Evaluating the potential of linked data sources for population estimates

The Integrated Data Infrastructure as an example

Sheree Gibb

Emily Shrosbree

New Zealand Government

Disclaimer
This paper represents the views of the authors. 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 1977.

Liability statement
While all care and diligence has been used in processing, analysing, and extracting data and information in this publication, Statistics NZ 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, & Shrosbree, E (2014). *Evaluating the potential of linked data sources for population estimates: The Integrated Data Infrastructure as an example.* Available from www.stats.govt.nz.

ISBN 978-0-478-42924-4 (online)

Published in September 2014 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 programme

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 socio-demographic 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).

The main emphasis of the longer-term strand is on the feasibility of producing census information from administrative data sources, as this aspect is the least understood. The investigations within the longer-term strand take a phased and iterative approach. The first phase is designed to provide sufficient evidence to inform decisions about the preferred direction for future development of the New Zealand census (see Statistics NZ, 2014a, for an overview). The early focus of the first phase is on developing an understanding of future census information requirements, and the ability of administrative data sources to meet those requirements.

Bycroft (2013) identified conceptual options for how future censuses might work in New Zealand and presented a set of working assumptions on which to base technical investigations into the options. In addition to the current five-yearly census, these options include a 10-yearly census, and several options based on administrative data.

About this paper

This paper is part of the first phase of the longer-term strand of the Census Transformation programme.

The paper describes a preliminary investigation into the feasibility of one of the administrative census options: linking multiple existing administrative data sources to produce a statistical population list. Under this option, the statistically constructed list of the New Zealand population would form the basis for estimating population counts without the need for a full census.

The linked administrative data sources available in Statistics NZ's Integrated Data Infrastructure (IDI) were used as a test environment to develop methods for constructing a statistical population list. Population estimates were then derived and compared with official estimated resident population figures.

While we identified clear limitations in the administrative sources available at the time of the study, the results show enough promise to continue with further investigations. The findings will help guide decisions about where to direct future work.

2 Introduction

Estimates of population size and structure underpin almost all social and economic statistics (United Nations, 2008). The five-yearly New Zealand Census of Population and Dwellings forms the basis for official population estimates. A key aim of the Census Transformation programme is to determine whether an alternative census model could produce accurate population estimates.

How are population estimates produced under the current census model?

Statistics NZ produces official national population estimates quarterly, and subnational population estimates annually, to territorial authority (TA) and area unit (AU) levels.[1] Population estimates are based on the estimated resident population (ERP) concept, which counts all individuals who usually live in a given area at a given time.[2]

The census is the starting point for deriving population estimates, although some adjustments are made to census counts. For the census date, the population estimates include:

- all residents present in New Zealand and counted by the census (the census usually resident population count)
- residents temporarily overseas (who are not included in the census)
- an adjustment for residents missed or counted more than once by the census (net census undercount).

Visitors from overseas are excluded.

In between censuses, the estimated resident population at a given date is produced using a cohort component methodology. This takes the base population estimate, adds births, adds net overseas migration (arrivals less departures), and subtracts deaths between census night and the given date. Birth registrations, death registrations, and overseas migration data provide good estimates of population change at a national level.

In addition, subnational population estimates must also estimate internal migration – the movement of people within New Zealand. Internal migration is estimated from symptomatic sources including:

- residential building consents
- information provided by territorial authority areas during an annual consultation round
- data on specific population groups: defence force personnel, prison populations, and tertiary students.

More recently, administrative data has played a more central role in the estimates process, as several additional datasets were used to estimate internal migration in 2012.

National population estimates are comparatively simple to produce because high-quality information on births, deaths, and international migration is available in New Zealand.

---

1. For definitions and explanations of Statistics NZ geographies, including territorial authorities and area units, see Geographic definitions (www.stats.govt.nz/Census/about-2006-census/2006-census-definitions-questionnaires/definitions/geographic.aspx)

2. See Demographic estimates (http://datainfoplus.stats.govt.nz/Item/nz.govt.stats/1802e0b8-3673-4d51-890c-5f2bed81f0a3)

Subnational population estimates are much more difficult, due to the lack of a direct measure of internal migration within New Zealand. The subnational population estimates present the greatest challenge when trying to produce population estimates without a full-enumeration census.

Aims and scope of this paper

This paper presents an investigation of the potential for linked administrative data sources to provide accurate population estimates by age, sex, and geographic area.

An existing linked data source, Statistics NZ's Integrated Data Infrastructure[3] (IDI), was used as a test environment. We took the IDI as it stood at April 2013 and did not attempt to change the structure or content of the IDI to suit our purposes.

Statistics NZ also produces population estimates by ethnicity, but ethnicity is not included in this paper due to the poor quality of ethnic information available in the IDI at the time of this study. Preliminary investigations revealed that only 65 percent of individuals had an ethnicity recorded and some ethnic groups were significantly over-represented.

The focus of this paper is solely population estimates. The potential for administrative data sources to produce other types of census information (for example, information about families, households, dwellings, education, or culture) is discussed in other work (O'Byrne et al, 2014).

This work is a preliminary investigation. It 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, this paper will provide broad information about the use of linked data sources that will guide future work.

The aims of this paper are:

- to evaluate the potential of linked administrative data sources for producing population estimates by age, sex, and geographic area, using the IDI as a test environment
- to identify limitations of the IDI, and linked data sources more generally, for producing population estimates.

3. See Integrated Data Infrastructure (www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure.aspx)

3 Data and methods

This section describes the data and methods used to evaluate the potential for linked administrative data sources to produce population estimates. It describes the structure of the IDI test environment, the methods used to produce population estimates from the IDI, and methods used to evaluate the quality of those population estimates.

Data source: The Integrated Data Infrastructure (IDI)

This subsection describes the structure of the IDI as at April 2013.

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 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. It can be used as a test environment for examining the potential of linked administrative data sources for producing population estimates.

The basic structure of the IDI (shown in figure 1) can be described as a central 'spine' to which a series of data collections are linked. The IDI spine consists of a list of all IRD numbers issued by Inland Revenue. Inland Revenue data works well as a spine because robust processes for issuing IRD numbers mean that we have high confidence that each IRD number corresponds to a unique individual, and that few individuals have more than one IRD number.

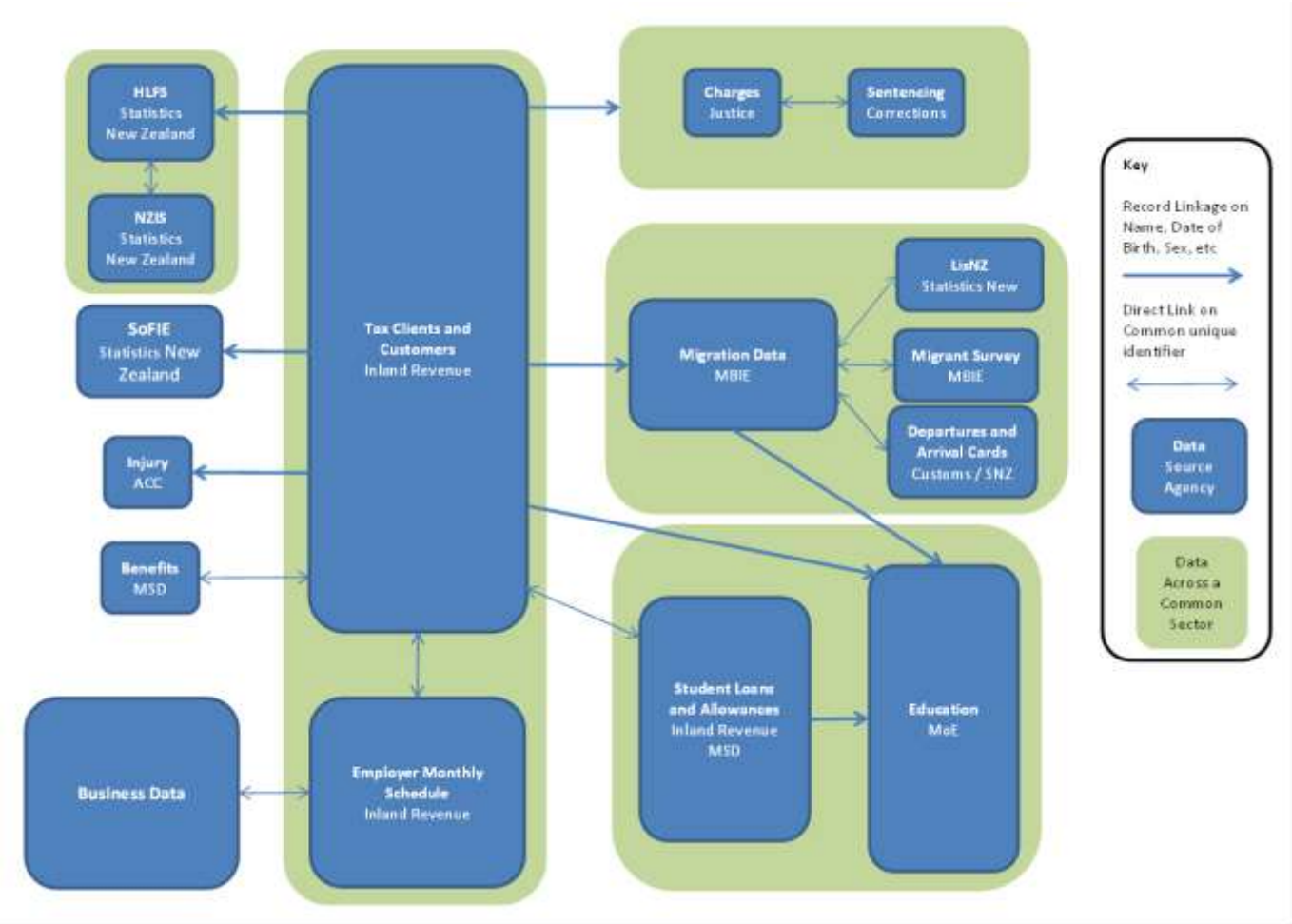
The datasets linked to the spine cover a wide range of subject areas and include:

- employer and employee job and earnings information from Inland Revenue
- education data from the Ministry of Education
- student loans and allowances data from several sources
- benefit dynamics data from the Ministry of Social Development
- migration movements data from the Ministry of Business, Innovation and Employment
- the Household Labour Force Survey and New Zealand Income Survey from Statistics NZ.

