Statistics New Zealand
TATAURANGA AOTEAROA

Experimental population estimates from linked administrative data: methods and results

Census Transformation

New Zealand Government

Liability
While all care and diligence has been used in processing, analysing, and extracting data and information in this publication, Statistics New Zealand 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.

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

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. 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.

Note: All figures in this paper were rounded to protect confidentiality.

Citation
Statistics New Zealand (2016). *Experimental population estimates from linked administrative data: methods and results.* Retrieved from www.stats.govt.nz

ISBN 978-0-908350-70-4 (online)

Published in September 2016 by
Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

Contact
Statistics New Zealand 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

1 Introduction

Census Transformation in New Zealand

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

1. A focus in the short-to-medium term on modernising the current census model and making it more efficient.
2. A longer-term focus on investigating alternative ways of producing small-area population and social and economic statistics. This includes the possibility of changing the census frequency to every 10 years, and exploring the feasibility of a census based on administrative data (Statistics NZ, 2012, 2014a).

The next census in 2018 will be significantly modernised, including an online completion target of 70 percent and re-use of administrative data to support collection and processing.

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

Census transformation – a promising future (a 2015 Statistics NZ Cabinet paper) recommended that Statistics 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. An experimental series of national population estimates by age and sex has been released, derived from linked administrative data in the Integrated Data Infrastructure (IDI).

This paper describes the method used to produce this series, including the improvements made since previous census transformation publications. We compare the results with official population estimates and summarise the possible factors contributing to any observed differences.

We are publishing this research to update you on our progress and invite your feedback to support future development.

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.

2 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) developed a method for constructing a resident population using linked administrative data sources available in Statistics NZ's Integrated Data Infrastructure (IDI) at April 2013. They derived and compared administrative-based population estimates with official estimated resident population (ERP) figures at 30 June 2010. While clear limitations were identified in the administrative sources available at the time of the study, the results showed enough promise to continue with further investigations.

Development of the IDI in following years overcomes some of these limitations. Birth and death registrations and health data are now available. An extended spine structure incorporating births and visa data provides better coverage of the population not paying tax, especially children (Black, 2016). Health data provides a much broader source of activity information than before.

Following these developments, the method for selecting the population was modified to incorporate the additional data sources. Gibb, Bycroft, and Matheson-Dunning (2016) described the approach and results, comparing them with the ERP at 30 June 2013. The resulting population estimates were a clear improvement and further highlighted the potential of using linked administrative data sources. However, they also found 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, incorrect classification of migrants, and individuals not selected because they were not active in administrative sources.

Aims and scope

This paper updates the methods described in Gibb et al (2016) for constructing an administrative-based population using the IDI. We compare the resulting estimates against the ERP at the national level for 2012–14 and discuss potential reasons for any differences. The final section provides a brief discussion of our findings.

This paper accompanies the release of an 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. The main aim is to update users of our progress in improving the quality of these administrative-based estimates. Another key aim of this release is to provide a mechanism for receiving input from users about the experimental series. To give feedback on any of the findings please complete this form.

Previous work detailing population estimates by subnational area is discussed in Gibb and Das (2015), and the quality of ethnic group information is discussed in Reid, Bycroft, and Gleisner (2016). Estimates for these subpopulations are not explored further within this paper but additional work is scheduled for both groups in late 2016.

The potential for producing census attribute information (eg information about education, income, families, and households) is discussed in other work (O'Byrne, Bycroft, & Gibb, 2014; Shrosbree, 2015; Bycroft, Reid, McNally, & Gleisner, 2016).

Future releases

This release is the first in an ongoing series of New Zealand population estimates produced from linked administrative data. We will publish updates annually, wherein we will extend the time series, incorporate more detailed estimates, and explain any further methodological developments.

The next experimental series release in 2017 is expected to include:

- an update to the national age-by-sex series discussed in this release
- population estimates by subnational area
- population estimates by level 1 ethnic group.

Improvements to the method will be incorporated into future releases. We will also continue to make previous versions available to allow for comparisons over time.

3 Data sources

This chapter presents the data sources we used for this research.

