Statistics New Zealand
TATAURANGA AOTEAROA

An initial investigation into the potential for administrative data to provide census long-form information

Census Transformation programme

Ellenor O'Byrne

Christine Bycroft

Sheree Gibb

New Zealand Government

Disclaimer
This report 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. Unless otherwise stated, results presented are the result of data analysis undertaken by the author.

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

Citation
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.* Available from www.stats.govt.nz.

ISBN 978-0-478-42908-4 (online)

Published in July 2014 by
Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

Contact
Statistics New Zealand Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4610
www.stats.govt.nz

Contents

List of tables and figures

List of tables

List of figures

1 Background

Census Transformation

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 creating efficiencies
- a longer-term focus on investigating alternative ways of producing small-area population and socio-demographic statistics, including the possibility of changing the census frequency to every 10 years, and exploring the feasibility of a census based on administrative data (Statistics New Zealand, 2012).

The main emphasis of the longer-term strand of the Census Transformation programme is on the feasibility of producing census information from existing administrative data sources, as this aspect is the least understood. This investigation takes a phased and iterative approach.

The first phase of the investigation (which this paper is part of) is designed to provide evidence that will inform decisions on the preferred direction for future development of the census (see Statistics NZ, 2014 for an overview). The early focus is on developing an understanding of future census information requirements, and whether existing administrative sources can meet those requirements.

While not excluding the possibility of using commercial, social media, or other big data sources, investigations in the first phase are based on administrative data generated through government activities.

About this paper

This paper forms part of the first phase of the Census Transformation programme. It provides a first broad look at the potential for administrative data to produce the long-form (social and economic) information currently provided by the census. This paper identifies administrative data sources that are related to these census topics, and uses quality measures to assess how likely these sources are to satisfy the information needs currently met by the census.

The purpose of this paper is not to make final decisions about the potential use of administrative data for census information, but rather to:

- provide an early indication of the likely ability of existing administrative data sources to produce census long-form information
- provide reference information about administrative data sources using key quality measures
- guide decisions about where to direct more in-depth analysis
- highlight any areas of considerable potential for using administrative data in the next census.

2 Introduction

What does the current census do?

The New Zealand Census of Population and Dwellings has been held since 1851, usually on a five-yearly cycle required by legislation (the Statistics Act 1975).

The census is the official count of people and dwellings, and aims to count everyone who is usually resident in New Zealand. The census produces counts of dwellings and people down to small geographic areas. In addition to basic information about population size and demographic make-up, the census gives an accurate and comprehensive picture of the social and economic characteristics of local communities and small population groups. Other data sources have not yet been able to provide reliable and consistent information at this geographic level (Statistics NZ, 2012).

The content of the census is classified into 'census topics'. Each of these topics relates to one or more questions from the census form and one or more output variables. These census topics can be grouped by subject category. For example, 'highest qualifications' and 'study participation' are census topics within the 'Education and training' subject category. The subject categories used for this study are:

- population structure
- ethnicity and culture
- education and training
- income
- work
- health and disability
- families and households
- housing.

Topics may change from census to census. Over the 20th century, census topics have changed significantly as society and its information needs have changed. Over time, census topics can be added or removed, or can appear periodically; for example, where information is needed every 10 years instead of every five.

Census topics are also grouped by 'quality levels', which helps guide resource and processing decisions for reaching data quality standards for particular topics and variables (Statistics NZ, 2009). The census topic and output variable quality-level groups are:

1. Foremost topics/variables: census output variables that the Statistics Act 1975 and the Electoral Act 1993 require Statistics NZ to produce. Their outputs are the key reasons for conducting a census, and inform population estimates.
2. Defining topics/variables: define key subject populations that the census measures. They are frequently used in cross-tabulations with foremost variables. The census has traditionally been the only detailed (ie subnational) source for this information.
3. Supplementary topics/variables: important to certain groups, but are not a primary purpose of the census.

What is an administrative census?

Administrative data is data that government agencies or private organisations collect in conducting their business or services. It is data that is not collected primarily for statistical purposes. An administrative census would provide census information using data

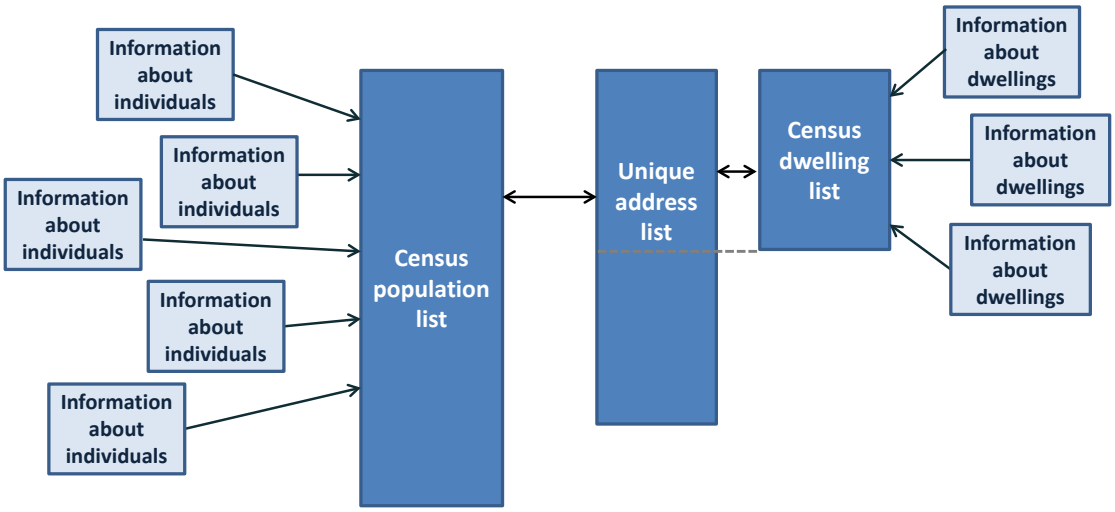
collected largely through administrative sources (for example, data that government agencies collect through taxation, or through registering births and deaths), instead of a survey of the whole population.

