed comparisons.

Our aim is to update users of our progress in producing these administrative-based estimates, and understanding their quality. We also hope to receive input from users of this data.

This series has been produced solely from the linked administrative data. The estimates do not incorporate any additional statistical modelling.

Future releases

This release is the second in an ongoing series of New Zealand population estimates produced from linked administrative data.

A third release is scheduled for later in 2017, focusing on estimates by ethnic group. We will also release a report detailing possible methods for producing modelled estimates of the New Zealand population. A series of experimental income estimates will also be published, making use of the populations described in this release.

Data sources and quality standards

This section presents the data sources used for this research.

Integrated Data Infrastructure

Stats NZ developed the IDI as an environment in which to link multiple data sources in a systematic and secure way. It was developed to produce official statistics and to allow Stats NZ staff and external researchers to conduct policy evaluation and research on people's transitions and outcomes. The IDI contains de-identified administrative and survey datasets, linked at the individual level. We use the IDI as a test environment for examining the potential of linked administrative data sources to produce population estimates.

The basic structure of the IDI consists of a central 'spine' to which a series of data collections are linked. The spine forms the conceptual centre of the IDI and all other datasets are linked to it. Broadly, the target population for the spine is all individuals who have ever been residents of New Zealand. Black (2016) provides more information on the formation of the IDI spine.

Other data sources are linked to the IDI spine (see Stats NZ, 2014a for a description of the linking process). The linked datasets cover a wide range of subject areas and include: employer and employee job and earnings information based on Inland Revenue data; health information, including GP enrolment and hospital visits from the Ministry of Health; education data from Ministry of Education; benefit dynamics data from Ministry of Social Development; migration movements data from Ministry of Business, Innovation and Employment; and data from Stats NZ's Household Labour Force Survey and 2013 Census of Population and Dwellings.

The IDI continues to change as new datasets are added. See Data in the IDI for current information.

The administrative population referred to in this paper was constructed from the data sources available in the June 2017 IDI refresh.

Address information in the IDI

The IDI contains address information from a range of administrative sources, which have been geocoded within the IDI. Address strings provided by each agency are linked to NZ Post's National Postal Address Database (NPAD). Where a successful link is made, a unique encrypted address identifier is made available to researchers, along with geographic information such as meshblock, area unit, and territorial authority (TA).

Two central geographic (or address notification) tables are derived within the IDI system. Seven sources currently contribute to these tables:

- 2013 Census – address of usual residence as reported on individual census form
- Inland Revenue (IR) – address for the contact residence supplied to IR
- National Health Index (NHI) – address of residence as recorded when visiting a hospital or outpatient clinic
- Primary Health Organisation (PHO) – address of residence as recorded when visiting a general practitioner

- Ministry of Social Development (MSD) – address of residence and postal address as reported when applying for working age benefits and superannuation
- Ministry of Education (MOE) – address of residence as reported when enrolling at primary or secondary school (but not tertiary education)
- Accident Compensation Corporation (ACC) – address of residence as reported when filing an ACC claim.

Each address notification has an associated date, reflecting when that information was provided to an agency. We use this date when selecting addresses for each individual at a given point in time. Within the IR data, there are a small number of dates with extremely high counts of IR notifications. In total, there were nine days with more than 100,000 address notifications. This compares to an average of around 7,000 updates per weekday from 2010 onwards.

It is highly improbable that all of the records from these specific dates reflect true address notifications or even re-verification of addresses. They instead appear to reflect an unrelated administrative process. While some notifications from these dates will contain genuine new information, many of these updates are likely to misrepresent the actual notification date, and so we have removed them.

Several other sources of address are also available in the IDI, but not incorporated in the central address notification tables, including:

- Household Labour Force Survey (HLFS) – address of usual residence for individuals responding to the survey
- New Zealand Transport Agency (NZTA) – address of usual residence for individuals issued with a New Zealand driver licence or motor vehicle registration.

‘Gold standard’ sources of address

2013 Census

The census is considered to provide the best indication of an individual’s usual residence on census day. We therefore treat the 2013 Census usual residence address as a ‘gold standard’ for assessing the quality of the administrative addresses.

As well as the IDI geocoded addresses, the original meshblocks of usual residence from the 2013 Census are available. Unlike the information in the address notification tables, these records do not necessarily include the specific address identifier. However, a meshblock exists for all individuals, not just those who were successfully geocoded to a particular address within the IDI.

Household Labour Force Survey

One limitation of the census is that it only represents a single point in time. The HLFS is a regular quarterly survey, which makes it useful for measuring changes in quality across time. Comparisons between the March 2013 quarter HLFS collection and the 2013 Census suggested a high level of consistency in the address information.

Given the relatively small sample size of the HLFS, it is likely to have limited value as an alternative source for allocating individuals to an address. Instead, we use the HLFS as a second ‘gold standard’ indicator of usual residence, particularly for assessing the quality of administrative addresses over time.

Estimated resident population

The estimated resident population (ERP) of New Zealand is the official estimate of all people who usually live in New Zealand at a given date (see Standard for population terms). We use the ERP to make comparisons with our administrative estimates at the aggregate level.

Official estimates of the national population are published by Stats NZ each quarter. The ERP by subnational area is produced annually at 30 June, and is published by a range of geographic boundaries (including regional councils, territorial authorities, and area units), five-year age group, and sex. The current methodology for producing the official ERP series relies on a periodic full-enumeration census. The ERP at a given date is derived by updating the census usually resident population count for estimates of:

- net census undercount (as estimated by the Post-enumeration Survey)
- residents temporarily overseas on census night
- natural increase (births less deaths) between census night and the given date
- net migration (arrivals less departures) between census night and the given date (Stats NZ, 2014b).

The ERP is generally most accurate immediately after the census, and accuracy tends to decrease over time the further away from the census. The ERP is revised when results from the next census are available.

Quality standards for comparing the IDI-ERP with the ERP

We compared the IDI-ERP with the official ERP at the aggregate level. These comparisons provide an indication of net undercoverage or net overcoverage, although we cannot make any conclusions about the gross errors. Overall similarities may also conceal individual sections of undercoverage and overcoverage.

Census Transformation previously developed a set of quality standards to assess the quality of population estimates produced from administrative data (McNally & Bycroft, 2015). These quality standards were determined through consultation with core customers, and provide a measure of the minimum accuracy acceptable for users. Separate standards were produced for both a survey-based and an administrative-based census model.

These standards apply to the final estimates, after implementing statistical models to adjust for any errors in the estimates produced from administrative data alone. Therefore, there is scope for improving estimates described in this paper which are not currently meeting the standards.

The standards have since been used to measure the performance of the official population estimates over recent intercensal periods (Stats NZ, 2016b).

Table 1 summarises the quality standards used in this paper, showing the proportion of estimates that should come within the fixed level of error. For example, of the national level five-year age group estimates, 90 percent should be within 1.5 percent of the ERP, and all should be within 5 percent. The age group breakdowns by territorial authority and Auckland local board area (TALB) are more detailed than those in the original report, and have been derived specifically for this analysis.

Table 1

Quality standards for an administrative census
By geographic area

Geographic area	Population unit	Percent within level of error	Level of error (within +/- percent)
Total New Zealand	Total population	100	0.5
	By five-year age group and sex	90 100	1.5 5
Territorial authorities & Auckland local boards, population 100,000 or more	Total population	100	2.5
	By five-year age group and sex	85 100	5 12.5
Territorial authorities & Auckland local boards, population less than 100,000	Total population	85 100	2.5 5
	By five-year age group and sex	80 100	5 20
Area units, population 500 or more	Total population	80 100	5 10
Area units, population 100–499	Total population	80 100	10 25

Source: Stats NZ

We also use the weighted mean absolute percentage error (WMAPE) to assess the quality of our administrative estimates. Unlike other measures, such as mean or median average errors, the WMAPE takes into account population size in its weighting, meaning a large population will have more impact than a smaller one (Lebel & Denis, 2016). It can be derived for any given population using the formula:

$$WMAPE = \sum_i \left(\left| \frac{IDIERP_i - ERP_i}{ERP_i} * 100 \right| * \frac{ERP_i}{\sum_i ERP_i} \right),$$

where ERP_i refers to the official population estimate and $IDIERP_i$ to the administrative estimate for a given subgroup of the population *i*.

Comparing the IDI-ERP with the ERP at the national level

Identifying a resident population in the IDI

Stats NZ (2016a) describes a method used to identify individuals who are resident within New Zealand at a given point in time. Activity in administrative data sources was used to indicate an individual's presence in New Zealand, with the selected sources considered to have high coverage of the population without overly increasing the risks of linkage errors. Anyone who had died or migrated overseas prior to the reference date was removed. The resulting population was called IDI-ERP v2. This method is currently implemented as a table in the IDI.

The first experimental series release also detailed a number of potential sources of error in the IDI-ERP v2. Improvements have since been made to reduce two of these sources of error – by extending the activity period from one to two years, and by applying a new method for identifying individuals no longer residing in New Zealand. We plan to implement this improved method in a future IDI refresh.

For this updated series we use the improved method, and refer to the new population as IDI-ERP v3. Specifically, the method has the following inclusions and exclusions.

Inclusion: Retain individuals whose presence is indicated by activity.

- For ages five years and over, the spine population is restricted to those individuals who had activity in one of the following IDI datasets in the two years before the reference date:
    - ACC claims
    - Inland Revenue tax (employer monthly summary of tax paid at source, or annual tax return data; receipt of taxable benefit payments is included)
    - Ministry of Health (pharmaceutical prescriptions, GP enrolment and attendance, hospital admissions, non-admission hospital visits)
    - Ministry of Education (school enrolment, tertiary enrolment or attainment).
- For ages under five years, having a New Zealand birth registration or visa approval (excluding visitor or transit visas) before the reference date is sufficient for inclusion in the population. For this age group there is no additional requirement of activity in the previous two years.

Exclusion: Remove those who have left the population.

- Linked death records are used to identify individuals with a date of death before the reference date.
- Linked migration data are used to identify individuals who were not New Zealand residents on the reference date, either because they had already migrated overseas or because they were short-term visitors to New Zealand.
    - For estimates through to 30 June 2014, individuals are classified as non-residents using the 12/16 rule (Stats NZ, 2017).
    - For estimates at 30 June 2015–16 individuals are classified as non-residents if the total length of time spent overseas is at least 6 of the 12 months spanning the reference date (that is, the six months either side of the reference date).

Improvements from IDI-ERP v2

This section describes the improvements made for IDI-ERP v3 in more detail. In addition to these changes, improvements to the quality of linkages in the IDI are intended for future IDI refreshes, including the removal of duplicates from the IDI spine and improved linkage between the IDI spine and deaths.