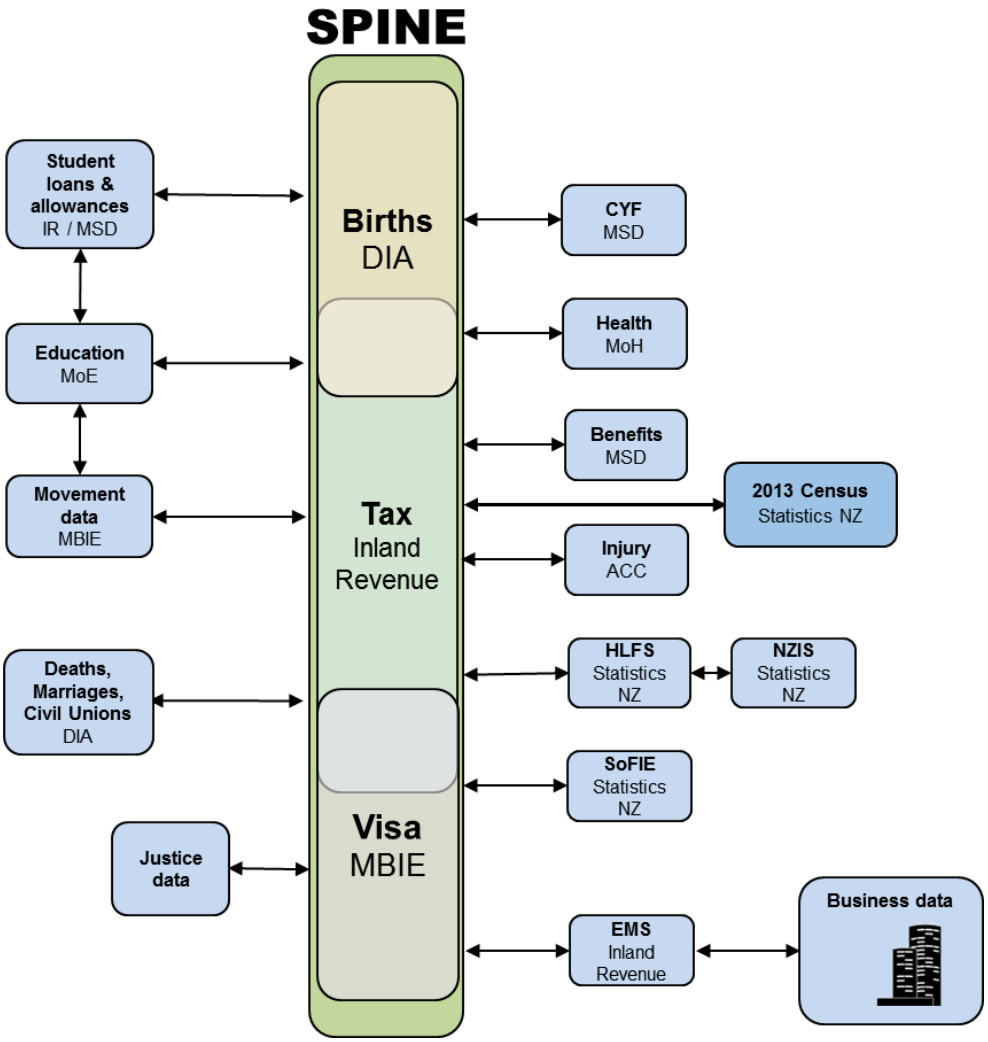
Integrated Data Infrastructure

Statistics 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 Statistics 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 IDI continues to change as new datasets are added (see Data in the IDI for current information).

The basic structure of the IDI is shown in figure 1. It 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.

Figure 1
Structure of the IDI



Three data sources are linked together probabilistically to create the spine:

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

The spine is the mathematical union of the three contributing data sources. People present in at least one source will be included in the spine. The linkages between the three contributing data sources aim to ensure that people present in any two data sources are included only once in the spine. Black (2016) provides more information on the formation of the IDI spine.

Other data sources are linked to the IDI spine (see Statistics NZ, 2014b 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; student loans and allowances data from several sources; migration movements data from Ministry of Business, Innovation and Employment; and data from Statistics NZ's Household Labour Force Survey, New Zealand Income Survey, and 2013 Census of Population and Dwellings.

The IDI also contains several summary tables that provide core information about individuals (age, sex, ethnicity, and geographic information) summarised from the available data sources.

Estimated resident population

The estimated resident population (ERP) of New Zealand is an estimate of all people who usually live in New Zealand at a given date (see Standard for population terms).

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 (Statistics NZ, 2014c).

In theory, the ERP is at its most accurate immediately after the most recent census, and accuracy generally decreases over time the further away from the census. The ERP is revised when results from the next census are available.

4 Methods

Producing population estimates from administrative data

The IDI spine includes over 9 million people, far more than the New Zealand ERP of about 4.7 million in 2016. In deriving the New Zealand population at a specific point in time, some people included in the IDI spine should be excluded, such as those who died or migrated overseas before the reference date, and those who were born or migrated to New Zealand after the reference date. Therefore, we need a method to restrict the IDI spine to individuals who were resident in New Zealand at a given point in time.

The developed method uses activity recorded in administrative data to indicate an individual's presence within New Zealand at a reference point. Individuals who have died before the reference date or were identified as being overseas residents are removed.

The method used to derive the IDI-ERP for the experimental series released alongside this paper is referred to as version 2 (see Gibb et al, 2016 for a description of version 1). 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 12 months 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 these ages there is no additional requirement of activity in the previous 12 months.

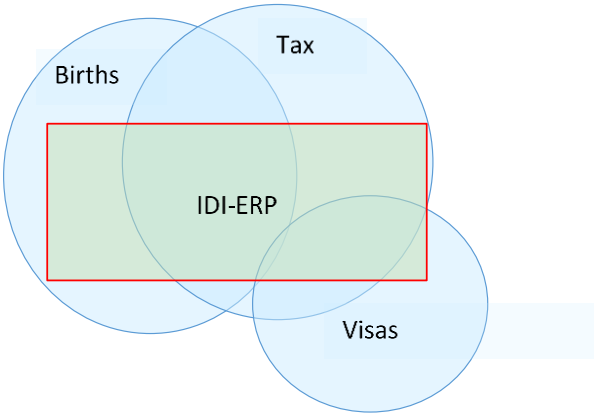
Exclusion: remove those who have left the population.

- 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. 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).

Figure 2 shows a simple diagram of the IDI-ERP as a subset of the IDI spine.

Figure 2

IDI-ERP as a subset of the IDI spine



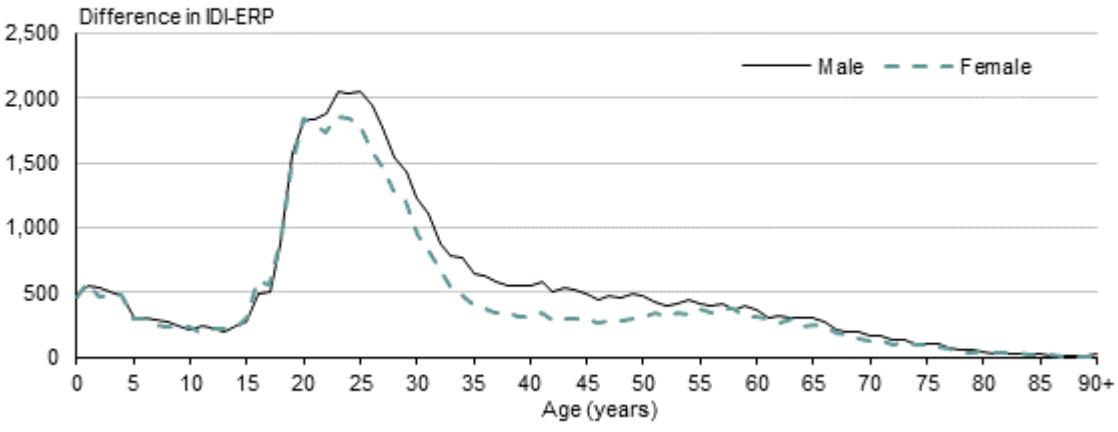
Improvement to migrant definition

The method used in version 1 of the IDI-ERP is described in Gibb et al (2016). The most significant change for version 2, as used in the experimental series, is the adjustment of the cut-off for classifying overseas residents. We refer to these people removed from the IDI-ERP as non-resident exclusions. Version 1 required an individual to be overseas for 10 of the 12 months spanning the reference date to be excluded from the IDI-ERP. Anyone with activity in administrative sources who spent two or more months in New Zealand was retained. For version 2, this cut-off was reduced to six months, meaning anyone spending six or more months overseas will be removed.