These links to the spine allow records for the same individual to be connected across different sectors.

The IDI continues to evolve as new datasets are added. Significant expansion is planned beginning in the second half of 2014.

Figure 1
Structure of the Integrated Data Infrastructure (IDI)
Source: Statistics NZ (2014b)



Description of IDI linking processes

This section describes how the links shown in figure 1 are formed.

The IDI contains many different data sources from different sectors across government. Unique identifiers are sometimes used within government sectors (for example the IRD number within the tax system, and National Student Number within the education system). Where possible, records in the IDI are linked directly using unique identifiers such as these. However, in most cases these unique identifiers are not held in common across different government agencies, and linkages to the IDI spine mainly use probabilistic record linkage techniques. The probabilistic linking uses several variables such as full name, date of birth, and sex to link records (Statistics NZ, 2014b).

When linking a data source like education to the spine, not all education records will be linked. This may be because a student does not have an IRD number and therefore does not have a record in the tax dataset. Or a student may have a tax number, but linking variables such as their name have been recorded differently in the tax and education systems, so a reliable link could not be made. The latter case is called a ‘false negative’ linkage error, and is important for the methods used in this paper. In both cases, the unlinked records are retained and are available for use within the IDI.

A detailed description of IDI linking procedures can be found in the IDI linking report (Statistics NZ, 2014b).

This paper focuses on the tax, education, and migration data linked in the IDI. A description of each of these data sources follows.

Tax data

Tax data from Inland Revenue plays an important role in the IDI. A list of all IRD numbers that have been issued to individuals (referred to as the Inland Revenue tax client register) forms the population spine of the IDI. This list of IRD numbers may include numbers that have not been used for some time (for example, for individuals who worked in New Zealand but now live overseas, or deceased individuals). In addition, other tax datasets are contained in the IDI, and are linked to the spine using IRD number.

The tax datasets used in this paper are:

- Employer Monthly Schedule (EMS) data, which includes all individuals who paid tax at source during the reporting month. 'Tax at source' is a concept used by Inland Revenue and includes:
    - tax paid on wages and salaries
    - tax paid on social assistance payments such as paid parental leave, student allowances, benefits, pensions, superannuation, and ACC payments
    - tax paid by overseas (ie non-resident) contractors and beneficiaries.

    Tax paid on self-employed income or investment income (such as interest from bank accounts) is not included in 'tax at source'. Individuals who earn all their income from self-employment or investment will not be included in the EMS
- annual tax return data, which provides information about all individuals who filed a tax return in a given year. This includes all self-employed individuals along with some individuals who paid tax at source.

Because the tax system includes most social assistance programs as well as earnings from wages and salaries and self-employment, most adults have registered for an IRD number at some time in their lives. More-recent incentives to register include Working for Families tax credits and the KiwiSaver retirement scheme.

Education data

IDI includes a series of tables containing individual-level information about secondary and tertiary education.

The education datasets used in this paper are:

- enrolment in modern apprenticeships, available from 2000 onwards
- enrolment in industry training, available from 2000 onwards
- enrolment in tertiary education courses, available from 1994 onwards
- achievement of NCEA credits (the New Zealand national secondary school qualification system), available from 2004 onwards.

Migration data

The IDI dataset includes individual-level data on all border movements in and out of New Zealand, sourced from the Ministry of Business, Innovation and Employment. The migration dataset contains information on date of birth, sex, movement type (in or out), and visa type. The data series is available from 1997 onwards.

Migration data is summarised in the 'person overseas spell' table, which lists overseas trips (periods of time spent outside of New Zealand), including the start and end dates of the trip, and the length of time spent outside New Zealand.

Method for constructing an estimated resident population from the IDI

This section describes the method used to construct a New Zealand resident population from the IDI. The population extracted from the IDI for the purposes of this paper is called the 'IDI-ERP'. The rules below are one of several possible ways of constructing a population from the IDI, and were chosen as the best, given the information available at the time of this study.

A population was constructed for a reference date of 30 June 2010. That is, the IDI-ERP was intended to estimate the resident population of New Zealand on 30 June 2010. The 30 June date was chosen so that the IDI-ERP can be compared with official ERP figures for subnational areas, which have a reference date of 30 June. The year 2010 was, as at April 2013, the most recent year for which complete sets of tax and education data were available in the IDI.

Previous work has shown that many individuals under age 20 are missing from the tax datasets (Statistics NZ, 2013). The education data held in the IDI is about tertiary student enrolments and high school achievement, and therefore can be expected to have good coverage across the ages where tax data is weakest.

The education and tax datasets in the IDI are the starting point for constructing a resident population from the IDI. That is, the initial population is individuals who are in either the IDI tax or education data sources. Because these datasets are linked in the IDI, in practice, this population includes:

- individuals in the education data who were linked to the tax spine
- education records not linked to the tax spine
- individuals in the tax spine not linked to the education data.

Over 5.6 million individuals have had activity recorded in the tax or education datasets in the IDI. Because this is greater than New Zealand's estimated resident population (ERP) of approximately 4.5 million, we need to restrict the IDI dataset to define a population that is closer (both conceptually and in number) to the ERP.

To reduce overcoverage in the IDI-ERP dataset, we took a 'signs of life' approach, with individuals only added to the population if they were active in one of the relevant datasets in the five years prior to the reference date. This method aimed to balance undercoverage (caused by using only a small number of datasets and/or a very short window for detecting 'signs of life') and overcoverage (caused by using a large number of datasets and/or a long window for detecting 'signs of life').

Using linked migration data to identify individuals who have left New Zealand, along with recorded dates of death in Inland Revenue data to identify deceased individuals, provides additional control for overcoverage. It allows a longer reference period to be used without accumulating individuals who are no longer living in New Zealand.

A five-year window was chosen because this produces lower levels of undercoverage than either one- or two-year windows, without the overcoverage seen with a longer 10-year window. These results held both for age-sex distributions (figure 8) and territorial authority areas (figure 9) in Appendix 1.

Extracting the IDI-ERP involves four steps:

1. Selecting individuals by identifying all those who, in the five years prior to the reference date, paid tax at source, filed an annual tax return, enrolled in tertiary education, or received NCEA credits. The resulting population can be described as the population contained in the union of the tax and education datasets. That is, unique individuals who appeared in the tax or the education datasets (or both).
2. Removing individuals if they were recorded as being overseas at the reference date, and the total duration of their time overseas was 10 months (305 days) or longer. A 10-month criterion was chosen to allow overseas-resident individuals to return to New Zealand for a brief visit each year, and therefore be more consistent with the current definition of permanent and long-term (PLT) migration used for population estimates.
3. Removing individuals if the IDI had a recorded year of death for them of 2010 or earlier. Year of death is sourced from Inland Revenue data. Inland Revenue records the year of death if they receive notification that a client is deceased. As a result, not all deaths are recorded in Inland Revenue data.
4. Obtaining basic demographic information (age, sex, and current meshblock of residence) from data held in the IDI for all individuals in the IDI-ERP.

This methodology effectively excludes children since very few individuals under age 15 will have paid tax in the previous five years, or earned NCEA credits (most students start to earn credits in year 11, when they are 15–16 years old).

A detailed description of the method used to construct the IDI-ERP for this paper, including a list of the IDI tables that were used, can be found in Appendix 2.

Criteria for assessing quality of population estimates produced from the IDI

This section describes the methods used to assess the quality of the population estimates produced from the IDI.

The quality of population estimates produced from the IDI was evaluated using a set of measures that have been used in previous work in the Census Transformation programme (O'Byrne et al, 2014). The quality measures are derived from the Statistics New Zealand Quality Model (Thomson, 2010), and are similar to quality dimensions used by other national statistics organisations. The quality measures, and how they were used in this evaluation, are shown in table 1.

Table 1

Quality measures used in the current study	
Quality measure	**Main questions used to evaluate this measure**
Relevance	How similar (conceptually) is the IDI population to the target population? Does the IDI dataset contain all variables necessary to produce population estimates (age, sex, and area)? Do the time periods in administrative datasets align with the reference time definitions used to produce population estimates?
Accuracy of coverage	How do IDI population totals compare to ERP totals, by age, sex, and territorial authority?

Accuracy of linking	What are the false positive and false negative error rates for the relevant links? What impact could false positive and false negative links have on population estimates produced from the IDI?
Timeliness	How frequently is the dataset updated? How long after the reference date is the data available for use in population estimates?
Accessibility	Not covered in this evaluation, as the IDI dataset is held by Statistics NZ

Relevance

A relevant administrative population will be similar to the estimated resident population (ERP) of New Zealand. The ERP is an estimate of all the people who usually live in New Zealand at a given date, including usual residents who are temporarily overseas, but excluding visitors to New Zealand.

The relevance of the administrative population will be assessed by examining metadata for the administrative datasets (information about how the administrative datasets are constructed and maintained) and comparing inclusion and exclusion criteria for the administrative population with that for the ERP.

A relevant administrative population will also contain all the demographic variables required to produce population estimates. Population estimates are published by five-year age group and sex at the territorial authority and area unit levels. Therefore, administrative datasets must contain age, sex, and location of usual residence, to the area unit level.

Relevance is also influenced by whether there is alignment between the reference time definition used for the ERP and how time is recorded in the linked data sources. Currently, population estimates provide a snapshot of the population at a point in time (for example, 30 June 2010). Administrative datasets may also use a ‘point in time’ approach, or count all individuals who appear over a period such as a week, month, or year. In a linked data source, different contributing datasets may have different time periods (some may cover a year, or a month, while others may measure a single point in time). This may make it difficult to estimate the population accurately at a specific point in time.

