Statistics New Zealand
TATAURANGA AOTEAROA

Quality of geographic information in the Integrated Data Infrastructure

Sheree Gibb and Susmita Das

New Zealand Government

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

Liability statement
While all care and diligence has been used in processing, analysing, and extracting data and information in this publication, Statistics New Zealand gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.

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

ISBN 978-0-908350-11-7 (online)

Published in December 2015 by
Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

Contact
Statistics New Zealand Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4600
www.stats.govt.nz

Contents

List of tables and figures

Tables

Figures

1 Background

Census Transformation in New Zealand

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

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

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

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

See Census Transformation in New Zealand for more information.

About this paper

The most important and fundamental reason for having a census is to provide population statistics that describe the size, structure and geographic distribution of the population. A central component of investigating a census based on administrative data is understanding whether administrative data sources can provide accurate information about where people live.

This paper examines the quality of location information in the administrative data sources held in Statistics New Zealand's Integrated Data Infrastructure (IDI) compared with the geographic information contained in the 2013 Census. This comparison will provide a better understanding of the quality of location information in government sources. Together with other Census Transformation work this will enhance our understanding of whether administrative data can meet critical information needs.

2 Introduction

Information about where people live is key to understanding population structure. In official statistics, location information is used to create regional and local population counts, and to understand the geographic distribution of other variables, such as ethnicity or income. Location information is also used to answer a wide range of policy and research questions.

The New Zealand Census of Population and Dwellings is a major source of information about where people live. The location information collected in the census is of high quality and forms the basis for population estimates and other official statistics.

Information about where people live is also contained in many administrative data sources, such as the health, tax and education data collections. Currently, this information is not widely used in official statistics, and only limited information is available about the quality of the location information in these administrative data sources. Understanding the quality of location information in administrative data sources is crucial for users of those sources.

Many national statistical agencies, including Statistics NZ, are currently attempting to increase their use of administrative data in the production of official statistics. Work underway at Statistics NZ to transform the census model relies on developing a better understanding of the quality of location information contained in administrative data sources. The next section provides an overview of the census transformation context at Statistics NZ.

Aims and scope

The work described in this paper was undertaken as part of Statistics NZ's Census Transformation project. The major aim of the work was to examine the quality of location of usual residence information in the IDI. This was done by comparing the geographic information recorded in the IDI with the geographic information recorded in the census.

The 2013 Census was linked to the IDI. This allowed us to compare the location that an individual reported in the census with the location reported to the agencies that have location information in the IDI. Census meshblocks of usual residence are recorded with high accuracy and so a comparison of an IDI meshblock against a census meshblock gives a good indication of the quality of the IDI meshblocks.

The analyses in this paper were based on the IDI as at May 2015. This paper only considers geographic information about where people live. Other types of geographic information, such as workplace address, are not considered in this paper. Furthermore, analyses in this paper are restricted to people who were usual residents of New Zealand at the time of the 2013 Census.

