Evaluation of administrative data sources for subnational population estimates
This report was prepared by Sheree Gibb and Alan Ambury of Statistics New Zealand.

Crown copyright ©
This work is licensed under the Creative Commons Attribution 3.0 New Zealand licence. You are free to copy, distribute, and adapt the work, as long as you attribute the work to Statistics NZ and abide by the other licence terms. Please note you may not use any departmental or governmental emblem, logo, or coat of arms in any way that infringes any provision of the Flags, Emblems, and Names Protection Act 1981. Use the wording 'Statistics New Zealand' in your attribution, not the Statistics NZ logo.

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


Published in September 2013 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 ................................................................. 4

1 Purpose and summary ................................................................. 5
   Purpose ....................................................................................... 5
   Summary ................................................................................... 5

2 Introduction .................................................................................. 6
   Why use administrative data to inform population estimates? ........ 6
   Project aims and scope ............................................................... 7

3 Methods ....................................................................................... 8
   Data sources ............................................................................. 8
   Data processing ........................................................................ 9
   Stocks and flows ...................................................................... 10
   Metadata collection .................................................................. 10
   Coverage assessment ............................................................... 10
   Criteria for assessing quality .................................................. 11

4 Results ......................................................................................... 16
   Primary health organisation (PHO) enrolment data .................. 16
   Inland Revenue (IR) tax data .................................................... 24
   School roll returns data ......................................................... 36
   Electoral enrolment data ......................................................... 41

5 Discussion .................................................................................. 47

6 References .................................................................................. 51
# List of tables and figures

## List of tables

1. PHO coverage rates for area units, 2012 ................................................................. 20
2. Quality summary of the PHO dataset ................................................................... 24
3. IR tax coverage rates for area units, 2012 ............................................................. 30
4. Quality summary of IR tax dataset ....................................................................... 35
5. Roll returns coverage rates for area units, 2012 ................................................... 40
6. Quality summary of roll returns dataset ................................................................. 41
7. Electoral enrolment coverage rates for area units, 2011 ....................................... 45
8. Quality summary for the electoral enrolment dataset ............................................. 46
9. Quality summary for the four administrative datasets ......................................... 48

## List of figures

1. Percentage of individuals who have been at their current usual residence less than one year, by age and sex, 2006 Census data ......................................................... 13
2. National PHO enrolments and ERP, by age and sex, 30 June 2012 ....................... 17
3. National PHO coverage, relative to ERP, by age and sex, 30 June 2012 ............... 18
4. PHO coverage rates, by territorial authority, 2012 ................................................. 19
5. Meshblock changes recorded in PHO data, by age and sex, 2012 .......................... 21
6. Moves out of Christchurch, by quarter, PHO data .................................................. 23
7. National IR tax enrolments and ERP, by age and sex, 30 June 2012 ..................... 27
8. National IR tax coverage, relative to ERP, by age and sex, 30 June 2012 ............. 27
9. IR tax coverage rates, by territorial authority, 2012 .............................................. 29
10. Territorial authority changes recorded in IR tax data, by age and sex, 2011 ......... 31
11. Net internal migration for selected areas in IR tax data, by territorial authority and age, 2011 ................................................................. 32
12. Moves out of Christchurch, IR tax data, by quarter ............................................... 34
15. July roll returns coverage rates, by territorial authority, 2012 ............................ 39
16. National electoral enrolment population and ERP, by age, 30 June 2012 .......... 43
17. National electoral enrolment coverage rate, relative to ERP, by age, 30 June 2012 ... 43
18. Electoral enrolment coverage rates, by territorial authority, 2011 ......................... 44
19. Comparison of national coverage rates for the four administrative datasets, by age, 2012 ................................................................. 49
1 Purpose and summary

Purpose

This report presents an evaluation of the potential for four administrative data sources to improve subnational population estimates. It builds on previous work at Statistics New Zealand (Statistics New Zealand, 2011) that identified administrative datasets that had potential to contribute to subnational population estimates.

Summary

• Estimating subnational populations in intercensal years (the years when no Census of Population and Dwellings is held) is difficult because there is no accurate measure of internal migration.

• The increasing availability of administrative data sources (data sourced from the administrative records of other agencies, rather than from Statistics NZ’s direct contact with respondents) provides an opportunity to improve subnational population estimates.

• This project evaluated the quality of four data sources for producing subnational population estimates:
  o primary health organisation (PHO) enrolments
  o Inland Revenue (IR) tax
  o school roll returns
  o electoral enrolments.

• No single administrative data source can currently produce sufficiently accurate subnational population estimates. All of the data sources have limitations. However, an aggregate or unit-record linked model combining several data sources may provide opportunities for further improving subnational estimates.

• None of the data sources have sufficient coverage of the 17–30-year age group. This leads to problems estimating populations for this age group, and for areas with large flows of young adults (eg Dunedin, Palmerston North, and to a lesser extent, Wellington and Auckland).

• Accurate and up-to-date address data is vital for subnational population estimates, but many administrative data sources contain low quality address data. Better quality of address data would be one of the most valuable improvements that could be made to administrative data sources.
2 Introduction

Why use administrative data to inform population estimates?

The Census of Population and Dwellings provides an accurate count of the resident population\(^1\), both at a national and subnational level, every five years.

During intercensal years, population change must be estimated and then added to the census base to derive accurate population estimates. The accuracy of intercensal population estimates at the national level is thought to be good, due to high-quality data on births, deaths, and international migration. Subnational estimates are difficult, however, because there is no direct measure of internal migration.

Traditionally, estimates of internal migration have been derived from data from the most recent census, along with symptomatic indicators such as building consents and consultation with local government authorities. However, the Canterbury earthquakes in 2010 and 2011 highlighted the need for a better measure of internal migration. The earthquakes caused the 2011 Census to be cancelled, so population estimates are now seven years out from the 2006 Census base. They also caused substantial migration in and around the Christchurch area. This migration was difficult to measure accurately using the traditional symptomatic indicators.

As a result, there has been increasing attention on whether administrative data can help to improve subnational population estimates in intercensal years.

Current methods for producing subnational population estimates

Statistics NZ currently produces subnational population estimates annually to territorial authority and area unit levels. Population estimates are based on the resident population concept – the estimated resident population (ERP) includes all individuals who usually live in a given area at a given time.

Subnational population estimates are produced using a component methodology, which takes the base population, adds births, net overseas migration, and net internal migration, and subtracts deaths.

Birth, death, and overseas migration data are sourced from datasets Statistics NZ holds. Internal migration is estimated from a range of symptomatic sources including:

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

In recent years, Statistics NZ has started using administrative data in the subnational population estimates process. In 2012, administrative data played a more central role in the estimates process, as several datasets were used to estimate internal migration.\(^2\)

---

\(^1\) Estimated resident population for 30 June in census year is based on the census population count adjusted for residents temporarily overseas, a measure of net undercount at census, and population change between census and 30 June of the same year.

\(^2\)
Subnational population estimates in DataInfo+ has a detailed description of the population estimation process.

Project aims and scope

The major aim of this project was to examine whether administrative data sources can improve the accuracy of subnational population estimates.

Statistics NZ has previously undertaken a brief overview of a range of administrative data sources to determine which data sources have the greatest potential to improve subnational population estimates (Statistics NZ, 2011). Four main data sources were identified in the review as having the greatest potential to inform subnational population estimates:

- primary health enrolment (PHO) data
- electoral enrolment data
- Inland Revenue (IR) tax data
- school roll return data.

The scope of the current project was to provide a more in-depth evaluation of the quality of these four data sources for producing subnational population estimates. The criteria for assessing quality are described in the ‘methods’ section.

Subnational population estimates are produced by age and sex at territorial authority and area unit level. Therefore, data sources have been evaluated for their potential to provide population estimates by age, sex, and detailed geographic area.

At this stage, the data sources have not been evaluated for their potential to provide population estimates by ethnicity or Māori descent. Future project work will examine these issues.

---

2 Flow data from the primary health organisation and tax datasets were used to estimate internal migration for the 2012 subnational population estimates. Stock data from electoral enrolments, primary health organisation enrolments, and school roll returns were used to check and adjust area unit stock populations.
3 Methods

This section describes the methods used to evaluate the quality of the administrative data sources. It includes a description of the data sources used, information about the ways data were processed, and a description of the methods used to obtain information about data quality, including metadata and coverage assessment.