Accuracy of coverage

National coverage

National coverage is assessed by comparing national stock totals for the population extracted from the IDI (the IDI-ERP) with the national ERP at a given date. Coverage rates above 100 percent indicate overcoverage, while coverage rates below 100 percent indicate undercoverage. National coverage rates are examined by age and sex to determine whether there are any age- or sex-specific patterns of undercoverage or overcoverage.

Ideally, a dataset will have close to 100 percent coverage of ERP at the national level. In practice, however, a dataset will likely have some undercoverage (where administrative data stocks are lower than ERP) or overcoverage (where administrative data stocks are higher than ERP). While a small amount of overcoverage or undercoverage is acceptable, extensive under- or overcoverage will limit the usefulness of a dataset. There may also be some bias in coverage, with different age or sex groups having different coverage rates.

Even if the aggregate total population of a given dataset is close to the ERP (that is, coverage is close to 100 percent), there may still be overcoverage and undercoverage of

individuals within that dataset. Some individuals may be missed, and others wrongly included, but these errors may cancel each other out to produce good coverage at an aggregate level. A more accurate method of assessing coverage would be to link the individuals in a given administrative population with a list of individuals in the ERP. While this approach to coverage assessment was not used in this paper, it may be used in future work.

Subnational coverage

A major focus of this paper is on producing subnational population estimates, which are currently produced at the territorial authority (TA) and area unit (AU) levels. This paper focuses on subnational coverage at the territorial authority level, but also presents some initial findings at the area unit level. Accurate coverage at the area unit level is a considerably more difficult goal because it requires more-accurate address data than at the TA level. In-depth evaluation of area unit coverage is planned for future project work.

Subnational coverage was calculated in the same way as for national coverage: by comparing subnational population totals against the ERP subnational totals. As with national coverage, a dataset will likely have some amount of undercoverage or overcoverage at the subnational level. Coverage may be lower overall at the subnational level, as some addresses may be unable to be coded to a meshblock, resulting in missing data. Incorrect or out-of-date addresses will mean that individuals are assigned to the wrong area, which may lead to overcoverage or undercoverage in some areas.

Coverage rates for subnational areas may also be affected by inaccuracies in the ERP. Subnational ERPs are most accurate in census year, but between censuses the accuracy decreases. Subnational population estimates are more affected by this process than national estimates because of uncertainty in the internal migration component of subnational population estimates.

To examine the impact of inaccuracies in the ERP on IDI coverage rates, subnational coverage rates were calculated for two different years: 2006 (immediately following a census) and 2010 (four years after a census). Separate IDI-ERPs were constructed for each year, and coverage rates calculated by comparing these with the ERP for that year.

Accuracy of linking

It is important to consider the quality of IDI linking, because linkage errors affect population estimates generated from the IDI.

Linkage errors can be of two types:

- false positive links – occur when a link is made between two records that are not a true match (that is, records for two different individuals are linked)
- false negative links – occur when no link is made between records that are a true match (that is, records for the same individual in two different datasets are not linked).

Linking quality was evaluated by using available information about IDI linking – no new evaluation of the linking was done for the purposes of this paper. Available measures of linking quality included link rates and false positive error rates. An estimate of false negative rates was not available. Linking quality was also evaluated by estimating the potential impact that false positive and false negative linking errors could have on population estimates.

Timeliness

Subnational population estimates currently measure the population as at 30 June each year. Estimates for the population as at 30 June are released in October each year; however, future release schedules may be flexible. Nonetheless, to release population estimates in a timely manner, IDI data must be available in time for the annual delivery of

population estimates, without introducing too much delay into the release schedule. Data should be available by age, sex, and detailed geographic location (at a minimum, the area unit level).

Several factors may influence the timeliness of a linked dataset:

- The frequency with which individuals are added to or removed from the population. Some agencies may update their membership regularly, others intermittently – or, in the case of removing individuals from the population, not at all.
- The frequency with which information about individuals is updated. Information about date of birth and sex is unlikely to change over time. However, address information changes frequently and requires updating when individuals change their address. It is reasonable to expect a delay between when an individual changes their details (such as address) and when this change is recorded in an administrative dataset. However, if the delay is substantial then the dataset may not provide accurate population estimates.
- There may be some delay before the source agency can pass data on to Statistics NZ. This delay may be due to long collection periods (for example, data is collected over a year and released together at the end of the year), processing (coding or cleaning may take a long time), or embargoes.

4 Results

Relevance

Comparison of IDI-ERP and ERP concepts

The IDI-ERP population constructed for this paper consisted of all individuals who had paid tax, enrolled in tertiary education, or earned secondary school NCEA credits in the five years before the reference date, and were not overseas for 10 months or longer at the reference date. Most individuals included in the usual resident population will be included in the IDI-ERP. However, overcoverage and undercoverage both occur in the IDI-ERP population.

Individuals who are in the IDI-ERP population, but would not be in the ERP (overcoverage), include:

- individuals who are not usually resident in New Zealand, but have paid tax, enrolled in tertiary education, or received NCEA credits in the last five years, and have not departed New Zealand by the reference date. For example, an individual who arrives in New Zealand in May 2010, works for two months, and departs permanently in July 2010 would be included in the IDI-ERP for 30 June 2010, but not in the ERP
- individuals who are deceased as at 30 June 2010, but their death has not been recorded by Inland Revenue.

Individuals who are not in the IDI-ERP, but would be in the ERP (undercoverage), include:

- most individuals under age 15. The method used to construct the IDI-ERP is likely to exclude almost all individuals below age 15. Because of this, in all analyses described in this paper the IDI-ERP was restricted to individuals aged 15 and over
- individuals aged 15 or older who are resident in New Zealand, but have not paid tax at source, enrolled in tertiary education, or earned NCEA credits in the last five years. This may include:
    - individuals who are caring for children full-time (if they are not employed and not receiving a benefit)
    - individuals who are retired but not receiving New Zealand superannuation (these individuals may be receiving superannuation from overseas, or may have retired before age 65)
    - individuals whose sole income is investment income (for example, interest and dividends)
    - high-school students who have not yet received any NCEA credits
- individuals who are excluded from Inland Revenue's dataset because their records are highly sensitive – including all Inland Revenue employees (around 8,000 individuals), and high-profile individuals such as members of parliament and high-court judges.

There are also differences between the definitions of permanent and long-term migration in the ERP and the IDI-ERP that may contribute to overcoverage and undercoverage:

- individuals must spend 10 consecutive months overseas to be removed from the IDI-ERP. In contrast, they must spend 12 months (does not need to be continuous) overseas to be removed from the ERP
- calculations of the length of overseas stay for the ERP are based on traveller intentions, whereas calculations in the IDI are based on actual traveller behaviour
- in migration statistics, individuals must be usually resident in New Zealand before they leave to be categorised as a permanent or long-term departure (and therefore removed from the ERP). In the IDI, they only need to have paid tax or enrolled in secondary or tertiary education before they leave. Therefore, permanent and long-term departures in the IDI may include individuals who would be considered 'visitors' in migration statistics.

Different datasets within the IDI also apply to different time periods, which affects relevance. The tax data used in this paper consists of EMS summaries, which count all relevant individuals over a month, and annual tax return data, which counts all relevant individuals over a tax year (April–March). Education data counts relevant individuals over the period of a calendar year (January–December). These time periods do not align with each other, and furthermore, do not align with the 'point in time' method currently used to produce population estimates as at 30 June annually. Therefore, to produce an ERP for 30 June 2010, we must use datasets that cover a period from 1 January 2010 to 31 March 2011, and may contain individuals who were not part of the resident population at the 30 June reference date.

Availability of key variables

The key variables required to produce population estimates are age, sex, and location of usual residence (to at least area unit level). The IDI contains a dataset with personal information about individuals, summarised from a range of other IDI datasets. This table includes date of birth and sex. Appendix 3 contains information about how the information in this table is derived. Date of birth was available for all individuals in the IDI-ERP, while sex was available for 99.97 percent of individuals in the IDI-ERP (1,129 individuals did not have a sex recorded).

Address information from across the IDI is summarised in an IDI dataset listing all address changes recorded in IDI datasets. This table can be used to determine a meshblock for each individual in the IDI at any given time.

Around 95 percent of individuals in the IDI-ERP for 30 June 2010 had meshblock information recorded. Of those individuals that had a meshblock available, almost all meshblocks (more than 99 percent) were sourced from Inland Revenue data.

Previous work (Statistics NZ, 2013) has identified quality problems with Inland Revenue address data. Given that almost all IDI addresses come from Inland Revenue, these quality problems apply to address data in IDI.

Problems include:

- self-employed individuals have regular contact with Inland Revenue, so are likely to update their address details, but most other taxpayers (such as wage and salary earners) do not, and therefore have few opportunities to update their details
- addresses are recorded by Inland Revenue in free text format, thus some addresses do not contain enough information to accurately geocode them to the meshblock level
- the address individuals provide to Inland Revenue may not match their residential address, as they may provide an accountant's address, or a contact address where they choose to receive correspondence from Inland Revenue (such as their parents' address).

Accuracy of coverage

National coverage

The IDI-ERP was drawn from several different sources: tax data, tertiary education enrolment data, and secondary school achievement data. Migration data and recorded dates of death were also used to remove individuals from the population if they were identified as being overseas for 10 months or more, or deceased, at the reference date. This method generates a total national IDI-ERP population (for ages 15+) of 3,545,556 as at 30 June 2010. This represents 102.1 percent of the ERP for ages 15 and over as at 30 June 2010.

Table 2 shows how much of the IDI-ERP came from each source. The table shows that most individuals were added from tax data, with another 93,542 individuals added from tertiary education data, and 118,151 from secondary school data. Using a 'signs of life' approach reduced the population substantially, from 5.6 million ever found in tax or education datasets, to 4.1 million showing activity over the previous five years.

The table also shows that using death and migration data removed many individuals from the population. Overall, using death data removed 68,597 individuals from the population, representing around 1.7 percent of the original IDI-ERP. This is far fewer than the deaths recorded with the Department of Internal Affairs over the five-year period from 2006 to 2010 (approximately 141,000 for ages 15 and over), suggesting that more than half of deaths are not recorded in Inland Revenue data.