This change in the migration cut-off reduces the version 2 IDI-ERP population by almost 90,000 compared with version 1. About 55 percent of these people are male, with a peak between ages 19 and 31 years (figure 3).

Figure 3

Difference in non-resident exclusions between IDI-ERP V1 and V2
By age and sex, at 30 June 2013



Source: Statistics New Zealand

This aggregate change, however, does not provide any evidence about the true resident status of the additional individuals being removed. It is important that we understand whether the individuals affected should in fact be excluded from the resident population.

To determine true resident status, we used an alternative migration series being developed within Statistics NZ. The approach uses the actual travel history of an individual after they enter or leave the country. A person is considered to be a migrant arrival if they spend 12 of the 16 months after entering the country within New Zealand. Conversely, they are a migrant departure if they spend 12 of the 16 months after leaving the country outside New Zealand. This is referred to as the 12/16 rule.

This method is currently in development and the time series is limited to 2007–12, so results could not be used to produce the IDI-ERP. However, the method is considered to provide a good measure of an individual's actual residence status at a given point in time. We used the 12/16 rule as a 'gold standard' for measuring true resident status.

Comparisons with the 12/16 rule allowed us to estimate the total level of undercoverage (people incorrectly excluded from the IDI-ERP) and overcoverage (people incorrectly included in the IDI-ERP) that occur due to the migration rules applied in deriving the IDI-ERP. Table 1 summarises the coverage errors for the methods used in version 1 and version 2, respectively. Compared with the 12/16 rule, the version 1 method using a 10-month cut-off resulted in the incorrect inclusion of 100,000 non-New Zealand residents and excluded a small number of New Zealand residents. Version 2 using a six-month cut-off greatly reduced the number of overseas resident inclusions, but also increased the number of incorrect New Zealand-resident exclusions. Both total and net coverage errors were reduced in the version 2 method, indicating that using a six-month cut-off for removing overseas residents improves the quality of the resulting IDI-ERP.

Table 1

Estimated coverage errors caused by migrant definitions, compared with 12/16 rule, at 30 June 2012

Definition of non-residents	Incorrect non-resident inclusions	Incorrect resident exclusions	Total coverage error	Net coverage error
V1 – 10 of 12 months overseas	102,600	1,500	104,100	101,100
V2 – 6 of 12 months overseas	25,200	16,800	42,000	8,400

We also tested variations to these rules with different cut-offs and different reference windows (rather than the 12 months currently used). We considered the version 2 method to be the best trade-off between the delay required to produce the estimates and the size of the resulting coverage errors.

We have also made additional minor changes from version 1 to improve the consistency of the method:

- inclusion of visa approvals for ages under five
- inclusion of individuals with negative tax income
- use of actual dates for the extraction of tertiary enrolments, rather than calendar-year information.

These three adjustments had only a small impact on the final IDI-ERP; for 2013, the total population increased by 1,200, with slight increases to ages 0–4 and 35+, and decreases for ages 15–34.

Comparing the IDI-ERP with the ERP and the quality standards

We compared the experimental estimates of the New Zealand population produced using administrative data with the official ERP at the aggregate level. These comparisons provide an indication of net undercoverage or net overcoverage. They are aggregate comparisons, so we cannot make any conclusions about the gross errors. Overall similarities may conceal individual sections of undercoverage and overcoverage.

Census Transformation has 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.

This paper makes use of the national-level quality standards for an administrative census model. The standards state that the total population estimate should be within 0.5 percent of the true population. All national population estimates by five-year age group and sex should be within 5 percent of the true population, and 90 percent of these estimates should be within 1.5 percent. For this paper, we extended these standards to single year of age by sex level; these have the same requirements as the five-year age group standards.

For the purposes of this paper, the ERP in census years (and specifically 2013) is assumed to represent the true population. We only applied the quality standards for 2013 because there is likely to be more uncertainty in the ERP itself outside census years.

5 Results

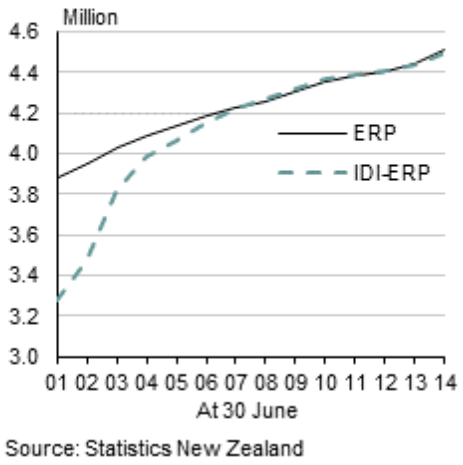
Comparing the IDI-ERP with the ERP at national level

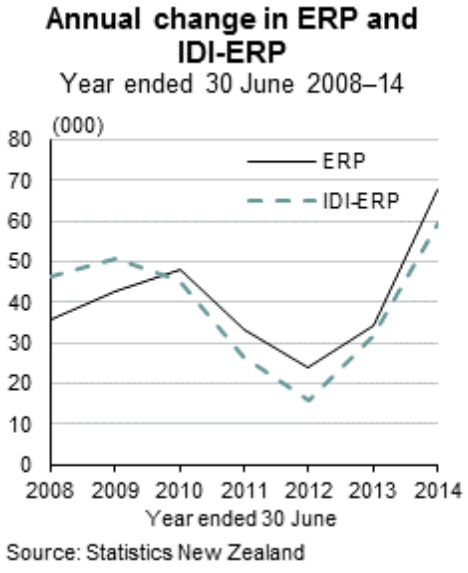
The IDI-ERP has been produced at 30 June for the years 2001–14. We evaluated these estimates by comparing them with the published ERP. Consistent estimates for 2015 could not be produced at the time of this release because not all of the administrative data were available.

Figure 4a compares the total New Zealand IDI-ERP with the ERP for 2001–14. From 2001 to 2006, the IDI-ERP is lower than the ERP, although it gets noticeably closer each year. The difference is primarily in ages 5–18, caused by full coverage of school enrolments not being available in the IDI until 2007. Due to these differences the experimental series has been restricted to estimates from 2007 onwards only. For that period, the IDI-ERP is very similar to the ERP, ranging from 9,800 (0.2 percent) lower to 18,300 (0.4 percent) higher. All eight years are within the 0.5 percent specified by the quality standards.

