Stats NZ
Tatauranga Aotearoa

The quality of administrative data for census variables: Strengths, limitations, and opportunities

Census Transformation

New Zealand Government

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 Stats NZ.

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

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

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

Citation

Bycroft, C, Miller, S, Gath, M, Matheson-Dunning, N, Simpson, K, & Das, S (2021). *The quality of administrative data for census variables: Strengths, limitations, and opportunities*. Retrieved from www.stats.govt.nz.

ISBN 978-1-99-003232-5 (online)

Published in January 2021 by
Stats NZ Tatauranga Aotearoa
Wellington, New Zealand

Contact
Stats NZ Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4600

www.stats.govt.nz

Contents

List of tables and figures

List of tables

List of figures

Purpose and summary

Purpose

This paper summarises the census transformation programme research to date on the quality of census attribute information derived from administrative (admin) data sources.

Summary of key points

Stats NZ's census transformation programme has undertaken a series of investigations based on the 2013 Census looking at the ability of admin data sources to provide census-type information. These investigations aimed to assess the potential for linked admin data sources to meet information requirements for social and economic characteristics (attributes) measured in the census. This is important in the context of moving towards a greater use of admin data in the census. The results of this programme of work have already provided benefits as they were integral to the use of admin data to fill in missing data due to non-response in the 2018 Census.

This report summarises the attribute investigations to date. We collate the key findings from comparisons of admin data and census responses where the assessments were based on a quality framework approach. Results are summarised in terms of accuracy, assessed by representativeness (does admin data include the right people or dwellings) and errors of measurement (are the right things being measured). The results are described at a high level, and general themes that have emerged are discussed. We found a broad spectrum of results across the variables considered.

The level of coverage for many of the variables that can be obtained from admin sources (sometimes combined with previous census data), reaches a similar level to that achieved by the full field enumeration census. Several of the admin-derived variables investigated were highly accurate, and a number of variables showed good potential for providing census-type information, but there are caveats either in coverage or measurement error. At the other extreme, around one-third of variables have limited, if any, admin data potential and in the absence of new data sources will continue to rely on survey collection.

This variation highlights the need to consider each census variable on its own merits, and to ensure the detail in each admin source is understood well enough to apply the data in the census context. Admin error structures are quite different from those from a field collection and may affect particular variable categories, specific age groups, or populations such as new migrants. A combination of imputation or statistical models combining admin and survey data will be needed to provide unbiased estimates.

The reasons for quality concerns are varied, and there may be potential for improving quality. Critical areas of focus for improvements in admin sources are the place of usual residence (particularly for young adults), family and household data, and iwi affiliation. The new Data and Statistics Bill, once passed into law, will facilitate the collection and use of administrative data for the benefit of the wider statistical system.

There is now a significant body of census-type information that can be collated from admin data for most of the New Zealand population. Next steps include the release of an experimental 'Admin Population Census' as an annual time series, progressively adding more census variables derived from admin sources to the admin-based resident population dataset. Our aim is to demonstrate the breadth of information available and to provide a focus for discussion with customers about the quality issues and benefits associated with an admin-based census.

Introduction

The Census of Population and Dwellings is an official count of how many people and dwellings there are in New Zealand. The census has been held every five years, with some exceptions. The first census was held in 1851, and the latest censuses were in 2013 and 2018.

The New Zealand census has followed a 'traditional' full field enumeration approach. By asking everyone to complete a set of questions about themselves and their household and dwelling, we can capture a snapshot of who is living in New Zealand. The census is the only survey in New Zealand that covers the whole population. It provides the most complete picture of life in our cities, towns, suburbs, and rural areas.

The data helps the government plan services. These include hospitals, kōhanga reo, schools, roads, and public transport. Councils, iwi, businesses, and other organisations also use the data to work out the needs in their area.

Census Transformation

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, including exploring the feasibility of a census based on admin data (Stats NZ, 2014a).

As outlined in the census transformation report Overview of progress on the potential use of admin data (Stats NZ, 2014a), a combined admin data and survey information approach could gather census information in future. After considering a range of options, government recommended in a 2015 Cabinet paper (Census transformation – a promising future) that Stats NZ work actively towards a future census based primarily on Government's administrative data, supported by redevelopment of its household surveys. Admin data could be used as the basis for population estimates (along with a coverage survey), and for some attribute information where available. A large-scale sample survey would be necessary for attribute information which cannot be obtained from admin sources.

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 admin data sources to meet those requirements. The attribute investigations summarised in this paper are an important contribution to further understanding the role that admin data may play in future censuses.

The modernised 2018 Census marks a significant step forward in the use of admin data in the New Zealand census (Stats NZ, 2019a; 2019b). This was partly as planned in the modernised collection model, but the role of administrative data was significantly expanded as a result of a lower than expected response rate. The admin data and new methods developed for the 2018 Census were based on the research undertaken by the census transformation programme.

Read more about the Census Transformation programme in New Zealand.

Admin data for census attributes

Administrative data is data collected by government agencies or private organisations in the course of conducting their business or services. It is data not collected primarily for statistical purposes. Rather, it is collected for operations such as delivering a 'service' (for example, health or education), or legal requirements to register events (for example, births, deaths, and marriages) or as a record of transactions or events (for example, tax payments and overseas travel journeys). The population and data content is defined by the collection organisation and they have primary control of the methods by which the administrative data are captured and processed. As a result, administrative data differs in nature, scope, and quality to data collected directly through the census or surveys, where control of who is asked for what information is in the hands of the statistical agency. To date, most of the admin sources investigated by census transformation are generated by government agencies.

The first requirement of an admin-based census is the construction of a suitable population from admin data. An admin New Zealand resident population has been constructed, which is a good approximation of the population as measured by official statistics (Stats NZ, 2017). Information about the social and economic characteristics (attributes) of the individuals in the admin population can be linked from admin data sources.

Stats NZ's Statistical Location Register has recently been developed and provides a reference list of addresses and dwellings in New Zealand. If individuals can be linked to the dwellings where they live, then people can be grouped into households, and admin information about the dwellings can be incorporated.

We used Stats NZ's Integrated Data Infrastructure (IDI) to access the admin data sources. The IDI is a linked (integrated) database created for research purposes. The IDI contains many admin datasets, Stats NZ surveys, and the 2013 and 2018 Censuses, linked at the individual level. The data includes information on education, work, income, benefits, migration, justice, and health gathered from a range of government agencies. The IDI matches admin supplied addresses to reference address IDs, and includes information about businesses in the Business Register. This range of integrated data starts to replicate the variety of information collected in the census and means the IDI can act as a test environment for examining the potential of linked admin data sources to produce census-type information.

The combination of data sources and census variables derived from them is shown in figure 1. The picture is complex, with some admin sources providing information about several variables, and variables that rely on several sources. The core demographic variables age, sex, usual residence, and ethnicity are collected by multiple agencies. This provides very high coverage of the population for these important attributes but introduces the need to develop methods for resolving conflicting information. Inland Revenue (IRD) is a key source for income, benefits, and work-related information, and linked employer and employee data derived from tax returns provides a link to the Business Register and workplace information. More detailed information on benefits can be derived from Ministry of Social Development (MSD) benefits data. The Ministry of Education (MOE) provides education and training data, and Department of Internal Affairs (DIA) registrations provide information on Māori descent, country of birth, family relationships, and number of children born. Ministry of Business, Innovation and Employment (MBIE) visa applications and border movements provide information about migrants. We also include historic censuses as a source of information for some variables.

Census housing variables can be obtained from several sources: MBIE tenancy bonds data and Housing New Zealand Corporation (HNZC) provide information about rented dwellings, while Building Consents, local council District Valuation Rolls and Quotable Value (QV) have information

about two potential new census variables: age of dwelling and floor area (figure 2). Vehicle registrations may provide information for number of vehicles, though this has yet to be investigated.

The system as a whole is stronger than might be apparent from the point of view of a single agency. As well as the more obvious benefit of increasing coverage, the use of multiple sources for a variable has other benefits. Multiple sources provide an in-built redundancy so that the key demographic variables, for example, are not dependent on any single data source. The use of multiple sources also provides protection against quality limitations. For example, lower quality sources may only be needed when higher quality sources are not available.

