Identifying the New Zealand resident population in the Integrated Data Infrastructure (IDI)

Census Transformation

Sheree Gibb, Christine Bycroft, Nathaniel Matheson-Dunning

New Zealand Government

Disclaimer
This paper represents the views of the author. It does not necessarily represent the views of Statistics NZ and does not imply commitment by Statistics NZ to adopt any findings, methodologies, or recommendations. Any data analysis was carried out under the security and confidentiality provisions of the Statistics Act 1975.

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.

Citation

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.

ISBN 978-0-908350-33-9 (online)

Published in April 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 4610
www.stats.govt.nz

Contents

List of tables and figures

List of tables

List of figures

1 Background

Census Transformation in New Zealand

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

- a focus in the short-to-medium term on modernising the current census model and making it more efficient
- 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 New Zealand, 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.

Continuing to meet critical information needs must underpin decisions on the future of census. Investigations into the long-term direction for census are focused on developing an understanding of future census information requirements, and the ability of administrative sources to meet those requirements.

See Census Transformation in New Zealand for more information.

About this paper

The fundamental reason for having a census is to provide population statistics that describe the size, structure, and geographic distribution of the population.

This paper describes a method for determining who is a resident in New Zealand at a given point in time using the linked administrative data sources held in Statistics New Zealand's Integrated Data Infrastructure (IDI).

We compare the resulting national population by age and sex with the official estimated resident population figures, and assess the accuracy against quality standards developed for Census Transformation.

2 Introduction

The New Zealand Census of Population and Dwellings is currently held every five years as legislated by the Statistics Act 1975. The primary role of the census is to provide population and dwelling counts for New Zealand, for regions and territorial authorities, and smaller geographic areas such as area units and meshblocks. The census is also the only comprehensive source of information about the social and economic characteristics of local communities and small population groups (eg Māori and iwi, youth, new migrants).

Population statistics are the most important requirement for a census to provide. Ensuring these statistics are fit for purpose is critical when considering future census models.

At the same time, achieving high quality in the population counts is a major driver of the costs of a census. The feasibility of obtaining sufficiently good population estimates, and this trade-off between quality and cost for population statistics will be key determinants in decisions about the long-term direction of census.

McNally and Bycroft (2015) developed quality standards for population estimates that reflect customer requirements for accuracy. These quality standards were designed to assess the statistical quality of population estimates produced by alternative approaches to census-taking, such as administrative-based models that use linked administrative data to produce population statistics.

Nordic countries and others that already produce their census information from administrative sources base population statistics on national population registers. The population registers serve administrative purposes and are designed to include everyone living in the country. There is typically a unique identifier assigned to each person which is widely used. These 'ready-made' administrative systems do not exist in New Zealand. Bycroft 2015 contrasts the administrative data available in New Zealand with that typically found in countries that have moved to register-based censuses.

Gibb and Shrosbree (2014) developed a method for constructing a statistical population list from the linked administrative data sources available in Statistics NZ's Integrated Data Infrastructure (IDI) at April 2013. Population estimates were then derived and compared with official estimated resident population figures for 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 the two years to May 2015 overcomes some of the earlier limitations. Birth and death registrations and health data are now available. An extended spine structure in the IDI provides better coverage of the population not paying tax, especially children. Health data provides a much broader source of activity information than before – the previous method had relied solely on payment of tax and enrolment in education.

Gibb and Shrosbree's investigation was undertaken before the 2013 Census, and the administrative data were only available up to 2010. We are now able to compare results against the 2013 Census and the official 2013 estimated resident population, and administrative sources have been updated to include information at least until 2013.

Aims and scope

The major aim of this paper is to examine the accuracy of national population estimates, produced from linked administrative data available in the IDI. We update the methods applied in 2014 in light of the recent changes to the IDI. The accuracy of the population estimates will be evaluated against the official 2013 estimated resident population using the quality standards developed by McNally and Bycroft.

The IDI is continually being expanded and developed. The analyses in this paper are based on the IDI as it stood in May 2015.

The scope of this paper is population estimates at the national level by age and sex. Statistics NZ also produces official population estimates for subnational geographies and for ethnicity. Subnational estimates are dependent firstly on the quality of national estimates, and secondly on information about where people live. The quality of geographic information in the IDI is examined in Gibb, 2015. The quality of ethnicity information in the IDI is examined in Reid et al, 2016.

The potential for administrative data sources to produce other types of census information (for example, information about education, income, families, households, or housing) is discussed in other work (O'Byrne et al, 2014; Shrosbree, 2015; Suei, in press).

This work is not intended to provide a final evaluation of the feasibility of using linked administrative data sources to produce population statistics in the absence of a full-enumeration census. Rather, we will provide information about the use of linked data sources to identify populations that will guide further work.

The remainder of this paper is organised as follows. Section 3 describes the data sources used: the census and official population estimates, the IDI, and the linked Census-IDI dataset. Section 4 describes the method used to construct an 'administrative' resident population from the IDI, and how we assess its accuracy. Section 5 provides results with comparisons at both the aggregate level and individual level. The paper concludes with a short discussion.