While there are a number of ways to conduct a census based on administrative data (Bycroft, 2013), our current investigations focus on a system of linked administrative data, similar to the 'register-based' census approach of the Nordic countries (United Nations Economic Commission for Europe, 2007). This approach aims to create a unit record (ie 'person-level') list of the whole population – a 'census population list', derived from administrative sources. The list would include attributes of people, such as education qualifications or income, from a range of sources linked to their list record. In addition to a census population list, a list of unique dwellings – a 'census dwelling list' – would also be needed.

Matching individuals in the census population list to dwellings in the census dwelling list, by address, so that individual information is linked to the dwelling where the individual lives, is also essential to this approach. This matching would allow us to produce statistics for subnational geographic areas, as well as allowing households to be formed by grouping people who live at the same address.

See figure 1 for a conceptual diagram of this approach to an administrative census.

Figure 1
Conceptual diagram of an administrative census



A census population list could be formed in different ways. Some countries that have an administrative census base their census population list on an official population register, which formally identifies all people resident in the country. New Zealand does not have such a register. One would either need to be developed for wider government use, or a statistical version of a census population list would need to be built from one or more existing data sources with good population coverage. Either of these methods would produce a census population list consistent with using linked administrative data sources for information about individuals and dwellings, as discussed in this paper.

Other work within the Census Transformation programme is considering the creation of a census population list and suitable address information to produce population statistics (Gibb and Shrosbree, 2014).

Assumptions

This assessment is about census social and economic information, which includes almost 50 topics. To remain focused and manageable this investigation requires us to make some assumptions. These are:

- a system of linked administrative data sources can be established that would provide the necessary conditions for linking additional sources of information about individuals and dwellings. As described above, these are:
    - a good census population list
    - a good census dwelling list
    - individuals in the census population list can be matched to the dwelling where they live.
- the current census is a reasonable proxy for the information census will need to provide in future.

This paper assesses administrative data sources against the current census topic requirements. It does not address whether users in future may require information that is markedly different from what the current census provides; for example, information that is more or less frequent, that is longitudinal rather than cross-sectional, of higher or lower quality, or about different census topics.

The Census Transformation programme is exploring the extent to which future users' needs are different from current needs. Further work on administrative sources can take any changes in user needs into account once these are better determined.

Scope

This assessment is a first-stage analysis and requires a defined set of census topics as a starting point. For this assessment, the starting point is the content of the latest New Zealand census in 2013.

Census topics in scope for this investigation are:

- social and economic ('long-form') census topics. Generally, this means all 2013 Census output topics except the core demographic variables (age, sex, and location)
- disability. Although not an output variable for census, disability is included because the census disability question is used to identify candidate respondents for the Disability Survey.

Topics out of scope for this investigation are:

- population counts, and census topics that are primarily demographic (age, sex, and location). Demographic topics and the ability to provide geographic location information are investigated elsewhere in the context of population statistics (Gibb and Shrosbree, 2014).
- dwelling counts
- other location-related census topics that rely on the address and identification of dwellings (such as number of occupants, absentees, address five years ago, and years at usual residence).
- census questions that are not outputs, and are used only to derive other variables (for example, living arrangements) or for operational purposes (for example, name).

Table 1 shows what census topics are in scope, and their priority status.

Table 1
Census topics that are in scope for this assessment

Subject category	Census topic	
	Defining (or foremost) census topics	**Supplementary census topics**
Population structure	Legally registered relationship status / Partnership status in current relationship	Number of children born
Ethnicity and culture	Ethnicity (foremost) Māori descent Iwi affiliation Birthplace	Language Religious affiliation Years since arrival in New Zealand
Education and training	Highest qualification Post-school qualification Highest secondary school qualification Study participation	…
Work	Status in employment Work and labour force status Hours worked in employment per week	Industry Workplace address Sector of ownership Unpaid activities Occupation Main means of travel to work
Income	Sources of personal income Total personal income	…
Health and disability	…	Cigarette smoking Disability
Families and households	Family type Extended families Household composition	…
Housing	Occupied dwelling type Weekly rent paid by households Sector of landlord Tenure of household	Tenure holder Number of bedrooms Number of rooms Access to telecommunication systems Fuel types used to heat dwelling Number of motor vehicles

Symbol:

… not applicable

Source: Statistics New Zealand

Aims

This paper asks whether administrative data can meet information needs that are usually satisfied by long-form census questions.

The aims of this project are to:

- identify the likelihood that each in-scope census topic could be satisfied by administrative data sources
- identify the subject categories and census priority areas that are more or less likely to be satisfied by administrative data sources
- identify the administrative data sources most critical for informing census topics.

3 Methods

This section describes the methods used to carry out this investigation. A wide-ranging search was undertaken to identify administrative data sources related to the census topics. The concept of quality was used to assess the suitability of administrative data sources for a census topic. The quality of statistical information can be assessed over a number of dimensions and in varying levels of detail. A set of five quality measures was developed and applied for this assessment.