Figure 4b shows annual population change for the years ended 30 June 2008–14. Relative to the ERP, the IDI-ERP slightly overestimates population change for 2008 and 2009, and underestimates population change from 2010 onwards. The differences are relatively small, however, further highlighting consistency between the two measures.

Figures 4a and 4b

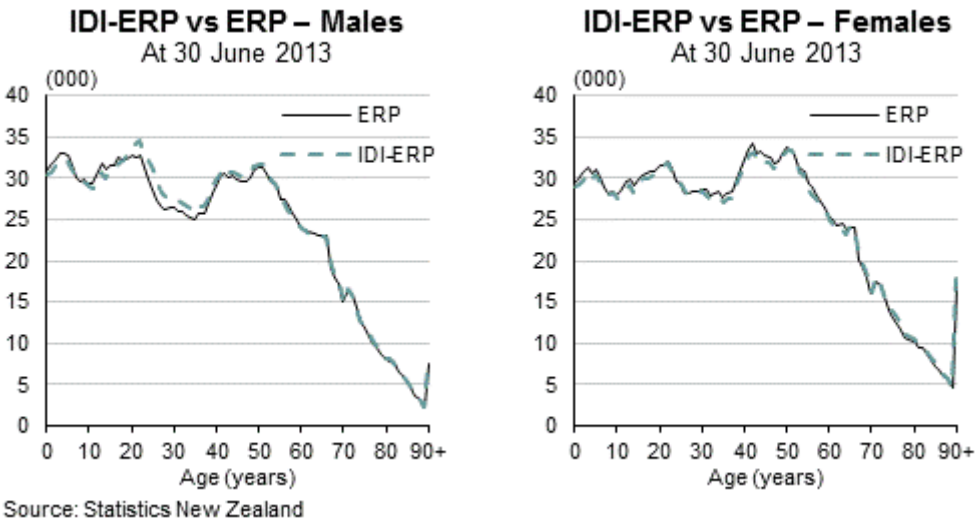




Comparisons at 30 June 2013

We have the highest level of confidence in the accuracy of the ERP in census years. The IDI-ERP at 30 June 2013 is 4,440,200, just 1,900 (less than 0.1 percent) lower than the respective ERP. Figure 5 shows the two measures have similar age distributions, although there is evidence of coverage errors at some ages. In particular, the number of males aged around 20–40 is higher in the IDI-ERP. There also appears to be some undercoverage for children and females aged 25–65.

Figures 5a and 5b



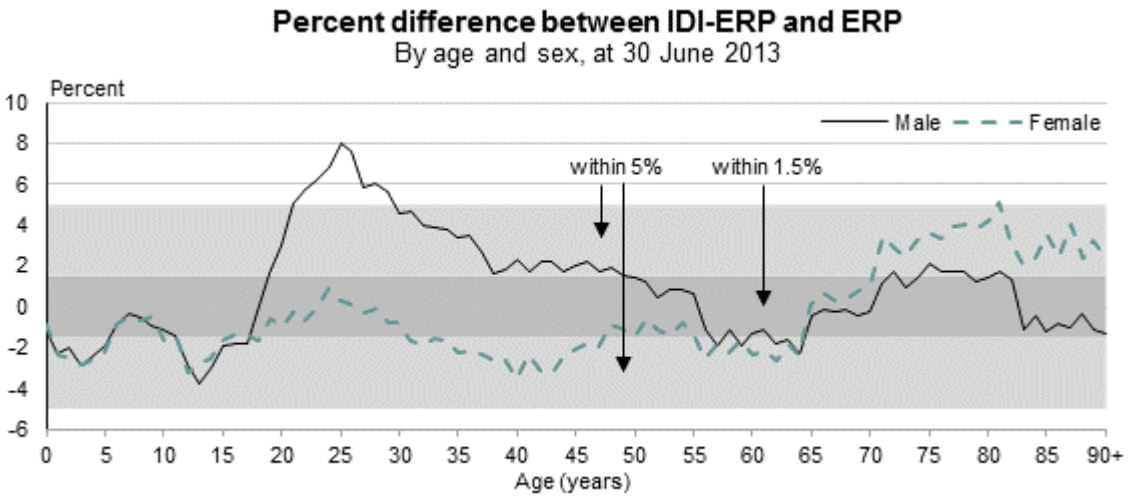
The IDI-ERP appears broadly similar to the official ERP, but we want to examine any differences more closely. Figure 6 shows the relative difference between the IDI-ERP and the ERP by single year of age and sex. Coverage differs noticeably between males and females and across ages.

Specifically, when compared with the ERP, IDI-ERP showed the following:

- undercoverage for children, particularly at ages 10–18
- clear separation between males and females for ages 19–55 – males have considerable overcoverage, peaking at 8.0 percent at age 25; females exhibit mostly undercoverage, as high as 3.5 percent at age 40
- coverage rates for the two sexes converge from 56–64, with undercoverage of about 2 percent.
- coverage increases from age 65 onwards, with females having higher rates of overcoverage than males.

The quality standards for national population estimates by single year of age are also included in figure 6 (represented by two grey grids). Of the estimates, 90 percent should be within 1.5 percent (dark grid) and all should be within 5 percent (lighter grid). Almost all the estimates were within the wider 5 percent limit, except for males aged 21–29 and females aged 81. However, less than half (40 percent) were within 1.5 percent. While final estimates will be adjusted using a coverage survey and statistical model, we would like to see whether improvements can be made to bring the IDI-ERP closer to the required quality standards.