Data sources

The four data sources evaluated in this project were:

**Primary health organisation (PHO) enrolments**

The primary health organisation (PHO) enrolment dataset is owned by the Ministry of Health. The Ministry of Health funds primary care (GP) services based on the number of individuals on GP enrolment lists. Funding is distributed to GPs via PHOs – umbrella organisations that collect enrolment information from member GP practices.

The PHO enrolment dataset contains a list of all patients enrolled with PHOs, along with basic demographic information about each patient. Patients enrol in PHOs via their GP. Enrolment is not compulsory, but there is a strong incentive for patients to enrol, as they pay substantially reduced GP consultation fees. Patients can only enrol with one PHO at any given time.

PHO enrolment data are available from 2004 onwards, although low enrolment rates in the early years mean data from 2005 onwards are most useful.

**Inland Revenue tax data**

Inland Revenue (IR) collects money to pay for public services under the Inland Revenue Act and other relevant laws. As part of administering the tax system, the department collects information about individuals and entities (businesses).

IR currently provides Statistics NZ with a monthly update of taxpayer information. These updates include basic demographic information about New Zealand taxpayers, along with information about recent interactions with IR (such as paying tax or filing a tax return). IRD number is used as a unique identifier, and IR invests considerable effort in ensuring individuals have only one IRD number.³

For this project, IR tax data were accessed via the linked employer-employee dataset (LEED) that is maintained by Statistics NZ. Data are available in LEED back to April 1999.

**Electoral enrolments**

The Electoral Commission (EC) maintains a database of all individuals eligible to vote in parliamentary, general, or local authority elections in New Zealand. This information is used to determine eligibility and to administer elections. The electoral enrolment database contains the name and address of all registered voters⁴, along with limited demographic information.

---

³ One of the few exceptions is individuals who have been declared bankrupt, who are issued with a new IRD number.

⁴ Statistics New Zealand received anonymised aggregate data on electoral enrolments. Personal details such as name and exact address were not part of the data provision.
It is mandatory for eligible individuals to be on the electoral roll. Multiple enrolments are not permitted and EC invests considerable effort in ensuring that duplicates do not occur.

**School roll returns data**

The Ministry of Education collects data on school enrolments two times a year for all schools, and two more times for secondary schools. Every school must submit school roll returns in March and July each year.

The July school roll return is the most timely for population estimates purposes. It provides a snapshot of the number of students recorded on school rolls in July each year.

Schools provide information either via their student management systems or by recording it on a standard form. All schools submit counts of student enrolments by sex and year level, and schools with an electronic student management system submit additional demographic information about their students.

Students are only permitted to be enrolled at one school at any given time.

The Ministry of Education also maintains the ENROL dataset, which contains information about all current school enrolments in New Zealand. However, the ENROL dataset is not covered as part of this evaluation.

Initial analysis of the ENROL dataset suggested that it was not suitable for population estimates due to the low relevance of the geographic information in the dataset (the addresses relate to school addresses rather than residential addresses). Student residential address information is also recorded, but the high level of missing data means this information is unlikely to be useful for population estimates purposes.

School roll returns contain information on student residential address with a much lower level of missing data, and therefore is more likely to be useful than ENROL data.

**Data processing**

Most of the administrative datasets require some processing before they are suitable for use in population estimation.

Much of this processing relates to geographic information. For three of the datasets (PHO, electoral enrolments, and roll returns), the data received by Statistics NZ includes geographic information that the data supplier has already coded to meshblock (the smallest geographic unit for which Statistics NZ collects statistical data). For these datasets, concordances are used to convert meshblocks into area units and territorial authorities.

Additional processing is required for the geographic information in the IR tax dataset. Raw address information was extracted from tax data and geocoded in-house using the Classification Coding System. Around 80 percent of raw addresses could be geocoded to meshblock. It was possible to assign almost all of the uncoded addresses to a territorial authority (but not to smaller geographies) using a combination of postcode and place name matching.

Additional processing not related to geographic information included removing data with missing values and calculating age from date of birth.

---

5 For the datasets where date of birth is available: PHO and IR tax.
Stocks and flows

There are two ways of measuring population change: stocks and flows. Stocks are part of Statistics NZ’s regular published outputs for population estimates, and flows are sometimes published along with stocks as a direct measure of population change.

Stocks are the number of individuals in a population at a given time point. Population change can be measured as the change in stocks from time one to time two.

Flows are the movements of individuals in and out of a population over a given time period. These can be inflows (individuals joining a population), outflows (individuals leaving a population), or net flows (the overall gain of individuals to a population, calculated as inflows minus outflows). Flows provide a direct measure of change in a population, as they measure the change in population size in any given time period.

Stocks and flows are closely related. In a given population, population stock at time two minus population stock at time one will be equal to the net flow for the period from time one to time two, plus natural increase (births minus deaths) and net overseas migration.

Where available, this project evaluated the quality of both stock and flow information for each of the data sources.

Metadata collection

A key aspect of this project was to understand the metadata – that is, how data providers collect, process, update, and maintain the data. This metadata can provide insight into the quality of the dataset. Metadata collection for this project was done in two stages, using a modified version of the Statistics NZ metadata template as a guide.

First, information was sourced from publicly available sources including data dictionaries, data guides, and web pages.

Second, data providers were consulted to verify the accuracy of information collected so far, and to answer any outstanding questions.

Coverage assessment

One way to examine coverage in an administrative dataset is to compare the administrative data population for a given date with the estimated resident population (ERP) for that date. This comparison may reveal undercoverage (where the administrative stock total is lower than the population estimate) or overcoverage (where the administrative stock total is greater than the population estimate).

For this project, comparisons between the administrative dataset and the estimated resident population were made using the most recent data available. For some datasets, although 2012 data are available, administrative data for the June 2012 year were used to calculate the ERP for that year (this applies to PHO and IR tax flow data, and PHO and school roll returns stock data). In these cases, comparing 2012 administrative data populations with 2012 ERP may result in artificially high coverage rates. To avoid this, coverage rates for these data sources were calculated using 2011 data.

The only exceptions are for PHO and roll return stock data, where 2011 data were not available. In these cases, 2012 data were used, even though these data were used to generate 2012 population estimates at the area unit level. It is therefore likely that area unit coverage rates for PHO and roll returns stocks appear higher than they would if 2011 data was used. This should be kept in mind when interpreting the area unit coverage rates for PHO and school roll returns stock data.
Criteria for assessing quality

Administrative data sources must be evaluated for their potential to produce accurate population estimates. This section discusses the quality measures that were used for this evaluation, and explains the features that distinguish a high quality data source from a poor quality data source for this purpose.

The quality measures in this project were drawn from major quality models and a set of quality guidelines used internationally in official statistics (Zhang 2012; ONS 2012; Statistics Canada, 2009; Thomson 2010). They are similar to those used in previous Statistics NZ work assessing administrative data sources (O’Byrne, 2012).

Key variables

For a data source to be used to produce subnational population estimates, it should contain several key variables.

Subnational population estimates are produced by age and sex, at the territorial authority and area unit levels. Therefore, administrative data source should ideally contain information about age, sex, and geographic location. Single year of age information is valuable as territorial authority estimates are produced by single year of age.

Accurate geographic information is needed to produce accurate subnational population estimates, so it is important to consider the quality of the geographic information in a data source. Geographic information should be available to at least the area unit level of detail.

It is advantageous if a dataset also contains address change information that can be used to extract a measure of annual flows.

Relevance of target population

A relevant administrative population will cover a population that is very similar to that included in the estimated resident population (ERP). Ideally it would include all individuals who are resident in New Zealand, and it would not include any individuals who are not living in New Zealand or are visiting on a temporary basis.

Coverage assessment

National stocks

National coverage was assessed by comparing national stock totals from an administrative dataset with the national ERP at a given date. Where a dataset covers only a section of the age range (such as school roll return data), coverage is assessed by comparing with ERP for the relevant ages only. National coverage rates are examined by age and sex to determine whether there are any age- 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 areas of undercoverage (where administrative data stocks are lower than ERP, coverage rate <100 percent) or overcoverage (where administrative data stocks are higher than ERP, coverage rate >100 percent).