Method for removing overseas residents

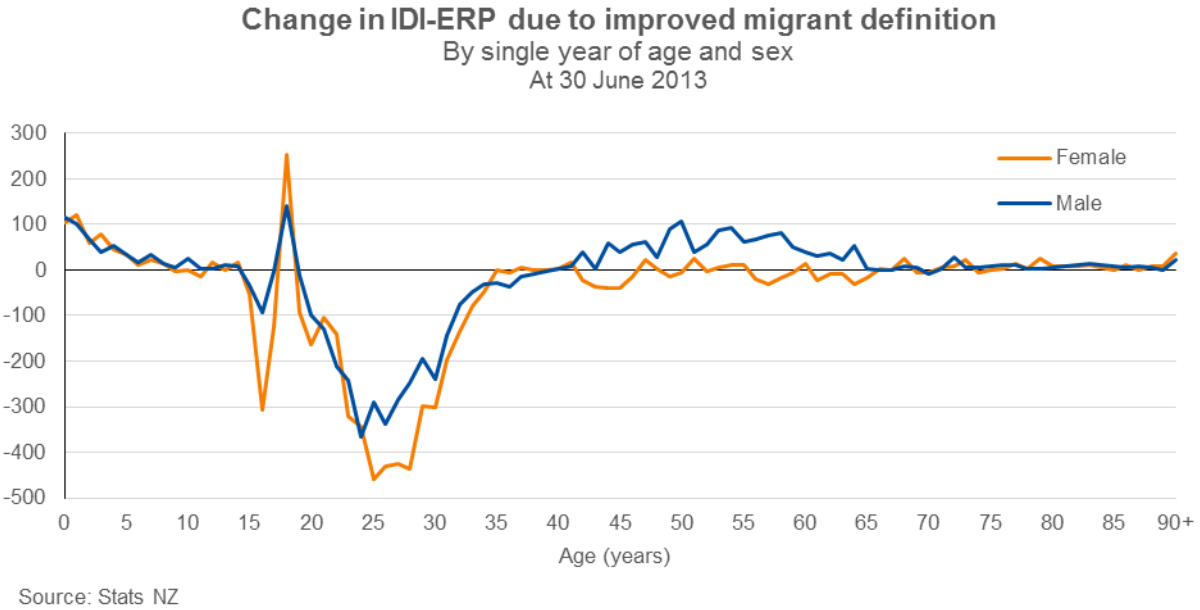
For IDI-ERP v3, we have improved the method used in v2 for classifying individuals as overseas residents. Stats NZ (2017) provides more details on the approach, referred to as the 12/16 rule. This measure uses linked travel histories to determine the resident status of each individual, based on the 16 months following a given journey. To be considered a long-term migrant, someone must spend 12 of the subsequent 16 months in New Zealand (for migrant arrivals) or outside New Zealand (for migrant departures).

Using these results, individuals were identified as being overseas residents if their most recent journey was classified as a long-term departure or a short-term visitor movement.

The 12/16 rule does require waiting 16 months after a given movement before determining an individual's resident status. For this release, results using the 12/16 rule were only available through to March 2015. Therefore, the estimates for June 2015 and 2016 make use of the 6 out of 12 rule used in IDI-ERP v2. Stats NZ is working to produce a more timely provisional measure of migration, which will also be implemented in future iterations of the IDI-ERP.

Overall, there is minimal change from V2 when applying the 12/16 rule, with only a slight decrease of 5,000 in the population for 2013 (due to more individuals being classified as overseas residents by the 12/16 rule). Figure 1 shows that most of the change occurs for ages 21–32, with a reduction in the IDI-ERP for these ages. There are two additional spikes, particularly for females, with a decrease in the IDI-ERP at age 16, and an increase in the IDI-ERP at age 18.

Figure 1



Source: Stats NZ

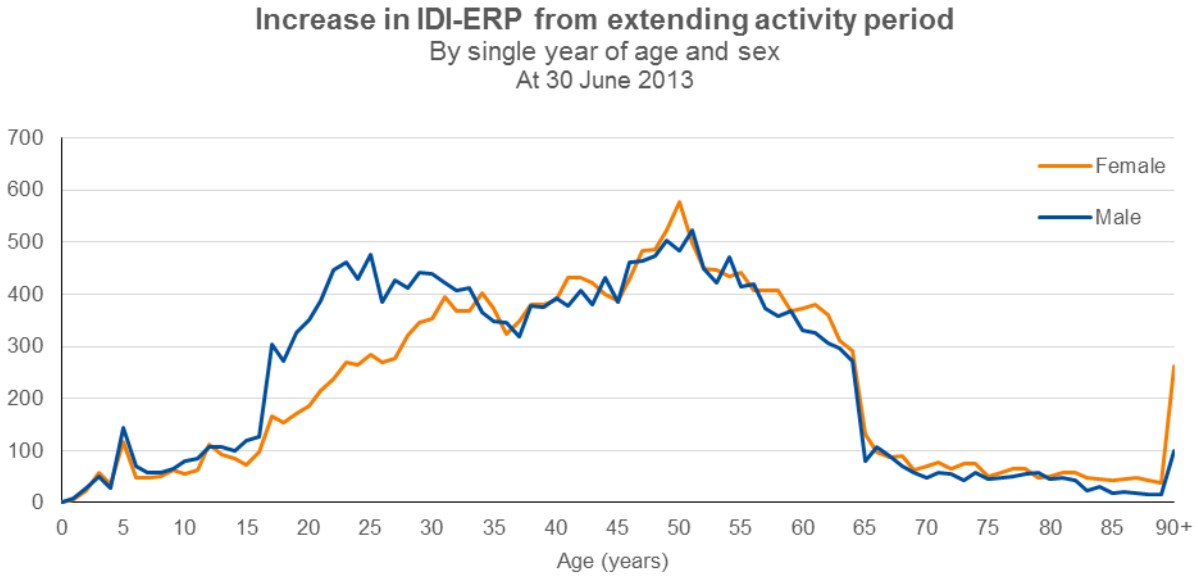
Lengthened activity period

We have also adjusted the period used to determine individuals who were active in administrative data from one to two years. This increase ensures that fewer people are missed from the population due to a lack of activity.

Overall, 42,000 additional people were included in the 2013 IDI-ERP with this adjustment. Figure 2 shows the distribution by age and sex. Very few individuals are added for ages 0–16 and 65+, when almost everyone is active in education or superannuation data, respectively. There is a relatively consistent increase across the remaining ages and by sex.

Ideally, an even longer activity period would be used to further reduce undercoverage. However, additional years were considered to have diminishing returns and to increase the likelihood of linkage errors, thereby leading to overcoverage.

Figure 2



Source: Stats NZ

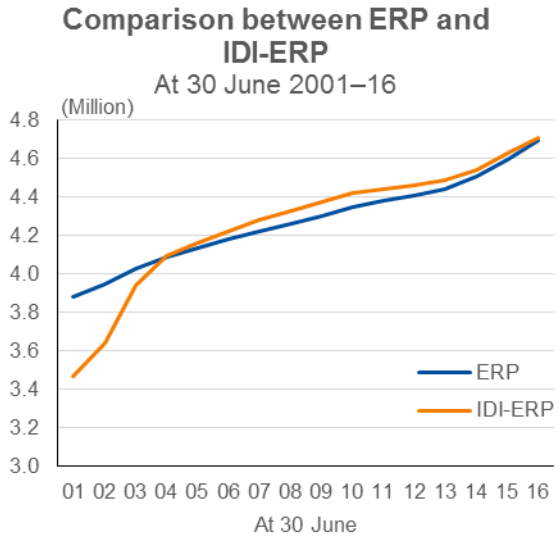
Results

We first produced the IDI-ERP at 30 June for the years 2001–16, and assessed the estimates against the ERP over time. Figure 3a compares the IDI-ERP and ERP for New Zealand. For 2001–03, the IDI-ERP is lower than the ERP, with not all activity data being available for these dates. From 2004 onwards, the IDI-ERP is higher than the ERP, peaking at 1.7 percent in 2010. Estimates from 2007–16 have been included in the published series.

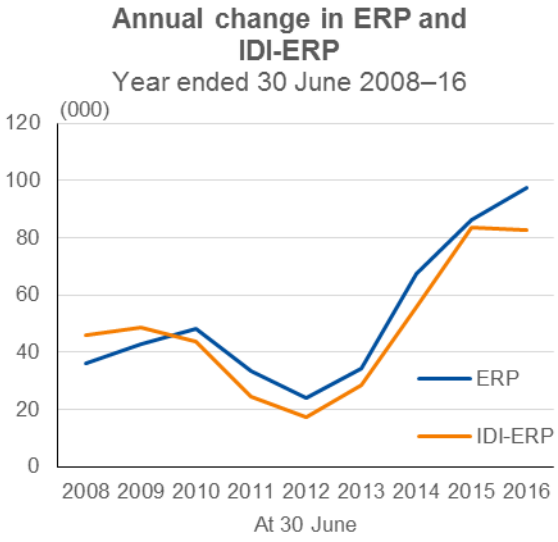
Figure 3b shows the annual change in both the ERP and IDI-ERP for 2008–16, the period covered by the experimental series. The patterns are generally very similar. The annual increase observed in the IDI-ERP is higher than the ERP for 2008 and 2009, but lower for the remaining years. These differences can partially be explained by the different migration measures being used, with the ERP making use of intentions data from arrival and departure cards. The larger difference for 2016 specifically is likely due to incomplete activity data through to June 2016 in the latest IDI refresh.

Compared with IDI-ERP v2, the population has increased for all years, due to the longer activity window. However, the annual change comparisons are generally consistent with those observed previously.