3 Data sources

The New Zealand Census of Population and Dwellings

The New Zealand Census of Population and Dwellings is the official count of people and dwellings in New Zealand. It provides a snapshot of New Zealand at a point in time, and measures social and economic change. The latest census was held in March 2013.

The census aims to count everyone who is in New Zealand on census night. Overseas visitors are included in the census, while New Zealand residents who are not in New Zealand on census night are not included.

Not all of those counted in the census returned census forms. The census count includes 4.8 percent (203,052) substitute records (Statistics NZ, 2014b). A substitute is a census record that is created where there is sufficient evidence received during the collection process that a person exists or a dwelling was occupied, but we obtained no corresponding form. As such, they form part of census non-response.

Some people are missed altogether or counted more than once in the census. Coverage in the census is measured by the Post-Enumeration Survey (Statistics NZ, 2014b). Net census undercount for the 2013 Census was estimated at 2.4 percent. Younger adults aged 15–29 years had a higher percentage undercount (4.8 percent) than other age groups.

The estimated resident population (ERP)

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

The estimated resident population of New Zealand is derived by adjusting the census usually resident population count for net census undercount (as estimated by the PES) and the estimated number of residents temporarily overseas on census night. To obtain the estimated resident population at a given date after census night, updates are made for natural increase (births less deaths) and net migration (arrivals less departures) between census night and the given date. The official estimated resident population (ERP) series provides the best measure of who is living in New Zealand at a given date.

The ERP is at its most accurate immediately after the most recent census, and accuracy generally decreases over time the further we move away from the census. For this reason we use the official ERP in the base census year at 30 June 2013 as the comparison for population estimates constructed from the linked administrative data in the Integrated Data Infrastructure (IDI).

The Integrated Data Infrastructure (IDI)

Statistics NZ has 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 outputs and to allow Statistics NZ staff and external researchers to conduct policy evaluation and research on people’s transitions and outcomes. The IDI contains 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 current information at stats.govt.nz/idi). This section describes the structure and content of the IDI as at May 2015.

The basic structure of the IDI is shown in figure 1. The structure of the IDI can be described as a central 'spine' to which a series of data collections are linked.

The spine forms the conceptual centre of the IDI. Broadly, the target population for the spine is all individuals who have ever been residents of New Zealand. The spine aims to include each individual only once.

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

- a list of all IRD numbers that have been issued by Inland Revenue
- 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 ensure that people present in any two data sources are included only once in the spine.

Other data sources are linked to the IDI spine (see Statistics NZ 2014c 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 the Ministry of Education; benefit dynamics data from the Ministry of Social Development; student loans and allowances data from several sources; migration movements data from the Ministry of Business, Innovation and Employment; and the Household Labour Force Survey and New Zealand Income Survey from Statistics NZ.