While a small amount of overcoverage or undercoverage is acceptable, extensive under- or overcoverage will limit the utility of a dataset. While there is no clear cut-off for what level of under- or overcoverage is acceptable, for the purposes of this project datasets with less than 85 percent coverage or greater than 115 percent coverage are considered unlikely to be useful for population estimates in their present form.
It should be noted that within any given dataset there may be areas of overcoverage and areas of undercoverage. In some cases, high coverage rates may disguise large pockets of overcoverage (with duplicate enrolments) and undercoverage (with missing individuals) that cancel each other out when the total coverage rate is considered.

**Subnational stocks**

The major focus of this project is on subnational population estimation, therefore it is important to consider whether coverage varies across territorial authorities and area units.

As for national coverage, a dataset will likely have some amount of undercoverage or overcoverage at the subnational level. Coverage rates are likely to be more variable at the area unit level than at the territorial authority level due to the small population sizes of some area units. Ideally, the majority of territorial authorities or area units should have coverage rates between 85 percent and 115 percent. If a large proportion of area units have coverage rates outside of these boundaries, the dataset is unlikely to be useful for population estimates at the territorial authority or area unit level.

**Flows**

Assessing coverage for flow data is more difficult than for stock data because there is no 'true' flow measure to compare it with. However, census data can provide some indication of what level of internal migration might be reasonable.

Data from the 2006 Census show that around 25 percent of individuals aged over one year old have been living at their current address for less than one year (suggesting they have moved at least once within the last year). The true rate of internal migration is likely to be slightly lower than this, as the census rate includes individuals whose last address was overseas.

Census data also show that around 50 percent of individuals changed usual residence within New Zealand between 2001 and 2006. If we assume that flows are approximately evenly distributed across the five-year period, this provides a lower limit of around 10 percent of the population changing residence per year. The actual figure is likely to be higher than this, as within the five-year intercensal period individuals may make multiple moves, or move away and then return to the same residence.

While these sources of information cannot provide an exact flow rate for comparison, they suggest that we would expect between about 10 percent and 25 percent of individuals in a dataset to move within New Zealand each year.

At the national level, the quality of flow data can also be examined by looking at age and sex trends in flow rates. Figure 1 shows the proportion of individuals who have been living at their current usual residence for less than one year, by age and sex (sourced from the 2006 Census).

The figure shows that flow rates vary with age. The proportion of individuals who have lived in their current usual residence for less than a year is high for one-year-olds, but declines throughout childhood until around age 16, when it rises sharply. The highest flow rates are observed in the 18–25-year age groups, and they gradually decline from age 25 onwards, rising slightly again after age 80.

Flow rates are very similar for males and females, with around 25.1 percent of males and 24.5 percent of females in the 2006 Census reporting that they have been living at their current usual residence for less than a year. We would expect to see similar age and sex patterns in flow data from an administrative source.
Evaluation of administrative data sources for subnational population estimates

Figure 1

Percentage of individuals who have been at their current usual residence less than one year
By age and sex
2006 Census

![Percentage of individuals who have been at their current usual residence less than one year](image)

Source: Statistics New Zealand

It is very difficult to assess flow coverage rates for subnational datasets due to the lack of a comparison measure for subnational areas.

We can, however, look for characteristic patterns of subnational movement to evaluate the accuracy of flow data. For example, university areas such as Dunedin and Palmerston North tend to have large net inflows of individuals aged 18–19 each year (students arriving to attend university), and large net outflows of individuals aged 21–22 each year (students leaving after completing their university studies).

Major urban areas such as Auckland, Wellington, and Christchurch tend to have large inflows of young adults (ages 18–25) who arrive seeking employment. Conversely, provincial areas tend to have large net outflows of individuals in the young adult ages, as young adults leave to travel to major centres for tertiary study or work – although net flows can be difficult to assess accurately in these territorial authorities due to small population sizes.

**Timeliness**

**Availability**

Subnational population estimates measure the population at 30 June each year. Ideally, updated administrative data counts will also be available at or around 30 June each year. Data should be available by age, sex, and detailed geographic location (at least area unit, but preferably meshblock).

Data sources need to be available for ongoing use by Statistics NZ. In some cases, Statistics NZ may already have supply agreements in place with the data providers; in others these agreements may need to be negotiated.

**Lag**

It is reasonable to expect a delay between when an individual changes address and when this address change is recorded in an administrative dataset. However, if the delay
is substantial then the dataset may not provide accurate subnational population estimates.

Lags are difficult to measure without knowing the exact date that an address change occurred. However, it is possible to gain some insight into lags by considering how data are collected and updated.

Datasets in which address information is checked regularly are likely to contain up-to-date addresses. Datasets where address updating occurs voluntarily are more likely to be out of date, especially if individuals do not have regular interactions with the data supplier.

It is also useful to know whether up-to-date addresses are a key requirement for the data supplier. If the data supplier has a strong need for up-to-date addresses, they will make an effort to maintain and update address information.

The Christchurch earthquakes also provide an opportunity to examine lags in the dataset. Following the February 2011 earthquake, there were a large number of moves away from Christchurch. By examining the number of recorded moves out of Christchurch in the datasets, we can see how long it takes for a spike in address changes to occur.

**Linkage potential**

While population estimates can be generated using individual datasets, or using aggregate totals from several different datasets, greater accuracy may be achieved by linking several datasets at the unit record level. This approach has been used to build large integrated datasets such as the IDI.

To enable linking at the unit record level the data must contain matching variables. For exact (deterministic) linking, the data must contain a unique identifier that is available for all individuals and has few duplicates. In the absence of a unique identifier, deterministic or probabilistic matching can be undertaken by matching variables such as name, date of birth, sex, and address of usual residence.

**Quality summary**

The quality of each dataset was summarized using nine indicators.

The first two indicators identify the potential age range of the dataset (noting any areas of especially low coverage) and state whether age and sex information is available.

The next seven indicators summarise the quality of the dataset using the categories:

- quality of geographic information
- relevance of target population
- accuracy (national)
- accuracy (subnational)
- accuracy (flows)
- linkage potential
- timeliness (availability to Statistics NZ)
- timeliness (lag).
Each indicator was given a rating of ‘excellent’, ‘good’, ‘moderate’, or ‘poor’. The broad definitions of these ratings are:

- excellent – no significant problems with this dataset
- good – some minor quality problems, but there is considerable potential for these to be improved
- moderate – several quality problems, and potential for improvement is limited
- poor – major quality problems, with little or no potential for improvement.
4 Results

Primary health organisation (PHO) enrolment data

**Key variables**

The dataset contains:
- date of birth
- sex
- residential address geocoded to meshblock.

Patients provide their residential addresses when they enrol. These addresses are geocoded to meshblock by GP practices using geocoding software provided by the Ministry of Health. In 2012, around 6 percent of addresses could not be geocoded to a meshblock.

Information about flows for 2012 can be obtained from the PHO dataset by comparing geocoded meshblock at time one with geocoded meshblock at time two to identify individuals who have changed meshblock during the year.

Statistics NZ currently receives data on quarterly flows from the PHO dataset. These have been summed to produce a measure of annual flows. However, this measure is likely to overestimate the true level of annual flows in the PHO dataset, as individuals moving twice within the same year will be counted twice in the annual total.

**Relevance of target population**

The target population for the PHO dataset is all individuals enrolled in PHOs. Most individuals who are included in the ERP are also eligible to enrol in a PHO. However, there are some exceptions, including:
- military personnel, defence force personnel, and prisoners, who receive their healthcare from other sources
- individuals on international student visas
- international visitors who are on working visas of less than two years’ duration.

Eligible patients have a strong incentive to enrol in a PHO because they pay lower consultation fees as an enrolled patient. However, there are several situations in which eligible patients may not enrol with a PHO:
- a patient attends a GP who is not registered with a PHO. PHO registration is not compulsory for GPs, but it is the best way for most GPs to receive funding, so most New Zealand GPs are registered
- a patient chooses, for privacy or other reasons, not to enrol with a PHO
- an individual has not used GP services in the last three years and is therefore not enrolled with a PHO.