3 Data and methods

Data source: the Integrated Data Infrastructure

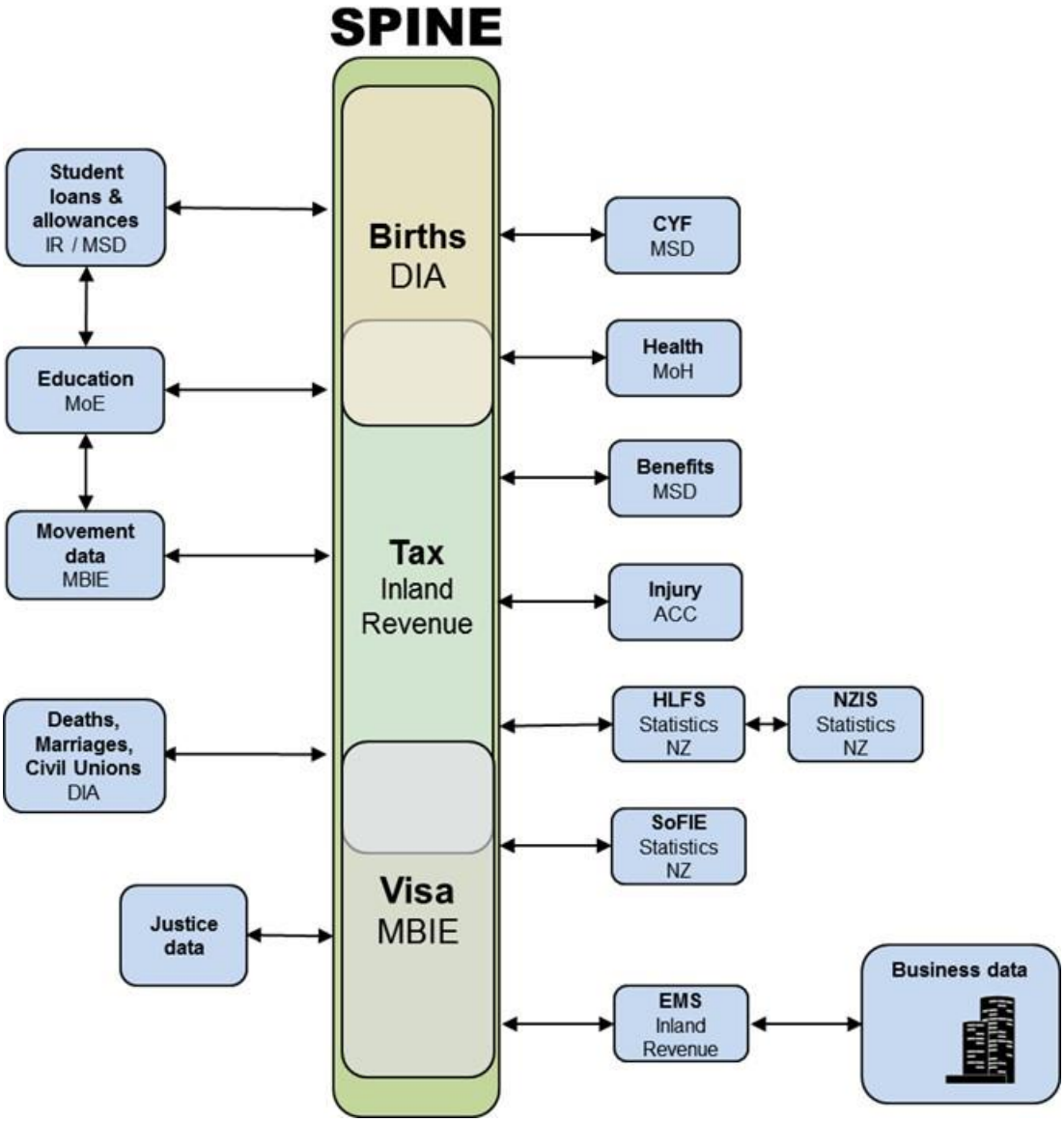
Statistics NZ developed the Integrated Data Infrastructure (IDI) as an environment in which to link multiple data sources in a systematic and secure way. It was developed to produce official statistics outputs and to allow Statistics NZ staff and external researchers to conduct policy evaluation and research on people's transitions and outcomes. The IDI contains administrative and survey datasets, linked at the individual level. The IDI continues to change as new datasets are added.

This section describes the structure and content of the IDI in May 2015.

See figure 1 for the basic structure of the IDI.

Figure 1

Structure of the Integrated Data Infrastructure in May 2015



Source: Statistics New Zealand

The structure of the IDI can be described as a central 'spine' to which a series of data collections are linked. The spine forms the conceptual centre of the IDI and all other datasets are linked to it. The target population for the spine is all individuals who have been residents of New Zealand. Three data sources are linked together probabilistically to create the spine: a list of all IRD numbers that have been issued by Inland Revenue (IR); a list of all births registered in New Zealand since 1920; and a list of all visas granted to migrants from 1997 (excluding visitor and transit visas).

Other datasets are linked to the IDI spine (Statistics New Zealand, 2014b). The linked datasets cover a wide range of subject areas and include: employer and employee job and earnings information based on Inland Revenue data; health information including GP enrolment and hospital visits from the Ministry of Health; education data from the Ministry of Education; benefit dynamics data from the Ministry of Social Development; student loans and allowances data from several sources; migration movements data from the Ministry of Business, Innovation and Employment; and the Household Labour Force Survey and New Zealand Income Survey from Statistics New Zealand.