Using migration data affected the population more, with 443,473 individuals (around 11 percent of the population) being removed from the population because migration data showed they were overseas at the reference date (for 10 continuous months or longer). Over the five-year period from 2006 to 2010 there were approximately 300,000 recorded permanent and long-term (PLT) departures from New Zealand, according to official migration statistics produced by Statistics NZ (based on arrival and departure cards collected at the border). This number is lower than the 443,473 departures recorded in the IDI. This difference can likely be explained by the difference between the migration definitions used for the ERP and the IDI-ERP.

Table 2

Number of individuals in the final IDI-ERP obtained from different IDI datasets					
Datasets used	**Total national population (age 15+), as at 30 June 2010**				
	Activity in tax or education datasets ever	Activity in tax or education data in last five years[1]	Deceased individuals[2]	Overseas individuals[3]	Final population with deceased and overseas removed
Tax data only	4,747,207	3,773,662	68,529	371,270	3,333,863
Tax data, tertiary education data	5,515,902	3,909,233	68,584	413,244	3,427,405
Tax data, tertiary education data, secondary school data	5,643,345	4,057,626	68,597	443,473	3,545,556

1. Population generated as described in the section 'Method for constructing an estimated resident population from the IDI'.

2. Individuals with a date of death prior to 30 June 2010 in Inland Revenue data.

3. Individuals who were not recorded as deceased and were overseas for 10 months or longer at 30 June 2010.

Source: Statistics New Zealand

Figure 2 shows the relationship between the populations in the tax and education (secondary and tertiary) datasets for the IDI-ERP. The figure shows that most individuals in the education dataset are also represented in the tax dataset (85.8 percent of individuals in the education dataset also appear in the tax dataset). Only 211,693 individuals in the education dataset do not appear in the tax dataset.

An unknown number of individuals appear in neither the tax nor the education dataset. These individuals include children, and other individuals who have not been working, on a benefit or superannuation, or in secondary or tertiary education in the last five years.

Figure 2
Relationship between tax and education populations in the IDI-ERP, 30 June 2010, for ages 15 years and over

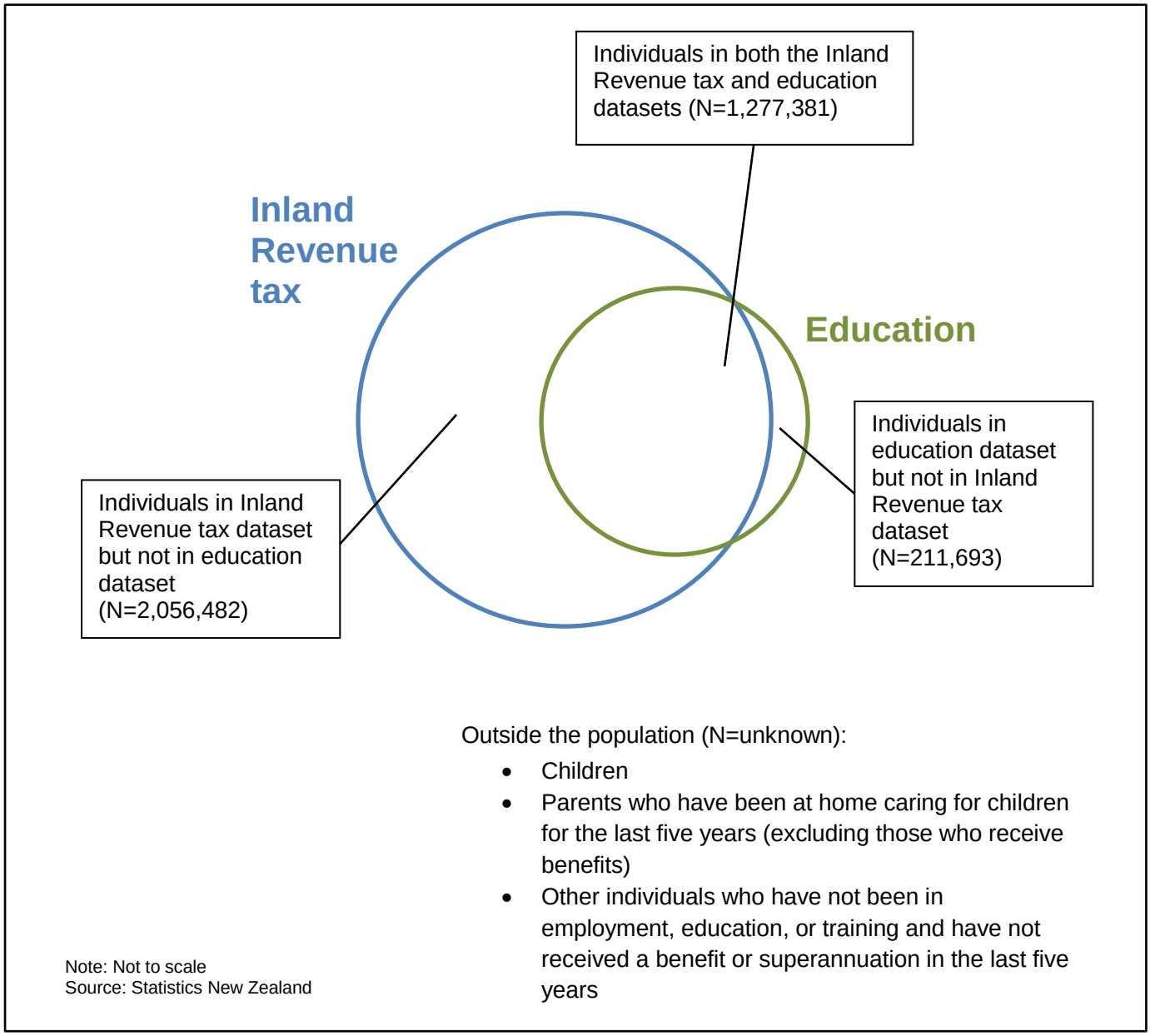
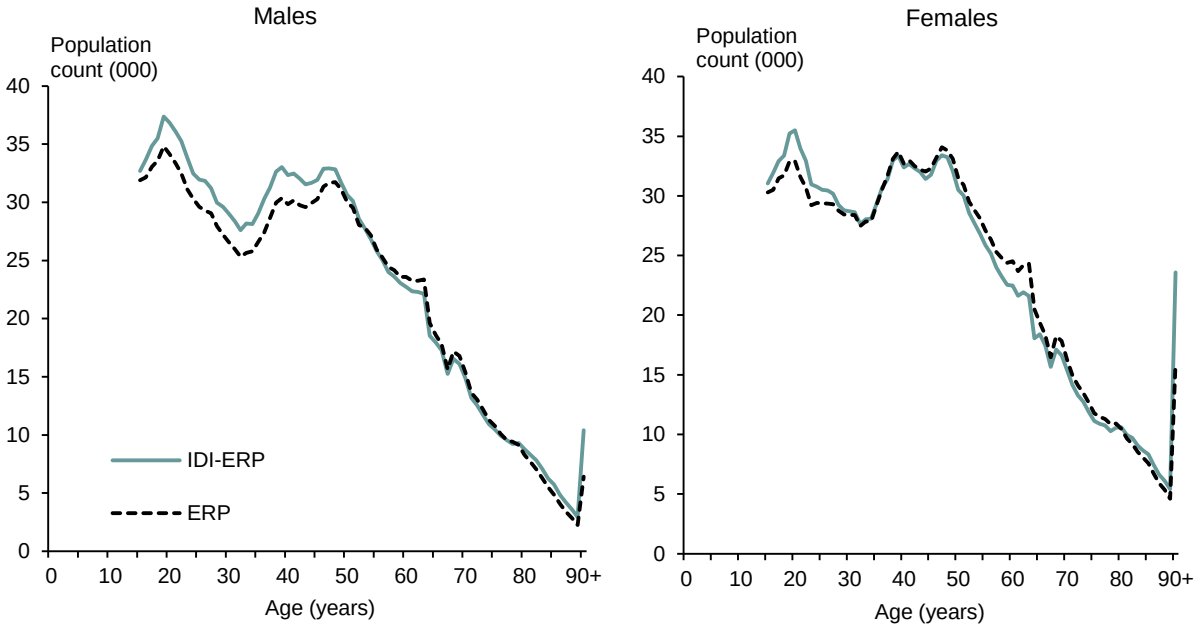


Figure 3 shows the IDI-ERP and ERP population counts, by age and sex, for ages 15 and over, as at 30 June 2010. The figure shows that coverage was high for all ages over age 15. Coverage was slightly lower for females than for males, particularly between the ages of 22 and 65. This is likely because some women in this age group have been caring for children full-time for the last five years, and have not been in paid employment or received a benefit, so are not included in the IDI-ERP.

There is some overcoverage (the IDI-ERP contains more individuals than the ERP) of males aged 15–50 years, and in the oldest age group (90 and over) for both sexes. Overcoverage in the oldest age groups may be due to recent deaths (within the last five years) that Inland Revenue has not recorded.