**Coverage assessment**

**National stocks**

Nationally, the PHO dataset contains a total population that is similar to ERP. In 2012, the PHO dataset contained a national total of 4,257,917 individuals, representing 96.1 percent of ERP for that year.
Figures 2 and 3 show PHO coverage (relative to ERP) by age and sex for 2012. The figures show that coverage varies by age and sex.

Coverage is high in the older age groups (55 and over) for both sexes. Amongst children (below age 15) coverage is also high, with a small amount of overcoverage amongst the youngest ages. Undercoverage is apparent in the young adult and adult ages (15–34) for both sexes, although males have especially low coverage (below 80 percent for males aged 20–29).

**Figure 2**

**National PHO enrolments and ERP**

*By age, males 30 June 2012*

**National PHO enrolments and ERP**

*By age, females 30 June 2012*

Source: Statistics New Zealand
Subnational stocks

PHO coverage varies between different territorial authority areas. Figure 4 shows the coverage rate (as a percentage of ERP) for different territorial authorities in 2012. The figure shows that, overall, there is undercoverage in most territorial authorities, with very few having coverage rates above 100 percent.

Undercoverage tends to be greater in largely rural territorial authorities such as Westland, Ruapehu, Rangitikei, Manawatu, Opotiki, and Selwyn, which all have coverage rates below 85 percent. This may be due to geocoding difficulties with rural addresses, which are more likely than urban addresses to be missing from subnational counts because they could not be geocoded.

Coverage is also low in Dunedin and Palmerston North, and may be related to the large proportions of university students in these areas. Students may use GP services less frequently than other groups, or may be enrolled with GPs in other areas, such as where their parents live.

In Dunedin, coverage rates amongst students are further reduced because the student health service does not belong to a PHO.
Figure 4
PHO coverage rates, by territorial authority, 2012

Source: Statistics New Zealand
At the area unit level, coverage varies substantially and may be explained by a range of factors including age and sex profiles, geocoding difficulties, socioeconomic profiles, and idiosyncratic factors such as a local GP not belonging to a PHO.

Table 1 shows the level of under- or overcoverage (relative to ERP) for area units in 2012. The table shows that more than one-third of all area units (34.4 percent) have good coverage rates, with stock populations that are within 5 percent of ERP. Very few area units have overcoverage, with only 1.8 percent of area units having coverage rates above 105 percent. Most area units have some level of undercoverage, with almost a quarter of area units (23.4 percent) having coverage rates below 85 percent.

**Table 1**

<table>
<thead>
<tr>
<th>Coverage rate (%)</th>
<th>% of area units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 85</td>
<td>23.4</td>
</tr>
<tr>
<td>85–94.99</td>
<td>40.4</td>
</tr>
<tr>
<td>95–104.99</td>
<td>34.4</td>
</tr>
<tr>
<td>105–114.99</td>
<td>1.3</td>
</tr>
<tr>
<td>115 or over</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Flows**

A total of 646,759 meshblock changes were recorded during the year to 30 June 2011. This represents 15.3 percent of the PHO stock population. Of these, 555,492 involved a change of area unit and 211,288 involved a change of territorial authority.

Figure 5 shows the number of meshblock changes in 2012, by age and sex. The distribution of flows across age groups is similar to what would be expected based on census data (see figure 1). There is a high flow rate for young infants, which decreases throughout childhood. Flows rise sharply after age 16, peaking around age 17–20, and gradually decline throughout adulthood.

The figure also shows more flows were recorded for females than for males in the PHO dataset, particularly in the young adult ages. This can be explained, in part, by the larger number of females in the dataset (due to higher female coverage rates).

However, when flows are calculated on a per-person basis they remain higher for females than for males, suggesting that males may be less likely to report their address changes than females. This may be because males have less frequent attendance at a GP than females do, particularly during the young adult ages, providing them with fewer opportunities to update their address.

---

6 PHO area unit stock data were used to produce area unit population estimates in 2012. As a result, coverage rates may be closer to 100 percent than would be expected if stock data had not been used to produce the area unit population estimates. See Methods for a more detailed explanation of the datasets used to calculate coverage rates.

7 Of a total of 1827 area units (area unit total excludes area units with zero population and those outside territorial authority boundaries).
Figure 6 shows the net internal migration from 2012 PHO enrolment data, by age, for a sample of territorial authorities. In general, the figure shows that net internal migration patterns in the selected territorial authorities were similar to what would be expected.

Major urban areas (Auckland, Wellington) had a net gain of young adults aged 18–22, as young adults move to these areas to study or seek employment.

Predominantly rural areas (Hastings and Marlborough) had net losses of young adults over the ages 18–22 as they move to urban areas for study or work.

Palmerston North, a university area, had a large net gain of young adults aged 18–20 (as students arrive to attend university), and a small net loss of students several years older, who have completed their studies and have left.

However, this pattern was not apparent in Dunedin, where we would expect a similar net gain, and then loss, of young adults. The poor coverage of young adult flows to and from Dunedin is thought to occur because the student health service at the University of Otago (used by many of the young adults in Dunedin) is not a member of a PHO, so young adults enrolling with this GP service are not captured in the PHO dataset.
Figure 6

**Net internal migration for selected areas in PHO enrolment data**
By territorial authority and age

Auckland city

Hastings district

Palmerston North city

Wellington city

Marlborough district

Dunedin city

Source: Statistics New Zealand
Timeliness

Availability
The dataset is updated at the start of each quarter (1 January, 1 April, 1 July, and 1 October). Data are available from the Ministry of Health approximately six weeks after the update. Statistics NZ has a memorandum of understanding in place for the quarterly supply of flow and stock data from the ministry.

Lag
Records for new patients and record changes for existing patients appear in the quarterly update following the record change. However, address updating is voluntary and it is likely that different GP practices vary in how often they check patient addresses. There may be some lag between when an individual changes address and when this change is reported to the GP (and thus updated in the dataset). This lag may be greater for individuals who visit their GP infrequently, such as young adult males.

Figure 7 shows the number of moves out of Christchurch (that is, moves from Christchurch to another territorial authority) recorded in PHO data for each quarter from 2009 to 2012. The dotted vertical line represents the approximate time of the February 22 earthquake.

The figure shows a large increase in moves away from Christchurch in the quarter immediately following the earthquake (quarter ended April 2011). This suggests that most moves out of Christchurch after the February 2011 earthquake were picked up in the PHO enrolment dataset with very little delay.

Figure 7

Moves out of Christchurch, by quarter
PHO data

Source: Statistics New Zealand

Linkage potential
The dataset contains National Health Index (NHI) numbers which enable linking with other health datasets. Name, sex, date of birth, and meshblock of usual residence are also available.

Quality summary
Table 2 summarises the results of the quality assessment of the PHO dataset.
The strengths of the PHO dataset include:

- a relevant target population
- high level of national coverage
- detailed geographic information.

The weaknesses of the PHO dataset include:

- low coverage in the young adult ages, particularly for males
- addresses may be out of date for individuals who do not visit a GP regularly
- subnational coverage rates are variable, and undercoverage is common, especially at the area unit level.

### Table 2
#### Quality summary of the PHO dataset

<table>
<thead>
<tr>
<th>Quality measure</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suitable age range</td>
<td>0–85+</td>
</tr>
<tr>
<td>Age and sex</td>
<td>Yes</td>
</tr>
<tr>
<td>Quality of geographic information</td>
<td>Good</td>
</tr>
<tr>
<td>Relevance of target population</td>
<td>Good to excellent</td>
</tr>
<tr>
<td>Accuracy – national coverage</td>
<td>Moderate (age 17–30), excellent (other ages)</td>
</tr>
<tr>
<td>Accuracy – subnational coverage</td>
<td>Good</td>
</tr>
<tr>
<td>Accuracy – flows</td>
<td>Good</td>
</tr>
<tr>
<td>Linkage potential</td>
<td>Good</td>
</tr>
<tr>
<td>Timeliness – availability</td>
<td>Good</td>
</tr>
<tr>
<td>Timeliness – lag</td>
<td>Good</td>
</tr>
</tbody>
</table>

**Inland Revenue (IR) tax data**

Several components of IR tax data can be used to estimate population. The tax client register is an exhaustive list of all individuals who have ever held an IRD number. It contains personal details for each of these individuals, including age, sex, and address.