Data sources

The first step was to identify administrative data sources that have information potentially related to the census topics. This was done by:

- reading both published and unpublished Statistics NZ documents
- talking with Statistics NZ topic experts and data custodians
- searching the Internet for related topics.

The next step was to evaluate the nature and content of potential administrative data sources. This was done by:

- reviewing published papers that used the administrative data sources
- reviewing published or internal data dictionaries and other metadata
- reviewing forms and other guidelines used by the source agencies during the collection process
- discussing the administrative data sources with Statistics NZ experts and data custodians
- questioning the source agencies.

This information about the administrative data sources (metadata) was then used to assess their suitability. No analysis using the administrative data sources themselves has been undertaken at this stage.

Quality measures

Five quality measures were used in this assessment: relevance, accuracy of coverage, accuracy of linkage, timeliness, and accessibility. This section outlines how the quality measures were developed and describes each measure in more detail.

Quality measures for assessing administrative data sources were identified by:

- reviewing quality models used internationally in official statistics
- determining which quality measures from these quality models were most critical for this first assessment stage.

The quality of statistical outputs is defined from a customer/user perspective. The most general and succinct definition of product quality is fitness for use (Eurostat, 2009). National statistics offices have developed a concept of statistical quality that measures the quality of data across a number of distinct, but interrelated, dimensions.

The quality model used by Statistics NZ has six dimensions: relevance, accuracy, timeliness, accessibility, consistency, and interpretability. This quality model is very similar to the quality frameworks used by other agencies that produce official statistics (Australian Bureau of Statistics, 2009; Statistics Canada, 2009; Office for National Statistics, 2013; Eurostat, 2009 and 2011). This was a starting point, but it was still

necessary to select specific quality measures for the assessment within some or all of these six dimensions.

The quality measures used to assess administrative data sources were chosen for their relevance in this initial assessment. Ideally, measures will be strongly discriminatory, in the sense that they are essential for the use of administrative data for census information, but will also be measures for which reasonable judgements can be made from metadata.

The focus of this first assessment is on understanding the main features of the administrative data sources themselves, before any statistical processing is undertaken. The ability to link to other administrative data sources is crucial for using administrative data for census, and this forms an additional criterion beyond the original administrative purpose.

Other quality measures not chosen may be important for the use of administrative data for census, but are not critical for this assessment at this stage. For example, interpretability and consistency dimensions are not considered here.

Table 2 shows the five quality measures used in this assessment and the questions that are most relevant to each measure.

Table 2
Quality measures for evaluating the potential for administrative sources

Quality measure	Main questions for assessment
Relevance	How close is the administrative data to the statistical concept? (the census topic is used as a proxy for the statistical concept) Who/what should be included in this data? (target population) Who/what is actually included in this data? (observed population)
Accuracy of coverage	Are there missing people or responses? (undercount) Are there duplicate records or other people who should not be included? (overcount)
Accuracy of linkage	Is it possible to link the data to the census population or dwelling lists?
Timeliness	How frequently is the data updated? How long after the reference date is the data available to Statistics NZ?
Accessibility	Are there privacy or legal issues around using this data? Are there any other barriers to access?

Source: Statistics New Zealand

The following subsections discuss each of these quality measures.

Relevance: Is the data close to the statistical concept?

For administrative data to replace the census as the method for satisfying information needs, the data needs to align closely with the relevant statistical concept. How close the data needs to be to the statistical concept will depend on user needs. Because users' interest in particular statistical concepts cannot be predicted, 2013 Census topics and relevant statistical standards were used as a proxy for the statistical concept of interest. The aim is to have high relevance for the census topic.

An example of where an administrative data source is close to, but not exactly the same as, the census topic is for language. The census asks about languages in which a person can have a conversation about a lot of everyday things. Administrative data is available for the languages a person has enrolled in a course in, or passed NCEA exams for. The difference between languages spoken and languages studied formally in New Zealand represents a relevance gap.

Accuracy of coverage: Who is included in the data?

For administrative data to replace the census, the coverage of the data needs to be good enough to meet customer needs. The aim is to have coverage similar to the current census for the census usually resident population counts and for individual variables.

In the current census model, coverage is maximised by attempting to:

- get everyone in the country to fill out a form once
- minimise non-response for particular questions.

In an administrative census, coverage would be maximised by:

- a census population list, constructed to represent the usual resident population
- sufficient coverage in source datasets, so that when integrated and combined, each census variable would have minimal non-response.

Complete coverage and zero non-response are not feasible, even in the current census. To assess whether an administrative source has adequate coverage, it is necessary to:

1. identify what its overall coverage is
2. assess the characteristics of the over-coverage or under-coverage (missing data).

Depending on the patterns of the missing data, under-coverage can be a severe problem for data use, not a problem at all, or somewhere in between. For example, the coverage of administrative datasets that collect smoking status is very problematic, because they only include those people who see a GP or are admitted to hospitals. This observed population is a small subset of the target population, and the likelihood of an individual being observed (or not) is related to the outcome variable (whether or not someone smokes).

Information about the user's necessary or ideal target population is also required for assessing whether data coverage is sufficient. In general, this investigation assumes that the target population for users of census information is the usually resident population of New Zealand.

Accuracy of linkage: Is it possible to link the data to a census population or dwelling list?

For administrative data to replace the census as the method for satisfying information needs, it must be possible to link the administrative data sources to a census population list or census dwelling list. As with non-response, the goal is not absolutely 'perfect' linkage, as this may never be fully achievable. The aim is to have linkage rates that result in data that is of sufficient quality to use.