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

Using the IDI to determine who is living in New Zealand

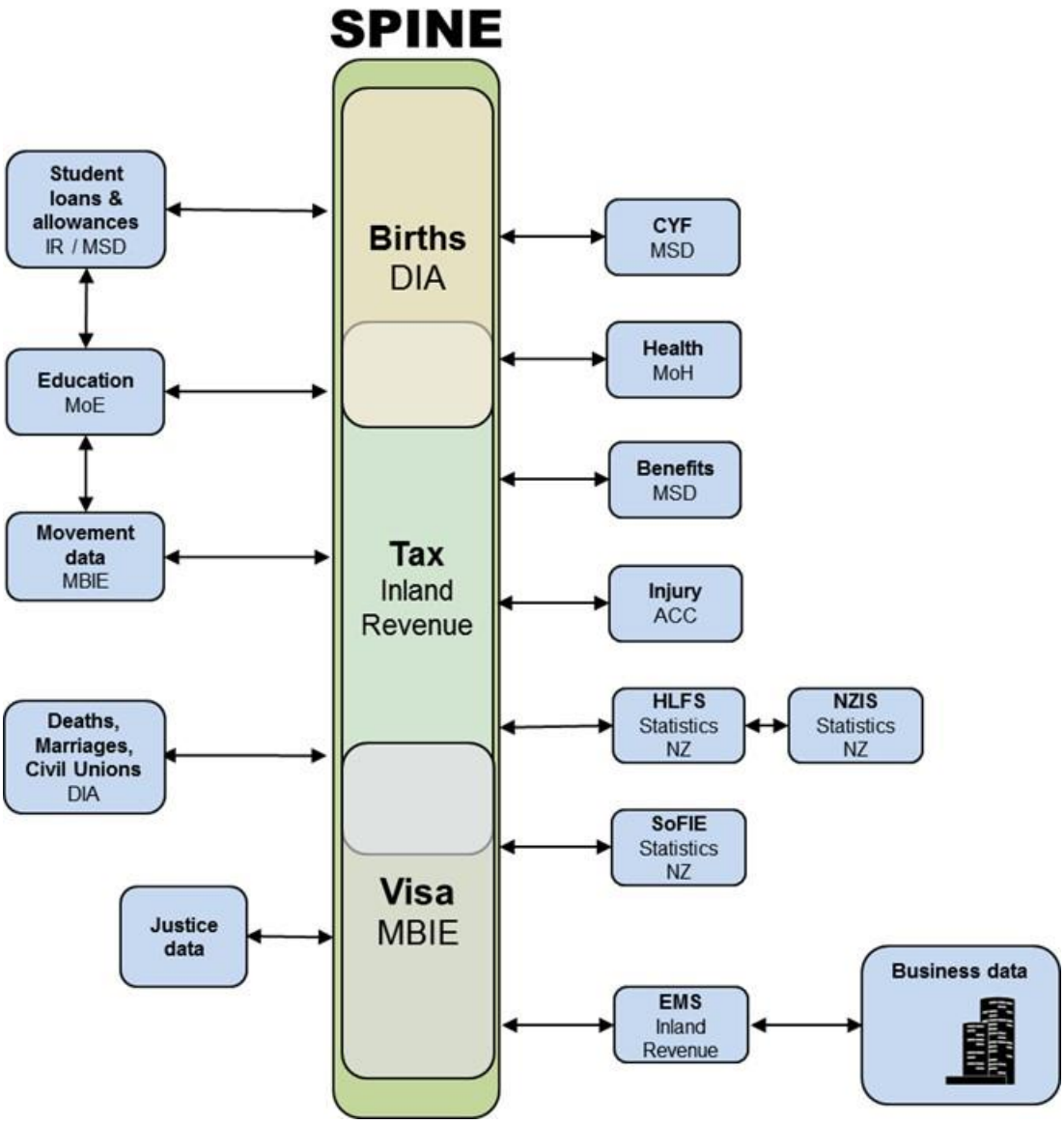
The IDI spine forms our starting point for determining who is living in New Zealand. Many of the other data sources contain information about events or activities where people interact with government, and can indicate who is present in the country at a given time.

External migration is very important in this context. The MBIE 'border movements' data in the IDI records travel journeys into and out of New Zealand. When linked to individuals in the spine, the border movements can help to indicate when people have left New Zealand on a long-term basis, and should no longer be counted among the New Zealand residents.

Deaths data is also linked to the spine, and allows us to remove those who have died from the population.

Figure 1

Structure of the Integrated Data Infrastructure in May 2015



The linked Census-IDI

The 2013 Census has been linked to the IDI. The linked Census-IDI dataset used for this study was created by the Census Transformation programme, and was created to better understand the coverage and quality of census information in the IDI. The linked data was only available to approved Statistics NZ staff working on Census Transformation.

The census was linked to the spine of the IDI in the May 2015 IDI refresh. Linking was completed in Quality Stage using probabilistic matching techniques. The variables full name, date of birth, sex, meshblock of usual residence, and country of birth were used in the linkage process.

Overall, 92.4 percent of the census usual resident population count were linked to the IDI. Of most interest here, 95.4 percent of people from responding households (ie excluding

those where substitute records were used for the entire household) were linked to the IDI. The linkage rate was better for individuals who had used electronic forms (98 percent linked) compared with paper forms (93 percent linked).

In any probabilistic linking process there will be linkage error. A false positive linking error occurs when two records are linked but they are not a true match (the linked records belong to different people). There is an estimated false positive rate of less than 1 percent in the Census-IDI dataset.

A false negative linking error occurs when two records are a true match (they belong to the same person) but they are not linked. False negative errors are more difficult to detect. The linkage rate of 95.4 percent of census respondents provides an upper bound of 4.6 percent for the false negative linkage rate.

4 Methods

Creating a usually resident population from the IDI

The IDI spine contains more than 9 million individuals – far more than the 2013 New Zealand usual resident population of approximately 4.5 million. Many individuals in the IDI spine are former usual residents of New Zealand who have since left or died. It is therefore necessary to restrict the IDI spine population to the subset of individuals who were usual residents of New Zealand at a given date.

The method used to select the resident population at a given date relies on identifying activity in New Zealand administrative systems that indicates an individual's presence in New Zealand over a period prior to the reference date. We then remove individuals who left the population by death or outmigration prior to the reference date.

In theory, we can remove migrants by observing their travel patterns. However, one problem is that border movement data is only available from 1998. Another is that determining a change in residency status through matching inward and outward travel journeys is difficult. There is currently no standard definition of what period of time out of the country determines when a person is considered **not** to be a New Zealand resident, and travel patterns of residents and visitors are complex.

The rule applied here is a somewhat pragmatic choice. It is relatively straightforward to apply and strikes a balance between retaining New Zealand residents who spend periods of time overseas, and removing genuine external migrants. Nevertheless, some short-term visitors may be incorrectly retained in the IDI resident population.