Figure 1. Administrative sources for census information about individuals

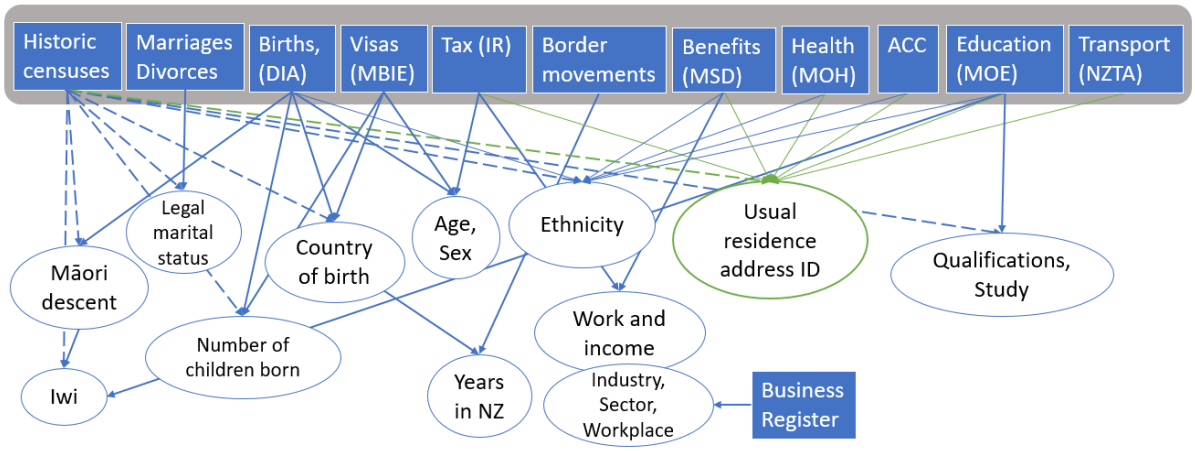
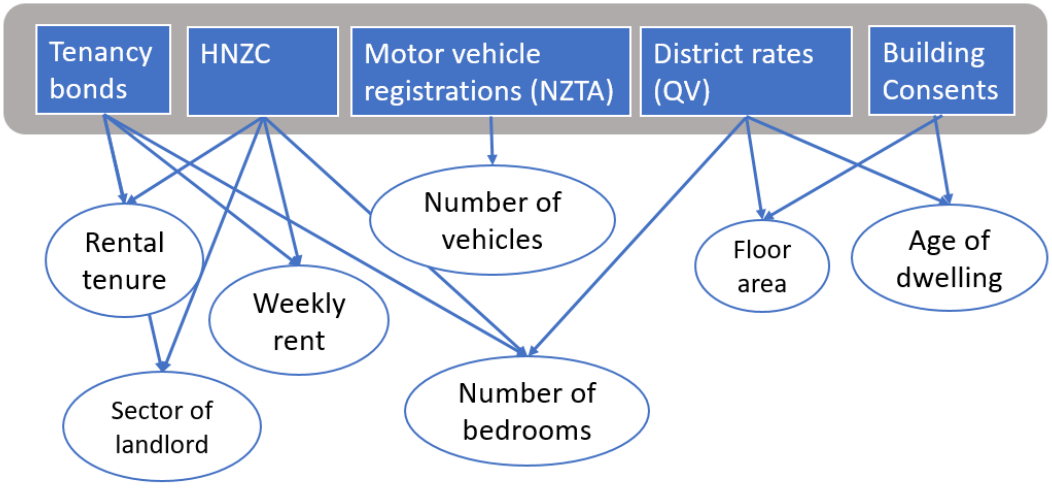


Figure 2. Administrative sources for census information about dwellings and households



Admin investigations

A range of investigations have been undertaken into the potential of linked admin data to meet information requirements for census social and economic characteristics (attribute variables) for individuals, households and families, and dwellings.

The key research questions guiding the attribute investigations were:

- What is the quality of the derived admin information in relation to the census information needs?
- Could the admin data satisfy the information need, in place of the census question?
- What improvements would be required to improve the potential for using the admin data for the census?

Research has been undertaken in two phases and is guided by Stats NZ's quality frameworks. The first phase considered admin data sources from a high-level, metadata perspective and considered several of the six dimensions of quality used by Stats NZ[1]. The purpose of phase 1 was to provide an early indication of the potential of admin sources to produce census attribute information and to guide decisions about where to direct more in-depth analyses. The second analysis phase focussed on a quantitative assessment of the accuracy dimension of quality.

O'Byrne et al (2014) provides a first broad look at the potential for admin data to produce the social and economic information currently provided by the census. The authors identified admin data sources related to census topics and assessed how likely these sources were to satisfy the information needs currently met by the census. These assessments were based on five quality measures: relevance, accuracy of coverage, accuracy of linkage, timeliness, and accessibility. The phase 1 results were reported as an indicative likelihood (likely, possible, or unlikely) of admin data to satisfy census information needs for different attributes. Fewer than half of the 39 attributes were assessed as 'likely' or 'possible' to be able to provide census-type information, with the majority assessed as 'unlikely'. The most common reason for the 'unlikely' rating was a lack of suitable admin data.

The phase 2 investigations are detailed analyses using the linked unit record data in the IDI. The assessments focus on accuracy in terms of representativeness (does admin data include the right people or dwellings) and errors of measurement (are the right things being measured). Assessments were based on comparisons between the admin data and the 2013 Census and carried out between 2015 and 2020 using the data available in the IDI at the time. Slight differences in data between different IDI refreshes are unlikely to have any material impact on results. More recently (March 2019), the IDI introduced improved processes for address matching, which may provide some improvement on our results for dwelling and household variables that used earlier refreshes.

All 2013 Census output variables were in scope for the attribute investigations except the core census usually resident population counts. New topics for the 2018 Census were also considered.

Since over 40 variables are output by the census and each detailed analysis involves a considerable amount of time and expertise, variables were prioritised in terms of the availability of admin data, the importance to census overall and to Māori, and the likely value to the 2018 Census. Analysis of the attribute variables focussed on national level results by age and sex. Breakdowns by subnational geographies and ethnicity were typically out of scope. To date, most of the variables highlighted by

[1] The six dimensions of quality are: relevance, accuracy, timeliness, accessibility, consistency/coherence, and interpretability (Stats NZ, 2007). Similar approaches are used by national statistics offices internationally.

O'Byrne et al (2014) as being 'likely' or 'possible' have been investigated, though further work remains to be done. Detailed research papers are published under Stats NZ Census Transformation Programme. Appendix 1 provides a list of the variables investigated, the admin sources used, and a reference link to the detailed research paper.

This paper summarises our current understanding of the quality of linked admin data and its potential to meet information needs for census attribute variables, bringing together the attribute research undertaken by the census transformation programme to date.

We first describe our approach to quality assessment, and then summarise the quantitative measures of representivity (mainly coverage) and errors of measurement obtained through the detailed variable analyses. We highlight common reasons for lack of coverage, and for measurement errors, and suggest some ways of improving quality. We conclude with a discussion.

Approach to quality assessment

While every variable is unique, each detailed investigation followed the same general process for assessing the quality of the admin sources. The first step was to build an understanding of the data sources, their purpose, collection process, and detailed information about the available data. For each variable, a method was developed for its derivation from one or more admin sources. Development of these methods included decisions about what data to include or exclude, and how to resolve conflicting information. This process is more or less complex depending on the nature of the variable.

The detailed analysis continued the use of quality frameworks, focussing on accuracy across the two dimensions of 'representation' and 'measurement' (Stats NZ, 2016; Zhang, 2012). Representation concerns whether the right population is being represented, whereas measurement is about whether the intended concept or variable is being measured correctly.

We evaluated the potential to produce census-type information by comparing the information between the 2013 Census and admin sources at three levels:

1. **concepts and definitions –** what is ideally being measured, as described in statistical standards
2. **aggregate distributions –** comparison of census distributions and distributions derived from the admin data
3. **individual-level information –** comparison of census responses with the equivalent admin data for the same person, household, or dwelling, using the 2013 Census linked to the admin data in the IDI.

Representivity: coverage of target populations

We measure representation mainly through **coverage** of the target population from which observations for a particular topic can be drawn. Coverage errors are the differences between the units actually available in practice and the full set of units we include in the (ideal) target population. If some groups in the target population are disproportionately missed from the data, this may lead to biased results unless there is some further statistical adjustment.

We compare the coverage obtained by the 2013 Census with coverage of admin-derived variables.