In principle, linking requires being able to uniquely identify individuals within each administrative data source, and identify the same individuals in the census population list. In practice, probabilistic techniques may be applied to link records, based on the likelihood that an individual in one source is the same as an individual in another.

At a minimum, a dataset will require at least a few stable characteristics to match across datasets. Stable characteristics for an individual might include name, date of birth, sex, or country of birth. Datasets might also have unique identifiers, such as an Inland Revenue tax number, that could be used to match individuals across datasets that share these

identifiers. However, the Privacy Act 1993 limits this practice in New Zealand. Ideally, a few key variables would be used that are:

- universally available
- fixed
- easily recorded
- unique to that individual
- readily verifiable (Gill, 2001).

If an administrative data source has no strong identifying variables in common with the census population list, it will remain outside the system of linked data. This means the data could not be cross-tabulated with variables in other datasets (a key strength of the census), but it could potentially be used to produce single aggregate distributions.

Timeliness: Is the data updated frequently and available to use quickly?

For administrative data to replace the census as the method for satisfying information needs, the data must be updated frequently and available soon after collection. The main considerations for timeliness are whether an individual's or dwelling's information is updated frequently enough in an administrative system to be relevant to statistical users, and whether the information is available to Statistics NZ quickly enough that relevant statistics can be produced. For example, the aim may be to receive updated information at least annually; however, data updated within two years may be acceptable.

Some administrative sources have rules about how frequently a person must update their data or report an event. For example, births must be registered within two years (in practice, approximately 95 percent are registered within six months of the birth). For other administrative sources, data for an individual may only be updated when that person interacts with the administrative data owner. For example, a person's records in the health system may only be updated when that person visits a GP or hospital.

Some information collected in the census might change rarely for most people, and for this kind of information, less frequent updates in administrative data may be less of a concern. However, it will be important to ensure that there are opportunities to update variables such as ethnicity, iwi, or educational qualifications as these might change for some people.

Accessibility: What would be required to use this administrative data?

For administrative data to replace the census as the method for satisfying information needs, Statistics NZ must have access to the data. The aim is to have legal permission, collaboration with source agencies, and general acceptance from the public to use administrative data sources for census statistical purposes.

While in general, it is mandatory to supply the information Statistics NZ requests under the Statistics Act 1975 (the Act), there may be some legal exceptions (for example, conflicts with the Electoral Act 1993). Furthermore, where Statistics NZ has used administrative data in the past, it has generally aimed to work closely with the source agency in a collaborative manner, rather than demanding information under the powers of the Act. Statistics NZ considers administrative data sources it already has and uses for statistical purposes as more accessible than others. Similarly, having an established relationship with the agencies that collect administrative data is a promising indicator of accessibility.

Determining the likelihood that administrative data could satisfy a census topic

The next step was to assess the group of administrative data sources for each census topic using the quality measures. This was done by assigning an assessment of 'excellent', 'good', or 'poor' to the group of administrative data sources for each census topic, using the guidelines in table 3. Table 3 describes how each of the quality measures would appear for excellent-, good-, or poor-quality data.

Table 3
Definitions of quality ratings

Quality measure	Definition of quality rating		
	Excellent	**Good**	**Poor**
Relevance	The data collected in the administrative sources is very close to the statistical concept.	The data collected in the administrative sources is not exactly the same as the statistical concept, but it is close, or related to a similar statistical concept that might be acceptable.	The data collected in the administrative sources is not at all relevant to the statistical concept we are interested in.
Accuracy of coverage	The coverage is similar to the census.	Most of the population is covered and those who are missing are 'missing at random'.	Coverage is very low, or there is bias in the distribution of missing values.
Accuracy of linkage	Data has excellent individual identifiers that can link the units in one dataset to other datasets.	Data has good individual identifiers.	Data has no individual identifiers. Data linkage is not possible.
Timeliness	Data is updated at least every year and available to Statistics NZ within two years.	Data is updated at least every two years and available to Statistics NZ soon after.	Data is updated sporadically, or with delays of more than two years.
Accessibility	No privacy or legal concerns exist. Statistics NZ understands the data and has a good relationship with the administrative owner.	Some privacy or legal concerns exist with one or more key datasets.	Serious privacy or legal concerns exist. No relationship with administrative owner or no history of using the data.

Source: Statistics New Zealand

This assessment was done by jointly assessing as many administrative data sources as may be needed to satisfy that census topic. For example, if one administrative data source contains information about languages spoken for everyone up to the age of 18, and another administrative data source contains information about languages spoken for everyone aged 18 and older, the final assessment will consider how well both datasets combined would satisfy information needs related to language spoken. In this case, either dataset individually would only have partial coverage of the population, but assessed together, they have high coverage of the population (assuming they each have high coverage of their target age groups).

If two datasets had opposing qualities (for example, one was very timely, the other not at all), the joint assessment usually reflects the lower quality, but gives greater weight to datasets with higher potential coverage. Datasets that cover more of the population are considered to be more important than smaller datasets that might just be supplementary to the analysis.