Specifically, the method used to identify the IDI resident population for a given date was as follows:

Inclusion: retain individuals whose presence is indicated by activity

- For ages five years and over, the spine population was restricted to those individuals who had activity in one of the following IDI datasets in the 12 months prior to 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)
    - Health (pharmaceutical prescriptions, GP enrolment and attendance, hospital admissions, non-admission hospital visits)
    - Education (school enrolment, tertiary enrolment or attainment).
- For ages under five years, having a record in the spine was sufficient for inclusion in the population. For these ages there was no additional requirement of activity in the previous 12 months.

Exclusion: remove those who have left the population

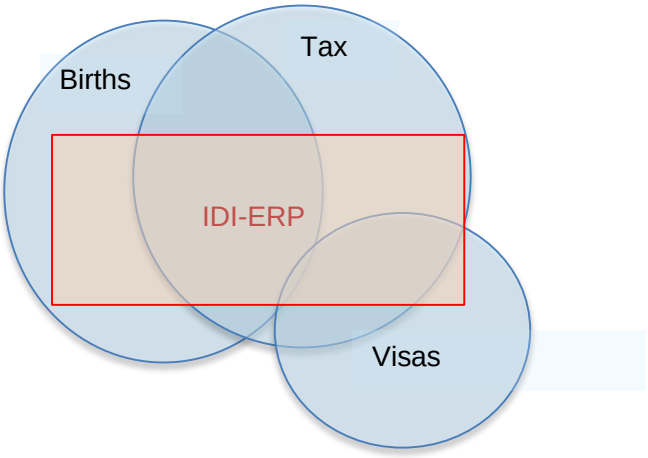
- Linked death records were used to identify individuals with a date of death prior to the reference date.
- Linked migration data were used to identify individuals who had moved overseas. Individuals were classified as having moved overseas if the total length of time spent overseas was at least 10 of the 12 months spanning the reference date (that is, the six months either side of the reference date).

The resulting resident population derived from the IDI is called the IDI-ERP.

Figure 2 shows a simple diagram (not to scale) of the administrative population derived from the IDI (the IDI-ERP) as a subset of the IDI spine.

Figure 2

The IDI-ERP shown as a subset of the IDI spine



Methods for assessing the coverage of the IDI resident population

The rules for inclusion and exclusion described above are used to create a list of the resident population from the IDI. This section describes the methods used to assess how well the rules are performing.

Aggregate comparison against the ERP and the quality standards

This administrative population derived from IDI (the IDI-ERP) can be compared to the official estimated resident population (the ERP) at an aggregate level, by age and sex. The result of this comparison will reveal net overcoverage or undercoverage.

Census Transformation has developed a set of quality standards to assess the quality of population estimates produced from alternatives to the current census model (McNally and Bycroft, 2015). The quality standards reflect a series of discussions held with core customers of population statistics. They provide a measure of the lowest acceptable accuracy for users of the estimated resident population data series.

The relevant standards for this paper are those for the national population applied to an administrative census model that produced independent estimates each year. According to the quality standards, the total national population estimate should be within 0.5 percent of the ERP. National population estimates by sex and five-year age group should all be within 5 percent of the ERP, and 90 percent of them should be within 1.5 percent of the ERP.

It should be noted that these quality standards apply to the final population estimates. The estimates presented in this paper are likely to be initial counts, which would be further improved by coverage adjustments and estimation methods. Nonetheless, they are a useful guide for evaluating the broad quality of the population counts in this paper.

Individual comparison against census

There are limitations to the aggregate comparison. Comparing populations at the aggregate level cannot reveal which individuals are missing from the population

(undercoverage) and which are erroneously included (overcoverage). Furthermore, aggregate comparisons may obscure patches of undercoverage and overcoverage that balance out to produce good net coverage.

To address these limitations of the aggregate-level comparisons, we can compare the IDI-ERP at the individual level against another list of the population – in this case, the census. While the census is not a completely accurate list of the New Zealand resident population (it contains undercoverage, overcoverage, and by design does not include residents overseas on census night), the results of an individual-level comparison against census may still reveal patterns of overcoverage and undercoverage that were not apparent in the aggregate-level analysis. This information could be used to understand and improve the IDI-ERP coverage of the resident population.

To enable a fair comparison of the IDI-ERP and census populations, we made the following adjustments:

- overseas visitors were removed from the census population
- New Zealand residents who were recorded in migration data as being overseas on census night (residents temporarily overseas) were removed from the IDI-ERP population
- babies born in March 2013 were removed from both populations (only month and year of birth were available in IDI, so it was not possible to distinguish babies born before 5 March from those born after).

The linked Census-IDI dataset is used for these individual-level comparisons.