Employer Monthly Schedule (EMS) tables are monthly updates that record all individuals who paid tax at source during the month (from wages, salaries, benefits, but not self-employment).

Information about self-employed individuals can be sourced from annual tax return tables, which contain income details for all individuals who file a tax return in a given year.
Key variables
The dataset contains:

- date of birth
- residential address – recorded by Inland Revenue in free text format
- sex, which is derived based on title and first name.

The IR tax data contains territorial authority information from a geocoding method based on postcode. This geographical information does not come from IR, but is added as part of LEED processing. However, many postcodes do not map directly to territorial authorities, and therefore the territorial authority information is of limited use for population estimates purposes.

To generate more useful geographic information, Statistics NZ extracted raw address data from the IR tax dataset and geocoded it in-house (see Methods for a more detailed description of this process). Almost 99 percent of records could be allocated to a territorial authority and around 84 percent to an area unit using this method.

Information about address changes can be extracted from IR tax data by comparing addresses at the start and end of a period of interest. Addresses are geocoded to meshblock where possible, and any meshblock changes are identified as flows.

Relevance of target population
Different components of IR tax data have different target populations.

The target population for the IR tax client register is all individuals who have ever held an IRD number. This population is likely to contain considerable overcoverage relative to ERP, as it contains large numbers of individuals who are no longer living in New Zealand.

The target population for EMS data is narrower, and consists of all individuals who receive income that has tax deducted at source. This includes:

- wage and salary earners
- individuals receiving social assistance payments such as paid parental leave, student allowances, benefits, pensions, and ACC payments
- overseas (i.e. non-resident) contractors and beneficiaries.

However, it excludes individuals who receive income from self-employment, and also individuals who have not paid tax within the given EMS time period.

The target population for the annual tax return dataset is all individuals who have filed a tax return in any given year. This is likely to include most self-employed individuals, and some wage and salary earners. However, anyone who has not filed an annual tax return in a given year is excluded from the target population. This is likely to include many individuals who pay tax at source, along with individuals who have not paid tax in a given year.

One strategy for generating a national population from all those registered with IR is to combine the EMS and annual tax return populations for the previous 12 months. Personal details including age, sex, and address can then be sourced from the client register. This strategy will capture all individuals who have paid tax at source during the previous 12 months, plus all self-employed individuals who filed a tax return for the previous financial year.

This strategy will likely lead to some undercoverage of ERP (as it will exclude individuals who have not received taxable income in the last 12 months), but it will yield a population that is conceptually closer to ERP than the EMS, client register, or annual tax return alone.
While there are other ways that the client register could be restricted, we chose this option as it maximises coverage of the estimated resident population while minimising undercoverage. Methods that use a longer time period would produce greater levels of overcoverage, while using a shorter time period or including fewer types of income would increase undercoverage.

When constructed using the strategy outlined above, the IR tax dataset has high coverage for most of the population. However, there are several groups of individuals who would be included in ERP, but are not included in the tax dataset:

- children, unless they are employed
- individuals who have not been in the labour force, or received a benefit or allowance, in the previous 12 months (such as international or domestic students, or parents who are caring for young children).

In addition, individuals are included in the IR tax population who would not normally be included in ERP. Individuals who have paid tax within the previous 12 months but have since moved overseas will appear in the tax population. This could include former New Zealand residents and temporary workers from overseas.

**Coverage**

**National stocks**

At a national level, the IR tax dataset has substantial overcoverage compared with ERP: in 2012 there were more than 15 million individuals in the tax client register, compared with an ERP of 4.5 million.

The overcoverage is probably due to records remaining in the database for individuals who are no longer residing in New Zealand, such as New Zealand residents who have moved overseas, and visitors to New Zealand who worked for a short time and then left. These individuals would be removed from ERP.

Because of the substantial overcoverage in the client register, it is necessary to restrict the dataset to individuals who have paid tax in the last 12 months (see above for a more detailed description of this process). For 2012, restricting the dataset in this way generates a national total of 3,167,161, which represents 94.3 percent of ERP for ages 18 and over. The rest of the analyses in this paper focus on the restricted IR tax dataset.

Figures 8 and 9 show the coverage for this restricted IR tax population (relative to ERP), by age and sex. The figures show some undercoverage for individuals aged under 20, males aged 50 to 65, and females aged 25 to 65. Undercoverage is more substantial for females than for males. This sex difference is likely because more females than males are out of the labour force, particularly during the primary childbearing ages.
Figure 8

**National IR tax enrolments and ERP**
**By age, males**
30 June 2012

![Graph showing IR tax enrolments and ERP for males by age.](image)

Source: Statistics New Zealand

Figure 9

**National IR tax coverage, relative to ERP**
**By age and sex**
30 June 2012

![Graph showing IR tax coverage relative to ERP by age and sex.](image)

Source: Statistics New Zealand

**Subnational stocks**

Figure 10 shows the level of IR tax coverage for each territorial authority in 2012. The figure shows that most territorial authorities have some undercoverage relative to ERP, with Southland having the lowest coverage rate (73.3 percent). The only territorial authority to have substantial overcoverage was Gore district, with a coverage rate of 112.0 percent.
Predominantly rural territorial authorities tended to have high levels of undercoverage, with Southland, Clutha, Westland, Waimate, Selwyn, Manawatu, Otorohonga, Waikato, Western Bay of Plenty, Kaipara, and Rodney all having coverage rates below 85 percent.

Dunedin and Wellington city also had coverage rates below 85 percent, which may be caused by large student populations in these areas failing to update their addresses with IR.

The undercoverage in some areas may be due to the geocoding process. Around 15 to 20 percent of IR tax addresses could not be geocoded to meshblock. These addresses were allocated to territorial authorities by using matching techniques based on postcode or place name, both of which can introduce biases into the coding. This is especially true for rural territorial authorities that share borders with large city territorial authorities, where city postcodes or place names may assign an address to the city territorial authority rather than the correct rural territorial authority.
Figure 10
IR tax coverage rates, by territorial authority, 2012

IR tax coverage rate (% of ERP)
- Less than 85%
- 86–94.89%
- 95–104.96%
- 105–114.99%

Source: Statistics New Zealand
Table 3 shows the coverage rate for the IR tax population relative to ERP for area units in 2012. The table shows that there is substantial undercoverage at the area unit level, with 60.1 percent of area units having coverage rates below 85 percent. Only 11.2 percent of area units have coverage rates that are within 5 percent of ERP. This undercoverage is due to the high level of missing data at the area unit level, as around 15 to 20 percent of addresses could not be geocoded to an area unit.

### Table 3
**IR tax coverage rates for area units**, 2012

<table>
<thead>
<tr>
<th>Coverage rate (%)</th>
<th>% of area units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 85</td>
<td>60.1</td>
</tr>
<tr>
<td>85–94.99</td>
<td>26.4</td>
</tr>
<tr>
<td>95–104.99</td>
<td>11.2</td>
</tr>
<tr>
<td>105–114.99</td>
<td>1.5</td>
</tr>
<tr>
<td>115 or greater</td>
<td>0.9</td>
</tr>
</tbody>
</table>

**Flows**

Around 508,799 meshblock flows were recorded in the IR tax dataset in the year to June 2011. This represents 16.1 percent of the total stock population for 2011. Of these, 266,062 involved a move from one area unit to another area unit, and 161,650 involved a move from one territorial authority to another territorial authority.

Figure 11 shows the number of territorial authority changes recorded in IR tax data, by age and sex, for the year to June 2011.

The pattern of flows by age is similar to that observed in census data (see figure 1) for ages 18 and over. Flows increase from age 18 and peak at around ages 20–24, then gradually decrease from age 25 onwards. The peak age for address changes in IR tax data (20–24) is slightly later than the peak age in census data (17–25). This may be due to low rates of coverage in the 18–21-year age group in IR tax data, or may reflect lags in the tax data.

At all ages, the number of territorial authority transitions recorded in IR tax data was similar for males and females.

---

8 Of a total of 1,827 area units (area unit total excludes area units with zero population and those outside territorial authority boundaries).
Figure 12 shows the net internal migration flows, by age, for a sample of territorial authorities in 2011. In general, the figure shows that net migration flows in subnational areas are similar to what would be expected.