Figure 3
IDI-ERP and ERP national populations, by age and sex (15 years and over), June 2010



Source: Statistics New Zealand

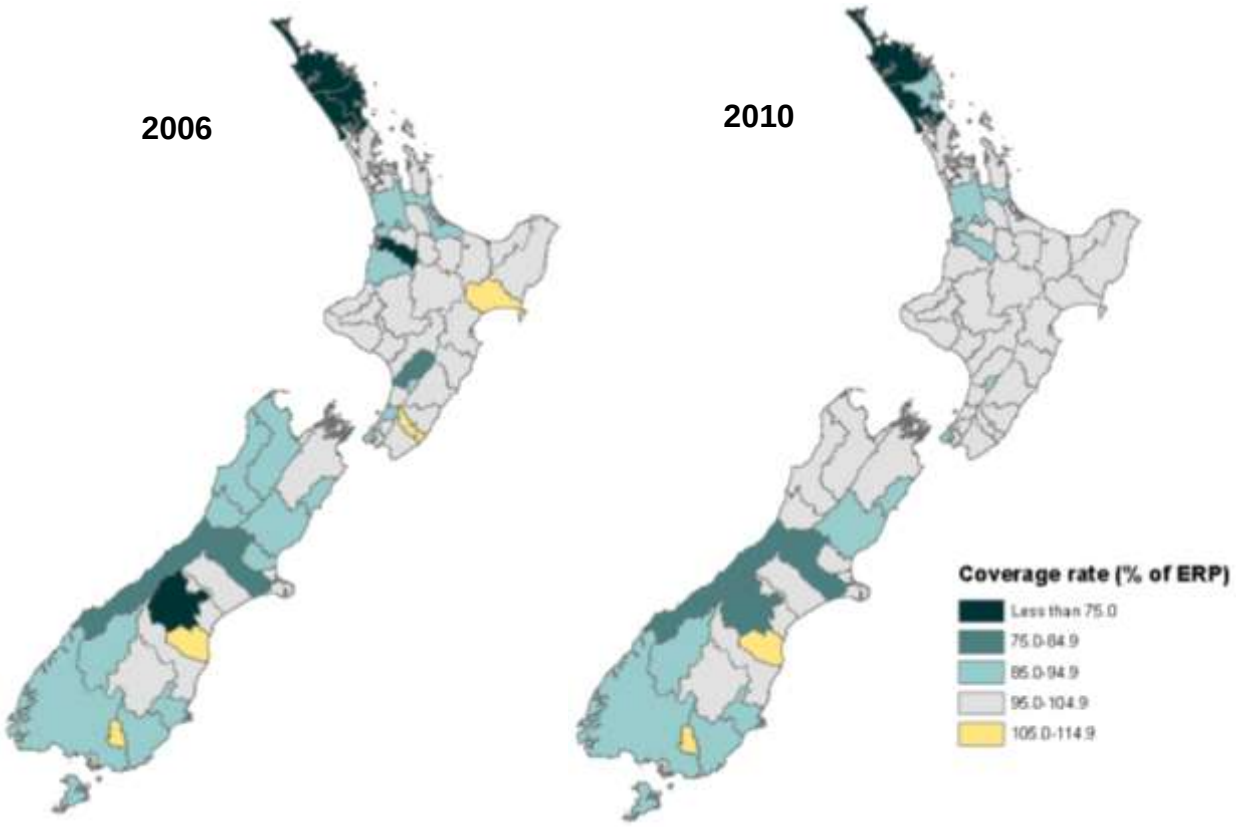
Subnational coverage

Due to missing meshblock information, the total IDI-ERP population available for analysis at the subnational level is smaller than the total IDI-ERP national population. Around 5 percent of individuals do not have a territorial authority recorded in the IDI, leaving a population of 3,362,495 individuals (age 15 and over) available for territorial authority-level subnational analysis. This population represents 96.8 percent of the national ERP for ages 15 and over.

Figure 4 shows the coverage of the IDI-ERP (as a percentage of ERP) for each territorial authority (TA). The IDI-ERP was extracted for two different time points: 30 June 2006 and 30 June 2010. The IDI-ERP for 2010 is the most recent available from IDI, but a limitation of assessing coverage in 2010 is uncertainty about the accuracy of the ERP for subnational areas four years from the 2006 census base. The 2006 census year ERP is likely to be more accurate.

The figure shows that, in both 2006 and 2010, coverage varied substantially across different subnational areas. Coverage was lowest in the Far North and Kaipara TAs, where it was below 75 percent. It was also low (below 85 percent) in Westland, Mackenzie, and Selwyn. In 2006, four areas had overcoverage relative to ERP (Gore, Waimate, Carterton, and Wairoa districts), while in 2010 only the Gore and Waimate districts did. Overall, however, coverage patterns across different TAs were similar for 2006 and 2010.

Figure 4
Coverage rates of the IDI-ERP (as a percentage of ERP), for population aged 15 years and over, by territorial authority
30 June 2006 and 30 June 2010

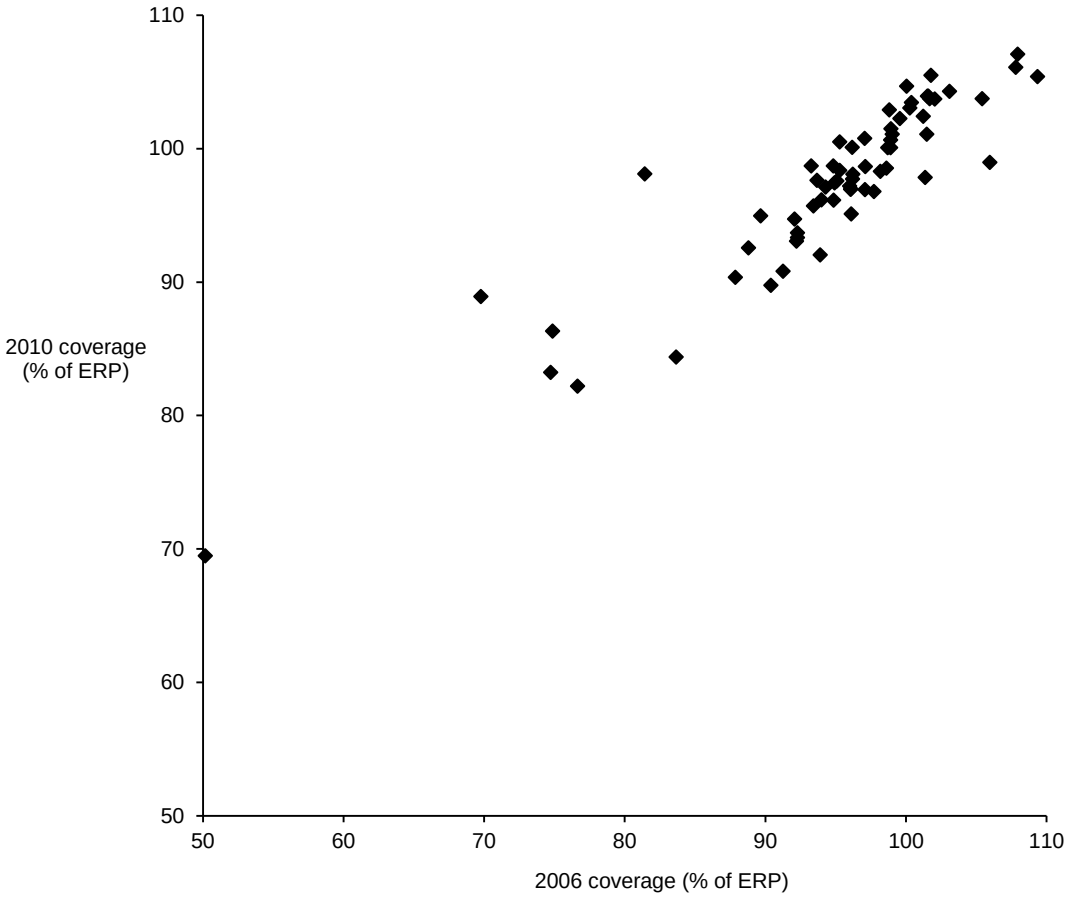


To examine whether coverage was similar in 2006 and 2010, we made a scatterplot of the 2006 coverage rate for each territorial authority (as a percentage of ERP) against the 2010 coverage rate for that territorial authority (as a percentage of ERP). The diagonal line represents exact correspondence between 2006 and 2010 coverage rates.

The figure shows that, in almost all cases, coverage was very similar in 2006 and 2010. The only exceptions are a few territorial authorities where coverage was considerably higher in 2010 than in 2006. Most of these territorial authorities had relatively low coverage rates in 2006, and changes in data collection procedures at the source agencies between 2006 and 2010 may have improved the coverage rates.

Overall, this comparison suggests that changes in the accuracy of the ERP between 2006 and 2010 do not significantly affect subnational coverage rates for the IDI-ERP.

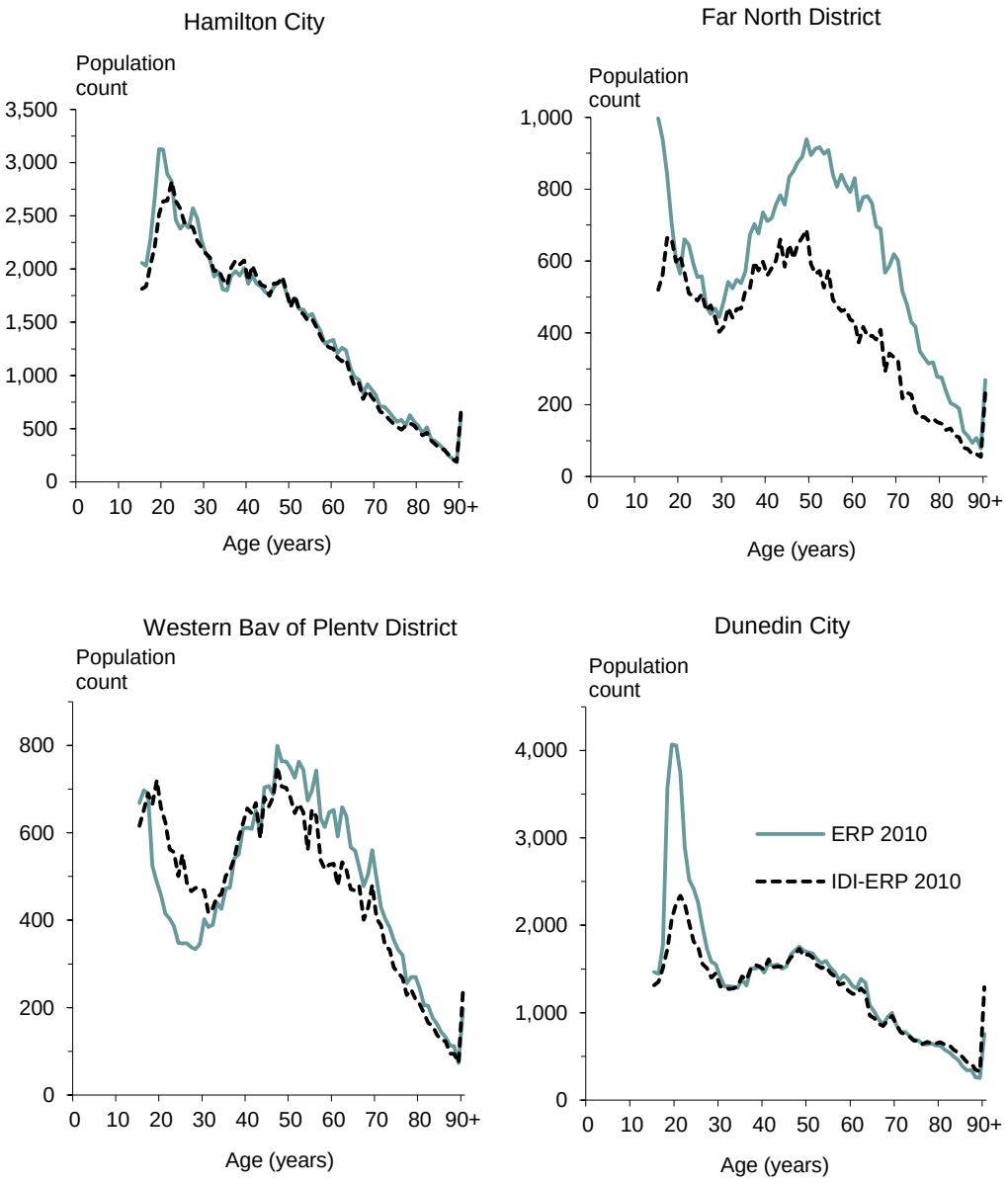
Figure 5
Comparison of 2006 and 2010 territorial authority coverage rates for the IDI-ERP For population aged 15 years and over