The 2013 Census target population

The census target population of interest here is all people who usually live in New Zealand and are present in New Zealand on census night[2]. New Zealand residents who are temporarily overseas on census night are excluded from the census by design. Some variables are collected for everyone, while other variables are restricted to a certain subpopulation (for example, those 15 years and over, or those born overseas).

All dwellings in New Zealand are included in the census. The census dwelling frame, a list of all dwellings counted by the census, is considered to be highly accurate. However, most housing and household variables are collected only for dwellings that are occupied on census night. The census

[2] The census target population also includes overseas visitors who are in New Zealand on census night, but they are not considered further.

does not collect information from vacant dwellings, or from dwellings where all the residents are away from home on census night.

Census coverage is affected by people who do not respond to the census at all (unit non-response) and the coverage of each variable also depends on how well questions are answered (item non-response).

The overall response rate of the 2013 Census counts was estimated as 92.9 percent by the 2013 post-enumeration survey (PES) (Stats NZ, 2014b). Non-response included a 2.4 percent net undercount and 4.7 percent of the 2013 population counted through substitute records in the census dataset (a unit imputation adjustment for people missed by the census). Response to the census differs across the population, with young adults, Māori, and Pacific peoples more likely to be missed by the census. Item non-response in the census dataset is variable specific and mainly between 1 and 5 percent. Combining all these factors, the overall census coverage for the attribute variables ranges between about 8 percent and 12 percent, depending on the variable, and could be higher for specific subgroups and some geographic areas.

Admin census target population

The official **estimated resident population (ERP)** is the best measure of who lives in New Zealand. The ERP adjusts for net census undercount and residents temporarily overseas, who are not included in the census usual resident population counts. For an admin-based census we are interested in an admin-based NZ resident population which estimates the ERP. This admin population is not constrained by the operational requirements of a census field collection. It includes residents who may be temporarily overseas on a given date and does not rely on households to be occupied on census night to obtain housing and household information.

For individual variables derived from admin data, the admin NZ resident population derived from the IDI (the IDI-ERP for the census reference date of 5 March 2013) formed the basis of the admin subject population. The IDI-ERP is a good approximation of the NZ resident population, though includes some under-coverage and over-coverage, with most uncertainty related to younger adults (Stats NZ, 2017). We expect some adjustment for under-coverage and over-coverage of this 'raw' admin population will be needed to achieve the accuracy required for official population estimates (as is the case for net undercount in the current census).

The admin populations for each variable aimed to reflect as closely as possible those who would also be included in the definition of the census subject population. The analysis results may remove residents temporarily overseas from the admin population for a closer comparison with the 2013 Census. Admin coverage can be assessed by the proportion of the relevant admin population for whom the admin variable can be derived. Aggregate comparisons of distributions show where coverage differs by variable categories.

Dwellings corresponding to the census night definition were derived from admin data using available dates and address information in the source data. For housing and household variables derived from admin sources, we used the size of the census subject population to assess the coverage rates.

Identifying statistical units

The incorrect identification of statistical units can be another source of representation errors. While the reporting unit in an admin context may be an event or transaction, it is often important for government agencies to correctly identify the individual associated with that event or transaction. For example, Inland Revenue issues IR tax numbers, and the health system maintains a National

Health Index number for each individual. IDI processes are designed to resolve any remaining duplicates, although a small percent may remain. Anonymised keys (snz_uid) are assumed to represent unique individuals in the IDI.

However, unit errors may be more likely to occur for dwellings since dwellings are a statistical unit that must be derived from address identifiers in the IDI. Unit errors can occur for example, if more than one dwelling is associated with a single address ID.

Errors of measurement

Errors of measurement occur when the value reported differs from the real value. Errors of measurement may occur at random but can also result in systematic bias when they are not random. Errors of measurement occur in both the census and admin sources, and for a number of different reasons.

Statistical standards and classifications provide the concepts and definitions the variable is aiming to measure. A statistical or data standard provides a comprehensive set of guidelines for surveys and admin sources collecting information on a particular topic. See Standards and classifications for more information. Validity error indicates misalignment between the ideal target information and the operational target measure used to collect it. Census questions have been designed to conform to the statistical standard, while this may not be the case for admin collections. The assessments compared how well the concepts underlying the admin variables aligned with the statistical concepts and definitions.

Measurement errors can occur through collection, coding, or other processing errors in both the census and admin contexts. Questions may be misinterpreted or answered incorrectly by respondents (or agency staff), or admin processes may not always operate correctly. Our methods for deriving an admin variable may also introduce measurement errors. Any of these reasons may result in incorrect values or values missing from the datasets.

While we cannot observe measurement error directly, comparison of individual-level information can inform our understanding. Close agreement of responses in admin data and the census provides strong support for good measurement in both sources. When there is disagreement between sources, however, it can be difficult to determine which source is more likely to be correct. This will depend on a range of factors and requires a deep understanding of the mechanisms underlying the particular admin data collection, and of how people respond to the survey questions.

Choice of indicators

Here we summarise errors of measurement with a single quantitative indicator. Consistency between a census response and the admin record for the same person (household, or dwelling) provides a proxy for measurement error. There is a choice of indicators that could be used, and inevitably a degree of subjectivity involved. For example, for nominal variables we look for exact agreement, but for ordinal variables we have taken into account the degree of difference that would be material for a majority of uses. For example, number of bedrooms requires exact agreement, while weekly rental amount is measured as agreement within one band.

The accuracy measures reported here are a single overall indicator that summarises more complex patterns. Accuracy may vary depending on the variable category. If a large majority fall into one category, then the measure will be dominated by the accuracy of the largest group. This may conceal lower accuracy for smaller categories which may be of higher policy interest. For example, employed is a much larger group and better able to be measured through admin sources than unemployed and not in the labour force.

For variables with detailed hierarchical classifications, a decision is made about which level of the classification to use, and for some variables we also need to consider multiple responses. For example, we report two indicators for ethnicity: ethnicity agreement at level 1 (least detailed) of the classification, calculated as the consistency of single and multiple combinations of level 1 ethnic groups, and also show total responses for ethnic groups at level 2 of the classification. Alternative approaches are possible.

The original research based on the 2013 Census reported the degree to which the admin values have the same information as the census, assuming the census response is correct. This was a conservative estimate of the accuracy of the admin data since there will be some level of error in the census responses. In some situations, it is reasonable to assume that the admin data is more likely to be correct, and this may provide a more realistic assessment of the accuracy of the admin-derived variable. Here we use both approaches depending on the variable, as was done in the 2018 Census assessments of admin variables (Stats NZ, 2019b).

Our aim in this paper is to provide a good overall indication of the level of measurement error for variables sourced from administrative data. For specific situations more detail may be needed, and the quality of the admin data may be better than, or worse than, the overall measure. Further detail is provided in the research papers listed in Appendix 1.

Linked data used to compare sources

Comparisons of individual-level information used the linked data in the IDI and compared admin records to the responses of the same individual, same household, or same dwelling, in the census. These comparisons can only be made for people (households or dwellings):

- that were linked (that is, were present in the 2013 Census and linked to the IDI-ERP dataset), and
- have valid responses in both census and derived from the admin data.

This means the individual-level comparisons were typically done for a subset of the full subject population.

Agreement between sources can also be affected by the methodology used to link individuals across data sources. Specifically, two types of linkage error will affect admin variables derived from linked data:

- links may be missed, for example, if the name of a person is recorded differently on different files; dwelling links may be missed if an admin text address is incomplete and cannot be matched to an address ID
- two different people may be wrongly linked, for example, if their names and dates of birth are very similar; dwellings may be incorrectly linked if a misspelled address is linked to the wrong address ID.

Linkage errors may reduce the coverage of an admin source (no information is available if links are not made when they should be), or they may introduce measurement errors if the wrong people (or dwellings) are linked together. Linkage processes in the IDI are designed to minimise incorrect links, with the trade-off that more correct links may be missed (Stats NZ, 2014c). For most IDI linkage projects, incorrect links are less than 2 percent of all links made[3]. Linkage rates vary depending on the data sources. Records that are not linked are a mix of those that cannot be linked because they

[3] Detailed reports of linkage accuracy are prepared for each IDI refresh and are available on request.

are not present on both sources, and incorrectly missed links. Incorrectly missed links are difficult to measure and are unknown.