The final step was to determine the overall likelihood that available administrative data sources could satisfy a census topic. This was done by assigning one of three overall scores: ‘likely’, ‘possible’, or ‘unlikely’ that administrative data sources will satisfy a census topic. The likelihood that administrative data sources will satisfy census information needs is based on evaluation of the data sources against the quality measures described in the previous section. The resulting assessments are:

- **Likely:** A census topic is ‘likely’ to be satisfied with administrative sources if one or more datasets combined are ‘good’ or ‘excellent’ for all five quality measures.
- **Possible:** It is ‘possible’ a census topic may be satisfied with administrative data if it relies on administrative data sources that would be suitable under certain conditions: For example, if certain changes happened (either changes to administrative collections, or changes to user needs), or modelling proved to be an acceptable approach to addressing quality issues. In some cases, these census topics may be ‘partly’ satisfied. A census topic with a ‘possible’ assessment usually has a mix of ‘excellent’, ‘good’, and ‘good to poor’ or ‘unknown’ quality measures.
- **Unlikely:** A census topic is ‘unlikely’ to be satisfied with administrative sources if the assessment of any quality measure for a census topic is poor, and there is little likelihood of improvement.

While attempts have been made to apply the quality measures in a consistent way across administrative data sources and topics, sometimes determinations have had a subjective element, particularly where an aspect of a data source was unknown.

These assessments are indicative only, and have used the information available at the time. They are not to be considered final verdicts on whether administrative sources could satisfy a particular census topic.

4 Results

This section presents the findings from this research.

Overall findings for census topics are presented below. For more detail about the quality measures and administrative data sources for each census topic, refer to Appendix 1. A full evaluation of individual census topics and their relevant administrative data sources is available on request from Statistics NZ.

Likelihood that administrative data sources satisfy census topics

One aim of this project was to identify the likelihood that each in-scope topic could be satisfied by administrative data sources. Overall, fewer than half of the census topics considered (16 of 39) were likely, or possibly able to be satisfied with administrative data sources. Specifically, of the 39 census topics considered:

- 9 are likely to be satisfied with administrative data sources
- 7 may possibly be satisfied with administrative data sources
- 23 are unlikely to be satisfied with administrative data sources.

Table 4 shows the assessment of how likely administrative data sources are to satisfy the census topics considered in this paper.

Table 4
Likelihood of a census topic being satisfied with administrative data sources

Subject category	Census priority	Census topic	Assessment
Population structure	Defining	Legally registered relationship status / Partnership status in current relationship	Unlikely
	Supplementary	Number of children born	Unlikely
Ethnicity and culture	Foremost	Ethnicity	Likely
	Defining	Māori descent	Likely
		Iwi affiliation	Possible
		Birthplace	Likely
	Supplementary	Language	Unlikely
		Religious affiliation	Unlikely
		Years since arrival in New Zealand	Likely
Education and training	Defining	Highest qualification	Unlikely
		Highest secondary school qualification	Possible
		Post-school qualification	Unlikely
		Study participation	Likely

Table continues below

Table 4 continued

Work	Defining	Status in employment	Possible
		Work and labour force status	Unlikely
		Hours worked in employment per week	Unlikely
	Supplementary	Industry	Likely
		Sector of ownership	Likely
		Workplace address	Possible
		Unpaid activities	Unlikely
		Occupation	Unlikely
		Main means of travel to work	Unlikely
Income	Defining	Sources of personal income	Likely
		Total personal income	Likely
Health and disability	Supplementary	Cigarette smoking	Unlikely
		Disability	Unlikely
Families and households	Defining	Family type	Unlikely
		Extended family type	Unlikely
		Household composition	Unlikely
Housing	Defining	Occupied dwelling type	Unlikely
		Weekly rent paid by households	Possible
		Sector of landlord	Possible
		Tenure of household	Unlikely
	Supplementary	Tenure holder	Unlikely
		Number of bedrooms	Unlikely
		Number of rooms	Unlikely
		Access to telecommunication systems	Unlikely
		Fuel types used to heat dwelling	Unlikely
		Number of motor vehicles	Possible

Source: Statistics New Zealand

Findings by subject category and census priority

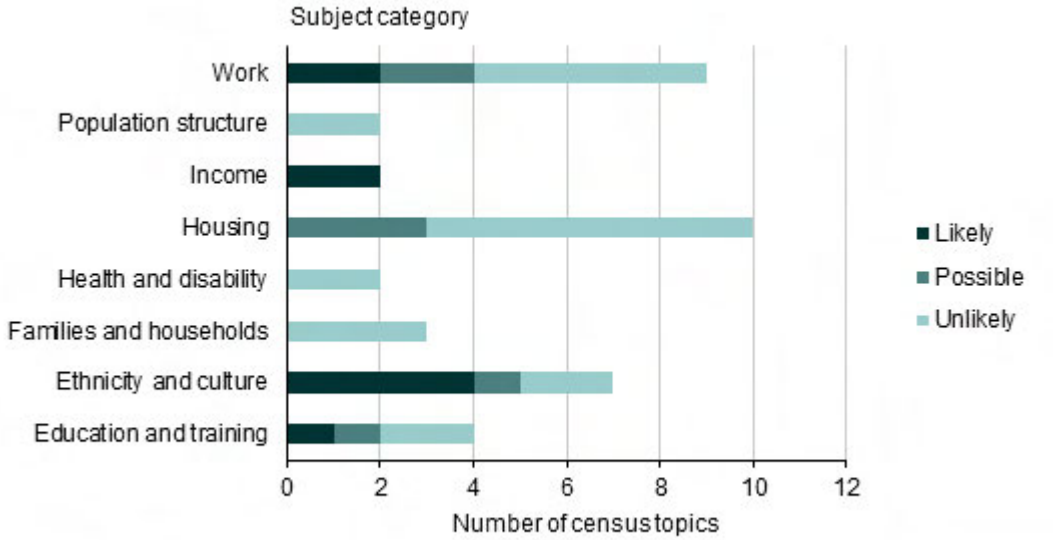
Another aim of this project was to identify the subject category and census priority areas that are more or less likely to be satisfied with administrative data sources.