Figures 3a and 3b



Source: Stats NZ



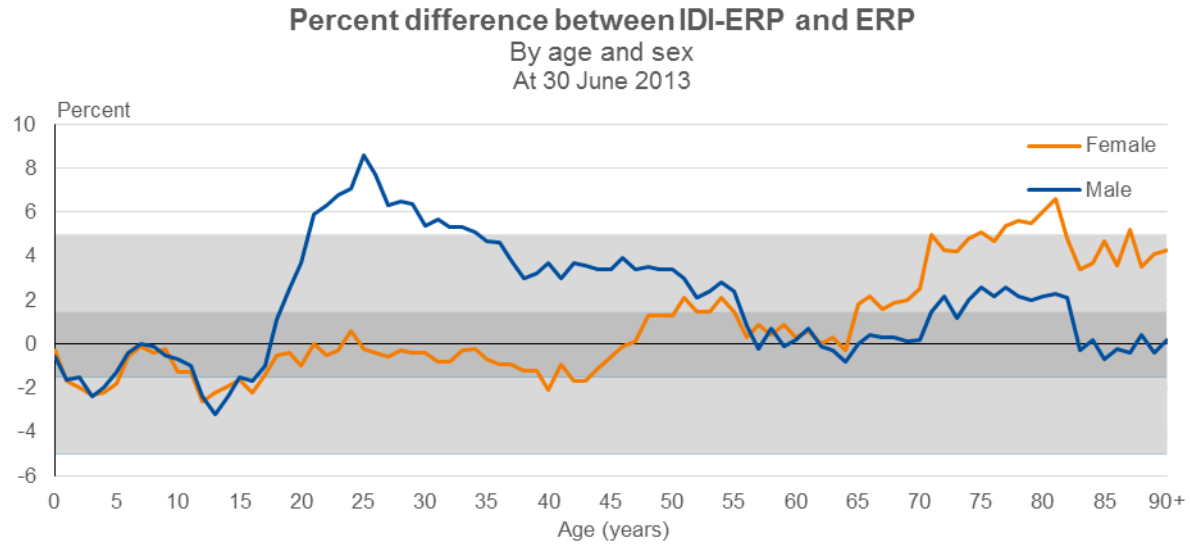
Source: Stats NZ

Differences between the ERP and IDI-ERP are not evenly distributed across age and sex. Figure 4 shows the percentage difference between the two measures at 30 June 2013, with the two grey boxes indicating the required quality standards. The IDI-ERP is considerably higher than the ERP for working-age males, particularly from 22–29. The IDI-ERP is lower than the ERP for children, and higher than the ERP for females aged 71 and over.

These patterns are similar to those observed in IDI-ERP v2. Undercoverage has been reduced for the female working ages, resulting in most of the estimates being within the strictest quality standards. However, overcoverage has increased for similarly aged males. For ages 0–16 and 65+, the results are almost identical, with neither of the improvements made having a significant impact on these age groups.

Overcoverage for the older ages, and particularly females, reflects the unresolved issue of duplicates in the IDI spine. Stats NZ (2016a) estimated that this could affect up to 67,000 individuals, mostly for females aged 45 and over. These duplicates will be resolved in an upcoming IDI refresh.

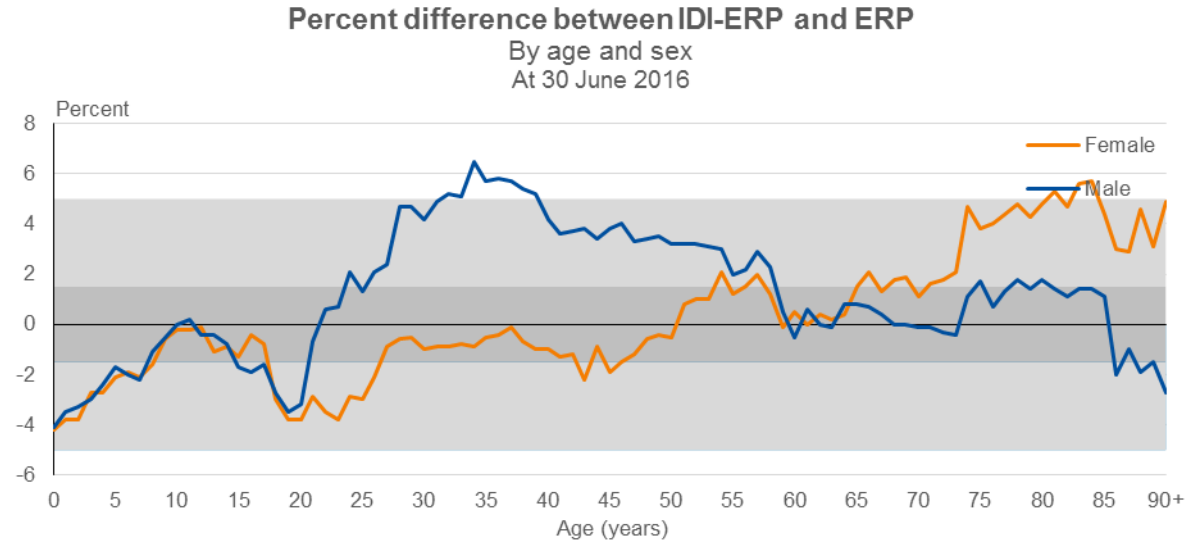
Figure 4



Note: The shaded areas represent the required quality standards for an administrative-based census model. 90 percent of estimates should be within the dark grey area, and all estimates should be within the light grey area.
Source: Stats NZ

Figure 5 shows the same comparison at 30 June 2016. While many of the broad trends are again similar, there are also some notable differences. To a large extent, the coverage patterns appear to shift by three years, reflecting the ageing of the population between the two sets of estimates. For example, the coverage for age 7 in 2013 closely resembles the coverage for age 10 in 2016. This indicates there may be underlying differences in the cohort populations. Given the IDI-ERP is constructed independently for each year, we wouldn't expect large cohort effects. Therefore, they may be indicative of some uncertainty in the ERP.

Figure 5



Note: The shaded areas represent the required quality standards for an administrative-based census model. 90 percent of estimates should be within the dark grey area, and all estimates should be within the light grey area.
Source: Stats NZ

There are additional changes, even after adjusting for this ageing trend. Coverage of the IDI-ERP compared with the ERP is lower in 2016 than 2013 for the youngest ages, reflecting births that have not yet been registered. An adjustment is made for unregistered births in the official population estimates, but not currently in the IDI-ERP.

The remaining changes between 2013 and 2016 are more complicated to assess. Most notably, coverage of the IDI-ERP compared with the ERP is:

- higher in 2016 than 2013 for ages 15–17
- considerably lower in 2016 than 2013 for females aged 18–27 and males aged 18–32.

There are a number of possible reasons for these changes. Overall, both the ERP and the IDI-ERP are likely to be less reliable than in 2013. The ERP tends to be the highest quality in census years and reduce in accuracy over time. For the IDI-ERP, some health activity data was not available through to June 2016, so individuals could potentially be missed. We were also unable to use the preferred migration method for the IDI-ERP in 2015 and 2016.

These differences in migration methods are likely to account for many of the changes. The young adult ages where we have observed the largest changes in relative coverage are also those with the greatest migration flows, which we know are difficult to reliably measure. The most recent migrants could also be missing from the IDI-ERP if they were not active in any administrative data sources prior to the end of the year.

Comparing the IDI-ERP with the ERP at the subnational level

This section compares the IDI-ERP with the ERP by territorial authority and Auckland local board area, and by area unit.

Selecting an address for individuals in the IDI-ERP

We first assessed the quality of the address information available in the IDI. Both coverage and accuracy differed across sources, although in general, more recent addresses tended to be of higher quality. Overall, 99 percent of the IDI-ERP had an address in at least one source by 2013. More detailed findings can be found in Appendix A: Quality of address information in the IDI.

Based on these investigations, we developed a set of rules to allocate a single address to each individual identified as being in the IDI-ERP at a given date. This method generally favours using the most recent address, regardless of source, similar to that described by Gibb and Das (2015). However, certain sources were deemed to be of lower quality, and therefore only used when other address information was not available. For ACC addresses specifically, we observed an increase in quality over time, so have moved these addresses into the higher quality group from 2014 onwards. By applying this method, 84 percent of individuals were assigned to the same meshblock as recorded in the 2013 Census.

For IDI-ERP v3, we apply the following method:

For address notifications through to 31 December 2013:

- select the most recent address notification from any of IR (excluding dates with high notification counts), NHI, PHO, MOE or MSD residential
- if no such address exists, select the most recent address notification from any of the remaining sources (ACC, MSD postal, and IR dates with high notification counts).

For address notifications from 1 January 2014 onwards:

- select the most recent address notification from any of IR (excluding dates with high notification counts), NHI, PHO, MOE, MSD residential or ACC
- if no such address exists, select the most recent address notification from any of the remaining sources (MSD postal and IR dates with high notification counts).

NZTA drivers licence and motor vehicle registrations were also identified as being high quality sources of address. However, no historical back series was available, so they have not been implemented in this experimental series.

Comparing the IDI-ERP with the ERP at the subnational level

Every individual in the IDI-ERP who had an address in any administrative data source was then allocated to the corresponding geographic area. No adjustment was made for individuals without any address information (5.8 percent of the IDI-ERP in 2007 decreasing to 0.7 percent in 2016). The resulting estimates were compared against the ERP.

Table 2 summarises the performance of the IDI-ERP at 30 June 2013 against the quality standards. The fourth column shows the required percentage of areas within the given level of error, while the final column shows the actual result for IDI-ERP v3. Currently, none of the breakdowns are meeting

the quality standards, although some are very close. In particular, the estimates for large TALBs by five-year age group are almost exactly meeting both of the required standards, with 83 percent of these estimates within 5 percent of the ERP (compared with the required 85 percent) and 99 percent are within 12.5 percent of the ERP (compared with 100 percent). There are larger discrepancies with the quality standards for the total populations, although this could be partially due to the coverage errors observed at the national level.

Table 2

Performance of IDI-ERP against quality standards
At 30 June 2013

Geographic area	Population unit	QS level of error (within +/- percent)	QS percent required within level of error	Percent of IDI-ERP v3 within level of error
Territorial authorities & Auckland local boards, population 100,000 or more	Total population	2.5	100	75
	By five-year age group and sex	5 12.5	85 100	83 99
Territorial authorities & Auckland local boards, population less than 100,000	Total population	2.5 5	85 100	52 86
	By five-year age group and sex	5 20	80 100	64 97
Area units, population 500 or more	Total population	5 10	80 100	62 86
Area units, population 100–499	Total population	10 25	80 100	56 83