The accuracy of the linkage between the 2013 Census and the IDI affects individual level comparisons. Overall, 89 percent of 2013 Census usual residents were linked to those in the IDI-ERP dataset, with less than 1 percent estimated as being incorrect links. The high accuracy of this linkage suggests that any bias in comparison results would be minimal.

Results

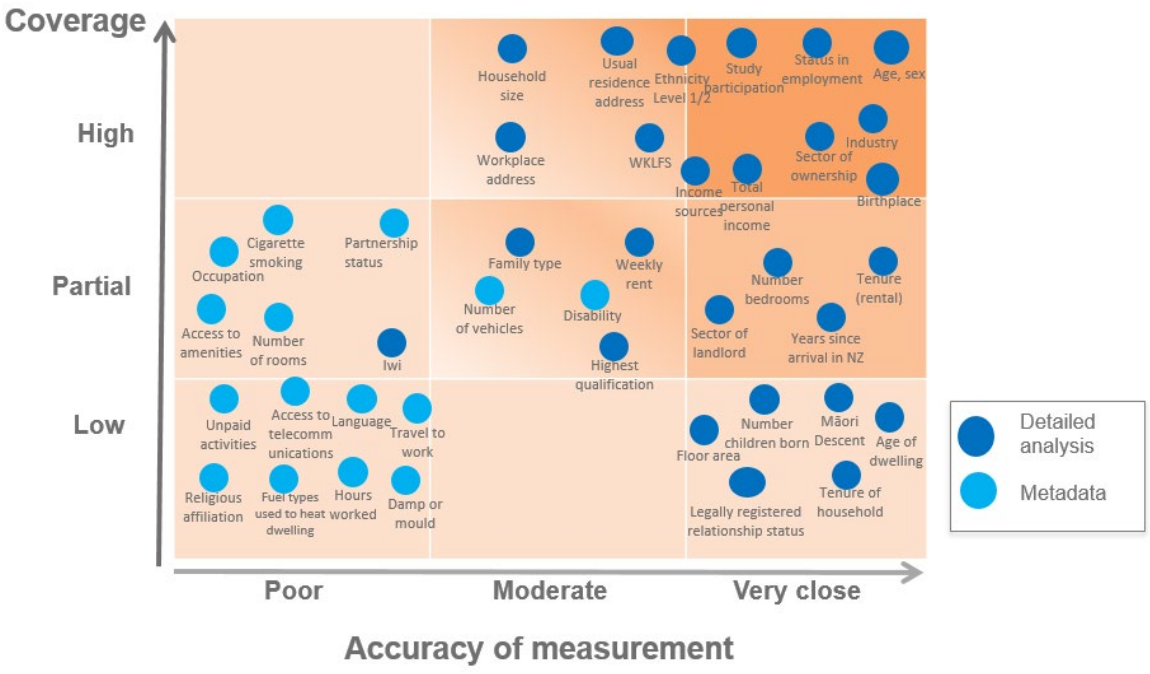
Summary of findings

Results are summarised along the dimensions of representativeness (through the estimated coverage of admin data) and errors of measurement (through the estimated consistency between admin and census values). Coverage tells us how much of the population we can obtain information for, while measurement error tells us whether we are measuring the right thing.

Figure 1 presents these results visually, placing variables in the section related to the level of coverage (low, partial, or high coverage), and the level of accuracy of measurement (poor, moderate, or very close). The light blue circles represent the results of the phase 1 metadata analysis, while the dark blue circles represent results of the phase 2 detailed data analysis. While the detailed analyses of the data provided much greater insight into the nature of the strengths and limitations of the admin data sources, results largely supported the initial phase 1 assessments.

The diagram is indicative only but is useful to illustrate the admin data potential. (We note that the relative placement of variables within a section is not meaningful.)

Figure 3 Accuracy of admin data as only source



Note: Subject to change; results as at November 2020.

Variables in the top right corner of figure 1 have close to full coverage and high accuracy, an indication that the admin data is likely to provide high quality census information.

In contrast, for variables in the bottom left corner we have been able to identify few, if any, relevant admin data sources, and the information that is available is typically not well-aligned with the statistical concept sought by the census. Some variables such as language, religion, or unpaid activities are personal information that is unlikely to ever be captured well by government agencies. Others, such as 'access to telecommunications' and 'fuel types used to heat dwellings', may have

related information in commercial datasets, but these do not necessarily meet the census concept of access or use. Other variables such as 'number of hours worked' and 'number of rooms' are factual concepts which are not collected by available admin systems at present. With changes to admin systems or use of sources such as commercial datasets, cellphone data, or social media, good statistical information from alternative data sources may be possible in the future. However, census-type information will need to be obtained through a survey collection for some time at least. The variables 'unlikely' to be satisfied by administrative data are listed in table 1.

This list may change over time. For example, from April 2020 employers are able to report hours paid on their payroll returns to Inland Revenue. In future this data may provide good coverage and act as a proxy for the hours worked variable.

Table 1

Census variables 'unlikely' to be satisfied by admin data

Variable	Information type	
	Personal[1]	Factual[2]
Access to telecommunications	√	
Disability activity limitations	√	
Fuel types used to heat dwellings	√	
Hours worked		√
Housing quality – damp, mould	√	
Housing amenities	√	
Language spoken	√	
Means of travel to work	√	
Means of travel to education	√	
Mortgage payments		√
Number of rooms		√
Religious affiliation	√	
Tenure of household (owned or family trust)		√
Unpaid activities	√	

1. These variables are unlikely to be suitable for collection by government agencies, and would require new approaches with alternative data sources.

2. These variables could potentially be provided through new collection by a government agency.

Source: Stats NZ

Between these extremes there is considerable variation, with some variables measured well by admin sources but lacking coverage of some parts of the population, and some with high coverage but some issues with measurement. Others have both coverage and measurement issues. For these middle group of variables, some combination of admin and survey data may be required to produce census information to an appropriate quality. As an example, the 2018 Census used admin data for variables with good measurement properties to fill gaps in census responses but did not require full coverage of the population (Stats NZ, 2019b).

Table 2 shows the results of coverage and measurement error assessments from the phase 2 detailed analysis underlying figure 1. Individual variables are provided in table 2a, and dwelling and household variables in table 2b. The results include coverage and accuracy from admin data sources calculated in research to date for the 2013 Census reference date. The middle column provides an updated coverage assessment for the 2018 reference date where past census data was used in addition to admin data. We return to the use of previous census data in a later section discussing low coverage caused by the lack of historical admin data.

Variables with potentially useful admin sources for which we have not undertaken this detailed analysis are number of vehicles (where vehicle registration data is available), occupation, and the derived household variables household income and household crowding. An indication of the reliability of household variables is given by the accuracy of household membership and family type.

Table 2a

Coverage and measurement error quantified for selected census individual variables				
Census variables	2013 Coverage: Admin sources only	2018 Coverage: Admin sources + 2013 Census	Admin accuracy (linked data only)	Derivation of accuracy score Comparisons are with 2013 Census unless otherwise stated
	%	%	%	
Age	100	…	98	Exact agreement
Sex	100	…	100	Exact agreement
Usual residence address				Exact agreement, 2018 data
Territorial authority and Auckland local board (TALB)	99	…	95	
Meshblock	99	…	89	
Address ID	99	…	87	
Ethnicity				Exact agreement
Level 1 Single and Combination	99	…	89	
Level 2 Total responses	99	…	86	
Māori descent(1): Overall	43	83	…	Exact agreement Yes/No
Aged 0-14 years in 2013	89	93	96	
Iwi affiliation	42	…		Range for 12 largest iwi
MOE schools	13	…	20-40	
MOE Tertiary	30	…	30-60	
Birthplace	87	98	97	Exact country agreement
Years since Arrival in NZ	45	94	92	Agreement within one year
Number of children born: All women 15+	42	78	77	Admin agreement or higher than census response
Women born since 1974	38	67	94	
Legally registered relationship status	17	72	88	Exact agreement
Highest qualification	47	61	77	Admin agreement or higher than census response
Study participation	99	…	87	Exact agreement
Status in employment	100	…	96	Exact agreement
Work and labour force status (WKLFS)	91	…	83	Exact agreement
Industry	90	…	100	
Sector of ownership	90	…	100	Exact agreement
Workplace address				Exact agreement
Regional council	90	…	70	
Territorial Authority	90	…	65	
Meshblock	90	…	41	
Total personal income	88	…	86	Admin value within one band, or higher than census
Sources of personal income	88	…	56	Income source agrees