Figure 6



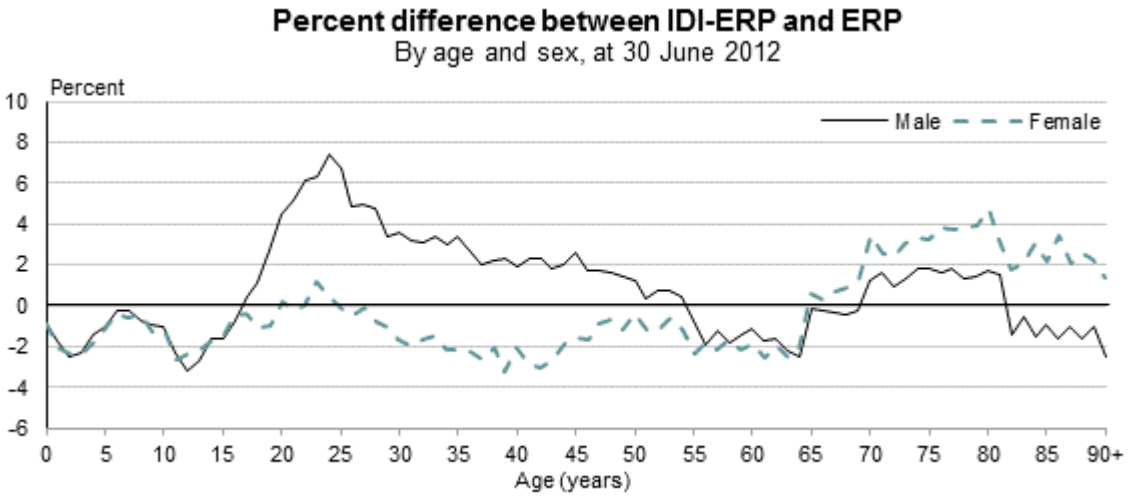
Source: Statistics New Zealand

Comparisons at 30 June 2012 and 30 June 2014

We also compared the IDI-ERP and ERP for 2012 and 2014. Results showed both years have similar patterns to 2013 (see figures 7 and 8).

For 2012, the total IDI-ERP is just 400 higher than the ERP. Males aged 25–34 are slightly closer to the ERP than in 2013, but there is still overcoverage in this age group.

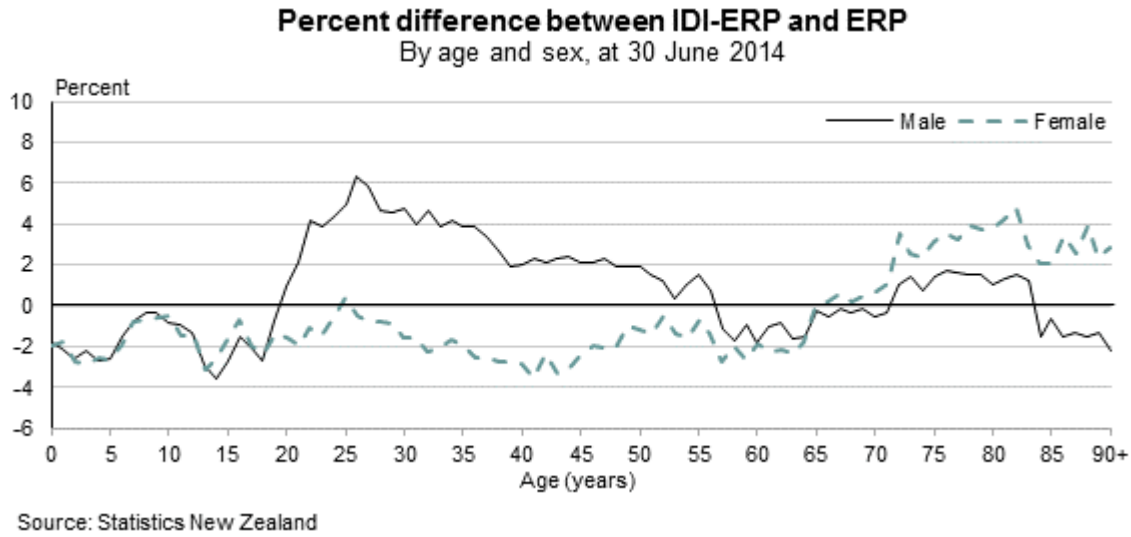
Figure 7



Source: Statistics New Zealand

Results are also very similar for 2014. The overall IDI-ERP is 9,800 (0.2 percent) lower than the ERP. As in 2012, there appears to be less overcoverage for the young adult males compared with 2013. There is evidence of slightly more undercoverage at age 0 than in either 2012 or 2013, likely caused by births that have not yet been registered.

Figure 8



Source: Statistics New Zealand

The similarity in coverage patterns across 2012–14 suggests that any underlying coverage errors are broadly consistent over time. This is likely to help in the treatment of any identified causes. If we can be confident that similar factors are contributing to undercoverage or overcoverage at each point in time, it will enable us to apply consistent adjustments.

Some of the observed coverage patterns do move as the population ages. For example, the peak level of overcoverage was observed for males aged 24 in 2012, aged 25 in 2013, and aged 26 in 2014. This indicates there could be additional factors relating to specific cohorts. It is also possible that this could highlight some level of overcoverage or undercoverage in the ERP itself.

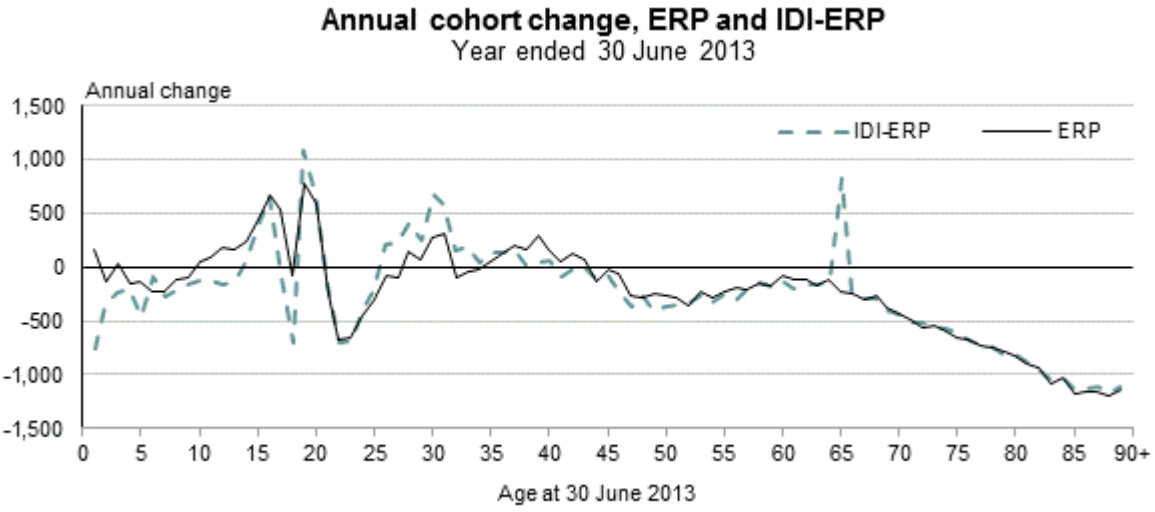
Annual change

To better understand the consistency of the IDI-ERP over time, we compared the annual population change with that for the ERP. Figure 9 shows how many people entered or left each population, by age, in the year ended 30 June 2013.