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

Identifying a resident population in IDI

The IDI spine contains more than 9 million individuals, far more than the current New Zealand usual resident population of approximately 4.5 million. Many individuals in the IDI spine are former usual residents of New Zealand who have since left or died. It is therefore necessary to restrict the IDI spine population to the subset of individuals who are usual residents of New Zealand at a given date (for the purposes of this paper, the date was census day, 5 March 2013).

The method used to select the usually resident population relies on the identification of activity in New Zealand administrative systems that would indicate an individual's presence in New Zealand over a period prior to the reference date. Individuals who leave the population by death or outmigration prior to the reference date are removed.

Specifically, the method used to identify the IDI resident population (IDI-ERP) for census day (5 March 2013) was as follows:

- For ages 5 and over, the spine population was restricted to those individuals who had activity in one of the following IDI datasets in the 12 months prior to census day:
    - ACC claims
    - Inland Revenue tax (employer monthly summary of tax paid at source, or annual tax return data; receipt of taxable benefit payments is included)
    - Health (pharmaceutical prescriptions, GP enrolment, hospital admissions, non-admission hospital visits)
    - Education (school enrolment, tertiary enrolment or attainment)
- For ages under 5 having a record in the spine was sufficient for inclusion in the population. There was no additional requirement of activity in the 12 months prior to census day.
- Linked death records were used to identify individuals with a date of death prior to census day. These individuals were removed from the population.
- Linked migration data were used to identify individuals who had moved overseas. Individuals were classified as having moved overseas if they were overseas on the reference date, and the total length of time spent overseas was at least 10 of

the 12 months spanning census day (that is, the six months either side of census day).

Location information in IDI

In May 2015 there were five data sources in the IDI that contained information about where individuals live. Not all addresses in these data sources represent where an individual actually lives. In some cases, addresses may represent postal or contact addresses provided to the administrative data supplier (Statistics NZ, 2013). For all of the data sources, address information was geocoded to a meshblock where possible. For Primary Health Organisation (PHO) data, address updates were available up until the end of November 2012. For all other sources, address updates were available at least up until census night (5 March 2013).

The five sources of address information were:

- Inland Revenue (IR): The address history table is a summary table in IDI containing a history of address updates (coded to meshblock level) for each individual in IDI. Almost all meshblocks in the address history table (99 percent) come from Inland Revenue data (Gibb & Shrosbree, 2014) and represent the meshblock for the contact address supplied to Inland Revenue. The remaining meshblocks (less than 1 percent) come from postcodes in student loan data, with a meshblock being randomly assigned from within the postcode.
- National Health Index (NHI): meshblock of residence as recorded when visiting hospital or outpatient clinic.
- Primary Health Organisation (PHO): meshblock of residence as recorded when visiting a general practitioner.
- Ministry of Social Development (MSD): meshblock of residence as reported when applying for a working age benefit (not including superannuation).
- Ministry of Education (MOE): meshblock of residence as reported when enrolling at primary or secondary school (but not tertiary education).

Timestamps in each of these sources were used to select the most recently updated meshblock from each source prior to the reference date. Meshblocks were converted to area units and territorial authorities using standard meshblock concordances from Statistics New Zealand.

Linking census data and the IDI

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

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

To enable individual-level comparisons between the geographic information in the IDI and the geographic information in the census, the census must be linked to the IDI at the individual level. This link was created by Census Transformation in May 2015 (Statistics NZ, 2014c). The linking was done for the purpose of better understanding the coverage and quality of census information in the IDI, and the linked data was only available to approved Statistics NZ staff working on the Census Transformation programme. The linking method used for this paper differed slightly from that being used to link the 2013 Census to the IDI spine for the September 2015 IDI refresh.

The census was linked to the spine of the IDI in the May 2015 refresh. Linking was completed in Quality Stage using probabilistic matching techniques. The variables full name, date of birth, sex, meshblock of usual residence, and country of birth were used in the linkage process. Overall, 94 percent of census records were linked to the IDI. The match rate was higher for NZ usual residents than for overseas visitors, and much better for individuals who had used e-forms (98 percent linked) compared to paper forms (93 percent linked).