Findings by subject category

Figure 2 presents the census topic findings by subject category. Figure 2 shows that:

- none of the housing, population structure, health and disability, or families and household census topics are likely to be satisfied with administrative data sources
- 5 out of 7 of the ethnicity and culture census topics are likely to be or may possibly be satisfied with administrative data sources
- 2 out of 4 education and training census topics are likely to be or may possibly be satisfied with administrative data sources.

Figure 2
Likelihood of a census topic being satisfied with administrative data sources, by subject category



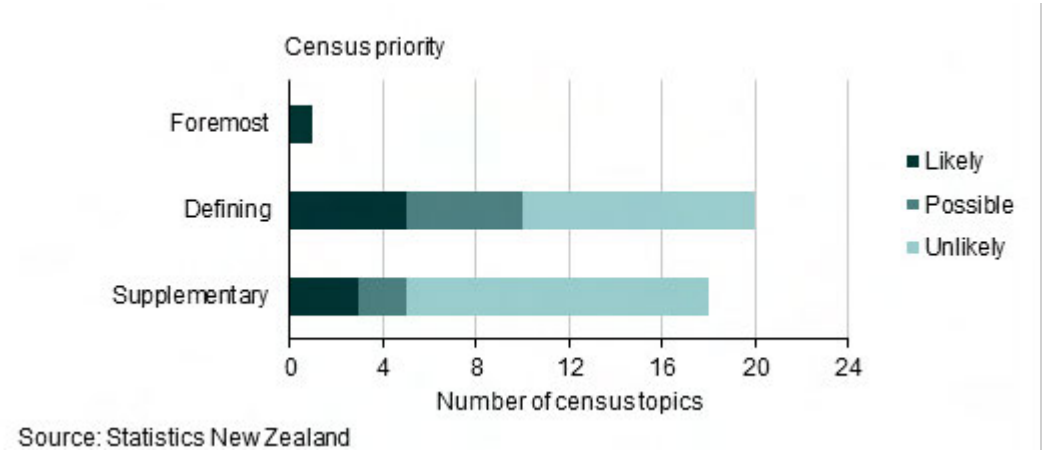
Source: Statistics New Zealand

Findings by census priority

Figure 3 presents the census topic findings by census priority level. Figure 3 shows:

- the only foremost variable considered (ethnicity) is likely to be satisfied with administrative data sources
- 5 of the 20 defining census topics are likely to be satisfied with administrative data sources, compared with 3 of the 18 supplementary census topics.
- When 'likely' and 'possible' categories are combined, 10 out of 20 defining census topics were likely to be or possibly able to be satisfied with administrative data sources, compared with 5 out of 18 supplementary census topics.

Figure 3
Likelihood of a census topic being satisfied with administrative data sources, by census priority



Source: Statistics New Zealand

For more information about the specific census topics and the quality assessments leading to their final scores, please refer to Appendix 1. The appendix provides the counts of census topics that are likely, possible, or unlikely to be satisfied with administrative data sources, which are the basis for figures 2 and 3. A full evaluation of

individual census topics and their relevant administrative data sources is available on request from Statistics NZ.

Key administrative data sources

Another aim of this project was to identify the administrative data sources that are most critical for informing census topics. This section highlights these data sources in two ways:

- administrative data sources that are the sole source for a particular census topic
- administrative data sources that many census topics are reliant upon.

Each situation has advantages and limitations. Relying on one administrative data source may mean fewer discrepancies in collection methods and definitions, but may mean greater vulnerability to loss of that data due to administrative or political changes. Having multiple administrative data sources may provide additional information to overcome shortcomings in one dataset, but may mean conflicting information needs to be resolved.

Administrative data sources that are the sole source for a particular census topic

Two administrative data sources have been identified as being the only source of information for particular census topics:

- Ministry of Education enrolments are the sole source of administrative data that might provide input into the language and study participation census topics.
- New Zealand Transport Authority motor vehicle registrations are the sole source of administrative data that might provide input into the number of motor vehicles census topic.

All other census topics that were assessed as ‘likely’ or ‘possible’ could be informed by multiple administrative data sources.

Administrative data sources that many census topics rely on

Some administrative data sources could inform many census topics. This section discusses some of the administrative data sources that might be most heavily relied on to produce social and economic census topics.

Vitals administrative data sources (especially birth and marriage registrations)

The Department of Internal Affairs collects this data, and Statistics NZ processes it for statistical outputs. Birth, death, and marriage registrations may be used to inform census topics such as legally registered relationship status, family type, extended family type, occupation, birthplace, number of children born, ethnicity, and Māori descent. Discussions may be needed about maintaining access to names, permission to code occupation, changes to questions about previous children, and other topics.

Arrival and departure cards

The Immigration Act 2009 requires all international travellers to fill out cards when they enter or exit New Zealand (except for air crew, transit passengers, military, and some other people generally not part of the census target population). The New Zealand Customs Service collects the information from these cards, and Statistics NZ processes it for international travel and migration statistics.

The traveller information on these cards could inform the following census topics: birthplace, years since arrival in New Zealand, and occupation.

Statistics NZ has traditionally used the information from these cards in an aggregate way, but not for unit-record linking. However, migration card data is now linked within the Integrated Data Infrastructure (IDI) at Statistics NZ. Customs also collects information

about border movements and passport details (held separately from arrival and departure card information), which overlaps in part with the arrivals and departure cards information.