Source: Stats NZ

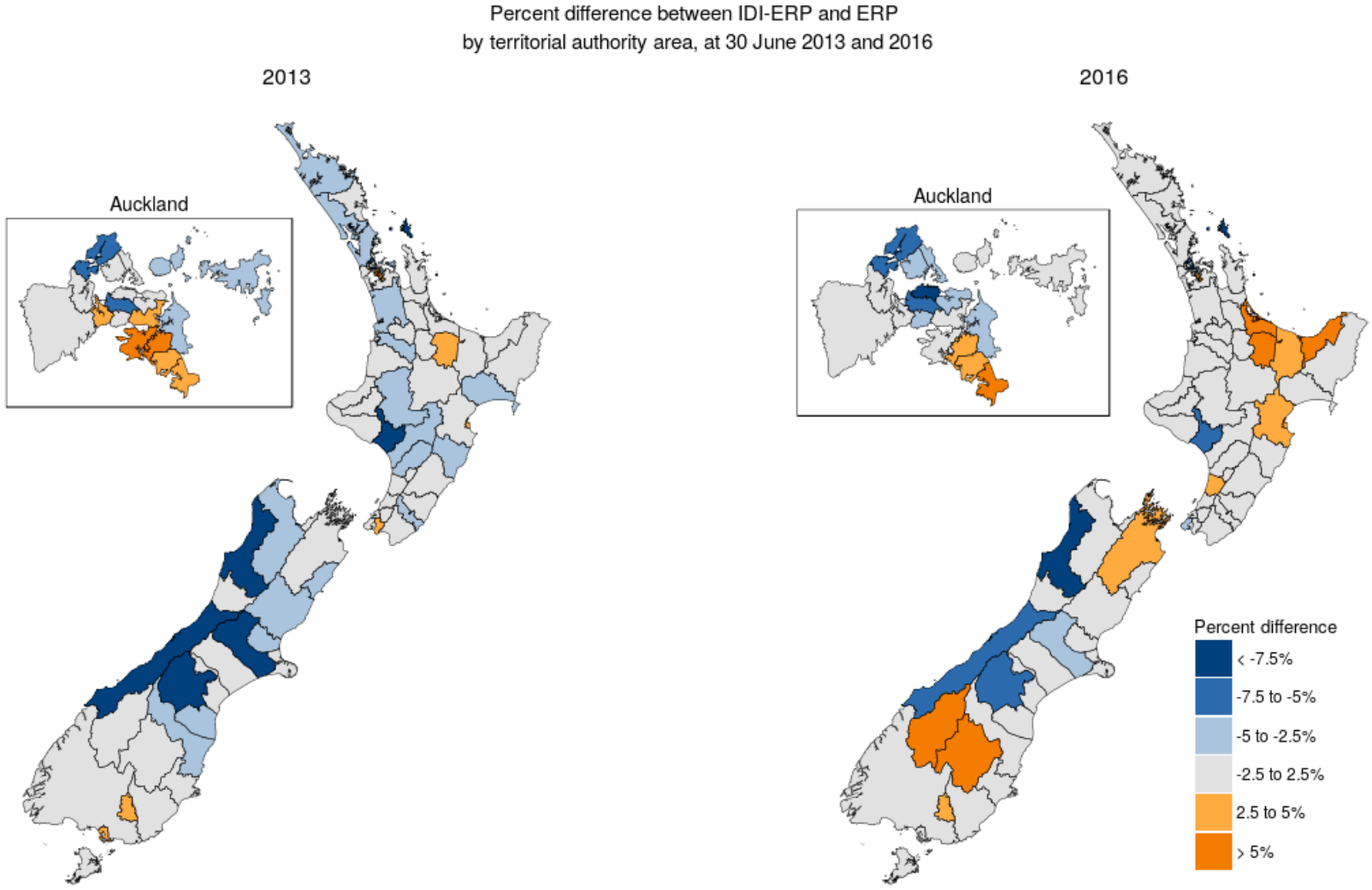
Territorial authority and Auckland local board areas

Total populations

Differences between the ERP and IDI-ERP vary across TALB areas. Figure 6 shows the relationship at 30 June 2013 and 30 June 2016. Areas shaded in white are within 2.5 percent of the ERP, meeting the most restrictive quality standard. Areas in blue are being underestimated and areas in orange are being overestimated.

Overall, 58 percent of territorial authorities and 43 percent of Auckland local boards were within 2.5 percent in 2013, compared with 70 percent and 43 percent in 2016, respectively. The patterns are quite different between the two years. In 2013, the majority of TAs outside the quality standards are being underestimated, while in 2016, more areas are being overestimated. For the Auckland local board areas, the trend is the opposite, with more areas being underestimated in 2016.

Figure 6


Percent difference between IDI-ERP and ERP
by territorial authority area, at 30 June 2013 and 2016
2013
2016
Auckland
Auckland
Percent difference
< -7.5%
-7.5 to -5%
-5 to -2.5%
-2.5 to 2.5%
2.5 to 5%
> 5%

The largest differences at 30 June 2013 are summarised in Appendix B1. The two worst performing areas are Chatham Islands (underestimated by 89 percent) and Great Barrier Island local board (underestimated by 36 percent). Other than these small island areas, five areas are underestimated by more than 8 percent, including four South Island districts, and Whanganui in the North Island. The three areas most overestimated in the IDI-ERP are local boards in South Auckland. Each of these areas exhibits similar coverage trends across other years.

More broadly, however, there is evidence that the IDI-ERP has been getting closer to the ERP over time. Appendix C: Relative errors between ERP and IDI-ERP shows the proportion of missing addresses, and the relative differences between the two measures for each year. Also included for comparison are the errors observed in the rolled-forward 2006-base estimates at 30 June 2013. These are used to represent the minimum quality achieved in the official population estimates.

By all three measures, the differences between our admin-based estimates and the official ERP decrease steadily through to 2014, before stabilising in 2015 and then increasing in 2016. . There is a noticeable reduction in the differences in 2009, corresponding to improvements in the NHI addresses. Many of the early gains are likely to represent the reducing number of individuals with no address information, but also point to some improvement in the administrative data. The growing differences between the ERP and IDI-ERP for 2015 and 2016 could once again indicate a reduction in the quality of the ERP, as the estimates get further from the 2013 Census.

Five-year age groups

As with the national level comparisons, similarities at the total level can disguise differences by age group. Table 3 shows the performance against the quality standards by five-year age group and sex. The cells shaded in grey are those meeting the required standards. For large TALBs, most of the groupings are within the required levels.

For the smaller TALBs there is more disagreement, although a number of groupings are again very close. If we excluded the Chatham Islands and Great Barrier Island areas, 24 of the 36 age by sex combinations have all estimates within 20 percent of the ERP.

Across all of the standards, males aged 15–29 are generally the worst performing, while children aged 5–14 and the older ages tend to be closest to the standards.

Table 3

Performance against quality standards for TALB areas
By five-year age group and sex
At 30 June 2013

Age group	TALBs >= 100,000				TALBs < 100,000			
	Within 5% (QS = 85%)		Within 12.5% (QS = 100%)		Within 5% (QS = 80%)		Within 20% (QS = 100%)	
	Male	Female	Male	Female	Male	Female	Male	Female
0–4	75	75	100	100	44	41	96	97
5–9	100	100	100	100	71	73	97	97
10–14	88	88	100	100	58	66	97	97
15–19	38	50	88	88	63	70	96	94
20–24	75	75	75	100	23	57	94	96
25–29	38	100	100	100	30	67	95	96
30–34	50	100	100	100	41	77	96	97
35–39	75	100	100	100	58	76	96	97
40–44	63	100	100	100	59	71	99	97
45–49	75	100	100	100	66	81	97	99
50–54	75	88	100	100	68	85	97	99
55–59	88	88	100	100	67	73	97	99
60–64	88	75	100	100	59	76	97	97
65–69	100	88	100	100	78	81	97	97
70–74	100	75	100	100	78	68	97	97
75–79	100	88	100	100	75	57	97	99
80–84	100	88	100	100	72	62	100	97
85+	100	75	100	100	59	53	95	96

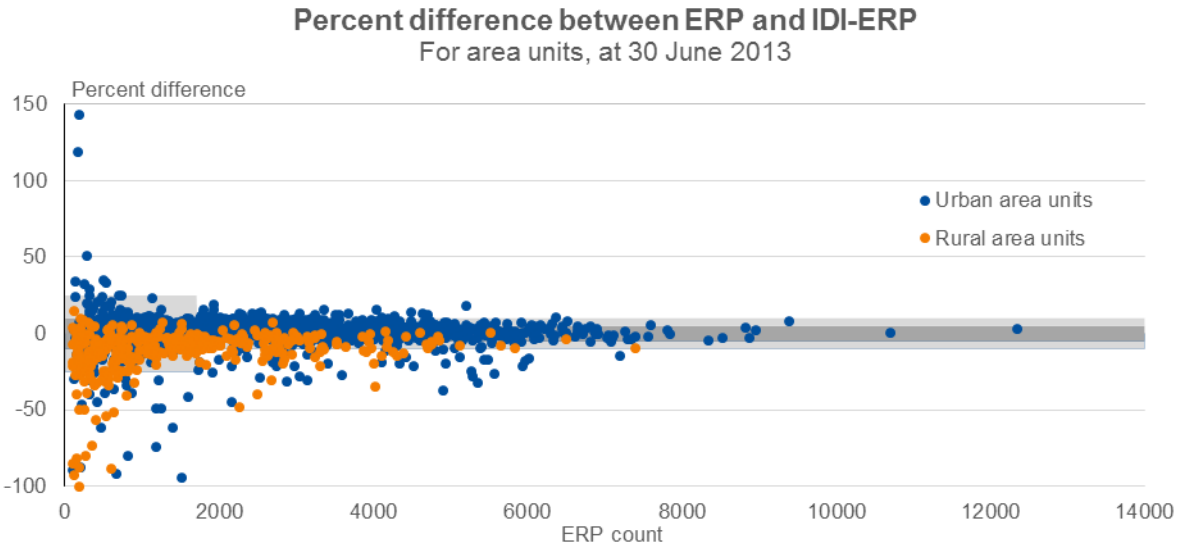
Note: Shaded values are those meeting the required quality standards for an administrative-based census model.
Source: Stats NZ

Area units

Comparisons between the ERP and IDI-ERP were also made for area units. Figure 7 shows that the relative differences vary even more for these areas than for TALBs. The two grey bars represent the required quality standards, with a wider allowance for areas with populations below 500.

Of the 1,826 area units with a population greater than 100 in 2013, 58 percent were within 5 percent of the ERP and 82 percent were within 10 percent of the ERP. The largest relative differences between the ERP and IDI-ERP tend to occur for smaller areas, and they are more likely to be considerably underestimated than overestimated. Compared with the ERP, there is also a clear tendency for rural areas to have lower coverage in the IDI-ERP than urban areas.

Figure 7



Note: The shaded areas represent the required quality standards for an administrative-based census model. 80 percent of estimates should be within the dark grey area, and all estimates should be within the light grey area.
Source: Stats NZ

Appendix C2 shows the average errors by area unit for 2007–16. Similar to the comparisons for TALBs, the measures generally decrease through to 2014, then increase again in 2015 and 2016. The errors are larger than for TALBs, reflecting the larger uncertainty involved for smaller areas.

Sources of error in subnational populations

We have observed notable differences between the ERP and IDI-ERP, particularly by geographic area. In general, these differences are likely to represent a number of sources of error, and interactions between the various factors can make it difficult to interpret the results. We are particularly interested in errors that are not evenly distributed across areas, because these will have a more significant effect on the final estimates. In this section, we describe some of the main sources of error. The following section presents four areas as examples.

National-level errors

Coverage errors at the national level will also affect estimates by geographic area, and will not necessarily indicate errors specific to that area. For example, we have observed the national IDI-ERP aged 25 is 8.6 percent larger than the ERP. This effect will naturally result in some areas also being overestimated compared with the ERP.