Populations for comparison

To enable better comparison of the IDI-ERP and census populations, the following adjustments were made to the populations:

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

From the above population, individual level comparisons can only be done for those individuals in the IDI-ERP who were able to be linked to their census record and who had a meshblock recorded in both the census and the IDI (n=3,787,700). The population available for individual-level comparisons represented 91 percent of the census population and 86 percent of the IDI-ERP. Unless otherwise stated, all comparisons in this paper are based on the linked population. The reference date used for individual-level comparisons was census night, 5 March 2013.

4 Results

Coverage of geographic information in the IDI

Table 1 shows the coverage of geographic information in the IDI. For each source of geographic information, the table shows the ages best covered by the data source, and the percentage of individuals in the IDI-ERP within those target ages who had a meshblock recorded in that data source.

Coverage of geographic information varied between different data sources. Health and Inland Revenue sources had the highest coverage, as most of the population has had some contact with these agencies and has, at some point, had an address recorded.

While most individuals in the target ages have had contact with the Ministry of Education (via a school enrolment), not all individuals have an address recorded, so coverage was low. The MSD benefits source had low coverage, because only a small proportion of the target population have had contact with MSD. Overall, when the data sources were combined, they had very good coverage of the population, with 99 percent of people having a meshblock recorded in at least one of the data sources.

Table 1

Coverage of meshblock information in IDI data sources

Source	Ages (years)	% with meshblock information
Health (NHI)	all	89
Health (PHO)	all	83
Inland Revenue	all	93
MOE[1] (education)	6–15	48
MSD[2] (benefits)	18–64	14
Any source	all	99

1. Ministry of Education.
2. Ministry of Social Development.

Source: Statistics New Zealand

Comparison of IDI and census meshblocks

Using the linked IDI and census dataset it was possible to compare the meshblock recorded for an individual in the IDI (as at census day, 5 March 2013) with the meshblock recorded for that same individual in the census. Census meshblocks have a high level of accuracy so can provide a reliable indication of the quality of the IDI meshblocks.

Table 2 shows the percentage of people who had the same geographic information recorded in IDI and census for three geographic levels: meshblock; area unit; and territorial authority (TA). The table shows that different sources had different levels of agreement with the census. The health sources (NHI and PHO) had the highest levels of agreement, with more than 70 percent of IDI meshblocks and more than 90 percent of territorial authorities being the same as in the census. The lowest levels of agreement were for MSD benefits, with 57 percent of IDI meshblocks and 85 percent of territorial authorities being the same as in the census. The different levels of agreement for different data sources may be due to differences in the frequency of contact and address updating procedures at different agencies. For example, many individuals do not have regular contact with Inland Revenue, so they may not update Inland Revenue when their address changes.

Table 2

Percentage of people with the same geographic information in IDI and census			
Source	**% of non-missing that are same as census**		
	Meshblock	**Area unit**	**TA[1]**
Health (NHI)	75	78	92
Health (PHO)	71	74	90
MOE (education)	67	73	92
Inland Revenue	63	68	89
MSD (benefits)	57	62	85

1. Territorial authority.
2. Ministry of Education.
3. Ministry of Social Development

Source: Statistics New Zealand

For all sources there was greater agreement between the census and the IDI at the TA level than at the area unit or meshblock level. This is because, although some individuals move house and do not update their address in administrative data sources, they often remain in the same TA, but not in the same meshblock or area unit.

A given individual may have a meshblock recorded in several different IDI data sources. These meshblocks may differ, for example if an individual has updated their address with some agencies but not others. It is therefore necessary to find a way to combine the geographic information from different sources and select the 'best' meshblock for each individual at any given date.

Table 3 shows the agreement between IDI and census geographic information for two simple methods of combining the information from different sources. The first is a 'prioritised' method in which the meshblock sources are ranked according to their agreement with census and then the meshblock from the highest-ranked available source is selected. The second is a 'most recent' method in which the meshblock that was updated most recently is selected.

Table 3 shows that the meshblocks selected using the 'most recent' method were more likely to agree with census meshblocks than those selected with the 'prioritised' method. Almost 80 percent of IDI meshblocks and 94 percent of IDI territorial authorities were the same as in census when the 'most recent' method was used, compared to 70 percent and 90 percent using the prioritised method. When meshblocks were selected using the 'most recent' meshblock method, around 46 percent of the meshblocks selected came from the NHI health data, 35 percent from Inland Revenue, 13 percent from PHO health, 5 percent from Education, and 2 percent from MSD working age benefits.