In university areas (Palmerston North and Dunedin) there is a net gain of young adults aged 18–20 (as students arrive to attend university) and a net loss of young adults several years later (as students finish their studies and leave).

In major urban areas (Auckland and Wellington) there is a net gain of young adults aged 18–24, as individuals move to the city to attend university or seek employment. Conversely, in predominantly rural areas (Hastings and Marlborough) there is a net loss of young adults aged 18–24, as these individuals move to other areas for study or employment.
Figure 12

Net internal migration for selected areas in IR tax data
By territorial authority and age
2011

Source: Statistics New Zealand
Timeliness

Availability
EMS tables are updated on a monthly basis and extractions can be made at any time. Historical tax data are available through the linked employer-employee dataset back to 1999.

Tax information for self-employed individuals is only available 18 months after the date of interest. Therefore, when restricting the IR tax client register to those individuals who have paid tax in the previous 12 months, it is necessary to wait 18 months for tax information to become available for self-employed individuals. To avoid this delay, an alternative is to use the most recent list of self-employed taxpayers as a proxy for the current year’s list.

Statistics NZ has well-established agreements in place for the regular supply of IR data. Access to IR tax data via LEED is available to Statistics NZ staff by application to the LEED team.

Lag
Residential addresses are collected at the time of first contact with IR, and are updated by individuals on a voluntary basis. Address details are likely also checked when individuals contact IR. These processes may result in lags because not all individuals will inform IR when they change address. It has been estimated previously that around 50 percent of IR addresses are updated within 7 months of an address change and around 75 percent are updated within 13 months (Graham, 2008).

The extent of lag probably varies between different groups of individuals. Salaried employees have fewer interactions with IR and therefore may take a long time to report a change of address. Self-employed individuals have more frequent interactions with IR and are required to provide up-to-date address details, so address changes will be reflected in the database more quickly for this group.

Figure 13 shows the number of moves out of Christchurch (that is, moves from Christchurch to another territorial authority) recorded in the IR tax dataset for each quarter from 2009 to 2012. The vertical dotted line shows the approximate time of the February 22 earthquake.

The figure shows a sharp increase in the number of moves away from Christchurch in the quarter ended June 2011 (the second quarter following the earthquake). This suggests that most moves away from Christchurch after the February 2011 earthquake were picked up in the IR tax dataset with only a slight delay.
**Linkage potential**

IR tax records contain IRD number, which enables linking to other datasets containing IRD number. The IR database also contains name, date of birth, and sex (imputed).

Although address of usual residence is available, the large number of addresses that are missing or unable to be geocoded would limit the use of address as a matching variable.

**Quality summary**

Table 4 summarises the quality assessment of the IR tax dataset.

The strengths of the tax dataset include:
- its high population coverage for adults
- its potential for linking with other datasets
- the availability of individual-level data and customized time periods.

The weaknesses of the tax dataset include:
- poor coverage for individuals aged under 18
- around 16 percent of addresses cannot be geocoded to area unit level. This creates substantial undercoverage at the area unit level. Additional matching has been undertaken to assign uncoded records to territorial authorities, but this introduces some coverage biases at the territorial authority level
- address details may be out of date for individuals who do not have regular contact with IR.
Table 4
Quality summary of IR tax dataset

<table>
<thead>
<tr>
<th>Quality measure</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suitable age range</td>
<td>20–85+</td>
</tr>
<tr>
<td>Age and sex</td>
<td>Yes</td>
</tr>
<tr>
<td>Quality of geographic information</td>
<td>Moderate (if coded in-house)</td>
</tr>
<tr>
<td>Relevance of target population</td>
<td>Good (after refinements)</td>
</tr>
<tr>
<td>Accuracy – national coverage</td>
<td>Good</td>
</tr>
<tr>
<td>Accuracy – subnational coverage</td>
<td>Moderate (territorial authority level) to poor (area unit level)</td>
</tr>
<tr>
<td>Accuracy – flows</td>
<td>Good</td>
</tr>
<tr>
<td>Linkage potential</td>
<td>Excellent</td>
</tr>
<tr>
<td>Timeliness – availability</td>
<td>Excellent</td>
</tr>
<tr>
<td>Timeliness – lag</td>
<td>Moderate</td>
</tr>
</tbody>
</table>
School roll returns data

**Key variables**

The July school roll returns dataset contains:

- sex
- date of birth (although Statistics NZ currently only receives year and month of birth).

Two types of geographic information are recorded in the July roll returns: information about the student’s residential address, and information about the location of the school attended. This geographic information is sourced from school Student Management Systems.

Information about school location is geocoded and readily available, but this information is not useful for informing subnational population estimates because it may differ from residential address.

Information about student residential address is not routinely geocoded as part of July roll returns processing. However, the information is being geocoded by another team within the Ministry of Education and the geocoded information is currently available to Statistics NZ.

Around 96 percent of student residential addresses are able to be coded to a meshblock.

Flow information is not currently available from the roll returns dataset.

**Relevance of target population**

The target population for July roll returns is all students currently enrolled in New Zealand schools.

In theory, this includes all New Zealand residents within the compulsory school ages (6–15), plus individuals who are voluntarily attending New Zealand schools (such as New Zealand residents outside the compulsory school ages and non-residents attending New Zealand schools).

A small number of New Zealand residents within the compulsory school ages do not attend school, and these individuals will not be recorded in the dataset (Leather, 2009). Nonetheless, the target population for school enrolment data is conceptually very close to ERP for the relevant ages.

**Coverage**

**National stocks**

If the population is restricted to the compulsory school ages (6 to 16), the roll returns dataset contained a national total of 630,703 individuals in July 2012. This represented 98.7 percent of the national ERP at 30 June 2012.

Figures 14 and 15 show the coverage of the roll returns data (relative to ERP), by age and sex, at 30 June 2012.

The figures show that coverage in roll returns data is high between the compulsory school ages of 6 and 16. For ages 6 to 15, coverage is close to 100 percent. Coverage is slightly lower at ages 5 and 16, but still exceeds 85 percent. After age 16, however, coverage rates decline rapidly, with almost no coverage by age 20.

At all ages, coverage rates are very similar for males and females.
Figure 14

**National July roll returns enrolments and ERP**
By age, males 2012

**National July roll returns enrolments and ERP**
By age, females 2012

Source: Statistics New Zealand

Figure 15

**National July roll returns coverage, relative to ERP**
By age and sex 2012

Source: Statistics New Zealand
Subnational stocks
The most accurate and relevant address information comes from the geocoded residential addresses from the July roll returns dataset.

Figure 16 shows the coverage rate (as a percentage of ERP) for each territorial authority at 30 June 2012. The figure shows that most territorial authorities had some level of undercoverage.

Coverage tends to be lower in the South Island than in the North Island, and is particularly low in territorial authorities in the West Coast, Otago, and Tasman regions, and in Kaikoura, Ruapehu, Carterton, and Kaipara territorial authorities.

Coverage is higher in cities and regional main centres, with Gore, Nelson city, Napier, Auckland city, Christchurch, and Invercargill all having coverage rates within 5 percent of ERP.
Figure 16
July roll returns coverage rates, by territorial authority, 2012

Roll return coverage rate (% of ERP)
- less than 85%
- 85–94.99%
- 95–104.99%
- 105–114.99%

Source: Statistics New Zealand

0 60 100 200 kilometres
At the area unit level, there was a mix of overcoverage and undercoverage relative to ERP in 2012. Table 5 shows the roll returns coverage rates for area units.

Overall, more than one-third (35.6 percent) of area units had stock populations that were within 5 percent of ERP. Just over half of all area units (51.6 percent) had some level of undercoverage, while 12.9 percent of area units had overcoverage (coverage rates greater than 105 percent).

Table 5
Roll returns coverage rates for area units, 2012

<table>
<thead>
<tr>
<th>Coverage rate (%)</th>
<th>% of area units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 85</td>
<td>21.8</td>
</tr>
<tr>
<td>85–94.99</td>
<td>29.8</td>
</tr>
<tr>
<td>95–104.99</td>
<td>35.6</td>
</tr>
<tr>
<td>105–114.99</td>
<td>10.2</td>
</tr>
<tr>
<td>115 or greater</td>
<td>2.7</td>
</tr>
</tbody>
</table>

**Timeliness**

**Availability**
July roll returns relate to a reference date near the end of June and are released in October each year. Geocoded information from the July roll returns is sourced from a separate team within the Ministry of Education and may not be available until several months after the roll return release date.