Educational enrolment data

The Ministry of Education collects enrolments for all levels of formal education, and qualifications gained at the secondary and tertiary level. Statistics NZ already has some of this enrolment and qualifications information. The enrolment data is believed to have high coverage of younger ages.

Enrolment data may inform the following census topics: ethnicity, Māori descent, iwi affiliation, language, highest secondary school qualification, post-school qualification, and study participation.

Tax and income records

Inland Revenue collects detailed information about people's taxable income. Inland Revenue tax data has a high-quality unique identifier (IRD number) that provides strong assurance the data does not contain duplicate records for an individual.

Inland Revenue data may be used to inform census topics such as sources of income, total income, and status in employment. Statistics NZ receives this data regularly from Inland Revenue.

Tenancy bond database

The Building and Housing department, within the Ministry of Business, Innovation and Employment, maintains this database. Each record represents a bond that is lodged for a tenancy, and it contains information about the property (for example, number of bedrooms), the landlord, the tenants, and the terms of the tenancy (for example, weekly rent paid).

This administrative data source may inform the following census topics: weekly rent paid by households, tenure of household, tenure holder, and number of bedrooms.

5 Discussion

This paper has evaluated the potential for administrative data to satisfy social and economic topics provided by the census. The major aims of this project were to:

- identify the likelihood that each in-scope census topic could be satisfied by administrative data sources
- identify the subject category areas and census priority areas that are more or less likely to be satisfied by administrative data sources
- identify the administrative data sources that are most critical for informing census topics.

The major conclusion of this paper is that the potential to replace census questions appears fairly limited. The results showed that fewer than half of the 39 census topics in this study scored as 'likely' or 'possible' to be satisfied by administrative data sources.

Reasons that administrative data sources may not be suitable include that:

- information is only available for parts of the population and misses significant groups (for example, post-school qualifications are available for New Zealand graduates since 2004, but not for qualifications gained previously, or from overseas institutions)
- only some of the categories are available (for example, legally registered marriages and civil unions are recorded, but not de facto relationships)
- the information is not collected in administrative systems (for example, language spoken)
- there is not enough information to construct more-complicated derived variables (for example, household and family arrangements).

The quality measures used are 'necessary but not sufficient' conditions: a definite 'poor' on any one measure is a strong indicator of failure. Therefore, our confidence in the assessment of census topics scored as 'unlikely' is high, barring significant changes to the administrative data sources (or important information we have missed).

We cannot rule out 'possible' assessments at present, but their usefulness is uncertain. While those census topics scored as 'likely' appear to offer a good chance of successfully using administrative data sources to satisfy census information needs, further testing using actual data is needed, and other factors, particularly coherence with statistical standards and consistency over time, will need to be examined.

The finding – that administrative data sources are likely to contribute only a small proportion of current census topics – reinforces the importance of understanding the essential information requirements provided by a census for small areas and small population groups. Certain census topics will be more or less critical when it comes to determining the shape of future censuses. The priority of different census topics will be reviewed in consultation with users of census data. Highest priority topics will be a focus of more in-depth analysis to test these preliminary findings.

On a more positive note, results from this study have highlighted where administrative data sources could contribute to improvements in the current census model. An additional focus of the next stage of analysis will be investigations of the way we might use administrative data sources in the next census; for example, to improve quality where responses are missing. Examples include questions about participation in study, income, and income sources. Eventually, administrative data sources might remove the need to ask some questions in the census questionnaire.

This investigation found that a relatively small number of administrative data sources are relied on for many of the census topics assessed as likely or possible. All of these

administrative data sources are already available to Statistics NZ, meaning that we can undertake further analysis of these census topics with relative ease.

Limitations of this investigation

This investigation was intended to provide early indications of the ability for administrative data sources to replace topics asked in the census. As such, it has some clear limitations which were imposed to manage the volume of work required.

Only current 2013 Census topics have been considered and no attempt has been made to consider changes to information needs that might occur in future censuses. However, it can be fairly safely assumed that the broad subject categories of census information are enduring.

This paper also does not consider potential advantages of using administrative data sources (for example, greater frequency of census information, the ability to carry out longitudinal analysis, or data that is more accurate or more relevant to the statistical concept), which would need to be weighed up against a likely reduction in the range of census topics that a purely administrative census could support.

Another limitation is that the findings are based on metadata, and no analysis of the data itself has been undertaken. Scoring is somewhat subjective, and further analysis will be needed to confirm the findings presented here. However, the metadata gathered and summarised in a systematic way over a large number of census topics is a valuable resource for anyone interested in the administrative data sources available for these topics.

This paper provides early-stage assessments of the extent to which administrative data sources might satisfy census information needs. It provides a reference point for further work, and will help to inform decisions about where to direct more in-depth analysis. Future work will provide a more detailed evaluation of the potential for administrative data sources to produce census information.

References

Australian Bureau of Statistics (2009). ABS data quality framework. Available from www.abs.gov.au.

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

Eurostat, (2009). ESS handbook for quality reports: 2009 edition. Available from http://epp.eurostat.ec.europa.eu.

Eurostat (2012). European statistics code of practice – revised edition 2011. Available from http://epp.eurostat.ec.europa.eu.

Gibb, S, Shrosbree, E. *Evaluating the potential of linked data sources for population estimates: IDI as an example.* Available from www.stats.govt.nz (forthcoming).

Gill, L (2001). *Methods for automatic record matching and linkage and their use in national statistics.* National Statistics Methodology Series No. 25. London: Office for National Statistics.

Office for National Statistics (2013). Guidelines for measuring statistical quality. Available from www.ons.gov.uk.