1. DIA birth registrations data only

Symbol: … not applicable

Source: Stats NZ

Table 2b

Coverage and measurement error quantified for selected census dwelling and household variables			
Census dwelling variables	Coverage	Admin accuracy (linked data only)	Derivation of accuracy score Comparisons are with 2013 Census
	%	%	
Tenure (overall)	18	…	
Tenure (rental only)			
Tenancy bonds (MBIE)	57	100	Admin tenure assumed correct
HNZC (housing NZ only)	99	100	
Weekly rent paid by households			
Tenancy bonds (MBIE)	54	82	Within one rental band
HNZC landlord sector	99	91	
Sector of landlord			
Tenancy bonds (MBIE)	57	97	Exact agreement
HNZC	99	100	
Number of bedrooms (rental only)			
Tenancy bonds (MBIE)	53	85	Exact agreement
HNZC (housing NZ only)	99	96	
Age of dwelling(1)	28	…	High quality
Floor area(1)	28	…	High quality for separate houses
Census household variables			
Number of usual residents in household	98	55	Exact agreement
Household membership	98	49	Exact agreement
Family type	59	68	Exact agreement

1 Building consents data only. No linked data comparison.

Symbol: … not applicable

Source: Stats NZ

Reasons for coverage and measurement errors

We now describe the main reasons for lack of coverage or measurement error in the admin sources. We first summarise the coverage gaps and measurement issues for the traditional field enumeration census and contrast these with what we find in the admin sources. Several common factors have emerged through the investigations of admin-derived variables. They highlight where there is potential to improve the quality of admin sources, and how survey-based information might support the use of admin data.

Coverage in the full field enumeration census

The census quality assurance framework (Stats NZ, 2019c) requires at least 90 percent non-missing responses in the census dataset for a moderate quality rating, and 95 percent for high. In the 2013 Census dataset, most, but not all, variables achieved a high or moderate rating. However as noted

earlier, the 2013 Census achieved between 88 and 92 percent coverage for most variables when net census undercount is included. Also, for operational reasons, the traditional full field enumeration census does not include residents who are temporarily overseas by design, who made up 1.8 percent of the ERP in 2013. For all these reasons combined, the field enumeration census typically achieves at best a 90 percent response rate compared with the full New Zealand resident population, and considerably lower for some variables and some groups of the population.

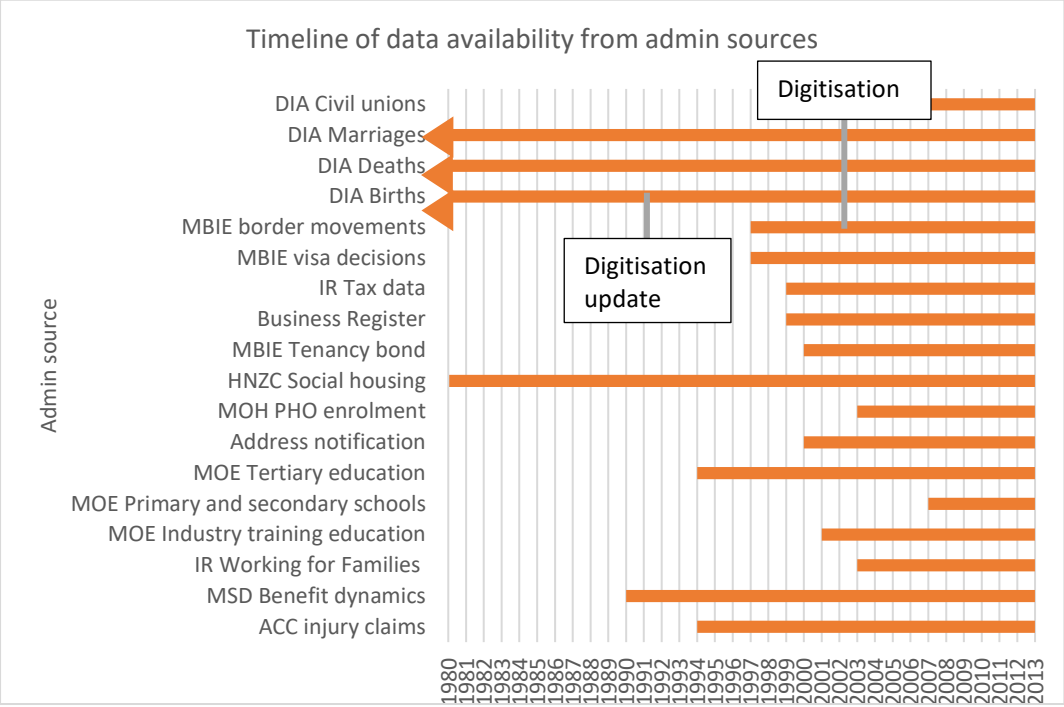
In contrast, for admin data, agencies collect information from people who interact with their service, and different sources can be combined to improve coverage. Resulting patterns of missing data are quite different from the traditional census. The admin information is not affected by those who are absent from New Zealand for short periods, so the target population for statistical purposes can be the full New Zealand resident population as measured by the ERP. The 2018 Census has shown that admin sources are able to count those who are typically missed by field collection (Stats NZ, 2019a). Rather, reasons for lack of coverage and measurement error for variables obtained from admin sources vary depending on the nature of those sources. Some common themes emerge.

Lack of historical admin data

One main reason for lack of coverage in admin sources is the lack of digitised information from earlier periods. Figure 2 summarises the timeline of data availability from admin sources. Digitisation of government admin systems occurred mainly from the late 1990s. For example, MBIE's immigration data is available from 1997, and DIA data (births, deaths, and marriages) was digitised from 1998. This affects many variables where historical information is not available for much of the adult population in 2013. Examples include birthplace and years in New Zealand for migrants who arrived before 1997, and child-parent relationships for women born before 1974. Historical data for Māori descent for those born in New Zealand was not collected by DIA before 1995, and most of our analyses included birth registration data only from 1998 when registration data was fully digitised.

Coverage due to lack of historical admin information may be improved in several ways. DIA recently digitised more information from birth registration records for another 10 years back until 1989. Digitisation of older records is resource intensive, and only likely to be undertaken if there is strong demand.

Figure 4



While many government agencies are able to provide data to Stats NZ for statistical purposes, the Electoral Act 1993 places restrictions on the use of information on the electoral roll. Comparisons using aggregate data suggested that the combination of birth registrations for those under 18 years of age, and electoral roll data for those of voting age, would provide high coverage of the population for Māori descent (Bycroft et al, 2016). However, current legislation precludes the use of unit record electoral roll data as an admin source for Māori descent for the census.

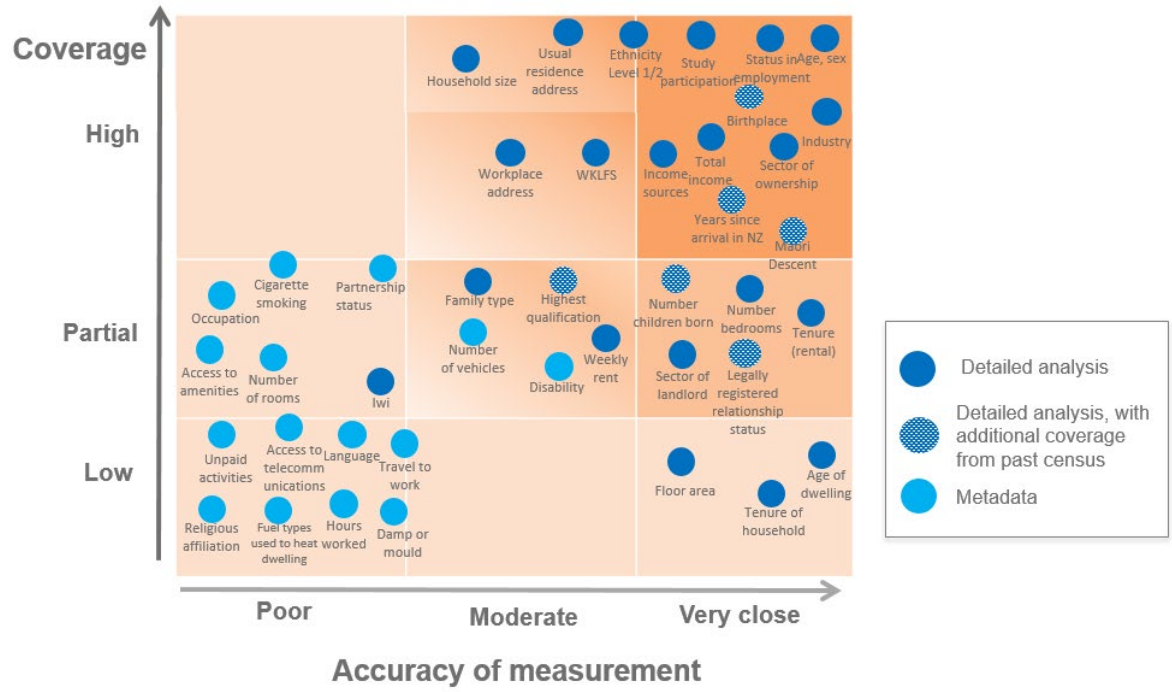
The census itself is another source of information for the whole population. For variables which do not change over time (for example, country of birth and year of arrival in New Zealand), or can be updated for more recent events (for example, highest qualification and number of children born), previous censuses can provide information for much of the population that is missed by the admin sources. A previous census is particularly beneficial for events that occurred prior to digitisation of the admin sources. This approach was used in the 2018 Census, where information from those who responded to the 2013 Census was used as a first option for filling in missing data for several variables (Stats NZ, 2019b).