Table 3

Percentage of people with the same geographic information in the IDI and the census By method used to combine meshblocks from different sources			
Method for combining information	**% of non-missing that are same as census**		
	Meshblock	**Area unit**	**TA[1]**
Prioritised	70	73	90
Most recent	79	82	94

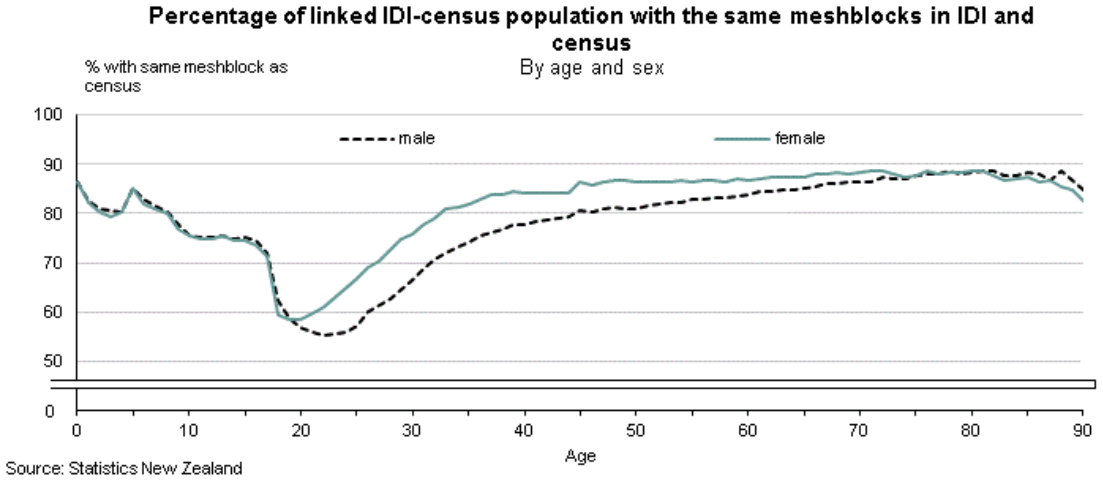
1. Territorial authority

Source: Statistics New Zealand

Further analysis revealed that, overall, 84 percent of individuals had a meshblock recorded in at least one IDI data source that was the same as their census meshblock. This provides an upper limit for the potential agreement rate that could be obtained by using a set of rules to select the 'best' meshblock in the IDI. The upper limit for area units was 86 percent and for territorial authorities it was 95 percent. For territorial authorities, the results from the 'most recent' selection method are close to the upper limit, suggesting that refining the method for selecting the 'best' meshblock would only improve territorial authority agreement by a small amount. For area units and meshblocks, it may be possible to get a slightly larger increase in agreement (up to 5 percent) by refining the selection method.

Figure 2 shows the proportion of individuals in the IDI-ERP who have the same meshblocks recorded in the IDI and the census, by age and sex. The 'most recent meshblock' method has been used to select a meshblock for these individuals. The figure shows that agreement between IDI and census meshblocks is lowest in the young adult ages (approximately ages 15–30) compared to other ages. Young adults are more mobile than other age groups and therefore may be more likely to have an administrative address that is out of date. Overall, agreement between IDI and census meshblocks is lower for males than for females at most ages, with the exception of ages under 15 and over 75, where levels of agreement are similar for males and females.

Figure 2



Using geographic information to create households

An additional use of address information in the IDI is to create households. Individuals who are living at the same address can be grouped together to form a household. Creating households is a more demanding test of the quality of address data as a 'correct' household requires that all individuals in the household are registered at the correct address, and that no additional individuals are incorrectly registered to that address.

We took several steps to create households in the IDI. First, a single address was allocated to each individual in the IDI-ERP using the 'most recent' method described previously. Second, the geocoded address identifiers associated with these addresses were used to group individuals into households. Individuals with the same address identifier were considered to be in the same household.

To examine the quality of household information in the IDI we compared the set of individuals living in an IDI-ERP household with the set of individuals living in a census household. All of the households (addresses) identified in the IDI-ERP were also identified in the census.