5 Results

Results for each step of the construction of the IDI-ERP are shown in table 1. Almost half the members of the spine have no activity apparent in the 12 months prior to June 2013. Another 200,000 of those reporting some activity were removed from the population because they died or left the country. The number of deaths removed is larger than the approximately 30,000 deaths that typically occur in a calendar year. This is at least partly because some people with activity in 2012/13 are recorded as a death prior to that time. There may be genuine reasons for this apparent anomaly.

Table 1
Results for construction of the IDI-ERP

Number in IDI spine	9,074,000
Retained through activity	4,722,600
Deaths removed	40,300
Outmigration removed	149,000
Total IDI-ERP	4,533,200
ERP	4,442,100

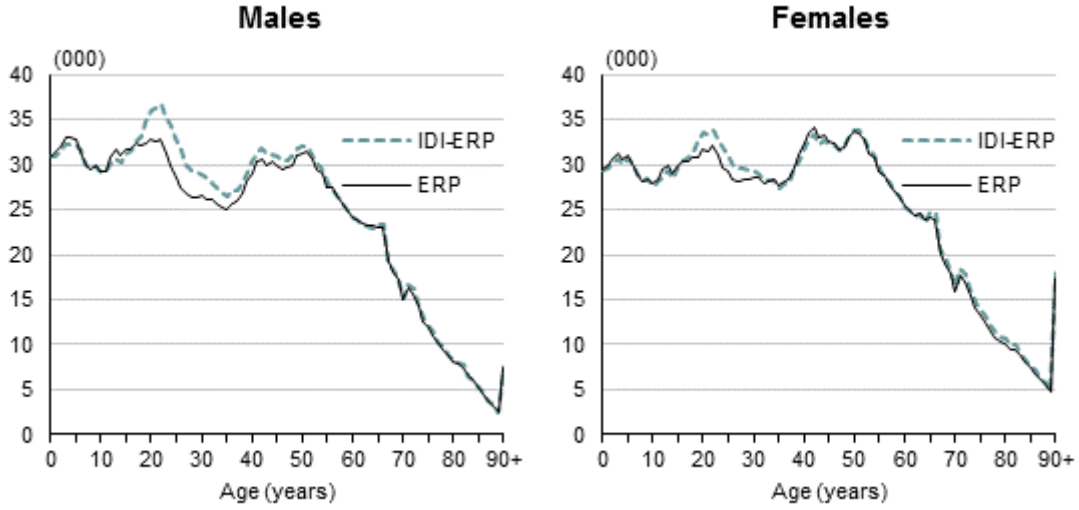
Aggregate comparison against the ERP

The total national population obtained using the activity-based method described above was 4,533,200 as at 30 June 2013. This represents 102 percent of the ERP for the same date – in other words, the IDI-ERP is 2 percent higher than the ERP. This is outside the quality standard which specifies that the total national population should be within 0.5 percent of the ERP.

The national IDI-ERP age-sex distribution pattern is largely similar to the ERP (see figure 3), suggesting that overall the approach to extracting the resident population from the administrative sources is working reasonably well.

Figure 3

National population distribution for IDI-ERP and ERP
By single year of age and sex
At 30 June 2013



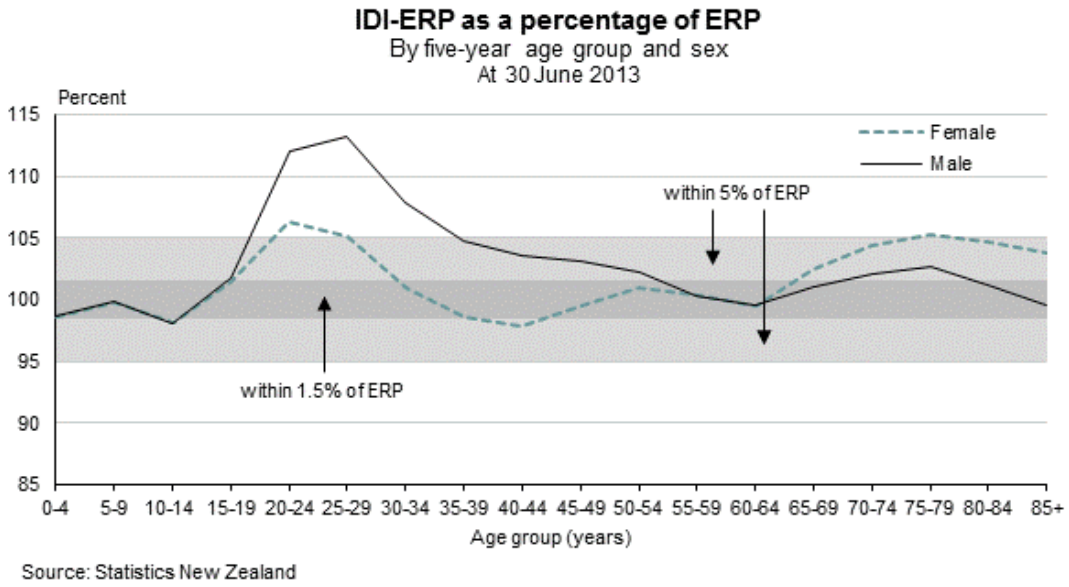
Source: Statistics New Zealand

However, this presentation may conceal important differences between the two sources. The quality standards are expressed in terms of relative difference from the official ERP. Figure 4 shows the IDI-ERP as a percentage of the ERP, by five-year age group and sex. The figure also shows the quality standards: 90 percent of the estimates should be within 1.5 percent of the ERP (the dark grey shaded area) and all should be within 5 percent (the lighter grey area).