Missing address information

Within the IDI-ERP for each year, some individuals have no address information available from any source. Without adjusting for the missing addresses, there will be undercoverage with individuals not assigned to their actual area of usual residence. The number of missing addresses has decreased over time, to around 0.7 percent in 2016.

Errors due to geocoding

The address matching process used in the IDI is not perfect and will therefore result in some valid addresses being incorrectly matched or not matched to a reference address at all. Both cases may result in individuals being included in the population of the wrong area. A thorough investigation of geocoding issues was not possible without access to the original address strings. However, our

analysis suggested that certain types of area could be more prone to unmatched addresses, including offshore islands and areas with certain types of non-private dwelling (eg prisons, defence establishments, selected universities, and boarding schools).

Delays in updating address information

None of the administrative data sources we have used require individuals to notify the agency immediately upon a change of address. Therefore, there tends to be a delay between a move occurring and the new address being recorded in the administrative data. In many cases, the address contained in the administrative data will accurately reflect an individual's usual residence at the time it was updated, but not their current location.

Differences between administrative and statistical purposes

For the purposes of producing population estimates, we wish to know the address at which an individual usually resides. However, this is not necessarily the case for all of the administrative collections. For example, some addresses in the IR data may be postal addresses. Similarly, individuals may list a parent's address or workplace address in some sources. These addresses may be perfectly valid for administrative purposes, but not representative of the individual's actual usual residence.

Errors in the ERP

The ERP itself will also be prone to error, meaning that differences between the ERP and IDI-ERP do not always reflect issues within the IDI-ERP. Bryant et al (2016) provided a measure of uncertainty for the 2013 base-ERP, largely reflecting sampling errors from the post-enumeration survey. At the national level, the relative uncertainty by single year of age (measured as half the width of the 95 percent credible intervals) is mostly between 0.5 and 1 percent. The uncertainty is larger when disaggregating by geographic area.

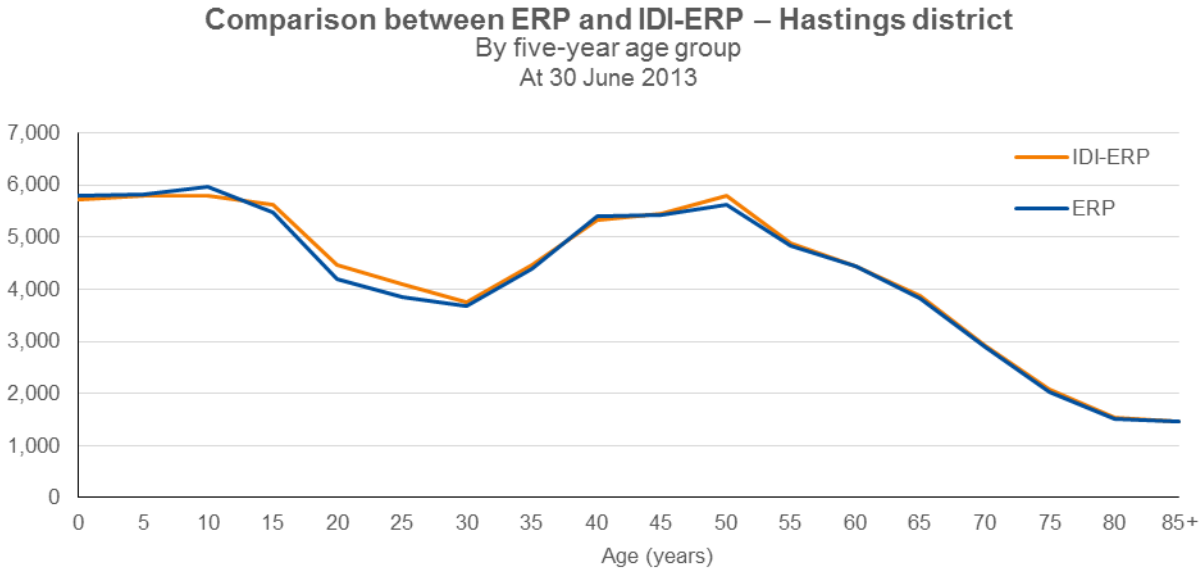
Case studies

In this section, we examine four specific TALB areas in more detail to understand the differences observed between the ERP and IDI-ERP. We look at the key types of errors that appear to be contributing in each case.

Hastings district

Many TALB areas exhibit differences which closely mirror those at the national level. Figure 8 shows that the ERP and IDI-ERP are broadly similar for Hastings district at 30 June 2013. The IDI-ERP is most noticeably higher than the ERP for ages 15–19 and 20–24.

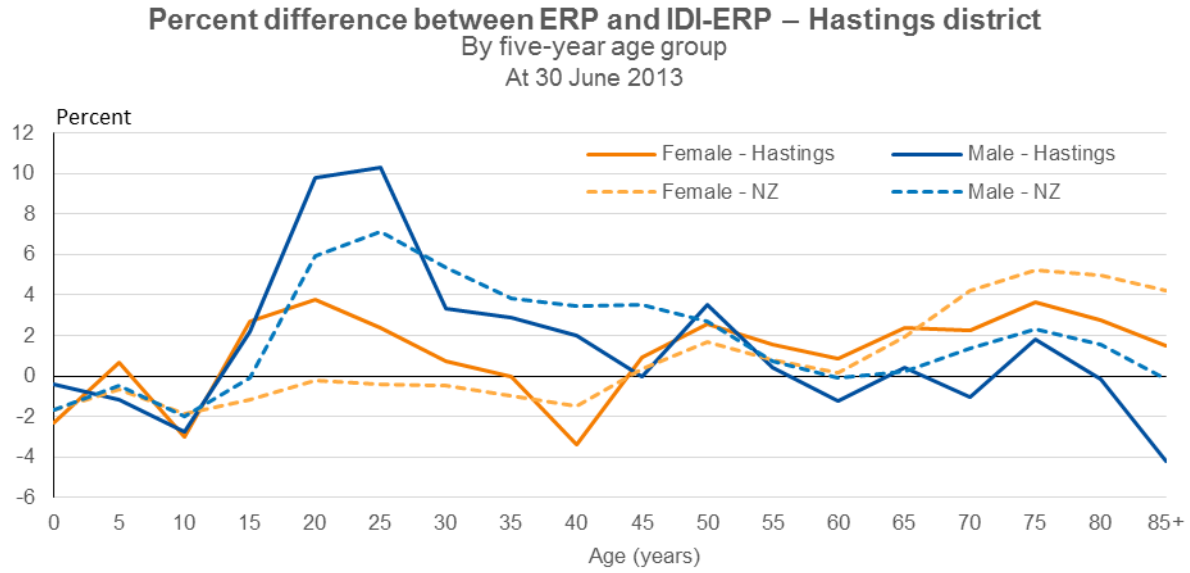
Figure 8



Source: Stats NZ

When we compare these differences with those observed at the national level, there are obvious similarities (figure 9). The peak overcoverage for young adult males, undercoverage for children, and overcoverage for older females are all apparent for both Hastings and the total New Zealand population. This may indicate that the errors observed are not specific to Hastings, but instead represent differences that exist throughout the entire IDI-ERP. It also highlights the difficulties of assessing area populations on their own.

Figure 9

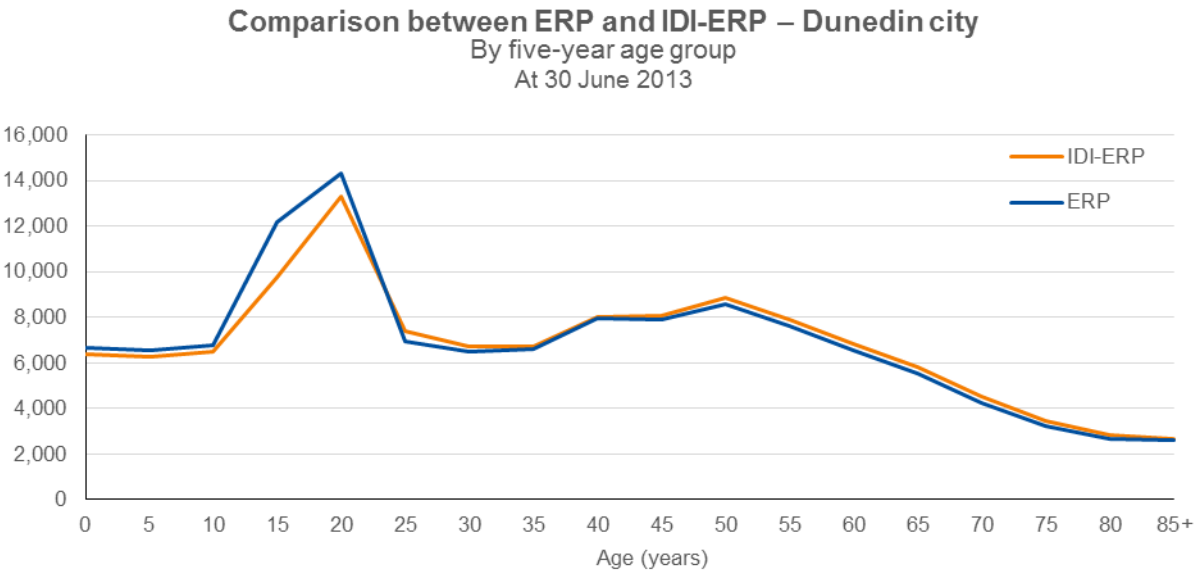


Source: Stats NZ

Dunedin city

University students are a population which cause sizeable differences between the ERP and IDI-ERP. Figure 10 shows a comparison of the two sets of estimates for Dunedin city, an area with one of the largest student populations. For most ages, the two measures are fairly similar, except for the 15–19 and 20–24 age groups, which are 20 percent and 7 percent lower in the IDI-ERP.

Figure 10



Source: Stats NZ

The differences are especially notable in those area units with the largest grouping of students. Otago University and North Dunedin, which had the highest proportion of students across the entire country according to the 2013 Census, are being underestimated by 28 percent and 27 percent respectively. This strongly suggests that the administrative data sources are not accurately recording the correct addresses of all of these students.