**Lag**
Information for roll returns is sourced from student management systems, which are likely to be up-to-date and accurate. Residential address is usually updated at the start of each year, and may also be updated voluntarily throughout the school year if students change address. Schools need to correspond with students regularly and therefore have an incentive to keep residential addresses updated.

**Linkage potential**
The roll returns dataset contains national student number, a unique identifier for students attending New Zealand schools. The datasets also contain sex, date of birth (although currently Statistics NZ only receives year and month of birth), and meshblock of usual residence.

---

9 Roll return area unit stock data were used to produce area unit population estimates in 2012. As a result, coverage rates may be closer to 100 percent than would be expected if stock data had not been used to produce the area unit population estimates. See Methods for a more detailed explanation of the datasets used to calculate coverage rates.

10 Of a total of 1,827 area units (area unit total excludes area units with zero population and those outside territorial authority boundaries).
Quality summary

Table 6 shows the summary of the quality assessment of the roll returns dataset.

The strengths of the roll returns dataset include:
- a relevant target population
- high coverage rates.

The weaknesses of the roll returns dataset are:
- it covers a narrow age band (ages 5–16)
- the availability of relevant geographic information currently relies on the activities of a separate team within the Ministry of Education, who may or may not continue to geocode student residential addresses.

Table 6

Quality summary of roll returns dataset

<table>
<thead>
<tr>
<th>Quality measure</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suitable age range</td>
<td>5–16</td>
</tr>
<tr>
<td>Age and sex</td>
<td>Yes</td>
</tr>
<tr>
<td>Quality of geographic information</td>
<td>Good</td>
</tr>
<tr>
<td>Relevance of target population</td>
<td>Excellent</td>
</tr>
<tr>
<td>Accuracy – national coverage</td>
<td>Excellent</td>
</tr>
<tr>
<td>Accuracy – subnational coverage</td>
<td>Good</td>
</tr>
<tr>
<td>Accuracy – flows</td>
<td>N/A</td>
</tr>
<tr>
<td>Linkage potential</td>
<td>Good</td>
</tr>
<tr>
<td>Timeliness – availability</td>
<td>Good</td>
</tr>
<tr>
<td>Timeliness – lag</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

Electoral enrolment data

Key variables

Electoral enrolment data contains:
- age (in 5-year age groups)
- residential address geocoded to meshblock.

It does not contain sex. It may be possible to impute sex based on title and first name, but this would have to be done in-house by Statistics NZ and was not explored as part of the current project.

The Electoral Commission (EC) invests considerable effort in geocoding. As a result, almost all addresses are geocoded to a meshblock.
In some cases, addresses recorded in the electoral roll may not represent an individual’s residential address. The EC encourages voters to register with the address they consider to be home, which may not be the same as their residential address. For example, university students may enrol using their parents’ address, even though they live in another town.

Flow information is not available from the electoral enrolment dataset at present. However, the EC has indicated that they retain some historical address details, which may allow flow information to be extracted in the future.

Relevance of target population

The target population for the electoral enrolment dataset is all individuals who are eligible to vote\footnote{To be eligible to vote in New Zealand, individuals must be: aged 18 years or older; a New Zealand citizen or permanent resident; and have lived in New Zealand for at least 12 months. In addition, individuals who are no longer living in New Zealand but wish to remain on the roll can do so if they have been in the country within the last three years (if a citizen) or within the last year (if a permanent resident). The following otherwise eligible individuals are excluded from enrolling: New Zealand citizens who are outside New Zealand and have not been in the country within the last three years (unless a public servant, a member of the Defence Force, a diplomat or NZ Trade and Enterprise employee, or accompanying dependent); permanent residents who have not been in New Zealand within the last 12 months (unless exempted as for a citizen, above); individuals who have been committed to a hospital or a secure facility under the Mental Health (Compulsory Assessment and Treatment) Act 1992 or the Intellectual Disability (Compulsory Care and Rehabilitation) Act 2003 and has been detained for a period exceeding three years; individuals sentenced to life, preventative detention, or imprisonment for a term that involves three years or more; and individuals included on the Corrupt Practices List.} in parliamentary, general, or local authority elections in New Zealand. The target population of the electoral enrolment dataset is similar to ERP for ages 18 and over. There may be some overcoverage, as individuals who have recently moved overseas may remain on the electoral roll.

Although it is mandatory for eligible individuals to be enrolled to vote, a number of eligible individuals are not on the electoral roll. In any given year, around 90 to 95 percent of eligible individuals are on the electoral roll. Non-enrolment is particularly high amongst young people and recent migrants to New Zealand (see, for example, Henderson, 2013).

Coverage

National stocks

In 2012, 3,065,752 individuals were on the electoral roll. This represents 91.3 percent of the estimated resident population aged 18 and over.

Figures 17 and 18 show the coverage of the electoral enrolment dataset (relative to ERP) for 2012, by age group. Coverage is lowest between the ages of 18 and 39. At all other ages, coverage is greater than 95 percent of ERP.
**Figure 17**

**National electoral enrolment population and ERP**  
By age  
30 June 2012

![Graph showing national electoral enrolment population and ERP by age, 30 June 2012.](image)

*Source: Statistics New Zealand*

**Figure 18**

**National electoral enrolment coverage rate, relative to ERP**  
By age  
30 June 2012

![Graph showing national electoral enrolment coverage rate, relative to ERP, by age, 30 June 2012.](image)

*Source: Statistics New Zealand*

**Subnational stocks**

Electoral roll coverage varies between different territorial authorities. Figure 19 shows the coverage rate (as a percentage of ERP) for each territorial authority in 2011. The figure shows that most territorial authorities had some level of undercoverage.

Areas with large proportions of students or young people tended to have low coverage, with Dunedin, Palmerston North, Queenstown Lakes, Auckland city, and Wellington all having coverage rates below 85 percent.

Some predominantly rural territorial authorities also tended to have low coverage, including Ruapehu, Westland, Waitomo, and Otorohanga.
Figure 19
Electoral enrolment coverage rates, by territorial authority, 2011

Electoral enrolment coverage rate (% of ERP)
- less than 95%
- 95–99.99%
- 100–104.99%
- 105–114.99%

Source: Statistics New Zealand
Table 7 shows the coverage rates (relative to ERP) for area units in 2011. The table shows that around a quarter of all area units (26.1 percent) had coverage rates within 5 percent of ERP.

There was considerable undercoverage in electoral enrolment data, with more than a quarter of all area units (28.2 percent) having coverage rates below 85 percent of ERP, and very little overcoverage, with just 0.8 percent of area units having coverage rates greater than 110 percent of ERP.

Table 7
Electoral enrolment coverage rates for area units¹², 2011

<table>
<thead>
<tr>
<th>Coverage rate (%)</th>
<th>% of area units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 85</td>
<td>28.2</td>
</tr>
<tr>
<td>85–94.99</td>
<td>44.3</td>
</tr>
<tr>
<td>95–104.99</td>
<td>26.0</td>
</tr>
<tr>
<td>105–114.99</td>
<td>1.0</td>
</tr>
<tr>
<td>115 or greater</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Timeliness

Availability
Data are updated continually and are available on request from the Electoral Commission. Data can usually be provided within a relatively short timeframe.

Lag
Records are entered and updated on an ongoing basis as new information becomes available. There is a one-month lag period before an address change can be recorded in the electoral roll. Residential addresses are recorded at the time of enrolment on the electoral roll and the EC attempts to update these before every election.

Outside of election years, the EC uses several sources to detect address changes, including voluntary updates and notifications from other agencies including the New Zealand Transport Authority, and the Ministry of Transport. However, it is likely that some addresses are out of date, and this number is likely to grow with increasing time from the last parliamentary election.

Individuals who move overseas may remain on the roll for some time after leaving New Zealand. Similarly, new migrants to New Zealand are not eligible to enrol until they have been in New Zealand for 12 months.

¹² Of a total of 1827 area units (area unit total excludes area units with zero population and those outside territorial authority boundaries).
Linkage potentials
No unique identifiers are included in the dataset, so any linking would need to be done on the basis of name, date of birth, and meshblock of usual residence.