Table 4 shows, for each IDI-ERP household size, the percentage of IDI-ERP households of that size that had the same household size in census, and the percentage that contained the exact same individuals in census. The population used for the household analysis in Table 4 was restricted to households where all household members (as specified in the IDI) had records in the census and the IDI, and those records were linked together. Households where one or more individuals were away from home on census night, or did not have an IDI address recorded, were excluded. Visitors (people who were in a household on census night but do not usually live there) were excluded from household counts. Census dwellings that were non-private (such as rest homes, boarding houses, university accommodation) were excluded from the analysis, as they are not considered to be 'households' in the census and as such they do not have a household size available.

Table 4 also shows that, overall, 55 percent of IDI-ERP households had the same household size in census, and 48 percent contained the same set of household members in census. Agreement between IDI-ERP and census household sizes was better for smaller households than for larger households.

Table 4

Comparison of household size and composition in the IDI and the census

IDI-ERP[1] household size (number of people)	Number of IDI-ERP households	% with same household size in the census	% with same household members in the census
1	261,300	69	64
2	328,200	70	62
3	212,900	40	32
4	170,300	50	42
5	90,000	36	29
6	40,800	24	17
7	18,900	16	10
8+	20,600	9	5
Total	1,142,900	55	48

1. Estimated resident population, see method used to identify the IDI resident population.

Source: Statistics New Zealand

If household sizes in the IDI-ERP are not correct, this will have an impact on the household size distribution.

Table 5 shows the distribution of household size in the IDI-ERP compared to census. The population used for the IDI-ERP distribution in Table 5 was the full IDI-ERP, not the restricted population used in Table 4.

Table 5 also shows that, compared to the census, the IDI-ERP contained substantially fewer two-person households (416,100 in the IDI-ERP compared to 527,700 in the census). Compared to the census, the IDI-ERP contained substantially more large households (six people or more). This may be due to non-private dwellings (such as university accommodation, rest homes, boarding houses) being included in the IDI-ERP household count, but not in the census household count.

Table 5

Household size distribution in the IDI and the census			
Household size (number of people)	**Number of IDI-ERP[1] households**	**Number of census households**	**Number of IDI-ERP[1] households as % of census**
1	368,600	355,300	104
2	416,100	527,700	79
3	273,800	252,900	108
4	216,200	235,300	92
5	117,700	106,300	111
6	55,700	41,200	135
7	27,200	15,400	177
8+	33,300	13,800	241
Total	1,508,600	1,549,900	97

1. Estimated resident population, see method used to identify the IDI resident population.
Source: Statistics New Zealand **Source:** Statistics New Zealand

5 Discussion

This paper examined the quality of geographic information in the IDI by comparing it with the geographic information contained in the 2013 Census.

Summary of main findings

Coverage of geographic information in the IDI was high, with 99 percent of individuals in the resident population having a meshblock recorded in at least one IDI data source.

Comparison of the meshblocks recorded in IDI against those recorded in the census revealed that different administrative data sources have different levels of agreement with the census, with health having the highest and MSD's working age benefits the lowest. Combining the geographic information from these individual sources produced geographic information that was more accurate than any single source alone. When the most recently updated meshblock from any source was selected, 79 percent of people had the same meshblock recorded in the IDI and the census, 82 percent had the same area unit, and 94 percent had the same territorial authority. Individuals in the young adult ages (15-30), and particularly males, were least likely to have agreement between IDI and census meshblocks.

The quality of geographic location information was also tested by using it to create households, and then comparing household size and composition against that recorded in the census. Agreement between households in IDI and census was lower than for individual geographic information. Overall, 55 percent of census households had the same household size in the IDI, and 48 percent contained exactly the same set of household members.

There are several possible reasons why the location information recorded about an individual in the IDI does not agree with that recorded in the census. Some individuals do not have frequent interactions with data providers, and may not update their address with the data provider when they move house. In addition, recording or geocoding errors may result in an individual's address being coded to the wrong meshblock. Finally, simple comparisons between IDI and census location information do not reflect the complex reality for some people. People who live across multiple residences may report different addresses in different sources. Some portion of the disagreement between IDI and census addresses may reflect these complex situations, rather than errors or outdated addresses.