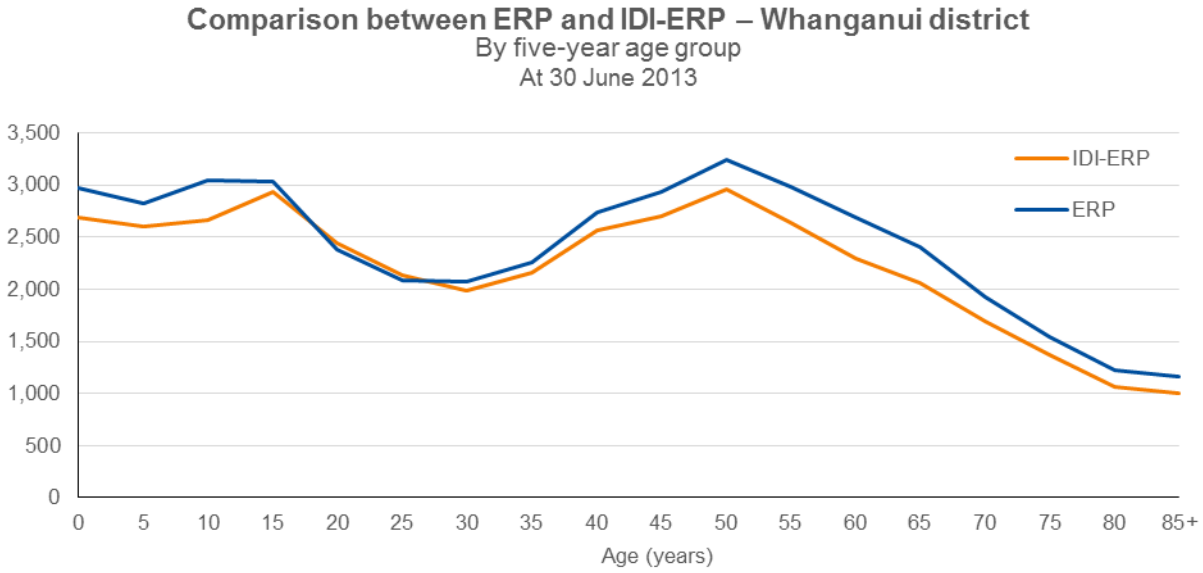
There are potentially multiple contributing factors. For students who have moved to Dunedin from other areas, they simply may have had no need to update their details in administrative data. They also may choose to use their parents' address, rather than a term-time address. For these reasons, more direct sources of address related to their tertiary study could be valuable. Within the IDI, there is information on the TA of the institution they are enrolled at. However, this is less precise than the other sources we have used, so has not been implemented at this point.

Differences across the student ages are notable across a range of areas. Other TALBs with large student populations, such as Palmerston North and Hamilton, show similar patterns to Dunedin. On the other hand, many of the other TALBs without universities overestimate the student ages compared with the ERP.

Whanganui district

Whanganui district is one of the areas with the largest differences between the ERP and IDI-ERP across all years. Overall, at 30 June 2013 the IDI-ERP was 8 percent lower than the ERP, with this difference apparent across most age groups (figure 11).

Figure 11



Source: Stats NZ

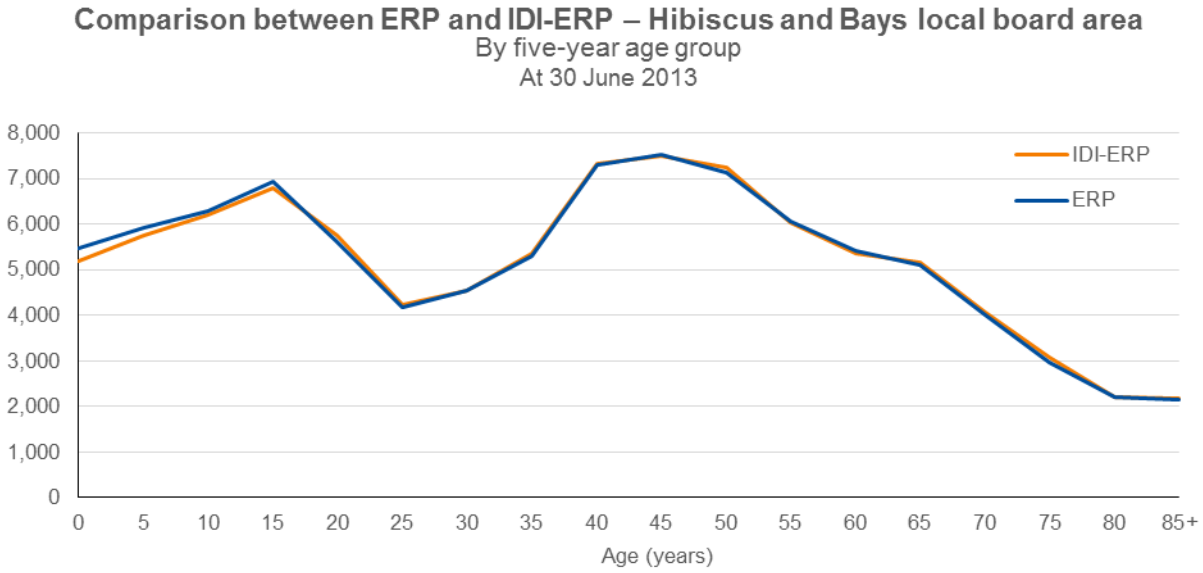
One of the key causes appears to be issues with geocoding addresses to certain areas within Whanganui. With the exception of the Chatham Islands, Whanganui had the highest proportion of usual residents (according to the census) whose most recent administrative address was unable to be geocoded. This was particularly noticeable across all sources in two specific area units (Bastia Hill and Durie Hill) and across the entire TA in the PHO data. As a result, a number of individuals are not being correctly allocated to Whanganui in the IDI-ERP.

Unlike many of the other types of error, this issue is not directly related to the quality of the administrative data, with many of the provided addresses likely to be accurate. Instead, it represents limitations of the various processes involved in making these addresses available for further use. For Whanganui specifically, these geocoding errors make interpretation of the actual address quality considerably more difficult. Therefore, it is important that we aim to minimise such errors.

Hibiscus and Bays local board area

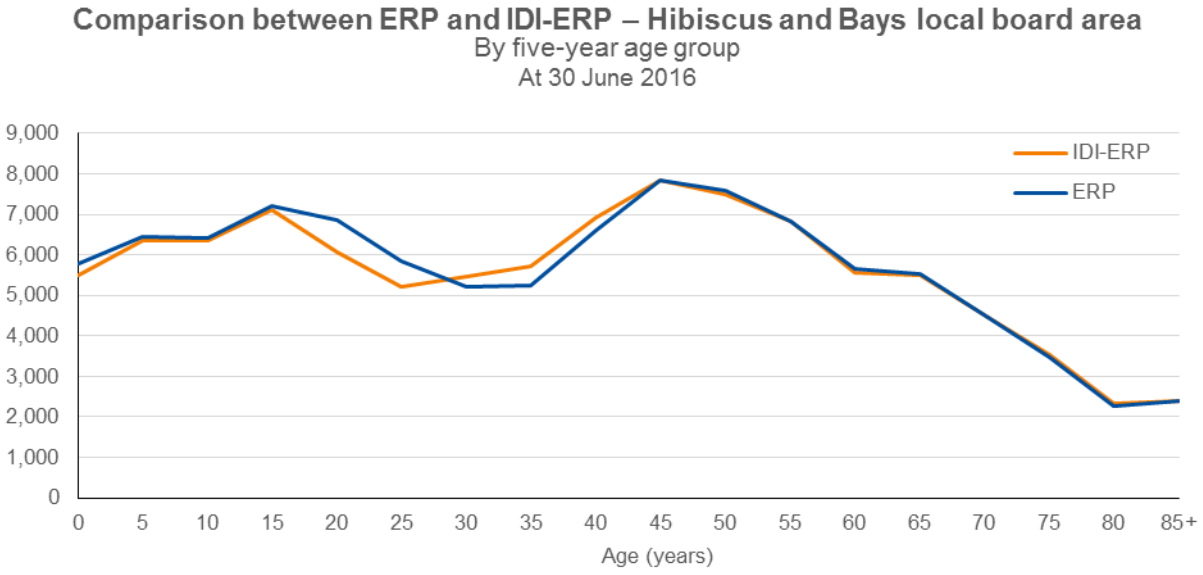
Finally, there are considerable changes in the relationship between the ERP and IDI-ERP for many areas over time. Figures 12 shows a comparison for the Hibiscus and Bays local board area in 2013, while figure 13 shows the same comparison for 2016.

Figure 12



Source: Stats NZ

Figure 13

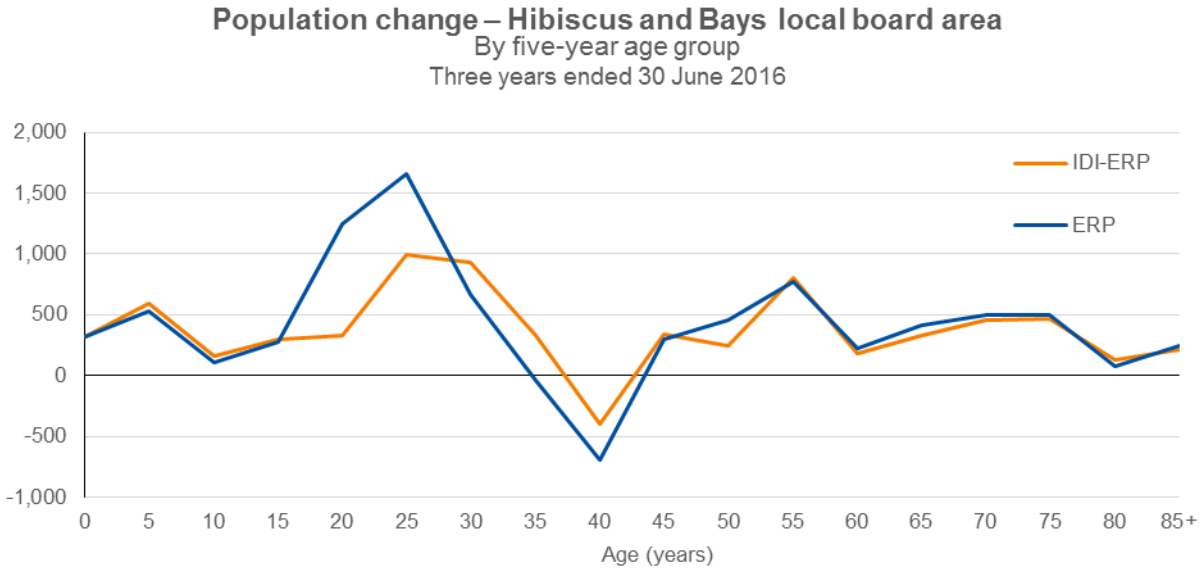


Source: Stats NZ

As we have observed with many of the other areas, the differences between the ERP and IDI-ERP is relatively minimal for children and for the older ages. For Hibiscus and Bays, the differences for these groups also appear to be fairly consistent in both 2013 and 2016. However, there are clear differences for ages 20–44. Specifically, in 2016 ages 20–29 are being underestimated compared with the ERP, while ages 30–44 are being overestimated compared with the ERP.

Figure 14 illustrates this point further, by comparing the respective population change between 2013 and 2016. These results are driven by different migration patterns – either from overseas or from other areas within New Zealand – between the ERP and IDI-ERP. There is uncertainty around each set of estimates, and in the case of the ERP specifically, this increases with each year further removed from the 2013 Census. Further analysis of the differences could be beneficial for improving the quality of both sets of estimates.