Table 2a shows how the coverage of selected variables can be increased by including previous census information. Here the reference date is the 2018 Census, with the 2013 Census used as an additional data source. This use of 2013 Census data increases coverage to various degrees. Birthplace and years in New Zealand reach coverage rates of 98 and 94 percent, respectively, providing very high rates of coverage for identifying migrant populations. Coverage of Māori descent overall increases significantly from 43 percent to 83 percent, with remaining missing data mainly affecting adults. There are also good increases in coverage for number of children born, legally registered relationship, and highest qualifications, to achieve between 61 and 78 percent coverage. The 2013 Census provides additional information mainly for older adults, and coverage gaps are concentrated among young adults who are not present in the admin sources because they have no

children, have not married, or have no qualifications. Figure 3 updates the coverage in the previous illustration combining admin data with 2013 Census where appropriate.

If an admin-based census were to be implemented, historical census data would likely be used as a data source where appropriate. As time goes on, the issues with historic data become less relevant as older people die. The admin sources will gradually be used for a greater proportion of the population, and any gaps in the ongoing collection of admin sources will have a larger impact on coverage.

Figure 5 Accuracy of admin data including previous census



Note: Subject to change; results as at November 2020.

Events that occur overseas

Some events are not captured in the New Zealand admin data, because they do not occur in New Zealand. These include

- children born overseas
- marriage, civil unions, or divorce that occur overseas
- qualifications obtained overseas
- income earned overseas.

This can lead to under-coverage, for example if a person's qualifications were all gained overseas, their qualifications will be missing. It can also lead to measurement error, for example if some qualifications are gained in New Zealand, but a higher qualification is gained overseas.

These data gaps affect new migrants to New Zealand more than those born in New Zealand. While previous censuses can provide some historical information for overseas events, they will continue to be a source of undercoverage in admin sources unless there is some means of capturing them on an

ongoing basis. Visa applications could potentially provide an avenue for collecting more information for new migrants. A small amount of qualifications data on principal applicants for some types of visas is already collected as part of the visa application process, although this is not available in the IDI.

Positive identification only

Some statistical information is derived from transactions or events recorded by government agencies, for example taxable income earned, qualifications received, enrolment in study, legal marriage or civil union, or the birth of a child. The admin sources provide a positive identification for these variables, but it is not possible to directly identify those with no income, no qualifications, not studying, those never married, or women who have had no children. These null categories are important for statistical purposes and can be asked directly in a survey questionnaire, however in the admin sources they cannot be distinguished from people with no information. Recording only positively identified events leads to some lack of coverage. For example, there is no admin information for those with no qualifications. We cannot assume that no admin data implies no qualifications since qualifications may have been gained before digital records are available, and even for recent events, missing data could also be due to qualifications gained overseas.

This situation can also be considered as a source of measurement error as the null category cannot be derived directly and needs to be estimated. If there is full coverage of those in the positive categories, the 'missing' data could be assumed to mean 'no' – for example almost all those who are studying at a given time are enrolled in a New Zealand educational institution and recorded in admin sources, so it could be assumed that those not enrolled are not studying. However, this is not the case for income, where just 40 percent of those who have no income information in the tax data also reported zero income in the 2013 Census (Suei, 2016). In general, an imputation model supported by some survey information is likely to be needed to allocate missing data between the null and other positive categories.

In a similar way, for some variables, very good admin data can be obtained for some parts of the classification, but other categories are not available. For example, marriage and civil unions registration systems, divorce records, and death registrations all provide high quality information for the census legal partnership status variable; however, none record information on those who are separated. For household tenure, Housing NZ Corporation and tenancy bonds data identify many of those renting, but we do not have sources that identify dwellings owned or held in a family trust.

The reverse argument also helps to understand measurement error in surveys. Where admin systems are robust (for example, when there are strong legal or financial requirements), we can be confident that where events are positively identified, these are likely to be correct, and a survey response less than the admin measure is likely to be an error. For example, total personal income is unlikely to be less than that reported in the tax system, and a mother is unlikely to have had fewer children than those recorded in birth registrations.

Linkage error

In the IDI, data sources are linked to the IDI central population spine using probabilistic linkage. Linkage is not perfect and linkage errors can affect the use of the admin data for statistical purposes. An incorrect link may mean that a person is assigned the wrong value for a variable, a form of measurement error. The low false link rate under 2 percent for IDI linkages means that any impact is likely to be small.

Missed linkages can result in missing data for a person. The linkage rate varies considerably by data source. For example, because tax registrations are used to form the IDI spine, there are minimal missed linkages for recording payment of tax over time. However, at the other extreme, the parents of children registered as born in New Zealand have very low linkage rates before 1989 because the information needed to determine the link has not been digitised. Linkage rates have tended to increase over time, for example the linkage rates for mothers on birth registrations was almost 100 percent by 2013 (Miller et al, 2019).

Address matching rates for admin sources related to dwellings can reduce the effective coverage of dwellings in the census context. For example, 30 percent of tenancy bonds records were not able to be matched to an address ID (Miller et al, 2018). From March 2019, IDI address matching has used the Statistical Location Register and new address matching software. This has led to an improvement in the quality of IDI address matching (Bycroft et al, 2021), however most of the census transformation analysis reported here was carried out before these improvements.

Unit error

Dwellings and households are statistical units that must be derived from admin address data. Where we have one address associated with two or more dwellings, this is a form of unit error. Similarly, we construct an administrative household by grouping people who report living at the same address. Unit errors may occur if we are in fact combining two or more households, or when an incorrect address results in the wrong people being associated with a household. While the admin address information provides good accuracy for larger geographic areas, accuracy decreases with smaller geographic areas, and the construction of admin households is problematic (Gath & Bycroft, 2018).

The variety of admin reporting units related to properties introduces more complexity when deriving a dwelling. Reporting units include bond lodgements, property IDs, houses, and buildings (Bycroft et al, 2021). The rules used to determine a dwelling may introduce unit error, particularly when the reporting unit includes multiple dwellings.

Time references and timeliness

A census questionnaire is answered with respect to a specific date, and results are generally released 9 to 18 months after census day. Both of these aspects related to timing can affect the coverage and measurement error for data derived from admin sources.

Most admin variables are able to be referenced to a specific date, with some exceptions. As a self-identified characteristic, ethnicity can change over time. If a respondent has not had a recent interaction with an agency it may mean that their ethnicity may be out of date. Weekly rent from tenancy bonds is another example as the rental amount is for the start of the tenancy and is not updated if rents are increased for the same tenancy. Time lags in the notification of address changes to agencies means that admin addresses can be out of date. In these cases, a person's information held in the IDI may therefore not match what they would have responded at the time of the census.

Time lags in receipt of admin data may affect what data is available for release in a timely manner. Lags can occur between the point of collection by government agencies, the supply of that data to Stats NZ, and integration to the IDI. Our experience with use of admin data in the 2018 Census gives an indication of how time lags affects availability (Stats NZ, 2019b). Admin sources were captured in September 2018, six months after census day. Most sources included data that covered census day in March 2018, with two exceptions. Ministry of Education qualifications and enrolments were available through to the end of 2017, but not for 2018. The delay was due to a lag in the standard

supply of information to Stats NZ, which could be overcome in future. Tax data provides information about income and employment, but not all tax information for the self-employed was available through to census night. This is more difficult to overcome as self-employed tax is mainly filed annually, and it can take more than a year before data from a processed annual tax return is supplied to Stats NZ.