The trend is generally close between the two measures. One exception is age 65 where the increase in IDI-ERP is about 1,000 larger than that for the ERP. It is the only age above 40 with a population increase observed in the IDI-ERP. At 65, most people become eligible for superannuation, which should result in activity in Inland Revenue data. It seems likely that the increase is due to people being excluded from the IDI-ERP at age 64 (and younger) because they were not active in any administrative sources, rather than genuine population growth.

Additional differences across other ages showed the IDI-ERP generally had a larger decrease at ages 0–18 and a larger increase at ages 25–35. These are mainly caused by the different classifications of migrants. Unlike the increase at age 65, these differences are more volatile and tend to be less consistent over time as migration patterns change.

Figure 9



Source: Statistics New Zealand

Components of population change

It is also possible to break the total population change into individual components. These reflect the specific ways an individual can enter or leave the population. Table 2 compares the components contributing to population change in the ERP and the IDI-ERP for the years ended 30 June 2012 and 2013. For the ERP, the components of change are births, deaths, and migration (both arrivals and departures).

Births and deaths in the IDI-ERP and ERP are very similar as both make use of DIA registrations.

Compared with the ERP, the IDI-ERP has higher levels of both migrant arrivals and departures. These differences occur mostly due to the definitions used in each series. In the ERP, permanent and long-term (PLT) migrants are largely based on self-reported intentions from passenger cards. The IDI-ERP removes any individuals spending six or more of the 12 months spanning the reference date outside New Zealand. This misclassifies those who spend six or more months outside their country of residence and will tend to overestimate the number of people whose residence status changes from year to year.

The final two components are specific to the IDI-ERP. ‘In-activity’ represents individuals who entered the IDI-ERP over the year, but with no evidence of either a birth or migrant arrival. ‘Out-activity’ represents individuals who exited the IDI-ERP over the year, but with no evidence of a death or migrant departure. Instead, they are people who were potentially usual residents at both points, but had activity in administrative data in only one of the two years. They therefore represent likely undercoverage in the year they were not selected.

Table 2

Components of population change – ERP and IDI-ERP At 30 June 2012 and 2013				
Component of population change	**Year ended 30 June 2012**		**Year ended 30 June 2013**	
	ERP	IDI-ERP	ERP	IDI-ERP
Births[1]	61,000	61,900	59,900	60,800
Deaths[1]	29,800	29,200	30,000	29,100
Migrant arrivals	84,400	97,200	88,200	103,800
Migrant departures	87,600	125,900	80,300	115,700
In-activity[2]	…	54,000	…	52,700
Out-activity[3]	…	42,200	…	40,700

1. ERP figures based on date of registration, IDI-ERP based on date of occurrence.
2. Entered IDI-ERP, but with no evidence of a birth or migrant arrival during year.
3. Exited IDI-ERP, but with no evidence of a death or migrant departure during year.

Symbol: … not applicable

Coverage by administrative source

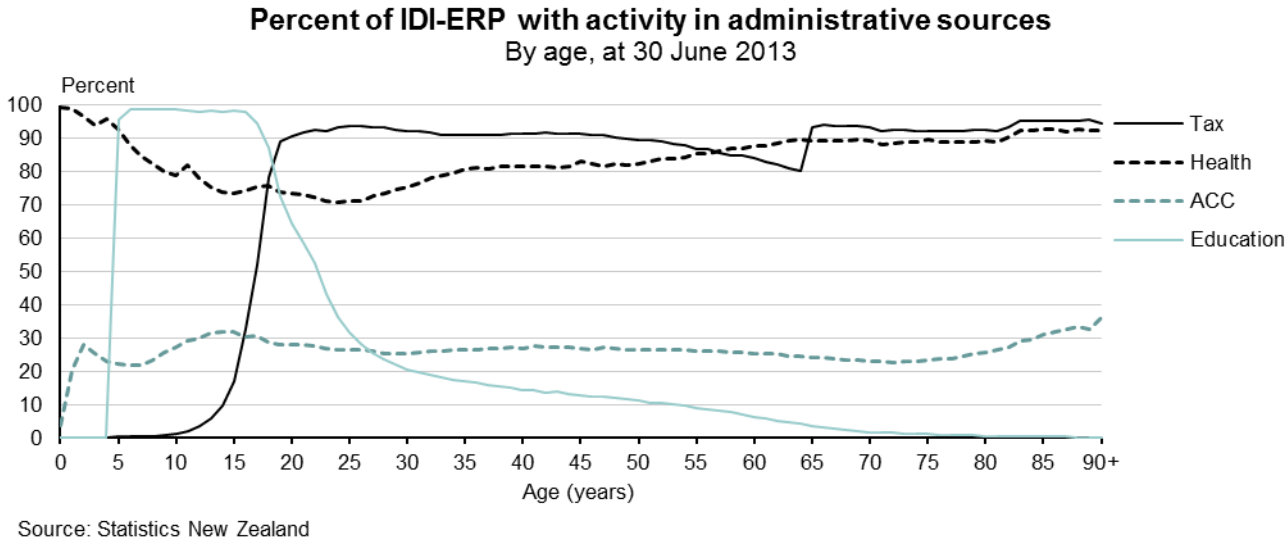
Each contributing administrative source has a different coverage pattern across ages, as seen in figure 10.

In particular:

- education has very high coverage (above 98 percent) for ages 6–16
- health has very high coverage at the youngest ages, dropping to about 70 percent at age 25; coverage then increases steadily to 90 percent at around age 65
- tax is consistently about 90 percent coverage from age 20 onwards, only dipping slightly in the ages leading up to 65
- ACC has consistent coverage at about 25 percent.