While some parts of the population have good coverage (for example, females aged 30 to 69), others have coverage that is outside the quality standards. Overall, 44 percent of the age-sex groups were within 1.5 percent of the ERP and 83 percent were within 5 percent of the ERP.

Where coverage is outside the quality standards, this was mostly overcoverage (where the IDI-ERP count is higher than the ERP) rather than undercoverage. The figure shows that overcoverage is greatest in the early adult ages (20–34 years), particularly for males. Possible reasons for this overcoverage are considered in the discussion section.

Figure 4



Source: Statistics New Zealand

Individual-level comparisons against census

The net coverage estimates from the aggregate comparison in figure 3 may conceal areas of overcoverage and undercoverage at the individual level. This section presents findings from the individual-level analysis of coverage in the IDI-ERP using the linked Census-IDI dataset.

Figure 5 shows the overlap between the census, IDI, and IDI-ERP populations. The figure is to scale and the size of the areas represents the relative size of the populations (Micallef & Rodgers, 2014).

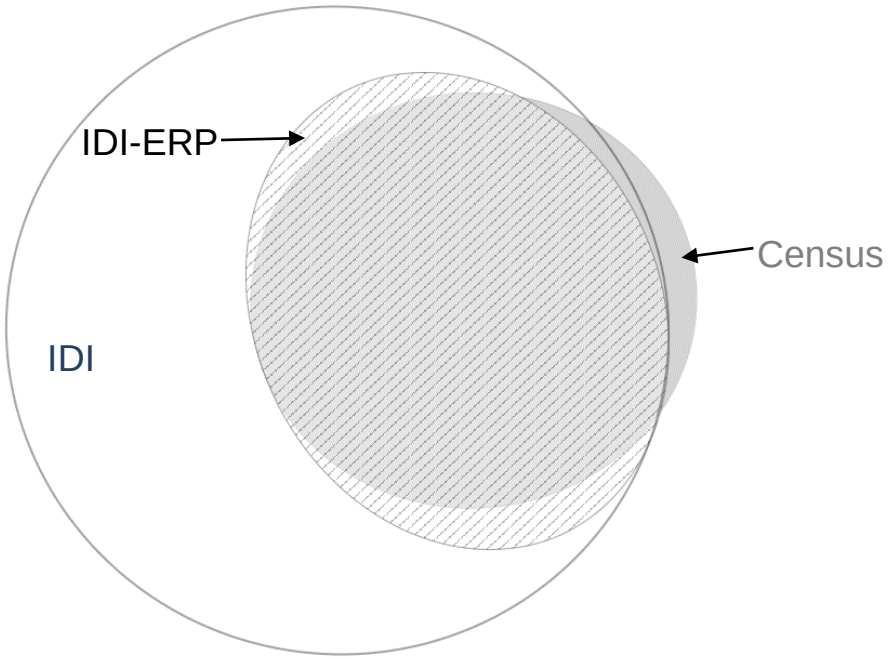
A total of 3,805,700 individuals were in both the IDI-ERP and census populations (the overlapping striped and shaded grey areas in figure 5). This represents 86.8 percent of the IDI-ERP population and 93.5 percent of the census population.

Some individuals in the census population were not found in the IDI-ERP (the non-overlapping grey areas in figure 5). Of the individuals who were in the census population, 6.5 percent (n=264,200) were not found in the IDI-ERP. As these individuals were identified as usual residents in the census, but were not included in the IDI-ERP, they can be thought of as potential undercoverage in the IDI-ERP population. Some of these individuals (1.9 percent, n=77,300) were in the IDI, but had not been selected in to the IDI-ERP population using the rules for defining a resident population from the IDI. The remainder (4.6 percent, n=186,900) were not found in the IDI at all.

There were also individuals who were in the IDI-ERP, but were not found in the census (the non-overlapping striped areas in figure 5). Of the individuals who were in the IDI-ERP, 13.2 percent (n=578,600) were not found in the census. As these individuals were included in the IDI-ERP population, but were not identified as usual residents in the census population, they can be thought of as potential overcoverage in the IDI-ERP population.

Figure 5

Overlap between the IDI, IDI-ERP, and census populations



This analysis suggests a first estimate of 6.5 percent undercoverage in the IDI-ERP and 13.2 percent overcoverage. However, we should be cautious of these estimates as there are several reasons why an individual may be in the IDI-ERP but not in census, or vice versa. Not all of these are 'true' undercoverage or overcoverage in the IDI-ERP. Other reasons for apparent coverage error include non-response in the census, and linkage errors in the Census-IDI link.