Source: Statistics New Zealand

The coverage of the IDI-ERP in June 2010 also varied by age across territorial authorities. Figure 6 presents four examples of this. For some TAs, such as Hamilton City, the IDI-ERP compares well with the ERP across most ages. Where overall coverage is low, such as in the Far North District, it is not uniform across all ages. Several areas, including Western Bay of Plenty, have overcoverage in younger age groups (around ages 20–35) and undercoverage in older age groups (over age 50). Student areas, such as Dunedin City, tend to show substantial undercoverage at student ages.

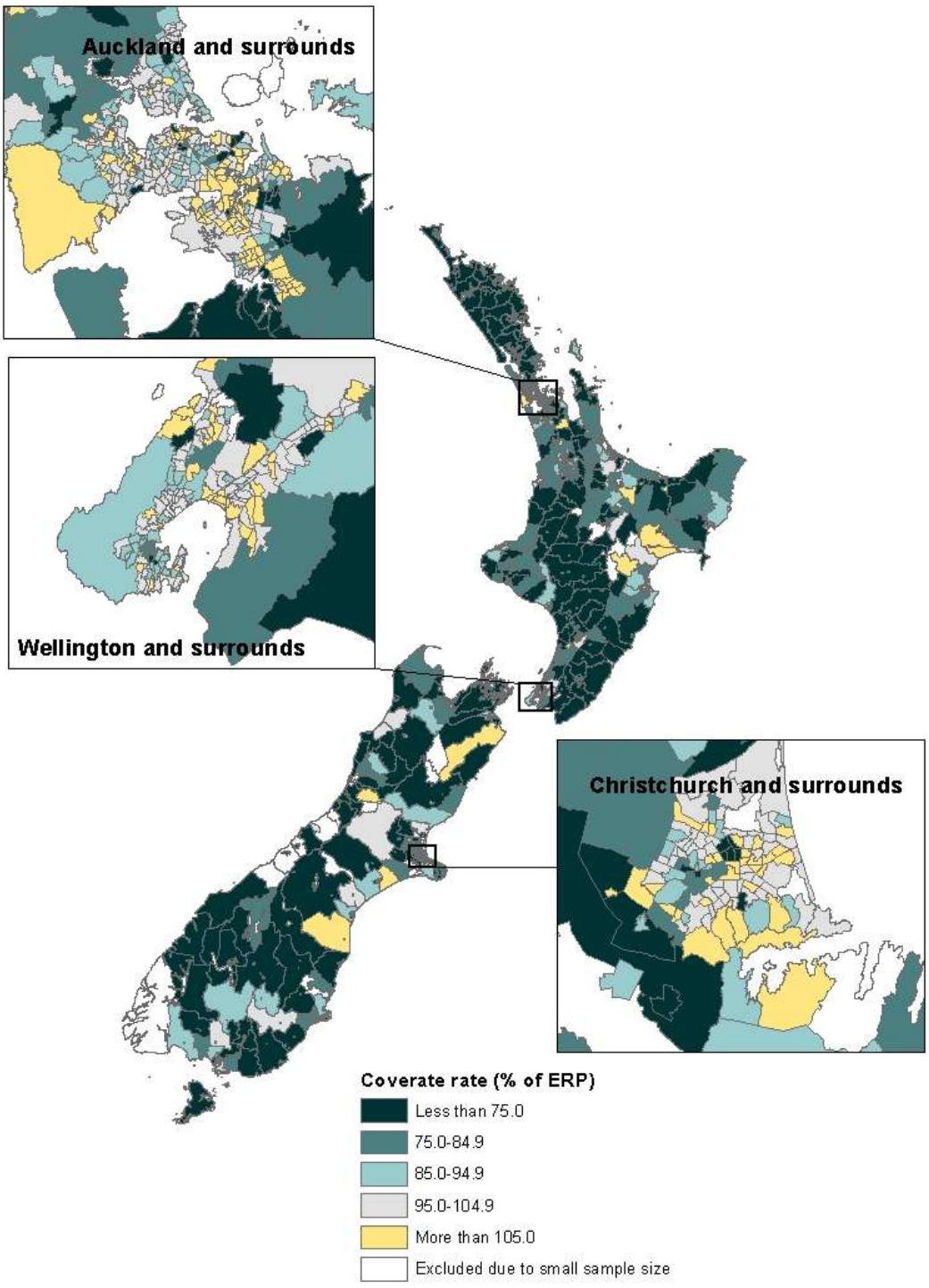
Figure 6
IDI-ERP and ERP populations for selected territorial authorities, by age, June 2010



Source: Statistics New Zealand

Further analysis of the June 2010 IDI-ERP at the area unit level revealed even greater variability. Figure 7 shows the variability of coverage of the IDI-ERP (as a percentage of ERP) for each area unit (AU). Coverage tends to be higher in urban area units and lower in rural area units.

Figure 7
Coverage rates of the IDI-ERP (as a percentage of ERP), for population aged 15 years and over, by area unit
30 June 2010



Accuracy of linking

The information in figure 2 is based on the assumption that all records are correctly linked between the tax and education datasets. In practice, some level of linking error is likely. Two IDI links are relevant to the current paper: the link between the education data and the tax spine, and the link between the migration data and the tax spine. For each of these links, two types of linking error may arise: false positives or false negatives.

False positive links

False positive links occur when two records that belong to different individuals are linked in error. Estimates of the false positive rate are available for all IDI links. These rates are estimated by clerical evaluation of a small subset of records. The false positive rates for the most recent linking were 0.3 percent for the education-tax link and 0.1 percent for the migration-tax link.

For the education-tax link, the false positive rate of 0.3 percent means there were approximately 3,832 incorrect links (of a total of 1,277,381 education-tax links in the IDI-ERP). For each incorrect link, two individuals have been counted as a single individual, which leads to undercoverage in the final population estimates. If all of these incorrect links were resolved (and the individuals concerned did not link to other unlinked records in the dataset), the total national IDI-ERP population would be increased by 3,832, or 0.1 percent.

For the migration-tax link, the false positive rate is estimated to be 0.1 percent. A total of 443,473 individuals were identified as not being resident in New Zealand on the reference date and were removed from the IDI-ERP. The false positive link rate suggests that 443 of these individuals (0.1 percent) were incorrectly linked. The impact of these incorrect links on population estimates depends on whether the relevant individuals were in New Zealand on the reference date. Nonetheless, the impact of such a small number is likely to be negligible.

False negative links

False negative links occur when two records for the same individual are not linked. False negative links lead to overcoverage in the final population estimate, as unlinked records for a single individual are counted as two different individuals.

Estimates of false negative link rates are not available for the IDI at present. The number of unlinked records provides an upper bound for the false negative rate. However, the rate is likely lower than this, because not all individuals in the education or migration datasets will have a record in the tax spine.

Although the false negative link rate in IDI is not known, it is still useful to consider the effect false negative links may have on the IDI-ERP. Table 3 shows the change in IDI-ERP for a range of false negative link rates for the tax-education link. The upper bound for the false negative rate for the education-tax link in the IDI-ERP is 14 percent, as 86 percent of education records are linked to a tax record.

Table 3 shows that even a moderate false negative rate (5 percent) would substantially affect the total national population, increasing it by 2.1 percent (74,454 individuals). In this example, the impact of false negative links on the total population count is minimised as the number of unlinked records in the education dataset is relatively small. If the number of unlinked records increased, false negative linking errors would have an even greater impact on the total population count.

Table 3

Projected impact of different false negative link rates on the IDI-ERP			
False negative rate in education dataset (%)[1]	Number of records affected[2]	Revised national population estimate	Percentage overcount in national population estimate
1	14,891	3,530,665	0.4
5	74,454	3,471,102	2.1
10	148,907	3,396,649	4.4
14	208,470	3,337,086	6.2

1. False negative rates are calculated with reference to the education dataset. Therefore, a false negative rate of 1 percent corresponds to 1 percent of the 1,489,074 records from the education dataset.

2. The calculation of the number of records affected assumes that each false negative link involves an unlinked record from the education dataset should have matched to an unlinked record from the tax dataset. In practice, the situation may be more complex.