Other errors of measurement

As well as the sources of measurement error noted above, there are a number of other reasons why a value recorded in a survey questionnaire or derived from an admin system may not correspond to the desired target measurement.

The target concepts for the admin-derived variables in table 2 in most cases do closely align with the target concepts described in the statistical standard, also used by census, and this provides a sound starting point for their statistical use. An exception is unemployment, where receipt of certain benefits did not map well to those who reported being unemployed in the census (Jer, 2020).

The operational collection, and subsequent processing of the data, can introduce errors that mean the recorded value is different from the target concept. As noted earlier, this is true of both a survey-based and an admin system. Errors can occur in a survey context, for example census respondents may find some questions difficult to answer precisely, they may misinterpret a question, or falsify an answer; responses on a paper form may not scan correctly; or incorrect coding of text responses may introduce errors. As one example, census responses to study participation underestimate those enrolled in industry training courses (Shrosbree, 2015). Errors in census responses will downgrade the admin consistency measures if the census is assumed to provide a correct response.

Admin systems are also subject to collection and processing errors which are more likely to affect variables that are not essential to the core operation of the agency. For example, ethnicity is collected by several government agencies which aim to adhere to the target statistical concept. However, for some of these agencies, issues in collection or processing resulted in a tendency to report fewer people with multiple ethnicities than the census (Reid et al, 2016). Some agencies aim to collect iwi using the appropriate statistical concept, but in practice the data collected did not compare well with the responses obtained by the census (Bycroft et al, 2016).

The methods used to derive the statistical variable from admin sources may also introduce error, particularly where there are multiple data sources and potentially conflicting values. Address of usual residence and ethnicity are two important census variables where multiple sources are combined. Rules-based methods have been developed to derive the ‘best’ address and ethnicity for an individual, however model-based methods would likely provide improved estimates.

From table 2 we see that for most variables, measurement accuracy ranges from around 80 percent to 99 percent. Exceptions are iwi affiliation, small area workplace, and household membership which drop to around 50 percent accuracy in comparison with census responses. There are particular issues associated with admin collection for these last variables.

Collection of iwi requires tight control of all aspects of data collection from questionnaire design, who fills out responses, and subsequent processing and coding systems. This can be difficult to achieve, especially in distributed collection systems. The coverage of iwi in government agencies is also limited at present. These issues have been highlighted by the low response for iwi in the 2018 Census, and the lack of suitable alternative sources. The Mana Ōrite relationship agreement was signed between Stats NZ and the Data Iwi Leaders Group in 2019. The associated work programme

includes a workstream to improve administrative data to ensure a sustainable and diversified flow of relevant iwi data for Māori [4].

Workplace address is not available through the tax system for enterprises with multiple geographic locations.

All household variables depend on placing the correct group of people at an admin address. More accurate address information for individuals is required to improve on the accuracy of admin-derived households. Some improvements have been seen between 2013 and 2018 with address information used in the 2018 Census, mainly due to improved address matching by Stats NZ (Stats NZ, 2019a). In recent work, the Social Wellbeing Agency (2020) finds that using the September 2019 IDI refresh, 63 percent of admin households have the same membership as the 2013 Census. However, at present we are unable to construct high quality admin households. Missing family relationship information also affects the ability to derive household variables.

Improvements to admin data quality across government

The need for high quality admin data, particularly for core demographic variables, has been recognised for some time and Stats NZ has been working with agencies to improve the quality of administrative data collection (Cabinet paper, 2016). A number of system improvements were already in place, or under development, in 2016. It was anticipated these would lift data quality over time, but that this would be a gradual process. Given the timeframes, any improvements are unlikely to be reflected in the investigations reported here.

The role of Government Chief Data Steward (GCDS) established in 2017, recognises the importance and value of government data. The GCDS has been supporting work across government to co-design, develop, and implement short data content standardisation, and has the authority to set mandatory requirements across government. Three data content requirements have to date been approved for Date of birth, Person name, and Street address and others are in progress or planned[5]. Data content requirements set out details of how the information is to be collected and applies to data being shared between organisations. They complement existing statistical and other related standards. Improving data content standardisation practices across government organisations will enable more efficient and effective sharing of data, within existing privacy and security settings.

One of the risks associated with the use of administrative sources for statistical purposes is that policy changes can affect the data collected. While some policy changes may be detrimental to the statistical use, others can be beneficial. For example, the transformation undertaken by Inland Revenue to modernise the taxation system has provided more detail for measures of employment and income. From April 2019, all employers have switched from filing the monthly EMS to filing pay period records which include pay period start and end dates. This could bring the admin paid employment measures more in line with the reference week concept used by the census. The full-time and part-time split for paid employees becomes possible after the addition of paid hours in pay period records from 1 April 2020. Paid hours may also provide proxy information for hours worked.

[4] **Mana Ōrite Relationship Agreement** was signed between Stats NZ and the Data Iwi Leaders Group of the National Iwi Chairs Forum in October 2019. The purpose of the relationship is to work together with iwi-Māori to realise the potential of data to make a sustainable, positive difference to outcomes for iwi, hapū, and whānau. A workstream of the Mana Ōrite Work Programme (see 2018 Census iwi data: October 2020 update) is focused on improving administrative data to ensure a sustainable and diversified flow of relevant iwi data for Māori.

[5] Register of government data content requirements. Retrieved from data.govt.nz. Last updated April 2020.

A New Data and Statistics Bill is expected to be introduced to Parliament in 2021, which will replace the Statistics Act 1975. The new Act will provide the tools necessary for the Government Statistician to lead and co-ordinate across the system, produce necessary statistics, and advise on and give effect to the data and statistical priorities of the Government, including the census. Once the bill becomes law, amendments will be made to a number of other Acts to remove legislative barriers that inadvertently restrict or prohibit the provision of data to Stats NZ (for example, amendments to the Electoral Act 1993 will enable Stats NZ to access electoral data). The new Act will also include provisions that enable the Government Statistician to authorise government agencies to collect data for official statistics on behalf of the Government Statistician.

Admin data opportunities

Admin-based data sources can also provide added value over the current census requirements. A common benefit is the more frequent supply of information. Admin sources are not restricted to a five-yearly cycle, and could be produced annually, for example.

Admin sources can be more detailed and precise than is possible in a census questionnaire context. For example, admin-derived personal income is available in dollar amounts and can be tabulated by each income source separately (Suei, 2016). Other examples include where admin sources can provide age at first birth for fertility studies (Miller et al, 2019), and dates of completion for educational qualifications (Shrosbree, 2015).

And as noted above, some admin sources are more accurate since they record actual events, and do not suffer from respondent recall, misinterpretation of questions, or unwillingness to answer. For example, census respondents tend to under-report receipt of income from government benefits that are recorded in administrative systems, or may report income net of tax instead of gross income.

Admin sources could be used to provide annual data for small areas to supplement official measures from existing household surveys such as the Household labour force survey. High quality admin labour market measures include estimates of employed and not employed based on taxable income sources, as well as those in receipt of work-ready benefits as a proxy for unemployment (Jer, 2020).

The linked admin data are inherently longitudinal by design, with known dates for many of the events that lead to changes in characteristics, providing a powerful means of analysing change over time. The five-yearly censuses have been linked to provide the longitudinal census dataset (Stats NZ, 2014d), but this does not have the granularity of linked admin data and relies on responses to each census to preserve the longitudinal analysis over time.

Admin sources can also provide some information currently not asked in census. One relevant example is the age and floor area of dwellings, which relate to housing quality and are not asked in the census but are available in admin sources (Bycroft et al, 2021).

Admin data can also be more flexible than a five-yearly census in meeting new information needs. When admin sources are already integrated into the wider statistical system, new variables may be derived from existing sources without the need to design and test new questions or wait for the next periodic census.

Discussion

The census transformation programme has undertaken a series of investigations aimed at assessing the potential for linked admin data sources to meet information requirements for social and economic characteristics measured in the census. The results summarised here show substantial promise for over half the variables currently collected by the census, while around one-third of the current census variables are clearly not available through admin data at present. Critical areas of focus for improvements in admin sources are the place of usual residence (particularly for young adults), family and household data, and iwi affiliation.