Some agencies have less operational imperative than others to update address information, particularly as the majority of services move to being offered online. This may be one explanation for the finding that different data sources have different levels of agreement with census geographic information. An additional possibility is that some different data sources cover population groups that are more likely to have outdated address. For example, individuals receiving welfare benefits may be more mobile than other groups, and this may explain the lower agreement between MSD and census addresses.

Limitations

There are some limitations to these results.

The analyses in this paper were restricted to individuals who had IDI and census records that were able to be linked together. The linking of IDI and census records relied in part on meshblock of usual residence, which was used as a blocking variable. Therefore, the results reported in this paper may overestimate the level of agreement between census

and IDI geographic information. It should be noted, however, that the link rate of census to IDI was high (94 percent) and the estimated rate of false positive errors was low (0.7 percent), suggesting that the linking was of high quality and the results in this paper are likely to be close to the 'true' levels of agreement.

The comparisons between IDI and census geographic information made in this paper could only be made for census day (5 March 2013). It is possible that the level of agreement between IDI and census geographic information has changed since census day, however this cannot be tested until the next census in 2018.

Improving the quality of location information

While the quality of geographic location information in the IDI varies by data source, it is possible to combine these sources in a way that provides accurate information for around 80 percent of people. While this result is promising, it leaves around 20 percent of individuals with an incorrect address. Given the importance of accurate location information to a range of analyses, attention should be given to improving the quality of the location information available in the IDI.

One strategy for improving the quality of address information in IDI is to refine the method for selecting a meshblock from multiple available meshblocks in the IDI. However, the analyses in this paper suggest that, at the present time, this strategy would only result in a small improvement in address quality.

There may be some small improvement in address quality when more up to date PHO address updates become available in the IDI. At the time that the analyses in this paper were conducted, PHO address updates were only available to the end of November 2012. If address updates were available right up to census night, this may capture additional updates and improve the quality of the location information.

Another strategy that could improve the quality of geographic information in the IDI could be to improve the method for geocoding addresses. This is currently under investigation.

The strategies mentioned here are likely to result in only small improvements in address quality. It is likely that greater gains would come from strategies to improve the quality of address information at source agencies. It is not mandatory for source agencies to collect residential addresses, and many agencies do not have a need to collect accurate and up-to-date location information. However, improvements in address quality could still be obtained by ensuring that addresses:

- are collected according to common standards
- include enough information to be accurately geocoded
- are updated regularly.

In particular, improving address quality for groups of individuals who are known to have poor quality addresses, such as tertiary students and other young adults, could be worthwhile. Some improvements are already in place. An improvement to Ministry of Health geocoding processes in 2013, for example, is likely to result in improved quality for newer health addresses.

Even with improvements in address quality, some individuals are likely to have an incorrect address recorded. There may be a role for modelling approaches that identify and correct likely address misclassifications. Imputing the small number of missing meshblocks in the IDI-ERP could also be useful in improving the quality of geographic information.

Accurate information about where individuals live is key to producing official statistics, and is central to many policy and research questions. If individuals are not placed in the right location, these errors will flow through to all regional statistics and analyses, including regional breakdowns of households and families, incomes, and educational achievement.

Further work is being undertaken as part of the Census Transformation programme to develop our understanding of the quality of geographic location information in administrative data sources. This includes examining the predictors of errors in geographic location, the impact of these errors on subnational population distributions, and a more in-depth investigation into the quality of household information in administrative data.

6 References

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

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

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

Statistics NZ (2014c). Privacy impact assessment for linking census and Post-enumeration Survey data to the Integrated Data Infrastructure. Retrieved from www.stats.govt.nz.

7 Disclaimer

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

The opinions, findings, recommendations, and conclusions expressed in this paper are those of the author(s), not Statistics NZ.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation and the results in this paper have been confidentialised to protect these groups from identification.

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

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes.

Any person who has had access to the unit-record data has certified that they have been shown, have read, and have understood section 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

- Reproduction of material
  Any table or other material in this report may be reproduced and published without further licence, provided that it does not purport to be published under government authority and that acknowledgement is made of this source.
- Copyright Information from Statistics NZ may be freely used, reproduced, or quoted unless otherwise specified. In all cases, Statistics NZ must be acknowledged as the source.
- Liability
  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.