Source: Statistics New Zealand

The impact of false negatives on population estimates would be greater if the false negatives were clustered in particular groups of individuals. For example, the education dataset contains a larger proportion of young people than the general population does: 52 percent of the education population is aged 15–29, compared with 28 percent of the total IDI-ERP. Therefore, false negatives in the education-tax link are likely to have a greater impact in younger age groups than in older age groups.

Table 4 shows the impact of false negatives in the education-tax link for broad age groups. The table shows that the impact of false negatives is greatest in the 15–29 age group, and smallest in the 60+ age group. A false negative rate of 5 percent would result in a 3.9 percent overcount in the 15–29 age group population, but only a 0.6 percent overcount in the 60+ population.

Table 4

Projected impact of different false negative link rates on the IDI-ERP, by age group				
False negative rate (%)	Overcount in age group population (%) due to false negatives			
	Age 15–29 (52% of education population)	Age 30–44 (24% of education population)	Age 45–59 (18% of education population)	Age 60+ (6% of education population)
1	0.8	0.4	0.3	0.1
5	3.9	1.9	1.6	0.6
10	7.9	3.9	3.1	1.1
14	11.0	5.5	4.4	1.6

Source: Statistics New Zealand

Timeliness

For producing population estimates, all datasets must be available in time to produce the estimates for any given reference year. As at April 2013, the datasets within the IDI were not current enough to produce population estimates by October each calendar year. In April 2013, the latest year for which all relevant tax and education datasets were available (and therefore for which population estimates could be produced) was 2010.

The timeliness of the IDI for producing population estimates is influenced by several factors.

Firstly, there may be delays in updating the source datasets. For example, individuals may delay informing a source agency about a change in their details (such as an address change, or a change in circumstances that means they would be added to or removed from the population). The extent of the delay will vary depending on the procedures source agencies use to maintain and update their datasets.

Secondly, there are delays in the supply of data from the source agency to the IDI team. The extent of this delay varies between different data sources. EMS tax data (used to identify individuals who have paid tax at source) are supplied to the IDI monthly, but there was an 18-month lag before they were incorporated into the IDI (this has since been shortened to three months). Tax data for individuals who pay tax annually, such as self-employed people, is provided to Inland Revenue in an annual tax return after the end of each financial year (31 March). Inland Revenue receives the largest volume of tax returns shortly after the end of the financial year. Information from these annual tax returns is passed on to IDI in the monthly updates. Education and migration data are currently supplied to IDI annually.

Lastly, IDI processing can cause delays. The IDI is updated every three months. Therefore, if the IDI receives data soon after an update, there may be a delay of up to three months before it is incorporated into the IDI system.

5 Discussion

Statistics NZ is investigating alternatives to the current census model. This paper discussed the potential for linked administrative data sources to produce population estimates, using Statistics NZ's Integrated Data Infrastructure (IDI) as a test environment. This paper is a first step in investigating the potential of linked data sources. Therefore, we took the IDI as it stood at April 2013 and made no changes to IDI content or structure for the purposes of this paper.

Summary of results

The population used for this analysis was generated from the union of the IDI tax and education datasets, with individuals removed from the population if they were overseas or known to be deceased at the reference date. Due to delays and embargoes on some data, the most recent year for which population estimates could be generated was 2010. The IDI had several strengths and limitations for producing population estimates.

Strengths

- It was possible to create an IDI-ERP that had good relevance and likely covered most of the same adult population as the ERP does.
- The national population count for ages 15 and over had good accuracy of coverage: the IDI-ERP was 102.1 percent of ERP as at 30 June 2010.
- The education-tax and migration-tax links had good accuracy of linking when measured by false positive error rates, meaning that false positive linking errors had minimal impact on the quality of population estimates produced from the IDI.

Limitations

- The 'signs of life' indicators used to select the population from IDI were not effective for children. While individuals aged less than 15 years may have IRD numbers and be present in the tax client register, most of them have no recent activity so will not be picked up.
- Accuracy of coverage was poorer at the subnational level than at the national level. The quality of address data was too low to produce sufficiently accurate estimates for subnational areas, particularly at the area unit level. Almost all IDI addresses come from Inland Revenue, and previous work (Statistics NZ, 2013) has identified a number of quality problems with Inland Revenue addresses.
- Accuracy of subnational coverage by age was also variable, with different subnational areas having different age patterns in coverage. In some areas these patterns may be explained by specific groups of individuals being missing from the administrative datasets. For example, the substantial undercoverage of those around age 20 in Dunedin City suggests that university students are underrepresented in the IDI-ERP, at least at their term-time address.
- Accuracy of coverage was poor for older age groups because only a small proportion of deaths are recorded by Inland Revenue, which leads to overcoverage.
- Accuracy of linking as measured by false negative error rates is unknown at present.
- Timeliness was poor due to lags in self-employed tax returns and education data that are currently too long to allow population estimates to be produced in a timely manner using the methods outlined in this paper. For example, under the current lags, population estimates for June 2013 cannot be produced until the end of 2014 at the earliest. This is much later than the current production date of October 2013.

What are the requirements for a linked data source to produce population estimates?

We need to consider the statistical features required in a linked data source to produce accurate population estimates. The findings from this paper suggest some initial requirements, described below.

Ability to select a relevant population from the linked data source

A key aim when generating population estimates from linked data sources is to generate a population that is conceptually close to the desired population measure. At present the most desirable target measure is the Estimated Resident Population (ERP), but other target populations (eg service populations) may also be useful.

If the target population of the linked data source does not map exactly to the desired target population, sufficient information should be available within the linked data source to allow users to define a more appropriate population from the linked data source. This information may include the criteria for inclusion and exclusion, how the population has been constructed, and how the datasets have been linked.

High-quality age and sex information

Having high-quality information on age (or date of birth) and sex is critical for producing accurate population estimates. Date of birth and sex are stable characteristics and are unlikely to change in an individual's life, so it is unnecessary to continually update these variables: collecting them once should be enough.

Accurate and up-to-date address information

Having high-quality address information for all or most of the resident population is a critical requirement for accurate subnational population estimates. Ideally, addresses should:

- represent the current usual residence of an individual at any given date
- contain enough information to be geocoded to at least the area unit level
- be updated regularly.

If several address sources are available, information will be needed to resolve conflicts between different sources, such as time stamps or address type information (eg postal or residential).

Measurements of linking accuracy: false positive and false negative links

Linking errors can substantially affect population estimates. The impact will be larger if the linking error rates are high, or if a large proportion of the population is being added through linking. It is vital that linked data sources have estimates of false positive and false negative link rates available to users of the datasets, and that these error rates are kept as low as possible.

Next steps / Future work

The current investigation is part of the first phase of Statistics NZ's research into alternative census models, and is a broad preliminary evaluation of the potential for linked data sources to produce population estimates. This first phase will guide decisions about where to focus further in-depth analysis.

The findings from the current paper suggest several areas for future analysis. These include the following areas.

Investigating whether changes to IDI could increase its potential to produce population estimates

Future work will need to consider whether changes can be made to the IDI to improve the quality of population estimates produced from it. Changes could include:

- improving address quality
- reducing the lag on tax and education data
- increasing coverage of children and adolescents
- estimating false negative linking errors
- linking death registrations into the IDI to identify deceased individuals.

More detailed evaluation of coverage

This paper measured coverage by comparing population totals (by age, sex, or area) with the ERP for the same time period. However, this approach has limitations:

- ERP figures are far out from the 2006 Census base and may be inaccurate, producing misleading coverage rates
- comparing aggregate totals does not provide any information about how closely the units (individuals) in the two populations match.

The 2013 census will provide an opportunity to address these limitations as it will deliver a more accurate ERP (available late 2014), and opportunities for more detailed coverage analysis. For example, linking an administrative data population (such as the IDI) to the 2013 census would provide useful information about individual-level coverage, and how coverage problems could be addressed (for example, through a coverage survey or imputation).

Evaluating the impact of linking errors on population estimates

In this analysis we estimated the potential impact of false positive links on population estimates. We couldn't estimate the impact of false negative links because no measure of these was available.

False negatives are difficult to identify using standard clerical review methods because a large number of records must be searched for potential matches. Nonetheless, to understand the full impact of linking errors on population estimates produced from the IDI it will be necessary to calculate an estimate of false negative linking errors.

Conclusion

Linked data sources are one option Statistics NZ is considering for an alternative to the current census model. For this paper, we used the IDI as a test environment to evaluate the quality of population estimates produced from a linked data source. The findings revealed that the IDI has several strengths in producing population estimates. But limitations were also revealed, particularly at the subnational level, and considerable work is required to address these.

Nonetheless, the preliminary findings suggest there is enough potential to continue investigating the linked-data-sources model as an alternative method of producing census information.

6 References

Bycroft, C (2013). Options for future New Zealand censuses: Census Transformation programme. Available 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. Available from www.stats.govt.nz.

Office for National Statistics (2012). Guidelines for measuring statistical quality. Available from www.ons.gov.uk.

Statistics NZ (2013). Evaluation of administrative data sources for subnational population estimates. Available from www.stats.govt.nz.

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

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

Thomson, S (2010). *Statistical quality model*. Statistics New Zealand: Unpublished report.

United Nations (2008). Principles and recommendations for population and housing censuses, version 2. Available from http://unstats.un.org.