Statistics Canada (2009). Statistics Canada quality guidelines: Fifth edition. Available from www.statcan.gc.ca.

Statistics New Zealand (2009). 2011 Census content report. Available from www.stats.govt.nz.

Statistics New Zealand (2012). Transforming the New Zealand Census of Population and Dwellings: Issues, options, and strategy. Available from www.stats.govt.nz.

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

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

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

United Nations Economic Commission for Europe (2007). Register-based statistics in the Nordic countries: Review of best practices with focus on population and social statistics. Available from www.unece.org.

Appendix 1: Detailed results

Table 5
Overall assessment of quality measures for specific census topics

Priority	Census topic	Assess-ment	Relevance	Accuracy: coverage	Accuracy: linkage	Timeliness	Accessibility
Population structure							
Defining	Legally registered relationship status / Partnership status in current relationship	Unlikely	Good	Poor	Good	Good	Poor
Supplemen-tary	Number of children born	Unlikely	Good	Poor	Good	Good	Poor
Ethnicity and culture							
Foremost	Ethnicity	Likely	Excellent	Good	Good	Good to unknown	Good
Defining	Māori descent	Likely	Excellent	Good	Good	Good	Poor
	Iwi affiliation	Possibly	Excellent	Poor to unknown	Good to unknown	Good	Poor
	Birthplace	Likely	Excellent	Excellent	Good	Excellent	Poor
Supplemen-tary	Language	Unlikely	Poor	Poor	Good	Good	Good
	Religious affiliation	Unlikely	Poor Good	Poor	Poor	Poor	Poor
	Years since arrival in NZ	Likely	Poor	Excellent	Good	Good	Good
Education and training							
Defining	Highest qualification	Unlikely	*	*	*	*	*
	Highest secondary school qualification	Possible	Excellent	Good to poor	Good	Good	Good
	Post-school qualification	Unlikely	Excellent	Poor	Good	Good	Good
	Study participation	Likely	Excellent	Good	Good	Good	Good
Work							
Defining	Status in employment	Possible	Good	Good	Good	Good	Good
	Work and labour force status	Unlikely	Poor	Poor	Good	Good	Good
	Hours worked in employment per week	Unlikely	Unknown	Poor	Good	Unknown	Unknown
Supplemen-tary	Industry	Likely	Good	Excellent	Excellent	Good	Excellent
	Workplace address	Possible	Good to poor	Good	Good	Good	Excellent
	Sector of ownership	Likely	Good	Excellent	Excellent	Good	Excellent
	Unpaid activities	Unlikely	…	…	…	…	…
	Occupation	Unlikely	Good	Good	Good	Poor	Poor
	Main means of travel to work	Unlikely	Poor	Poor	Good to poor	Good	Poor

Table continues below

Table 5 continued

Priority	Census topic	Assess-ment	Relevance	Accuracy: coverage	Accuracy: linkage	Timeliness	Accessibility
Income							
Defining	Sources of personal income	Likely	Good	Excellent	Excellent	Good	Good
	Total personal income	Likely	Excellent	Excellent	Excellent	Good	Good
Health and disability							
Supplemen-tary	Cigarette smoking	Unlikely	Good	Poor	Good	Poor	Good
	Disability	Unlikely	Poor	Poor	Good	Good	Poor
Families and households							
Defining	Family type	Unlikely	Poor	Poor	Good to poor	Poor	Good to poor
	Extended family type	Unlikely	Poor	Poor	Good	Poor	Good
	Household composition	Unlikely	Poor	Poor	Good to poor	Poor	Good
Housing							
Defining	Occupied dwelling type	Unlikely	Poor	Poor	Unknown	Poor	Good
	Weekly rent paid by households	Possible	Good	Good	Good to unknown	Good	Good
	Sector of landlord	Possible	Good	Good	Good to unknown	Unknown	Good
	Tenure of household	Unlikely	Poor	Poor to unknown	Good to unknown	Unknown	Good
Supplemen-tary	Tenure holder	Unlikely	Poor to unknown	Unknown	Unknown	Unknown	Good
	Number of bedrooms	Unlikely	Poor to unknown	Unknown	Good to unknown	Unknown	Good
	Number of rooms	Unlikely	…	…	…	…	…
	Access to telecommunication systems	Unlikely	Poor	Good to poor	Good to poor	Good to poor	Good to poor
	Fuel types used to heat dwelling	Unlikely	Poor	Poor	Good to poor	Good to poor	Poor
	Number of motor vehicles	Possible	Good	Good	Good	Good	Poor

Symbols:
* Highest educational qualification is derived from other variables (highest secondary school qualification and post-school qualification); therefore, a separate quality assessment was not undertaken.

… not applicable

Source: Statistics New Zealand

Table 6
Likelihood of census topics being satisfied with administrative data sources
By subject category, and census priority

All topics Number of census topics				
	Likely	**Possible**	**Unlikely**	**Total**
Total	9	7	23	39

By subject category Number of census topics				
	Likely	**Possible**	**Unlikely**	**Total**
Population structure	0	0	2	2
Ethnicity and culture	4	1	2	7
Education and training	1	1	2	4
Work	2	2	5	9
Income	2	0	0	2
Health and disability	0	0	2	2
Families and households	0	0	3	3
Housing	0	3	7	10

By census priority Number of census topics				
	Likely	**Possible**	**Unlikely**	**Total**
Foremost	1	0	0	1
Defining	5	5	10	20
Supplementary	3	2	13	18

Source: Statistics New Zealand