Figure 10 does not show those aged under five selected for the IDI-ERP due to a birth record or visa approval. Combined, the two sources record activity for almost everyone in the IDI-ERP aged 0–4 and nobody five or older.

Figure 10



Source: Statistics New Zealand

Reasons for differences between ERP and IDI-ERP

To continue improving the IDI-ERP we need to understand the underlying reasons for observed differences between the IDI-ERP and the ERP. Investigations have been carried out examining possible sources of these coverage errors.

In this section we summarise four of the main error sources:

1. duplicate records in the IDI spine
2. false positive links between the IDI spine and deaths
3. definition of migrants
4. individuals with no activity in administrative sources within 12-month period.

These four causes represent the major contributing factors we have been able to identify and at least partly quantify. This list is not exhaustive and other sources of coverage error exist. For example, some usual residents may not be included in the IDI spine, which will always cause undercoverage in the IDI-ERP.

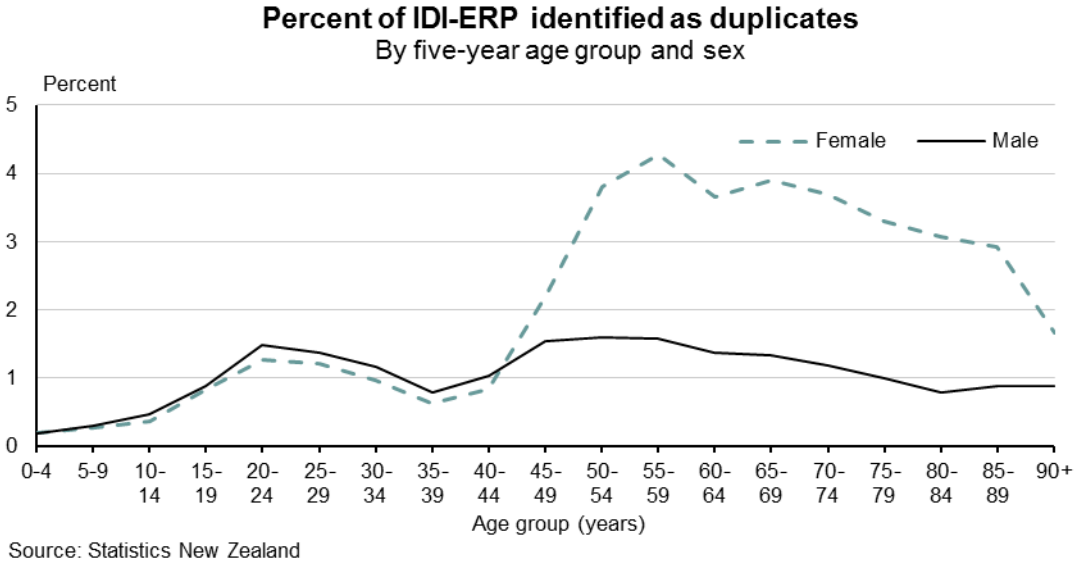
Duplicates records in the IDI spine

Some individuals may appear in the IDI spine more than once. These are referred to as duplicates and can occur in two ways. First, duplicates may be present within any of the contributing administrative data sources (IR, births, or visas). Second, links between the same individual in multiple sources could be missed. Both result in overcoverage in the IDI spine, and potentially in the IDI-ERP.

We identified up to 67,000 duplicates in the IDI-ERP (1.5 percent of the population). The majority of duplicates identified were either duplicates within IR data, or missed links between IR and either of the other two sources. Figure 11 shows that there are relatively few duplicates through to age 44. At about age 45, there is a spike in the percentage of females identified as duplicates.

Given the clear age differences, it is likely that this is at least partially a systematic issue resulting from legacy IR data, and not an ongoing problem. If this hypothesis is correct, the IR duplicates could potentially be removed as a one-off exercise. This would create a permanent solution for the IDI, and remove the largest source of duplicate records in the IDI-ERP and the IDI spine more generally.

Figure 11



False positive links between the IDI spine and deaths

A number of individuals were identified as having activity in administrative sources, or being included in the 2013 Census, well after their date of death in the IDI. This inconsistency indicates a strong likelihood of a false positive link, either between deaths and the IDI spine, or between the IDI spine and the source of activity.

For the IDI-ERP at 30 June 2013, about 5,000 individuals had activity in administrative sources but also had a death record before 2012. With no additional information, we assumed the death record to take precedence and excluded these individuals from the IDI-ERP. Some of these 5,000 will be false links between the IDI spine and deaths. Removing these individuals is currently causing some IDI-ERP undercoverage. Further work is planned to assess the quality of these particular links.

Definition of migrants

Improvement to migrant definition describes the improved method for removing non-residents for the experimental series. However, some individuals will still be incorrectly classified, leading to instances of both undercoverage and overcoverage.

We estimated that at 30 June 2012, undercoverage from the incorrect exclusion of New Zealand residents was around 17,000; overcoverage from the incorrect inclusion of non-New Zealand residents was around 25,000 (see table 1). These coverage errors appear to roughly offset at the national level, with similar distributions by age and sex.

Certain groups were identified as contributing to overcoverage. For example, we found evidence of approximately 1,000 individuals who entered New Zealand on seasonal work visas incorrectly included in the IDI-ERP. For this first experimental series release, we made no specific changes based on these findings. Any improvement to the resulting population was outweighed by the added complexity. This will be reassessed for ongoing work.

In the longer term we hope to make use of an alternative method for classifying migrants currently in development. Any method is unlikely to completely remove classification errors, particularly given the need for timely estimates. However, we hope that resulting

coverage errors will be smaller than for the current methods used in the IDI-ERP. We will continue to monitor these developments and assess whether any improvements can be made to our rules.