Impact of census non-response

Census non-response contributes to apparent overcoverage in the IDI-ERP. Individuals who are part of the usual resident population but did not fill out a census form may appear in the IDI-ERP but not in the census. As these individuals are part of the resident population, their inclusion in the IDI-ERP is correct. Total census non-response due to substitutes and undercount is 7.1 percent (Statistics NZ, 2014b).

Census overcount contributes to an apparent undercount in the IDI-ERP. Some individuals may be counted more than once in the census, or counted when they were not usual residents. Overcount was estimated to be less than 1 percent in the 2013 Census.

Impact of Census-IDI linkage errors

Linkage errors in the Census-IDI link are problematic because they inflate estimates of overcoverage and undercoverage in the IDI-ERP. If the records for an individual are not linked when they should have been (false negative link), the records for that individual will appear as two unlinked records – one in the census, and one in the IDI-ERP. The unlinked IDI-ERP record will be counted as IDI-ERP overcoverage (because it was in the IDI-ERP but not in the census). The unlinked record in the census will be counted as IDI-ERP undercoverage (because it was in the census but not in the IDI-ERP). Thus the false negative link will contribute both to apparent undercoverage and overcoverage in the IDI-ERP.

Conversely, if records for two different individuals are linked when they should not be (false positive link), the records will not be counted towards IDI-ERP undercoverage or

overcoverage totals, when in fact they should have been. False positive and false negative linkage error may affect some population groups more than others.

While the false positive linkage rate is estimated at less than 1 percent, it is difficult to distinguish between people in the census who were incorrectly missed in the linkage process and those genuinely not found in the IDI. The presence of linkage error in the Census-IDI link may lead us to incorrect conclusions about coverage of the IDI-ERP. Further work is needed to adjust for the effect of Census-IDI linkage errors before we have a better understanding of the undercoverage and overcoverage of the IDI-ERP.

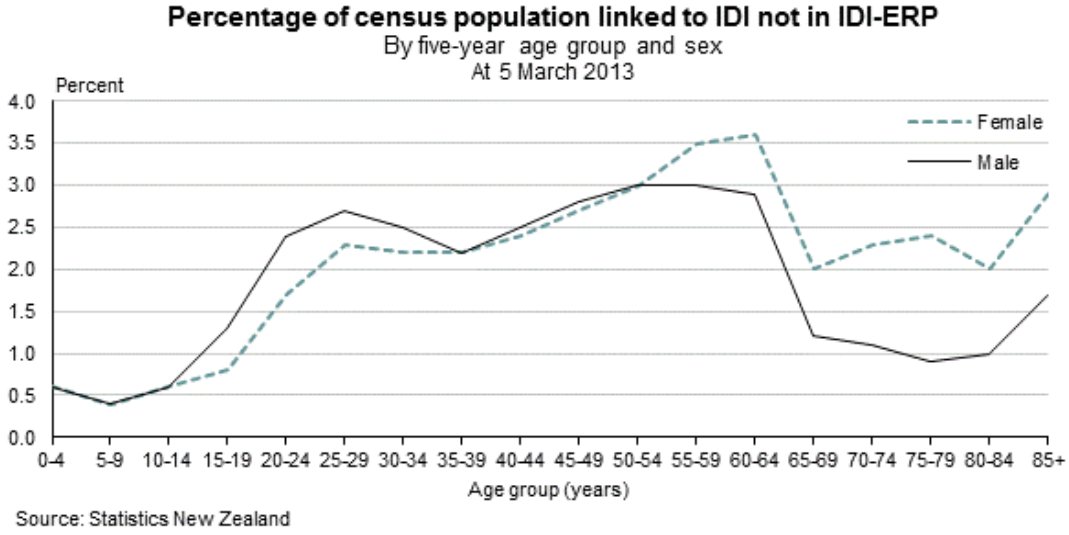
An example of IDI-ERP undercoverage

One subset of the IDI-ERP undercoverage population is of particular interest and less affected by these problems. Around 77,000 individuals were in the census population, and were linked to the IDI, but were not included in the IDI-ERP due to not meeting the requirements of the rules. Those individuals are likely to represent genuine undercoverage as they have been identified as residents in the census, but have not been included in the IDI-ERP.

Figure 6 shows, by five-year age group, the percentage of the linked Census-IDI population (the group of people who were found in both the IDI and the census) that was not included in the IDI-ERP. The figure shows that the percentage of people who were in the IDI but not in the IDI-ERP increases throughout adulthood. It reaches a peak at the 60–64-year age group, before declining.

At most age groups, the percentages are similar for males and females. The exception is the oldest age groups, where females are less likely to be selected into the IDI-ERP than males.