Figure 14



Source: Stats NZ

Discussion

The ability to produce population estimates for New Zealand, and by subnational area, is crucial for any census model. This paper discusses our progress in developing these estimates using the linked administrative data in the IDI. We have compared our administrative-based estimates with the official population estimates series, and assessed them against the required quality standards.

Summary of results

Since the first release of these experimental population estimates, we have made improvements to the method used to identify the New Zealand resident population from administrative data sources. By incorporating improved rules for classifying overseas residents and by lengthening the activity period, individual components of both overcoverage and undercoverage have been reduced. Despite these adjustments, the national-level estimates are not meeting the required quality standards. For example, the number of young adult males, as well as females aged 70 and older, are higher in the IDI-ERP than the ERP. However, we have confidence that further improvements, most notably related to improved linkages within the IDI, will bring the estimates closer to the ERP.

There is considerable variability in the quality of estimates by subnational area. We have applied a method to select an address for each individual from a range of administrative data sources. As a result, more than 99 percent of individuals have an address, and 84 percent of these individuals are assigned to a meshblock consistent with the 2013 Census. Currently, our IDI-ERP by subnational area is not meeting the quality standards required of an administrative census, although the results are promising for many areas and age groups.

There are potentially a number of factors that contribute to differences between the ERP and the IDI-ERP for subnational areas. The types of error observed at the national level will also affect subnational populations. Some errors appear to be caused by addresses being incorrectly geocoded or not geocoded at all. Perhaps the largest source of error is due to the delay in reporting a change of address, with the administrative addresses far less reliable for individuals who have recently moved. In such cases, there is often no incentive to immediately notify any government agencies, resulting in the movements not being recorded for some time. Differences between statistical and administrative purposes mean that the addresses listed across sources may disagree, and won't necessarily reflect an individual's actual usual residence.

The combined effects of these sources of error are more apparent in certain groups. Those who are more mobile, such as young adults, will be more prone to having outdated addresses. Others who live in certain areas or non-private dwelling types may also have less reliable information. These result in many of the observed differences at higher geographic levels.

There is also uncertainty in the ERP itself, which increases as we get further from the 2013 Census. Comparisons over time are therefore useful, both for assessing the quality of the IDI-ERP and for understanding any limitations of the official ERP. The IDI-ERP population is constructed from administrative data independently for each year, using a consistent method. If data quality remained the same over time, we would expect to see similar patterns of undercoverage and overcoverage repeated over time.

For the national comparisons we do largely see consistency over time. The total IDI-ERP population is higher than the ERP for each year. Some differences, such as those caused by linkage errors, appear in much the same way each year, although there is also evidence of some cohort effects. For subnational areas we have seen agreement between the IDI-ERP and the ERP improve steadily over

time, most likely due to improvements in the coverage and accuracy of the administrative address data. The relative trends are largely similar across years, with the same areas consistently overestimated or underestimated. However, there are more considerable differences between 2013 and 2016, likely driven by high levels of migration, and the different methods being applied.

Timeliness

The focus of this paper has primarily been on assessing whether the administrative data is of high enough quality to produce reliable population estimates. However, for any potential census model, these estimates would also have to be produced in a much more timely manner.

At present, delays in the availability of administrative data within the IDI limit how quickly the IDI-ERP can be derived, with the most recent estimates available more than a year out of date. Some of these delays are an effect of events being registered and are largely unavoidable, such as late birth registrations and the 16 months required to use the 12/16 month migration rule. These examples will likely require alternative estimation methods, similar to those already implemented in the official ERP processes.

However, other delays could be reduced with a more regular supply of data. For example, in the June 2017 IDI refresh, some health data used to determine activity were only available through to March 2016. Unlike previous examples, these delays do not come from lags in the underlying data being provided, but instead, in that data making its way through various systems. There is more potential for improvements to minimise these delays.

More work should be undertaken to assess the timeliness requirements, and to understand which delays can be reduced to an acceptable level.

Improvements to data sources

Improvements to linkages in the IDI are crucial to the production of high quality population estimates. Most importantly, the presence of duplicates in the IDI spine needs to be resolved to reduce the overcoverage of older females.

We will continue to reassess our methods, in terms of the sources used and the length of time applied in the activity rules. The ACC data is currently adding only a small number of individuals, and we would not expect many people to have these claims without any other forms of activity. On the other hand, sources targeted towards specific populations, such as corrections data, could be added if there is evidence that these people are being missed otherwise. The migration data could be used more explicitly as a source of activity, rather than simply to remove overseas residents. This would ensure that all individuals who have moved to New Zealand are included, even if they have not yet made use of other administrative services.

At the subnational level, further improvement is needed as well, in terms of both the administrative addresses themselves, and of our methods for combining multiple sources.

One key improvement would be further consistency in the standard of addresses collected across the system. Many government agencies are already moving towards better address validation, thereby reducing errors from poorly recorded details. A number of commercial address validation tools are being used across government. Consistency would be improved if a comprehensive New Zealand address list were openly available and used by all agencies. Ideally, this information should then flow directly through the system, avoiding the need for additional geocoding within the IDI.

Improvements to the IDI geocoding would also be beneficial for any addresses that require this. Work is ongoing to implement a new tool in the IDI which should improve the quality of address matching. As with any matching process, it is important that we understand and monitor any quality issues.

There are a number of additional sources of address which already exist within government. Two additional tables containing addresses are now being supplied by NZTA – covering drivers' licences and motor vehicle registrations. No historical back series is available, and each refresh includes only a single snapshot of the population, restricting our ability to make earlier comparisons. However, an assessment of the data compared with the HLFS suggested that the address information could be higher quality than any of the other sources we are currently using, particularly for the age groups that are otherwise difficult to locate. These sources will be incorporated into future iterations of the IDI-ERP.

The electoral roll is another source which is likely to contain high quality address information, although there are currently legislative barriers to accessing that register for production purposes. More targeted sources could also be implemented. Information on students from tertiary enrolments or from student loans data could provide more reliable addresses for a group that is currently difficult to accurately locate. Similarly, information from institutions such as prisons or defence establishments could be incorporated for those specific populations.

Conclusion

Overall, our progress in producing population estimates from administrative data is encouraging. Although the estimates are not currently meeting all of the quality standards, there are broad similarities with the official estimates, and there is evidence that the administrative populations are improving over time. Further improvements to the data and methods have been identified and will continue to increase the quality of these estimates.

In addition, we would not expect to fully meet accuracy requirements from administrative data alone, and are also developing statistical models to adjust for any errors. Resulting modelled estimates are the ones which need to meet the quality standards. More detail on this work will be released later in 2017. Any ongoing development will also contribute to an assessment of the quality of the administrative estimates against the ERP based off the 2018 Census, leading towards recommendations on a future census model.

We welcome your feedback

This paper presents the latest findings from our research. We are publishing these findings to update you of our progress and to invite your feedback, which will help us improve our methods. We welcome input on any of the methods or results discussed, and suggestions for other improvements or possible explanations for the observed differences.

To send your feedback, please complete this form.

References

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

Bryant, J, Dunstan, K, Graham, P, Matheson-Dunning, N, Shrosbree, E, & Speirs, R (2016). Measuring uncertainty in the 2013-base estimated resident population (Statistics New Zealand Working Paper No 16-04). Retrieved from www.stats.govt.nz.

Gibb, S, Bycroft, C, Matheson-Dunning, N (2016). Identifying the New Zealand resident population in the Integrated Data Infrastructure (IDI). Retrieved from www.stats.govt.nz.

Gibb, S. J. & Das, S. (2015). Quality of geographic information in the Integrated Data Infrastructure. Retrieved from www.stats.govt.nz.

Gibb, S, & Shrosbree, E (2014). Evaluating the potential of linked data sources for population estimates: The Integrated Data Infrastructure as an example. Retrieved from www.stats.govt.nz.

Lebel, A & Denis, J (2016). Assessing the usability of a statistical population register for the Census of Population in Canada. Paper presented at Meeting of the Group of Experts on Population and Housing Censuses, 2016, Geneva. Retrieved from http://www.unece.org.

McNally, J, & Bycroft, C (2015). Quality standards for population statistics: Accuracy requirements for future census models. Retrieved from www.stats.govt.nz.

Stats NZ (2012). Transforming the New Zealand Census of Population and Dwellings: Issues, options, and strategy. Retrieved from www.stats.govt.nz.

Stats NZ (2014a). Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project. Retrieved from www.stats.govt.nz.

Stats NZ (2014b). Estimated resident population 2013: Data sources and methods. Retrieved from www.stats.govt.nz.

Stats NZ (2016a). Experimental population estimates from linked administrative data: methods and results. Retrieved from www.stats.govt.nz.

Stats NZ (2016b). How accurate are population estimates and projections? An evaluation of Statistics New Zealand population estimates and projections, 1996–2013. Retrieved from www.stats.govt.nz.

Stats NZ (2017). Outcomes versus intentions: Measuring migration based on travel histories. Retrieved from www.stats.govt.nz.

Stats NZ (nd). Geographic data. In Data in the IDI. Retrieved from www.stats.govt.nz.

Appendix A: Quality of address information in the IDI

The ability to locate individuals within New Zealand is crucial for producing population estimates by geographic area. In order to understand the quality of address information available from administrative data, we looked at the levels of coverage across key sources and of consistency with our gold standard sources of address – the 2013 Census and the HLFS.

Coverage of geographic information in the IDI

Table A1 shows the coverage of geographic information from a number of administrative sources. We can see that coverage varies between sources, and has increased over time in all sources. IR, NHI, and PHO all have geocoded address information for more than 85 percent of the IDI-ERP by 2013, while the remaining sources cover less than half of the total population. In combination, more than 99 percent of individuals included in the IDI-ERP had a geocoded address from at least one administrative source.

Table A1

Coverage of geographic information
By administrative source

Data source	Age range	% of IDI-ERP with geographic information at 30 June			
		2007	2010	2013	2016
IR	All	78	87	90	92
IR (excluding outlier dates)	All	78	85	88	91
NHI	All	49	79	87	89
PHO	All	75	82	85	88
MOE	6–15	0	7	32	83
ACC	All	31	34	36	51
MSD	18+	38	49	56	59
Any	All	94	98	99	99

Source: Stats NZ

Comparison with the 2013 Census

The 2013 Census is considered to contain the most reliable source of address available for individuals at a specific point in time. Although not perfect, these census addresses provide a useful measure for comparing the administrative sources. We assessed the accuracy of the administrative data by comparing address information in the various sources at 5 March 2013 with that from the 2013 Census. Table A2 shows the match rates for individuals with an address recorded in both datasets (3,763,900 in total). The first column is based on the geocoded addresses from the address notification table, while the area comparisons are based on the original meshblock as recorded in the census itself.