Individuals with no activity in administrative sources within 12-period

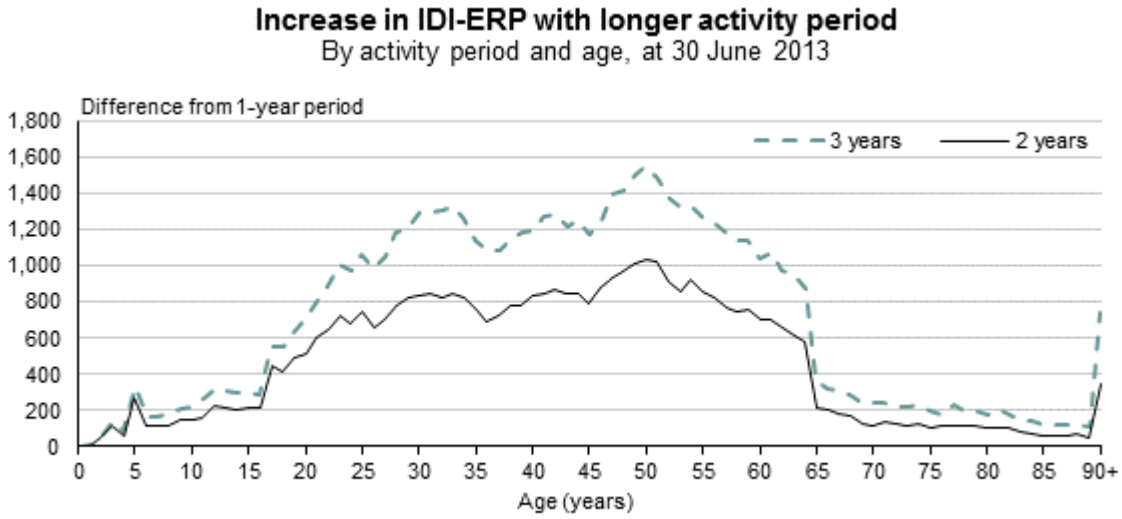
One source of undercoverage is individuals who are usual residents, but have no activity in administrative sources in the 12 months before the reference date. Gibb et al (2016) described this group using the 2013 Census. They found that the percentage of people in the IDI and Census, but not included in the IDI-ERP, increased through the adult ages and peaked at the 60–64 age group.

Without any activity-based restriction, the IDI-ERP would be far larger than the resident population of New Zealand. However, relying on a one-year activity period leads to undercoverage, as shown in figure 9 and table 2, which suggests that a group of people are being incorrectly excluded from the IDI-ERP because they were not active in administrative data.

We identified prisoners as one group contributing to this undercoverage. About 1,000 people were recorded as being in prison at 30 June 2013, but had no activity in the administrative sources used to select the IDI-ERP. Corrections data could be used as an additional indicator of activity in future versions.

Another possible way to reduce this source of undercoverage is to widen the activity period from the current 12 months. The effect of using a two- or three-year period is shown in figure 12. At the youngest and oldest ages there is minimal impact due to high coverage from school enrolments and tax data, respectively. The remaining ages had a larger effect, with average increases of about 800 per year of age using two years of activity, and 1,200 per year of age using three years of activity. Implementing a two-year activity window increased the total population by 43,000 (1.0 percent); a three-year activity window gave an increase of 65,000 (1.5 percent). Currently, this apparent undercoverage is being offset by other factors causing overcoverage.

Figure 12



An increased activity period could potentially be used with other improvements to create a new version of the IDI-ERP. However, there are complicated interactions between the various sources of coverage error that we do not yet fully understand. Widening the activity period could also have side effects, such as increasing overcoverage resulting from linkage errors. Adjusting the current rules to make them more selective is also a possibility. For example, we could use different activity periods for different administrative sources, or different age groups.

6 Discussion

The national population estimates produced from administrative data discussed in this paper represent an improvement on previous estimates. This is the first time we have released outputs alongside our research paper. We also extended the analysis from a single point in time to eight years of data.

For 2007–14, the IDI-ERP generally agrees closely with the ERP, particularly at the total level, with all eight years within the required quality standards. Coverage errors differ by age and sex – the highest overcoverage is for males aged 20–29. Only 40 percent of the single year of age by sex estimates are within the acceptable quality standards for an administrative-based census. This indicates that further work is required to improve the quality of the IDI-ERP.

We identified a number of factors contributing to the discrepancies between the IDI-ERP and ERP. Some usual residents are being excluded from the IDI-ERP due to inactivity in administrative sources within the 12-month period. Duplicates in the spine appear to be causing overcoverage in the IDI-ERP. Other linkage errors within the IDI and the misclassification of migrants are both leading to pockets of undercoverage and overcoverage. These errors appear to have a fairly consistent impact on the IDI-ERP over time.

We do not yet fully understand the potential interactions between all these factors. Because we have not yet been able to implement all the suggested changes, we do not know precisely how accurate these administrative estimates could be. The trade-offs between reducing undercoverage at the expense of overcoverage (or vice versa) require further investigation.

Future work

We intend to continue developing the methods discussed in this paper, with any improvements incorporated into future experimental series releases. We identified specific areas for improvement, including:

1. exploring strategies for resolving duplicates within the IDI spine
2. further understanding potential false-positive linkages between deaths and the IDI spine
3. reconsidering the length of the activity period once we have a better understanding of the potential interactions and the needs for estimation modelling.

The 2017 release is scheduled to include more detailed population estimates, by subnational area and by level 1 ethnic group. For both subpopulations, further investigation is required to understand how the coverage errors at the national level translate to these groups.

We anticipate that even with further development, administrative data alone will not be sufficiently accurate. We are continuing to progress work developing a coverage survey and statistical models that will adjust for any remaining discrepancies. These will allow us to improve the estimates for particular groups with coverage errors that cannot be fully reduced using simple rules.

The current method for deriving the IDI-ERP will be incorporated as a table in the September 2016 IDI refresh, allowing users to easily identify the New Zealand resident population for each year.

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.

Bycroft, C, Reid, G, McNally, J, & Gleisner, F (2016). Identifying Māori populations using administrative data: A comparison with the census. 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, SJ, & 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.

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

O'Byrne, E, Bycroft, C, & Gibb, S (2014). An initial investigation into the potential for administrative data to provide census long-form information: Census Transformation programme. Retrieved from www.stats.govt.nz.

Reid, G, Bycroft, C, & Gleisner, F (2016). Comparison of ethnicity information in administrative data and the census. Retrieved from www.stats.govt.nz.

Shrosbree, E (2015). Comparing education and training information in administrative data sources and census. Retrieved from www.stats.govt.nz.

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

Statistics New Zealand (2014a). An overview of progress on the potential use of administrative data for census information in New Zealand: Census Transformation programme. Retrieved from www.stats.govt.nz.

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

Statistics New Zealand (2014c). Estimated resident population 2013: Data sources and methods. Retrieved from www.stats.govt.nz.