Quality summary
Table 8 shows the summary of the quality evaluation for the electoral enrolment dataset. The strengths of the electoral enrolment dataset include:
- a relevant target population
- high national coverage rates.

The limitations of the electoral enrolment dataset include:
- age is recorded in five-year age groups, rather than single year of age
- sex is not available
- coverage is low for younger age groups (ages 18–39).
- encouraging voters to register with the address that they consider to be home may distort subnational migration patterns, particularly amongst younger age groups who may consider their home to be different from their current residential address.

Table 8
Quality summary for the electoral enrolment dataset

<table>
<thead>
<tr>
<th>Quality measure</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suitable age range</td>
<td>18–85+</td>
</tr>
<tr>
<td>Age and sex</td>
<td>Age in 5-year groups; no sex</td>
</tr>
<tr>
<td>Quality of geographic information</td>
<td>Good</td>
</tr>
<tr>
<td>Relevance of target population</td>
<td>Excellent</td>
</tr>
<tr>
<td>Accuracy – national coverage</td>
<td>Moderate (age 18–25), excellent (other ages)</td>
</tr>
<tr>
<td>Accuracy – subnational coverage</td>
<td>Good</td>
</tr>
<tr>
<td>Accuracy – flows</td>
<td>N/A</td>
</tr>
<tr>
<td>Linkage potential</td>
<td>Moderate</td>
</tr>
<tr>
<td>Timeliness – availability</td>
<td>Good</td>
</tr>
<tr>
<td>Timeliness – lag</td>
<td>Moderate to good</td>
</tr>
</tbody>
</table>
5 Discussion

This paper evaluated the quality of four administrative data sources and their potential for producing subnational population estimates. A summary of the findings from this evaluation, and the major conclusions, are outlined below.

Summary of findings

Table 9 provides a summary of the quality evaluation for the four datasets. The table shows that several quality criteria had high ratings across all four datasets. Relevance of target population and accuracy of flows data were rated at least 'good' in all four datasets. In addition, the accuracy of national coverage was rated at least 'good' in all four datasets, except for the young adult ages which were rated as 'moderate' in PHO and electoral enrolment datasets.

Overall, school roll return data received the highest ratings, being rated at least 'good' in all categories. However, school roll return data has the most limited age range of all of the datasets, covering only ages 5 to 16.

IR tax data had the lowest quality scores, with ratings of 'moderate' or lower for three quality criteria.
Table 9
Quality summary for the four administrative datasets

<table>
<thead>
<tr>
<th>Quality measure</th>
<th>PHO</th>
<th>IR tax</th>
<th>School roll return</th>
<th>EC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suitable age range</td>
<td>0–85+</td>
<td>20–85+</td>
<td>5–16</td>
<td>18–85+</td>
</tr>
<tr>
<td>Age and sex</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>5-year age groups; no sex</td>
</tr>
<tr>
<td>Quality of geographic information</td>
<td>Good</td>
<td>Moderate (if coded in-house)</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Relevance of target population</td>
<td>Good to excellent</td>
<td>Good (after refinements)</td>
<td>Excellent</td>
<td>Excellent</td>
</tr>
<tr>
<td>Accuracy – national coverage</td>
<td>Moderate (age 17–30), excellent (other ages)</td>
<td>Good</td>
<td>Excellent</td>
<td>Moderate (age 18–25), excellent (other ages)</td>
</tr>
<tr>
<td>Accuracy – subnational coverage</td>
<td>Good</td>
<td>Moderate (territorial authority level) to poor (area unit level)</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Accuracy – flows</td>
<td>Good</td>
<td>Good</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Linkage potential</td>
<td>Good</td>
<td>Excellent</td>
<td>Good</td>
<td>Moderate</td>
</tr>
<tr>
<td>Timeliness – availability</td>
<td>Good</td>
<td>Excellent</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Timeliness – lag</td>
<td>Good</td>
<td>Moderate</td>
<td>Excellent</td>
<td>Moderate to good</td>
</tr>
</tbody>
</table>

The table also points to a gap in the suitable age ranges covered by the datasets. None of the datasets has suitable cover for age 17, and no dataset has ‘good’ or better coverage for ages 18–19.

To examine this further, Figure 20 shows the national coverage rates (as a percentage of ERP), by age group, for each of the data sources.

The figure shows that few datasets have coverage between the ages of 17 and 20. School roll return and IR tax data are not available for these ages, and electoral enrolment data has a very low coverage rate. While the PHO coverage rate is above 80 percent, assessing coverage for males and females together obscures the fact that PHO coverage rates for males in this age group are especially low.

Above age 35, coverage is high in all datasets, with both electoral enrolment and PHO data having coverage rates close to 100 percent.
No single data source will be sufficient for estimating subnational populations

None of the data sources in this evaluation can generate sufficiently accurate subnational population estimates. It will likely be necessary to combine several data sources to generate accurate subnational population estimates.

One approach to combining several administrative data sources was taken in 2012 for the production of the subnational estimates. The approach involved using different data sources for different age groups, with IR tax data used for the primary working ages and PHO used at all other ages.

A weakness of this approach is that it cannot generate estimates for the young adult age group, where coverage is low in all data sources (see below). To address this issue in the 2012 subnational estimates, historical census migration patterns were used instead of administrative data for the young adult age group. Work underway at Statistics NZ using Bayesian statistical techniques to generate subnational population estimates may provide additional options for aggregate models in the future.

An alternative to an aggregate model is a unit record linked dataset, in which records from several different data sources are linked at the individual level. Further investigation is required to determine whether a unit record linked dataset could generate more accurate population estimates than an aggregate model.

Poor coverage of the young adult age group in all datasets

None of the datasets evaluated as part of this project had high coverage of the young adult age group (age 17–25). This is especially problematic given that this age group is highly mobile and accounts for a disproportionate amount of internal migration.

Subnational population estimates based on administrative data sources are likely to have lower accuracy for the young adult age group. Estimates for areas that have large flows of young adults (such as Dunedin and Palmerston North) are also likely to be less accurate under an administrative data-based model.
It may be possible to address this problem by identifying additional datasets that have high coverage for young adults. This may include tertiary education enrolment data or driver licensing data. Further evaluation is required to determine whether there are additional data sources that may be suitable for estimating young adult populations.

**Better-quality address data are needed**

Accurate and up-to-date address information is needed to produce high-quality subnational population estimates.

There are several quality issues with regard to address data in administrative datasets. In some cases, the poor quality of address data means many addresses cannot be geocoded, resulting in missing data and low coverage rates at the territorial authority and area unit levels. This is particularly problematic for rural addresses, which are more likely to fail an automated geocoding process.

Addresses can also be problematic if they are out of date. The available information about how datasets are updated and maintained suggests that there are likely to be lags in the PHO, IR tax, electoral enrolment, and roll returns datasets. However, we do not know the extent of these lags, and accurately measuring lags in a dataset is difficult without a comparison dataset containing the actual time of an address change.

Addresses are a major source of error in administrative datasets. Improving how address data is collected and processed could be one of the biggest steps to improving the quality of administrative data sources for population estimates. Strategies to improve the quality of address data could include:

- using standard address lists
- more frequent updating of addresses
- data matching to share updated addresses between agencies.

**The 2013 Census will provide an opportunity to re-evaluate coverage**

Coverage was assessed in this evaluation by comparing administrative data populations with ERP. However, there are limitations to this approach.

In some cases, administrative data have been used to generate ERP. Although comparisons between administrative datasets and ERPs based on administrative data were avoided where possible, administrative data has been used to check and guide subnational population estimates at Statistics NZ for several years. This may lead to coverage rates being higher than they would be if administrative data had not been used to generate ERP.

An additional problem relates to the accuracy of the ERP estimates being used to derive coverage rates. ERPs for 2011 are five years out from the 2006 Census population base, so there may be some inaccuracy in these estimates. A better alternative would be to compare administrative stock populations for census years with ERP counts from census year, which are likely to be more accurate than our current ERP estimates. Most datasets do not have quality data extending far enough back to compare with 2006 Census counts.

The 2013 Census will provide an opportunity to look more closely at coverage rates using 2013 Census population counts, and may give us a clearer picture of subnational coverage differences.
6 References