NHI and PHO had the highest levels of agreement, with 75 percent and 73 percent respectively having the same exact address as census. IR, ACC, MOE, and MSD residential all had similar match rates of 66–67 percent. The only source with notably low match rates was MSD postal addresses, with less than one-third of individuals having the same address recorded as in the census.

Applying our method to select a single address for each individual results in 82 percent of people having the same combined address as census, considerably higher than any source on its own.

By comparison, 86 percent of individuals had a match with the census address in at least one source. This represents an upper limit on the level of accuracy currently achievable by selecting only from the available addresses. The remaining 14 percent will have no chance of being assigned the correct address without additional information. The difference between the final two rows indicates that 4 percent of individuals had the correct address available, but this address was not selected under our current method.

For all of the data sources, match rates increase as the size of the geographic area increases. Agreement was on average 2 percent higher for meshblocks than individual addresses, 4 percent higher for area units than meshblocks, and 12 percent higher for TALBs than area units.

Table A2

Percent of individuals with address information matching 2013 Census
By administrative data source

Data source	% of non-missing matching 2013 Census				
	Address id	Meshblock	Area unit	TALB	TA
IR	66	68	72	86	91
IR (excluding outlier dates)	67	69	73	86	91
NHI	75	77	80	90	93
PHO	73	75	79	88	93
MOE	66	68	73	87	91
ACC	67	69	73	86	90
MSD residential	66	68	72	84	89
MSD postal	27	29	36	61	68
Combined address	82	84	87	93	96
Any source[1]	86	89	91	95	97

1. Percent of total linked IDI-census population (including those with no address)

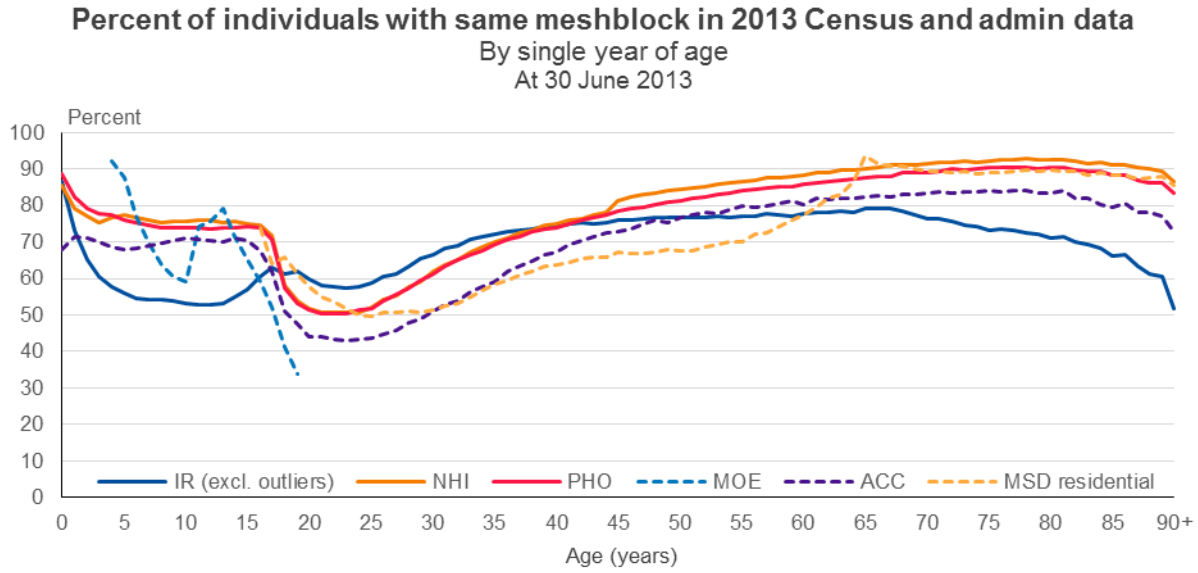
Source: Stats NZ

Figure A1 shows that these match rates differ considerably by age and across sources. Generally, accuracy is relatively high for young children, dropping significantly at age 18. The match rates remain low throughout the twenties, before steadily increasing.

The two health sources have the highest level of agreement with the census for most ages, with exceptions at ages 4–5 and 13 (where MOE is higher), 18–35 (IR), and 65–66 (MSD). These results appear reasonable based on the ages where people are likely to have more contact with certain agencies. The MOE addresses have high match rates specifically for ages 5 and 13, corresponding to enrolment at primary and secondary school respectively. The MSD addresses appear to be of consistently high quality for ages 65 and older when most will be receiving superannuation. The IR addresses are comparatively good for the young working ages where there are low match rates in all sources, but look less useful for other age groups.

Although not shown, match rates are slightly higher for females than males, particularly in the two health sources, and between ages 25 and 40.

Figure A1



Source: Stats NZ

We also explored other variables which could affect the likelihood of having high quality address information in administrative data. These will be explored in more detail in an upcoming paper focused on the quality of administrative address information. The key findings suggested consistency with the 2013 Census was:

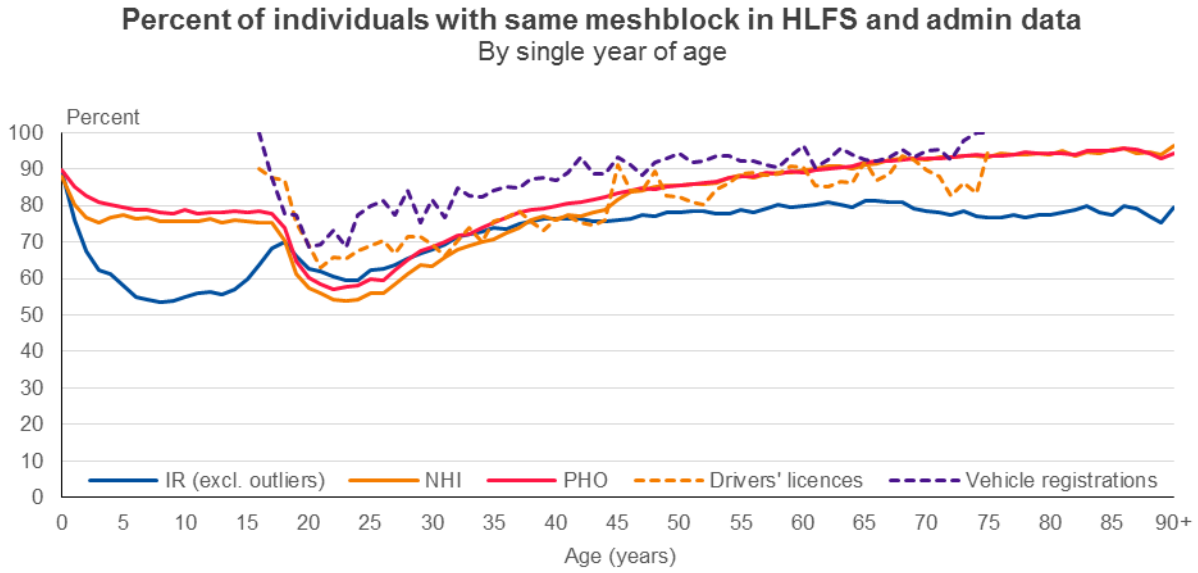
- higher for individuals who stated they had been at their current address for a longer period of time
- higher for administrative addresses which had been updated more recently
- higher for individuals in private dwellings and residential care facilities. Certain non-private dwelling types, such as prisons, defence establishments, and educational institutions, had particularly low match rates.

Comparison with the HLFS

Given the observed high quality of HLFS addresses, we also performed comparisons with the administrative sources for each HLFS quarter. These provide a measure of how quality is changing over time. We found that:

- The relationship between sources was relatively consistent with the comparisons against the 2013 Census. Most sources also remained reasonable steady over time.
- ACC addresses improved considerably across the comparison period.
- The IR addresses have notable decreases in quality, coinciding with the occurrence of the previously identified outlier dates. With those records excluded, the IR accuracy remains considerably higher. However, it is still lower than other sources.
- NZTA addresses were more consistent with the HLFS than any of the other sources, particularly for young adult males (figure A2). However, no historical series was available so these have not been included in our estimates.

Figure A2



Source: Stats NZ

Appendix B: Largest differences between ERP and IDI-ERP

Table B1

TALB areas with largest differences between ERP and IDI-ERP At 30 June 2013			
TALB area	**ERP**	**IDI-ERP**	**Difference (%)**
Chatham Islands territory	600	72	-88.0
Great Barrier local board area	950	612	-35.6
Mackenzie district	4,300	3,693	-14.1
Westland district	8,570	7,644	-10.8
Selwyn district	46,700	42,690	-8.6
Whanganui district	43,500	39,948	-8.2
Buller district	10,650	9,783	-8.1
Gore district	12,400	12,909	4.1
Rotorua district	68,400	71,670	4.8
Kawerau district	6,650	6,969	4.8
Papakura local board area	48,200	50,511	4.8
Otara-Papatoetoe local board area	80,300	84,858	5.7
Mangere-Otahuhu local board area	75,300	80,151	6.4
Source: Stats NZ			

Appendix C: Relative errors between ERP and IDI-ERP

Table C1

Absolute percentage error (APE) between ERP and IDI-ERP
Territorial Authority and Auckland local board areas
At 30 June 2006–16

At 30 June	Missing (%)	Mean APE (%)	Median APE (%)	WMAPE (%)
2007	5.8	10.1	5.1	6.0
2008	4.8	8.7	4.4	5.1
2009	2.8	6.5	3.7	3.6
2010	2.3	5.5	3.2	3.0
2011	1.9	5.1	2.9	2.8
2012	1.6	4.6	2.4	2.6
2013	1.4	4.1	2.0	2.2
2014	1.1	3.6	1.8	2.0
2015	0.9	3.6	1.8	2.2
2016	0.7	3.8	1.9	2.6
Rolled-forward 2006-base ERP (2013)	..	2.6	2.4	2.8

Source: Stats NZ

Table C2

Absolute percentage error (APE) between ERP and IDI-ERP
Area units
At 30 June 2006–16

At 30 June	Missing (%)	Mean APE (%)[1]	Median APE (%)	WMAPE (%)
2007	5.8	16.2	9.3	12.4
2008	4.8	14.7	8.9	11.2
2009	2.8	11.8	7.6	8.7
2010	2.3	10.3	6.6	7.7
2011	1.9	9.5	6.1	7.1
2012	1.6	8.6	5.4	6.4
2013	1.4	7.1	4.2	5.2
2014	1.1	5.9	3.3	4.2
2015	0.9	5.8	3.4	4.3
2016	0.7	6.3	3.8	4.9
Rolled-forward 2006-base ERP (2013)	...	5.3	3.6	4.5

1. Area units with ERP >= 100 only

Source: Stats NZ