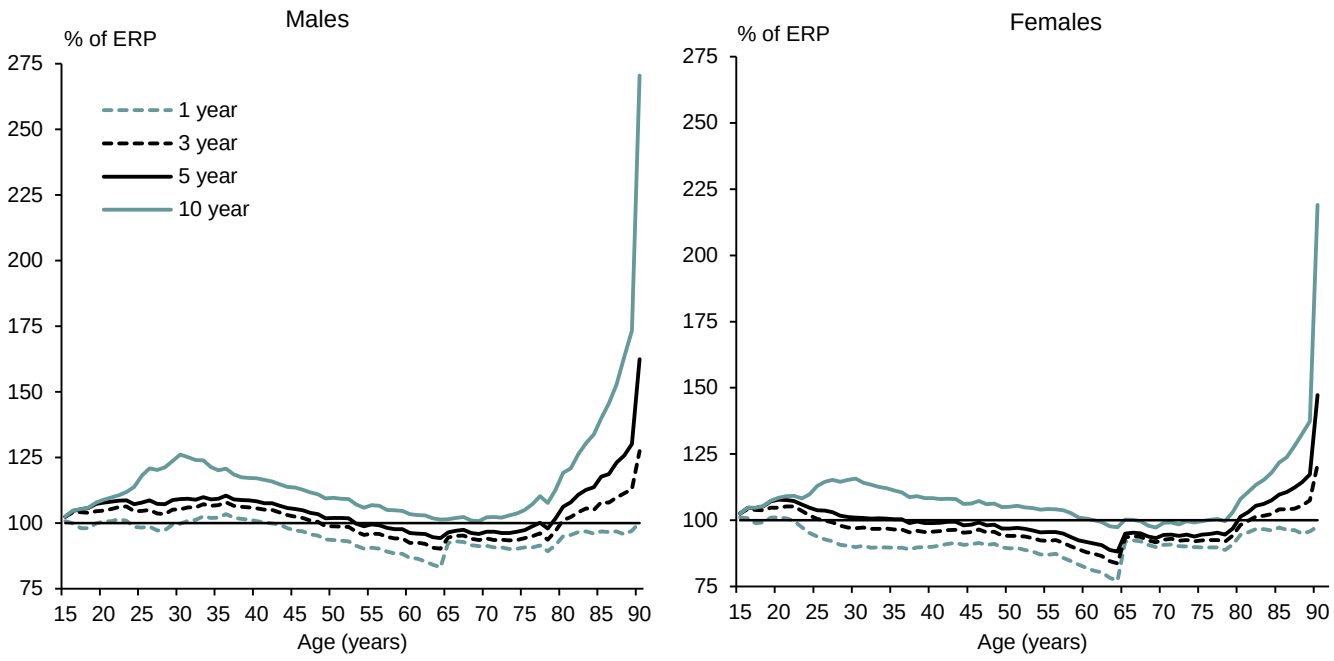
Zhang, L (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica 66*(1), 41–63.

Appendix 1: Impact of changing the window for generating the IDI-ERP

Figure 8 shows the coverage of the national IDI-ERP (as a percentage of the estimated resident population for ages 15 and over, as at June 2010), by age and by the length of the window used to detect 'signs of life'. Coverage increased with longer time windows across all ages.

The increase is particularly large for the older ages, with the coverage rate for ages 90 and over with a 10-year signs-of-life window being more than 230 percent of the estimated resident population for that age group. It is likely that the overcoverage in this oldest age group is due to deceased individuals whose records have not been removed from the population. The number of such individuals accumulates over time, producing more overcoverage in longer signs-of-life windows.

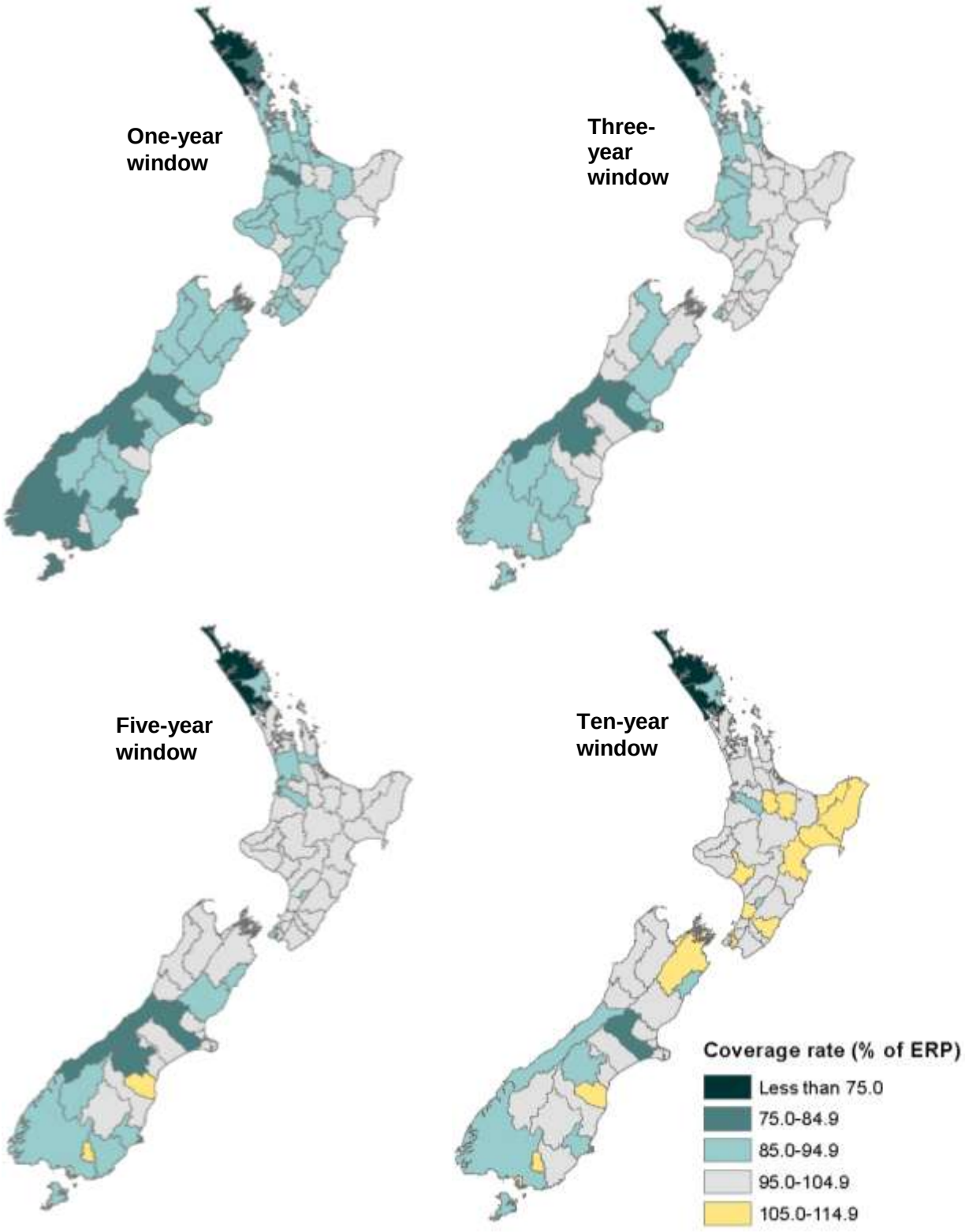
Figure 8
Coverage of IDI population (percentage of the estimated residential population), by age and by length of window used to detect signs of life



Source: Statistics New Zealand

Figure 9 shows the coverage of the IDI-ERP (as a percentage of ERP) for territorial authorities in New Zealand, by the length of the time window used to detect signs of life. The figure shows that short time windows (one year and three years) led to substantial undercoverage, with most territorial authorities having coverage rates of less than 95 percent. The longest (10-year) signs-of-life window produced overcoverage in some areas, although other areas continued to have low coverage even with this long time window. The five-year signs-of life-window produces lower levels of undercoverage than either the one- or the two-year windows, without the overcoverage seen with the 10-year window.

Figure 9
Coverage rates (percentage of ERP) for territorial authorities, by length of time window used to detect signs of life
As at 30 June 2010



Appendix 2: Method for constructing a population from IDI

The following rules describe the construction of an estimated resident population from IDI for the June 2010 year.

The following individuals[4] were added to the population:

- individuals who appeared in the employer monthly schedules (fact_job tables) from July 2005 to June 2010 (dim_return_period_key between 200507 and 201006 inclusive)
- individuals who appeared in the annual tax summary (dim_ann_employee table) from 2006 to 2010 (tax_year_nbr between 2006 and 2010 inclusive)
- individuals who appeared in the transitions dataset (trns_transitions table) who were recorded as attending school between 2006 and 2010 (moe_trn_at_school_ind=1 and moe_trn_year_nbr between 2006 and 2010 inclusive)
- individuals who appeared in the tertiary education enrolment dataset (enrolment table) between 2006 and 2010 (moe_enr_year_nbr between 2006 and 2010 inclusive)
- individuals who appeared in the modern apprenticeships dataset (tec_ma_learner table) between 2006 and 2010 (moe_mal_year_nbr between 2006 and 2010 inclusive)
- individuals who appeared in the industry training dataset (tec_it_learner table) between 2006 and 2010 (moe_itl_year_nbr between 2006 and 2010 inclusive).

From the total population described above, the following individuals were removed:

- individuals who were overseas on 30 June 2010, if their total length of stay overseas was 10 months or longer. These individuals were identified from the person_overseas_spell table (individuals with pos_applied_date on or before 30 June 2010, pos_ceased_date on or after 30 June 2010, and pos_day_span_nbr of 305 or greater) were considered to be overseas as at 30 June 2010).
- individuals who were deceased (according to Inland Revenue records) before or during 2010 (snz_deceased_year_nbr less than or equal to 2010).

Prioritisation rules for date of birth and sex

Date of birth and sex are available from many different datasets in the IDI and are summarised in the Personal Detail table of the IDI according to the following rules:

1. Take the most common value from across all sources, counting only unique values from each source.
2. If there is no 'most common' value from step 1, take the most common value from the following collections in order of priority (counting only unique values from each source):
    - moe_qs.moe_master
    - sla_qs.msd_master
    - leed_qs.msd_master
    - sla_qs.ird_master

4. An individual is defined as a unique identification number (snz_uid) in the IDI.

- leed_qs.ird_master
- dol_qs.dol_master
- sla_qs.slam_moe_master
- sla_qs.slam_ird_master
- hlfs_qs.hlfs_master.

If the first prioritised source has no 'most common' value or is null, then take from the second prioritised source (and so on).

3. If no 'most common' value from step 2, randomly assign a value from all the unique values from across all sources.