The GCDS is leading coordinated efforts across government to improve the quality and use of government data by collectively managing data as a shared asset within existing privacy and security settings. System changes will lift data quality over time. However, agencies are at different stages of development, and do not all have the funding necessary to update complex IT systems. The new Data and Statistics Bill, once passed, will facilitate the collection and use of administrative data for the benefit of the wider statistical system.

Any method of obtaining census information will include some missing data and reported values which differ from the target concept. The full field enumeration census in New Zealand measures and reports on census under-coverage and provides information about the quality of census variables. Prior to the 2018 Census, missing data was not imputed for most variables, in effect applying an implicit model that non-respondents were similar to respondents. The 2018 Census was the first time that alternative sources and imputation were used to adjust for missing census responses. For example, use of admin data in the 2018 Census has improved the quality of the count for those typically missed by the field enumeration, particularly Māori and Pacific populations.

The level of coverage for many of the variables that can be obtained from admin sources (sometimes combined with previous census data), reaches a similar level to that achieved by the full field enumeration census. However, the error structures of admin sourced variables are quite different from the traditional census. Different groups are more likely to be missing, and sources of measurement errors are also quite different. As with the traditional census, an estimation process is needed to adjust for biases due to missing data or other errors. Imputation methods will need to be developed possibly combined with information from a survey to estimate for specific groups, for example new migrants, or specific categories like ‘null’ responses.

There is now a significant body of census-type information that can be collated from admin data for most of the New Zealand population. An admin-based NZ resident population provides a good approximation to the official population estimates, and around half the attribute variables have high or partial coverage from administrative sources. Experimental population estimates have been released for the core demographic variables age, sex, and ethnicity by subnational geography for the years 2006 to 2016 (Stats NZ, 2017; 2018). We are now in a position to progressively add more census variables derived from admin sources to this population dataset, and to continue to update it over time. This experimental ‘Admin Population Census’ will allow the admin variables to be produced by various demographic breakdowns including ethnicity, and cross tabulated with other variables providing more of the multi-variate and small area information that is the heart of a census. Our aim is to demonstrate the breadth of admin information available and to provide a focus for discussion with customers about the quality issues and benefits associated with the admin-based variables.

We envisage that this will become a stand-alone source of admin-based census information, regularly updated, and independent of the five-yearly census. It is also clear that admin data alone cannot provide the full range of topics traditionally collected by the census, and there will continue to be a requirement for collection of some variables through a questionnaire.

References

Bycroft, C, Reid, G, McNally, J, & Gleisner, F (2016). Identifying Māori populations using administrative data: A comparison with the census. Retrieved from www.stats.govt.nz.

Bycroft, C, Miller, S, Das, S, & Goodyear, R (2021) Administrative sources for census housing information: An overview. [forthcoming]

Cabinet paper, (2016). Census transformation progress update: Realising the potential of government data. Retrieved from www.stats.govt.nz.

Gath, M, & Bycroft, C (2018). The potential for linked administrative data to provide household and family information. Retrieved from www.stats.govt.nz.

Gath, M, & Das, S (2019). Potential for admin data to provide country of birth and years since arrival in New Zealand information. Retrieved from www.stats.govt.nz.

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

Jer, R (2020). Comparing labour force variables from 2013 Census and administrative data. Retrieved from www.stats.govt.nz.

Miller, S, Didham, R, & Bycroft, C (2019). Comparing 2013 Census and admin data for number of children born. Retrieved from www.stats.govt.nz.

Miller, S, Suie, S, & Bycroft, C (2018). Comparing housing information from census and tenancy bond data. Retrieved from www.stats.govt.nz.

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

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

Social Wellbeing Agency (2020). Constructing households from linked administrative data: An attempt to improve address information in the IDI (opens a PDF, 1.02MB, 30 pages). Retrieved from https://swa.govt.nz.

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

Stats NZ (2007). Principles and protocols for producers of Tier 1 statistics. Retrieved from www.stats.govt.nz.

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

Stats NZ (2014b). Coverage in the 2013 Census based on the New Zealand 2013 Post-enumeration Survey. Retrieved from www.stats.govt.nz.

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

Stats NZ (2014d). Linking censuses: New Zealand longitudinal census 1981–2006. Retrieved from www.stats.govt.nz.

Stats NZ (2016). Guide to reporting on administrative data quality. Retrieved from www.stats.govt.nz.

Stats NZ (2017). Experimental population estimates from linked administrative data: 2017 release. Retrieved from www.stats.govt.nz.

Stats NZ (2018). Experimental ethnic population estimates from linked administrative data. Retrieved from www.stats.govt.nz.

Stats NZ (2019a). Overview of statistical methods for adding admin records to the 2018 Census dataset. Retrieved from www.stats.govt.nz.

Stats NZ (2019b). Data sources, editing, and imputation in the 2018 Census. Retrieved from www.stats.govt.nz.

Stats NZ (2019c). Data Quality Assurance for 2018 Census. Retrieved from www.stats.govt.nz.

Suei, S (2016). Comparing income information from census and administrative sources. Retrieved from www.stats.govt.nz.

Zhang, L-C (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica 66*: 41–63. https://doi.org/10.1111/j.1467-9574.2011.00508.x

Appendix 1: Admin sources and references by variable

Table A1

A1 Admin data sources and references for census transformation attribute investigations

Admin data sources and references for census transformation attribute investigations			
Variables by census topic		**Admin sources**	**References**
Location	Usual residence	Addresses from MOH (NHI, PHO), IR tax, MOE schools, MSD benefits, ACC[1], NZTA	Stats NZ (2017), Stats NZ (2019a)
Population Structure	Birthplace	DIA births, MBIE Border Movements	Gath and Das (2019)
	Years since Arrival in NZ	MBIE Border Movements	Gath and Das (2019)
	Legally registered relationship status	DIA Marriages, civil unions and deaths	Internal Stats NZ report
	Number of children born	DIA Births, MBIE visas and border movements	Miller et al (2019)
Ethnicity and culture	Ethnicity L1, L2	DIA births, MOH NHI, MOE, MSD benefits, ACC	Reid et al (2016), Stats NZ (2018)
	Māori descent	DIA births since 1998; Electoral enrolments	Bycroft et al (2016)
	Iwi affiliation	MOE	Bycroft et al (2016)
	Language	…	Bycroft et al (2016)
Education and training	Highest qualification	MOE	Shrosbree (2015)
	Study participation	MOE	Shrosbree (2015)
Income	Sources of personal income	IR tax: EMS, annual tax returns	Suei (2016)
	Total personal income	IR tax: EMS, annual tax returns	Suei (2016)
Work	Status in employment	IR tax: EMS, annual tax returns	Jer (2020)
	Work and labour force status	IR tax: EMS, annual tax returns; MSD benefit dynamics and WFF; MOE, ACC	Jer (2020)
	Industry	IR tax EMS; Stats NZ BF	Internal Stats NZ report
	Workplace address	IR tax EMS; Stats NZ BF	Internal Stats NZ report
	Sector of ownership	IR tax EMS; Stats NZ BF	Internal Stats NZ report
Families and households	Household counts	IDI addresses as above	Gath & Bycroft (2018)
	Number of usual residents in household	IDI addresses as above	Gibb & Das (2015) Gath & Bycroft (2018)
	Family type	DIA (births, marriages, civil unions), WFF (MSD and IR), MBIE visas and border movements	Gath & Bycroft (2018)
	Household composition	IDI addresses and family type information	Gath & Bycroft (2018)
Housing variables	Weekly rent paid by households	MBIE tenancy bonds, HNZC social housing	Miller et al (2018) Bycroft et al, 2021
	Sector of landlord	MBIE tenancy bonds, HNZC social housing	Miller et al (2018)
	Tenure (rental only)	MBIE tenancy bonds, HNZC social housing	Bycroft et al, 2021
	Number of bedrooms	MBIE tenancy bonds, HNZC social housing, Quotable Value	Miller et al (2018) Bycroft et al, 2021
	Age of dwelling	Building consents, Quotable Value	Bycroft et al, 2021
	Floor area	Building consents, Quotable Value	Bycroft et al, 2021

1. National Health Index Number (NHI), Primary Health Organisation (PHO), Accident Compensation Commission (ACC)

Source: Stats NZ