Figure 6



Source: Statistics New Zealand

6 Discussion

A New Zealand resident population (the IDI-ERP) has been derived from the linked administrative sources in the IDI. The IDI-ERP population is 2 percent larger than the official ERP population estimate. The overall pattern of the national age-sex distribution is similar to the ERP distribution, suggesting that the approach taken to deriving the IDI-ERP works reasonably well. However, coverage patterns vary by age and sex, with high net overcoverage in early adult ages (20–34 years), especially for males.

A key finding is that the accuracy of linkages becomes critical when we wish to count populations using linked data sources. It is just as vital to minimise missed linkages as it is to avoid linking two different individuals.

Coverage errors

Net coverage typically conceals underlying undercoverage and overcoverage. Linking the census to the IDI has provided some insight into the individuals who may be wrongly included, or wrongly excluded from the IDI-ERP. However, our initial estimates are inflated by linkage errors in the Census-IDI link and by census non-response.

Coverage errors in the IDI-ERP may be due to the rules we have used to define New Zealand residents, and to linkage errors in the construction of the IDI.

Sources of overcoverage in the IDI-ERP include:

- People who are not part of the resident population but were erroneously included in the IDI-ERP, for example short- or medium-term migrants.
- False negative links between component datasets of the IDI spine. IDI datasets are linked together probabilistically and are subject to linking error in the same way as the Census-IDI link. False negative links in the spine lead to an individual appearing twice, and therefore contribute to overcoverage in the IDI-ERP.

Sources of undercoverage in the IDI-ERP include:

- People who are part of the resident population but were not selected into the IDI-ERP because they did not have recent activity in the administrative data sources used here.
- People who are part of the resident population but do not appear in the IDI spine. For example those born overseas whose visa is before 1997 (or who do not require visas) and have no tax records.
- False positive links between component datasets of the IDI spine. False positive links lead to two individuals being counted as one and therefore contribute to undercoverage in the IDI-ERP.
- False negative links between the IDI spine and activity data sources. People may have been recently active, but a failure to link any record of activity to the spine would mean they are not included in the IDI-ERP.

At the aggregate level there appears to be considerable overcoverage in the IDI-ERP, suggesting that we are erroneously including individuals who are not New Zealand residents. Many of these erroneous inclusions are young adult males. They may be due to linkage error within the IDI spine (resulting in duplicate records for an individual) or to short-term visitors to New Zealand who are not identified as such from migration data. Errors in identifying migrants may be a result of the rules we have used to identify migrants, or to linkage errors involving the border movements data in the IDI.

In addition, the rules are failing to select some people who are usual residents. In particular there is a group of individuals who are in the IDI and census but are not being selected into the IDI-ERP. Many of these individuals are in the ages leading up to

retirement. They may not have activity in any of the relevant datasets in the IDI (for example, they have retired early and are not visiting a doctor regularly). Or they may have been active, but their records were not linked in the IDI. Or they may be absent from the spine, for example, if they migrated to New Zealand before 1997 (or do not require a visa to live in New Zealand), and have not worked or received a taxable benefit.

Further work

This work has been undertaken in the context of Census Transformation. The IDI-ERP administrative population estimates presented in this paper are likely to be initial counts, which would be further improved by coverage adjustments and estimation methods to fully meet the quality standards. Some level of coverage error in the IDI-ERP seems inevitable. However, larger discrepancies will require a larger coverage survey and greater reliance on models in the final estimation, with consequently higher costs and higher levels of uncertainty.

We anticipate that a method for identifying New Zealand residents at a given time will also be useful more generally for research using the IDI.

In conclusion, the structure of the linked administrative data available in the IDI, with a spine that targets those 'ever resident' in New Zealand linked to records of activity in health, taxation and education, international border movements and deaths, provides a solid basis for identifying a New Zealand resident population at a given time. However, further work is needed to understand the causes of undercoverage and erroneous inclusions that are apparent from this study.

Disclaimer

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

The opinions, findings, recommendations, and conclusions expressed in this paper are those of the author(s), not 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 have been confidentialised to protect these groups from identification.

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

Note: All figures presented in this paper have been rounded to the nearest hundred to protect confidentiality.

References

Bycroft, C (2015). Census transformation in New Zealand: Using administrative data without a population register. *Statistical Journal of the IAOS 31* (2015), 401–411. DOI: 10.3233/sji-150916.

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.

Gibb, S (2015). Quality of geographic information in the Integrated Data Infrastructure. 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.

Micallef, L, & Rodgers, P (2014). eulerAPE: Drawing Area-proportional 3-Venn Diagrams Using Ellipses. *PLoS ONE 9*(7): e101717. Retrieved from http://www.eulerdiagrams.org/eulerAPE.

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

Reid G, Bycroft, C, & Gleisner, F (in press). Comparison of ethnicity in administrative data and the census. To be available 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.

Suei, S (in press). Comparing income information in administrative data and the census [working title]. To be available 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). Coverage in the 2013 Census based on the New Zealand 2013 Post-enumeration Survey. Retrieved from www.stats.govt.nz.

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